Robust probabilistic PCA with missing data and contribution analysis for outlier detection

Tao Chen [a,*], Elaine Martin [b], Gary Montague [b]

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

[b] *School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK*

Abstract

Principal component analysis (PCA) is a widely adopted multivariate data analysis technique, with interpretation being established on the basis of both classical linear projection and a probability model (i.e. probabilistic PCA (PPCA)). Recently robust PPCA models, by using the multivariate t distribution, have been proposed to consider the situation where there may be outliers within the data set. This paper presents an overview of the robust PPCA technique, and further discusses the issue of missing data. An expectation-maximization (EM) algorithm is presented for the maximum likelihood estimation of the model parameters in the presence of missing data. When applying robust PPCA for outlier detection, a contribution analysis method is proposed to identify which variables contribute the most to the occurrence of outliers, providing valuable information regarding the source of outlying data. The proposed technique is demonstrated on numerical examples, and the application to outlier detection and diagnosis in an industrial fermentation process.

Key words: EM algorithm, missing data, multivariate t distribution, principal component analysis, probability density estimation, robust model.

1 Introduction

Principal component analysis (PCA) (Jolliffe, 2002) is a general multivariate statistical projection technique for dimension reduction, and it has seen a wide spectrum of applications in various areas, including exploratory data analysis, pattern recognition, quality monitoring and control. The traditional approach to the implementation of PCA is based on the linear projection of the original data

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

onto a space where the variance is maximized. Let $\{\mathbf{x}_n, n = 1, \cdots, N\}$ be the d-dimensional data set, then the first step in PCA is to compute the sample covariance matrix, $\mathbf{S}$, of order $d \times d$. The eigenvectors $\mathbf{w}_j$ and eigenvalues λ_j of $\mathbf{S}$ are then calculated ($j = 1, \cdots, d$). By retaining those eigenvectors corresponding to the q largest eigenvalues, the q-dimensional PCA score vectors, $\mathbf{t}_n$, are calculated as: $\mathbf{t}_n = \mathbf{W}^{\mathrm{T}}(\mathbf{x}_n - \boldsymbol{\mu})$, where $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_q)$, and $\boldsymbol{\mu}$ is the mean of the data set. Therefore the original data can be represented as a linear combination of the scores plus a noise vector: $\mathbf{x}_n = \mathbf{W}\mathbf{t}_n + \boldsymbol{\mu} + \mathbf{e}_n$.

More recently Tipping and Bishop (1999b) proposed a probabilistic formulation of PCA (PPCA) from the perspective of a Gaussian latent variable model. The PPCA model is realized by specifying a Gaussian noise model $\mathbf{e} \sim G(\mathbf{0}, \sigma^2\mathbf{I})$, which implies that the conditional distribution of data given PCA scores is: $\mathbf{x}|\mathbf{t} \sim G(\mathbf{W}\mathbf{t} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$. By adopting a prior Gaussian distribution for the PCA score vector, $\mathbf{t} \sim G(\mathbf{0}, \mathbf{I})$, the marginal distribution of the data $\mathbf{x}$ is also shown to be Gaussian: $\mathbf{x} \sim G(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I})$. PPCA degenerates to traditional PCA if $\sigma^2 \to 0$. Within the PPCA framework, the principal components are essentially the maximum likelihood estimates of the model parameters, which can be implemented using either the eigen-decomposition of the sample covariance matrix (as in the traditional PCA), or an expectation-maximization (EM) algorithm (Dempster et al., 1977). Tipping and Bishop (1999b) argued that PPCA attains several advantages over traditional PCA, including its extendability to handling missing data and to forming a mixture model (Tipping and Bishop, 1999a), and its potential application in probability density estimation and multivariate statistical process monitoring (Chen and Sun, 2009; Kim and Lee, 2003).

It is a well known issue that the conventional PCA is sensitive to anomalous observations because the calculation of sample mean and covariance matrix can be significantly influenced by a small number of outliers. Similarly, PPCA is not robust to outliers since the data are assumed to follow a multivariate Gaussian distribution that is easily affected by deviant observations. There is a rich literature on robust PCA methods to obtain principal components that are insensitive to outliers. The first category of methods are based on a robust estimate of the covariance matrix (Cambell, 1980; Devlin et al., 1981; Huber, 1981; Ruymagaart, 1981). The idea is to give different weights to the observations where the weight is a function of Mahalanobis distance. The observations with large Mahalanobis distance are automatically down-weighted since they tend to be outliers. The major difficulty with these weighting methods is due to high computation. Note that computing the covariance matrix requires $O(Nd^2)$ operations, which is infeasible for high dimensional data. Recently, more robust estimates of the covariance structure have been proposed, including positive-breakdown estimator (Croux and Haesbroeck, 2000) and that based on a convex loss function (Ibazizen and Dauxois, 2003). Nevertheless, these methods are still limited to moderate dimensions due to computational cost.

An alternative approach to robust PCA is based on projection pursuit (Li and Chen, 1985; Hubert et al., 2002). Dealing with high dimensional data, projection pursuit seeks low dimensional projections that maximize a robust measure of spread. By obtaining the projections sequentially, projection pursuit is computationally efficient particularly when $q \ll d$. More recently, Hubert et al. (2005) proposed a robust PCA method that is aimed at combining the advantages of robust covariance estimation and projection pursuit. It should be noted that all the robust PCA methods reviewed above are based on conventional PCA, and thus they do not fall into the family of probability models.

This paper discusses a robust PCA method within a probabilistic framework, based on replacing the Gaussian distribution utilized in the original PPCA by a heavy-tailed t distribution that is robust to the presence of outliers. The idea of using t distribution was originally proposed by Archambeau et al. (2006) for both robust PPCA and robust probabilistic canonical correlation analysis (PCCA), and was later extended to developing robust latent variable regression models (Fang and Jeong, 2008). This paper will initially review the rationale of the multivariate t distribution in Section 2, and the formulation of robust PPCA in Section 3. In Section 4 we address the issue of missing data, which is common in many data analysis tasks due to sensor fault or human errors, within the framework of robust PPCA. An EM algorithm will be developed for the estimation of model parameters in the presence of missing data. Section 5 gives two numerical examples to illustrate the effectiveness of the proposed robust PPCA model. Subsequently we demonstrate the application of the robust PPCA for outlier detection in Section 6. Specifically we present a contribution analysis method, which was previously proposed for multivariate statistical process control using non-robust PPCA and mixture models (Chen and Sun, 2009), to identify which variables contribute the most to the occurrence of outliers and provide useful information regarding the source of outlying data. Finally Section 7 concludes this paper.

