Probabilistic Contribution Analysis for Statistical Process Monitoring: A Missing Variable Approach

Tao Chen^{a*}; Yue Sun^b

^a*School of Chemical and Biomedical Engineering,
Nanyang Technological University, Singapore 637459*

^b*College of Automation, Chongqing University, Chongqing 400044, China*

Abstract: Probabilistic models, including probabilistic principal component analysis (PPCA) and PPCA mixture models, have been successfully applied to statistical process monitoring. This paper reviews these two models and discusses some implementation issues that provide alternative perspective on their application to process monitoring. Then a probabilistic contribution analysis method, based on the concept of missing variable, is proposed to facilitate the diagnosis of the source behind the detected process faults. The contribution analysis technique is demonstrated through its application to both PPCA and PPCA mixture models for the monitoring of two industrial processes. The results suggest that the proposed method in conjunction with PPCA model can reduce the ambiguity with regard to identifying the process variables that contribute to process faults. More importantly it provides a fault identification approach for PPCA mixture model where conventional contribution analysis is not applicable.

Key words: Contribution analysis, missing variable, mixture model, multivariate statistical process monitoring, probabilistic principal component analysis.

*Corresponding author. Email: chentao@ntu.edu.sg; Tel.: +65 6513 8267; Fax: +65 6794 7553.

1 Introduction

Multivariate statistical process monitoring (MSPM) is a critical approach to the safe and efficient operation of a wide range of manufacturing and chemical processing plants. The basis of MSPM is the development of a statistical model from normal operating data, followed by the calculation of monitoring statistics (e.g. Hotelling’s T^2 and squared prediction error (SPE)) and their confidence bounds. Then process faults can be detected when the on-line monitoring statistics exceed the confidence bounds. In the past decades, extensive research on process monitoring has resulted in the application of multivariate statistical projection approaches, notably principal component analysis (PCA) (Wold et al., 1987), partial least squares (PLS) (Geladi and Kowalski, 1986) and their non-linear and dynamic variants (Ku et al., 1995; Wilson et al., 1999; Zhao et al., 2006), to extract relevant process information, and to attain an enhanced understanding of process behavior.

The PCA/PLS based MSPM techniques assume that the scores are multivariate Gaussian distributed, which could be invalid particularly when data are collected from a complex process (Thissen et al., 2005), or when non-linear projection techniques are utilized (Wilson et al., 1999). This issue can be addressed through semi-parametric and non-parametric statistical methods, including kernel density estimation (Martin and Morris, 1996), wavelet-based density estimation (Safavi et al., 1997), adaptive local modeling (Ge and Song, 2008), and the Gaussian mixture model (GMM) (Choi et al., 2004; Thissen et al., 2005; Chen et al., 2006). The other interesting research topic is that of formulating PCA in a probabilistic framework, resulting in probabilistic PCA (PPCA) model (Tipping and Bishop, 1999b; Kim and Lee, 2003). The probabilistic models provide a unified likelihood-based statistic for process monitoring as opposed to multiple control charts of T^2 and SPE (Chen et al., 2006). In addition probabilistic models can handle missing variables in a systematic way (Tipping and Bishop, 1999b). Furthermore, the PPCA has

been extended to the PPCA mixture model (Tipping and Bishop, 1999a) to account for non-Gaussian distributions, and this PPCA mixture model has been applied to process monitoring that demonstrated improved fault detection results (Choi et al., 2005).

In practice, process monitoring is usually followed by contribution analysis, which investigates the contribution of individual variable to the T^2 and/or SPE that violate the confidence bounds (Miller et al., 1998). Although contribution analysis does not explicitly indicate the mode of the fault, it does provide important information as to which variables are the most responsible for the MSPM system to raise alarm. However since both the T^2 and SPE are monitored in conventional MSPM, contribution analysis is typically performed for these two statistics (e.g. (Shao et al., 1999)). There is no consensus in the literature regarding how to resolve the conflicts if the two contribution analysis procedures identify different sets of responsible faulty variables. One approach is to examine SPE first (e.g. (Nelson et al., 2006)): a large SPE indicates that the data is far from the score space and thus the T^2 , calculated from the PCA scores, should not be used for contribution analysis. Arguably this approach ignores the information contained in the PCA space that could be helpful in fault diagnosis.

An alternative route is to use a single monitoring statistic and thus a single contribution analysis. Several researchers have attempted to combine the T^2 and SPE in some way (Chen et al., 2004; Yue and Qin, 2001), whilst the PPCA and PPCA mixture models naturally provide the single likelihood-based statistic. However to the best knowledge of the authors, a unified contribution analysis method for these probabilistic models has not been explored in the literature.

The primary focus of this paper is thus to develop a fully probabilistic framework for contribution analysis to be used with PPCA and PPCA mixture models. Section 2 gives an overview of the PCA and PPCA mixture models with a discussion on some im-

plementation issues, including model selection and the construction of confidence bounds based on a single statistic, the likelihood value. Then contribution analysis is developed in Section 3 through the methodology of missing variables. This is motivated by the reconstruction-based fault identification method (Yue and Qin, 2001; Dunia et al., 1996) that was proposed within the conventional PCA framework. In Section 4, the proposed contribution analysis methodology is demonstrated through its application to both PPCA and PPCA mixture models for the monitoring of two industrial processes. Finally Section 5 concludes this paper.

