International Journal of Systems Science

ISSN: 0020-7721 (Print) 1464-5319 (Online) Journal homepage: http://www.tandfonline.com/loi/tsys20

Robust PLS approach for KPI-related prediction and diagnosis against outliers and missing data

Shen Yin, Guang Wang & Xu Yang

To cite this article: Shen Yin, Guang Wang & Xu Yang (2014) Robust PLS approach for KPI-related prediction and diagnosis against outliers and missing data, International Journal of Systems Science, 45:7, 1375-1382, DOI: 10.1080/00207721.2014.886136

To link to this article: http://dx.doi.org/10.1080/00207721.2014.886136

International Journal of Systems Science, 2014
Vol. 45, No. 7, 1375–1382, http://dx.doi.org/10.1080/00207721.2014.886136

Taylor & Francis
Taylor & Francis Group

Robust PLS approach for KPI-related prediction and diagnosis against outliers and missing data

Shen Yin, Guang Wang* and Xu Yang

Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Harbin, China

(*Received 4 July 2013; accepted 26 September 2013*)

In practical industrial applications, the key performance indicator (KPI)-related prediction and diagnosis are quite important for the product quality and economic benefits. To meet these requirements, many advanced prediction and monitoring approaches have been developed which can be classified into model-based or data-driven techniques. Among these approaches, partial least squares (PLS) is one of the most popular data-driven methods due to its simplicity and easy implementation in large-scale industrial process. As PLS is totally based on the measured process data, the characteristics of the process data are critical for the success of PLS. Outliers and missing values are two common characteristics of the measured data which can severely affect the effectiveness of PLS. To ensure the applicability of PLS in practical industrial applications, this paper introduces a robust version of PLS to deal with outliers and missing values, simultaneously. The effectiveness of the proposed method is finally demonstrated by the application results of the KPI-related prediction and diagnosis on an industrial benchmark of Tennessee Eastman process.

Keywords: PLS; partial least squares; data-driven; KPI; key performance indicator; prediction; diagnosis; outlier; missing value; robust

1. Introduction

In the past few decades, the rapid development of modern technology has brought many benefits to the industrial process, e.g. a higher degree of automation, higher productivity and lower production costs. In the meanwhile, the industrial processes become more and more complicated. As a result, process performance-related issues such as safety and reliability of the complicated process become the most critical factors in process design (Chiang, Braatz, & Russell, 2001; Patton & Frank, 2000) and draw much attention both from industry and academia. If the analytical process model is available, the model-based fault diagnosis techniques can be successfully implemented in the framework of modern control theory and serve as powerful tools for online process monitoring (Ding, 2008, 2009; Dong, Wang, & Gao, 2012; Dong, Wang, Lam, & Gao, 2012; Du, Jiang, & Shi, 2012; Gao, Breikin, & Wang, 2007, 2008; Yu, Liu, Yang, Shi, & Tong, 2013). Actually, because of the difficulty in obtaining the precise analytical model of the process, it is still an enormous challenge to apply the model-based techniques in practical, especially in large-scale industrial process.

For modern industrial process, the goal of producing is not just the pursuit of production but the pursuit of quality. Consistent product quality has become a major concern due to the fierce competition in the worldwide market. Nowadays, a wide range of indusial applications require the quality variable or key performance indicator (KPI)-related prediction and diagnosis (Gao, Jian, & Luo, 2012; Wang, 2011). The relationship between the operating conditions and the KPI is very important which can help the operators to keep the production process running under economic operation while maintaining product quality consistency. In order to predict the KPI, much effort has been devoted to obtaining the analytical model of the production process. Unfortunately, hardly any results can be applied successfully in practice because of the complexity of the underlying process. In actual operation, the prediction of the KPI is usually completed by skilled operators with their experiences, which, however, also brings potential threats to the reliability and safety of the production process.

Therefore, techniques that can predict the KPI based on the measured process data are highly desired. In the meantime, detecting any process fault which has serious effect on the KPI as early as possible is very necessary for the successful operation of the process. Data-driven prediction and diagnosis approaches are such techniques that have developed quickly in recent years (Ding, Zhang, Naik, Ding, & Huang, 2009; Dong, Verhaegen, & Gustafsson, 2012; Li & Qin, 2001; Yin, Ding, Abandan Sari, & Hao, 2013; Yin, Ding, Haghani, Hao, & Zhang, 2012). Among these approaches, partial least squares (PLS) (de Jong, 1993; Geladi & Kowalski, 1986) is one of the most popular data-driven methods due to its simplicity and easy implementation in large-scale industrial process. The applications of PLS in

* Corresponding author. Email: guang.wang@hit.edu.cn

prediction and fault detection have been reported in many existing literatures (Yin & Wang, 2013; Yin, Wei, Gao, & Peng, 2012). The basic idea of PLS in prediction is to identify the regression coefficient between the measurable variables and the product quality variables. Based on it, the KPI can be predicted using the online measured process data. As PLS is totally based on the measured process data, the characteristics of the data are then quite critical for the success of PLS. In practical industrial processes, outliers and missing values are two common characteristics of the measured data that are caused by variety of reasons like hardware failure, formatting errors, non-representative sampling, etc.