2 Multivariate t distribution

Assume the random variable $\mathbf{x}$ follows a multivariate Gaussian distribution: $\mathbf{x} \sim G(\mathbf{m}, \mathbf{\Sigma})$. In the presence of outliers, a two-component Gaussian mixture model can be employed to account for the relatively large variance of the outliers:

$$\mathbf{x} \sim (1-\epsilon)G(\mathbf{m}, \mathbf{\Sigma}) + \epsilon\, G(\mathbf{x}; \mathbf{m}, b\mathbf{\Sigma}) \tag{1}$$

where b is a positive large factor, and $\epsilon \in [0, 1]$ is a small value to reflect the prior knowledge that a small portion of the data may be outliers. This two-component model has seen various applications in outlier detection and measurement rectifi-

cation (Chen et al., 2008; Schick and Mitter, 1994). The mixture model in Eq. (1) can be extended to an *infinite* Gaussian mixture model as (Peel and McLachlan, 2000):

$$\mathbf{x} \sim \int G(\mathbf{m}, b\mathbf{\Sigma})p(b)db \tag{2}$$

where the integration is performed over the scaling factor b. Suppose $u = 1/b$ is a chi-square random variable with degrees of freedom v: $u \sim \text{Ga}(v/2, v/2)$, where the Gamma probability density function is given by: $\text{Ga}(\alpha, \beta) = \beta^{\alpha} u^{\alpha-1} e^{-\beta u}/\Gamma(\alpha)$ Then the marginal distribution of $\mathbf{x}$ can be obtained by performing the integration in (2), resulting in a multivariate t distribution with degrees of freedom v (Lange et al., 1989; Liu, 1997; Peel and McLachlan, 2000): $\mathbf{x} \sim t_v(\mathbf{m}, \mathbf{\Sigma})$. An alternative perspective on the t distribution is to treat u as a latent variable, and the conditional distribution of $\mathbf{x}|u$ is Gaussian: $\mathbf{x}|u \sim G(\mathbf{m}, \mathbf{\Sigma}/u)$.

The Gaussian distribution is a special case of the t distribution when $v \to \infty$. In general the t distribution has significantly heavier tails than the Gaussian distribution, which is a desirable property to handle data sets in the presence of outliers.

3 Robust PPCA

This section reviews the robust PPCA model originally proposed in (Archambeau et al., 2006; Fang and Jeong, 2008), including the probability model and detailed parameter estimation method using EM algorithm.

3.1 The probability model

To consider the presence of outliers in the data set, the Gaussian distribution in PPCA is replaced by the t distribution to achieve a robust model. Specifically the conditional distribution of the data $\mathbf{x}$ given PCA scores $\mathbf{t}$ is

$$\mathbf{x}|\mathbf{t}, u \sim G\left(\mathbf{W}\mathbf{t} + \boldsymbol{\mu}, \sigma^2\mathbf{I}/u\right) \tag{3}$$

where $u \sim \text{Ga}(v/2, v/2)$. Thus $\mathbf{x}|\mathbf{t} \sim t_v(\mathbf{W}\mathbf{t} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$. If the prior of the scores is also a t distribution: $\mathbf{t}|u \sim G(\mathbf{0}, \mathbf{I}/u)$, or equivalently $\mathbf{t} \sim t_v(\mathbf{0}, \mathbf{I})$, the distribution of the data given u is:

$$\mathbf{x}|u \sim \int p(\mathbf{x}|\mathbf{t},u)p(\mathbf{t}|u)d\mathbf{t} = G(\boldsymbol{\mu}, \mathbf{C}/u) \tag{4}$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$ is a $d \times d$ matrix. The marginal distribution of the data $\mathbf{x}$ is then a multivariate t distribution: $\mathbf{x} \sim t_v(\boldsymbol{\mu}, \mathbf{C})$.

For the estimation of model parameters presented subsequently, the conditional distributions, $u|\mathbf{x}$ and $\mathbf{t}|\mathbf{x}, u$, are required. The former was shown (Lange et al., 1989) to be a Gamma distribution:

$$u|\mathbf{x} \sim \mathrm{Ga}\left(\frac{v+d}{2}, \frac{v+p}{2}\right) \tag{5}$$

where $p = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. The conditional distribution of the scores can be calculated by using Bayes' rule, resulting in:

$$\mathbf{t}|\mathbf{x}, u \sim G\left(\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}/u\right) \tag{6}$$

where $\mathbf{M} = \mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}$ is a $q \times q$ matrix. Similarly $\mathbf{t}|\mathbf{x}$ is also t distributed: $\mathbf{t}|\mathbf{x} \sim t_v(\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$. In summary, the Gaussian distributions primarily used in PPCA have been replaced by the t distributions in the proposed robust PPCA model. For the purpose of inference, it is required to invert the covariance matrix $\mathbf{C}$. This can be efficiently performed by using the Woodbury matrix identity if $q \ll d$ (which is often the case if d is large):

$$\mathbf{C}^{-1} = \left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}\right)^{-1} = \mathbf{I}/\sigma^2 - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}/\sigma^2 \tag{7}$$

The objective of dimension reduction, one of the major motivation behind PCA, can be achieved by utilizing the mean of the latent variables:

$$\langle \mathbf{t}|\mathbf{x} \rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}) \tag{8}$$

where $\langle\rangle$ is the expectation operator. This is also the projection of the original d-dimensional data to the (lower) q-dimensional PCA scores.

3.2 Maximum likelihood estimation

The model parameters to be estimated are: $\{\boldsymbol{\mu}, \mathbf{W}, \sigma^2, v\}$. Given a set of training data, the maximum likelihood estimation of these parameters can be achieved

by using the expectation-maximization (EM) algorithm (Dempster et al., 1977). To apply the EM algorithm, the latent variables $\{\mathbf{t}_n, u_n\}$ are treated as "missing data", and the "complete" data comprise the latent variables and the observed data $\mathbf{x}_n$. The corresponding complete-data log-likelihood is:

$$\mathcal{L}_C = \sum_{n=1}^{N} \ln\{p(\mathbf{x}_n, \mathbf{t}_n, u_n)\} \tag{9}$$

where the joint distribution can be factorized as:

$$p(\mathbf{x}_n, \mathbf{t}_n, u_n) = p(\mathbf{x}_n|\mathbf{t}_n, u_n)p(\mathbf{t}_n|u_n)p(u_n) \tag{10}$$

In the E-step, the expectation of the complete-data log-likelihood with respect to the conditional distribution $\mathbf{t}_n, u_n|\mathbf{x}_n$ is calculated as follows:

$$\begin{aligned}\langle\mathcal{L}_C\rangle = -\sum_{n=1}^{N}\Bigg\{&\frac{d}{2}\ln(\sigma^2) + \frac{\langle u_n\rangle}{2\sigma^2}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{x}_n-\boldsymbol{\mu}) - \frac{1}{\sigma^2}\langle u_n\mathbf{t}_n\rangle^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_n-\boldsymbol{\mu}) \\ &+\frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{W}\langle u_n\mathbf{t}_n\mathbf{t}_n^{\mathrm{T}}\rangle) + \frac{v}{2}\log\frac{v}{2} + \left(\frac{v}{2}-1\right)\langle\log u_n\rangle - \log\Gamma\left(\frac{v}{2}\right) - \frac{v}{2}\langle u_n\rangle\Bigg\}\end{aligned} \tag{11}$$

where the terms that are independent of the model parameters are omitted. The expectation terms in (11) are:

$$\langle u_n\rangle = \frac{v+d}{v+p_n} \tag{12}$$

$$\langle\mathbf{t}_n\rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_n-\boldsymbol{\mu}) \tag{13}$$

$$\langle u_n\mathbf{t}_n\rangle = \langle u_n\rangle\langle\mathbf{t}_n\rangle \tag{14}$$

$$\langle u_n\mathbf{t}_n\mathbf{t}_n^{\mathrm{T}}\rangle = \sigma^2\mathbf{M}^{-1} + \langle u_n\rangle\langle\mathbf{t}_n\rangle\langle\mathbf{t}_n\rangle^{\mathrm{T}} \tag{15}$$

$$\langle\log u_n\rangle = \psi\left(\frac{v+d}{2}\right) - \log\left(\frac{v+p_n}{2}\right) \tag{16}$$

where $p_n = (\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$, and $\psi()$ is the digamma function.

