2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

Process Monitoring based on Kullback Leibler Divergence

Jiusun Zeng, Lei Xie, Uwe Kruger, Jie Yu, Jingjing Sha, Xuyi Fu

Abstract— **This article proposes to monitor industrial process faults using Kullback Leibler (KL) divergence. The main idea is to measure the difference between the distributions of normal and faulty data. Sensitivity analysis on the KL divergence under Gaussian distribution assumption is performed, which shows that the sensitivity of KL divergence increases with the number of samples. For non-Gaussian data, a recently proposed kernel method for density ratio estimation is used to estimate the KL divergence. The density ratio estimation method does not involve direct estimation of probability density functions, hence is fast and efficient. For monitoring of non-Gaussian data, the confidence limits are obtained through a window based strategy. Application studies involving a simulation example and an industrial melter process show that the performance of the proposed monitoring strategy is better than the principal component analysis (PCA) based statistical local approach.**

I. INTRODUCTION

The requirements of reliability, safety and efficiency of modern industrial processes have made the problem of fault detection and diagnosis increasingly important. It has been recognized that statistical-based techniques have many attractive advantages in dealing with large variable sets encountered in complex industrial systems. Among various statistical-based techniques, multivariate projection-based methods like principal component analysis (PCA) and partial least squares (PLS) [1],[2],[3],[4] are perhaps the most popular ones. It is found that PCA is an optimal procedure when the errors in variables are independently and identically distributed (i.i.d.). However, the structure of errors in variables is generally unpredictable, making the estimation by traditional PCA inconsistent in some cases [5]. Reference [6] proposed to use maximum likelihood PCA (MLPCA) to get a consistent estimation of the model matrix by applying Cholesky decomposition to the error covariance matrix. Moreover, the reference also proposed an algorithm for simultaneously estimating the error covariance matrix and the model matrix. Recently, Fietal *et al.* [7] presented a revised version of this algorithm by introducing a method for estimation of the number of source signals. The general procedure of process monitoring by PCA is to use T^2 and Q statistics to detect process faults exhibited in the latent variables and the residuals. By using T^2 and Q statistics, the underlying assumption is that the data considered follow Gaussian distribution. However, due to different kinds of practical reasons, many process variables exhibit significantly non-Gaussian behaviors. The latent variables may exhibit significant non-Gaussian information, so that monitoring by T^2 statistic may not be appropriate. It is shown in Liu *et al.* [8] that the non-Gaussian components are encapsulated in the first k dominant principal components, the authors suggested to use independent component analysis (ICA) to extract non-Gaussian components, similar lines are employed by [9], [10].

As the problem of monitoring Gaussian components is well defined in literature, monitoring of non-Gaussian signal components is not so straightforward. Several different approaches have been proposed, including kernel density estimation (KDE) [11], support vector data description (SVDD) [8], statistical local approach [12] etc. The main problem of KDE is that the identification of confidence limits is laborious, especially when the dimension of non-Gaussian components is high. Moreover, KDE is not suitable for sparse or clustered data distribution [13]. In contrast, the confidence limits and regions can be obtained by solving a quadratic programming problem using SVDD. Comparing to KDE, SVDD is computationally efficient; however, it is difficult to determine the parameters (kernel width and misclassification probability) of SVDD. Different from KDE and SVDD, the statistical local approach transforms the problem of fault detection of non-Gaussian processes to the monitoring of the mean of a Gaussian vector, hence significantly simplifying the monitoring task. However, according to the central limit theorem, the monitoring statistic follows a Gaussian distribution only when the length of samples tends to infinity, which is always not the case in practice.

In this paper, MLPCA is used to get a consistent estimation of the model and the error covariance matrix. Unlike previous work which uses ICA to the extracted dominant principal components, this paper addresses the problem of determining the confidence limits for non-Gaussian distributed principal components by inspecting the deviation between probability density functions (PDF) of normal and faulty data, which is measured by Kullback Leibler (KL) divergence. The sensitivity analysis of the KL divergence as a test statistic is

This work was supported in part by National Science Foundation of China under Grant No. 61203088 and Natural Science Foundation of Zhejiang Province under Grant No. LQ12F03015.

