Preprints of the 8th IFAC Symposium on Advanced Control of Chemical Processes
The International Federation of Automatic Control
Furama Riverfront, Singapore, July 10-13, 2012

Adaptive Outlier Detection and Classification for Online Soft Sensor Update

Hector J. Galicia* Q. Peter He,[1]. Jin Wang*,[1]**

**Department of Chemical Engineering, Auburn University, Auburn, AL 36849 USA*
(Tel: 334-844-2020; e-mail: hjg0002, wang@auburn.edu).
*** Department of Chemical Engineering, Tuskegee University, Tuskegee, AL 36088 USA,*
(qhe@mytu.tuskegee.edu)
[1] correspondance authors

Abstract: Data-driven soft sensors that predict the primary variables of a process by using the secondary measurements have drawn increased research interests recently. They are easy to develop and only require a good historical data set. Among them, the partial least squares (PLS) based soft sensor is the most commonly used approach for industrial applications. 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 a multivariate approach for online outlier detection. In addition, to differentiate outliers caused by erroneous readings from those caused by process changes, we propose a Bayesian supervisory approach to analyze and classify the detected outliers. Finally, to address time-varying nature of industrial processes, we proposed a simple yet effective scheme to update the detection threshold. Both simulated and industrial case studies of the Kamyr digesters are used to demonstrate the effectiveness of the proposed approaches.

1. 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). To address frequent changes in industrial processes, 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 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. Both univariate methods (Pearson, 2002, Davies and Gather, 1993) and multivariate methods (Jolliffe, 2002) have been developed, and a general review of the outlier detection problem and outlier detection algorithms was provided by Hodge and Austin (2004). Despite existing research, outlier detection remains a challenging problem. For online adaptation of soft sensor models, outlier detection is even more challenging because outliers could either be erroneous readings, or normal samples of new process states.

In this work, we propose a multivariate method for online outlier detection. In addition, we propose a novel Bayesian approach to differentiate the outliers that represent a process change from those of erroneous readings. To address the time-varying nature of industrial processes, we further propose a robust scheme to update the detection threshold. The effectiveness of the proposed outlier detection and classification methods is demonstrated using both simulated and industrial case studies.

2. ONLINE OUTLIER DETECTION FOR RECURSIVE MODEL UPDATE

Outliers are sensor values which deviate from the typical, sometimes also meaningful, ranges of the measured values (Kadlec et al., 2009). For online outlier detection, we use the SPE_x and SPE_y (squared prediction error for $\mathbf{X}$ and $\mathbf{Y}$) indices that are generated from an PCA model to monitor the independent variable and dependent variable space, respectively. Specifically, if a SPE index (i.e., either SPE_x or SPE_y) violates its corresponding control limit we declare that the sample is an outlier and should be analyzed further. For a new sample, its SPE_x and SPE_y indices are calculated as

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

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

where m is the number of independent variables, p the number of dependent variables, $\mathbf{x}_{new,i}$ and $\mathbf{y}_{new,i}$ the new samples of independent and dependent variables, and $\hat{\mathbf{x}}_{new,i}$ and $\hat{\mathbf{y}}_{new,i}$ the corresponding soft sensor 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).

3. BAYESIAN SUPERVISORY APPROACH FOR ONLINE OUTLIER CLASSIFICATION

Once a new sample is detected as an outlier, it is desirable 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 online outlier classification is much more challenging than off-line outlier classification. Because for off-line outlier classification, plenty of data collected after the outlier(s) are available to help make the decision. While for online outlier classification, decision needs to be made as soon as possible – significant delay will be resulted if decision can only be made after many more samples become available once the outlier(s) is(are) detected.

In the proposed Bayesian supervisory approach, 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, and only one or a couple more measurements after the outlier(s) are needed to perform the classification. In the proposed approach, instead of using the values of SPE indices directly which is 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.

By definition, Bayes' Theorem is a simple mathematical formula used to calculate 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

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B})} \quad (3)$$

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

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 for 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 detected 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 from 2 to 5 samples 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.

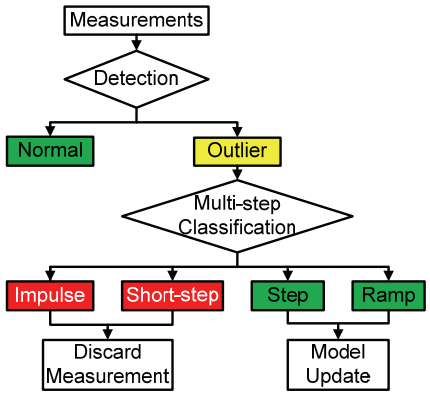


Figure 1 Schematic diagram of the Bayesian-statistics based disturbance detection and classification

Depending on the pre-assumed disturbance type, the posterior probabilities in the post-change window form different patterns. In the proposed Bayesian approach, the pattern of the posterior probabilities in the post-change window is compared to the predefined patterns in order to classify the type of the 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. 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 as listed in **Error! Not a valid bookmark self-reference.**, the classification decision is made when the minimum required information becomes available, which minimize the classification delay.

