Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Process-Quality Monitoring Using Semi-supervised Probability Latent Variable Regression Models

Le Zhou*, Zhihuan Song*, Junghui Chen, Zhiqiang Ge*, Zhao Li***

**State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, 310027, China*
(*e-mail: zjuzhoule@zju.edu.cn; songzhihuan@zju.edu.cn; gezhiqiang@zju.edu.cn; zhaoli@zju.edu.cn*)
***Department of Chemical Engineering, Chung-Yuan Christian University,*
Chung-Li, Taiwan, 320, R.O.C. (*e-mail: jason@wavenet.cycu.edu.tw*)

Abstract: In order to sustain safety and high product quality, the data-driven fault detection tools are increasingly used in industrial processes. The quality variables are the key index of the final product. Obtaining them in high frequency is time-consuming in the laboratory because they require the efforts of experienced operators. Meanwhile, process variables such as the temperature, the pressure, and the flow rate can be readily sampled in high frequency; hence the sample size between the process and the quality data is quite unequal. To effectively integrate two different observation sources, the high-rate process measurements and low-rate quality measurements, a semi-supervised regression model with probabilistic latent variables is proposed in this article to enhance the performance monitoring of the variations of the process and the quality variables. The corresponding statistics are also systematically developed and a TE benchmark problem is presented to illustrate the effectiveness of the proposed method.

1. INTRODUCTION

In modern industries, it is essential to produce value-added products of the high quality. To maintain the operating safety and quality consistency in various processes, multivariate statistical process monitoring has become the most popular direction and its approaches have been widely used in industrial processes, including chemicals, polymers, semi-conductor manufacturing and biology industries (Qin, 2012, Ge et al., 2013). Among them, principal component analysis (PCA) and its extensions have been firstly developed (Ge et al., 2009, Lee et al., 2004, Kim and Lee, 2003, Ge and Song, 2012). Of great importance is its ability to divide the original measured data into two orthogonal spaces, low-dimensional model subspace which contains the system behavior and the residual space which includes the uncertain patterns of the model, such as noises or outliers. By using the multivariate control charts, the variations of processes are then monitored in these two spaces(Choi et al., 2005).

PCA is based on a predefined model from the normal operating data of process variables(LI et al., 2009). Hence, it can detect the variations and the abnormal status of the process variables. However, producing a product based monitoring is crucial not only to the process operation but also to quality improvement. When the measured quality variables are incorporated into the monitoring model, due to the existing constraint relationships between the process variables (inputs) and the quality variables (outputs), the detectability of the abnormal situations resulting from the key quality variables will be enhanced. In chemometric techniques, projections to latent structures or partial least squares (PLS) and its other forms are increasingly used based on the process and the quality data collected from normal operations (Kruger et al., 2001, LI et al., 2009, Qin and Zheng, 2013). However, in the traditional PLS-based model, it is assumed that the sample size between process variables and quality variables are equal. Indeed, most process variables, like temperatures, levels, flow rates and pressures, are easily observed and recorded on a second or minute basis. Nevertheless, the quality variables that are the key indicators of the process performance are often measured off-line in the laboratory and are available infrequently on hourly or daily basis. Hence, what we observed is a small amount of quality data at several particular intervals and much more samples of the process variables.

For effective monitoring of the process performance, the statistic model should be developed based on complete data samples (both input and output variables), directly using the high-rate process measurements and low-rate quality measurements. Thus, the whole dataset can be divided into two parts. The one that contains both input and output data samples is denoted as the labelled dataset; the other that only consists of input data samples, as the unlabelled dataset. Model training with both labelled and unlabelled data samples is termed as the semi-supervised learning, which is an area in machine learning and more generally, artificial intelligence. Because semi-supervised learning requires less human effort and gives higher accuracy, its theory and practical applications are of great interest(Zhu, 2006). The common semi-supervised learning methods include the EM algorithm with generative mixture models, self-training, co-training, transductive support vector machines, and graph-based methods (Chapelle et al., 2008, Belkin et al., 2006, Ge and Song, 2011). In this article, a probabilistic generative model-based method called semi-supervised probabilistic latent variable regression (SSPLVR) is proposed. In a

SSPLVR model, the process variables and quality variables with unequal sample sizes are used to develop a semi-supervised model in the probabilistic framework. Compared with the conventional methods using labelled data only, the control decisions of the SSPLVR model using a small number of labelled data and a huge number of unlabelled process variables have been improved because all the data information has been sufficiently utilized.