The effectiveness of PLS can be seriously affected by outliers and missing values; therefore, these two abnormal data characteristics should be considered during the actual applications. In statistical sense, outliers are samples with extreme values that from a different population rather than the data majority. Generally speaking, two categories of outliers can be distinguished, i.e. (1) high leverage points that are located far from the data centre in the measurable variables space and, (2) high residual points with large absolute difference between the observed value and its prediction in the product quality variables space. To eliminate the impact of outliers, many robust versions of PLS have been proposed in the literatures (Cummins & Andrews, 1995; Hubert & Branden, 2003; Wakelinc & Macfie, 1992). Nevertheless, all these methods suffer from significant problems such as non-robust to leverage points (Cummins & Andrews, 1995), or high computational cost (Hubert & Branden, 2003; Wakelinc & Macfie, 1992). To obtain an efficient and robust method against both high leverage points and high residual points, Serneels, Croux, Filzmoser, and Van Espen (2005) proposed a partial robust M-regression (PRM) method that devotes to down-weighting the outliers by choosing a proper weighting scheme with relative less computational load. On the other hand, samples with missing values should be carefully treated as PLS also can not deal with such cases. A good strategy to solve this problem is to incorporate the PLS method in the expectation maximisation (EM) framework. EM (Dempster, Laird, & Rubin, 1977) is an iterative approach for dealing with missing elements, which has been widely used in data analysis (Kruger, Zhou, Wang, Rooney, & Thompson, 2008; Smirnov & Egbert, 2012; Turkmen, 2008). A methodology that integrates the robust multiple regression techniques into the EM framework has been reported in Stanimirova, Serneels, Van Espen, and Walczak (2007). More information about outliers and missing values as well as the relevant stochastic control problems can be found in Hubert and Branden (2003), Serneels et al. (2005), Kadlec, Gabrys, and Strandt (2009), Serneels and Verdonck (2009), Wang, Shen, Shu, and Wei (2012) and Wang, Shen, and Liu (2012).

From the application point of view, it has an important practical significance to develop an efficient data-driven KPI-related prediction and diagnosis approach against outliers and missing values, simultaneously. This paper aims to develop such an approach. Based on the PRM method proposed by Serneels et al. (2005) and the integration strategy suggested in Stanimirova et al. (2007), we first realise an EM-PRM method to predict the KPI, and furthermore we develop the EM-PRM-based fault detection method for the KPI-related fault diagnosis. The proposed approach will be only based on the available process measurements while completing the online monitoring tasks, i.e. predicting the KPI of the underling process, and detecting any abnormal faults related to the KPI, with less implementation efforts.

The rest of this paper is organised as follows. Section 2 first reviews the basic PLS algorithm, the PRM algorithm as well as the EM framework. Then, the EM-PRM-based KPI-related prediction and diagnosis approach is proposed based on these algorithms. Section 3 introduces an industrial benchmark of Tennessee Eastman (TE) process and presents the detailed process description and the fault scenarios. The proposed approach is applied on the benchmark in Section 4. Finally, we draw conclusion in Section 5.

2. KPI-related prediction and diagnosis

2.1. PLS-based prediction and diagnosis

PLS is a linear regression technique which aims to describe the relationship between the measurable variables matrix X and the product quality variable vector y (a univariate output is considered here). By projecting X and y onto the score matrix

$$T = [\, t_1 \cdots t_\gamma \,] \in \mathcal{R}^{N\times\gamma}$$

the PLS model can be presented as follows:

$$X = TP^T + \tilde{X} = \hat{X} + \tilde{X} \tag{1}$$

$$y = Tq^T + \tilde{y} = Xb + \tilde{y} \tag{2}$$

where $X = [\, x_1 \cdots x_N \,]^T \in \mathcal{R}^{N\times n}$ ($x_i \in \mathcal{R}^n$) records N observations of n measurable variables, $y = [\, y_1 \cdots y_N \,]^T \in \mathcal{R}^{N\times 1}$ ($y_i \in \mathcal{R}$) contains N observations of one product quality variable (i.e. KPI), γ is the number of latent variables. P and q are the loading matrixes of X and y, respectively. b is the regression coefficient vector, $\tilde{X}$ and $\tilde{y}$ are the residual matrix and residual vector, respectively. To calculate T and b we use the nonlinear iterative partial least squares (NIPALS) algorithm (Dayal & Macgregor, 1997) as follows:

(1) Normalise data matrix X and y into zero mean and unit variance
(2) For $i = 1, \ldots, \gamma$,

$$(w_i^*, q_i^*) = \arg \max_{||w_i||=1, ||q_i||=1} w_i^T X_i^T y q_i,\ X_1 = X$$

$$t_i = X_i w_i^*,\ p_i = \frac{X_i^T t_i}{||t_i||^2},\ X_{i+1} = X_i - t_i p_i^T$$
$$r_1 = w_1^*,\ r_i = \prod_{j=1}^{i-1}(I_{n\times n} - w_j^* p_j^T) w_i^*,\ i > 1$$

where γ is determined using cross validation (Wold, Sjöström, & Eriksson, 2001)

(3) Calculate the following matrixes:

$$P = [\, p_1 \cdots p_\gamma \,],\ T = [\, t_1 \cdots t_\gamma \,]$$
$$q = [\, q_1 \cdots q_\gamma \,],\ R = [\, r_1 \cdots r_\gamma \,]$$
$$b = Rq^T$$

Then, the predicted value of y_i (i.e. KPI) can be calculated using the online observation x_i as follows:

$$y_i = x_i b \tag{3}$$

Furthermore, the PLS-based fault diagnosis can be achieved by using the following steps under the assumption that x_i and y_i follow normal distribution. For each observation x_i, we calculate the test statistic using the Hotelling's T^2 or the squared prediction error (SPE) statistical algorithm, i.e.