2 PPCA and mixture models for process monitoring

2.1 PPCA and mixture models

Principal component analysis (PCA) is a general multivariate statistical projection technique for dimension reduction. The central idea of PCA is to project the original D dimensional data, $\mathbf{x}$, onto a space where the variance is maximized: $\mathbf{x} = \mathbf{W}\mathbf{t} + \boldsymbol{\mu} + \mathbf{e}$. Here $\mathbf{W}$ refers to the eigenvectors of the sample covariance matrix corresponding to the Q ($Q \leq D$) largest eigenvalues, $\mathbf{t}$ is the Q dimensional scores, $\boldsymbol{\mu}$ is the mean of the data, and $\mathbf{e}$ is the noise term.

More recently Tipping and Bishop (1999b) proposed a probabilistic formulation of PCA (PPCA) from the perspective of a Gaussian latent variable model. Specifically the noise is assumed to be Gaussian: $\mathbf{e} \sim G(\mathbf{0}, \sigma^2 \mathbf{I})$, which implies $\mathbf{x}|\mathbf{t} \sim G(\mathbf{W}\mathbf{t} + \boldsymbol{\mu}, \sigma^2 \mathbf{I})$. Furthermore, by adopting a Gaussian distribution for the scores, $\mathbf{t} \sim G(\mathbf{0}, \mathbf{I})$, the marginal distribution of the data is also Gaussian: $\mathbf{x} \sim G(\boldsymbol{\mu}, \mathbf{C})$, where the covariance matrix is $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}$. In contrast to the sample covariance $\mathbf{S}$, the PPCA model defines $\mathbf{C}$ in terms of the auxiliary parameters $\mathbf{W}$ and σ^2 . Note the number of free parameters in $\mathbf{C}$ is $DQ + 1 - Q(Q - 1)/2$, which is smaller than the $D(D + 1)/2$ parameters in $\mathbf{S}$

if $Q < D$ (Tipping and Bishop, 1999b). Therefore PPCA provides a way to constrain the model complexity via the selection of Q . The model parameters, $\{\boldsymbol{\mu}, \mathbf{W}, \sigma^2\}$, can be estimated using the maximum likelihood algorithm implemented through either eigen-decomposition or the expectation-maximization (EM) algorithm; see (Tipping and Bishop, 1999b) for details.

The probabilistic formulation offers the potential to extend the scope of conventional PCA. For example, the issue of missing values in the data set, which is a common situation in practice, can be systematically addressed within the model through the EM algorithm. Furthermore, multiple PPCA models can be combined to form a mixture model to characterize the local properties of the data set.

Specifically a PPCA mixture model with K mixture components is formulated as (Tipping and Bishop, 1999a):

$$p(\mathbf{x}) = \sum_{k=1}^K \alpha_k G(\boldsymbol{\mu}_k, \mathbf{C}_k) \quad (1)$$

Within each mixture a local PPCA model is adopted: $\mathbf{x}|k \sim G(\boldsymbol{\mu}_k, \mathbf{C}_k)$, where $\mathbf{C}_k = \mathbf{W}_k \mathbf{W}_k^T + \sigma_k^2 \mathbf{I}$. The sum of the mixing weights, $\{\alpha_k, k = 1, \dots, K\}$, must be equal to unity. Here $\mathbf{W}_k$ is a loading matrix of order $D \times Q_k$, reflecting the fact that the number of principal components, Q_k , can vary for different mixtures. The maximum likelihood estimation of the model parameters, $\{\boldsymbol{\mu}_k, \mathbf{W}_k, \sigma_k^2, \alpha_k; k = 1, \dots, K\}$, can be implemented through the EM algorithm as presented in (Tipping and Bishop, 1999a).

The major motivation to introduce mixture models in MSPM is to monitor the processes with multiple operating modes (Choi et al., 2004, 2005; Ge and Song, 2008). Indeed, as a general semi-parametric approach to probability density estimation, mixture modeling has a broad spectrum of applications in statistical process monitoring, especially when the process data cannot be accurately approximated by a global Gaussian distribution

(Thissen et al., 2005; Chen et al., 2006).

2.2 Model selection

For a PPCA model, model selection refers to deciding the number of principal components (PCs) to be retained. According to the criterion of variance ratio, the retained PCs should explain in excess of say, 90%, of the total variance in the data (Musa et al., 2004). Alternatively the Bayesian information criterion (BIC) (Schwarz, 1978) can be used to select the model for which $L - (H/2) \log N$ is the largest, where L is the log-likelihood of the data, H and N are the number of parameters and training data points, respectively. The motivation of BIC is that a good model should be able to sufficiently explain the data (the log-likelihood) with low model complexity (the number of parameters). In this study the criterion of variance ratio is adopted for the selection of principal components for PPCA models.

