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

A Systematic Approach for Soft Sensor Development

Bao Lin[a], Bodil Recke[b], Philippe Renaudat[b], Jørgen Knudsen[b], Sten Bay Jørgensen[a*]

[a] CAPEC, Department of Chemical Engineering, DTU

Lyngby 2800, Denmark

[b] FLS Automation

Valby 2500, Denmark

Abstract

This paper presents a systematic approach for development of a data-driven soft sensor using robust statistical technique. Data preprocessing procedures are described in detail. First, a template is defined based on a key process variable to handle missing data related to severe operation interruption. Second, a univariate, followed by a multivariate principal component analysis (PCA) approach, is used to detecting outlying observations. Then, robust regression techniques are employed to derive an inferential model. The proposed methodology is applied to a cement kiln system for realtime estimation of free lime, demonstrating improved performance over a standard multivariate approach.

Keywords: Multivariate regression analysis, Soft sensing, Robust statistics

1. Introduction

Soft sensors have been developed as supplement to online instrument measurements for process monitoring and control. Early work on soft sensor development assumed that a process model was available. The inferential model is developed using Kalman filter (Joseph and Brosilow, 1978). In case the process mechanisms are not well understood, empirical models, such as neural networks (Qin and McAvoy, 1992; Radhakrishnan and Mohamed, 2000), multivariate statistical methods may be used to derive the regression model (Kresta et al., 1994; Park and Han, 2000; Zhao, 2003). A model-based soft sensor can be derived if a first principle model (FPM) describes the process sufficiently accurately. However, modern measurement techniques enable a large amount of operating data to be collected and stored, thereby rendering data-driven soft sensor development a viable alternative.

Multivariate regression techniques have been extensively employed to develop data-driven soft sensors. Principal component regression (PCR) and partial least squares (PLS) address collinearity issues of process data by projecting the original process variables into a smaller number of orthogonal latent variables. Process measurements are often contaminated with data points that deviate significantly from the real values due to human errors, instrument failure or changes of operating conditions. Since

[*] Author to whom correspondence should be addressed: sbj@kt.dtu.dk

outlying observations may seriously bias a regression model, robust statistical approaches have been developed to provide a reliable model in the presence of abnormal observations. This paper presents a systematic approach for building a soft sensor. The proposed method using robust statistical techniques is applied to the estimation of free lime for cement kilns.

The paper is organized as follows. Section 2 describes data preprocessing which includes both univariate and multivariate approaches to detect outlying observations. The robust PCR and PLS approaches are presented in section 3, followed by the illustrative application on development of a free lime soft sensor for a cement kiln.

2. DATA PREPROCESSING

Outliers are commonly defined as observations that are not consistent with the majority of the data (Pearson, 2002; Chiang et al., 2003), including missing data points or blocks, and observations that deviate significantly from the normal values. A data-driven soft sensor derived with PCR or PLS deteriorates even in the presence of a single abnormal observation, resulting in model misspecification. Therefore, outlier detection constitutes an essential prerequisite step for a data-driven soft sensor design.

A heuristic procedure has been implemented in the paper to handle missing data related to sever operating interruptions. A template is defined by the kiln drive measurement to identify missing observations, since near zero drive current data corresponds to a stop of cement kiln operation. During such a period, other process measurements will not be reliable or meaningful. In case a small block (less than 2 hour) of data is missing, interpolated values based on neighbouring observations will be inserted. If a large segment of missing data is detected, these blocks will be marked and not used to build the soft sensor.

Both univariate and multivariate approaches have been developed to detect these outlying process observations. The *3σ edit rule* is a popular univariate approach to detect outliers (Ratcliff, 1993). This method labels outliers when data points are three or more standard deviations from the mean. Unfortunately, this procedure often fails in practice because the presence of outliers tends to inflate the variance estimation, causing too few outliers to be detected. The *Hampel identifier* (Davies and Gather, 1981) replaces the outlier-sensitive mean and standard deviation estimates with the outlier-resistant median and median absolute deviation from the median (MAD). The MAD scale estimate is defined as:

$$MAD = 1.4826 \ median \ \{|x_i - x^*|\} \tag{1}$$

where x^* is the median of the data sequence. The factor 1.4826 was chosen so that the expected MAD is equal to the standard deviation σ for normally distributed data.

