A BAYESIAN SUPERVISORY APPROACH OF OUTLIER DETECTION FOR RECURSIVE SOFT SENSOR UPDATE

Hector J. Galicia[a], Q. Peter He[b,*] and Jin Wang[a,*]
[a] Department of Chemical Engineering, Auburn University, Auburn, AL 36849
[b] Department of Chemical Engineering, Tuskegee University, Tuskegee, AL 36088

Abstract

Partial least squares (PLS) based soft sensors that predict the primary variables of a process by using the secondary measurements have drawn increased research interests recently. Such data-driven soft sensors are easy to develop and only require a good historical data set. As industrial processes often experience time-varying changes, it is desirable to update the soft sensor model with the new process data once the soft sensor is implemented online. Because the PLS algorithms are sensitive to outliers in the dataset, outlier detection and handling plays a critical role in the development of the PLS based soft sensors. In this work, we develop multivariate approaches for both off-line and online outlier detection. For online application, to differentiate outliers caused by erroneous readings from those caused by process changes, we propose a Bayesian supervisory approach to further analyze and classify the detected outliers. Both simulated and industrial case studies of the Kamyr digesters are used to demonstrate the effectiveness of the proposed approaches.

Keywords

Outlier detection and classification, Bayesian statistics, partial least squares, soft sensor, recursive update.

Introduction

In many industrial processes such as distillation columns and pulping digesters, the primary product variables that are required for feedback control are either not measured online or not measured frequently. To address this challenge, many data-driven soft sensors have been developed and implemented in process industry (see comprehensive reviews by Kadlec et al. 2009, Fortuna et al. 2010 and references cited therein). Various adaptation techniques have been published to update data-driven soft sensors online, and Kadlec et al. (2011) provide a comprehensive review on the adaptation mechanisms for data-driven soft sensors.

In our previous work (Galicia et al., 2011a), a reduced-order dynamic PLS (RO-DPLS) soft sensor was developed to address some limitations of the traditional DPLS soft sensor when applied to processes with large transport delays. By taking the process characteristics into account, RO-DPLS soft sensor can significantly reduce the number of regressor variables and improve prediction performance. More recently we extended the RO-DPLS soft sensor to its online adaptation version in order to track process changes (Galicia et al., 2011b). Since our focus in Galicia et al. (2011b) was to investigate the properties of different recursive updating schemes and data scaling

* Corresponding authors
Email addresses: hjg0002@tigermail.auburn.edu (Hector J. Galicia), qhe@tuskegee.edu (Q. Peter He), wang@auburn.edu (Jin Wang)

methods, we preprocessed the industrial datasets to remove all outliers before subjecting them to different experiments.

However, it should be noted that the PLS algorithms are sensitive to outliers in the dataset (Hubert and Branden, 2003). Therefore, outlier detection and handling plays a critical role in the development of the PLS based soft sensors, and there exist extensive studies on outlier detection for off-line model building (Hodge and Austin, 2004). Typical approaches are based on statistics of historical data such as the 3σ outlier detection method (Pearson, 2002) and the more robust Hampel identifier method (Davies and Gather, 1993). In addition, multivariate outlier detection methods have also been proposed, such as the PCA-based Jolliffe parameter (Jolliffe, 2002). Although many methods have been published, outlier detection remains a challenging problem. Therefore, it is often recommended to accompany any outlier detection method with a graphical inspection of the residual space and model parameters to eliminate any possible outlier masking effect (i.e., outliers are classified as consistent samples) and outlier swamping effect (i.e., consistent samples are classified as outliers) (Martens and Naes, 2002). For online adaptation of soft sensor models, if erroneous readings are used to update the soft sensor model, future predictions from the updated model may deteriorate significantly. In addition, online outlier detection is even more challenging since while outliers could be erroneous readings, they could also be normal samples of new process states.

It is worth mentioning that an alternative approach to reduce the effect of outliers on soft sensor performance is to use the robust versions of PLS algorithms (Wakelinc and Macfie, 1992, Cummins and Andrews, 1995, Gil and Romera, 1998, Pell, 2000, Hubert and Branden, 2003). However, the robust versions of PLS algorithms usually involve much more complicated computation, and may have various limitations, such as applicable to one-dimensional response variable, not resistant to leverage point, and not applicable to high dimensional regressors. More importantly, unlike the conventional PLS algorithm which has various recursive versions available for online updating, currently there are no recursive versions available for the robust PLS algorithms. Therefore, they are not desirable for online model updating, and in this work we only consider the conventional PLS algorithms.