The M-step maximizes $\langle\mathcal{L}_C\rangle$ with respect to the model parameters, resulting in the update equations as:

$$\tilde{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N} \langle u_n \rangle \left(\mathbf{x}_n - \mathbf{W} \langle \mathbf{t}_n \rangle \right)}{\sum_{n=1}^{N} \langle u_n \rangle} \tag{17}$$

$$\tilde{\mathbf{W}} = \left(\sum_{n=1}^{N} (\mathbf{x}_n - \tilde{\boldsymbol{\mu}}) \langle u_n \mathbf{t}_n \rangle^{\mathrm{T}} \right) \left(\sum_{n=1}^{N} \langle u_n \mathbf{t}_n \mathbf{t}_n^{\mathrm{T}} \rangle \right)^{-1} \tag{18}$$

$$\tilde{\sigma}^2 = \frac{1}{Nd} \sum_{n=1}^{N} \left\{ \langle u_n \rangle ||\mathbf{x}_n - \tilde{\boldsymbol{\mu}}||^2 - 2 \langle u_n \mathbf{t}_n \rangle^{\mathrm{T}} \tilde{\mathbf{W}}^{\mathrm{T}} (\mathbf{x}_n - \tilde{\boldsymbol{\mu}}) + \mathrm{tr}(\tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{W}} \langle u_n \mathbf{t}_n \mathbf{t}_n^{\mathrm{T}} \rangle) \right\} \tag{19}$$

and $\tilde{v}$ can be updated using a scalar non-linear maximization routine that is available in most computation software packages. The EM algorithm does not explicitly calculate the sample covariance matrix that requires $O(Nd^2)$ operations. An inspection of (18)(19) reveals that the computational complexity is only $O(Ndq)$. When $q \ll d$, considerable computational cost can be saved.

Furthermore, the EM algorithm can be simplified by using a two-stage procedure, where the PCA scores $\mathbf{t}_n$ are not considered in the first stage (Tipping and Bishop, 1999a). Hence the objective of the first stage is to estimate $\boldsymbol{\mu}$. More specifically, the complete-data log-likelihood in the first stage is:

$$\mathcal{L}_{C_1} = \sum_{n=1}^{N} \ln\{p(\mathbf{x}_n, u_n)\} = \sum_{n=1}^{N} \ln\{p(\mathbf{x}_n | u_n) p(u_n)\} \tag{20}$$

where $p(\mathbf{x}_n | u_n)$ is given by (4). The expectation of $\mathcal{L}_{C_1}$ with respect to $u_n | \mathbf{x}_n$, as given in (5), is:

$$\langle \mathcal{L}_{C_1} \rangle = -\sum_{n=1}^{N} \langle u_n \rangle (\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \tag{21}$$

Maximization of (21) with respect to $\boldsymbol{\mu}$ gives

$$\tilde{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N} \langle u_n \rangle \mathbf{x}_n}{\sum_{n=1}^{N} \langle u_n \rangle} \tag{22}$$

In the second stage, the latent variables $\mathbf{t}_n$ are introduced, and the log-likelihood in (20) is increased through the EM algorithm to update $\mathbf{W}$, σ^2 and v. It should be noted that in the second stage $\mathcal{L}_C$ is not actually maximized, because $\tilde{\boldsymbol{\mu}}$ is kept fixed. In this sense, the second stage corresponds to the generalized EM algorithm (Gelman et al., 1995). This two-stage EM algorithm leads to improved

convergence speed (Tipping and Bishop, 1999a), since the expectation terms in (12)-(15) are calculated using the updated mean, $\tilde{\boldsymbol{\mu}}$, to update $\mathbf{W}$, σ^2 and v.

In summary, the two-stage EM algorithm operates as follows.

(1) Stage 1:
- **E-Step:** Given current parameters $\{\boldsymbol{\mu}, \mathbf{W}, \sigma^2, v\}$, calculate the expected value $\langle u_n \rangle$ as in (12).
- **M-Step:** Update $\tilde{\boldsymbol{\mu}}$ as in (22).

(2) Stage 2:
- **E-Step:** Given current parameters $\{\tilde{\boldsymbol{\mu}}, \mathbf{W}, \sigma^2, v\}$, re-calculate the expected value as in (12)-(16).
- **M-Step:** Update $\tilde{\mathbf{W}}$ and $\tilde{\sigma}^2$ as in (18)(19), followed by the updating of $\tilde{v}$.

(3) Repeat Stage 1 and 2 until convergence is reached.

3.3 Post-processing of $\mathbf{W}$

In general, the loading matrix $\mathbf{W}$ at convergence is not necessarily orthogonal (Tipping and Bishop, 1999b), and a rotation of $\mathbf{W}$ through an arbitrary $q \times q$ orthogonal matrix $\mathbf{R}$, i.e. $\mathbf{WR}$, is still the maximum likelihood estimate of the robust PPCA model. The rotational ambiguity can be resolved if necessary by computing the eigen-decomposition of $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{R}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{R}$ where $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_q)$, and rotating $\mathbf{W}$ according to $\mathbf{WR}$. Based on this eigen-decomposition the percentage of explained variance by the PCA model, a widely used measure to assess the effectiveness of the PCA, can be calculated as $\sum_{j=1}^{q} \lambda_j / (\sum_{j=1}^{q} \lambda_j + \sigma^2)$

4 Missing data

The issue of missing data refers to the situations where the d-dimensional data $\mathbf{x}$ has some missing values. By assuming the missing-data mechanism does not depend on the missing values, i.e. missing at random, the conditional distribution of the missing values given observed data can be formulated, and it forms the basis of the EM algorithm for parameter estimation.