Since process measurements from the cement kiln system are not independent from each other, detecting outliers using univariate diagnostics is not sufficient, resulting in masking and swamping. Masking refers to the case that outliers are incorrectly identified as normal samples; while swamping is the case when normal samples are classified to be outliers. Effective outlier detection approaches are expected to be based on multivariate statistical techniques.

Principal component analysis (PCA) is a multivariate analysis tool that projects the predictor data matrix to a lower dimensional space. The loading vectors corresponding to the k largest eigenvalues are retained to optimally capture the variations of the data and minimize the effect of random noise. The fitness between data and the model can be calculated using the residual matrix and Q statistics that measures the distance of a sample from the PCA model. Hotellings T^2 statistics indicates that how far the estimated sample by the PCA model is from the multivariate mean of the data, thus provides an indication of variability within the normal subspace.

The combined Q and T^2 tests are used to detect abnormal observations. Given the significance level for the Q (Jackson and Mudholkar, 1979) and T^2 statistic (Wise, 1991), measurements with Q or T^2 values over the threshold are classified as outliers. In this paper the significance level, α has the same value in the two tests, however finding a compromise between accepting large modelled disturbances and rejecting large unmodelled behaviours for outlier detection clearly needs further investigation.

3. ROBUST STATISTICS

Scaling is an important step in PCA. Since numerically large values are associated with numerically large variance, appropriate scaling methods are introduced such that all variables will have approximately equal weights in the PCA model. In the absence of a prior knowledge about relative importance of process variables, autoscaling (mean-centering following by a division over the standard deviation) is commonly used. Since both mean and standard deviation are inflated by outlying observations, autoscaling is not suitable for handling data which are especially noisy. This paper applies robust scaling to cement kiln data before performing PCA (Chiang et al., 2003) which replace mean by median and the standard deviation by MAD.

There are two types of approaches for rendering PCA robust. The first detects and removes outliers using a univariate approach then carries out a classic PCA on the new data set; the second is multivariate and is based on robust estimation of covariance matrix. An elliposidal multivariate trimming (MVT) (Devlin et al., 1981) approach is used. It iteratively detects bad data based on the squared Mahalanobis distance:

$$d_i^2 = (x_i - x_i^*)^T S^{*-1} (x_i - x_i^*) \tag{2}$$

where x_i^* is the current robust estimation of the location and S^* is the robust estimation of the covariance matrix. Since the data set has been preprocessed with a *Hampel identifier*, 95% of data with smallest Mahalanobis distance are retained in the next iteration. The iteration proceeds till both x_i^* and S^* converge. In this paper, the iteration stops at the 10th iteration such that at least 60% of the data is retained for the estimation of covariance matrix. Chiang et al (2003) suggested the closest distance to center (CDC) approach that $m/2$ observations with the smallest deviation from the center of the data is used to calculate the mean value. The CDC method is integrated such that the covariance matrix from the initialization step is not disrupted by outlying observations.

Principal component regression (PCR) derives an inferential model with score vectors and free lime measurements from the lab. During the regression step, zero weights are assigned to outlying observations identified by the PCA model; a weight value of one is assigned to normal data.

PLS is a multivariate statistical approach for relating input and dependent data matrices. The input data is projected onto a k -dimensional hyper-plane such that the coordinates are good predictors of dependent variables. The outlying measurements identified with an also downweighted PCA model before PLS analysis. The proposed approaches, robust PCR and weighted PLS, are applied to a data set collected from the log system of a cement kiln.

4. CASE STUDY

The product quality of a cement kiln is indicated by the amount of CaO (free lime) in clinker. The direct measurement is generally only available with a time delay of about an hour. In addition, the measurement also suffers from operating perturbations within the kiln and the cooler, which result in uncertain indication of the average quality. It is desirable to develop a soft sensor that is able to accurately predict the content of free lime in real time, and can be employed for effective quality control.

