Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Chemometrics and Intelligent Laboratory Systems 67 (2003) 109–123
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

# Process monitoring based on probabilistic PCA

Dongsoon Kim, In-Beum Lee*

*School of Environmental Science and Engineering, Pohang University of Science and Technology, San 31 Hyoja-Dong, Pohang, Kyungbuk 790-784, South Korea*

Received 25 April 2002; accepted 12 April 2003

**Abstract**

This paper proposes a multivariate process monitoring method based on probabilistic principal component analysis (PPCA). First we will summarize several well-known statistical process monitoring methods, e.g. univariate/multivariate Shewhart charts, and the PCA-based method, i.e. Q and Hotelling's $T^2$ charts. And then the probabilistic method will be proposed and compared to the existing methods. In essence, the univariate Shewhart chart, multivariate Shewhart chart, Q chart, and $T^2$ chart are unified to the probabilistic method. The PPCA model is calibrated by the expectation and maximization (EM) algorithm similar to PCA by NIPALS algorithm; EM algorithm will be explained briefly in the article. Finally, through an illustrative example, we will show how the probabilistic method works and is applied to the process monitoring.


*Keywords:* EM algorithm; Monitoring; PCA; Probabilistic PCA; Shewhart chart

## 1. Introduction

Consider the zero-meaned Gaussian process expressed by Eq. (1).[1]

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^P$ signifies measurement variable, $\sim$ denotes 'distributed according to', and $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ symbolizes Gaussian probability density function (pdf) with zero-means and $\boldsymbol{\Sigma}$ covariance matrix. Henceforth $\mathcal{N}(\boldsymbol{x}: \boldsymbol{\mu}, \boldsymbol{\Sigma})$ will be used to express $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ due to its simplicity. Notice that $p(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x}: \boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that the pdf of $\boldsymbol{x}$ is characterized by a Gaussian parametric function with two parameters: mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; and hence if $p(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x}: \boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $p(\boldsymbol{x}) = (2\pi)^{-0.5\cdot\dim(\boldsymbol{x})}\cdot\det(\boldsymbol{\Sigma}^{-1})^{0.5}\cdot \exp(-0.5\cdot(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\cdot(\boldsymbol{x}-\boldsymbol{\mu}))$.

All vector variables used in this article are column vectors represented by bold-italic style, scalar variables are denoted by italic fonts, and matrix variables are expressed by bold-capital letters. For example, let us denote the $P\times N$ calibration sample set sampled from Eq. (1) as $\mathbf{X}_{\mathrm{cali}}=\{\boldsymbol{x}_n\}_{n\in N}$. Then the $n$th column of the set, $\boldsymbol{x}_n=\{x_{pn}\}_{p\in P}$, indicates the $n$th sample of $\boldsymbol{x}$, where $x_{pn}$ represents the $p$th row and $n$th column element in $\mathbf{X}_{\mathrm{cali}}$.[2] In the same manner, we can also

* Corresponding author. Tel.: +82-54-279-2274; fax: +82-54-279-3499.
*E-mail address:* iblee@postech.ac.kr (I.-B. Lee).

[1] If characteristic of a process can be described by a Gaussian probability density function (pdf), then we can rewrite it to have zero-means via mean-centering.

[2] For concise expression, unusual notation $n\in N$ is used to express $n=1, 2,\ldots, N$. This notation should be distinguished from the 'element' symbol, e.g. $1\in\{1, 2, 3\}$, $\boldsymbol{x}\in\mathbb{R}^P$, etc.

0169-7439/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0169-7439(03)00063-7

signify the $P \times \mathbb{N}$ test sample set as $\mathbf{X}_{\text{test}}=\{\boldsymbol{x}_{\text{n}}\}_{\text{n}\in\mathbb{N}}$ with $\boldsymbol{x}_{\text{n}}=\{x_{p\text{n}}\}_{p\in P}$.

The objective of process monitoring is detecting abnormal event(s) of processes concerned with respect to the process model. In other words, it is the test whether or not $\boldsymbol{x}_{\text{n}}$ comes from the normal condition of the process regarding the calibrated model by $\mathbf{X}_{\text{cali}}$. The only admittable variation of $\boldsymbol{x}_{\text{n}}$ is caused by white-type noise with negligible amplitude or variance, which is inevitable. There may be three types of abnormal for the Gaussian process: (1) abnormal means of variables, (2) abnormal variances of variables, and (3) abnormal correlations among variables.

The univariate Shewhart chart [1] has been widely used to detect types (1) and (2). It judges the abnormality by statistical testing for all the elements in $\boldsymbol{x}$, $\{x_p\}_{p\in P}$, separately. But the separated charts-based method cannot address type (3). It is a serious defect of the method as will be discussed in Section 2.1.

The multivariate Shewhart chart [2,3] was devised to overcome the defect. It perceives type (3) as well as both the types (1) and (2) in one chart. In brief, theoretic foundation on the chart-based method is that the squared Euclidian norm of whitened $\boldsymbol{x}$,[3] i.e. squared Mahalanobis norm of $\boldsymbol{x}$, follows central $\chi^2$ distribution with dimension of $\boldsymbol{x}$, dim($\boldsymbol{x}$), degrees of freedom (dof). However, when insufficient samples were used to estimate the Gaussian parameters, the test is substituted by $\mathcal{F}$ test which checks the centrality of the $\chi^2$ distribution. This is the Hotelling's $\mathcal{T}^2$ test [4], which will be explained in Section 2.2.

If some elements in $\boldsymbol{x}$ are strongly correlated to each other, then inversion of $\boldsymbol{\Sigma}$, which is essential to the whitening, is often intractable due to its singularity. Principal components analysis (PCA) [5] is a well-known countermeasure for the singularity; thus PCA-based process monitoring method [6] is applicable to the case. PCA converts high dimensional strongly correlated $\{x_p\}_{p\in P}$ to low dimensional mutually uncorrelated principal components $\{z_l\}_{l\in L}$ which contain most of variances in $\{x_p\}_{p\in P}$.[4] In essence, the PCA-based method was devised to utilize projected $\boldsymbol{x}$ onto the significant $L$ eigen-subspaces of $\boldsymbol{\Sigma}$, $\boldsymbol{z}\in\mathbb{R}^{\text{L}}$, instead of $\boldsymbol{x}\in\mathbb{R}^{\text{P}}$ itself.[5] But this substitution, $\boldsymbol{x}\rightarrow\boldsymbol{z}$, must accompany some information loss because the neglected $(P-L)$ eigen-subspaces of $\boldsymbol{\Sigma}$ can also have important information in the process. So, we should check whether the loss is negligible or not; Q test [7] has been used for checking. Therefore, the PCA-based method should encompass both tests: Q test for 'in-model', which checks whether the PCs' score obtained from $\boldsymbol{z}_{\text{n}} \leftarrow (\text{PCA of } \mathbf{X}_{\text{cali}}) - \boldsymbol{x}_{\text{n}}$ is still valid or not; and $\mathcal{T}^2$ test for 'in-control', which checks whether the $\boldsymbol{z}_{\text{n}}$ implies normal condition of the process or not. Hence, Q test is the necessary condition to use $\mathcal{T}^2$ test. We will inspect the tests carefully in Section 3.

However, the PCA-based method has some subtle but crucial demerits. First of all, PCA itself is not achieved in probability density space at all; it just assumes the uniform probability densities for all the variables in the model, in other words, Euclidian distance is used as its measuring unit. But notice that all statistical judgments or decisions must be made based on probability density space, and hence Mahalanobis distance is essential to the measuring unit. By this reason, two totally different stages should be passed when we apply PCA to the process monitoring: (1) develop PCA model using Euclidian distance measure between $\boldsymbol{x}$ and $\boldsymbol{z}$, i.e. NIPALS algorithm, and then (2) judge condition of the process based on the Mahalanobis norm of $\boldsymbol{z}$, i.e. Hotelling's $\mathcal{T}^2$ test. Moreover, the in-model test (Q test) utilizes the Euclidian norm of estimation errors. These different measuring units make the statistical decision complicated. We will explain this in Section 3 in more detail.

Suppose there is a novel method which finds the principal axes via probabilistic way, i.e. each variable in the model has specified its pdf, and the model is calibrated by the probability density information of the variables. Then the statistical decision will be the direct answer of the probabilistic model calibration.

[3] Suppose $\mathcal{N}(\boldsymbol{x}: \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Whitened $\boldsymbol{x}$ is defined by $\boldsymbol{x}^{\text{w}} \equiv \boldsymbol{\Sigma}^{-0.5}\cdot(\boldsymbol{x}-\boldsymbol{\mu})$, and hence $\mathcal{N}(\boldsymbol{x}^{\text{w}}: \boldsymbol{0}, \mathbf{I})$.

[4] Under $\boldsymbol{x}=\mathbf{A}\cdot\boldsymbol{z}+\boldsymbol{e}$ model structure, PCA is the solution of $\arg_{\mathbf{A}}\min: \Sigma_{p\in P}\,\text{var}(x_p) - \Sigma_{l\in L}\,\text{var}(z_l)$.

[5] In PCA, the significance is determined by the magnitudes of eigenvalues of $\boldsymbol{\Sigma}$.

PPCA [8,9] is one such method. We will investigate the model more closely in Sections 4 and 5. Besides, PPCA has two other good properties over PCA: it can handle incomplete data well, even data with some missing values by using the conditional probability density of the variables, and extending this to the mixture model [10]. However, we will not deal with these two properties in this article; Ref. [11] is helpful to understand them.

## 2. Shewhart chart

### 2.1. Univariate Shewhart chart

If all elements in $\boldsymbol{x}$, $\{x_p\}_{p\in P}$, in Eq. (1) were uncorrelated, i.e. all off-diagonals in $\boldsymbol{\Sigma}$ were zeros, then the elements-based process abnormal test might be applicable. Let us denote the variance of the $p$th element in $\boldsymbol{x}$, $x_p$, as $\lambda_p = \mathrm{diag}(\boldsymbol{\Sigma})_p$. Since $\mathcal{N}(\lambda_p^{-0.5}\cdot x_p: 0, 1)\ \forall p$, if $\{x_p\}_{p\in P}$ satisfies Eq. (2.1), then we regard the process which generates $\boldsymbol{x}$ as normal with $\alpha$ level of significance (LoS). And the happening probability density of $x_p$ from the process is given by Eq. (2.2).

$$\lambda_p^{-0.5}\cdot x_p \in \left[\mathcal{N}^{-1}_{s(0.5\cdot\alpha)},\ \mathcal{N}^{-1}_{s(1-0.5\cdot\alpha)}\right] \tag{2.1}$$

$$p(x_p) = (2\cdot\pi\cdot\lambda_p)^{-0.5}\cdot\exp(-0.5\cdot\lambda_p^{-1}\cdot x_p^2), \tag{2.2}$$

where $\mathcal{N}_s^{-1}{}_{(\beta)}$ denotes the inverse of cumulative $\mathcal{N}(0, 1)$ corresponding to $\beta\cdot 100\%$ probability. When insufficient samples were used to estimate $\lambda_p$, the sample variance, $s_p \equiv N^{-1}\cdot\Sigma_{n\in N}\ x_{pn}^2$, based test is used instead of Eq. (2.1). This is the Student's $\mathcal{T}$ test:

$$s_p^{-0.5}\cdot x_p \in \left[\mathcal{T}^{-1}_{(0.5\cdot\alpha;N)}, \mathcal{T}^{-1}_{(1-0.5\cdot\alpha;N)}\right], \tag{2.3}$$

because $\lambda_p^{-0.5}\cdot x_p\cdot(\Sigma_n(\lambda_p^{-0.5}\cdot x_{pn})^2/N)^{-0.5} = s_p^{-0.5}\cdot x_p \sim \mathcal{T}_{(N)}$. Notice that the relevant dof is $N$, not $(N-1)$, since zero-mean process was assumed. The elements' variances, $\{\lambda_p\}_{p\in P} = \mathrm{diag}(\boldsymbol{\Sigma})$, can also be tested by $\chi^2_{(P)}$ test. Since $\Sigma_{n\in N}(\lambda_p^{-0.5}\cdot x_{pn})^2 = \lambda_p^{-1}\cdot(N\cdot s_p) \sim \chi^2_{(N)}$, if $s_p$ satisfies Eq. (2.4), then $s_p$ is considered as $\lambda_p$ with $\alpha$ LoS.

$$s_p \le (\lambda_p/N)\cdot\chi^{-2}_{(1-\alpha;N)} \tag{2.4}$$

But the elements-based tests do not account for the correlation abnormality among $\{x_p\}_{p\in P}$. For this problem, a good example is found in Ref. [6].