In this work, we develop multivariate monitoring methods for both off-line and online outlier detection. In addition, to differentiate the samples that represent a process change from those of erroneous readings, we propose a Bayesian supervisory approach to further analyze and classify the detected outliers.

Outlier detection and classification methods

In this section, the proposed outlier detection and classification methods are described in detail.

Off-line outlier detection for initial model building

We propose to combine leverage and y-studentized residuals for off-line outlier detection. Leverage of an observation is a concept developed in the ordinary regression theory (Cook and Weisberd, 1982), which defines the influence that a given sample will have on a model and is related to the Hotelling's T^2 statistic (Martens and Naes, 2002). The leverage in terms of the **T** scores (Walczak and Massart, 1995) is computed as follows:

$$h_i = \frac{1}{n} + \sum_{j=1}^{a} \frac{t_{i,j}^2}{t_j^T t_j} \quad (1)$$

where n is the total number of samples, $\mathbf{t}_{i,a}$ is the vector of scores for object i, and a is the number of principal components retained. For a given sample, it would be classified as an outlier if its leverage $h_i > \beta(1+a)/n$, where β is a constant (usually 2 or 3). In this work we choose β = 3 as our experience shows that this setting provides a balanced performance between desired sensitivity and specificity.

For a given sample, the studentized residual is an indication of the lack of fit of the y-value, which is defined as the following (Walczak and Massart, 1995)

$$r_i = \frac{f_i}{s(1-h_i)^{1/2}} \quad (2)$$

where $f_i = y_i - \hat{y}_i$ is the residual of the dependent variable, $\hat{y}_i$ is the i^{th} prediction of the dependent variable provided by the soft sensor model, and $s^2 = \frac{1}{n-a-1}\sum_{i=1}^{n} f_i^2$ is the sample estimate variance of the residual. To detect outliers, the studentized residuals are usually compared to a normal distribution $N(0,1)$ (Martens and Naes, 2002). In this work, if $|f_i| > 3$, we classify the sample as an outlier.

Online outlier detection for recursive model update

For online outlier detection, we use the SPE_x and SPE_y (squared prediction error for **X** and **Y**) indices (MacGregor and Kourti, 1995, Qin, 2003) to monitor the independent variable and dependent variable space, respectively. It is worth noting that we do not use Hotelling's T^2 index to detect outliers, although it is commonly used in conjunction with SPE index for fault detection. Our choice of SPE over T^2 is mainly due to their different roles in process monitoring. It has been suggested that for the samples whose T^2 indices exceed the threshold but not the SPE indices, in many cases they correspond to process operation changes instead of outliers (Qin, 2003). Our own experiences also confirm that when a real outlier occurs, it is usually identified by both the T^2 index and SPE index. Therefore, in this work, we use the SPE index alone to detect outliers in the independent and dependent variable

space. Specifically, if the SPE index (i.e., SPE_x or SPE_y) violates its corresponding control limit we declare that the sample is an outlier and should be analyzed further. SPE_x and SPE_y for a new sample to be used for soft sensor update are defined as:

$$\mathrm{SPE}_{\mathrm{x}} = \sum_{i=1}^{m} \left(\mathbf{x}_{new,i} - \hat{\mathbf{x}}_{new,i} \right)^2 \tag{3}$$

$$\mathrm{SPE}_{\mathrm{y}} = \sum_{i=1}^{p} \left(\mathbf{y}_{new,i} - \hat{\mathbf{y}}_{new,i} \right)^2 \tag{4}$$

where m is the number of independent variables, p the number of dependent variables, $\mathbf{x}_{new,i}$ and $\mathbf{y}_{new,i}$ are the new samples of independent and dependent variables, and $\hat{\mathbf{x}}_{new,i}$ and $\hat{\mathbf{y}}_{new,i}$ are the corresponding predictions. The thresholds for SPE_x and SPE_y can be determined based on the theorems developed by Box (1954) using the training data, or they can be determined empirically using the training or validating data under normal operating conditions (Wise et al., 1999, Russell, 2000).