$$T^2 = x_i^T R \left(\frac{T^T T}{n-1}\right)^{-1} R^T x_i \tag{4}$$
$$\text{SPE} = ||(I_{n\times n} - PR^T)x_i||^2 \tag{5}$$

Then we calculate the threshold corresponding to the Hotelling's T^2 or the SPE test statistic as follows:

$$J_{\text{th},T^2} = \frac{\gamma(n^2-1)}{n(n-\gamma)} F_\alpha(\gamma, n-\gamma) \tag{6}$$
$$J_{\text{th,SPE}} = g\chi^2_{h,\alpha} \tag{7}$$

where $F_\alpha(\gamma, n-\gamma)$ is F distribution with γ and $n-\gamma$ degrees of freedom with significance level α. $g\chi^2_{h,\alpha}$ is χ^2 distribution under significance level α with scaling factors $g = S/2\mu$ and $h = 2\mu^2/S$, where μ and S are the mean and variance of SPE statistic, respectively (Nomikos & MacGregor, 1995; Tracy, Young, & Mason, 1992). The appearance of a fault is confirmed when the test statistic exceeds the threshold.

2.2. *EM-PRM-based prediction and diagnosis*

PRM is a robust version of PLS devoting to reduce the impact of outliers by choosing a proper weighting scheme. Outliers that are far from the data centre will get small weighting coefficients, so that their impact will be weakened. As mentioned previously, high leverage points and high residual points are different types of outliers existing in the measurable variables space and the product quality variables space, respectively. Therefore, in the PRM algorithm, two types of weighting strategies are considered, i.e. the leverage weights w_i^x and the residual weights w_i^r, which are calculated as follows:

$$w_i^x = f\left(\frac{||t_i - \text{med}_{L_1}(T)||}{\text{med}_i(||t_i - \text{med}_{L_1}(T)||)}, c\right) \tag{8}$$
$$w_i^r = f\left(\frac{r_i}{\hat{\sigma}}, c\right) \tag{9}$$

with

$$r_i = y_i - t_i q \tag{10}$$
$$\hat{\sigma} = \text{med}(|r_i - \text{med}(r_j)|) \tag{11}$$
$$f(z,c) = \frac{1}{1 + |\frac{z}{c}|^2} \tag{12}$$

where med$(\cdot)$ denotes the median estimate and med$(\cdot)_{L_1}$ denotes the L_1-median estimate (Møller, von Frese, & Bro, 2005). f is the 'fair' function and c is a tuning constant (Serneels et al., 2005). The global weights w_i are calculated by

$$w_i = w_i^x w_i^r \tag{13}$$

The solving of PRM comes down to an iterative re-weighted PLS algorithm (Serneels et al., 2005). In each iteration, the observations (x_i, y_i) will be re-weighted by w_i to $(\sqrt{w_i}x_i, \sqrt{w_i}y_i)$, then PLS regression is performed on the re-weighted model.

Although PRM can effectively deal with outliers, it can not handle data with missing values which is another common problem in practice. A straightforward way to solve this problem is to incorporate the PRM method in the EM framework (Stanimirova et al., 2007). Each iteration of EM consists of two steps, i.e. (1) the expectation step, in which the missing elements are filled in by the expected values, and (2) the maximisation step, in which the expected values are updated using the data in which missing elements have been filled in. The iterative process stops if the estimates of the missing elements between two runs of the algorithm do not differ considerably (Serneels & Verdonck, 2009). By integrating PRM method into the EM framework, we get the following EM-PRM algorithm.

(1) Estimate the initial $X^{(0)}$ and $y^{(0)}$ where the missing elements are filled in by the column's medians of X and y, respectively.
(2) Set $l = 1$.
(3) If convergence is not attained, let $X = X^{(l-1)}$, $y = y^{(l-1)}$ and do:

(a) Compute the initial value w_i^r using Equation (9) with $r_i = y_i - \text{med}(y_j)$, and compute w_i^x using Equation (8) with t_i replaced by x_i. Then compute w_i using Equation (13).
(b) Multiply each row of X and y by $\sqrt{w_i}$. Perform PLS regression on the re-weighted model. Then correct score matrix T by dividing each row by $\sqrt{w_i}$.
(c) Recompute residual r_i using Equation (10) and update w_i using Equations (8), (9) and (13).
(d) Go back to 3(b) until the relative difference in norm between two consecutive approximations of regression coefficients is smaller than a specified threshold, e.g. 10^{-3}.
(e) Obtain score matrix $\hat{T}^{(l)}$, loading matrix $\hat{P}^{(l)}$ and $\hat{q}^{(l)}$.

(4) Set $\hat{X}^{(l)} = \hat{T}^{(l)}\hat{P}^{(l)^T}$ and $\hat{y}^{(l)} = \hat{T}^{(l)}\hat{q}^{(l)^T}$.
(5) Subtract the estimated missing elements of $\hat{X}^{(l)}$ and $\hat{y}^{(l)}$ from the corresponding elements of $X^{(l-1)}$ and $y^{(l-1)}$. Then, calculate the sum of all squared differences and divide it by the number of missing elements.
(6) If the calculated result is smaller than a specified threshold, e.g. 10^{-4}, then convergence is attained. Otherwise, create $X^{(l)}$ and $y^{(l)}$ using the non-missing elements of X and y and the estimated missing elements of $\hat{X}^{(l)}$ and $\hat{y}^{(l)}$.
(7) Set $l = l + 1$ and go to step (3).
(8) Get the final regression coefficient vector b from the last PLS step in the final iteration.

