Calibration, Prediction and Process Monitoring Model Based on Factor Analysis for Incomplete Process Data

Dong Soon KIM[1], Chang Kyoo YOO[1], Young Il KIM[2],
Jae Hak JUNG[3] and In-Beum LEE[1]

[1]*School of Environmental Engineering/Department of Chemical Engineering, Pohang University of Science and Technology, San 31 Hyoja Dong, Pohang 790-784, Korea*

[2]*Energy Research Department, Research Institute of Industrial Science & Technology (RIST), San 32 Hyoja Dong, Pohang 790-330, Korea*

[3]*School of Chemical Engineering Technology, Yeungnam University, Daedong, Gyeongsan, 712-749, Korea*

***Keywords*:** Calibration Model, Expectation and Maximization (EM) Algorithm, Factor Analysis (FA), Incomplete Data, Missing Values, Principal Component Analysis (PCA), Process Monitoring

Unspecific missing of values in real chemical and biological industries have been found. Regardless of the incompleteness of the measured sample, a monitoring system should be designed to tackle the missing data problem and be applied to on-line systems immediately. A calibration method of a factor analysis (FA) model for incomplete data sets is proposed. And a prediction method based on the calibrated model is suggested in order to estimate missing values in incomplete calibration sets and incomplete test sets. An expectation and maximization (EM) algorithm is used to calibrate the model and expectation of conditional density is used to predict the model result. The proposed method is compared with the well-known iterative singular values decomposition (iSVD) method, i.e. a principal component analysis (PCA) based method; and a simple data set is tested as an illustrative example. The proposed method gives better estimation results for the missing values than the well-known PCA based method. There are several advantages of the proposed method over the PCA based projection methods: (1) data pretreatment is not an essential step since the FA model is scale invariant whereas the PCA model is not, (2) since the proposed method utilizes probability information of all variables directly, to apply it as a statistical process monitoring technique is preferable to others, and (3) the single model can be extended to a mixture of such models by the virtue of the EM algorithm.

Introduction

The real-world multivariate data set can have missing values at random by some reasons, e.g. troubles of sensors for industrial data, incomplete answers of respondents for statistical survey data, etc. The online measured sample can also have missing values in general since different sensing periods among sensors may exist; some measurements not yet received in the sample are regarded as missing values. Some techniques can be used to handle such types of data set or data, like a calibration technique of a multivariate model from an incomplete calibration set, and a prediction method for the missing values in every sample. Suppose that there is an incomplete calibration set. Calibrating a multivariate model from the set, should we discard a whole sample just for small portion of missing in it? Again suppose a partial sample from missing data would be received successively. When diagnosing the sample, should we wait all measurements' scores to be gathered? This paper primarily aims at answering these two questions.

A principal component analysis (PCA) based regression methods to estimate the missing values are well known (Grung and Mamme, 1998; Walczak and Massart, 2001). However, in spite of their algorithmic simplicity, they have a critical defect that the PCA model is scale variant inherently, that is, differently scaled data sets result in totally different models. It is the natural consequence of PCA which is designed to focus on variances of variables. When working with PCA, a pretreatment step on variables to have equal importance is essential.

The factor analysis (FA) model by an EM algorithm is scale invariant, where its parameters are the factor loading matrix, mean vector and noise variance matrix. The FA pursues learning the most probable model parameters that can best explain correlations between the measurements and the factors under Gaussian probability density assumption on the

Received on January 4, 2005. Correspondence concerning this article should be addressed to C. K. Yoo (E-mail address: ckyoo@postech.ac.kr or ChangKyoo.Yoo@biomath.ugent.be).

measurement space. When making the FA model, we may not worry about a scale problem of the measurements. Mathematically, PCA is a special case of FA in which set noise variances infinitesimal (Tipping and Bishop, 1997a, 1997b; Kim and Lee, 2003).