The issue of model selection is more complex for PPCA mixture model, since both the number of mixtures, and the number of PCs within each mixture, require to be determined. The major difficulty is the large number of candidate models to be considered. If the appropriate number of mixtures is considered to be between one and $K_{\max}$, for a D dimensional data set, the total number of candidate models is $D(D^{K_{\max}} - 1)/(D - 1)$ if $D \neq 1$, otherwise it is equal to $K_{\max}$. As an example 3,905 models must be evaluated in terms of some optimality criterion (such as BIC) even for small numbers: $D = K_{\max} = 5$. Therefore an exhaustive search over all possible candidates is infeasible, and some sub-optimal algorithms are required. Choi et al. (2005) developed a resolution-based framework in conjunction with a Bayesian complexity measure for model selection; however the resolution-based method still requires high computation. In (Kim et al., 2003) a fast model selection algorithm was proposed with the restrictive assumption that all mixtures

have the same number of PCs (Q_k). To allow different Q_k 's while attaining computational efficiency, Musa et al. (2004) proposed to apply the criterion of variance ratio for the selection of PCs within each mixture, and developed a greedy training algorithm to determine the number of mixtures. Despite the fact that this two-stage algorithm is sub-optimal, it was shown (Musa et al., 2004) to achieve promising performance in practice.

Motivated by Musa et al. (2004), this paper adopts the criterion of variance ratio to determine Q_k within each mixture, which is a very fast procedure, and then apply BIC to select the appropriate number of mixtures. Specifically the model selection algorithm for PPCA mixture models is as follows.

1. For $k = 1 : K_{\max}$
 - Estimate the model parameters, $\{\boldsymbol{\mu}_j, \mathbf{W}_j, \sigma_j^2, \alpha_j\}$, $j = 1, \dots, k$, using the EM algorithm. At this stage D principal components are retained within each mixture.
 - Within each mixture, select Q_j principal components such that they explain in excess of $100\gamma\%$ of the total variance. Then retain the corresponding Q_j column vectors in $\mathbf{W}_j$.
 - Calculate the BIC value of current model with k mixtures.

End

2. Select K such that the model has the largest BIC value.

This is also a sub-optimal model selection algorithm where the compromise has been made to restrict the computational cost. The variance ratio $100\gamma\%$ is taken to be 90% according to the literature of statistical process monitoring.

2.3 Confidence bounds for process monitoring

One of the advantages of the probabilistic models for process monitoring is that they provide confidence bounds based on a single statistic, the likelihood value, to detect process faults, as opposed to the confidence bounds for T^2 and SPE in conventional methods. In practice a single monitoring statistic will reduce the work load of data analyst and plant operators as they will only be exposed to one monitoring chart. This is crucial for the wider acceptance of the process monitoring techniques in industry (Chen et al., 2006).

On the basis of the probability density function $p(\mathbf{x})$ for normal operating data, the $100\beta\%$ confidence bound is defined as a likelihood threshold h that satisfies the following integral (Chen et al., 2006):

$$\int_{\mathbf{x}:p(\mathbf{x})>h} p(\mathbf{x})d\mathbf{x} = \beta \quad (2)$$

For a PPCA model, $p(\mathbf{x})$ is a multivariate Gaussian distribution and the above confidence bounds are readily available (see e.g. (Berger, 1985, Chapter 4.3.2)). In particular the data $\mathbf{x}$ is considered as out-of-control when

$$M^2 = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}) > \chi_D^2(\beta) \quad (3)$$

where $\chi_D^2(\beta)$ is the β -fractile of the chi-square distribution with degree of freedom D .

When $p(\mathbf{x})$ is a mixture of Gaussians as in the PPCA mixture model, the integral in Eq. (2) can be calculated conveniently using numerical methods such as the following Monte Carlo approach (Chen et al., 2006):

1. Generate N_s random samples, $\{\mathbf{x}^j, j = 1, \dots, N_s\}$, from $p(\mathbf{x})$.
2. Calculate the likelihood of these samples as $p(\mathbf{x}^j)$.
3. Sort $p(\mathbf{x}^j), j = 1, \dots, N_s$, in descending order.

4. The confidence bound is given by $h = p(\mathbf{x}^l)$, where $l = N_s\beta$.

Then the data point $\mathbf{x}$ is considered to be faulty if $p(\mathbf{x}) < h$. Typically the confidence bound is expressed as a upper bound. Hence the process fault is detected if the negative likelihood exceeds the confidence bound: $-p(\mathbf{x}) > -h$.

The algorithm to generate random samples from a mixture of Gaussians can be found in (Bishop, 2006, Chapter 9.2). The number of Monte Carlo samples required (N_s) to approximate the confidence bounds can be determined heuristically. Alternatively if the nominal process data set used to train the PPCA mixture model is sufficiently large, the confidence bounds can be calculated based on the above procedure by replacing the Monte Carlo samples with the training data points.

3 Missing variable based contribution analysis

The objective of contribution analysis is to identify which variables are the most responsible for the occurrence of the process faults. In general contribution analysis may not explicitly reveal the root-cause of the onset of faults, but it is undoubtedly helpful in pinpointing the inconsistent variables that should undergo further diagnosis procedures. The contribution analysis for PCA model is to decompose T^2 and SPE into the sum of D contribution terms for the variables, and then the magnitude of these terms indicates the relative responsibility of the variables (Miller et al., 1998). A similar method was developed for PPCA model (Kim and Lee, 2003). Alternatively reconstruction-based methods have been proposed (Yue and Qin, 2001; Dunia et al., 1996), where each variable is treated as if it were missing and is reconstructed in turn, and the variables corresponding to the largest reconstruction errors are considered to contribute the most to the occurrence of process faults.

This paper extends the reconstruction methods in (Yue and Qin, 2001; Dunia et al.,

1996) to develop a missing variable based contribution analysis in a probabilistic framework. For PPCA model, the proposed approach only requires to analyze the contribution to a single likelihood-based statistic, as opposed to investigating both T^2 and SPE. More importantly, the probabilistic framework fills a gap to provide a contribution analysis method for PPCA mixture model, a topic that has not been explored in the literature.