With the calculated final regression coefficient vector b, we can realise online KPI prediction by Equation (3). Next, we will develop the EM-PRM-based fault detection approach.

As pointed out in Yin, Ding, Zhang, Hagahni, and Naik (2011), the drawback of standard PLS algorithm comes from the orthogonal variations among interacting subspaces. It is desired that $\tilde{y}$ is uncorrelated with X. According to Yin et al. (2011), we have the following orthogonal decomposition algorithm for the EM-PRM-based fault detection.

- Perform the singular value decomposition (SVD) on bb^T

$$bb^T = \begin{bmatrix} P_b & \tilde{P}_b \end{bmatrix} \begin{bmatrix} \Lambda_b & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} P_b^T \\ \tilde{P}_b^T \end{bmatrix}$$

  where $P_b \in \mathcal{R}^{n\times 1}$, $\tilde{P}_b \in \mathcal{R}^{n\times(n-1)}$, $\Lambda_b \in \mathcal{R}^{1\times 1}$
- Construct Π_b and $\Pi_b^\perp$, which orthogonally project on span$\{b\}$ and span$\{b\}^\perp$, respectively.

$$\Pi_b = P_b {P_b}^T, \Pi_b^\perp = \tilde{P}_b \tilde{P}_b^T$$

- Decompose X into orthogonal subspaces $\hat{X}$ and $\tilde{X}$,

$$\hat{X} = X\Pi_b = X P_b P_b^T \in S_{\hat{X}} \equiv \text{span}\{b\}$$

$$\tilde{X} = X\Pi_b^\perp = X \tilde{P}_b \tilde{P}_b^T \in S_{\tilde{X}} \equiv \text{span}\{b\}^\perp$$

We expect to distinguish the faults that are related to the KPI and the faults that are unrelated to the KPI. Therefore, we should design test statistics on subspaces $\hat{X}$ and $\tilde{X}$, respectively. First, we use ${P_b}^T x_i \in \mathcal{R}$ for T^2 statistic for monitoring $\hat{X}$, i.e.

$$T_{\hat{X}}^2 = x_i^T P_b \left(\frac{P_b^T X^T X P_b}{N-1} \right)^{-1} P_b^T x_i. \tag{14}$$

The corresponding threshold for $T_{\hat{X}}^2$ is

$$J_{\text{th},T_{\hat{X}}^2} = \frac{(N^2-1)}{N(N-1)} F_\alpha(1, N-1) \tag{15}$$

and the fault detection logic is

$T_{\hat{X}}^2 > J_{\text{th},T_{\hat{X}}^2} \Longrightarrow$ a fault appears which will influence the KPI, otherwise fault-free.

Similar to $\hat{X}$, we use $\tilde{P}_b^T x_i \in R^{n-1}$ as the T^2 statistic for monitoring $\tilde{X}$ which is not correlated to KPI, i.e.

$$T_{\tilde{X}}^2 = x_i^T \tilde{P}_b \left(\frac{\tilde{P}_b^T X^T X \tilde{P}_b}{N-1} \right)^{-1} \tilde{P}_b^T x_i \tag{16}$$

The corresponding threshold for $T_{\tilde{X}}^2$ is

$$J_{\text{th},T_{\tilde{X}}^2} = \frac{(n-1)(N^2-1)}{N(N-n+1)} F_\alpha(n-1, N-n+1) \tag{17}$$

and the fault detection logic is

$T_{\tilde{X}}^2 > J_{\text{th},T_{\tilde{X}}^2} \Longrightarrow$ a fault appears which has no effect on the KPI, otherwise fault-free.

In the end, we summarise the EM-PRM-based KPI-related prediction and diagnosis algorithm into Algorithm 1:

Algorithm 1: Based on the normalised data X and y:

(1) Calculate the regression coefficient vector b using EM-PRM algorithm;
(2) Predict KPI using $\hat{y} = Xb$;
(3) Perform SVD on $b * b'$;
(4) Calculate the test statistics on $\hat{X}$ subspace and the corresponding threshold using Equations (14) and (15), respectively;
(5) Calculate the test statistics on $\tilde{X}$ subspace and the corresponding threshold using Equations (16) and (17), respectively.