Jiusun Zeng is with the College of Metrology & Measurement Engineering, China Jiliang University, Hangzhou, 310018, China. email:zjs1020@gmail.com

Lei Xie is with the Institute of Cyber Systems & Control, Zhejiang University, Hangzhou, 310027, China phone: 86-571-87952268; fax: 86-571-87952279, email: leix@csc.zju.edu.cn

Uwe Kruger is with Department of Mechanical and Industrial Engineering, Sultan Oaboos University, P. O. Box 33, Sultanate of Oman email:uwe.kruger@gmail.com

Jie Yu is with the Department of Chemical Engineering, McMaster University, Hamilton L8S 4L7, Canada. email:jieyu@mcmaster.ca

Jingjing Sha is with the Institute of Cyber Systems & Control, Zhejiang University, Hangzhou, 310027, China. email:shajj@zju.edu.cn

Xuyi Fu is with the College of Metrology & Measurement Engineering, China Jiliang University, Hangzhou, 310018, China. emal:xuyifu001@gmail.com

also presented. It is found that direct estimation of PDF can be avoided and only the density ratio between normal and faulty data needs to be estimated [14], [15]. With the density ratio function, an empirical estimation of Kullback-Leibler divergence between the distributions of normal and faulty data can be easily obtained. A window based monitoring scheme is then explored by using KL divergence as a test statistic.

Comparing to existing work in the literature, the monitoring strategy by KL divergence is conceptually more straightforward and easier to understand. The proposed framework shares the advantages of statistical local approach; it only requires univariate monitoring statistics. Through a window-based approach, it is also sensitive in detecting incipient fault conditions. Moreover, since the proposed framework considers the difference between distributions, it can detect more general kinds of process faults.

The remainder of the paper is organized as follows. In Section 2, statistic properties of KL divergence as a test statistic is explored. Section 3 presents a kernel based density ratio estimation method for estimation of KL divergence. Sections 4 develop a monitoring strategy for non-Gaussian components by using a permutation test procedure. Section 5 and Section 6 consider two application studies involving a simulation example and a glass melter process. Finally, concluding summaries of this article are presented in Section 7.

II. Kullback Leibler divergence as a test statistic

In multivariate statistical process control (MSPC), an important task is to monitor the difference between normal data and test data to check whether faulty operation arises. A natural idea would be to measure the difference between the probability density functions of normal data and test data, which can be achieved by observing the KL divergence between the two distributions. The KL divergence provides a means to compare between different distributions and has a wide range of applications from text and image retrieval[16], pattern recognition[17], change detection[18] etc. The KL divergence has several favorable properties, e.g., it shares some properties of distance metric, it is invariant to monotonic transformation of data and easily applicable to multivariate cases etc.

The general procedure of statistical monitoring is to collect a large number of data samples operating on normal condition and use the normal data as reference data set. The test data is then compared with the normal data to check whether abnormal conditions occur. Hence it is reasonable to regard the distribution of normal data as reference distribution when computing the KL divergence. Assume d dominant principal components $\mathbf{s} = \hat{\mathbf{P}}\hat{\mathbf{P}}^{\mathbf{T}}\mathbf{x}$ have been obtained by MLPCA and there is a reference/normal data set $\mathbf{S} = \left\{\mathbf{s}_i | \mathbf{s}_i \in \mathbb{R}^d\right\}_{i=1}^{n}$ with the distribution of $p(\mathbf{s})$ and a test set $\tilde{\mathbf{S}} = \left\{\tilde{\mathbf{s}}_j | \tilde{\mathbf{s}}_j \in \mathbb{R}^d\right\}_{j=1}^{m}$ with the distribution of $\tilde{p}(\mathbf{s})$. The KL divergence between $p(\mathbf{s})$ and $\tilde{p}(\mathbf{s})$ can be expressed as

$$KL = \int \tilde{p}(\mathbf{s}) \log \frac{\tilde{p}(\mathbf{s})}{p(\mathbf{s})} \tag{1}$$

The KL divergence is equal to zero if and only if $p(\mathbf{s}) = \tilde{p}(\mathbf{s})$, hence we can introduce the following hypothesis test to decide whether the test data are sampled from the same distribution as the reference data

$H_0:$ $\mathbf{S}$ and $\tilde{\mathbf{S}}$ are sampled from the same distribution, such that $p(\mathbf{s}) = \tilde{p}(\mathbf{s})$;
$H_1:$ $\mathbf{S}$ and $\tilde{\mathbf{S}}$ are sampled from different distributions, such that $p(\mathbf{s}) \neq \tilde{p}(\mathbf{s})$.

The above hypothesis can be expressed in terms of KL divergence as

$$H_0: \ KL = 0;$$
$$H_1: \ KL \neq 0.$$

This hypothesis test, however, can only be used when the critical value of the null hypothesis can be obtained. For Gaussian distributed data, Refs. [21] showed that the KL divergence is related to the χ^2 when $\mathbf{\Sigma} = \mathbf{\Sigma}_{\tilde{s}}$. Here, $\mathbf{s} \sim \mathbb{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ and $\tilde{\mathbf{s}} \sim \mathbb{N}(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{\Sigma}})$, $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\tilde{s}}$ are the mean vector and $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{\tilde{s}}$ the covariance matrix of reference and test data. Ref.[22] also confirmed that the KL divergence follows a χ^2 distribution. However, none of them have investigated its distribution and properties under finite samples, i.e., how the distribution is related to sample number m. While in many practical settings, this is an important issue. Proposition 1 gives the distribution of KL divergence under finite samples.