4.1 Conditional distribution of missing data

The data can be divided as $\mathbf{x}^{\mathrm{T}} = [\mathbf{x}^{o\mathrm{T}}, \mathbf{x}^{u\mathrm{T}}]$, where $\mathbf{x}^o$ and $\mathbf{x}^u$ are sub-vectors of observed and unobserved (missing) data respectively. According to the robust PPCA model, $\mathbf{x}|u \sim G(\boldsymbol{\mu}, \mathbf{C}/u)$. For ease of derivation the mean and covariance are also organized into blocks:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^o \\ \boldsymbol{\mu}^u \end{bmatrix}; \qquad \mathbf{C} = \begin{bmatrix} \mathbf{C}_{oo} & \mathbf{C}_{ou} \\ \mathbf{C}_{uo} & \mathbf{C}_{uu} \end{bmatrix}$$

The conditional distribution of missing data given observed data, $\mathbf{x}^u|\mathbf{x}^o, u$, is also Gaussian (Little and Rubin, 1987) with mean $\boldsymbol{\mu}^u + \mathbf{C}_{uo}\mathbf{C}_{oo}^{-1}(\mathbf{x}^o - \boldsymbol{\mu}^o)$, and covariance $(\mathbf{C}_{uu} - \mathbf{C}_{uo}\mathbf{C}_{oo}^{-1}\mathbf{C}_{ou})/u$. For the development of maximum likelihood estimation, it is convenient to utilize the distribution of the complete vector $\mathbf{x}$, which is again a Gaussian distribution: $\mathbf{x}|\mathbf{x}^o, u \sim G(\mathbf{z}, \mathbf{Q}/u)$, or equivalently $\mathbf{x}|\mathbf{x}^o \sim t_v(\mathbf{z}, \mathbf{Q})$, where

$$\mathbf{z} = \begin{bmatrix} \mathbf{x}^o \\ \boldsymbol{\mu}^u + \mathbf{C}_{uo}\mathbf{C}_{oo}^{-1}(\mathbf{x}^o - \boldsymbol{\mu}^o) \end{bmatrix}; \qquad \mathbf{Q} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{C}_{uu} - \mathbf{C}_{uo}\mathbf{C}_{oo}^{-1}\mathbf{C}_{ou}) \end{bmatrix} \tag{23}$$

4.2 Maximum likelihood estimation

The EM algorithm for the maximum likelihood estimation of model parameters is similar to that presented in Section 3.2, the difference being the handling of missing data represented by a Gaussian distribution $\mathbf{x}|u \sim G(\mathbf{z}, \mathbf{Q}/u)$. In the first stage of the EM algorithm, the PCA scores is not considered, and thus the expectation in the E-step is:

$$\langle \mathcal{L}_{C_1} \rangle = -\sum_{n=1}^{N} \mathrm{tr}\left(\langle u_n (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \rangle \mathbf{C}^{-1} \right) \tag{24}$$

Due to the presence of missing data, the expectation in (24) is taken with respect to $\mathbf{x}_n, u_n|\mathbf{x}_n^o$, as opposed to $u_n|\mathbf{x}_n$ in (21). $p(\mathbf{x}_n, u_n|\mathbf{x}_n^o)$ further factorizes as:

$$p(\mathbf{x}_n, u_n|\mathbf{x}_n^o) = p(\mathbf{x}_n|u_n, \mathbf{x}_n^o)\; p(u_n|\mathbf{x}_n^o) \tag{25}$$

where the first term is formulated in (23) as a Gaussian distribution: $G(\mathbf{z}_n, \mathbf{Q}_n/u)$. The second term is a Gamma distribution:

$$u_n|\mathbf{x}_n^o \sim \mathrm{Ga}\left(\frac{v + d_n^o}{2}, \frac{v + p_n^o}{2} \right) \tag{26}$$

where d_n^o is the dimension of observed data $\mathbf{x}_n^o$, and $p_n^o = (\mathbf{x}_n^o - \boldsymbol{\mu}^o)^{\mathrm{T}} \mathbf{C}_{oo}^{-1} (\mathbf{x}_n^o - \boldsymbol{\mu}^o)$. Therefore the expectations can be obtained as

$$\langle u_n \rangle = \frac{v + d_n^o}{v + p_n^o} \tag{27}$$

$$\langle u_n (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \rangle = \mathbf{Q}_n + \langle u_n \rangle (\mathbf{z}_n - \boldsymbol{\mu})(\mathbf{z}_n - \boldsymbol{\mu})^{\mathrm{T}} \tag{28}$$

Substituting (28) into (24) and letting $\partial \langle \mathcal{L}_{C_1} \rangle / \partial \boldsymbol{\mu} = \mathbf{0}$ results in the updating formula for $\boldsymbol{\mu}$:

$$\tilde{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N} \langle u_n \rangle \mathbf{z}_n}{\sum_{n=1}^{N} \langle u_n \rangle} \tag{29}$$

In the second stage of the EM algorithm, the PCA scores $\mathbf{t}_n$ are considered to estimate $\mathbf{W}$, σ^2 and v. Now the expectation of the complete-data log-likelihood in (9)(10) must be calculated with respect to $p(\mathbf{x}_n, \mathbf{t}_n, u_n | \mathbf{x}_n^o)$, which can be factorized as: $p(\mathbf{x}_n, \mathbf{t}_n, u_n | \mathbf{x}_n^o) = p(\mathbf{t}_n | \mathbf{x}_n, u_n) p(\mathbf{x}_n, u_n | \mathbf{x}_n^o)$, where the two terms are given in (6) and (25) respectively. Therefore the expected log-likelihood can be expanded as:

$$\begin{aligned} \langle \mathcal{L}_C \rangle = -\sum_{n=1}^{N} \Bigg\{ & \frac{d}{2} \ln(\sigma^2) + \frac{1}{2\sigma^2} \mathrm{tr} \left[\langle u_n (\mathbf{x}_n - \tilde{\boldsymbol{\mu}})(\mathbf{x}_n - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \rangle \right] \\ & - \frac{1}{\sigma^2} \mathrm{tr} \left[\langle u_n (\mathbf{x}_n - \tilde{\boldsymbol{\mu}}) \mathbf{t}_n^{\mathrm{T}} \rangle \mathbf{W}^{\mathrm{T}} \right] + \frac{1}{2\sigma^2} \mathrm{tr} \left[\mathbf{W}^{\mathrm{T}} \mathbf{W} \langle u_n \mathbf{t}_n \mathbf{t}_n^{\mathrm{T}} \rangle \right] \Bigg\} \end{aligned} \tag{30}$$

and the expectation terms are

$$\langle u_n (\mathbf{x}_n - \tilde{\boldsymbol{\mu}}) \mathbf{t}_n^{\mathrm{T}} \rangle = \langle u_n (\mathbf{x}_n - \tilde{\boldsymbol{\mu}})(\mathbf{x}_n - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \rangle \mathbf{W} \mathbf{M}^{-1} \tag{31}$$

$$\langle u_n \mathbf{t}_n \mathbf{t}_n^{\mathrm{T}} \rangle = \sigma^2 \mathbf{M}^{-1} + \mathbf{M}^{-1} \mathbf{W}^{\mathrm{T}} \langle u_n (\mathbf{x}_n - \tilde{\boldsymbol{\mu}})(\mathbf{x}_n - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \rangle \mathbf{W} \mathbf{M}^{-1} \tag{32}$$

where $\langle u_n (\mathbf{x}_n - \tilde{\boldsymbol{\mu}})(\mathbf{x}_n - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \rangle$ is given in (28) by replacing $\boldsymbol{\mu}$ with $\tilde{\boldsymbol{\mu}}$. Maximization of $\langle \mathcal{L}_C \rangle$ with respect to $\mathbf{W}$ and σ^2 results in the following updating formula:

$$\tilde{\mathbf{W}} = \left(\sum_{n=1}^{N} \langle u_n (\mathbf{x}_n - \boldsymbol{\mu}) \mathbf{t}_n^{\mathrm{T}} \rangle \right) \left(\sum_{n=1}^{N} \langle u_n \mathbf{t}_n \mathbf{t}_n^{\mathrm{T}} \rangle \right)^{-1} \tag{33}$$

$$\begin{aligned} \tilde{\sigma}^2 = \frac{1}{Nd} \sum_{n=1}^{N} \Big\{ & \mathrm{tr} \left[\langle u_n (\mathbf{x}_n - \tilde{\boldsymbol{\mu}})(\mathbf{x}_n - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \rangle \right] - 2\mathrm{tr} \left[\langle u_n (\mathbf{x}_n - \tilde{\boldsymbol{\mu}}) \mathbf{t}_n^{\mathrm{T}} \rangle \mathbf{W}^{\mathrm{T}} \right] \\ & + \mathrm{tr}(\tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{W}} \langle u_n \mathbf{t}_n \mathbf{t}_n^{\mathrm{T}} \rangle) \Big\} \end{aligned} \tag{34}$$

and $\tilde{v}$ can be updated using a non-linear maximization algorithm. The two-stage EM algorithm operates by alternating the E-step and M-step until convergence, similar to the procedure summarized in Section 3.2.