Bayesian supervisory approach for online outlier classification

For online outlier detection, once a new sample is identified as an outlier, we have to determine whether it corresponds to an erroneous reading, or it represents a new process state. In this work, we propose a Bayesian supervisory approach to perform this task. It should be noted that this task is less difficult for off-line data, because the data collected after the outlier(s) are available to help make the decision.

In the proposed Bayesian supervisory approach, once an outlier is detected by the SPE indices, we wait for few more measurements to become available and apply Bayesian inference to make the decision. The basic assumption is that if an outlier is due to erroneous readings, the increase in the SPE indices will not be sustainable and will result in an impulse or short step disturbance in the time series of SPE indices. On the other hand, if an outlier is caused by a process change, the following samples will all deviate from the previous model and will result in a sustained step or ramp disturbance in the time series of SPE indices. Therefore, when an outlier is identified, we try to classify whether the change in the SPE index belongs to an impulse/short step or a ramp/step disturbance in order to determine whether the outlier is due to an erroneous reading or a process change. In this work, the classification is achieved through a Bayesian approach.

By definition, Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities. Simply put, it gives the probability of a random event *A* occurring given that we know a related event *B* occurred. This probability is denoted as P(A|B) and is called the posterior probability, since it is computed after all other information on *A* and *B* is known. Using Bayes' Theorem, the posterior probability can be computed as

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B})} \tag{5}$$

In our case, event *B* corresponds to the measurements collected after the outlier is detected and event *A* corresponds to a specific disturbance type.

In the proposed approach, instead of using the values of SPE indices directly which are often affected by the stochastic nature of the process, we transform the index values into a more robust statistic-based probability description using the Bayesian statistics. In this way, different processes with different dynamic characteristics can be analyzed using a unified statistical framework. Figure 1 shows the schematic diagram of the proposed Bayesian outlier classification procedure, which is a modification of the previously developed Bayesian approach for detection and classification of different disturbances in the semiconductor processes (Wang and He, 2007). The classification algorithm is triggered by the identification of an outlier through SPE indices, and a brief description is provided below.

1. Denote the time index of the outlier as *k*; construct the pre- and post-change windows around the outlier *k*. The pre-change window contains a few samples' SPE indices prior to the identified outlier; while the post-change window contains the SPE indices after (and including) the outlier. In this work, the width of the pre-change window is fixed to *5* samples, while the width of the post-change window varies depending on the assumed type of disturbance.
2. Wait until sample *k+1* is available, then perform hypothesis testing using Bayes' Theorem to determine whether the disturbance is an impulse.
3. If the hypothesis is rejected, we wait for more future samples to determine whether the sample is part of a short step disturbance (with duration 2, 3 or 4).
4. If all previous hypotheses are rejected, we conclude that a real process change has occurred.

It is worth noting that the posterior probabilities in the post-change window form different patterns depending on the pre-assumed disturbance type and observation values in the post-change window. Instead of putting a threshold on a single posterior probability, which is a univariate method (Hu, 1992), the proposed Bayesian approach compares the pattern of the posterior probabilities in the post-change window to the predefined patterns in order to classify the type of disturbance. This pattern matching approach is a multivariate method, which is more robust compared to the univariate method, and greatly improves classification accuracy and reduces classification delay. In addition, it makes the classification results not sensitive to the a priori probability (in this work, we set a priori probability to 0.5). Detailed description of different patterns of posterior probability in the post-change window and pattern matching procedure can be found in Wang and He (2007).

In addition, by specifying different post-change window widths for different disturbances, we can minimize the classification delay. In other words, the classification decision is made when the minimum required information becomes available.

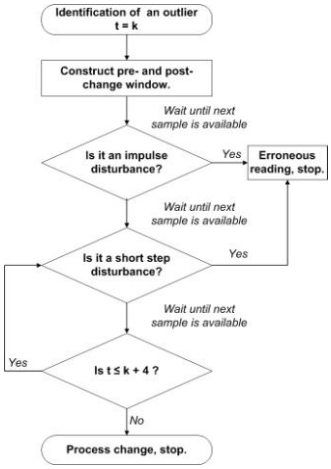


Figure 1: Bayesian disturbance classification procedure

Soft sensor recursive update with outlier detection and classification

In this section, the proposed outlier detection and classification methods are integrated into the previously developed RO-DPLS soft sensor (Galicia et al., 2011a, 2011b) for online recursive update. Based on the results of a comprehensive comparison of different recursive PLS update schemes (Galicia et al., 2011b), regular recursive PLS updating scheme is chosen to update the RO-DPLS soft sensor model. Both simulated case study and industrial case studies of the single vessel Kamyr digester are used to demonstrate the performance of the proposed approach. In this section, the performance of the RO-DPLS soft sensor is compared for four scenarios.