Proposition 1: Under the assumption that $\mathbf{s} \sim \mathbb{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, $\tilde{\mathbf{s}} \sim \mathbb{N}(\boldsymbol{\mu}_{\tilde{\mathbf{s}}}, \mathbf{\Sigma}_{\tilde{\mathbf{s}}})$ and $m \to \infty$, the KL divergence between $p(\mathbf{s})$ and $\tilde{p}(\mathbf{s})$ asymptotically converges to a χ^2 distribution, i.e.,

$$\widehat{KL} \sim \frac{1}{2(m-1)} \chi^2_{2d} \tag{2}$$

In Proposition 1, the distribution of normal data is used to replace the true distribution as the true data generating distribution is generally unavailable. The larger the sample size n of the normal data set is, the closer the estimated distribution is from the true distribution. In most applications, a large number of data operating under normal condition would be collected and a relatively small number of faulty data can be obtained. Hence this replacement is reasonable.

With Proposition 1, the sensitivity of KL divergence can be compared with that of Hotelling's T^2 statistic. Again, consider the case of Gaussian variables. If there is no fault, the KL divergence between $p(\mathbf{s})$ and $\tilde{p}(\mathbf{s})$ will be inside the critical value determined in Proposition 1. For the case of faulty conditions, consider the additive fault in the following form

$$\tilde{\mathbf{s}} = \mathbf{s} + \Delta\mathbf{s} \tag{3}$$

where $\Delta\mathbf{s}$ is vector containing the fault magnitude. The fault only affects the mean value while the covariance remains unchanged

$$\boldsymbol{\mu}_{\tilde{\mathbf{s}}} = \boldsymbol{\mu} + \Delta\mathbf{s}, \quad \mathbf{\Sigma}_{\tilde{\mathbf{s}}} = \mathbf{\Sigma} \tag{4}$$

For Gaussian distributed data, the KL divergence between the distribution of normal data $p(\mathbf{s})$ and the distribution of faulty data $\tilde{p}(\mathbf{s})$ can be reduced to

$$KL(\mathbf{s}+\Delta\mathbf{s}) = \frac{1}{2}\Delta\mathbf{s}^T\boldsymbol{\Sigma}^{-1}\Delta\mathbf{s} \tag{5}$$

Define the following sensitivity index as

$$J = \frac{KL\left(\mathbf{s}+\Delta\mathbf{s}\right) - KL\left(\mathbf{s}\right)}{KL_{d,0.95}} \tag{6}$$

where $KL_{d,0.95}$ is the critical value of KL divergence under the 95% significance and $KL(\mathbf{s})$ is the KL divergence for $\Delta\mathbf{s} = \mathbf{0}$. Combining Proposition 1 and Eqn (6) we have

$$J = (m-1)\,\frac{\Delta\mathbf{s}^T\boldsymbol{\Sigma}^{-1}\Delta\mathbf{s}}{\chi^2_{2d,0.95}} \tag{7}$$

It can seen from Eqn (7) that the sensitivity index is determined by both the fault magnitude $\Delta\mathbf{s}$ and the number of test samples. By selecting a larger test set, higher sensitivity can be obtained. For comparison, the sensitivity index of T^2 statistic is also examined. T^2 statistic for the faulty condition can be expressed as

$$T^2\left(\mathbf{s}+\Delta\mathbf{s}\right) = \left(\mathbf{s}+\Delta\mathbf{s}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{s}+\Delta\mathbf{s}\right) \tag{8}$$

In contrast, the T^2 statistic for $\Delta\mathbf{s} = \mathbf{0}$ is

$$T^2\left(\mathbf{s}\right) = \mathbf{s}^T\boldsymbol{\Sigma}^{-1}\mathbf{s} \tag{9}$$

The sensitivity index for T^2 statistic can now be obtained as

$$J = \frac{\Delta\mathbf{s}^T\boldsymbol{\Sigma}^{-1}\Delta\mathbf{s}}{\chi^2_{d,0.95}} \tag{10}$$

Comparing Eqn (7) and Eqn (10) it can be seen that by including more test samples, the sensitivity of KL divergence will be much higher than that of the T^2 statistic.

III. Kernel based density ratio estimation

Provided the distribution functions of normal data and test data can be estimated, the monitoring problem using KL divergence reduces to examine the difference between distribution functions. However, estimation of probability density function is difficult, especially for high dimensional non-Gaussian data sets. To alleviate this difficulty, density ratio estimation methods [23] have been developed and explored in the field of machine learning. In this section, the least squares approach to density ratio estimation proposed by Kanamori *et al.* [14] is presented.