5 Numerical examples

We first consider a simple numerical example that comprises a data set with 90 random samples generated from a two-dimensional Gaussian distribution with zero mean and the following covariance matrix:

$$\begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix} \tag{35}$$

Another 10 data points were generated from a uniform distribution over the range $[-5, 5]$ to simulate the presence of outliers. These 100 data points were utilized for the development of the PPCA and robust PPCA model. In addition, a separate data set with missing values was simulated by randomly removing each value in the original data with probability 0.2.

Fig. 1 illustrates the first principal component and the 2.58 standard deviation contour obtained by using the PCA models. The "true results" were obtained by eigen-decomposition of the covariance matrix in Eq. (35), which was used to generate the data. It can be seen that the PPCA is sensitive to outliers, both the principal component and the contour being significantly different from the true ones. In contrast, the robust PPCA is much less susceptible to outliers. Furthermore, the presence of a reasonable level of missing data appears to have only small impact on the results of robust PPCA. Fig. 2(a) shows the projections of the first 20 data points by using robust PPCA whilst in Fig. 2(b) 20% of the values are missing. Despite some variations, the two plots are largely similar, indicating the effectiveness of the proposed approach to the handling of missing data.

To demonstrate the computational efficiency of the proposed method, we further consider five numerical data sets each having 500 data points ($N = 500$) with varying dimensions: $d = (10, 50, 200, 500, 1000)$. For each d, the data matrix $\mathbf{X}$ of order $N \times d$ is formed such that each element is a random sample generated from a univariate Gaussian distribution with zero mean and standard deviation. Note the PCA model is not appropriate for analyzing these data sets since the variables are independent. In this study the data are purely utilized for the illustration of computational time. In all cases we fix the number of principal components to five ($q = 5$).

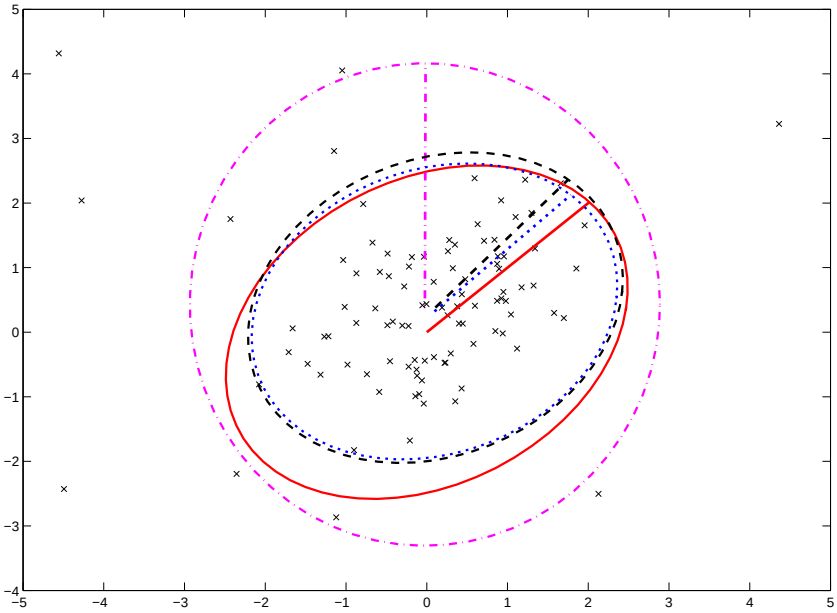
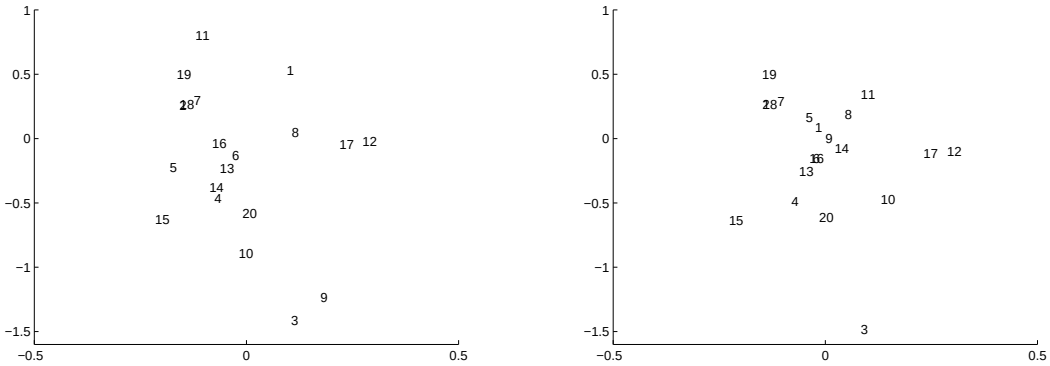


Fig. 1. The first principal component (straight line) and the 2.58 standard deviation contour (ellipsoid) obtained by using the PCA models. Original data ($\times$); True results (——); PPCA ($- \cdot -\cdot$); robust PPCA ($--$); robust PPCA with 20% missing values ($\cdots$). The results of the latter two situations are very similar, indicating small impact of the missing data on the robust PPCA model.



(a) (b)

Fig. 2. Projections of the first 20 data points by using robust PPCA (a) on the full data set and (b) with 20% missing values.

Table 1 gives the CPU time (s) for each iteration of the EM algorithm for the parameter estimation. The algorithm was implemented within the Matlab environment under Windows XP system equipped with a Pentium 2.8 GHz CPU. The results show the proposed method is reasonably efficient in computation.

Table 1
CPU time (s) for the parameter estimation (per iteration) of the robust PPCA models.

d	10	50	200	500	1000
Time (s)	0.05	0.06	0.11	0.77	2.67

The CPU time at 1000 variables is still feasible from the practical perspective. Clearly the total computation time is also dependent on the number of EM iterations that are required to converge. We have observed that in all the cases presented in Table 1, the EM algorithm appears to converge within 10 iterations. If necessary, the algorithm can run significantly faster by compiling the Matlab script into a binary executable file.

6 Application to outlier detection and diagnosis

As a general multivariate statistical tool, PCA has been applied to many practical problems in science, engineering and econometrics. As an example this section considers the application in outlier detection in an industrial fermentation process.