The remainder of the paper is organized as follows. The supervised probabilistic latent variable regression model is briefly reviewed in Section 2. Then the detailed SSPLVR model is developed and how to train the model using the EM algorithm is discussed in Section 3. Its corresponding monitoring approaches are also proposed. The TE benchmark is carried out as a case study to evaluate the proposed method in Section 4. Finally, some conclusions are made.

2. PROBABILISTIC LATENT VARIABLE REGRESSION MODEL

2.1 Supervised probabilistic PCA

As a widely used technique for dimension reduction, PCA has been extended to its probabilistic form, which is called probabilistic PCA (PPCA). In a PPCA model, each observed sample is given as

$$\mathbf{x} = \mathbf{P}\mathbf{t} + \mathbf{e} \tag{1}$$

in which $\mathbf{x} \in R^J$ has been scaled to zero mean and $\mathbf{P} \in R^{J\times R}$ is the loading matrix. The latent variables $\mathbf{t} \in R^R$ are defined to follow standard normal distribution. J is the number of process variables. R is the number of latent variables. The noise of the process $\mathbf{e} \in R^J$ is set to be isotropic Gaussian with zero mean and its variance is calibrated by $\sigma^2\mathbf{I}$. $\mathbf{I}$ is an identity matrix. Like the other unsupervised method, only process variables are incorporated into the training model. Hence, PPCA cannot make clear that if the variations of process variables are relevant to the product quality.

2.2 Probabilistic Latent Variable Regression (PLVR)

To monitor quality variables more efficiently, the measured quality variables need to be utilized for modelling. When the quality data are measured, a supervised model can be constructed to link the relationships between process variables and quality variables. Given the input data (process variables, $\mathbf{x}$) and the output data (quality variables, $\mathbf{y}$), the probabilistic latent variable regression (PLVR) model is represented by

$$\mathbf{x} = \mathbf{W}\mathbf{t} + \mathbf{e} \tag{2}$$

$$\mathbf{y} = \mathbf{Q}\mathbf{t} + \mathbf{f} \tag{3}$$

where $\mathbf{W} \in R^{J\times D}$ and $\mathbf{Q} \in R^{M\times D}$ are loading and regression matrix respectively, in which M is the number of quality variables. Similar to PPCA, the latent variables $\mathbf{t} \in R^D$ which are shared by $\mathbf{x}$ and $\mathbf{y}$ follow Gaussian distribution with zero mean and unit variance. D is the number of the latent variables. The process noises of the process and the quality data are $\mathbf{e} \in R^J$ and $\mathbf{f} \in R^M$ respectively and the noises take isotropic Gaussian as $\mathbf{e} \sim N(0, \sigma_{\mathbf{x}}^2\mathbf{I})$ and $\mathbf{f} \sim N(0, \sigma_{\mathbf{y}}^2\mathbf{I})$. Given the latent variables, it is assumed that all the input and output data are conditionally independent to each other (Yu et al., 2006). Hence, the jointly marginal distribution of the observation $(\mathbf{x}, \mathbf{y})$ can be given

$$p(\mathbf{x}, \mathbf{y}) = \int p(\mathbf{x}|\mathbf{t})\, p(\mathbf{y}|\mathbf{t}) p(\mathbf{t}) d\mathbf{t} \tag{4}$$

For a sample set of N samples, the complete log-likelihood is then calculated as

$$L = \sum_{n=1}^{N} \ln p(\mathbf{x}_n, \mathbf{y}_n, \mathbf{t}_n \,|\, \mathbf{W}, \mathbf{Q}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2) \tag{5}$$

The model parameters $\{\mathbf{W}, \mathbf{Q}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2\}$ for the PLVR model can be estimated utilizing the EM algorithm. In the PLVR model, the process variables and quality variables are observed in the same frequency; that is, their quantity is equal. In real processes, however, the assumption of the PLVR model is hard to satisfy. Since most quality variables are often measured at a lower sampling rate because measuring the variables is time-consuming. The quality variables make offline examinations through particular instruments. Thus, only a small number of quality variables with plentiful process variables are collected. This means that the PLVR model can be only trained after some process variable data are removed.

3. SEMI-SUPERVISED PLVR MODEL

In this section, a semi-supervised probabilistic latent variable regression (SSPLVR) model for the unequal sample sizes of the process variables and the quality variables is proposed. Then SSPLVR will be applied to process monitoring.