The core idea of the least squares density ratio estimation is to use a kernel function to approximate the density ratio between an i.i.d reference set $\mathbf{S} = \left\{\mathbf{s}_i|\mathbf{s}_i \in \mathbb{R}^d\right\}_{i=1}^n$ and test set $\tilde{\mathbf{S}} = \left\{\tilde{\mathbf{s}}_j|\tilde{\mathbf{s}}_j \in \mathbb{R}^d\right\}_{j=1}^m$ as follows

$$f(\mathbf{s}) = \frac{p(\mathbf{s})}{\tilde{p}(\mathbf{s})} \tag{11}$$

where $p(\mathbf{s})$ and $\tilde{p}(\mathbf{s})$ are the probability density functions of the reference and test set. The density ratio function $f(\mathbf{s})$ can be approximated using the following kernel model

$$\hat{f}(\mathbf{s}) = \sum_{i=1}^{n}\alpha_i K(\mathbf{s},\mathbf{s}_i) = \alpha^T\mathbf{k}(\mathbf{s}) \tag{12}$$

Here, $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]$ are unknown parameters to be estimated from data

$$\mathbf{k}\left(\mathbf{s}\right) = \left(K\left(\mathbf{s},\mathbf{s}_1\right), ..., K(\mathbf{s},\mathbf{s}_n)\right)^T \tag{13}$$

are kernel basis functions. Generally, the kernel is chosen as the Gaussian function

$$K\left(\mathbf{s},\mathbf{s}_i\right) = \exp\left(-\frac{\|\mathbf{s}-\mathbf{s}_i\|}{2\sigma^2}\right) \tag{14}$$

where σ denotes the width of Gaussian kernel.

The parameter α can be determined by minimizing the following squared error function

$$\begin{array}{l}\mathfrak{J}_0\left(\alpha\right) = \frac{1}{2}\int\left(f\left(\mathbf{s}\right) - \hat{f}\left(\mathbf{s}\right)\right)^2\tilde{p}\left(\mathbf{s}\right)d\mathbf{s} \\ = \frac{1}{2}\int\hat{f}^2\left(\mathbf{s}\right)\tilde{p}\left(\mathbf{s}\right)d\mathbf{s} - \int\hat{f}\left(\mathbf{s}\right)p\left(\mathbf{s}\right)d\mathbf{s} \\ + \frac{1}{2}\int f^2\left(\mathbf{s}\right)\tilde{p}\left(\mathbf{s}\right)d\mathbf{s} \\ = \frac{1}{2}\int\hat{f}^2\left(\mathbf{s}\right)\tilde{p}\left(\mathbf{s}\right)d\mathbf{s} - \int\hat{f}\left(\mathbf{s}\right)p\left(\mathbf{s}\right)d\mathbf{s} \\ + \frac{1}{2}\int f\left(\mathbf{s}\right)p\left(\mathbf{s}\right)d\mathbf{s}\end{array} \tag{15}$$

It should be noted that the last term is a constant and therefore can be ignored. Denote the first two terms by $\mathfrak{J}$ we have

$$\begin{array}{l}\mathfrak{J}(\alpha) = \frac{1}{2}\int\hat{f}^2\left(\mathbf{s}\right)\tilde{p}(\mathbf{s})d\mathbf{s} - \int\hat{f}\left(\mathbf{s}\right)p(\mathbf{s})d\mathbf{s} \\ = \frac{1}{2}\alpha^T\mathbf{H}\alpha - \mathbf{h}^T\alpha\end{array} \tag{16}$$

where $\mathbf{H}$ is the $n \times n$ matrix defined by

$$\mathbf{H} = \int\mathbf{k}\left(\mathbf{s}\right)\mathbf{k}\left(\mathbf{s}\right)^T\tilde{p}\left(\mathbf{s}\right)d\mathbf{s} \tag{17}$$

and $\mathbf{h}$ is the n-dimensional vector

$$\mathbf{h} = \int\mathbf{k}\left(\mathbf{s}\right)p\left(\mathbf{s}\right)d\mathbf{s} \tag{18}$$

The matrix $\mathbf{H}$ and vector $\mathbf{h}$ are expectations over density functions $p(\mathbf{s})$ and $\tilde{p}(\mathbf{s})$, which are unavailable. However, it is possible to approximate the expectations by their empirical averages as follows

$$\hat{\mathbf{H}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{k}\left(\tilde{\mathbf{s}}_j\right)\mathbf{k}\left(\tilde{\mathbf{s}}_j\right)^T \tag{19}$$

$$\tilde{\mathbf{h}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{k}\left(\mathbf{s}_i\right) \tag{20}$$