Table 1 Post-change window settings

Disturbance type	Post-change window width
Impulse	2
Short step of duration 2	3
Short step of duration 3	4
Short step of duration 4	5
Step	5
Ramp	5

4. OUTLIER DETECTION WITH THRESHOLD UPDATING

It is worth noting that after a major process change occurs, such as a wood type change in a pulp digester, the process may settle to a completely new state, and the normal SPE indices may switch to a different level. Therefore, it is necessary to update the thresholds of both SPE indices. Otherwise, the performance of outlier detection will deteriorate. In this section, we propose a robust way to update the thresholds using an exponentially weighted moving average (EWMA) filter as shown below.

$$\text{Th\_new} = \lambda \cdot \text{Th\_old} + (1-\lambda) \cdot \text{Th\_current} \quad (4)$$

Th_old denotes the previous threshold for outlier detection before the update; Th_new the threshold after the update; Th_current the threshold estimated using the reconstructed SPE indices of new measurements. The initial thresholds are determined using historical data under normal operation condition. λ is a tuning parameter which controls how fast the thresholds are updated.

For normal process operations, i.e., no outliers are detected, the thresholds are updated every 20 samples, and a relatively conservation setting is used ($0.9 < \lambda < 0.7$); for detected process changes, i.e., detected outliers are classified as caused by a process change, a more conservative setting is used ($0.95 < \lambda < 0.99$). It is also worth noting that usually different settings are used for SPE_x and SPE_y threshold update, due to different levels of variabilities in the independent and dependent variable spaces. Due to the limited number of samples (20 for normal process operation and even less for process changes), Th_current is estimated through an empirical way as shown below

$$\text{Th\_current} = \text{mean (SPE)} + a \cdot \text{std (SPE)} \quad (5)$$

where $a = 2 \sim 3$ is a tuning parameter.

5. SIMULATED AND INDUSTRIAL CASE STUDIES

In this section, both simulated and industrial case studies of a single-vessel Kamyr digester are used to demonstrate the effectiveness of the proposed outlier detection and classification methods. In addition, the case studies are used as benchmarks to compare four different soft sensor update schemes:

a) Recursive update without outlier detection, i.e., all samples are used to update the model;

b) Recursive update with outlier detection only, i.e., all outliers identified by SPE indices are excluded from model update;

c) Recursive update with outlier detection and classification, i.e., erroneous readings are excluded from model update, while process changes are used for model update;

d) Recursive update with adaptive outlier detection and classification, i.e., thresholds for outlier detection are updated recursively.

We first develop an initial soft sensor model to predict the product quality variable, i.e., the Kappa number, using historical data. Next, we compare the four update schemes when new data become available. The recursive update is implemented using the regular recursive PLS (RPLS) updating algorithm (Galicia, et al., 2011b). The soft sensor is updated every 10 new measurements. But for the case of c) and d), if a process change is detected, the soft sensor model is updated immediately after the process change is confirmed. In this case, the model is updated with the new measurements in the post-change window only, i.e., 5 samples.

4.1 Simulated Case Study

In this subsection, 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. Both single and consecutive multiple outliers (i.e. impulses and short steps) are manually added to the measured Kappa number as shown in Figure 2.

It should be noted that the dramatic change in Kappa number during the transition period (sample 100 to 230 shown in Figure 2) is due to the wood type change, and the samples during the transition should be used for model update; while the outliers that occur before and after the transition period are introduced erroneous readings and should not be used for model update. Figure 2 compares the soft sensor predicted Kappa number with measured Kappa number for four model update schemes. In addition, the performance of the different model update schemes are evaluated quantitatively using three indices, which are defined below:

Mean squared prediction error (MSE): $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

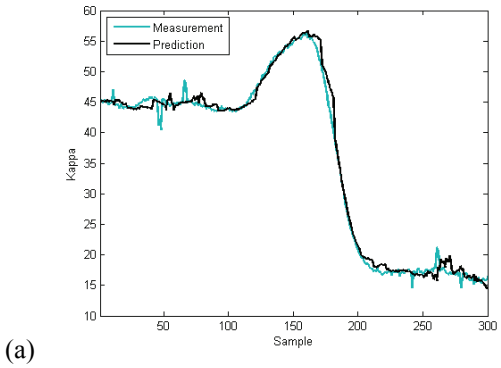
Mean prediction error (ME): $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$