Specifically assume the data $\mathbf{x}$ is identified as faulty because its monitoring statistic exceeds the confidence bound. Then each time one variable of $\mathbf{x}$ is regarded as missing. Let x_d ($d = 1, \dots, D$) be the missing variable, and $\mathbf{x}_{-d}$ be the vector of other observed variables in $\mathbf{x}$. For $d = 1 : D$, the proposed contribution analysis executes as follows:

1. Calculate the conditional distribution of the missing variable given the other variables, $p(x_d|\mathbf{x}_{-d})$. The conditional distribution depends on the choice of models and will be discussed in Section 3.1 (PPCA model) and Section 3.2 (PPCA mixture model).
2. Re-calculate the monitoring statistic (M^2 in PPCA and negative likelihood in PPCA mixture model) with the d -th variable being missing. This step is to calculate the expected monitoring statistic with respect to $p(x_d|\mathbf{x}_{-d})$.
3. If the d -th variable contributes significantly to the data being detected as faulty, then the re-calculated statistic will be much smaller than the original monitoring statistic. Therefore the difference between the original and re-calculated statistics is an indication of the responsibility of the d -th variable for the fault.

The above procedure is different from the algorithms proposed in (Yue and Qin, 2001; Dunia et al., 1996) in that the missing variables are not *reconstructed*. Instead a probabilistic method is developed to account for the missing variables when re-calculating the monitoring statistic.

The rest of this section discusses how the conditional distribution of the missing variables is developed and utilized for the calculation of the monitoring statistics using the PPCA and PPCA mixture models. Although this paper is primarily concerned with the contribution analysis of individual variables, the discussion takes a more general perspective where there may be multiple missing variables. Therefore the developed framework will be applicable for the analysis of the contribution of a group of variables by treating these variables as missing. To facilitate the general discussion, the following notation is adopted: $\mathbf{x}$ denotes the original data point whilst $\tilde{\mathbf{x}}$ is the corresponding data with some missing values. Furthermore $\tilde{\mathbf{x}}$ can be divided as $\tilde{\mathbf{x}}^T = [\tilde{\mathbf{x}}_o^T, \tilde{\mathbf{x}}_m^T]$, where $\tilde{\mathbf{x}}_o$ and $\tilde{\mathbf{x}}_m$ are sub-vectors of observed and missing variables respectively.

3.1 Missing variables in PPCA model

Given a PPCA model, the probability distribution of $\tilde{\mathbf{x}}$ is Gaussian: $\tilde{\mathbf{x}} \sim G(\boldsymbol{\mu}, \mathbf{C})$. For ease of derivation the mean and covariance are organized into blocks:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_m \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_{oo} & \mathbf{C}_{om} \\ \mathbf{C}_{mo} & \mathbf{C}_{mm} \end{bmatrix}. \quad (4)$$

Hence the conditional distribution of $\tilde{\mathbf{x}}_m$ given $\tilde{\mathbf{x}}_o$ is also Gaussian with mean and covariance matrix being $\mathbf{z}_m = \boldsymbol{\mu}_m + \mathbf{C}_{mo}\mathbf{C}_{oo}^{-1}(\tilde{\mathbf{x}}_o - \boldsymbol{\mu}_o)$ and $\mathbf{Q}_m = \mathbf{C}_{mm} - \mathbf{C}_{mo}\mathbf{C}_{oo}^{-1}\mathbf{C}_{om}$ respectively, i.e. $\tilde{\mathbf{x}}_m|\tilde{\mathbf{x}}_o \sim G(\mathbf{z}_m, \mathbf{Q}_m)$. To calculate the M^2 monitoring statistic in Eq. (3) in the presence of missing values, it is convenient to utilize the conditional distribution of the complete vector $\tilde{\mathbf{x}}$, which is again a Gaussian: $\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o \sim G(\mathbf{z}, \mathbf{Q})$, where

$$\mathbf{z} = \begin{bmatrix} \tilde{\mathbf{x}}_o \\ \mathbf{z}_m \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_m \end{bmatrix}. \quad (5)$$

Therefore the expectation of M^2 statistic with respect to $\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o \sim G(\mathbf{z}, \mathbf{Q})$ can be calculated

as follows:

$$\begin{aligned}
E[M^2] &= E[(\tilde{\mathbf{x}} - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\tilde{\mathbf{x}} - \boldsymbol{\mu})] \\
&= \text{Tr}[\mathbf{C}^{-1} \{(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^T + \mathbf{Q}\}].
\end{aligned} \tag{6}$$

Note if the missing variables are simply replaced by the mean $\mathbf{z}_m$ to calculate the M^2 statistic, the uncertainty of $\tilde{\mathbf{x}}$ in terms of the covariance matrix $\mathbf{Q}$ is ignored. The ignorance of the uncertainty may result in a monitoring statistic smaller than it should be, and thus lead to incorrect contribution analysis.

The contribution of the variables that are treated as missing can be quantified as $M^2 - E[M^2]$. A large value of $M^2 - E[M^2]$ indicates that by eliminating the variables $\tilde{\mathbf{x}}_m$, the monitoring statistic is considerably decreased; in other words these variables are responsible for the fault. Furthermore if $E[M^2]$ is even smaller than the confidence bound $\chi_D^2(\beta)$ (or equivalently $M^2 - E[M^2] > M^2 - \chi_D^2(\beta)$), the process operators will be very confident about the responsible variables, since the elimination of them would bring the process back to normal operating region.