### 2.2. Multivariate Shewhart chart

Any Gaussian random variable $\boldsymbol{x}\in\mathbb{R}^{\mathrm{P}}$, $\mathcal{N}(\boldsymbol{x}: \boldsymbol{\mu}, \boldsymbol{\Sigma})$, can be transformed to the standard Gaussian via whitening; e.g. $\boldsymbol{x}^{\mathrm{w}} \equiv \boldsymbol{\Sigma}^{-0.5}\cdot(\boldsymbol{x}-\boldsymbol{\mu})$ then $\mathcal{N}(\boldsymbol{x}^{\mathrm{w}}: \boldsymbol{0}, \mathbf{I})$. The squared Mahalanobis norm of $\boldsymbol{x}$, which is identical to the squared Euclidian norm of whitened $\boldsymbol{x}$, follows $\chi^2_{(P)}$ distribution by the definition of $\chi^2$ distribution.[6] Therefore, if the sample obtained from Eq. (1) satisfies Eq. (3.1), then the process which generates the sample is considered normal with $\alpha$ LoS; and the corresponding probability density of the sample is given by Eq. (3.2).

$$\left\|\boldsymbol{\Sigma}^{-0.5}\cdot\boldsymbol{x}\right\|^2 \le \chi^{-2}_{(1-\alpha;P)} \tag{3.1}$$

$$p(\boldsymbol{x}) = (2\cdot\pi)^{-0.5\cdot P}\cdot\det(\boldsymbol{\Sigma}^{-1})^{0.5}\cdot\exp(-0.5\cdot\boldsymbol{x}^{\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{x}), \tag{3.2}$$

where $||\boldsymbol{\Sigma}^{-0.5}\cdot\boldsymbol{x}||^2$ in Eq. (3.1), and $\boldsymbol{x}^{\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{x}$ in the exponential term in Eq. (3.2) represent the squared Mahalanobis norm of $\boldsymbol{x}$.

When insufficient samples were used to estimate $\boldsymbol{\Sigma}$, the sample covariance matrix of $\boldsymbol{x}$, defined by $\mathbf{S} \equiv N^{-1}\cdot\Sigma_{n\in N}\ \boldsymbol{x}_n\cdot\boldsymbol{x}_n^{\mathrm{T}}$, based normality test can be applicable. Hotelling's $\mathcal{T}^2$ statistics is the test. If a sample $\boldsymbol{x}$ is generated from Eq. (1), then $||\mathbf{S}^{-0.5}\cdot\boldsymbol{x}||^2 \sim P\cdot(N^2-1)\cdot(N^2-N\cdot P)^{-1}\cdot\mathcal{F}_{(P,N-P)}$ is expected. So, Eq. (3.1) is substituted by Eq. (4) for the sample set with small $N$ [12].

$$\left\|\mathbf{S}^{-0.5}\cdot\boldsymbol{x}\right\|^2 \le P\cdot(N^2-1)\cdot(N^2-N\cdot P)^{-1}\cdot\mathcal{F}^{-1}_{(1-\alpha;\ P,N-P)}, \tag{4}$$

where $\mathbf{S}$ is a matrix and $\mathcal{F}^{-1}_{(\beta;\ P,N-P)}$ denotes the inverse of cumulative $\mathcal{F}_{(P,N-P)}$ at $\beta\cdot 100\%$ probability. Notice that the right half term in Eq. (4) is identical to $\chi^{-2}_{(1-\alpha;\ P)}$ in Eq. (3.1) if $N \approx \infty$ [13].

Eigen-decomposition of $\mathbf{S}$ is helpful to analyze its structure. Let us denote the $p$th large eigenvalue of $\mathbf{S}$ as $d_p$, and its paired eigenvector as $\boldsymbol{u}_p$. Then the $p$th

[6] $||\boldsymbol{x}||^2_{\mathrm{Mahalanobis}} \equiv \|\boldsymbol{\Sigma}^{-0.5}\cdot(\boldsymbol{x}-\boldsymbol{\mu})\|^2 = \|\boldsymbol{x}^{\mathrm{w}}\|^2_{\mathrm{Euclidian}} = \Sigma_{p\in P}\,(x_p^{\mathrm{w}})^2 \sim \chi^2_{(\dim(\boldsymbol{x}))}$ since $\mathcal{N}(\boldsymbol{x}^{\mathrm{w}}: \boldsymbol{0}, \mathbf{I})$.

eigen-subspace of $\mathbf{S}$ is defined by $\mathbf{S}_p \equiv d_p \cdot \boldsymbol{u}_p \cdot \boldsymbol{u}_p^{\mathrm{T}}$, and $\mathbf{S} = \Sigma_{p \in P}\, \mathbf{S}_p$. Therefore, it is obvious that

$$\begin{aligned}\mathbf{S}^{-1} &= \Sigma_{p\in P}\mathbf{S}_p^{-1} = \Sigma_{p\in P} d_p^{-1} \cdot \boldsymbol{u}_p \cdot \boldsymbol{u}_p^{\mathrm{T}} \\ &= \left(\{d_p^{-0.5} \cdot \boldsymbol{u}_p\}_{p\in P}\right) \cdot \left(\{d_p^{-0.5} \cdot \boldsymbol{u}_p\}_{p\in P}\right)^{\mathrm{T}} \\ &= \left(\mathbf{S}^{-0.5}\right)^{\mathrm{T}} \cdot \left(\mathbf{S}^{-0.5}\right)\end{aligned}$$

If some elements in $\boldsymbol{x}$ are strongly correlated to each other, then obtained $\mathbf{S}$ will be nearly singular, i.e. some of $\{d_p\}_{p\in P}$ are practically zero, or some of $\{d_p^{-1}\}_{p\in P}$ are nearly infinite. It implies that there are over weightings on the insignificant eigen-subspaces of $\mathbf{S}$. To prevent the problem, the PCA-based process monitoring method was devised.

## 3. PCA-based process monitoring

PCA is a well-established dimensionality reduction technique. Numerous applications, e.g. data compression, image processing, visualization, exploratory data analysis, pattern recognition, time series prediction as well as multivariate process monitoring [14–17], have been reported. PCA is achieved via eigen-decomposition of the sample covariance matrix of $\boldsymbol{x}$, $\mathbf{S}$. To ease description, let us partition $\mathbf{S}$ into two parts: its systematic part, $\mathbf{S}_{\mathrm{S}}$, consisting of significant $L$ eigen-subspaces, and its noise part, $\mathbf{S}_{\mathrm{N}}$, constructed with remaining $(P-L)$ eigen-subspaces

$$\mathbf{S} = \mathbf{S}_{\mathrm{S}} + \mathbf{S}_{\mathrm{N}}, \tag{5}$$

where $\mathbf{S}_{\mathrm{S}} \equiv \Sigma_{l\in L} d_l \cdot \boldsymbol{u}_l \cdot \boldsymbol{u}_l^{\mathrm{T}}$, and $\mathbf{S}_{\mathrm{N}} \equiv \Sigma_{i=L+1,\ldots,P} d_i \cdot \boldsymbol{u}_i \cdot \boldsymbol{u}_i^{\mathrm{T}}$. PC-based Hotelling's $\mathcal{T}^2$ test is based on the fact that $\|\mathbf{S}_{\mathrm{S}}^{-0.5} \cdot \boldsymbol{x}\|^2 \sim L \cdot (N^2-1) \cdot (N^2 - N \cdot L)^{-1} \cdot \mathcal{F}_{(L, N-L)}$ where $(\mathbf{S}_{\mathrm{S}}^{-0.5})^{\mathrm{T}} = \{d_l^{-0.5} \cdot \boldsymbol{u}_l\}_{l\in L} \in \mathbb{R}^{\mathrm{P}\times \mathrm{L}}$. So, it is the test of the systematic part of $\mathbf{S}$. Therefore, for the sample which is satisfied by Eq. (6.1), the process which generates the sample is regarded as in-control with $\alpha$ LoS.

$$\left\|\mathbf{S}_{\mathrm{s}}^{-0.5} \cdot \boldsymbol{x}\right\|^2 \le L \cdot (N^2-1) \cdot (N^2 - N \cdot L)^{-1} \cdot \mathcal{F}^{-1}_{(1-\alpha;\, L, N-L)}, \tag{6.1}$$

where $\mathbf{S}_{\mathrm{S}}^{-0.5}$ is no longer problematic since $\{d_l^{-0.5}\}_{l\in L}$ are properly bounded.

Notice that Eq. (6.1) is only applicable to the monitoring when the tested sample shares the significant eigen-subspaces of $\mathbf{S}$ spanned by $\{\boldsymbol{u}_l\}_{l\in L}$. That means we cannot decide whether the sample is in-control or not if the sample does not accord with the developed PCA model. This sharing is tested by Q test; it is the measure of squared Euclidian distance between $\boldsymbol{x}$ and its projection onto $\{\boldsymbol{u}_l\}_{l\in L}$. We regard a sample shares the principal subspaces with $\alpha$ LoS if the sample satisfies

$$\begin{aligned}\boldsymbol{x}^{\mathrm{T}} \cdot (\mathbf{I} - \Sigma_{l\in L} \boldsymbol{u}_l \cdot \boldsymbol{u}_l^{\mathrm{T}}) \cdot \boldsymbol{x} \le \theta_1 \cdot [&(\mathcal{N}^{-1}_{s(1-\alpha)} \cdot (2 \cdot \theta_2 \cdot h_0^2)^{0.5} \\ &\cdot \theta_1^{-1}) + (\theta_2 \cdot h_0 \cdot (h_0 - 1) \cdot \theta_1^{-2}) + 1]^{1/h0},\end{aligned} \tag{6.2}$$

where $\theta_j = \Sigma_{i=L+1,\ldots,P} d_i^j$ for $j \in \{1, 2, 3\}$, and $h_0 = 1 - (2 \cdot \theta_1 \cdot \theta_3) \cdot (3 \cdot \theta_2^2)^{-1}$. The left-half term in the equation indicates $\|\boldsymbol{x} - (\Sigma_{l\in L}\ \boldsymbol{u}_l \cdot \boldsymbol{u}_l^{\mathrm{T}}) \cdot \boldsymbol{x}\|^2$ since $(\mathbf{I} - \Sigma_{l\in L} \boldsymbol{u}_l \cdot \boldsymbol{u}_l^{\mathrm{T}})$ is an idempotent matrix.[7] So, it is apparent that the test is governed by Euclidian measuring unit. But all statistical decisions must be made by the normalized unit, e.g. Mahalanobis norm used in $\mathcal{T}^2$ test. It is the reason why the right-half term in Q test is so complicated.

Moreover, since $\mathcal{T}^2$ and Q use different measuring units, they cannot be unified to diagnose a sample; in other words, we cannot produce the multivariate Shewhart chart using $\mathcal{T}^2$ and Q charts. For instance, suppose we have two samples, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, obtained from Eq. (1) where the estimate of $\boldsymbol{\Sigma}$ was given by $\mathbf{S}$. We may decide which sample is more abnormal than the other by comparing $\|\mathbf{S}^{-0.5} \cdot \boldsymbol{x}_i\|^2$ and $\|\mathbf{S}^{-0.5} \cdot \boldsymbol{x}_j\|^2$ for invertible $\mathbf{S}$, i.e. $\min(\mathrm{eig}(\mathbf{S})) \ge \varepsilon > 0$ for small enough $\varepsilon$.[8] But if $\min(\mathrm{eig}(\mathbf{S})) < \varepsilon$, and hence $\mathbf{S}^{-1}$ is hard to be tractable, then the PCA-based method would be utilized as mentioned before. However, the comparison is not clear in the PCA-based method; e.g. if $\|\mathbf{S}_{\mathrm{S}}^{-0.5} \cdot \boldsymbol{x}_i\|^2 > \|\mathbf{S}_{\mathrm{S}}^{-0.5} \cdot \boldsymbol{x}_j\|^2$ but $\boldsymbol{x}_i^{\mathrm{T}} \cdot (\mathbf{I} - \Sigma_{l\in L} \boldsymbol{u}_l \cdot \boldsymbol{u}_l^{\mathrm{T}}) \cdot \boldsymbol{x}_i < \boldsymbol{x}_j^{\mathrm{T}} \cdot (\mathbf{I} - \Sigma_{l\in L} \boldsymbol{u}_l \cdot \boldsymbol{u}_l^{\mathrm{T}}) \cdot \boldsymbol{x}_j$, then we cannot decide which sample is more abnormal than the other.