Mean absolute percentage error(MAPE): $\frac{100}{N} \times \sum_{i=1}^{N}\left|\left(\frac{y_i - \hat{y}_i}{yi}\right)\right|$

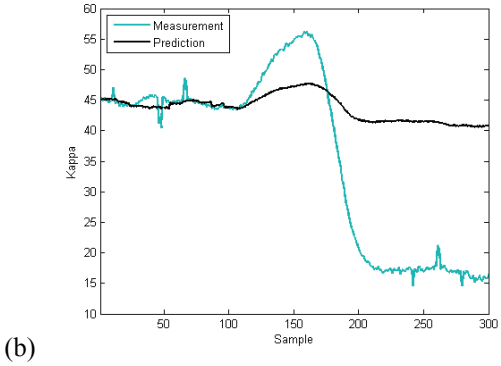
The performance indices for different model update schemes are listed in Table 2. Both Figure 2 and Table 2 demonstrate the important role of outlier detection and classification, and their impact on soft sensor performance. Figure 2 (b) shows that outlier detection alone may even deteriorate the performance of a soft sensor if process changes were treated the same way as erroneous measurements. On the other hand, if the proposed outlier classification mechanism is integrated into outlier detection, the soft sensor can be made more robust to erroneous measurements and at the same time be able to track process changes. The prediction performance of the soft sensor is further improved when the detection thresholds are updated recursively to reflect the process changes. Figure 3 plots the SPE_x indices together with their thresholds and identified outliers for update schemes c) and d). Figure 3 shows that with adaptive update, the threshold of SPE_x index track the process changes well and provide a reasonable boundary for outlier detection.

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

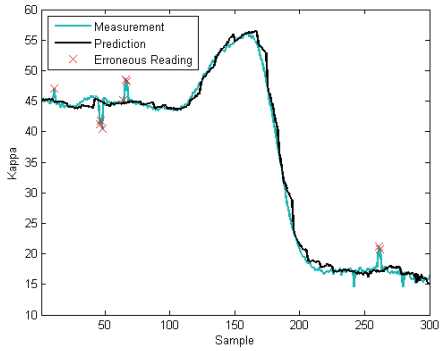
Soft Sensor	MSE	ME	MAPE(%)
(a)	1.22	-0.30	2.65
(b)	221.84	-8.01	54.72
(c)	0.99	-0.27	2.45
(d)	0.94	-0.28	2.34



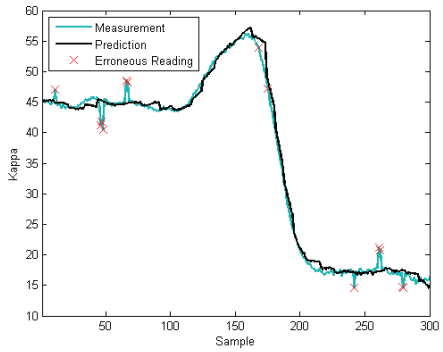
(a)



(b)

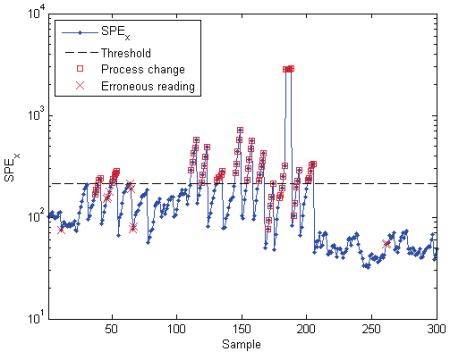


(c)

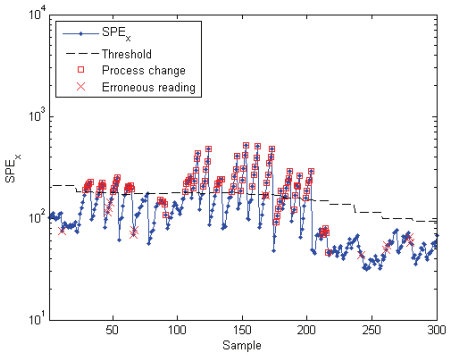


(d)

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



(a)



(b)

Figure 3: SPE_x indices for the simulated case study; (a) fixed threshold; (b) adaptive threshold

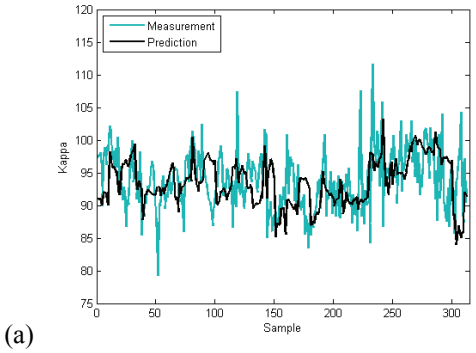
4.2 Industrial Case Study

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 more challenging problem. The soft sensor setup for this case is the same as that reported in Galicia et al. (2011a).