a) Static model without update;
b) Recursive update without outlier detection, i.e., all samples are used to update the model;
c) Recursive update with outlier detection, i.e., all outliers identified by SPE indices are excluded from model update;
d) Recursive update with Bayesian supervisory approach for outlier detection, i.e., erroneous readings are excluded from model update, while process changes are used for model update;

Simulated case study

In this subsection, the simulated Kamyr digester is used to illustrate how the Bayesian supervisory approach works. The extended Purdue model (Wisnewski et al., 1997) is implemented to simulate a single vessel high yield Kamyr digester. The RO-DPLS soft sensor set up can be found in Galicia et al. (2011a). In this case study, we consider a very challenging problem: tracking the disturbance of a wood type change. It is worth noting that wood type change (from softwood to hardwood and vice versa) is a major disturbance in pulping processes, and usually results in off-spec product during the transition.

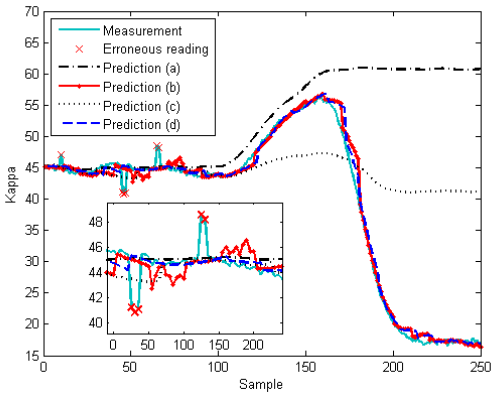
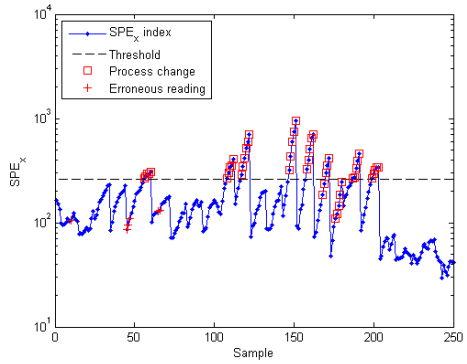


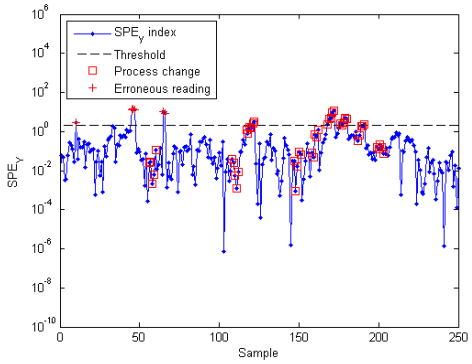
Figure 2: Prediction comparison of different approaches applied to a simulated case study

Both single and consecutive multiple outliers (i.e. impulses and short steps) are added to the process, and the Kappa number measurements are plotted in Figure 2. From Figure 2 we can see that the dramatic change in Kappa number during the transition period is due to the wood type change, and the samples during the transition should be used to update the model; while the changes that occur at samples 10 (impulse), 45 (short step with duration 3) and 65 (short step with duration 2), are introduced outliers and should not be used for model update. The soft sensor predicted Kappa number values are also plotted in Figure 2, with the corresponding mean squared prediction error (MSE) and mean prediction error (ME) given in Table 1. Both Figure 2 and Table 1 demonstrate the important roles of outlier detection and classification, and their impact on soft sensor performance. It is shown that outlier detection alone may even deteriorate the performance of a soft sensor if process changes were treated the same as erroneous measurements. On the other hand, if the proposed Bayesian outlier classification mechanism is integrated into outlier detection, the soft sensor can be made more robust to erroneous measurements and at the same time is able to track process changes. Figure 3 (a) and (b) show the SPE_x and SPE_y indices for the Bayesian supervisory approach, together with the classified outliers. Figure 3 shows that SPE_x and SPE_y indices can promptly identify the outliers caused by both erroneous reading and process change, and the Bayesian supervisory approach is effective in classifying the identified outliers. Without the Bayesian supervisory approach, all identified outliers will be excluded from updating the model, which results in poor prediction performance of approach (c), similar to the static model. For approach (b) where all new samples are used for model update, the negative impact of using

erroneous readings for model update are illustrated more clearly with the insert in Figure 2.