3.1 SSPLVR model

It is assumed that the process and the quality variables are recorded as $\mathbf{X} \in R^{K\times J}$ and $\mathbf{Y} \in R^{N\times M}$, where $N < K$ owing to the lower sample frequency of quality data. With the assumption that each sample of the process is independent of each other, the order of the process variable can be adjusted so that the first N samples of the process variables will have their homologous quality variables (Fig. 1). After the measured data have been reordered and normalized, the whole observations (V) can be written as the union of the two parts, the labelled dataset (V_1) and the unlabelled dataset (V_2),

$$V = V_1 \bigcup V_2 = \left\{ \left(\mathbf{x}_n, \mathbf{y}_n\right) \middle| n = 1, \cdots, N \right\} \bigcup \left\{ \mathbf{x}_k \middle| k = N+1, \cdots, K \right\} \tag{6}$$

Thus, the complete log-likelihood is separated into two parts,

$$L = L_1 + L_2 = \sum_{n=1}^{N} \ln p(\mathbf{x}_n, \mathbf{y}_n) + \sum_{k=N+1}^{K} \ln p(\mathbf{x}_k) \tag{7}$$

The marginal probability can be estimated separately, but the latent variables link with the two parts.

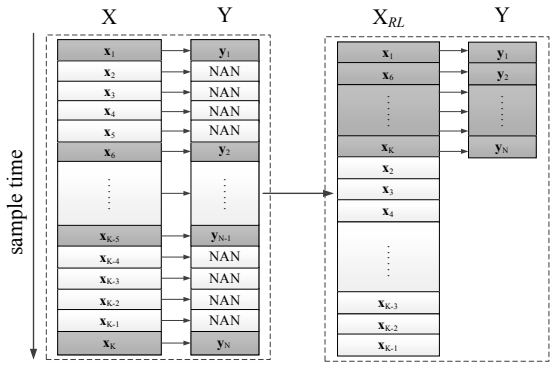


Fig.1. Collected samples are divided into the labelled dataset (in shade) and the unlabelled dataset (in un-shade).

3.2 EM training method for SSPLVR

The model parameters $\{\mathbf{W},\mathbf{Q},\sigma_{\mathbf{x}}^2,\sigma_{\mathbf{y}}^2\}$ for the SSPLVR model can be estimated using the EM algorithm. The general framework for EM iterates the expectation step (E-step) and the maximization step (M-step) until convergence.

In E-step, the old model parameters are fixed and are used to estimate the likelihood of all the observations. In M-step, the new parameters are calculated through maximizing the likelihood with respect to each of them. In E-step, given two partitions of the observed data $V_1=\{(\mathbf{x}_n,\mathbf{y}_n)|n=1,\cdots,N\}$ and $V_2=\{\mathbf{x}_k|k=N+1,\cdots,K\}$, the posterior distributions of the latent variables are calculated respectively. Then, the expected sufficient statistics of the two parts latent variables are written as

$$E(\hat{\mathbf{t}}_n|\mathbf{x}_n,\mathbf{y}_n)=\mathbf{M}^{-1}(\sigma_{\mathbf{x}}^{-2}\mathbf{W}^T\mathbf{x}_n+\sigma_{\mathbf{y}}^{-2}\mathbf{Q}^T\mathbf{y}_n) \tag{8}$$

$$E(\hat{\mathbf{t}}_n\hat{\mathbf{t}}_n{}^T|\mathbf{x}_n,\mathbf{y}_n)=\mathbf{M}^{-1}+E(\hat{\mathbf{t}}_n|\mathbf{x}_n,\mathbf{y}_n)E^T(\hat{\mathbf{t}}_n|\mathbf{x}_n,\mathbf{y}_n) \tag{9}$$

$$E(\hat{\mathbf{t}}_k|\mathbf{x}_k)=\sigma_{\mathbf{x}}^{-2}\mathbf{L}^{-1}\mathbf{W}^T\mathbf{x}_k \tag{10}$$

$$E(\hat{\mathbf{t}}_k\hat{\mathbf{t}}_k{}^T|\mathbf{x}_k)=\mathbf{L}^{-1}+E(\hat{\mathbf{t}}_k|\mathbf{x}_k)E^T(\hat{\mathbf{t}}_k|\mathbf{x}_k) \tag{11}$$

where $\mathbf{M}=\sigma_{\mathbf{x}}^{-2}\mathbf{W}^T\mathbf{W}+\sigma_{\mathbf{y}}^{-2}\mathbf{Q}^T\mathbf{Q}+\mathbf{I}$ and $\mathbf{L}=\sigma_{\mathbf{x}}^{-2}\mathbf{W}^T\mathbf{W}+\mathbf{I}$.