3.2 Missing variables in PPCA mixture model

Similarly the conditional distribution of the data with missing variables, $\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o$, can be derived under the PPCA mixture model. By definition the conditional distribution is

$$p(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o) = \frac{p(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_o)}{p(\tilde{\mathbf{x}}_o)} = \frac{p(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_o)}{\int p(\tilde{\mathbf{x}}_m, \tilde{\mathbf{x}}_o) d\tilde{\mathbf{x}}_m}. \tag{7}$$

Under the PPCA mixture model $p(\tilde{\mathbf{x}}_m, \tilde{\mathbf{x}}_o)$ (and also $p(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_o)$ since $\tilde{\mathbf{x}}^T = [\tilde{\mathbf{x}}_o^T, \tilde{\mathbf{x}}_m^T]$) is a Gaussian mixture model as in Eq. (1). Therefore

$$\begin{aligned}
p(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o) &= \frac{\sum_{k=1}^K \alpha_k p(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_o | \boldsymbol{\mu}_k, \mathbf{C}_k)}{\sum_{k=1}^K \alpha_k \int p(\tilde{\mathbf{x}}_m, \tilde{\mathbf{x}}_o | \boldsymbol{\mu}_k, \mathbf{C}_k) d\tilde{\mathbf{x}}_m} \\
&= \frac{\sum_{k=1}^K \alpha_k p(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o, \boldsymbol{\mu}_k, \mathbf{C}_k) p(\tilde{\mathbf{x}}_o | \boldsymbol{\mu}_k, \mathbf{C}_k)}{\sum_{k=1}^K \alpha_k p(\tilde{\mathbf{x}}_o | \boldsymbol{\mu}_k, \mathbf{C}_k)}
\end{aligned} \tag{8}$$

where $p(\tilde{\mathbf{x}}_o | \boldsymbol{\mu}_k, \mathbf{C}_k)$ and $p(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o, \boldsymbol{\mu}_k, \mathbf{C}_k)$ are the marginal probability of $\tilde{\mathbf{x}}_o$ and the conditional distribution of $\tilde{\mathbf{x}}$ under the k -th mixture component, respectively. Since each mixture component is a Gaussian distribution, $p(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o, \boldsymbol{\mu}_k, \mathbf{C}_k)$ is also a Gaussian and it can be obtained by following the procedure in Section 3.1 for the PPCA model, where $\boldsymbol{\mu}$ and $\mathbf{C}$ are replaced by $\boldsymbol{\mu}_k$ and $\mathbf{C}_k$ respectively.

By introducing a new set of weights $\eta_k, k = 1, \dots, K$:

$$\eta_k = \frac{\alpha_k p(\tilde{\mathbf{x}}_o | \boldsymbol{\mu}_k, \mathbf{C}_k)}{\sum_{k=1}^K \alpha_k p(\tilde{\mathbf{x}}_o | \boldsymbol{\mu}_k, \mathbf{C}_k)}, \tag{9}$$

the conditional distribution of $\tilde{\mathbf{x}}$ can then be re-written as

$$p(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o) = \sum_{k=1}^K \eta_k p(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o, \boldsymbol{\mu}_k, \mathbf{C}_k) = \sum_{k=1}^K \eta_k p(\tilde{\mathbf{x}}|\mathbf{z}_k, \mathbf{Q}_k) \tag{10}$$

where $\mathbf{z}_k$ and $\mathbf{Q}_k$ are the mean and covariance of the k -th Gaussian distribution that are calculated as in Eq. (5). Eq. (10) clearly shows that the conditional distribution is still a mixture of Gaussian distributions.

For the purpose of contribution analysis, it is required to calculate the expectation of the monitoring statistic, the likelihood value of $\tilde{\mathbf{x}}$ under the PPCA mixture model, with respect to $\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o$. The expectation can be expressed as:

$$\begin{aligned}
E[p(\tilde{\mathbf{x}})] &= \int p(\tilde{\mathbf{x}}) p(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_o) d\tilde{\mathbf{x}} \\
&= \sum_{k=1}^K \sum_{j=1}^K \alpha_k \eta_j \int p(\tilde{\mathbf{x}}|\boldsymbol{\mu}_k, \mathbf{C}_k) p(\tilde{\mathbf{x}}|\mathbf{z}_j, \mathbf{Q}_j) d\tilde{\mathbf{x}}
\end{aligned} \tag{11}$$

The above integral can be analytically calculated by noting that the two integrands are both Gaussian distributions. The multiplication of two Gaussians (Quiñonero-Candela et al., 2003) is proportional to a Gaussian distribution:

$$G(\boldsymbol{\mu}_k, \mathbf{C}_k)G(\mathbf{z}_j, \mathbf{Q}_j) = c_{kj} G(\mathbf{p}, \mathbf{P}) \quad (12)$$

where $\mathbf{P} = (\mathbf{C}_k^{-1} + \mathbf{Q}_j^{-1})^{-1}$, $\mathbf{p} = \mathbf{P}(\mathbf{C}_k^{-1}\boldsymbol{\mu}_k + \mathbf{Q}_j^{-1}\mathbf{z}_j)$, and c_{kj} is the normalizing constant given by

$$c_{kj} = (2\pi)^{-\frac{D}{2}} |\mathbf{C}_k + \mathbf{Q}_j|^{-\frac{1}{2}} \cdot \exp \left[-\frac{1}{2}(\boldsymbol{\mu}_k - \mathbf{z}_j)^T (\mathbf{C}_k + \mathbf{Q}_j)^{-1} (\boldsymbol{\mu}_k - \mathbf{z}_j) \right]. \quad (13)$$