Let us consider the probabilistic model: $\boldsymbol{x} = \mathbf{A} \cdot \boldsymbol{z} + \boldsymbol{e}$ where $\mathcal{N}(\boldsymbol{z} \in \mathbb{R}^{\mathrm{L}}: \boldsymbol{0}, \mathbf{I})$ and $\mathcal{N}(\boldsymbol{e} \in \mathbb{R}^{\mathrm{P}}: \boldsymbol{0}, \boldsymbol{\Lambda} = \mathrm{diag}\ \{\lambda_p\}_{p\in P})$; and hence $\mathcal{N}(\boldsymbol{x} \in \mathbb{R}^{\mathrm{P}}: \boldsymbol{0}, \boldsymbol{\Sigma} = \mathbf{A} \cdot \mathbf{A}^{\mathrm{T}} + \boldsymbol{\Lambda})$.

[7] Matrix $\mathbf{P}$ is idempotent if $\mathbf{P}^2 = \mathbf{P}$.

[8] $\min(\mathrm{eig}(\mathbf{S}))$ denotes the minimum value of the eigenvalues of $\mathbf{S}$.

If we have a method to estimate the model parameters, $\mathbf{A}$ and $\mathbf{\Lambda}$, from $\mathbf{X}_{\text{cali}}=\{\boldsymbol{x}_n\}_{n\in N}$, then the in-control and in-model tests will be simplified to $\|\boldsymbol{z}\|^2 \le \chi^{-2}_{(1-\alpha;\ L)}$, and $\|\mathbf{\Lambda}^{-0.5}\cdot\boldsymbol{e}\|^2 \le \chi^{-2}_{(1-\alpha;\ P)}$, respectively. And both tests will be unified to $\|(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\mathbf{\Lambda})^{-0.5}\cdot\ \boldsymbol{x}\|^2 \le \chi^{-2}_{(1-\alpha;\ P)}$; and hence to simplify the comparison, e.g. compare between $\|(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\mathbf{\Lambda})^{-0.5}\cdot\boldsymbol{x}_i\|^2$ and $\|(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\mathbf{\Lambda})^{-0.5}\cdot\boldsymbol{x}_j\|^2$.[9] Moreover, there are several interesting properties in this probabilistic approach as will be explained in the following section.

## 4. PPCA-based process monitoring

Before we explain the PPCA, let us briefly summarize the probabilistic generative model. In this model, we consider the measurement variable $\boldsymbol{x}\in\mathbb{R}^{\mathrm{P}}$ as just the output of the linear combinations of mutually uncorrelated input variable $\boldsymbol{z}\in\mathbb{R}^{\mathrm{L}(\le \mathrm{P})}$ plus additive noise $\boldsymbol{e}$, and hence $\boldsymbol{x}=\mathbf{A}\cdot\boldsymbol{z}+\boldsymbol{e}$. This input–output model is called as the generative model. And if we set some specified probability densities to all variables in the model, e.g. Gaussians, then the model is named as the probabilistic model. The probabilistic generative models, e.g. PPCA, factor analysis (FA), independent component analysis (ICA), hidden Markov model (HMM), etc., consider the measurements, no matter how many there are, just as the shadow of the latent variables' (linear) combination.

The projection models, e.g. PCA, PLS, continuum regression (CR), cyclic subspace regression (CSR), etc., are neither probabilistic nor generative; all variables in the models do not have probability densities, and there is no input and output. Normalized PC scores extracted from $\boldsymbol{x}$ by PCA model, $\boldsymbol{z}=\mathbf{S}_{\mathrm{S}}^{-0.5}\cdot\boldsymbol{x}$, may look like the input corresponding to the output $\boldsymbol{x}$; but the scores are not the input, these are just the normalized projection results of $\boldsymbol{x}$ onto some interesting eigen-subspaces of $\mathbf{S}$.

In brief, the essential difference between the projection model, e.g. PCA, and the probabilistic generative model, e.g. PPCA, is that the former focuses on the measurement variable $\boldsymbol{x}$, and the latent variable $\boldsymbol{z}$ is treated just as the projection results of $\boldsymbol{x}$ onto some interesting covariance or correlation subspaces of $\boldsymbol{x}$. But the latter concentrates on $\boldsymbol{z}$, and $\boldsymbol{x}$ is treated just as the result of linear combination $\boldsymbol{z}$ plus small additive white-type noise. This subtle but important difference is the starting point to devise independent component analysis (ICA) [18], which is one of the famous techniques to solve blind sources separation problem.

### 4.1. PPCA model

PPCA aims to find the most probable parameter set $\mathbf{\Theta}\equiv\{\mathbf{A},\ \lambda\}$ in the model structure:

$$\boldsymbol{x}=\mathbf{A}\cdot\boldsymbol{z}+\boldsymbol{e}, \tag{7}$$

where $\mathbf{A}\in\mathbb{R}^{\mathrm{P}\times\mathrm{L}}$ denotes loading matrix, $\mathcal{N}(\boldsymbol{z}\in\mathbb{R}^{\mathrm{L}}:\ \boldsymbol{0},\ \mathbf{I})$ represents the pdf of latent variable, and $\mathcal{N}(\boldsymbol{e}\in\mathbb{R}^{\mathrm{P}}:\ \boldsymbol{0},\ \lambda\cdot\mathbf{I})$ signifies the pdf of noise variable. If $\mathcal{N}(\boldsymbol{e}:\ \boldsymbol{0},\ \mathbf{\Lambda}=\mathrm{diag}\{\lambda_p\}_{p\in P})$ were assumed, i.e. $\{x_p\}_{p\in P}$ were permitted to have different noise levels, then the factor analysis (FA) model [19] could be built. In fact, the PPCA model is a special case of FA model subjected to $\lambda_p=\lambda\ \forall p$, and PCA model is a special case of PPCA model restricted to $\lambda\approx 0$ [8].

The parameter set, $\mathbf{\Theta}$, can be estimated by the expectation and maximization (EM) algorithm, which is an iterative likelihood maximization algorithm, as will be explained in Section 5.1. In EM algorithm, the likelihood value of the estimated $\mathbf{\Theta}$ given $\mathbf{X}_{\text{cali}}$ can never decrease as the iteration proceeds; it is learning. And there is no guide or supervisor for the learning; it is unsupervised learning. Therefore, the PPCA model is classified to the unsupervised learning family.

Singular value decomposition (SVD) of $\mathbf{A}$ is useful to show how principal axes are retained in the model. Decomposed $\mathbf{A}$ by SVD results $\mathbf{A}=\sum_{l\in L}\boldsymbol{u}_l\cdot s_l\cdot\boldsymbol{v}_l^{\mathrm{T}}$.[10] So, Eq. (7) is rewritten by the sum of weighted $\boldsymbol{u}_l$, i.e. $\boldsymbol{x}=\sum_l \boldsymbol{u}_l\cdot w_l+\boldsymbol{e}$ where $w_l=s_l\cdot\boldsymbol{v}_l^{\mathrm{T}}\cdot\boldsymbol{z}$. Here, $\boldsymbol{u}_l$ is the $l$th principal axis found by probabilistic way. PCA is achieved by SVD of $\mathbf{X}_{\text{cali}}=\{\boldsymbol{x}_n\}_{n\in N}$: $\mathbf{X}_{\text{cali}}=\sum_{p\in P}\boldsymbol{u}'_p$;

[9] Since $\dim(\boldsymbol{z})<\dim(\boldsymbol{x})$, $\mathbf{A}$ is a thin matrix. So, $(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})^{-1}$ is intractable but $(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\mathbf{\Lambda})^{-1}$ is tractable. When $\dim(\boldsymbol{z})\gg\dim(\boldsymbol{x})$, $(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\mathbf{\Lambda})^{-1}$ is sometimes also intractable. In this case, the matrix inversion lemma, $(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\mathbf{\Lambda})^{-1}=\mathbf{\Lambda}^{-1}-\mathbf{\Lambda}^{-1}\cdot\mathbf{A}\cdot(\mathbf{I}+\mathbf{A}^{\mathrm{T}}\cdot\mathbf{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^{\mathrm{T}}\cdot\mathbf{\Lambda}^{-1}$, is very useful.

[10] Where $\boldsymbol{u}_l\in\mathbb{R}^{\mathrm{P}}$ $s_l\in\mathbb{R}^1$, $\boldsymbol{v}_l\in\mathbb{R}^{\mathrm{L}}$ denote the $l$th left-singular vector, singular value, and right-singular vector of $\mathbf{A}$, respectively.

$s'_p \cdot \boldsymbol{v}'^{\mathrm{T}}_p$.[11] Rewrite it to $\mathbf{X}_{\mathrm{cali}} = \Sigma_{l \in L} \boldsymbol{u}'_l \cdot w'_l + \mathbf{E}$ where $w'_l = s'_l \cdot \boldsymbol{v}'^{\mathrm{T}}_l$ and $\mathbf{E} = \Sigma_{j=L+1,\ldots,P}\ \boldsymbol{u}'_j \cdot s'_j \cdot \boldsymbol{v}'^{\mathrm{T}}_j$, then it indicates that $\mathbf{X}_{\mathrm{cali}}$ can be interpreted as the sum of weighted $\boldsymbol{u}'_l$ plus insignificant term $\mathbf{E}$. Here, $\boldsymbol{u}'_l$ is the $l$th principal axis found by PCA, and is identical to $\boldsymbol{u}_l$ if $\mathbf{A}$ was properly estimated by PPCA. This is the reason that the probabilistic approach to finding $\{\boldsymbol{u}_l\}_{l \in L}$ is called probabilistic PCA.

### 4.2. Probability densities in PPCA model

The most essential, in both PPCA model calibration and the model-based process monitoring, is the evaluation of probability densities of all variables in the model, Eq. (7); not only $p(\boldsymbol{z})$, $p(\boldsymbol{e})$ and $p(\boldsymbol{x})$, but also the posteriors, $p(\boldsymbol{z}\,|\,\boldsymbol{x})$ and $p(\boldsymbol{e}\,|\,\boldsymbol{x})$. Since PPCA already assumes that $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}: \boldsymbol{0}, \mathbf{I})$ and $p(\boldsymbol{e}) = \mathcal{N}(\boldsymbol{e}: \boldsymbol{0}, \lambda \cdot \mathbf{I})$, the remainders are $p(\boldsymbol{x})$, $p(\boldsymbol{z}\,|\,\boldsymbol{x})$ and $p(\boldsymbol{e}\,|\,\boldsymbol{x})$. Evaluation of $p(\boldsymbol{x})$ is simple. Since Gaussian pdf is closed to linear operations, that is, linear transformation of a Gaussian produces another Gaussian, and the sum of two Gaussians generates a new Gaussian, it is obvious that: $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}: \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)$ where $\boldsymbol{\mu}^x = \mathbf{A} \cdot \mathcal{E}[\boldsymbol{z}] + \mathcal{E}[\boldsymbol{e}] = \boldsymbol{0}$, and $\boldsymbol{\Sigma}^x = \mathbf{A} \cdot \mathcal{E}[\boldsymbol{z} \cdot \boldsymbol{z}^{\mathrm{T}}] \cdot \mathbf{A}^{\mathrm{T}} + \mathcal{E}[\boldsymbol{e} \cdot \boldsymbol{e}^{\mathrm{T}}] = \mathbf{A} \cdot \mathbf{A}^{\mathrm{T}} + \lambda \cdot \mathbf{I}$. Thus

$$\mathcal{N}(\boldsymbol{x} : \boldsymbol{0},\ \mathbf{A} \cdot \mathbf{A}^{\mathrm{T}} + \lambda \cdot \mathbf{I}) \tag{8.1}$$

Evaluation of $p(\boldsymbol{z}\,|\,\boldsymbol{x})$ is somewhat tricky. Since the conditional density of two Gaussians is a Gaussian, the posterior of $\boldsymbol{z}$ is also Gaussian: $p(\boldsymbol{z}\,|\,\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}\,|\,\boldsymbol{x}: \boldsymbol{\mu}^{z|x}, \boldsymbol{\Sigma}^{z|x})$ where

$$\begin{aligned}
\boldsymbol{\mu}^{z|x} &= \boldsymbol{\mu}^z + \Sigma^{zx} \cdot \Sigma^{x-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu}^x) \\
&= \mathbf{A}^{\mathrm{T}} \cdot (\mathbf{A} \cdot \mathbf{A}^{\mathrm{T}} + \lambda \cdot \mathbf{I})^{-1} \cdot \boldsymbol{x} \\
&= (\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A} + \lambda \cdot \mathbf{I})^{-1} \cdot \mathbf{A}^{\mathrm{T}} \cdot \boldsymbol{x}
\end{aligned}$$

$$\begin{aligned}
\Sigma^{z|x} &= \Sigma^z - \Sigma^{zx} \cdot \Sigma^{x-1} \cdot \Sigma^{xz} \\
&= \mathbf{I} - \mathbf{A}^{\mathrm{T}} \cdot (\mathbf{A} \cdot \mathbf{A}^{\mathrm{T}} + \lambda \cdot \mathbf{I})^{-1} \cdot \mathbf{A} \\
&= \lambda \cdot (\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A} + \lambda \cdot \mathbf{I})^{-1}
\end{aligned}$$

[11] Where $\boldsymbol{u}'_p \in \mathbb{R}^{\mathrm{P}}$, $s'_p \in \mathbb{R}^1$ and $\boldsymbol{v}'_p \in \mathbb{R}^{\mathrm{N}}$ denote the $p$th left-singular vector, singular value, and right-singular vector of $\mathbf{X}_{\mathrm{cali}}$, respectively.