So that the cost function in Eqn (16) can be approximated by

$$\begin{array}{l}\hat{\mathfrak{J}}\left(\alpha\right) = \frac{1}{2m}\sum\limits_{j=1}^{m}\hat{f}^2\left(\mathbf{s}_j\right) - \sum\limits_{i=1}^{n}\hat{f}\left(\mathbf{s}_i\right) \\ = \frac{1}{2}\alpha^T\hat{\mathbf{H}}\alpha - \alpha^T\hat{\mathbf{h}}\end{array} \tag{21}$$

The optimization problem for the kernel based density ratio estimation can be formulated by including a regularization term as follows

$$\hat{\alpha} = \arg\min\left[\frac{1}{2}\alpha^T\hat{\mathbf{H}}\alpha - \alpha^T\hat{\mathbf{h}} + \frac{\lambda}{2}\alpha^T\alpha\right] \tag{22}$$

Here $\lambda(\geqslant 0)$ is the regularization parameter and $\frac{1}{2}\alpha^T\alpha$ is the regularization term. Differentiating over the above objective

function with respect to α and equating it to zero we can get the solution as

$$\hat{\alpha} = \left(\hat{\mathbf{H}} + \lambda \mathbf{I}_n\right)^{-1} \hat{\mathbf{h}} \tag{23}$$

where $\mathbf{I}_n$ is the identity matrix. Finally, the density ratio estimator can be obtained as

$$\hat{f}(\mathbf{s}) = \hat{\alpha}^T \mathbf{k}(\mathbf{s}) \tag{24}$$

With the analytic form expression, the kernel based density ratio estimator is computationally efficient and it is possible to select the kernel width σ and the regularization parameter λ by cross validation [14].

IV. Fault detection based on KL divergence

In this section, a fault detection scheme is formulated using the KL divergence. An empirical estimation of KL divergence using the density ratio estimator is proposed and the confidence limit of the KL divergence for non-Gaussian data is obtained by a moving window approach.

With the density ratio estimator, it is possible to measure the difference between the distributions of training data $\mathbf{S}$ and test data $\tilde{\mathbf{S}}$ using the KL divergence, which can be approximated by the following empirical average

$$\hat{KL} = -\int \tilde{p}(\mathbf{s}) \log \frac{p(\mathbf{s})}{\tilde{p}(\mathbf{s})} d\mathbf{s} = -\frac{1}{m}\sum_{j=1}^{m} \log \frac{p(\tilde{\mathbf{s}}_j)}{\tilde{p}(\tilde{\mathbf{s}}_j)} = -\frac{1}{m}\sum_{j=1}^{m} \log f(\tilde{\mathbf{s}}_j) \tag{25}$$

Replace the density ratio function $f(\mathbf{s})$ by the estimation in Eqn (24) the following KL divergence estimator can be obtained

$$\hat{KL} = -\frac{1}{m}\sum_{j=1}^{m} \log\left(\hat{\alpha}^T \mathbf{k}(\tilde{\mathbf{s}}_j)\right) \tag{26}$$

$\hat{KL}$ vanishes if and only if $p(\mathbf{s}) = \tilde{p}(\mathbf{s})$.

As is shown in Proposition 1, the sample distribution of KL divergence for Gaussian distributed data follows a χ^2 distribution and the confidence limits can be obtained easily for the monitoring task. For non-Gaussian data, the confidence limits in Proposition 1 cannot be used anymore. In this section, a window based procedure is developed to obtain the confidence limits for non-Gaussian data.

Assume there are L normal non-Gaussian data samples. The normal data is then divided into two data sets with equal length, i.e., $\frac{L}{2}$ samples. The first $\frac{L}{2}$ samples are used as the reference data set $\mathbf{S}_{ref}$. The validation set (with the size of m, $m \ll \frac{L}{2}$) is then constructed from the remaining $\frac{L}{2}$ samples (denoted to be $\bar{\mathbf{S}}$) as follows

1) Select the first m samples of the data set $\bar{\mathbf{S}}$ as the test set, denoted to be $\mathbf{S}_{test}$;
2) Obtain the estimation of KL divergence $\hat{KL}(\mathbf{S}_{ref}, \mathbf{S}_{test})$ between the reference data set and the test data set using the kernel based density ratio estimation;
3) Select the 2 to $m+1$ samples of the data set $\bar{\mathbf{S}}$ as the new test set $\bar{\mathbf{S}}_{test}$;
4) Compute the estimation of KL divergence $\hat{KL}(\mathbf{S}_{ref}, \bar{\mathbf{S}}_{test})$;
5) Repeat the above steps by including a new sample and discarding the oldest sample until $\frac{L}{2} - m + 1$ estimations of KL divergence are estimated.