Then the expectation in Eq. (11) is obtained as:

$$E[p(\tilde{\mathbf{x}})] = \sum_{k=1}^K \sum_{j=1}^K \alpha_k \eta_j c_{kj} \quad (14)$$

Similar to the PPCA model, the contribution of the variable under investigation is quantified as $E[p(\tilde{\mathbf{x}})] - p(\mathbf{x})$, i.e. the increase in likelihood (equivalent to the decrease in the monitoring statistic that is the negative likelihood) if the variable is treated as missing. If $E[p(\tilde{\mathbf{x}})]$ is greater than the confidence bound h , the corresponding variable can be considered to contribute significantly to the detected process fault.

4 Case studies

This section first demonstrates the probabilistic contribution analysis on the slurry-fed ceramic melter data set (SFCM) (Wise et al., 1991) that is available from the PLS Toolbox developed by Eigenvector Research Inc. (<http://www.eigenvector.com/>). Then the proposed method is applied to the monitoring of an industrial propylene polymerization process.

Figure 1: Trends of the process measurements being monitored.

4.1 The slurry-fed ceramic melter process

In the SFCM process slurry is formed through the combination of nuclear waste from fuel reprocessing with glass forming materials. This slurry is fed into a high temperature glass melter, and then a stable vitrified product is produced for disposal in a long term repository. The process data consists of 20 temperature values at different locations within the melter, and the measured molten glass level, resulting in 21 variables in total. A training data set with 450 samples is utilized to model the nominal process condition, and then the model is used for the monitoring of a test set with 54 samples (Fig. 1). As a pre-processing step, the data set is auto-scaled to zero mean and unit variance within each variable.

First the conventional PCA is applied to the training data. To explain in excess of 90% of the total variance within the data, six principal components (PCs) are retained. The same number of PCs are retained in PPCA model based on the same criterion. According to previous analysis (Wise et al., 1991) the data can be well modeled by a single Gaussian distribution, and thus mixture model is not considered.

Figure 2: Monitoring charts from PCA model with 99% (dash-dotted line) confidence bound.

The process monitoring charts for PCA and PPCA models are shown in Fig. 2 and 3 respectively, where 99% confidence bound is utilized for the detection of process anomaly. It is clearly seen that the T^2 (Fig. 2(a)) is not sensitive to the process faults present in this data set, where the only identified abnormality is at sample 54, whilst the SPE chart (Fig. 2(b)) detects a larger number of faulty samples. Therefore practitioners must consider the two charts together, and by combination the identified faulty samples are: 33 and 50-54. Potentially using two monitoring charts may be desirable in that they give different information about the fault, i.e. T^2 to test “in-control” and SPE to test “in-model” (Kim and Lee, 2003). However there exist practical incentives to use a single monitoring chart, including it being more sensitive to detect abnormal process behavior than the individual statistics (Chen et al., 2004) and the reduction of plant operators workload by observing only one chart (Chen et al., 2006). Fig. 3 shows that the likelihood-based monitoring chart based on PPCA model is capable of detecting the same set of faulty samples with a single chart.

As an illustration the contribution plots for sample number 54 are given by Fig. 4 for PCA model. Since both T^2 and SPE charts identify this sample as faulty, the contribu-

Figure 3: Monitoring chart from PPCA model with 99% (dash-dotted line) confidence bound.

tion analysis are executed for the two monitoring statistics. The contribution plot of T^2 indicates a relatively large number of responsible variables for the fault, whereas variable 5 is dominant in the contribution plot of SPE. Therefore the challenge is to combine the information contained in both contribution plots to reach a decision as to which variables are the most responsible and should be selected for further investigation. In practice plant operators or data analysts may draw a subjective conclusion from Fig. 4 that variable 5 has made the largest contribution to the fault. However this *ad hoc* approach to the incorporation of two plots involves unnecessary human intervention (rather than a rigorous analysis), and thus the appropriateness of results is subject to human experience.

Fig. 5 gives the missing variable based contribution analysis for sample number 54 using PPCA model, where the contribution is plotted in terms of $M^2 - E[M^2]$, i.e. the decrease in the monitoring statistic if the corresponding variable is treated as missing. In addition the 99% confidence bound $\chi_D^2(0.99)$ is illustrated in terms of $M^2 - \chi_D^2(0.99)$ in the plot. If the contribution exceeds this confidence bound, i.e. $M^2 - E[M^2] > M^2 - \chi_D^2(0.99)$,

Figure 4: Contribution plots for PCA model. Left: T^2 ; Right: SPE.

the corresponding variable is considered to be significantly responsible for the process fault. (Note it is also possible to establish confidence limits for conventional contribution plots; see for example (Conlin et al., 2000; Qin et al., 2001; Westerhuis et al., 2000).)