To simplify the notation, let us define $\mathbf{M} \equiv (\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A} + \lambda \cdot \mathbf{I})^{-1}$. Then we can rewrite it to

$$\mathcal{N}(\boldsymbol{z}\,|\,\boldsymbol{x} : \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}} \cdot \boldsymbol{x},\ \lambda \cdot \mathbf{M}) \tag{8.2}$$

So, $\mathcal{E}[\boldsymbol{z} \cdot \boldsymbol{z}^{\mathrm{T}}\,|\,\boldsymbol{x}] = \boldsymbol{\Sigma}^{z|x} + \boldsymbol{\mu}^{z|x} \cdot (\boldsymbol{\mu}^{z|x})^{\mathrm{T}} = \lambda \cdot \mathbf{M} + \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}} \cdot \boldsymbol{x} \cdot \boldsymbol{x}^{\mathrm{T}} \cdot \mathbf{A} \cdot \mathbf{M}$. Probability density of $\boldsymbol{e}\,|\,\boldsymbol{x}$ is derived from Eqs. (7), (8.1) and (8.2). According to the conditional Gaussian property, it is obvious that $p(\boldsymbol{e}\,|\,\boldsymbol{x}) = \mathcal{N}(\boldsymbol{e}\,|\,\boldsymbol{x}: \boldsymbol{\mu}^{e|x}, \boldsymbol{\Sigma}^{e|x})$ where $\boldsymbol{\mu}^{e|x} = (\mathbf{I} - \mathbf{A} \cdot \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}}) \cdot \boldsymbol{x}$ and $\boldsymbol{\Sigma}^{e|x} = \lambda \cdot \mathbf{A} \cdot \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}}$. Therefore

$$\mathcal{N}(\boldsymbol{e}\,|\,\boldsymbol{x} : (\mathbf{I} - \mathbf{A} \cdot \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}}) \cdot \boldsymbol{x},\ \lambda \cdot \mathbf{A} \cdot \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}}) \tag{8.3}$$

Notice that all the covariance matrices in the model, i.e. $\boldsymbol{\Sigma}^z$, $\boldsymbol{\Sigma}^e$, $\boldsymbol{\Sigma}^x$, $\boldsymbol{\Sigma}^{z|x}$ and $\boldsymbol{\Sigma}^{e|x}$, are not affected by the measurements of $\boldsymbol{x}$; they depend only on the model parameters to be estimated. It implies that $\boldsymbol{x}$ cannot affect the confidence levels of all random variables in the model. Another good property is that the measurement scores can be decomposed to its systematic part and noise part:

$$\begin{aligned}
\boldsymbol{x} &= \mathbf{A} \cdot \mathcal{E}[\boldsymbol{z}\,|\,\boldsymbol{x}] + \mathcal{E}[\boldsymbol{e}\,|\,\boldsymbol{x}] \\
&= (\mathbf{A} \cdot \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}}) \cdot \boldsymbol{x} + (\mathbf{I} - \mathbf{A} \cdot \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}}) \cdot \boldsymbol{x},
\end{aligned}$$

where $\boldsymbol{x} = \mathcal{E}[\boldsymbol{x}\,|\,\boldsymbol{x}]$.

### 4.3. Monitoring chart of latent variable

Since $\mathcal{N}(\boldsymbol{z} \in \mathbb{R}^{\mathrm{L}}: \boldsymbol{0}, \mathbf{I})$ was assumed, the Mahalanobis norm of $\boldsymbol{z}$ is identical to the Euclidian norm. But since the latent variable cannot be measured directly, let us substitute $\boldsymbol{z}$ by its estimate $\underline{\boldsymbol{z}} \equiv \mathcal{E}[\boldsymbol{z}\,|\,\boldsymbol{x}] = \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}} \cdot \boldsymbol{x}$. Then since the squared Mahalanobis norm of $\boldsymbol{z}$ follows $\chi^2_{(L)}$ distribution, we expect $||\underline{\boldsymbol{z}}||^2 \sim \chi^2_{(L)}$. Therefore, for sample $\boldsymbol{x}$, which satisfies Eq. (9.1), we regard the process which generates the sample as in-control with $\alpha$ LoS, and the corresponding happening probability of $\underline{\boldsymbol{z}}$ from the model is given by Eq. (9.2).

$$\|\underline{\boldsymbol{z}}\|^2 \le \chi^{-2}_{(1-\alpha;\ L)} \tag{9.1}$$

$$p(\underline{\boldsymbol{z}}) = (2 \cdot \pi)^{-0.5 \cdot L} \cdot \exp(-0.5 \cdot \|\underline{\boldsymbol{z}}\|^2), \tag{9.2}$$

where Eq. (9.1) is comparable to Hotelling's $\mathcal{T}^2_{(L)}$ test in Eq. (6.1). Moreover, for the sample which

does not hold Eq. (9.1), we can test which elements in $\underline{\boldsymbol{z}}$ are responsible for the out-of-control because all elements in $\underline{\boldsymbol{z}}$, $\{\underline{z}_l\}_{l\in L}$, are mutually uncorrelated, and $\mathcal{N}(\boldsymbol{z}_l: \boldsymbol{0}, 1)\ \forall l$.

$$\underline{z}_l \in \left[\mathcal{N}^{-1}_{s(0.5\cdot\alpha)},\ \mathcal{N}^{-1}_{s(1-0.5\cdot\alpha)}\right] \forall l \tag{9.3}$$

$$p(\underline{z}_l) = (2\cdot\pi)^{-0.5}\cdot\exp(-0.5\cdot\underline{z}_l^2), \tag{9.4}$$

where $\mathcal{N}^{-1}_{s(0.5\cdot\alpha)}$ and $\mathcal{N}^{-1}_{s(1-0.5\cdot\alpha)}$ represent individual latent elements' lower in-control limit and upper in-control limit, respectively.

The contribution of $z_l$ to the out-of-control event $\underline{\boldsymbol{z}}=\{\underline{z}_l\}_{l\in L}$ is inversely proportional to $p(\underline{z}_l)$. Therefore, the contribution of $\underline{z}_j$ to $\underline{\boldsymbol{z}}$, denoted by $\mathbb{C}^z_j$, is given by

$$\mathbb{C}^z_j = p(\underline{z}_j)^{-1}\Sigma_{l\in L}p(\underline{z}_l)^{-1}, \tag{9.5}$$

where $\Sigma_{l\in L}\mathbb{C}^z_j = 1$. It is the test of the latent elements' contributions to the out-of-control event. For example, suppose two-dimensional out-of-control PC score vector $\underline{\boldsymbol{z}}=\{\underline{z}_1, \underline{z}_2\}$ is given, and assume $p(\underline{z}_1)=\mathcal{P}_1 < p(\underline{z}_2)=\mathcal{P}_2$. Then the first PC score, $\underline{z}_1$, is more responsible for the event than $\underline{z}_2$; and hence, $\mathbb{C}^z_1 = \mathcal{P}_1^{-1}\cdot(\mathcal{P}_1^{-1}+\mathcal{P}_2^{-1}) > \mathbb{C}^z_2 = \mathcal{P}_2^{-1}\cdot(\mathcal{P}_1^{-1}+\mathcal{P}_2^{-1})$, and $\mathbb{C}^z_1 + \mathbb{C}^z_2 = 1$.

### 4.4. Monitoring chart of noise variable

Since $\mathcal{N}(\boldsymbol{e}\in\mathbb{R}^{\mathrm{P}}: \boldsymbol{0}, \lambda\cdot\mathbf{I})$ was assumed, the squared Mahalanobis norm of $\boldsymbol{e}$, or the squared Euclidian norm of whitened $\boldsymbol{e}$, $\|\lambda^{-0.5}\cdot\boldsymbol{e}\|^2$, follows $\chi^2_{(P)}$ distribution. As before, we can substitute undetectable $\boldsymbol{e}$ by $\underline{\boldsymbol{e}} \equiv \mathcal{E}[\boldsymbol{e}\,|\,\boldsymbol{x}] = (\mathbf{I}-\mathbf{A}\cdot\mathbf{M}\cdot\mathbf{A}^{\mathrm{T}})\cdot\boldsymbol{x}$, then $\|\lambda^{-0.5}\cdot\underline{\boldsymbol{e}}\|^2 \sim \chi^2_{(P)}$. Using this, in-model test for a sample is derivable. For instance, the developed PPCA model is suitable to explain the process condition by the sample with $\alpha$ LoS if Eq. (10.1) is satisfied; and the happening probability density of $\boldsymbol{e}$ from the model is given by Eq. (10.2).

$$\left\|\lambda^{-0.5}\cdot\underline{\boldsymbol{e}}\right\|^2 \le \chi^{-2}_{(1-\alpha;\ P)} \tag{10.1}$$

$$p(\underline{\boldsymbol{e}}) = (2\cdot\pi\cdot\lambda)^{-0.5\cdot P}\cdot\exp\left(-0.5\cdot\lambda^{-1}\cdot\left\|\underline{\boldsymbol{e}}\right\|^2\right) \tag{10.2}$$

Eq. (10.1) is comparable to the Q test expressed in Eq. (6.2). For an out-of-model event, we can also test which elements are responsible for the event. Since all elements in $\underline{\boldsymbol{e}}$, $\{\underline{e}_p\}_{p\in P}$, are mutually uncorrelated, and $\mathcal{N}(\lambda^{-0.5}\cdot\underline{\boldsymbol{e}}_p: 0, 1)\ \forall p$, if the $p$th noise element's score does not hold Eq. (10.3), then the element is responsible for the out-of-model event with $\alpha$ LoS, and the corresponding happening probability density is given by Eq. (10.4).

$$\underline{e}_p \in \left[\lambda^{0.5}\cdot\mathcal{N}^{-1}_{s(0.5\cdot\alpha)},\ \lambda^{0.5}\cdot\mathcal{N}^{-1}_{s(1-0.5\cdot\alpha)}\right] \tag{10.3}$$

$$p(\underline{e}_p) = (2\cdot\pi\cdot\lambda)^{-0.5}\cdot\exp(-0.5\cdot\lambda^{-1}\cdot\underline{\boldsymbol{e}}_p^2), \tag{10.4}$$

where $\lambda^{0.5}\cdot\mathcal{N}^{-1}_{s(0.5\cdot\alpha)}$ and $\lambda^{0.5}\cdot\mathcal{N}^{-1}_{s(1-0.5\cdot\alpha)}$ indicate the lower in-model limit and upper in-model limit of $\{\underline{e}_p\}_{p\in P}$, respectively. Similar to Eq. (9.5), the contribution of $\underline{e}_j$ to $\underline{\boldsymbol{e}}=\{\underline{e}_p\}_{p\in P}$, denoted by $\mathbb{C}^e_j$, is given by

$$\mathbb{C}^e_j = p(\underline{\boldsymbol{e}}_j)^{-1}\cdot\Sigma_{p\in P}p(\underline{\boldsymbol{e}}_p)^{-1}, \tag{10.5}$$

where $\Sigma_{p\in P}\ \mathbb{C}^e_p = 1$. Notice that the individual elements-based normality test discussed in the univariate Shewhart chart section (Eq. (2.1)) is decomposed into the test of its systematic part (Eq. (9.3)) and noise part (Eq. (10.3)). And the probability density which occurs, $x_p$, in Eq. (2.2), is partitioned into the density of its systematic part (Eq. (9.4)) and noise part (Eq. (10.4)).