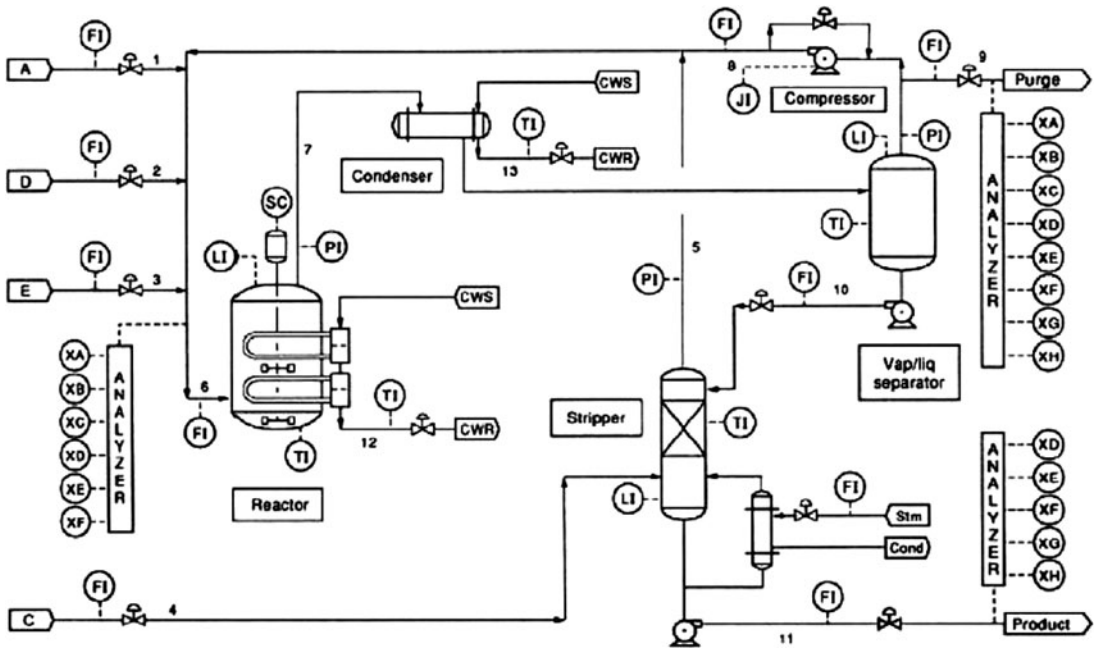


Figure 1. The Tennessee Eastman process.

3. Tennessee Eastman process

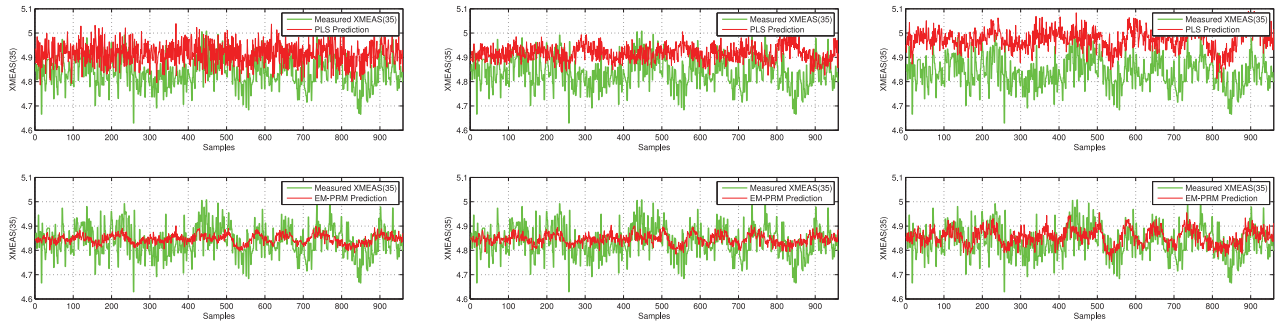
The TE process is a realistic simulation program of a chemical plant developed by Eastman Chemical Company to serve as a benchmark for control and monitoring studies. The simulator code and the simulated data can be download from the Internet. Figure 1 shows the diagram of the TE process which contains five major units: the reactor, the condenser, a vapour–liquid separator, a recycle compressor and a product stripper (Downs & Vogel, 1993). In the process, two produces are produced from four reactants. In addition, an inert and a by-product are also present making a total of eight components (Ding et al., 2009), which are named as A, B, C, D, E, F, G and H (Downs & Vogel, 1993). There are six operating modes mentioned in Downs and Vogel (1993), of which the base mode is considered here. As a highly instrumented process, a total of 53 variables are available, out of which 41 are process variables (XMEAS(1-41)) and 12 are manipulated variables (XMV(1-12)) (Downs & Vogel, 1993). In addition, 21 faults are predefined for fault detection purpose (Chiang et al., 2001) just as shown in Table 1.

4. Benchmark simulation

In this section, we will demonstrate the effectiveness of the proposed approach on the TE process. Two major tasks are involved in the simulations, i.e. (1) KPI-related prediction and (2) KPI-related fault diagnosis. For this purpose, 22 process measurements (XMEAS(1-22)) and 11 manipulated variables (XMV(2-12)) are selected as the data matrix X. The analyser for component G (XMEAS(35)) is taken as the product quality variable, i.e. the KPI. Both PLS-based and EM-PRM-based schemes are implemented for comparison purpose. All these schemes are designed based on 480 samples obtained from the normal process operation. However, these samples are clean data without any outliers and missing data; we should manually add a certain percentage of contaminated data in the normal samples. In actual, limited by the physical characteristics of the sensors, outlier is

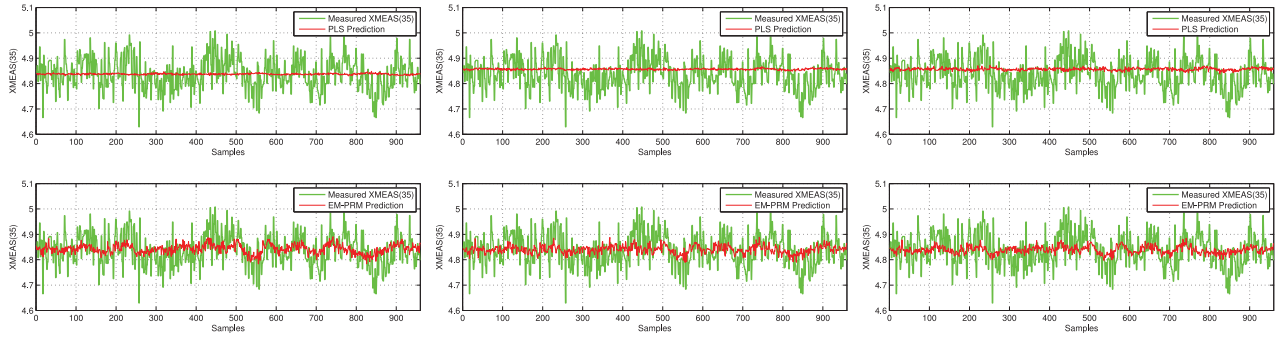
Table 1. Predefined faults in TE.