(a)



(b)

Figure 3: SPE indices for the simulated case study; (a) SPE_x; (b) SPE_y

Table 1: Performance of different soft sensors for the simulated case study

Soft Sensor	MSE	ME
(a)	405.4800	-10.8817
(b)	1.1476	-0.1695
(c)	120.6689	-3.7614
(d)	0.8505	-0.1526

Industrial case study

In the industrial case study, the process data were collected from a Kamyr digester at a pulp mill located in Mahrt, Alabama run by MeadWestvaco Corp. The training data were collected in 2006 which contain 1100 samples, while the testing data for online update were collected in 2010 which contain 300 samples. Clearly, this case study presents a challenging problem as training and testing data sets were collected about 4 years apart. The soft sensor setup for the industrial case is the same as that reported in Galicia et al. (2011a).

Figure 4 plots a segment of the testing data to illustrate the prediction performance of different soft sensors and compare them with the process measurements, and Table 2

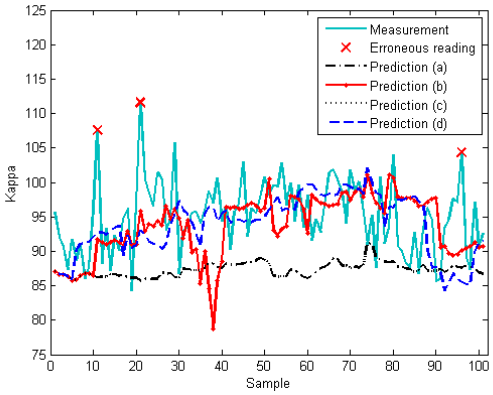
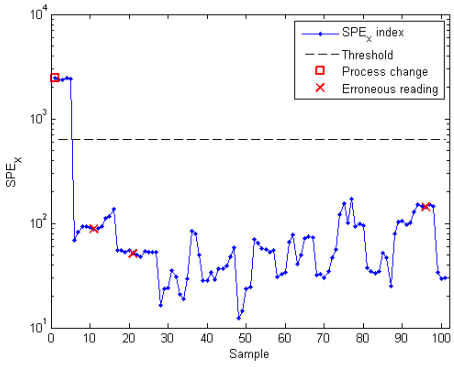
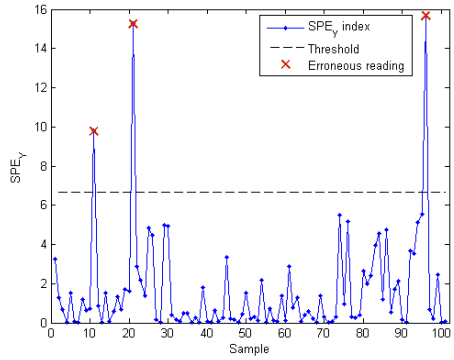


Figure 4: Comparison of predictions of different approaches for the industrial case study



(a)



(b)

Figure 5: SPE indices for the industrial case study; (a) SPE_x; (b) SPE_y

lists the MSE and ME of different approaches for the whole testing data set. Figure 5 (a) and (b) plot the SPE_x and SPE_y indices for the Bayesian supervisory approach, together with the classified outliers. It should be noted that for this case study the soft sensor that updates recursively with outlier detection (but without classification) performs exactly the same as the static soft sensor. This is due to the big difference between the training data and testing data, which causes all new data to be classified as outliers. From this case study, it is clear that the Bayesian supervisory approach is effective and robust in determining whether an outlier is caused by erroneous reading or process change.

Conclusions

Outlier detection and handling plays a critical role in data-driven soft sensor development. In this work, we propose multivariate approaches for both off-line outlier detection (for initial soft sensor model building) and online outlier detection (for soft sensor model recursive update). Specifically, for off-line outlier detection we combine leverage and y-studentized residuals; while for online outlier detection, we use squared prediction error indices for **X** and **Y** to monitor the independent variable and dependent variable space, respectively. For online outlier detection, to differentiate the outliers caused by erroneous reading from those caused by process changes, we propose a Bayesian supervisory approach to further analyze and classify the identified outliers. Both simulated and industrial case studies demonstrate the superior performance of the soft sensor with Bayesian supervisory approach for outlier detection and classification.