The operating data from a cement kiln log system are used to derive a soft sensor of free lime in the clinker. There are 13 process measurements available, including kiln drive current, kiln feed, fuels to calciner and kiln, plus several temperature measurements within the kiln system. The standard measurements are logged every 10 min, whereas the laboratory analysis of free lime content of the clinker is logged approximately every 2 hours. A data block of 12500 samples is selected in this study: 6500 samples to derive the model and 6000 samples for validation.

One step ahead prediction residual sum of squared errors (PRESS) between the model and measured lime content is used to select the number of principal components (PCs):

$$PRESS = \sum_{i=1}^{N_V} \left(\hat{y}(i) - y_m(i) \right)^2 \tag{3}$$

where N_V is the total number of samples during the validation period. It is calculated only when a new lab measurement is available.

The PRESS of regression models derived with PCR and PLS are shown in Figure 1. The PCR model with 5 PCs has the minimum PRESS (23.443). The PLS analysis shows a minimum of PRESS for 2 latent variables (LVs), because PLS finds LVs that describe a large amount of variation in X and are correlated with dependent variables, Y, while the PCs in PCR approach are selected only on the amount of variation they explain in X.

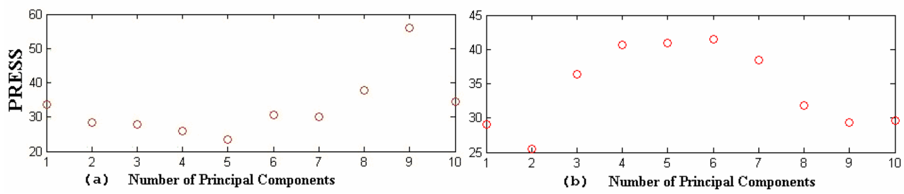

Figure 1. PRESS of (a) – PCR model ; (b) – PLS model during validation period

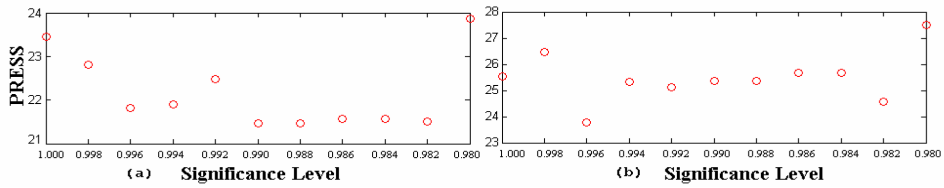

Figure 2. PRESS with significance level varied from 100% to 98% (a)-PCR (b) PLS model

Given the PCA decomposition, weights of 0 are assigned to abnormal points to downweight these observations before a regression model is derived. 95% significance level is commonly used for Q and T^2 tests. The lower the significance level, the higher the chance to reject outlying points. Figure 2 shows the PRESS of robust PCR model of 5 PCs and PLS model with 2 LVs while significance level varies from 100% to 98%. When the significance level is 100%, robust PCR approach is the same as the standard version applied to operating data preprocessed by the univariate approach.

As shown in Figure 2, downweighting outlying observations improves the predictability of both models. With the choice of an optimal significance level 99.0% for a PCR model, the minimum PRESS of 21.471 is obtained, which is about 20% less than that of a standard PCR model (25.754). The PRESS of a PLS model drops from 27.435 to 23.786 by incorporating robust statistical components and choosing the significance level of 99.6%. Figure 2 also reveals that an optimal significance level depends on the quality of the modelling data block.

Comparisons of the PCR and PLS model with lab measurements during the validation period are shown in Figure 3 and 4 respectively (only 1000 samples during the validation period is shown). PLS model is able to capture more of the relevant information than PCR model with a smaller number of LVs. Although the robust PCR approach has a smaller PRESS than that of the PLS model, it is obtained at the cost of using 3 more principal components and higher noise contents in the regression model.