Since the reference data and test data are assumed to be sampled from the same distribution, the KL divergence should be very close to zero. The sample distribution of the KL divergence under the null hypothesis that normal data are drawn from the same distribution can be constructed for the $\frac{L}{2} - m + 1$ estimates. The confidence limit can then be obtained by choosing the δ percentile of the $\frac{L}{2} - m + 1$ estimates of KL divergence.

With the confidence limit obtained, a window based strategy can then be developed for online monitoring.

1) For the test sample collected at time instance $L+1$, construct a test set as $\bar{\mathbf{S}}_{\text{test}} = \left\{\mathbf{s}_i | \mathbf{s}_i \in \mathbb{R}^d\right\}_{i=L-m+2}^{L+1}$;
2) Compute the estimation of KL divergence $\hat{KL}(\mathbf{S}, \bar{\mathbf{S}}_{test})$; compare it with the confidence limit. If the KL divergence exceeds the confidence limit, then a fault occurs;
3) For a new test sample, repeat the above steps by including the new sample and discarding the oldest sample to determine whether a faulty condition can be observed.

V. Simulation Example

In this section, a simulation example involving 5 input variables is considered. The 5 input variables are linear combination of 3 non-Gaussian sources, corrupted by Gaussian distributed noises with zero mean and variance of 0.04. The data generating model is as follows

$$\mathbf{x}_0 = \mathbf{A}\mathbf{s} + \mathbf{e} \tag{27}$$

where

$$\mathbf{A} = \begin{bmatrix} 0.15 & 0.50 & 0.35 \\ 0.50 & -0.62 & 0.46 \\ 0.30 & 0.50 & 0.10 \\ -0.25 & 0.55 & 0.20 \\ 0.32 & -0.18 & 0.24 \end{bmatrix} \tag{28}$$

The 3 non-Gaussian source signals are as follows

$$\begin{cases} s_1 = 0.2\cos(0.08k)\sin(0.06k) \\ s_2 = 1.5\sin k + 2\cos k \\ s_3 = \text{sign}\left(\sin(0.03k) + 9\cos(0.01k)\right) \end{cases} \tag{29}$$

The proposed method is compared with the PCA based local approach [24]. A total of 5000 samples are generated as the training data. MLPCA model is then determined using the 5000 samples and 3 principal components are extracted using the MLPCA. To test the proposed monitoring strategy based on KL divergence, a fault involving rotation of mixing matrix

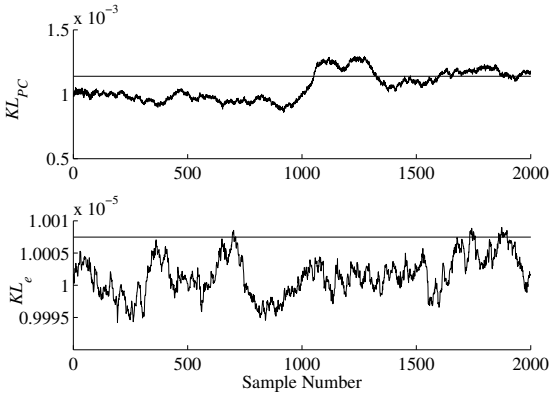

Fig. 1. Monitoring results for rotation of source variables using KL divergence

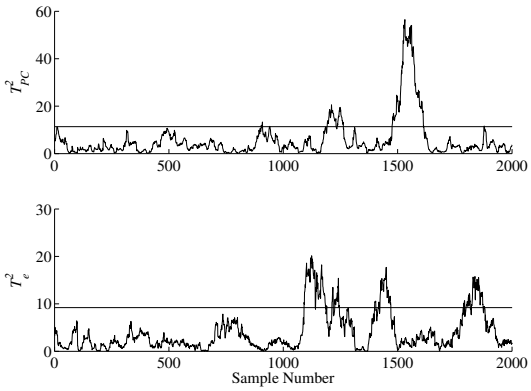

Fig. 2. Monitoring results for rotation of source variables using local PCA

by 15° to the first two non-Gaussian sources is considered as follows

$$\mathbf{A}_{rot} = \mathbf{A} \begin{bmatrix} 0.966 & -0.259 & 0 \\ 0.259 & 0.966 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{30}$$