Clearly Fig. 5 indicates without ambiguity that variable 5 is the dominant source of the data being detected as abnormal. The fact that the contribution from variable 5 exceeds the 99% confidence bound reinforces the finding that the data would have been normal should variable 5 be eliminated. Compared with the conventional contribution plots in Fig. 4, the major advantage of the proposed methodology is that the contribution of variables require to be analyzed only for a single monitoring statistic, i.e. the M^2 in the case of PPCA model.

A close inspection of Figs. 2 – 5 may reveal that the probabilistic monitoring chart and contribution analysis lead to the same results as the SPE. However this observation should not be generalized. In fact, the likelihood statistic encapsulates the information of both T^2 and SPE, and the similarity between the SPE and probabilistic results is because the SPE is dominant in detecting the fault in this example. We have observed in practice (not reported here) that when T^2 and SPE are both over the confidence bounds but the effect of T^2 is more substantial, the probabilistic chart and contribution analysis resemble the

Figure 5: Missing variable based contribution plot for PPCA model with 99% confidence bound (dash-dotted line).

results of T^2 . Therefore the proposed method appears to be a promising way to provide unified monitoring chart and contribution plot.

4.2 Monitoring of an industrial propylene polymerization process

As a second example the proposed method is applied to the monitoring of an industrial propylene polymerization process. The whole polymerization process consists of two continuous stirred tank reactors (CSTRs) and two fluid bed reactors, with 36 process variables being measured on-line, including the feed rate of materials (catalysts, hydrogen and propylene), reactor conditions (hydrogen concentration, temperature and pressure) and cooling water inlet/outlet temperature and flow rate values. In this study the performance monitoring of the first CSTR is presented for illustration where 12 process variables were recorded. There are 1150 data points available from this reactor. The first 1000 samples are known to be collected when the process is running under nominal condition, and thus they are used for model development. The remaining 150 samples, depicted in Fig. 6, are used to demonstrate the process monitoring methodology. The data is pre-processed

Figure 6: Trends of the process measurements being monitored.

Figure 7: Scatter plot of the scores on the first two principal components (PCs).

using auto-scaling before further analysis.

As a preliminary analysis the conventional PCA is applied to the training data. Fig. 7 shows the scatter plot of the scores on the first two principal components. Clearly this data set exhibits some multi-modal characteristic, and thus a single PCA or PPCA model is not expected to adequately approximate the distribution of the nominal process conditions.

Figure 8: Bayesian information criterion (BIC) to determine the number of mixtures.

This issue can be addressed using mixture modeling methodology (Chen et al., 2006; Choi et al., 2004, 2005; Thissen et al., 2005). In the rest of this section the PPCA mixture model will be developed and the proposed contribution analysis will be performed.

Fig. 8 gives the Bayesian information criterion (BIC) used for the selection of number of mixtures, where within each mixture component the number of principal components is determined such that they explain in excess of 90% of the total variance. According to the BIC, the PPCA mixture model with three mixture components (the number of principal components is 5, 6 and 6) is utilized for the on-line process monitoring.

Fig. 9 depicts the monitoring chart with 99% confidence bound for the on-line data. It is known from post-inspection that the process underwent significant drift in product quality index after sample number 50. This primary fault is successfully detected by examining the monitoring chart. However it is clearly seen that a number of early samples also exceed the 99% confidence bound. The contribution analysis is employed to investigate the responsibility of the variables for the confidence bound being violated.

Figure 9: Monitoring chart from PPCA mixture model with 99% confidence bound (dash-dotted line).

Fig. 10 demonstrates the contribution plots based on the proposed missing variable approach. The contribution is quantified in the logarithm scale as $\log E[p(\tilde{\mathbf{x}})] - \log p(\mathbf{x})$ for ease of illustration. Accordingly the 99% confidence bound is shown in terms of $\log(h) - \log p(\mathbf{x})$. The responsible variable in sample number 18 (Fig. 10(a)) is identified as variable 3 (catalyst feed rate), whereas the responsible variables in sample number 37 (Fig. 10(b)) are variable 7 (cooling water outlet temperature) and 12 (reactor level). Clearly the faulty patterns of these two samples are different from that of sample number 50 where the primary process fault occurs. Fig. 10(d) identifies variable 5 (inlet cooling water temperature) as the most responsible variable for the onset of the primary fault. Hence the contribution analysis suggests that the samples 18 and 37 exceed the confidence bound not because they are early indications of the primary process fault, but as the result of some other process disturbance. Finally Fig. 10(c) locates variables 5-7 and 9 as the contributing factors of the detected fault at sample number 43, with variable 5 being most

Figure 10: Missing variable based contribution plots from PPCA mixture model with 99% confidence bound (dash-dotted line). (a): Sample No. 18; (b): Sample No. 37; (c): Sample No. 43; (d): Sample No. 50.

influential. Therefore sample 43 could be an early indication of occurrence of the fault from sample 50 onwards.

5 Conclusions

This paper proposes a missing variable based contribution analysis methodology in conjunction with multivariate statistical process monitoring for process fault detection and identification. In particular two probabilistic approaches to process monitoring are considered: the probabilistic PCA (PPCA) and PPCA mixture models. The application of the PPCA and PPCA mixture models to process monitoring is discussed in detail, including model selection and the development of a single likelihood based confidence bound. The

probabilistic framework provides a natural way of handling the missing variables that form the basis of the contribution analysis. When applied with PPCA model, the missing variable based contribution plots appear to be more indicative with regard to identifying the most responsible variables for the detected faults. Furthermore, the proposed methodology provides a fault identification approach for PPCA mixture model where conventional contribution analysis is not applicable. The case studies with two industrial processes demonstrate that the proposed contribution analysis can provide significant information to facilitate process fault diagnosis.