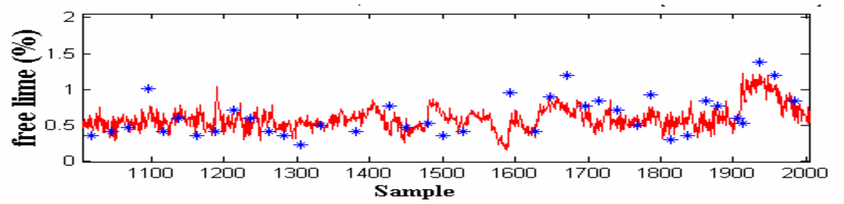

Figure 3. Validation of robust PCR model with 5 PCs (- lab measurements; solid line - PCR)*
PRESS = 21.471

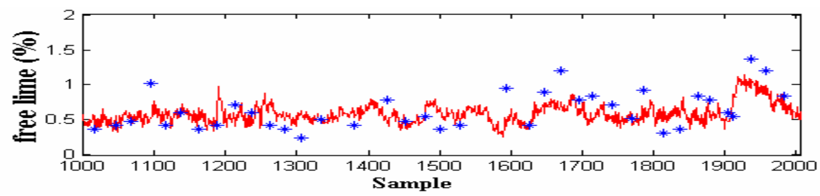

Figure 4. Validation of robust PLS model with 2 LVs (- lab measurements; solid line - PLS)*
PRESS = 23.786

Although deviations are observed when fast dynamics take place in the process (see Figure 3 and 4), the CaO soft sensors developed with a systematic robust statistical approach capture the slow change and the trend of lab measurements, which are important for process operation and control.

5. CONCLUSIONS

This paper presents a systematic approach to build a soft sensor using robust statistical approaches. The proposed methodology is applied to predict free lime of cement kiln systems. Due to the low signal-to-noise ratio in operating data, data preprocessing demonstrates to be an essential step for development of a data-drive soft sensor.

A case study demonstrates the improved performance of a robust PCA model for the detection outliers. The real-time estimation of free lime can be obtained with the data-driven soft sensor, which shows some potential to be used in closed loop control. Due to contamination of process measurements in operating data, downweighting outlying observations is beneficial to enhance the predictability of a regression model. The case study indicates the existence of a optimal downweighting vector determined by the significance level of Q- and T^2 - statistics. The issue of finding the optimal significance levels for regression model development and integrating the information from irregularly-sampled low quality measurements into the weighting vector need further investigation.

References

Chiang, L.H., R.J. Pell and M.B. Seasholtz (2003). "Exploring Process Data with the Use of Robust Outlier Detection Algorithms." Journal of Process Control, 13(5): 437-449.

Davies, L. and U. Gather (1981). "The Identification of Multiple Outliers." Journal of the American Statistical Association, 88: 782-801.

Devlin, S.J., R. Gnanadesikan and J.R. Kettenring (1981). "Robust Estimation of Dispersion Matrices and Principal Components." Journal of the American Statistical Association, 76: 354-362.

Jackson, J.E. and G.S. Mudholkar (1979). "Control Procedures for Residuals Associated With Principal Component Analysis." Technometrics, 21(3): 341-349.

Joseph, B. and C. Brosilow (1978). "Inferential Control of Processes." AIChE J., 24: 485-509.

Kresta, J.V., T.E. Marlin and J.F. MacGregor (1994). "Development of Inferential Process Models using PLS." Computers and Chemical Engineering, 18(7): 597-611.

Park, S. and C. Han (2000). "A Nonlinear Soft Sensor based on Multivariate Smoothing Procedure for Quality Estimation in Distillation Columns." Computers and Chemical Engineering, 24: 871-877.

Pearson, R.K. (2002). "Outliers in Process Modeling and Identification." IEEE Transactions on Control Systems Technology, 10(1): 55-63.

Qin, S.J. and T.J. McAvoy (1992). "Nonlinear PLS Modeling using Neural Networks." Computers and Chemical Engineering, 16(4): 379-391.

Radhakrishnan, V.R. and A.R. Mohamed (2000). "Neural Networks for the Identification and Control of Blast Furnace hot Metal Quality." Journal of process control, 10(6): 509.

Ratcliff, R. (1993). "Methods for Dealing with Reaction Time Outliers." Psychological Bulletin, 114: 510-532.

Wise, B.M. (1991). Adapting Multivariate Analysis for Monitoring and Modeling Dynamic Systems, PhD Thesis, University of Washington.

Zhao, Y.-H. (2003). A Soft Sensor Based on Nonlinear Principal Component Analysis. 2003 International Conference on Machine Learning and Cybernetics, Xian China.