In M-step, the parameters $\{\mathbf{W},\mathbf{Q},\sigma_{\mathbf{x}}^2,\sigma_{\mathbf{y}}^2\}$ are similarly updated by maximizing the log likelihood function, which can be calculated as follows

$$\hat{\mathbf{W}}=\left[\sum_{n=1}^{N}\mathbf{x}_nE^T(\hat{\mathbf{t}}_n|\mathbf{x}_n,\mathbf{y}_n)+\sum_{k=N+1}^{K}\mathbf{x}_kE^T(\hat{\mathbf{t}}_k|\mathbf{x}_k)\right]\left[\sum_{n=1}^{N}E(\hat{\mathbf{t}}_n\hat{\mathbf{t}}_n{}^T|\mathbf{x}_n,\mathbf{y}_n)+\sum_{k=N+1}^{K}E(\hat{\mathbf{t}}_k\hat{\mathbf{t}}_k{}^T|\mathbf{x}_k)\right]^{-1} \tag{12}$$

$$\hat{\mathbf{Q}}=\left[\sum_{n=1}^{N}\mathbf{y}_nE^T(\hat{\mathbf{t}}_n|\mathbf{x}_n,\mathbf{y}_n)\right]\left[\sum_{n=1}^{N}E(\hat{\mathbf{t}}_n\hat{\mathbf{t}}_n{}^T|\mathbf{x}_n,\mathbf{y}_n)\right]^{-1} \tag{13}$$

$$\hat{\sigma}_{\mathbf{x}}^2=\frac{\left\{\begin{array}{l}\sum_{k=1}^{K}\mathbf{x}_k{}^T\mathbf{x}_k\\-2tr\left(\sum_{n=1}^{N}E^T(\hat{\mathbf{t}}_n|\mathbf{x}_n,\mathbf{y}_n)\hat{\mathbf{W}}^T\mathbf{x}_n+\sum_{k=N+1}^{K}E^T(\hat{\mathbf{t}}_k|\mathbf{x}_k)\hat{\mathbf{W}}^T\mathbf{x}_k\right)\\+tr\left(\left(\sum_{n=1}^{N}E(\hat{\mathbf{t}}_n\hat{\mathbf{t}}_n{}^T|\mathbf{x}_n,\mathbf{y}_n)+\sum_{k=N+1}^{K}E(\hat{\mathbf{t}}_k\hat{\mathbf{t}}_k{}^T|\mathbf{x}_k)\right)\hat{\mathbf{W}}^T\hat{\mathbf{W}}\right)\end{array}\right\}}{JK} \tag{14}$$

$$\hat{\sigma}_{\mathbf{y}}^2=\frac{\sum_{n=1}^{N}\left\{\begin{array}{l}\mathbf{y}_n{}^T\mathbf{y}_n-2E^T(\hat{\mathbf{t}}_n|\mathbf{x}_n,\mathbf{y}_n)\hat{\mathbf{Q}}^T\mathbf{y}_n\\+tr(E(\hat{\mathbf{t}}_n\hat{\mathbf{t}}_n{}^T|\mathbf{x}_n,\mathbf{y}_n))\hat{\mathbf{Q}}^T\hat{\mathbf{Q}}\end{array}\right\}}{MN} \tag{15}$$

in which tr $(\cdot)$ is a calculator for the trace value of the matrix. After some iterations between the E-step and the M-step, the likelihood will converge and the final model parameters can be obtained.