Since there is no formal definition of "outlier", this paper relies on the informal and intuitive statement that outliers are observations that are in some way inconsistent with the remainder of a data set (Barnett and Lewis, 1994). A large number of univariate outlier detection approaches have been suggested in the literature; see (Barnett and Lewis, 1994) for a review. However, these approaches, if applied to multivariate problems by investigating one variable at a time, would ignore the covariance structure of the data. Therefore multivariate techniques must be employed, and PCA has been one of the most accepted methods for outlier detection (Daszykowski et al., 2007; Jolliffe, 2002).

The procedure of using PCA for outlier detection is closely related to that of multivariate statistical process control (MSPC) (Qin, 2003), which is to monitor the performance of a manufacturing process to ensure process safety and delivery of consistent product. Both methods require the modeling of data using PCA (MSPC further requires the data being collected under normal operating conditions), followed by the development of confidence bound. If the confidence bound is exceeded, the occurrence of an outlier (or abnormal behavior in MSPC) is detected. Furthermore for diagnosis purpose, a *contribution analysis* procedure can be applied to identify which variables contribute the most to the occurrence of the outlier or abnormal process. These issues will be discussed in more detail subsequently.

One of the advantages of a probabilistic PCA model, in place of conventional PCA, for outlier detection (and MSPC), is that it provides a single likelihood-based confidence bound to detect outliers, as opposed to the confidence bounds for two metrics, i.e. Hotelling's T^2 and squared prediction error (SPE). In the literature of PCA, outliers are classified into two categories: leverage and orthogonal. Leverage outliers are far from the data majority on the score space (i.e. large T^2), whilst orthogonal outliers have large residuals from the PCA model (i.e. large SPE) (Daszykowski et al., 2007). In the community of process fault detection and diagnosis based on conventional PCA, the combination of T^2 and SPE metrics has attracted significant attention (Yue and Qin, 2001); however this is not an issue with probabilistic models. The likelihood value is a sufficient metric to measure how far one observation is from the majority of the data that is represented by the probabilistic model. Therefore the distinction between leverage and orthogonal outliers is unnecessary in the context of probabilistic models. In practice a single monitoring metric 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 outlier detection and MSPC in practice (Chen et al., 2006).

On the basis of the probability distribution $p(\mathbf{x})$, the $100\alpha\%$ confidence bound can be defined as a likelihood threshold h that satisfies the integral (Chen et al., 2006):

$$\int_{\mathbf{x}:p(\mathbf{x})>h} p(\mathbf{x})d\mathbf{x} = \alpha \tag{36}$$

For PPCA model $p(\mathbf{x})$ is a multivariate Gaussian distribution, and the equivalent confidence bound to (36) is based on the squared Mahalanobis distance M^2: the data point $\mathbf{x}$ is considered as an outlier if

$$M^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}) > \chi^2_d(\alpha) \tag{37}$$

where $\chi^2_d(\alpha)$ is the α-fractile of the chi-square distribution with degrees of freedom d. For robust PPCA model where $p(\mathbf{x})$ is multivariate t distribution, M^2/d has a F-distribution with d and v degrees of freedom (Kotz and Nadarajah, 2004). Therefore a data point is detected as an outlier if M^2/d exceeds the α-fractile of the corresponding F-distribution. However as the result of the heavy-tail characteristic of the t distribution, the confidence bound is observed in extensive preliminary studies (not reported) to be larger than is required and thus will fail to identify potential outliers. An alternative approach is to regard the robust PPCA as a method to robustly estimate the PCA projections (and thus $\boldsymbol{\mu}$ and

$\mathbf{C}$), and the outlier detection is performed based on the chi-square distribution as in (37). This method was used in (Peel and McLachlan, 2000) for noise detection.

Note the confidence bound based on the chi-square distribution is approximate, since the mean and covariance are not exactly known but estimated. Indeed, the determination of an appropriate threshold for the confidence bound has been a difficult task in the literature of outlier detection. Wilks (1962) showed that if the mean and covariance are estimated using the standard method (i.e. $\boldsymbol{\mu} = \sum_{n=1}^{N} \mathbf{x}_n/N$ and $\mathbf{C} = \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}/(N-1)$), then the squared Mahalanobis distance has a Beta distribution. However, the confidence bound derived from the Beta distribution is also approximate, as the mean and covariance are not estimated from the above standard equations in robust methods. Hardin and Rocke (2005) developed an improved F approximation to the distribution of M^2 for the robust minimum covariance determinant (MCD) estimator (Rousseeuw and van Driessen, 1999); how this improved approach can be adapted for the proposed robust PPCA method is an interesting topic to explore in the future.

In the presence of missing data, M^2 can be calculated as the expected value with respect to the conditional distribution of the missing data (Section 4.1):

$$E[M^2] = \mathrm{tr}\left[\mathbf{C}^{-1}\left\{(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}} + \mathbf{Q}\right\}\right] \tag{38}$$

6.2 Contribution analysis

The objective of contribution analysis is to identify which variables contribute the most to the occurrence of outliers. In general contribution analysis may not explicitly reveal the root-cause of the presence of outliers, but it is undoubtedly helpful in pointing out the inconsistent variables that may undergo further diagnosis procedures.

The traditional contribution analysis (Miller et al., 1998) is to decompose the Hotelling's T^2 and SPE obtained from PCA model into the sum of d contributing terms, each corresponding to a variable. A similar method was developed for PPCA model (Kim and Lee, 2003). These techniques are limited by the requirement of investigating the contribution to two metrics. It is not clear how to resolve the conflicts if the two contribution analysis procedures reach different conclusion about the responsible variables. Alternatively reconstruction based methods have been proposed (Dunia et al., 1996; Yue and Qin, 2001), 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 outlier.

Motivated by the reconstruction based techniques, this paper adopts a missing variable based contribution analysis method, which was originally proposed for MSPC using non-robust PPCA and mixture models (Chen and Sun, 2009), for robust PPCA model. Assume $\mathbf{x}$ is identified as a candidate outlier, for the j-th variable of $\mathbf{x}$ ($j = 1, \ldots, d$), the proposed method operates as follows:

(1) Let $\mathbf{x}^j$ be the candidate outlier with the j-th variable missing. Calculate the conditional distribution of $\mathbf{x}^j$ given the observed variables (Section 4.1).
(2) Calculate $E[M^2]$ as in (38).
(3) The contribution of the j-th variable can be quantified as $M^2 - E[M^2]$, i.e. the decrease in the monitoring metric if the variable is eliminated.

Furthermore, in step (c) if $E[M^2]$ is smaller than the confidence bound $\chi^2_d(\alpha)$, the corresponding variable can be regarded as being significantly influential, since its elimination would bring the data back to normal.