### 4.5. Monitoring chart of measurement variable

In the PPCA model, the covariance matrix of $\boldsymbol{x}$ is expressed by $\mathbf{A}$ and $\lambda$, i.e. $\mathcal{N}(\boldsymbol{x}: \boldsymbol{0}, \boldsymbol{\Sigma}^x = \mathbf{A}\cdot\mathbf{A}^{\mathrm{T}} + \lambda\cdot\mathbf{I})$. And the squared Mahalanobis norm of $\boldsymbol{x}$ follows $\chi^2_{(P)}$ distribution. So, for the sample satisfied by Eq. (11.1), we regard the process which generates the sample as normal with $\alpha$ LoS; and the corresponding probability density of the sample is given by Eq. (11.2).

$$\left\|(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\lambda\cdot\mathbf{I})^{-0.5}\cdot\boldsymbol{x}\right\|^2 \le \chi^{-2}_{(1-\alpha;\ P)} \tag{11.1}$$

$$p(\boldsymbol{x}) = (2\cdot\pi)^{-0.5P}\cdot\det[(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\lambda\cdot\mathbf{I})^{-1}]^{-0.5} \cdot\exp(-0.5\cdot\boldsymbol{x}^{\mathrm{T}}\cdot(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\lambda\cdot\mathbf{I})^{-1}\cdot\boldsymbol{x}) \tag{11.2}$$

Here, $(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\lambda\cdot\mathbf{I})^{-1}$ is often intractable if $\dim(\boldsymbol{x})\gg\dim(\boldsymbol{z})$. But there is an efficient way to the type of inversion using the following matrix inversion lemma:[12]

$$(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\boldsymbol{\Lambda})^{-1}=\boldsymbol{\Lambda}^{-1}-\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot(\mathbf{I}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}$$

Using the lemma, the following results:

$$(\boldsymbol{\Sigma}^{x})^{-1}=(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\lambda\cdot\mathbf{I})^{-1}=\lambda^{-1}\cdot(\mathbf{I}-\mathbf{A}\cdot(\mathbf{A}^{\mathrm{T}}\cdot\mathbf{A}+\lambda\cdot\mathbf{I})^{-1}\cdot\mathbf{A}^{\mathrm{T}})$$

Notice that $(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\lambda\cdot\mathbf{I})^{-1}\in\mathbb{R}^{\mathrm{P}\times\mathrm{P}}$ but $(\mathbf{A}^{\mathrm{T}}\cdot\mathbf{A}+\lambda\cdot\mathbf{I})^{-1}\in\mathbb{R}^{\mathrm{L}\times\mathrm{L}}$. If estimated $\mathbf{A}$ and $\lambda$ are proper, then Eqs. (11.1) and (11.2) are identical to Eqs. (3.1) and (3.2), respectively; in addition, if sufficient samples were used to estimate $\boldsymbol{\Sigma}^{x}$, then Eqs. (3.1), (4) and (11.1) are all equivalent.

Since we were given all probabilistic information on every variable in the model, various types of control charts are possible. For example, we may introduce $\mathcal{F}$ distribution to the process control chart. Since $\|\underline{\boldsymbol{z}}\|^2\sim\chi^2_{(L)}$, $\|\lambda^{-0.5}\cdot\underline{\boldsymbol{e}}\|^2\sim\chi^2_{(P)}$, and $(\chi^2_{(L)}/L)\cdot(\chi^2_{(P)}/P)^{-1}\sim\mathcal{F}_{(L,P)}$ by definition of $\mathcal{F}$ distribution, it is apparent that $(P\cdot\|\underline{\boldsymbol{z}}\|^2)\cdot(L\cdot\|\lambda^{-0.5}\cdot\underline{\boldsymbol{e}}\|^2)^{-1}\sim\mathcal{F}_{(L,P)}$. Using the distribution, we can decide whether the abnormality of $\boldsymbol{x}$ comes from its systematic part or noise part, e.g, if Eq. (11.3) is satisfied, then the abnormality is mainly due to the systematic part of $\boldsymbol{x}$ with $0.5\cdot\alpha$ LoS, and if Eq. (11.4) holds, then the abnormality is caused by the noise part of $\boldsymbol{x}$ with $0.5\cdot\alpha$ LoS.

$$\lambda\cdot(P/L)\cdot\|\underline{\boldsymbol{z}}\|^2\cdot\|\underline{\boldsymbol{e}}\|^{-2}>\mathcal{F}^{-1}_{(1-0.5\cdot\alpha;\ L,P)}\qquad(11.3)$$

$$\lambda\cdot(P/L)\cdot\|\underline{\boldsymbol{z}}\|^2\cdot\|\underline{\boldsymbol{e}}\|^{-2}<\mathcal{F}^{-1}_{(0.5\cdot\alpha;\ L,P)},\qquad(11.4)$$

where $\underline{\boldsymbol{z}}=(\mathbf{A}^{\mathrm{T}}\cdot\mathbf{A}+\lambda\cdot\mathbf{I})^{-1}\cdot\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{x}$, and $\underline{\boldsymbol{e}}=(\mathbf{I}-\mathbf{A}\cdot(\mathbf{A}^{\mathrm{T}}\cdot\mathbf{A}+\lambda\cdot\mathbf{I})^{-1}\cdot\mathbf{A}^{\mathrm{T}})\cdot\boldsymbol{x}$.

## 5. PPCA model calibration

All statistical decisions in the PPCA model (Eqs. (9.1)–9.5) Eqs. (10.1)–10.5) Eqs. (11.1)–(11.4)) are wholly dependent on the adequate estimate of the model parameters, $\mathbf{A}$ and $\lambda$; under appropriate decision of latent variable dimension, $\dim(\boldsymbol{z})=L$. In the article, the parameters are estimated by the expectation and maximization (EM) algorithm [20]; and the dimension is decided by the variance explanation ratios, Eqs. (14.1) and (14.2), according to the increase in the number of PCs retained in the model.

### 5.1. EM for PPCA

McLachlan and Krishnan [21] devote an entire book to EM. In brief, EM is an iterative maximization algorithm of complete data log likelihood function. To explain how EM works, let us denote the parameter set in the model as $\Theta$, log likelihood of the $i$th estimated $\Theta$ as $\mathcal{L}(\Theta_i)$, log likelihood of a new estimated $\Theta$ as $\mathcal{L}(\Theta)$, and variation of the two log likelihoods as $\Delta\mathcal{L}$. Since $\mathcal{L}(\cdot)\equiv\log\ p(\mathbf{X}\colon\cdot)$, the following are easily derived using Bay's rule.

$$\begin{aligned}\Delta\mathcal{L}&=\mathcal{L}(\Theta)-\mathcal{L}(\Theta_i)\\&=\log[p(\mathbf{X}:\Theta)\cdot\ p(\mathbf{X}:\Theta_i)^{-1}]\\&=\log\int p(\boldsymbol{z}\,|\,\mathbf{X}:\Theta_i)\cdot p(\boldsymbol{z},\mathbf{X}:\Theta)\cdot p(\boldsymbol{z}\,|\,\mathbf{X}:\Theta_i)^{-1}d\boldsymbol{z}\\&=\log\mathcal{E}_{\boldsymbol{z}\,|\,\mathbf{X};\Theta_i}[p(\boldsymbol{z},\mathbf{X}:\Theta)\cdot p(\boldsymbol{z},\mathbf{X}:\Theta_i)^{-1}]^2\end{aligned}\qquad(12.1)$$

[12] Proof: for any invertible $\mathbf{U}$, and diagonal matrix $\boldsymbol{\Lambda}$: (in this case $\mathbf{U}=\mathbf{I}$, $\boldsymbol{\Lambda}=\lambda\cdot\mathbf{I}$):

$$\begin{aligned}\mathbf{I}&=[\boldsymbol{\Lambda}^{-1}-\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}]\cdot[\boldsymbol{\Lambda}+\mathbf{A}\cdot\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}]\\&=\mathbf{I}+\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}-\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^{\mathrm{T}}-\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}\\&\quad-\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^{\mathrm{T}}\\&=\mathbf{I}+\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot[\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}-(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^{\mathrm{T}}\\&\quad-(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}]\\&=\mathbf{I}+\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot[\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}-(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{U}^{-1}\cdot\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}\\&\quad-(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}]\\&=\mathbf{I}+\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A}\cdot[\mathbf{I}-(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})^{-1}\cdot(\mathbf{U}^{-1}+\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{A})]\cdot\mathbf{U}\cdot\mathbf{A}^{\mathrm{T}}\\&=\mathbf{I}+\mathbf{0}\end{aligned}$$

Notice that Eq. (12.1) includes probability density information of latent variable $z$ to evaluate $\Delta\mathcal{L}$. Let us define $Q^{o}(\Theta \mid \Theta_i)$ as follows:

$$\begin{aligned} Q^{o}(\Theta \mid \Theta_i) &\equiv \mathcal{E}_{z \mid \mathbf{X}:\Theta_i}[\log(p(z, \mathbf{X} : \Theta) \\ &\quad \cdot p(z, \mathbf{X} : \Theta_i)^{-1})] \\ &= \int p(z \mid \mathbf{X} : \Theta_i) \cdot \log[p(z, \mathbf{X} : \Theta) \\ &\quad \cdot p(z, \mathbf{X} : \Theta_i)^{-1})] dz \end{aligned} \tag{12.2}$$

Then $Q^{o}$ is to be the lower bound of $\Delta\mathcal{L}$, i.e. $\Delta\mathcal{L} \geq Q^{o} \forall\Theta$; Jensen's inequality [22] can prove it.[13] Suppose $\mathbf{X}=\{\boldsymbol{x}_n\}_{n\in N}$ is an independent and identically distributed (iid)[14] sample set. And the next estimate of $\Theta$, $\Theta_{i+1}$, is the solution of the following maximization problem:

$$\begin{aligned} \Theta_{i+1} &= \arg_{\Theta}\max : \ Q^{o}(\Theta \mid \Theta_i) \\ &= \arg_{\Theta}\max : \ \int p(z \mid \mathbf{X} : \Theta_i) \cdot \log p(z, \mathbf{X} : \Theta) dz \\ &= \arg_{\Theta}\max : \ \Sigma_{n\in N} \int p(z \mid \boldsymbol{x}_n : \Theta_i) \\ &\quad \cdot \log p(z, \boldsymbol{x}_n : \Theta) dz \\ &= \arg_{\Theta}\max : \ \Sigma_{n\in N} \mathcal{E}_{z \mid x_n:\Theta_i}[\log p(z, \boldsymbol{x}_n : \Theta)] \end{aligned} \tag{12.3}$$

It is obvious that $0 = Q^{o}(\Theta_i \mid \Theta_i) \leq Q^{o}(\Theta_{i+1} \mid \Theta_i)$; moreover, $Q^{o}(\Theta_{i+1} \mid \Theta_i) \leq \Delta\mathcal{L} = \mathcal{L}(\Theta_{i+1}) - \mathcal{L}(\Theta_i)$. Therefore $\mathcal{L}(\Theta_{i+1})$ cannot be smaller than $\mathcal{L}(\Theta_i)$ if $\Theta_{i+1}$ is the solution of Eq. (12.3). That means the next estimate, $\Theta_{i+1}$, must be the better solution candidate than the previous estimate, $\Theta_i$, $\forall i$.