3.3 On-line monitoring scheme

For on-line monitoring of the measurement variables, the control limits should be built up at each sampling time point. Similar to the monitoring statistic of PLS, two commonly used measures T^2 and *SPE* statistics can be computed based on the SSPLVR model to monitor the variations of the latent variables subspace and the residual subspace. When a new sample of the process variables $\mathbf{x}_{new}$ is collected, the mean projection of the latent variables is estimated as

$$\mathbf{t}_{new}=(\mathbf{W}^T\mathbf{W}+\sigma_x^2\mathbf{I})^{-1}\mathbf{W}^T\mathbf{x}_{new} \tag{16}$$

Based on the latent variables of the process, the major variations of the model can be monitored through T^2 statistics, which is constructed as

$$T_{new}^2=\mathbf{t}_{new}{}^T\,\mathrm{var}^{-1}(\mathbf{t}_{new}|\mathbf{x}_{new})\mathbf{t}_{new} \tag{17}$$

in which $\mathrm{var}(\mathbf{t}_{new}|\mathbf{x}_{new})=(\sigma_x^{-2}\mathbf{W}^T\mathbf{W}+\mathbf{I})^{-1}$ represents the variance of the latent variables. Also, it is interesting to construct *SPE* statistics. Because the final quality indexes determine the value of the product, the variations of the quality variables should be considered. Hence, it is straightforward to calculate the prediction error of the quality variables and monitor the model residual subspace for the PLVR model. However, only a small number of the quality variables have been collected. For the labelled data whose process and quality variables are recorded simultaneously, SPE_1 built based on the prediction errors of the quality variables is given as

$$\begin{aligned}\mathbf{e}_{ynew}&=\mathbf{y}_{\mathrm{new}}-\hat{\mathbf{y}}_{new}\\&=\mathbf{y}_{\mathrm{new}}-\mathbf{Q}(\mathbf{W}^T\mathbf{W}+\sigma_x^2\mathbf{I})^{-1}\mathbf{W}^T\mathbf{x}_{new}\end{aligned} \tag{18}$$

$$SPE_1=\mathbf{e}_{ynew}{}^T\mathbf{e}_{ynew} \tag{19}$$

For the unlabelled data whose process variables are obtained, SPE_2 developed based on the prediction error of the process variables can be given.

$$\begin{aligned}\mathbf{e}_{xnew} &= \mathbf{x}_{new} - \hat{\mathbf{x}}_{new} \\ &= \mathbf{x}_{new} - \mathbf{W}(\mathbf{W}^T\mathbf{W} + \sigma_x^2\mathbf{I})^{-1}\mathbf{W}^T\mathbf{x}_{new}\end{aligned} \tag{20}$$

$$SPE_2 = \mathbf{e}_{xnew}{}^T\mathbf{e}_{xnew} \tag{21}$$

SPE_2 can be treated as the supplement formation when the quality data are not sampled during this period. With the developed statistic distribution that reflects the normal operation, control limits are required to detect any departure of the process from its standard behaviour. The confidence limit for T^2 can be approximated by means of an *F*-distribution(Tracy et al., 1992)

$$T_{new}^2 \sim \frac{D(K^2-1)}{K(K-D)}F_\alpha(D, K-D) \tag{22}$$

where $F_\alpha(D, K-D)$ is the upper $100\alpha\%$ critical point of *F* distribution with D and $K-D$ degrees of freedom. The confidence bound of SPE_1 and SPE_2 can be determined by a χ^2 distributed approximation: $SPE \sim g \cdot \chi_h^2$, where *g* and *h* can be calculated as(Wen et al., 2012)

$$\begin{aligned} g \cdot h &= mean(SPE_{normal}) \\ 2g^2h &= \mathrm{var}(SPE_{normal}) \end{aligned} \tag{23}$$

Among the three statistics of SSPLVR, T^2 statistic can reflect the variations of both process and quality variables, especially the process variables that is relevant to the final quality variables, which is because not only the intra-covariance of the process and the quality variables but also the inter-covariance between them is taken into account in the latent variables subspace. SPE_1 statistic, which is based on the prediction errors of quality variables, will reveal that if the fault is quality relevant. Similarly, SPE_2 statistic mainly reflects the variations of process variables. After a huge number of unlabelled process variables are incorporated into SSPLVR, the performance of T^2 and SPE_2 statistic will be improved naturally When more information of the process is provided. However, the detection ability of SPE_1 statistic will also be enhanced because the increase of the prediction accuracy of the model when all the data information has been sufficiently utilized in SSPLVR.