Table 2: Performance of different soft sensors for the industrial case study

Soft Sensor	MSE	ME
(a)	75.3031	7.3024
(b)	34.8632	1.0535
(c)	75.3031	7.3024
(d)	29.4228	0.7468

Acknowledgments

The authors gratefully acknowledge the financial support from NSF (QPH under Grant CBET-0853748, HJG and JW under grant CBET-0853983). HJG would also like to thank the Alabama Center for Paper and Bioresource Engineering (AC-PABE) for financial support. Finally, the authors thank Dr. Russell Hodges at R.E. Hodges, LLC and Mr. Charles Hodge at MeadWestvaco Corporation for providing the data and digester process knowledge.

References

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. *The Annals of Mathematical Statistics, 25(2): 290-302*.

Cook, R. D. and Weisberd, S. (1982). Residuals and influence in regression. *Chapman and Hall*, London.

Cummins, D. J. and Andrews, C. W. (1995). Iteratively reweighted partial least squares: A performance analysis by monte carlo simulation. *Journal of Chemometrics 9(6): 489-507*.

Davies, L. and Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association 88(423): 782-792*.

Fortuna, L., Graziani, S., Rizzo, A., Xibilia, M. (2010). Soft Sensors for Monitoring and Control of Industrial Processes, *Springer*, London.

Galicia, H. J., He, Q. P., Wang, J. (2011a). A reduced order soft sensor approach and its application to a continuous digester. *Journal of Process Control 21(4): 489-500*.

Galicia, H. J., He, Q. P., Wang, J. (2011b). Comparison of the performance of a reduced-order dynamic PLS soft sensor with different updating schemes for digester control. submitted to *Control Engineering Practice*.

Gil, J. A. and Romera, R. (1998). On robust partial least squares (PLS) methods. *Journal of Chemometrics 12(6): 365-378*.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review 22(2): 85-126*.

Hu, A. (1992). An optimal Bayesian process control for flexible manufacturing processes. *Massachusetts Institute of Technology*, MA.

Hubert, M. and Branden, K. V. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics 17(10): 537-549*.

Jolliffe, I. T. (2002). Principal Component Analysis, *Springer*.

Kadlec, P., Gabrys, B., Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers and Chemical Engineering 33(4): 795-814*.

Kadlec, P., Gabrys, B., Strandt, S. (2011). Review of adaptation mechanisms for data-driven soft sensors. *Computers and Chemical Engineering 35(1): 1-24*.

MacGregor, J. F. and Kourti, T. (1995). Statistical process control of multivariate processes. *Control Engineering Practice 3(3): 403-414*.

Martens, H. and Naes, T. (2002). Multivariate Calibration, *John Wiley and Sons Ltd*.

Pearson, R. K. (2002). Outliers in process modeling and identification. *Control Systems Technology, IEEE Transactions on 10(1): 55-63*.

Pell, R. J. (2000). Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems 52(1): 87-104*.

Qin, S. J. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics 17(8-9): 480--502*.

Russell, E. L., Chiang L. H., Braatz, R. D. (2000). Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems 51(1): 81-93*.

Wakelinc, I. N. and Macfie, H. J. H. (1992). A robust PLS procedure. *Journal of Chemometrics 6(4): 189-198*.

Walczak, B. and Massart, D.L. (1995). Robust principal components regression as a detection tool for outliers. *Chemometrics and Intelligent Laboratory Systems 27(1): 41-54*.

Wang, J. and He, Q. P. (2007). A Bayesian Approach for Disturbance Detection and Classification and Its Application to State Estimation in Run-to-Run Control. *IEEE Transactions on semiconductor manufacturing 20(2): 126-136*.

Wise, B. M., Gallagher, N.B., Butler, S., White, D., Barna, G.G. (1999). A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. *Journal of Chemometrics 13(3-4): 379-396*.

Wisnewski, P. A., Doyle, F., Kayihan, F. (1997). Fundamental continuous-pulp-digester model for simulation and control. *AIChE Journal 43(12): 3175-3192*.