6.3 Application in a batch fermentation process

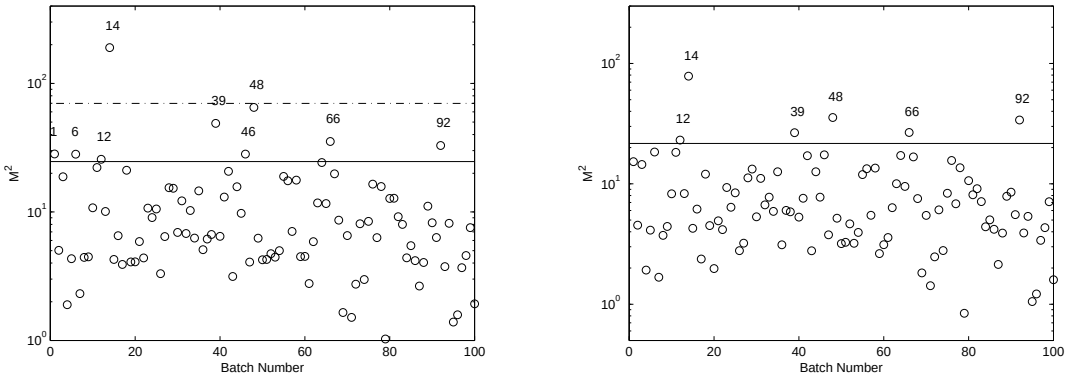
In the beer production industry, the fermentation process has the greatest influence over the product quality and production time variability. The process under investigation is operated in batch mode, where the wort from previous operations is fed into the fermenter. Then yeast is added to metabolize sugars and amino acids and alcohol is produced. As the sugars are used up the fermentation slows down, and a cooling operation can stop the process at the desired gravity (density of beer) and/or diacetyl concentration.

Since it is difficult to produce beer consistently in industrial scale, the quality control in beer industry relies significantly on the consistency of raw materials and process conditions. In this study data was collected from an industrial fermenter comprising 100 batches (Basabe, 2004). Each batch is treated as one data point that consists of 9 variables (Table 1), including initial conditions and process parameters. As opposed to including the temperature trajectory throughout each batch, only initial and mean temperatures were used for analysis. This is because the process temperature was manually controlled, and its large variation during the batch is not directly related to product quality (Basabe, 2004). Of all the 900 values (100 batches $\times$ 9 variables) 59 (6.56%) are missing. The data are preprocessed to zero mean and unit standard deviation on each dimension.

Both PPCA and robust PPCA were performed on the data and the dimension of the problem was reduced to 5 principal components, which explained 76.4% (PPCA) and 77.9% (robust PPCA) of the total variance. Fig. 3(a) gives the outlier detection chart for robust PPCA. Due to the heavy-tail effect of the t distribution, the F-distribution based confidence bound is significantly larger than the bound based on chi-square distribution, and the F-distribution identifies only one outlier.

Table 2
Variables that were used for the analysis of the fermentation process.

Variable	*Definition*	*Variable*	*Definition*
1	Dissolved oxygen	6	End gravity
2	Pitch weight	7	Time to reach desirable gravity
3	Viability	8	Time to reach desirable diacetyl
4	Initial gravity	9	Batch mean temperature
5	Initial temperature		



(a) (b)

Fig. 3. Outlier detection by using (a) robust PPCA and (b) PPCA with 99% confidence bound based on chi-square distribution (——) and F-distribution ($- \cdot -\cdot$). The batch numbers of the detected outliers are labeled.

An post-inspection revealed that the F-distribution missed a number of batches that are clearly regarded as outliers. Therefore the chi-square distribution based confidence bound is recommended and will be used subsequently.

Fig. 3 shows that PPCA detects 6 outliers (batches 12, 14, 39, 48, 66 and 92), whilst robust PPCA identifies an additional three (batches 1, 6, 46). Batch 14 is the most obvious outlier, and the contribution analysis in Fig. 4 (d) clearly indicates the first variable (dissolved oxygen) is responsible for the batch being detected as outlying. Variable 1 in batch 14 has the value of 99 ppm that appears to be the result of measurement error as opposed to abnormal process condition, since the concentration of dissolved oxygen at 99 ppm is not physically possible in this fermentation process. Fig. 4 (a) depicts that batch 1 is also detected as outlying by robust PPCA due to the first variable, whose value is 40 ppm and less extreme than batch 14. However due to the dominant influence of batch 14, the PPCA significantly over-estimates the variance of the first variable, and it failed to identify batch 1 as outlier. In general PPCA tends to be sensitive to a small number of influential data points. Therefore robust PPCA is more appropriate for the modeling of the data where outliers are present.

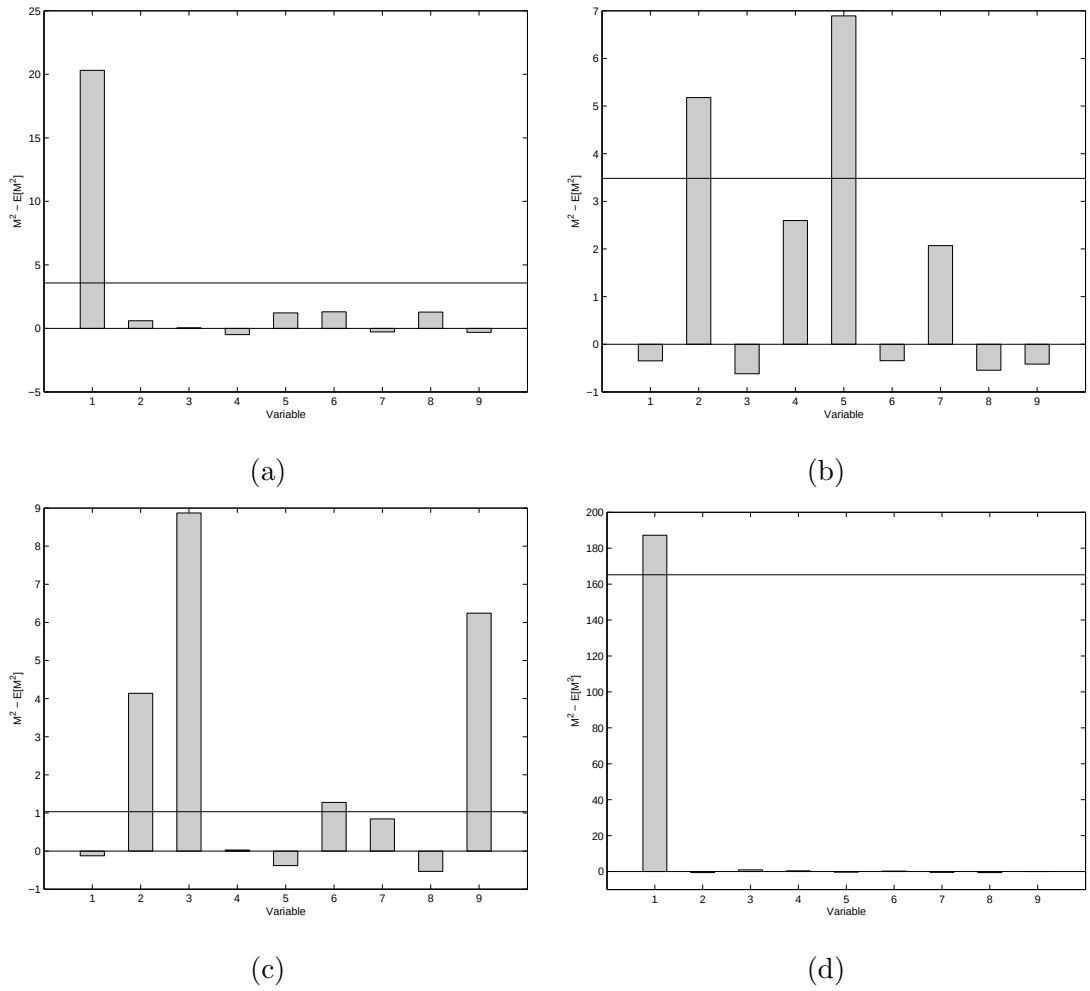


Fig. 4. Contribution analysis for robust PPCA with 99% confidence bound (——) that was calculated as $M^2 - \chi^2_d(0.99)$. (a) Batch 1; (b) Batch 6; (c) Batch 12; (d) Batch 14.