4 CASE STUDY

The Tennessee Eastman (TE) process is a real industrial plant often used for developing and evaluating the multivariable control technology. The process involves five major operation units: a reactor, a condenser, a compressor, a separator, and a stripper. More detailed discussion of the process and its control structures can be obtained in the reference (Downs and Vogel, 1993). In the TE process, the sample frequencies between the process and the quality variables are dissimilar. In this study, 16 easy-measured process variables are selected as the input data, which are tabulated in Table 1. All the process variables are collected per 3 minutes, whereas in the real process, the quality variable is sampled at the lower frequency. Hence, the composition of Stream 9 is assumed to be collected per 30 minutes in this paper. Then, 500 process data samples and 50 quality data samples are incorporated into the SSPLVR model to build up a semi-supervised projection. This means that 10% of the process variables are labelled by the corresponding quality data.

Table 1. Monitoring variables in the TE process

No.	Measured variables	No.	Measured variables
1	A feed	9	Product separator temperature
2	D feed	10	Product separator pressure
3	E feed	11	Product separator underflow
4	A and C feed	12	Stripper pressure
5	Recycle flow	13	Stripper temperature
6	Reactor feed rate	14	Stripper steam flow
7	Reactor temperature	15	Reactor cooling water outlet temperature
8	Purge rate	16	Separator cooling water outlet temperature

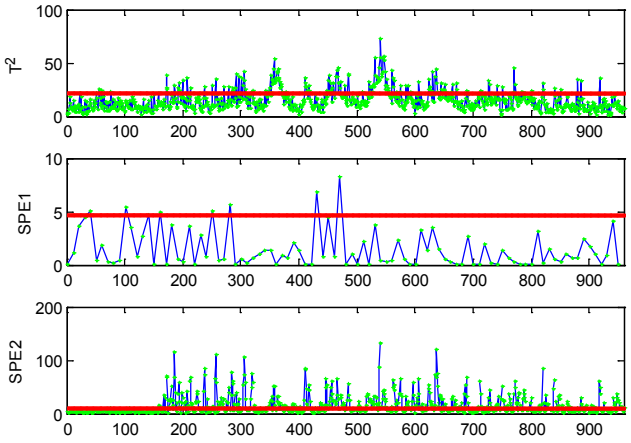
Table 2. Missing detection rates in TE Process

Fault types	***SSPLVR***			***PLVR***	
	T^2	SPE_1	SPE_2	T^2	*SPE*
1	0.0063	0.225	0.0014	0.0125	0.8625
2	0.0175	0.0375	0.0153	0.025	0.1875
3	0.8962	0.9375	0.9375	0.9625	0.975
4	0.9638	0.9375	0.9681	0.9875	1
5	0.7013	0.875	0.7431	0.75	0.925
6	0.0262	0.0375	0	0	0.0625
7	0.5425	0.7625	0.6014	0.5875	0.9
8	0.025	0.575	0.0264	0.0375	0.575
9	0.9287	0.95	0.9417	0.975	0.9875
10	0.5062	0.875	0.3639	0.6	0.925
11	0.7688	0.9375	0.4778	0.875	0.9375
12	0.0138	0.425	0.0167	0.05	0.425
13	0.0563	0.4375	0.0486	0.0625	0.325
14	0.0362	0.9625	0	0.3	0.9875
15	0.8612	0.9125	0.9514	0.9375	0.975

The monitoring performance of the proposed method is tested using 15 known faults of the TE process. All the faults consist of 960 samples and the faults all happened after 160 sampling time. Similarly, only 10% of process variables, i.e. 96 process variables are labelled. To make a comparison, the PLVR model is also applied based on the labelled data only. Take Fault 2 and Fault 11 for examples. For Fault 11, Fig. 2 shows that the T^2 statistic of SSPLVR performs better than the results of PLVR, because the T^2 statistic is constructed based on the latent variables subspace and it reflects both the variations of the process and the quality variables. The SPE_2 statistic is built based on the prediction error of the unlabelled process variables and it mainly reveals the changes of the

process variables. Hence, the monitoring performance of the T^2 and SPE_2 statistic is improved when many unlabelled process data are incorporated into the SSPLVR model. For Fault 2, Fig 3 also indicates that SPE_1 of SSPLVR performs better than SPE of PLVR. The SPE_1 statistic is used to indicate if the fault affects the final product quality, because it is based on the prediction error of the quality variables. It is also seen that SPE_1 of SSPLVR can detect the fault earlier, which is crucial for the consistency of the product quality because the quality performance deterioration is monitored timely. For all the 15 faults, the missing detection rates of SSPLVR and PLVR are compared and listed in Table 2. It is found that both T^2 and the SPE statistics of SSPLVR are more efficient in fault detection in most cases. The PLVR model just utilized limited labelled process variables and the corresponding quality variables. Even though the constrain relationship of PLVR between the input and the output data can be constructed, the useful data are too few to mine the valuable information among them. On the contrary, the huge unlabelled process variables are helpful to detecting the abnormal situations of the process with higher accuracy.