Eq. (12.3) can be rewritten by introducing the energy functions, $Q$ for $\mathbf{X}$, and $q_n$ for $\boldsymbol{x}_n$:

$$Q(\Theta \mid \Theta_i) \equiv \Sigma_{n\in N} q_n(\Theta \mid \Theta_i),$$

where $q_n(\Theta \mid \Theta_i) \equiv \mathcal{E}_{z \mid xn:\Theta i}[\log p(z, \boldsymbol{x}_n : \Theta)] = \mathcal{E}_{z \mid xn:\Theta i}[\log p(\boldsymbol{x}_n \mid z : \Theta_x) \cdot p(z : \Theta_z)]$. Then the maximization problem, Eq. (12.3), is solved via two steps: expectation step, E-step; and maximization step, M-step. The E-step aims to evaluate $q_n(\Theta \mid \Theta_i)$, and the M-step seeks to find $\Theta$ which maximizes $Q(\Theta \mid \Theta_i)$. If we iterate E-step and M-step until $\Theta$ converges, then all parameters in the model should have their most probable values regardless of their initial values since $\Delta\mathcal{L}$ cannot have negative value for all the iterations. It may be the most simplified explanation of the EM algorithm.

In the PPCA model, since $p(z : \Theta_z) = \mathcal{N}(z : \boldsymbol{0}, \mathbf{I})$, and $p(\boldsymbol{x} \mid z : \Theta_x) = \mathcal{N}(\boldsymbol{x} \mid z : \boldsymbol{\mu}^{x \mid z}, \boldsymbol{\Sigma}^{x \mid z})$ where $\boldsymbol{\mu}^{x \mid z} = \mathcal{E}[\boldsymbol{x} \mid \boldsymbol{z}] = \mathbf{A} \cdot \boldsymbol{z}$, and $\boldsymbol{\Sigma}^{x \mid z} = \mathcal{E}[\boldsymbol{x} \cdot \boldsymbol{x}^{\mathrm{T}} \mid \boldsymbol{z}] - \mathcal{E}[\boldsymbol{x} \mid \boldsymbol{z}] \cdot \mathcal{E}[\boldsymbol{x} \mid \boldsymbol{z}]^{\mathrm{T}} = \lambda \cdot \mathbf{I}$, the energy function is to be

$$\begin{aligned} Q(\Theta \mid \Theta_i) &= \Sigma_{n\in N} \mathcal{E}_{z \mid x_n:\Theta_i}[\log \mathcal{N}(\boldsymbol{x}_n \mid z : \mathbf{A} \cdot \boldsymbol{z}, \lambda \cdot \mathbf{I}) \\ &\quad \cdot \mathcal{N}(z : \boldsymbol{0}, \mathbf{I})] \\ &= \Sigma_{n\in N} \mathcal{E}_{z \mid \boldsymbol{x}_n:\Theta_i}[\log \det(\lambda \cdot \mathbf{I})^{-1} \\ &\quad -(\boldsymbol{x}_n - \mathbf{A} \cdot \boldsymbol{z})^{\mathrm{T}} \cdot (\lambda \cdot \mathbf{I})^{-1} \\ &\quad \cdot (\boldsymbol{x}_n - \mathbf{A} \cdot \boldsymbol{z})] + \delta, \end{aligned}$$

where $\delta$ signifies $\Theta$ independent terms in the function. So, E-step in PPCA model is the evaluation of both $\mathcal{E}_{z \mid xn:\Theta i}[\boldsymbol{z}] = \mathcal{E}[z \mid \boldsymbol{x}_n : \mathbf{A}_i, \lambda_i]$ and $\mathcal{E}_{z \mid xn:\Theta i}[\boldsymbol{z} \cdot \boldsymbol{z}^{\mathrm{T}}] = \mathcal{E}[\boldsymbol{z} \cdot \boldsymbol{z}^{\mathrm{T}} \mid \boldsymbol{x}_n : \mathbf{A}_i, \lambda_i]$. However, we have already evaluated them in Eq. (8.2), i.e. $\mathcal{N}(z \mid \boldsymbol{x} : \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}} \cdot \boldsymbol{x}, \lambda \cdot \mathbf{M})$. Thus

$$\underline{z}_n \equiv \mathcal{E}_{z \mid xn}[\boldsymbol{z}] = \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}} \cdot \boldsymbol{x}_n \tag{13.1}$$

$$\underline{zz}_n \equiv \mathcal{E}_{z \mid x_n}[z \cdot z^{\mathrm{T}}] = \lambda \cdot \mathbf{M} + \mathbf{M} \cdot \mathbf{A}^{\mathrm{T}} \cdot \boldsymbol{x}_n \cdot \boldsymbol{x}_n^{\mathrm{T}} \cdot \mathbf{A} \cdot \mathbf{M}, \tag{13.2}$$

where $\mathbf{M} = (\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A} + \lambda \cdot \mathbf{I})^{-1}$. In the M-step, optimal $\Theta = \{\mathbf{A}, \lambda\}$ which maximizes $Q(\Theta \mid \Theta_i)$ is given by

$$\begin{aligned} &\boldsymbol{0} = (\partial / \partial \mathbf{A}) \cdot [Q(\Theta \mid \Theta_i)] \Rightarrow [15] \\ &\mathbf{A} = \left(\Sigma_{n\in N} \boldsymbol{x}_n \cdot \underline{z}_n^{\mathrm{T}}\right) \cdot \left(\Sigma_{n\in N} \underline{zz}_n\right)^{-1} \\ &0 = (\partial / \partial \lambda^{-1}) \cdot [Q(\Theta \mid \Theta_i)] \Rightarrow [16] \\ &\lambda = (\dim(\boldsymbol{x}) \cdot N)^{-1} \cdot \Sigma_{n\in N} \mathrm{tr}(\boldsymbol{x}_n \cdot \boldsymbol{x}_n^{\mathrm{T}} - \mathbf{A} \cdot \underline{z}_n \cdot \boldsymbol{x}_n^{\mathrm{T}}), \end{aligned} \tag{13.3, 13.4}$$

where tr[·] signifies the trace operator, and $\mathbf{A}$ in Eq. (13.4) should be the resultant of Eq. (13.3). EM is

[13] Jensen's inequality indicates that $\mathcal{E}[f(\boldsymbol{w})] \geq f(\mathcal{E}[\boldsymbol{w}])$ for any random variable $\boldsymbol{w}$ if $f$ is a convex function. But since 'log' is a concave function, $\mathcal{E}[\log(\boldsymbol{w})] \leq \log(\mathcal{E}[\boldsymbol{w}])$ results.

[14] Mathematically, iid on $\mathbf{X}$ implies that $p(\boldsymbol{x}_i \mid \mathbf{X} \backslash \boldsymbol{x}_i)^3 = p(\boldsymbol{x}_i) = p(\boldsymbol{x}_j) \forall i, j \in N$; where first equality indicates *independent* and second identity represents *identically distributed*.

[15] $(\partial/\partial \mathbf{A}) \cdot (\boldsymbol{x} - \mathbf{A} \cdot \boldsymbol{z})^{\mathrm{T}} \cdot \mathbf{\Lambda}^{-1} \cdot (\boldsymbol{x} - \mathbf{A} \cdot \boldsymbol{z}) = -2 \cdot \mathbf{\Lambda}^{-1} \cdot (\boldsymbol{x} - \mathbf{A} \cdot \boldsymbol{z}) \cdot \boldsymbol{z}^{\mathrm{T}}$ for symmetric $\mathbf{\Lambda}$. Here, $\mathbf{\Lambda} = \lambda \cdot \mathbf{I}$.

[16] $(\partial/\partial \mathbf{\Lambda}^{-1}) \cdot [\log \det(\mathbf{\Lambda}^{-1})] = \mathbf{\Lambda}^{\mathrm{T}}$, and $(\partial/\partial \mathbf{\Lambda}^{-1}) \cdot [(\boldsymbol{x} - \mathbf{A} \cdot \boldsymbol{z})^{\mathrm{T}} \cdot \mathbf{\Lambda}^{-1} \cdot (\boldsymbol{x} - \mathbf{A} \cdot \boldsymbol{z})] = (\boldsymbol{x} - \mathbf{A} \cdot \boldsymbol{z}) \cdot (\boldsymbol{x} - \mathbf{A} \cdot \boldsymbol{z})^{\mathrm{T}}$.

an iterative algorithm, so E-step, Eqs. (13.1) and (13.2), and M-step, Eqs. (13.3) and (13.4), are calculated iteratively until both $\mathbf{A}$ and $\lambda$ converge.

In summary, when we were given a calibration sample set $\mathbf{X}_{\text{cali}}=\{\boldsymbol{x}_n\}_{n\in N}$, PPCA model is calibrated by EM algorithm as follows:

Set initial $\Theta = \{\mathbf{A}, \lambda\}$

Until $\Theta$ converge,

$\mathbf{M} = (\mathbf{A}^{\mathrm{T}}\,\mathbf{A} + \lambda\,\mathbf{I})^{-1}$

For $n \in N$,

$\underline{z}_n = \mathbf{M}\cdot\mathbf{A}^{\mathrm{T}}\cdot\boldsymbol{x}_n$ ; $\underline{zz}_n = \lambda\cdot\mathbf{M} + \underline{z}_n\cdot\underline{z}_n^{\mathrm{T}}$ [E-step]

End for

$\mathbf{A} = (\sum_{n\in N}\boldsymbol{x}_n\cdot\underline{z}_n^{\mathrm{T}})\cdot(\sum_{n\in N}\underline{zz}_n)^{-1}$ ; $\lambda = (\dim(\boldsymbol{x})\cdot N)^{-1}\cdot\sum_{n\in N}\ \mathrm{tr}(\boldsymbol{x}_n\cdot\boldsymbol{x}_n^{\mathrm{T}} - \mathbf{A}\cdot\underline{z}_n\cdot\boldsymbol{x}_n^{\mathrm{T}})$ [M-step]

End until

### 5.2. Number of principal components

As Eq. (8.1) indicates, PPCA model approximates the covariance matrix of $\boldsymbol{x}$ as the sum of its systematic part $\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}$, and noise part $\lambda\cdot\mathbf{I}$, that is, $\mathbf{\Sigma}^{x}=\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\lambda\cdot\mathbf{I}$. It is obvious that if we increase the number of PCs retained in the model, then the systematic part will converge to $\mathbf{\Sigma}^{x}$, whereas the noise part will go to $\mathbf{0}$. Using the property, we may decide $\dim(\boldsymbol{z})=l$ by one of the following three ways. After setting a small enough decision criterion $\varepsilon>0$,

(1) Since $\Delta\lambda(l) \equiv \lambda(l) - \lambda(l+1) \geq 0$ for $l\in(\dim(\boldsymbol{x}) - 1)$, we may decide the number via:

Calculating $\lambda(l)$ for $l\in\dim(\boldsymbol{x})$, then select $l$ at which $\Delta\lambda(l) \leq \varepsilon$.

(2) Let us denote systematic part of the $p$th variable in $\boldsymbol{x}$ as $x_p^{\mathrm{s}}$, then the variance explanation ratio of $x_p$ is defined by

$$r_p \equiv \mathrm{var}(x_p^{\mathrm{s}})\cdot\mathrm{var}(x_p)^{-1} = \mathrm{diag}(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})_p\cdot[\mathrm{diag}(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})_p + \lambda]^{-1}, \qquad (14.1)$$

where $\mathrm{diag}(\cdot)_p$ and $\mathrm{diag}(\cdot)_p^{-1}$ denote the $p$th diagonal element, and its inverse, respectively. Since $\Delta r_p(l) \equiv r_p(l+1) - r_p(l) \geq 0 \forall p$ and $l\in(\dim(\boldsymbol{x}) - 1)$, we may decide the number via:

Calculate $\{r_p(l)\}_{p\in P}$ for $l\in\dim(\boldsymbol{x})$, then select $l$ at which $\max(\{\Delta r_p(l)\}_{p\in P}) \leq \varepsilon$.

(3) Using Eq. (14.1), let us define the average variance explanation ratio such as

$$r_{\mathrm{avg}} \equiv P^{-1}\cdot\sum_{p\in P} r_p \qquad (14.2)$$

Since $\Delta r_{\mathrm{avg}}(l) \equiv r_{\mathrm{avg}}(l+1) - r_{\mathrm{avg}}(l) \geq 0$ for $l\in(\dim(\boldsymbol{x})- 1)$, we can decide the number via:

Calculate $r_{\mathrm{avg}}(l)$ for $l\in\dim(\boldsymbol{x})$, then select $l$ at which $\Delta r_{\mathrm{avg}}(l) \leq \varepsilon$.