A data set containing the fault with 2000 samples is considered, with the first 1000 samples under normal operational condition while the last 1000 samples are faulty data. For the proposed monitoring strategy, the first step would be to determine the window length, i.e., the length of test data set m. As is shown in Section 3.2, larger m value will increase the sensitivity of KL divergence, however, at the cost of larger delay for monitoring of faults. An appropriate window size is selected as $m = 100$. The first 2500 samples of the training set is selected as the reference set and the remaining 2500 samples are used to generate the confidence limit of the KL divergence. Two separate statistics are constructed, i.e., KL_{PC} and KL_e, where KL_{PC} is the test statistic constructed for the 3 retained principal components and KL_e is the test statistic for the residual space. The monitoring results for the rotation fault is shown in Figure 1. For comparison, Fig 2 presents the monitoring results using local PCA. For local PCA, the window length is also set as 100. Two separate statistics T^2_{PC} and T^2_e are constructed for the retained principal components and the residual space. It is shown in Figure 1 that rotation of the first 2 non-Gaussian sources can be detected by the KL_{PC} of the posed KL divergence based approach, however, the statistic KL_e is not sensitive to this fault. In contrast,both the T^2_{PC} and T^2_e statistics do not show significant violations. The applications to the rotation of source variables confirmed that the monitoring strategy based on KL divergence yields better monitoring results than local PCA.

VI. Application Study

This section presents the comparison results between the monitoring strategy based on KL divergence and the local PCA on an industrial melter process. The industrial melter process is part of a a disposal procedure that clad waste radioactive material in the form of powder by molten glass. The powder continuously fill in the vessel and the raw material in the form of glass frit is introduced every 120 seconds. This composition is then heated by four induction coils, positioned around the vessel. During the heating procedure, the glass becomes molten homogeneously. The filling and heating procedure results in an increased liquid column. The process continuous until a desired height of the liquid column is achieved. At this stage, the molten mixture is poured out through an exit funnel until the vessel is emptied. After that the next cycle of filling and heating begins. A detailed description of this process can be found in Liu *et al.*[8]

In this section, the recorded data collected from a melter unit is used to test the proposed monitoring strategy. The data set include 8 temperature variables, denoted by $T_1, ..., T_8$, the power in the induction coils, denoted by $P_1, ..., P_4$, the viscosity of the molten glass P_5 and the voltage supplied to the induction coils, denoted by P_6. The collected variable data exhibit strong non-Gaussian behaviors, similarly to the observation in Liu *et al.*[8] The data is collected at an interval of 5 minutes. Similar to the work in Ge *et al.*[21], a linear relationship between the source signals and the recorded signals are assumed. A data set with 1050 samples (87.5h) are considered, where a developing crack was observed in the melter vessel. This data set included normal process data over a period of 60h and a faulty condition resulting from a developing crack, which is manifested in several sensor readings. The last 300 samples corresponds to the developing crack. Application of MLPCA to the normal data reveals that 4 principal components are sufficient to extract more than 95% variance. Again, a window length of 100 is adopted and the monitoring strategy based on KL divergence is used. Figure 3 gives the results. It is shown in Figure 3 that the fault is successfully detected by both the KL_{PC} and KL_e statistics.

For comparison, Figure 4 presents the monitoring results using local PCA, which shows that the fault is only detected at the later stage. This further shows the superiority of our monitoring strategy based on KL divergence.

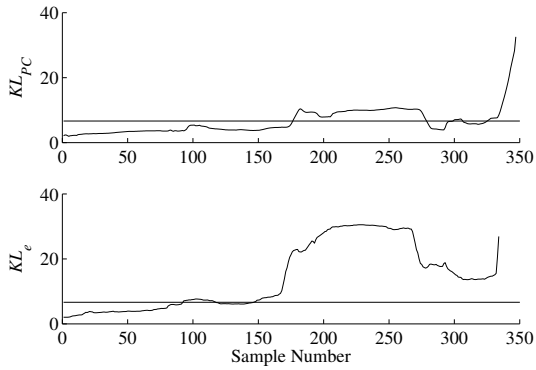

Fig. 3. Monitoring results for the developing crack using KL divergence

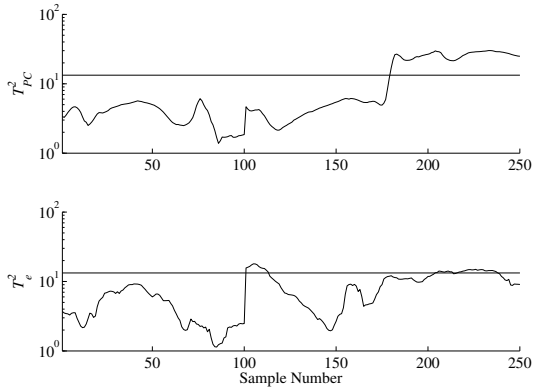

Fig. 4. Monitoring results for the developing crack using local PCA

VII. Conclusions

In this paper, a monitoring strategy based on KL divergence is developed for fault detection of industrial systems. Maximum likelihood PCA is firstly used to reduce the dimension of the original data and encapsulate the non-Gaussian information. Statistical properties of KL divergence as a monitoring statistic is investigated by using Monte Carlo simulations. It is found that KL divergence between two Gaussian data sets follows a χ^2 distribution and a confidence limit can be easily obtained. The sensitivity of KL divergence is then compared with that of T^2 statistic. The sensitivity of KL divergence increases as the number of test samples increases. A window based monitoring strategy is then developed for non-Gaussian distributed variables and the monitoring strategy is then test on fault monitoring of a simulation example and an industrial melter process. The monitoring strategy based on KL divergence is then compared with that of local PCA and the results show that the developed monitoring strategy has better sensitivity and less false alarms. Comparing the monitoring methods in the literature, this method is conceptually more straightforward.