Fault number	Location
IDV(1)	A/C feed ratio, B composition constant
IDV(2)	B composition, A/C ration constant
IDV(3)	D feed temperature
IDV(4)	Reactor cooling water inlet temperature
IDV(5)	Condenser cooling water inlet temperature
IDV(6)	A feed loss
IDV(7)	C header pressure loss-reduced availability
IDV(8)	A,B,C feed composition
IDV(9)	D feed temperature
IDV(10)	C feed temperature
IDV(11)	Reactor cooling water inlet temperature
IDV(12)	Condenser cooling water inlet temperature
IDV(13)	Reaction kinetics
IDV(14)	Reactor cooling water valve
IDV(15)	Condenser cooling water valve
IDV(16)	Unknown
IDV(17)	Unknown
IDV(18)	Unknown
IDV(19)	Unknown
IDV(20)	Unknown
IDV(21)	The valve fixed at steady state position



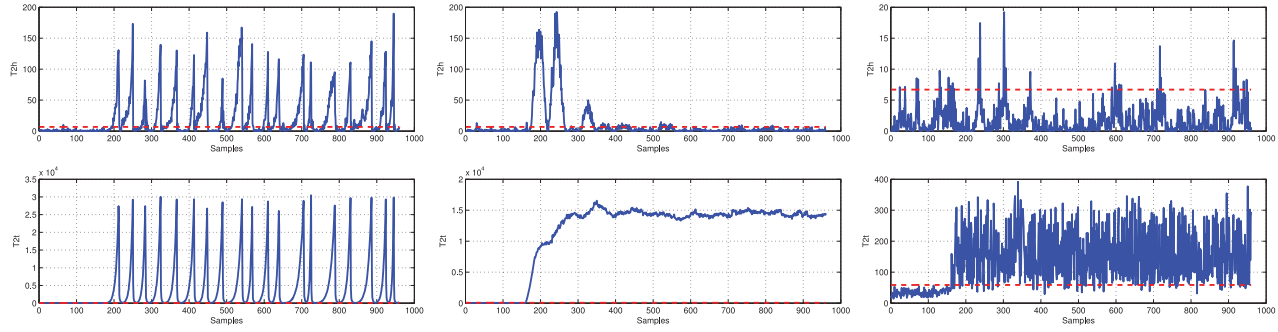
(a) 5% outliers (b) 10% outliers (c) 15% outliers

Figure 2. KPI-related prediction using PLS and EM-PRM with 1% missing data and (a) 5% outliers, (b) 10% outliers, (c) 15% outliers.



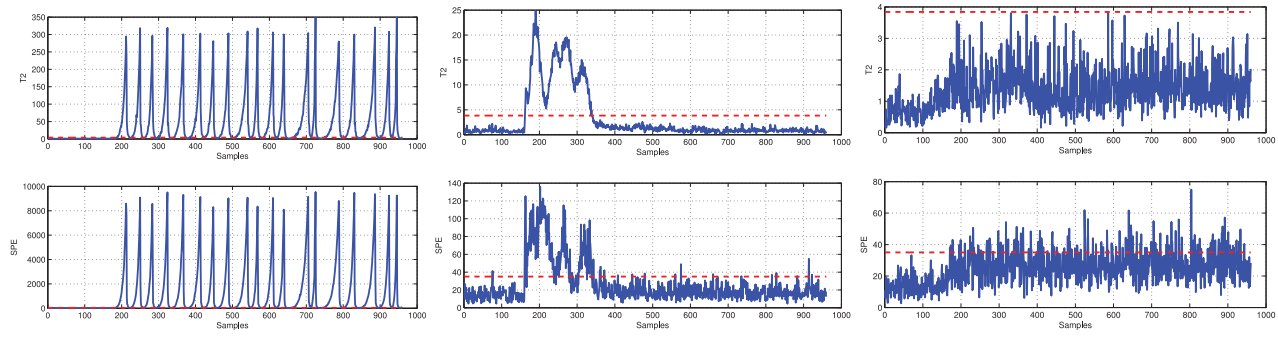
(a) 5% missing data (b) 10% missing data (c) 15% missing data

Figure 3. KPI-related prediction using PLS and EM-PRM with 1% outliers and (a) 5% missing data, (b) 10% missing data, (c) 15% missing data.



(a) IDV17 (b) IDV05 (c) IDV19

Figure 4. KPI-related fault detection using EM-PRM. (a) IDV(17), (b) IDV(5), (c) IDV(19).



(a) IDV17 (b) IDV05 (c) IDV19

Figure 5. KPI-related fault detection using PLS. (a) IDV(17), (b) IDV(5), (c) IDV(19).

often the data with higher or lower value than the normal data, and missing value is often 0 or a saturation value of the sensor. Therefore, in our simulations, outliers are the data with mean value of 1.5 times of the normal value, and all of the missing data are treated as zeros.