Fig. 4 also depicts the contribution analysis for batch 6 and 12. Clearly the contribution plots provide important information for the diagnosis of the source of outliers.

7 Conclusions and discussions

This paper presents a robust PCA model within a probabilistic framework, with specific focus on handling missing data and its applications in outlier detection and diagnosis. The idea is to replace the Gaussian distribution utilized by the probabilistic PCA with the heavy-tailed and more robust multivariate t distribution. The EM algorithm is implemented for the maximum likelihood parameter estimation and the handling of missing data. Numerical example has shown that the presence of a reasonable level of missing data appears to have only small impact on the results of robust PPCA. Furthermore, a contribution analysis method

has been developed to help identify the influential variables that contribute to the occurrence of outliers, and it has been successfully applied to the analysis of the data collected from an industrial fermentation process.

One limitation of the robust PPCA model is that it does not consider the situation where outliers are clustered and thus the overall distribution of data is multi-modal. In general, clustered outliers are more difficult to detect. The presented robust PPCA, which assumes a uni-modal distribution of the data, is not specifically designed to address this issue. A potential solution, suggested by Rocke and Woodruff (2001), is to utilize cluster analysis to first identify the clusters, and then develop a metric from the largest identified cluster(s) for outlier detection. Following this idea, the robust PPCA can be extended to a mixture model, similar to the mixture of (non-robust) PPCA (Tipping and Bishop, 1999a). Alternatively, the methodology of "forward search" (Atkinson et al., 2004) can be adopted to incrementally include the data for analysis, and thus both the isolated and clustered outliers can be identified sequentially. Currently these methods are under investigation.

References

Archambeau, C., Delanney N., Verleysen, M., 2006. Robust probabilistic Projection. Proc. 23rd International Conference on Machine Learning, Pittsburgh, USA.

Atkinson, A. C., Riani, M., Cerioli, A., 2004. Exploring Multivariate Data with the Forward Search. Springer-Verlag, New York.

Barnett, V., Lewis, T., 1994. Outliers in Statistical Data, 3rd Edition. John Wiley, New York.

Basabe, X. L., 2004. Towards improved fermentation consistency using multivariate analysis of process data. Master's thesis, University of Newcastle upon Tyne, UK.

Cambell, N. A., 1980. Robust procedures in multivariate analysis. Applied Statistics 29, 231–237.

Chen, T., Morris, J., Martin, E., 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.

Chen, T., Morris, J., Martin, E., 2008. Dynamic data rectification using particle filters. Computers and Chemical Engineering 32, 451–462.

Chen, T., Sun, Y., 2009. Probabilistic contribution analysis for statistical process monitoring: a missing variable approach. Control Engineering Practice 17, 469–477.

Croux, C., Haesbroeck, G., 2000. Principal components analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. Biometrika 87, 603–618.

Daszykowski, M., Kaczmarek, K., Heyden, Y. V., Walczak, B., 2007. Robust statistics in data analysis - a review basic concepts. Chemometrics and Intelligent Laboratory Systems 85, 203–219.

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

Devlin, S. J., Gnanadesikan, R., Kettenring, J. R., 1981. Robust estimation of dispersion matrices and principal component. Journal of the American Statistical Association 12, 136–154.

Dunia, R., Qin, S., Edgar, T., McAvoy, T., 1996. Identification of faulty sensors using PCA. AIChE Journal 42, 2797–2812.

Fang, Y., Jeong, M. K., 2008. Robust probabilistic multivariate calibration model. Technometrics 50, 305–316.

Gelman, A. B., Carlin, J. S., Stern, H. S., Rubin, D. B., 1995. Bayesian data analysis. Chapman & Hall/CRC.

Hardin, J., Rocke, D. M., 2005. The distribution of robust distances. Journal of Computational and Graphical Statistics 14, 910–927.

Huber, P. J., 1981. Robust Statistics. Wiley, New York.

Hubert, M., Rousseeuw, P. J., Branden, K. V., 2005. ROBPCA: a new approach to robust principal component analysis. Technometrics 47, 64–79.

Hubert, M., Rousseeuw, P. J., Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics. Chemometrics and Intelligent Laboratory Systems 60, 101–111.

Ibazizen, M., Dauxois, J., 2003. A robust principal component analysis. Statistics 37, 73–83.

Jolliffe, I. T., 2002. Principal Component Analysis, 2nd Edition. Springer.

Kim, D., Lee, I.-B., 2003. Process monitoring based on probabilistic PCA. Chemometrics and intelligent laboratory systems 67, 109–123.

Kotz, S., Nadarajah, S., 2004. Multivariate t distributions and their applications. Cambridge University Press.

Lange, K. L., Little, R. J. A., Taylor, J. M. G., 1989. Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84, 881–896.

Li, G., Chen, Z., 1985. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. Journal of the American Statistical Association 80, 759–766.

Little, R. J. A., Rubin, D. B., 1987. Statistical Analysis with Missing Data. Wiley, Chichester.

Liu, C., 1997. ML estimation of the multivariate t distribution and the EM algorithm. Journal of Multivariate Analysis 63, 296–312.

Miller, P., Swanson, R. E., Heckler, C. F., 1998. Contribution plots: a missing link in multivariate quality control. International Journal of Applied Mathematics and Computer Science 8, 775–792.

Peel, D., McLachlan, G. J., 2000. Robust mixture modelling using the t distribution. Statistics and Computing 10, 339–348.

Qin, S. J., 2003. Statistical process monitoring: basics and beyond. Journal of Chemometrics 17, 480–502.

Rocke, D. M., Woodruff, D. L., 2001. Multivariate outlier detection and robust covariance matrix estimation - discussion. Technometrics 43, 300–303.

Rousseeuw, P. J., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Ruymagaart, F. H., 1981. A robust principal component analysis. Journal of Multivariate Analysis 11, 485–497.

Schick, I. C., Mitter, S. K., 1994. Robust recursive estimation in the presence of heavy-tailed observation noise. Annals of Statistics 22, 1045–1080.

Tipping, M. E., Bishop, C. M., 1999a. Mixtures of probabilistic principal component analysers. Neural Computation 11, 443–482.

Tipping, M. E., Bishop, C. M., 1999b. Probabilistic principal component analysis. Journal of the Royal Statistical Society B 61, 611–622.

Wilks, S., 1962. Mathematical Statistics. Wiley, New York.

Yue, H., Qin, S., 2001. Reconstruction based fault identification using a combined index. Industrial and Engineering Chemistry Research 40, 4403–4414.