Notice that the total variance explanation ratio in PCA model can be expressed by

$$r_{\mathrm{PCA}} = \left[\sum_{p\in P}\mathrm{var}(x_p^{\mathrm{s}})\right]\cdot\left[\sum_{p\in P}\mathrm{var}(x_p)\right]^{-1} = \left[\sum_{p\in P}\mathrm{diag}(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})_p\right]\cdot\left[\sum_{p\in P}\mathrm{diag}(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})_p + \lambda\right]^{-1}$$

So, there is an important difference to decide the number between PPCA and PCA, that is

$$r_{\mathrm{avg}} = P^{-1}\cdot\sum_{p\in P}[\mathrm{var}(x_p^{\mathrm{s}})\cdot\mathrm{var}(x_p)^{-1}] \text{ vs.}$$

$$r_{\mathrm{PCA}} = \left[\sum_{p\in P}\mathrm{var}(x_p^{\mathrm{s}})\right]\cdot\left[\sum_{p\in P}\mathrm{var}(x_p)\right]^{-1}$$

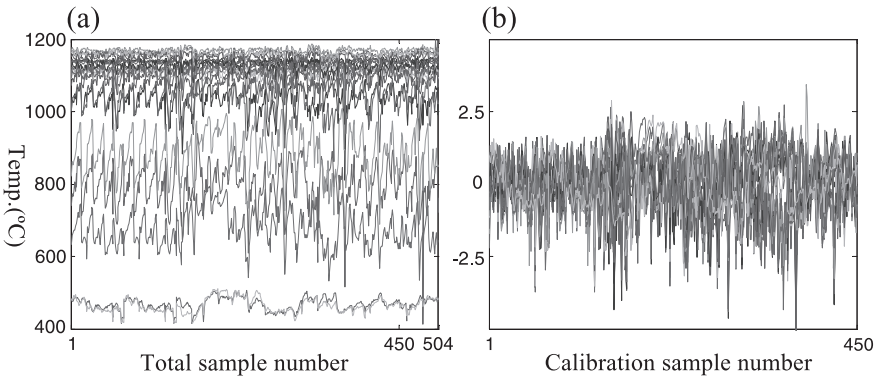


Fig. 1. Data patterns (a) Raw data. (b) auto-scaled calibration data set.

By the difference, PPCA can predict how much variances of the individual elements in $\boldsymbol{x}$ will be explained by the model while PCA cannot.

## 6. Example

As an illustrative example, we applied the proposed method to the slurry-fed ceramic melter (SFCM) data set[17] [23]. SFCM data set consists of 504 samples, each of which has 20 temperature measurements; let us denote the set as $\mathbf{X}^{o}=\{\boldsymbol{x}_n^{o}\}_{n\in 504}\in\mathbb{R}^{20\times 504}$. We partitioned the set into two subsets: calibration set $\mathbf{X}_C^{o}=\{\boldsymbol{x}_{pn}^{o}\}_{p\in 20;\ n\in 450}\in\mathbb{R}^{20\times 450}$, and test set $\mathbf{X}_T^{o}=\{\boldsymbol{x}_{pn}^{o}\}_{p\in 20;\ n=451,\ldots,504}\in\mathbb{R}^{20\times 54}$; and hence $\mathbf{X}_C^{o}\cup\mathbf{X}_T^{o}=\mathbf{X}^{o}$ and $\mathbf{X}_C^{o}\cap\mathbf{X}_T^{o}=\{\phi\}$. Auto-scaled calibration data set, $\mathbf{X}_C=\{x_{pn}\}_{p\in 20;\ n\in 450}$,[18] was used to calibrate both PCA and PPCA models. Fig. 1(a) and (b) show data patterns of the sets.

The test set, $\mathbf{X}_T^{o}$, was also rescaled to $\mathbf{X}_T=\{x_{pn}\}_{p\in 20;\ n=451,\ldots,504}$[19] regarding the means and variances of the calibration set. Finally, let us define a mixed sample set $\mathbf{X}'\equiv\{\boldsymbol{x}_n\}_{n=401,\ldots,450}\cup\mathbf{X}_T$. Notice that $\{\boldsymbol{x}_n\}_{n=401,\ldots,450}\subset\mathbf{X}_C$, and $\mathbf{X}'\in\mathbb{R}^{20\times 104}$.

An appropriate number of PCs in the PCA model can be decided by the scree-plot, which is the plot of large size-ordered eigenvalues of $(449^{-1}\cdot\mathbf{X}_C\cdot\mathbf{X}_C^{T})$. Fig. 2(a) is the plot; however, the proper number is not clear in the plot.

In case of PPCA model, the number, $\dim(\boldsymbol{z})$, can be determined by one of the three methods proposed in Section 5.2: (1) $\Delta\lambda(l)\le\varepsilon$, (2) $\max(\{\Delta r_p(l)\}_{p\in P})\le\varepsilon$, or (3) $\Delta r_{\mathrm{avg}}(l)\le\varepsilon$, for $l\in\dim(\boldsymbol{x})$. Fig. 2(b), (c), and (d) are the result of the first, second, and third methods, respectively. The number is also not clear in both the first and third methods as shown in Fig. 2(b) and (d). But the second method, which is the elements' variances explanation ratios method, apparently suggests the number four as disclosed in Fig. 2(c).

As Fig. 2(c) indicates, the 10th and 20th measurement variables, $x_{10}$ and $x_{20}$, will not be regressed properly if the number is less than four. Notice that if we are not concerned with either $x_{10}$ or $x_{20}$, then a three-PC model would be enough, but if we regard the two elements, at least a four-PC model should be selected. When we use the four-PC model, we can expect that about 80% of the variances of $\{x_p\}_{p\in 20}$ are explained by the model.

Under a four-PC model assumption, PPCA-based process monitoring charts are compared with the PCA-based charts; i.e. Q chart to test in-model, and $\mathcal{T}^2$ chart to test in-control. Fig. 3(a) and (b) are the results of Q and $\mathcal{T}^2$ chart of $\mathbf{X}'$, respectively. In a similar manner, there are also two charts in the PPCA-based method: $\chi^2_{(20)}$ chart using Eq. (10.1) to check the in-model, Fig. 3(c); and $\chi^2_{(4)}$ chart using Eq. (9.1) to

[17] Temperature variable of 'repdata.mat' in PLS toolbox ver.2.1 were used to the data set.

[18] Let us define the $p$th mean and variance of $\boldsymbol{x}^{o}$ as $m_p\equiv 450^{-1}\cdot\Sigma_{n\in 450}x_{pn}^{o}$, and $v_p\equiv 449^{-1}\cdot\Sigma_{n\in 450}(x_{pn}^{o}-m_p)^2$, respectively; then $\mathbf{X}_C$ is obtained from $x_{pn}=v_p^{-0.5}\cdot(x_{pn}^{o}-m_p)$ for $p\in 20$, $n\in 450$.

[19] The rescaled set, $\mathbf{X}_T$, is obtained from $x_{pn}=v_p^{-0.5}\cdot(x_{pn}^{o}-m_p)$ for $p\in 20$, $n\in[451, 452,\ldots, 504]$.

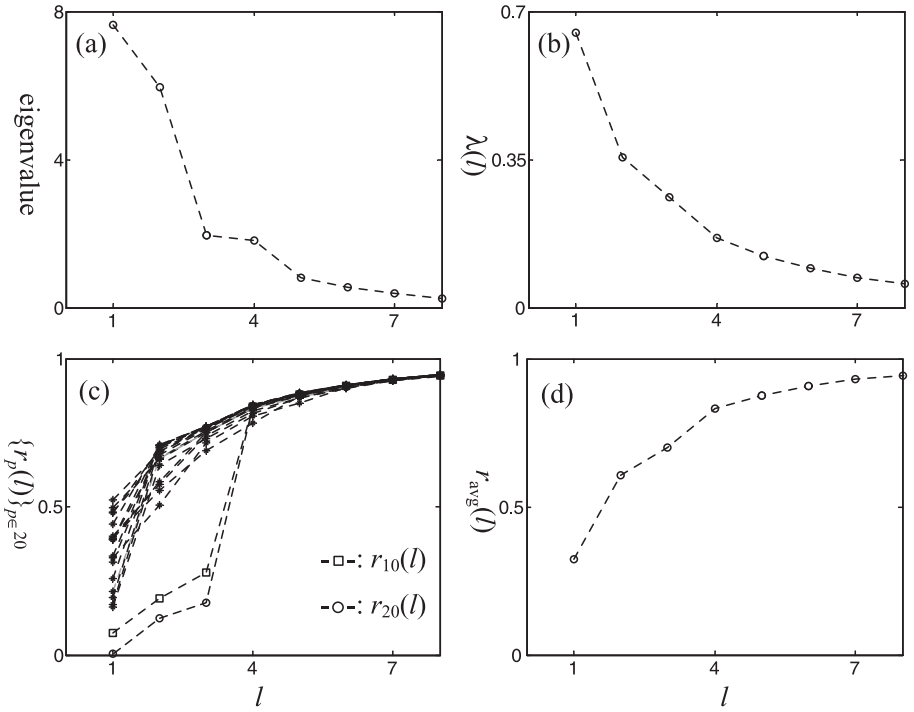


Fig. 2. Number of PCs retained in PCA and PPCA models. (a) PCA: eigenvalues of sample covariance matrix of auto-scaled calibration data set. (b) PPCA: variance of noise variable, $\lambda(l)$, according to increase the number of PCs retained, $l$. (c) PPCA: variance explanation ratios of $\{x_p\}_{p\in 20}$. (d) PPCA: averaged variance explanation ratio, $r_{\text{avg}}=20^{-1}\cdot\Sigma_{p\in 20}r_p$.

check the in-control (Fig. 3(d)). Notice that the $\chi^2_{(20)}$ chart has an equivalent pattern to the Q chart, and the $\chi^2_{(4)}$ chart is practically identical to the $T^2$ chart. And both the $\chi^2$ charts are worked with the same measuring unit, i.e. squared Mahalanobis norm, while Q and $T^2$ charts are govern by different measuring units, i.e. squared Euclidian norm in Q and squared Mahalanobis norm in $T^2$.

Fig. 3(a) or (c) indicates that there are several out-of-model samples in $\mathbf{X}'$, e.g. $\boldsymbol{x}_{484}$ and $\boldsymbol{x}_{501-504}$. And Fig. 3(b) or (d) displays that there are two out-of-control events in $\mathbf{X}'$, i.e. $\boldsymbol{x}_{423}$ and $\boldsymbol{x}_{504}$.

Since the proposed method uses the same measuring unit for all variables in the model, various types of control charts are possible. For instance, we can test which components are responsible for the out-of-model using Eq. (10.3), and calculate the relative responsibilities of individual elements to the abnormal event by Eq. (10.5). Fig. 4(a) indicates that the first out-of-model event, $\boldsymbol{x}_{484}$, comes from the ninth element in $\boldsymbol{x}$, $\boldsymbol{x}_9$; and the other out-of-model samples, $\boldsymbol{x}_{501-504}$, are mainly due to $\boldsymbol{x}_5$. Fig. 4(b) shows that the first out-of-control event, $\boldsymbol{x}_{423}$, is the result of both the third and fourth PCs; and the out-of-control of the 504th event, $\boldsymbol{x}_{504}$, is caused by the second PC. But notice that the PC scores extracted from the out-of-model samples, $\underline{\boldsymbol{z}}_{484}$ and $\underline{\boldsymbol{z}}_{501-504}$, are not suitable for application in the in-control test; in other words, we cannot believe that $\underline{\boldsymbol{z}}_{484}$ and $\underline{\boldsymbol{z}}_{501-504}$ are the true PC scores as discussed in Section 3.

Fig. 4(c), obtained from Eq. (11.1), is the control chart for the measurement variable $\boldsymbol{x}$. In fact, the chart is the combined expression of Fig. 4(a) and (b), or Fig. 3(c) and (d). In theory, the chart is identical to the multivariate Shewhart chart; however, we may not worry about the inverse problem of the covariance matrix of $\boldsymbol{x}$ to the chart as we explained in Section 4.5.