First, we consider the KPI-related prediction. Two groups of simulations are designed, where different proportions of outliers and missing data are mixed in the training samples. In the first group, the proportion of the outliers is fixed while the proportions of missing data are 5%, 10% and 15%. In the second group, the proportion of the missing data is fixed while the proportions of the outliers are 5%, 10% and 15%. Figures 2 and 3 show the results, respectively. In Figure 2, we can see that the EM-PRM-based method has a better prediction performance than PLS-based method for all considered proportions of outliers. With the increase of the outliers, the PLS-based method has a larger prediction mean error while the EM-PRM-based method keeps quite stable. In Figure 3, the EM-PRM-based method also performs better than the PLS-based method for all considered proportions of the missing data. Comparing Figures 2 and 3, we conclude that the EM-PRM-based method has a more accurate predicted output and a smaller predicted offset than the PLS-based method for all considered situations.

Next, we will compare the KPI-related diagnosis. Since the faults in the measurable variables space may occur in different subspaces, it is meaningful to design a monitoring scheme based on the joint use of the related test statistics, i.e. design test statistics in KPI-related space and KPI-unrelated space. In the sense of KPI-related classification of faults (Zhou, Li, & Qin, 2010), the joint use of the related test statistics can significantly reduce the false alarm rate. The EM-PRM-based fault detection approach is such a scheme that can distinguish whether the occurred fault has effect on the KPI or not. The detection results of the faults IDV(5), IDV(17) and IDV(19) using the EM-PRM-based approach and the PLS-based approach are presented in Figures 4 and 5, respectively, in which T2h is the T^2 statistic in the $\hat{X}$ subspace which is related to the KPI, T2t is the T^2 statistic in the $\tilde{X}$ subspace which is unrelated to the KPI, T2 is the T^2 statistic and SPE is the SPE statistic. As can be seen from Figure 4(a), fault IDV(17) occurs both in the $\hat{X}$ subspace and the $\tilde{X}$ subspace which has effect on the KPI. However, in Figure 4(c), fault IDV(19) occurs only in the $\tilde{X}$ subspace which has almost no influence on the KPI. Although fault IDV(5) occurs both in the $\hat{X}$ subspace and the $\tilde{X}$ subspace, IDV(5) has no influence on the KPI after 350 s. As we can see from Figure 5, the PLS-based fault detection approach only tells whether a fault appears but can not distinguish whether this fault has influence on the KPI.

5. Conclusion

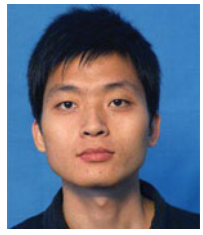
This paper presents an EM-PRM-based KPI-related prediction and diagnosis approach against outliers and missing data, simultaneously. Based on the PRM method and the EM framework, we first realise the EM-PRM-based KPI-related prediction approach. Afterwards, we develop the EM-PRM-based fault detection approach which can distinguish the fault related to the KPI and the fault unrelated to the KPI, so that the false alarm rate can be significantly reduced in the sense of KPI-related classification of faults. The effectiveness of the proposed approach is finally demonstrated on an industrial benchmark of TE process.

Funding

The authors acknowledge the support of China postdoctoral science foundation [grant number 2012M520738]; Heilongjiang postdoctoral fund [grant number LBH-Z12092].

Notes on contributors

Shen Yin received his BE degree in automation from Harbin Institute of Technology, China, in 2004, MSc degree in control and information system and the PhD degree in electrical engineering and information technology from University of Duisburg-Essen, Germany, in 2007 and 2012, respectively. His research interests are model-based and data-driven fault diagnosis, fault-tolerant control and their application to large-scale industrial processes.



Guang Wang received his BE degree in automation from Harbin Engineering University, Harbin, China, and the ME degree in control science and engineering from Harbin Institute of Technology, Harbin, China, in 2010 and 2012, respectively. He is currently working toward the PhD degree in control science and engineering with the Research Institute of Intelligent Control and Systems. His research interests include data-driven fault detection and diagnosis, performance monitoring, fault-tolerant control and their applications in the industrial process.

Xu Yang received his BS degree in automation from Harbin Institute of Technology, Harbin, China, in 2012. He is currently a master in control theory and engineering at Harbin Institute of Technology. His current research interests include modelling and analysis of mechatronics, reliability analysis of process control system, application of subspace identification method and application of data-driven method.

References

Chiang, L.H., Braatz, R.D., & Russell, E. (2001). *Fault detection and diagnosis in industrial systems*. London: Springer.

Cummins, D.J., & Andrews, C.W. (1995). Iteratively reweighted partial least squares: A performance analysis by Monte Carlo simulation. *Journal of Chemometrics, 9*(6), 489–507.

Dayal, B., & Macgregor, J.F. (1997). Improved PLS algorithms. *Journal of Chemometrics, 11*(1), 73–85.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems, 18*(3), 251–263.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.

Ding, S.X. (2008). *Model-based fault diagnosis techniques: Design schemes, algorithms, and tools*. Berlin: Springer.

Ding, S. (2009). Integrated design of feedback controllers and fault detectors. *Annual Reviews in Control, 33*(2), 124–135.

Ding, S., Zhang, P., Naik, A., Ding, E., & Huang, B. (2009). Subspace method aided data-driven design of fault detection and isolation systems. *Journal of Process Control, 19*(9), 1496–1510.