Future work is focused on extending the proposed contribution analysis for the monitoring of batch manufacturing processes.

Acknowledgment

Y. Sun is grateful for the support from National Natural Science Foundation of China (Grant Number: 50777071).

References

- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Chen, Q., U. Kruger, M. Meronk, and A. Y. T. Leung (2004). Synthesis of T^2 and Q statistics for process monitoring. *Control Engineering Practice* 12, 745–755.
- Chen, T., J. Morris, and E. Martin (2006). Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring. *Journal of the Royal Statistical Society C (Applied Statistics)* 55, 699–715.

- Choi, S. W., E. B. Martin, and A. J. Morris (2005). Fault detection based on a maximum-likelihood principal component analysis (PCA) mixture. *Industrial and Engineering Chemistry Research* 44, 2316–2327.
- Choi, S. W., J. H. Park, and I.-B. Lee (2004). Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. *Computers and Chemical Engineering* 28, 1377–1387.
- Conlin, A. K., E. B. Martin, and A. J. Morris (2000). Confidence limits for contribution plots. *Journal of Chemometrics* 14, 725–736.
- Dunia, R., S. Qin, T. Edgar, and T. McAvoy (1996). Identification of faulty sensors using PCA. *AIChE Journal* 42, 2797–2812.
- Ge, Z. and Z. Song (2008). Online monitoring of nonlinear multiple mode processes based on adaptive local model approach. *Control Engineering Practice* 16, 1427–1437.
- Geladi, P. and B. R. Kowalski (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 185, 1–17.
- Kim, D. and I.-B. Lee (2003). Process monitoring based on probabilistic PCA. *Chemometrics and intelligent laboratory systems* 67, 109–123.
- Kim, H.-C., D. Kim, and S. Y. Bang (2003). An efficient model order selection for PCA mixture model. *Pattern Recognition Letters* 24, 1385–1393.
- Ku, W., R. H. Storer, and C. Georgakis (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 30, 179–196.
- Martin, E. B. and A. J. Morris (1996). Non-parametric confidence bounds for process performance monitoring charts. *Journal of Process Control* 6, 349–358.

- Miller, P., R. E. Swanson, and C. F. Heckler (1998). Contribution plots: a missing link in multivariate quality control. *International Journal of Applied Mathematics and Computer Science* 8, 775–792.
- Musa, M. E. M., D. de Ridder, R. P. W. Duin, and V. Atalay (2004). Almost autonomous training of mixtures of principal component analyzers. *Pattern Recognition Letters* 25, 1085–1095.
- Nelson, P. R. C., J. F. MacGregor, and P. A. Taylor (2006). The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemometrics and Intelligent Laboratory Systems* 80, 1–12.
- Qin, S. J., S. Valle-Cervantes, and M. Piovoso (2001). On unifying multiblock analysis with applications to decentralized process monitoring. *Journal of Chemometrics* 15, 715–742.
- Quiñonero-Candela, J., A. Girard, and C. E. Rasmussen (2003). Prediction at an uncertain input for gaussian processes and relevance vector machines - application to multiple-step ahead time-series forecasting. Technical Report IMM-2003-18, Technical University of Denmark, Lyngby, Denmark.
- Safavi, A. A., J. Chen, and J. A. Romagnoli (1997). Wavelet-based density estimation and application to process monitoring. *AIChE Journal* 43, 1227–1241.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
- Shao, R., F. Jia, E. B. Martin, and A. J. Morris (1999). Wavelets and non-linear principal components analysis for process monitoring. *Control Engineering Practice* 7, 865–879.
- Thissen, U., H. Swierenga, A. P. de Weijer, R. Wehrens, W. J. Melssen, and L. M. C. Buy-

- dens (2005). Multivariate statistical process control using mixture modelling. *Journal of Chemometrics* 19, 23–31.
- Tipping, M. E. and C. M. Bishop (1999a). Mixtures of probabilistic principal component analysers. *Neural Computation* 11, 443–482.
- Tipping, M. E. and C. M. Bishop (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society B* 61, 611–622.
- Westerhuis, J. A., S. P. Gurden, and A. K. Smilde (2000). Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems* 51, 95–114.
- Wilson, D. J. H., G. W. Irwin, and G. Lightbody (1999). RBF principal manifolds for process monitoring. *IEEE Transactions on Neural Networks* 10, 1424–1434.
- Wise, B. M., D. J. Veltkamp, N. L. Ricker, B. R. Kowalski, S. M. Barnes, and V. Arakali (1991). Application of multivariate statistical process control (mspc) to the west valley slurry-fed ceramic melter process. In *Proceedings of Waste Management’91*, Tucson, AZ, USA.
- Wold, S., K. Esbensen, and P. Geladi (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 37–52.
- Yue, H. and S. Qin (2001). Reconstruction based fault identification using a combined index. *Industrial and Engineering Chemistry Research* 40, 4403–4414.
- Zhao, S. J., J. Zhang, Y. M. Xu, and Z. H. Xiong (2006). A nonlinear projection to latent structures method and its applications. *Industrial and Engineering Chemistry Research* 45, 3843–3852.