Fig. 4(d), made by Eqs. (11.3) and (11.4), is another expression of Fig. 4(c); at the same time, the chart is

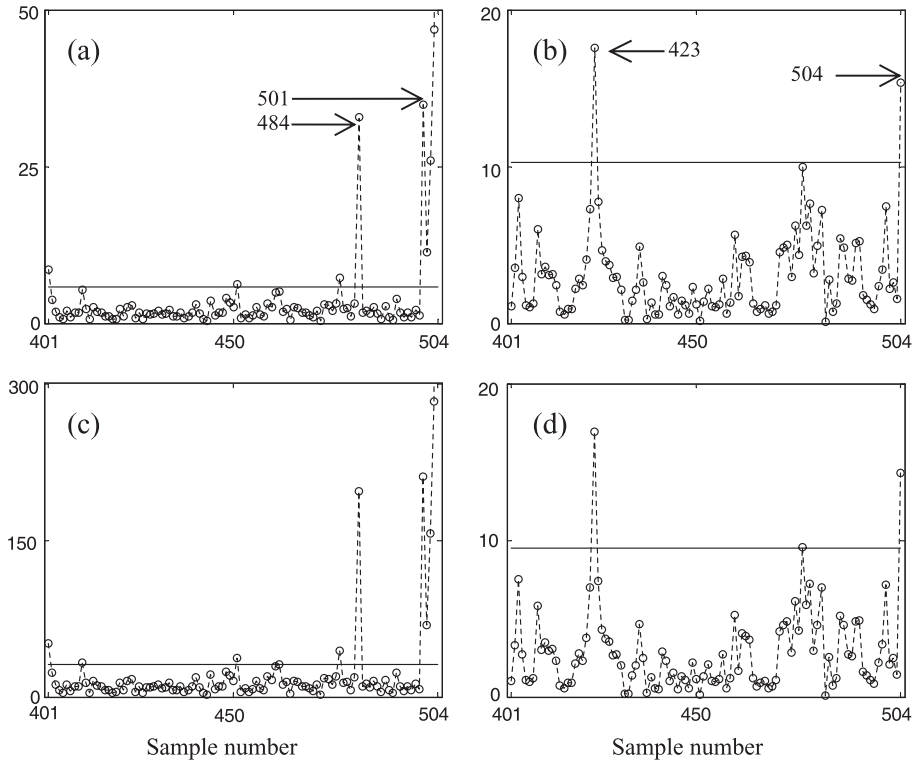


Fig. 3. PCA- vs. PPCA-based control charts with 0.05 LoS. (a) PCA in-model chart: Q chart chart. (b) PCA in-control chart: $T^2$ chart. (c) PPCA in-model chart: $\chi^2_{(20)}$ chart, $\{\|\lambda^{-0.5}\cdot\underline{e}_n\|^2;\ \chi^{-2}_{(0.95;\ 20)}\}_{n=401,\ldots,504}$. (d) PPCA in-control chart: $\chi^2_{(4)}$ chart, $\{\|\underline{z}_n\|^2;\ \chi^{-2}_{(0.95;\ 4)}\}_{n=401,\ldots,504}$.

the combined representation of Fig. 4(a) and (b), or Fig. 3(c) and (d). By the chart, we can monitor the ratio of out-of-control over out-of-model, i.e. $(\lambda\cdot P/L)\cdot\|\underline{z}\|^2\cdot\|\underline{e}\|^{-2}$. Notice that if $(\lambda\cdot 20/4)\cdot\|\underline{z}_n\|^2\cdot\|\underline{e}_n\|^{-2}<\mathcal{F}^{-1}_{(0.025;\ 4,\ 20)}$ then the $n$th sample can be considered as out-of-model $\gg$out-of-control, else if $(5\cdot\lambda)\cdot\|\underline{z}_n\|^2\cdot\|\underline{e}_n\|^{-2}>\mathcal{F}^{-1}_{(0.975;\ 4,\ 20)}$ then the sample can be regarded as out-of-model $\ll$out-of-control with $0.5\cdot\alpha$ LoS. As an example, let us investigate the 450th sample. The sample is almost out-of-model as shown in Fig. 3(c), but is completely in-control as indicated in Fig. 3(d); therefore, the ratio of the sample is nearly zero as disclosed in Fig. 4(d). Thus, by analyzing Fig. 4(d), we can notice that the 450th sample has a property that out-of-model$\gg$ out-of-control. But for the 423rd sample, it is obvious that out-of-model$\ll$ out-of-control as shown in Fig. 4(d), and the result is easily verified by checking both Fig. 3(c) and (d).

## 7. Conclusions

In this paper, we have proposed the PPCA-based process monitoring method, and compared the method with Shewhart charts and the PCA-based method through an illustrative example.

The projection model, PCA, seeks to find the least square sense optimal solution to the model; however, as we know all kinds of statistical decisions of the process condition should be made regarding to the statistics of all variables in the model. Therefore, when we apply PCA to the process monitoring, the model calibration is a totally different problem in the decision making.

The probabilistic generative model, PPCA, aims to find the likelihood sense optimal solution to the model under the assumption that all variables in the model have their specified probability densities. So, the statistical decision is the direct consequence of the model calibration itself. Therefore, when we apply the

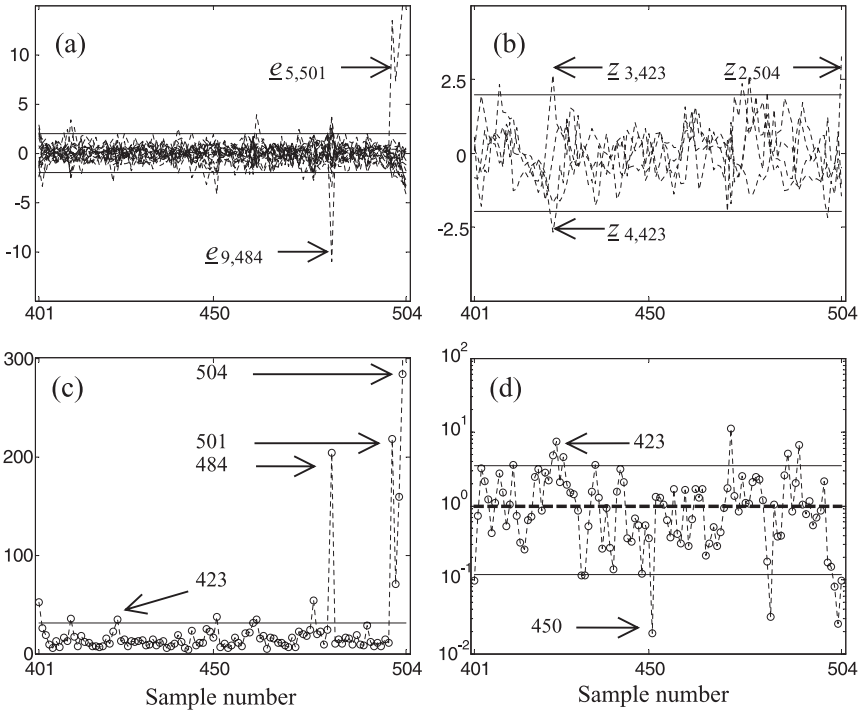


Fig. 4. Extensions of PPCA-based control charts. (a) Individual error scores chart, $\{\lambda^{-0.5}\cdot\underline{\boldsymbol{e}}_{pn};\mathcal{N}^{-1}_{s(0.025)},\ \mathcal{N}^{-1}_{s(0.975)}\}_{p\in 20,n=401,\ldots,504}$. (b) Individual PC scores chart, $\{\underline{z}_{ln};\mathcal{N}^{-1}_{s(0.025)},\ \mathcal{N}^{-1}_{s(0.975)}\}_{l\in 4,n=401,\ldots,504}$. (c) In-normal chart for measurement score, $\{\|(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}+\lambda\cdot\mathbf{I})^{-0.5}\cdot\boldsymbol{x}_n\|^2;\chi^2_{(0.95;20)}\}n$=401,...,504. (d) Out-of-control/out-of-model chart, $\{5\cdot\lambda\cdot\|\underline{\boldsymbol{z}}\|^2/\|\underline{\boldsymbol{e}}\|^2;\mathcal{F}^{-1}_{(0.025;4,20)},\mathcal{F}^{-1}_{(0.975;4,20)}\}_{n=401,\ldots,504}$.

probabilistic generative model to the process monitoring, most of the statistical control charts, e.g. univariate Shewhart chart, multivariate Shewhart chart, Q chart and $\mathcal{T}^2$ charts, are unified into the probabilistic framework. In summary,

$$\mathrm{PCA} = \{\mathrm{Euclidian},\ \mathrm{LSE},\ \mathrm{NIPALS}\}\ \text{vs.}$$

$$\mathrm{PPCA} = \{\mathrm{Mahalanobis},\ \mathrm{MLE},\ \mathrm{EM}\}.$$

Moreover, there are two potentially important aspects to the probabilistic approach, which come from the virtue of the EM algorithm: (1) handling incomplete data set, e.g. missing values at random in the set, and (2) extending the model to mixture of such models, e.g. mixture of PPCA.

## Acknowledgements

This work was supported by the Brain Korea 21 project.

## References

[1] W. Shewhart, Economic Control of Quality of Manufactured Product, Van Nostrand, Princeton, NJ, 1931.
[2] R. Sparks, Quality control with multivariate data, Australian Journal of Statistics 34 (1992) 88–95.
[3] S. Wierda, Multivariate statistical process control—recent results and directions for future research, Statistica Neerlandica 48 (1994) 147–168.
[4] H. Hotelling, Multivariate quality control, illustrated by the air testing of sample bombsights, in: C. Eisenhart, M. Hastay, W. Wallis (Eds.), Techniques of Statistical Analysis, Mc Graw, New York, 1947, pp. 111–184.
[5] I. Jolliffe, Principal component analysis, Springer Series in Statistics, Springer Verlag, New York, 1986.
[6] J. MacGregor, T. Kourti, Statistical process control of multivariate processes, Control Engineering Practice 3 (1995) 403–414.
[7] J. Jackson, G. Mudholkar, Control procedures for residuals associated with principal component analysis, Technometrics 21 (1979) 341–349.
[8] S. Roweis, EM algorithms for PCA and SPCA, Advances in Neural Information Processing System, vol. 10, MIT Press, Cambridge, MA, 1998, pp. 626–632.

[9] M. Tipping, C. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society. Series B 61 (1997) 611–622 (part 3).
[10] M. Tipping, C. Bishop, Mixtures of probabilistic principal component analyzers. Technical Report NCRG/97/003. Neural Computing Research Group, Aston University, June (1997) pp. 1–29.
[11] Z. Ghahramani, M. Jordan, Supervised learning from incomplete data via an EM approach, Advances in Neural Information Processing Systems 6 (1994) 120–127.
[12] N. Tracy, J. Young, R. Mason, Multivariate control charts for individual observations, Journal of Quality Technology 24 (1992) 88–95.
[13] T. Anderson, An Introduction to Multivariate Statistical Analysis, 2nd ed., Wiley, New York, 1984, 163 pp.
[14] K. Mardia, J. Kent, J. Bibby, Multivariate Analysis, Academic Press, London, 1989.
[15] J. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
[16] P. Miller, R. Swanson, F. Heckler, Contribution plots: the missing link in multivariate quality control, 37th Annual Fall Conference ASQC, Rochester, NY, 1993.
[17] J. Westerhuis, S. Gurden, A. Smilde, Generalized contribution plots in multivariate statistical process monitoring, Chemometrics and Intelligent Laboratory Systems 51 (2000) 95–114.
[18] A. Hyvarinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (2000) 411–430.
[19] T. Anderson, An Introduction to Multivariate Statistical Analysis, 2nd ed., Wiley, New York, 1984, 550 pp.
[20] A. Dempster, N. Laird, D. Rubin, Maximum likelihood form incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B 39 (1977) 1–38.
[21] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley, New York, 1997.
[22] S. Krantz, Jensen's Inequality, Handbook of Complex Analysis, Birkhäuser, Boston, MA, 1999, 118 pp.
[23] B. Wise, N. Gallagher, PLS_Toolbox Version 2.1, Eigenvector Research Inc., 2000, p. 42.