(a)



(b)

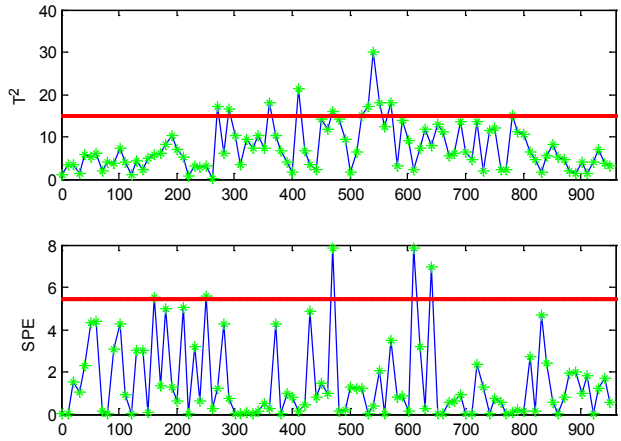
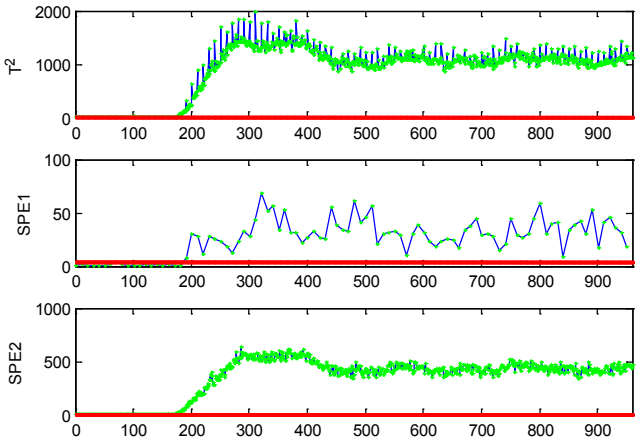


Fig.2. Monitoring results of Fault 11 by (a) SSPLVR; (b) PLVR

The monitoring performance affected by the sample sizes is evaluated. With more labelled data, the performance of the supervised model should be improved and vice versa. The missing detection rates using SSPLVR and PLVR for different sample sizes of quality data are collected and listed in Table 3. It is found that the missing detection rates of PLVR increase when the sample time of quality data change from 15 minutes to 60 minutes. As the labelled data are too few to extract the features of the whole process accurately, more false alarms and detection delays occur. However, the monitoring performance and detection delays of SSPLVR are almost unchanged for the quality data in different sampling frequencies. Thus, the SSPLVR monitoring model performs better after incorporating all the process and quality data into the model, especially when the sample size of the quality data are few.

(a)



(b)

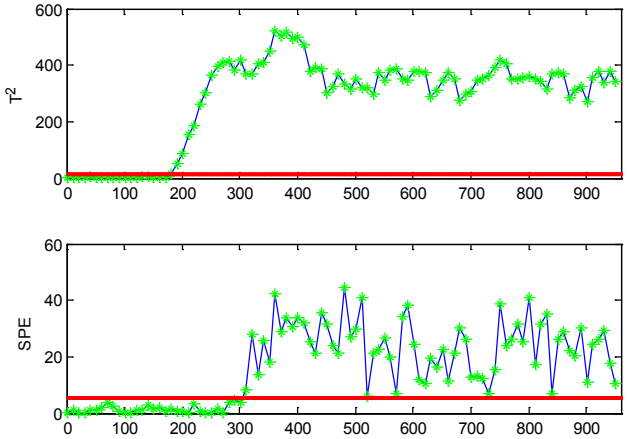


Fig.3. Monitoring results of Fault 2 by (a) SSPLVR; (b) PLVR

Table 3: Missing detection rates using SSPLVR and PLVR for different sample size of quality data

Sample time(min)	*SSPLVR*			*PLVR*	
	T^2	SPE_1	SPE_2	T^2	*SPE*
6	0.0175	0.075	0.0175	0.02	0.0525
15	0.0175	0.0375	0.0172	0.0313	0.1812
30	0.0175	0.0375	0.0153	0.025	0.1875
60	0.0175	0.075	0.0158	0.025	0.9