References

[1] M.E. Tipping and C.M. Bishop, Probabilistic principal component analysis. *Journal of Royal Statistical Society Series B*, vol. 61, 1999, pp 611∼622.

[2] Edgar T. F. Dunia, R. and T. J. McAvoy. Identification of faulty sensors using principal component analysis. *AIChE Journal*, vol. 42, 1996, pp 2797∼2812.

[3] S. J. Qin. Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, vol. 17, 2003, pp 480∼502, 2003.

[4] Li G. Zhou, D. G. and S. J. Qin. Total projection to latent structures for process monitoring. *AIChE Journal*, vol. 55, 2009, pp 168∼178.

[5] Andrews D. T. Hamilton D. C. Faber K.Wentzell, P. D. and B. R. Kowalski. Maximum likelihood principal component analysis. *Journal of Chemometrics*, vol. 11, 1997, pp 339∼ 366.

[6] S. Narasimhan and S. L. Shah. Model identification and error covariance matrix estimation from noisy data using pca. *Control Engineering Practice*, vol. 16, 2008, pp 146∼155.

[7] T. Feital, L. Kruger, U. amd Xie, U. Schubert, E. L. Lima, and J. C. Pinto. A unified statistical framework for monitoring multivariate systems with un-known source and error signals. *Chemometrics and Intelligent Laboratory Systems*, vol. 104, 2010, pp 223∼232.

[8] Liu, X. Q., Xie L., Kruger U. Littler T. and S. Q. Wang. Statistical-based monitoring of multivariate non-gaussian systems. *AIChE Journal*, vol. 54, 2008, pp 2379∼2391.

[9] J. M. Lee, C. K. Yoo, and I. B. Lee. Statistical process monitoring with independent component analysis. *Journal of Process Control*, vol.14, 2004, pp 467∼485.

[10] J. M. Lee, S. J. Qin, and I. B. Lee. Fault detection and diagnosis based on modified independent component analysis. *AIChE Journal*, vol. 52, 2006, pp 3501∼3514.

[11] Q. Chen, U. Kruger, and A. Y. T. Leung. Regularized kernel density estimation for clustered process data. *Control Engineering Practice*, vol. 12, 2004, pp 267∼274.

[12] M. Basseville. On-board component fault detection and isolation using the statistical local approach. *Automatica*, vol.34, 1998, pp 1391∼1415.

[13] U. Kruger and G. Dimitriadis. Diagnosis of process faults in chemical systems using the local partial least squares approach. *AIChE Journal*, vol. 54, 2008, pp 2581∼2596.

[14] S. Kanamori, S. Hido, and M. Sugiyama. A least squares approach to direct importance estimation. *Journal of Machine Learning Research*, vol. 10, 2009, pp 1391∼1445.

[15] M. Sugiyama, T. Suzuki, Y. Itoh, and T. Kanamori. Least-squares two-sample test. *Neural Networks*, vol. 24, 2011, pp 735∼751.

[16] F. Perronnin and J. Llados, and G. Sanchez. A similarity measure between vector sequences with application to handwritten word image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1722∼1729, 2009.

[17] P. A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 317∼320, 2007.

[18] J. Inglada. Change detection on sar images by using a parametric estimation of the kullback-leibler divergence. In *Proceedings of IEEE International Symposium on Geoscience and Remote Sensing*, pages 4104∼4106, 2003.

[19] G. E. P. Box and W. G. Hunter and S. J. Hunter. *Statistics for experimenters*. New York, NY: John Wiley & Sons, Inc, 1978.

[20] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Applications*. Prentice Hall, London, 1993.

[21] T. Dasu and S. Krishnan and S. Venkatasubramanian, and K. Yi. An information-theoretic approach to detecting changes in multi-dimensional data streams. In *Proceedings of Symposium on the Interface of Statistics, Computing Science and Applications*, pages 172∼179, 2006.

[22] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall, London, 2004.

[23] S. Bickel and T. Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In *Proceedings of Advances in Neural Information Processing Systems*, 2007, pages 161∼168.

[24] U. Kruger and S. Kumar and T. Littler. Improved principal component monitoring using the local approach. *Automatica*,vol. 43, 2007, pp 1532 ∼ 1542.