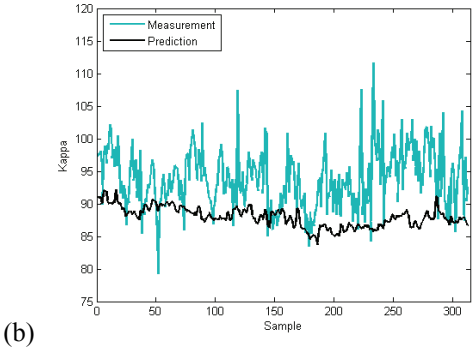
Figure 4 compares the soft sensor predicted Kappa number with the measured Kappy number for the four model update schemes, with Table 3 compared their performance indices.

Table 3: Performance of different soft sensors for the industrial case study

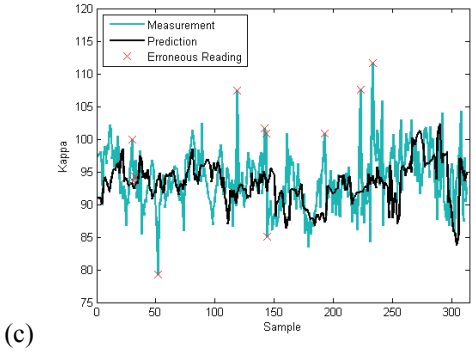
Soft Sensor	MSE	ME	MAPE(%)
(a)	22.85	0.17	4.02
(b)	46.99	5.50	6.01
(c)	21.52	0.33	3.94
(d)	18.30	-0.05	3.71



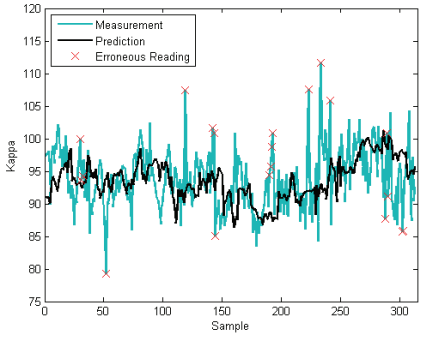
(a)



(b)



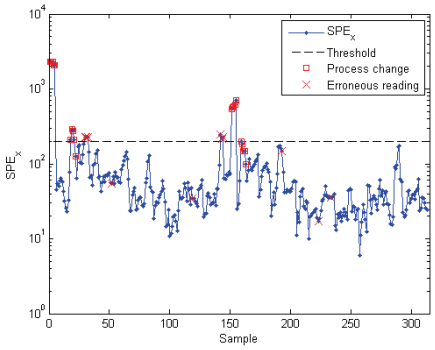
(c)



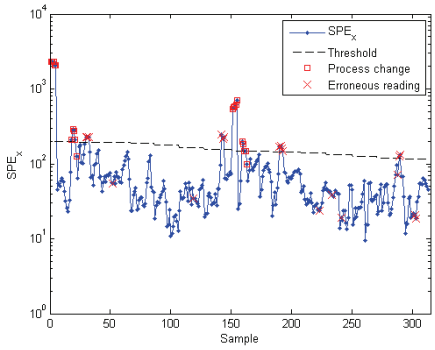
(d)

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

Figure 5 plots the SPE_x indices for update schemes c) and d), together with the classified outliers and the corresponding thresholds. It should be noted that for this case study the soft sensor that updates recursively with outlier detection performs did not update the model at all. This is due to the big difference between the training data and testing data, which causes all new data to be classified as outliers. This industrial case study once again confirms that the proposed outlier detection and classification approach is effective and robust in determining whether an outlier is caused by erroneous reading or process change.



(a)



(b)

Figure 5: SPE_x indices for the industrial case study; (a) fixed threshold; (b) adaptive threshold

6. CONCLUSIONS

Outlier detection and handling plays a critical role in data-driven soft sensor development. In this work, we propose a multivariate approach for online outlier detection, which is necessary for soft sensor recursive update. Specifically, we use squared prediction error indices for X and Y to detect the outliers in the independent variable and dependent variable spaces, respectively. In addition, to differentiate the outliers caused by erroneous reading from those caused by process changes, we propose a novel Bayesian approach to further classify the identified outliers. Finally, a robust way to update the outlier detection thresholds is developed to track process changes. Both simulated and industrial case studies of Kamyr digester demonstrate the superior performance of the soft sensor with proposed approaches for outlier detection and classification.

7. 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*.

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*.