5 CONCLUSION

Monitoring the variations of the key quality variables is more important for on-line fault detection. It is hard to achieve because the quality data are examined offline and infrequently. It is straightforward using the process and the quality variables of the same sample size to construct the relationship between them, but the monitoring performance becomes much worsen when too few quality data are available. In this paper, a SSPLVR model is proposed using the unequal sample size between the process and the quality variables. When a mass of process variables are incorporated into the model, most of them do not have corresponding measured quality variables, but it is helpful to improve a more accurate regression model.

SSPLVR is developed for process monitoring. Similar to the supervised probabilistic model, the T^2 and two *SPE* statistics of the SSPLVR model are developed to monitor the variations of the latent variable subspace, process and quality data regression errors respectively. To evaluate the feasibility of the proposed SSPLVR model, a TE benchmark is illustrated. The results disclose that the proposed method is superior to the supervised model which uses the labelled data only.

ACKNOWLEDGEMENT

This work is supported by the National Basic Research Program (973 Program) of China Grant Number 2012CB720505, the National Natural Science Foundation of China (61273167) and the National Science Council, R.O.C.

REFERENCES

BELKIN, M., NIYOGI, P. & SINDHWANI, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research,* 7**,** 2399-2434.

CHAPELLE, O., SINDHWANI, V. & KEERTHI, S. S. 2008. Optimization techniques for semi-supervised support vector machines. *The Journal of Machine Learning Research,* 9**,** 203-233.

CHOI, S. W., LEE, C., LEE, J.-M., PARK, J. H. & LEE, I.-B. 2005. Fault detection and identification of nonlinear processes based on kernel PCA. *Chemometrics and intelligent laboratory systems,* 75**,** 55-67.

DOWNS, J. J. & VOGEL, E. F. 1993. A plant-wide industrial process control problem. *Computers & Chemical Engineering,* 17**,** 245-255.

GE, Z. & SONG, Z. 2011. Semisupervised Bayesian method for soft sensor modeling with unlabeled data samples. *AIChE Journal,* 57**,** 2109-2119.

GE, Z. & SONG, Z. 2012. *Multivariate Statistical Process Control: Process Monitoring Methods and Applications*, Springer.

GE, Z., SONG, Z. & GAO, F. 2013. Review of recent research on data-based process monitoring. *Industrial & Engineering Chemistry Research,* 52**,** 3543-3562.

GE, Z., XIE, L., KRUGER, U., LAMONT, L., SONG, Z. & WANG, S. 2009. Sensor fault identification and isolation for multivariate non-Gaussian processes. *Journal of Process Control,* 19**,** 1707-1715.

KIM, D. & LEE, I.-B. 2003. Process monitoring based on probabilistic PCA. *Chemometrics and intelligent laboratory systems,* 67**,** 109-123.

KRUGER, U., CHEN, Q., SANDOZ, D. & MCFARLANE, R. 2001. Extended PLS approach for enhanced condition monitoring of industrial processes. *AIChE journal,* 47**,** 2076-2091.

LEE, J.-M., YOO, C., CHOI, S. W., VANROLLEGHEM, P. A. & LEE, I.-B. 2004. Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science,* 59**,** 223-234.

LI, G., QIN, S.-Z., JI, Y.-D. & ZHOU, D.-H. 2009. Total PLS based contribution plots for fault diagnosis. *Acta Automatica Sinica,* 35**,** 759-765.

QIN, S. J. 2012. Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*.

QIN, S. J. & ZHENG, Y. 2013. Quality‐relevant and process‐relevant fault monitoring with concurrent projection to latent structures. *AIChE Journal,* 59**,** 496-504.

TRACY, N., YOUNG, J. & MASON, R. 1992. Multivariate control charts for individual observations. *Journal of Quality Technology,* 24.

WEN, Q., GE, Z. & SONG, Z. Nonlinear dynamic process monitoring based on kernel partial least squares. American Control Conference (ACC), 2012, 2012. IEEE, 6650-6654.

YU, S., YU, K., TRESP, V., KRIEGEL, H.-P. & WU, M. Supervised probabilistic principal component analysis. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006. ACM, 464-473.

ZHU, X. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison,* 2**,** 3.