Dong, J., Verhaegen, M., & Gustafsson, F. (2012). Robust fault detection with statistical uncertainty in identified parameters. *IEEE Transactions on Signal Processing, 60*(10), 5064–5076.

Dong, H., Wang, Z., & Gao, H. (2012). Fault detection for Markovian jump systems with sensor saturations and randomly varying nonlinearities. *IEEE Transactions on Circuits and Systems I: Regular Papers, 59*(10), 2354–2362.

Dong, H., Wang, Z., Lam, J., & Gao, H. (2012). Fuzzy-model-based robust fault detection with stochastic mixed time delays and successive packet dropouts. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 42*(2), 365–376.

Downs, J.J., & Vogel, E.F. (1993). A plant-wide industrial process control problem. *Computers & Chemical Engineering, 17*(3), 245–255.

Du, D., Jiang, B., & Shi, P. (2012). Sensor fault estimation and compensation for time-delay switched systems. *International Journal of Systems Science, 43*(4), 629–640.

Gao, Z., Breikin, T., & Wang, H. (2007). High-gain estimator and fault-tolerant design with application to a gas turbine dynamic system. *IEEE Transactions on Control Systems Technology, 15*(4), 740–753.

Gao, Z., Breikin, T., & Wang, H. (2008). Reliable observer-based control against sensor failures for systems with time delays in both state and input. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 38*(5), 1018–1029.

Gao, C., Jian, L., & Luo, S. (2012). Modeling of the thermal state change of blast furnace hearth with support vector machines. *IEEE Transactions on Industrial Electronics, 59*(2), 1134–1145.

Geladi, P., & Kowalski, B.R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta, 185*, 1–17.

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

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

Kruger, U., Zhou, Y., Wang, X., Rooney, D., & Thompson, J. (2008). Robust partial least squares regression: Part I, algorithmic developments. *Journal of Chemometrics, 22*(1), 1–13.

Li, W., & Qin, S.J. (2001). Consistent dynamic PCA based on errors-in-variables subspace identification. *Journal of Process Control, 11*(6), 661–678.

Møller, S.F., von Frese, J., & Bro, R. (2005). Robust methods for multivariate data analysis. *Journal of Chemometrics, 19*(10), 549–563.

Nomikos, P., & MacGregor, J.F. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics, 37*(1), 41–59.

Patton, R.J., & Frank, P.M. (2000). *Issues of fault diagnosis for dynamic systems*. London: Springer.

Serneels, S., Croux, C., Filzmoser, P., & Van Espen, P.J. (2005). Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems, 79*(1), 55–64.

Serneels, S., & Verdonck, T. (2009). Principal component regression for data containing outliers and missing elements. *Computational Statistics & Data Analysis, 53*(11), 3855–3863.

Smirnov, M.Y., & Egbert, G. (2012). Robust principal component analysis of electromagnetic arrays with missing data. *Geophysical Journal International, 190*(3), 1423–1438.

Stanimirova, I., Serneels, S., Van Espen, P.J., & Walczak, B. (2007). How to construct a multiple regression model for data with missing elements and outlying objects. *Analytica Chimica Acta, 581*(2), 324–332.

Tracy, N., Young, J., & Mason, R. (1992). Multivariate control charts for individual observations. *Journal of Quality Technology, 24*(2), 88–95.

Turkmen, A. (2008). Robust partial least squares for regression and classification, VDM Verlag Dr. Müller: Saarbrücken.

Wakelinc, I., & Macfie, H. (1992). A robust PLS procedure. *Journal of Chemometrics, 6*(4), 189–198.

Wang, D. (2011). Robust data-driven modeling approach for real-time final product quality prediction in batch process operation. *IEEE Transactions on Industrial Informatics, 7*(2), 371–377.

Wang, Z., Shen, B., & Liu, X. (2012). H filtering with randomly occurring sensor saturations and missing measurements. *Automatica, 48*(3), 556–562.

Wang, Z., Shen, B., Shu, H., & Wei, G. (2012). Quantized H-infinity control for nonlinear stochastic time-delay systems with missing measurements. *IEEE Transactions on Automatic Control, 57*(6), 1431–1444.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109–130.

Yin, S., Ding, S.X., Abandan Sari, A.H., & Hao, H. (2013). Data-driven monitoring for stochastic systems and its application on batch process. *International Journal of Systems Science, 44*(7), 1366–1376.

Yin, S., Ding, S., Haghani, A., Hao, H., & Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control, 22*(9), 1567–1581.

Yin, S., Ding, S.X., Zhang, P., Haghani Abandan Sari, A., & Naik, A. (2011, August). *Study on Modifications of PLS Approach for Process Monitoring*. In World Congress (Vol. 18, No. 1, pp. 12389–12394).

Yin, S., & Wang, G. (2013, May). A modified partial robust M-regression to improve prediction performance for data with outliers. *2013 IEEE International Symposium on Industrial Electronics (ISIE)* (pp. 1–6). Taipei, Taiwan: IEEE.

Yin, S., Wei, Z., Gao, H., & Peng, K. (2012, October). Data-driven quality related prediction and monitoring. *IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society* (pp. 3874–3879). Taipei, Taiwan: IEEE.

Yu, J., Liu, M., Yang, W., Shi, P., & Tong, S. (2013). Robust fault detection for Markovian jump systems with unreliable communication links. *International Journal of Systems Science, 44*(11), 2015–2026.

Zhou, D., Li, G., & Qin, S.J. (2010). Total projection to latent structures for process monitoring. *AIChE Journal, 56*(1), 168–178.