After setting several notations in Section 1, we will summarize the PCA based calibration and prediction method from incomplete data in Section 2. The FA model based probabilistic approach will be proposed in Section 3. Comparisons of the two methods will be presented in Section 4 through the case study of a simple example. Finally, the conclusions will be addressed.

1. Incomplete Data

Suppose that there is a calibration data set $\mathbf{X} = \{\boldsymbol{x}_n\}_{n\in N} \in \mathbb{R}^{P\times N}$ with $\boldsymbol{x}_n = [\boldsymbol{x}^{o}_{n}; \boldsymbol{x}^{m}_{n}] \in \mathbb{R}^{P}$. Here, subscript '$n$' represents a calibration sample index, and superscript 'o' and 'm' symbolize the observed and the missing data, respectively; the dimension of the observed vector of the n-th sample, $\dim(\boldsymbol{x}^{o}_{n})$, can have one of $\{1, 2, \cdots, \dim(\boldsymbol{x})\}$ according to the sample index n but $\dim(\boldsymbol{x}^{o}_{n}) + \dim(\boldsymbol{x}^{m}_{n}) = \dim(\boldsymbol{x})\ \forall\ n$. Notation $n \in N$ to express $n = 1, 2, \cdots, N$ just for its conciseness is used in this paper. This unusual notation should be distinguished from an 'element' symbol, e.g. $\boldsymbol{x} \in \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ or $\boldsymbol{x} \in \mathbb{R}^{P}$. In parenthesis '[·]', a semicolon is the symbol to denote the next row while comma represents the next column, e.g. $[1; 2] \in \mathbb{R}^{2\times1}$ and $[1, 2] \in \mathbb{R}^{1\times2}$. Let us denote the observed part in $\mathbf{X}$ as $\mathbf{X}^{o} = \{\boldsymbol{x}^{o}_{n}\}_{n\in N}$, and the missing part in $\mathbf{X}$ as $\mathbf{X}^{m} = \{\boldsymbol{x}^{m}_{n}\}_{n\in N}$. Hence $\mathbf{X}^{o} \cup \mathbf{X}^{m} = \mathrm{X}$ and $\mathbf{X}^{o} \cap \mathbf{X}^{m} = \{\phi\}$.

For a newly measured sample, $\boldsymbol{x}_n$, it can also have a missing value. Let us denote $\boldsymbol{x}_n = [\boldsymbol{x}^{i}_{n}; \boldsymbol{x}^{t}_{n}]$, where subscript '$n$' represents a newly measured sample index, and superscript 'i' and 't' symbolize the input and the target, respectively. Here, to distinguish $\boldsymbol{x}_n$ from $\boldsymbol{x}_n$ clearly, we gave the names 'target' and 'input' for the new sample; the target and the input in the test sample are equivalent to the missing and the observed data in the calibration sample, respectively. **Figure 1** shows an example of missing values in the multivariate data, where the hollowed squares represent the observed element and the grey squares indicate the missing element in multivariate data. For instance, in this figure, a calibration data set $\mathbf{X} = \{\boldsymbol{x}_n\}_{n\in 10} \in \mathbb{R}^{7\times10}$ in which $\boldsymbol{x}_1 = [\boldsymbol{x}^{o}_{1} \in \mathbb{R}^{6}; \boldsymbol{x}^{m}_{1} \in \mathbb{R}^{1}]$, and a test data vector $\boldsymbol{x}_n = [\boldsymbol{x}^{i}_{n} \in \mathbb{R}^{5}; \boldsymbol{x}^{t}_{n} \in \mathbb{R}^{2}]$. Then our tasks are clear: (1) to calibrate the FA model only from $\mathbf{X}^{o}$, and (2) to predict $\boldsymbol{x}^{t}_{n}$ from $\boldsymbol{x}^{i}_{n}$ using the calibrated FA model.

2. PCA from Incomplete Data

First of all, the well-known iterative PCA based regression method is addressed.

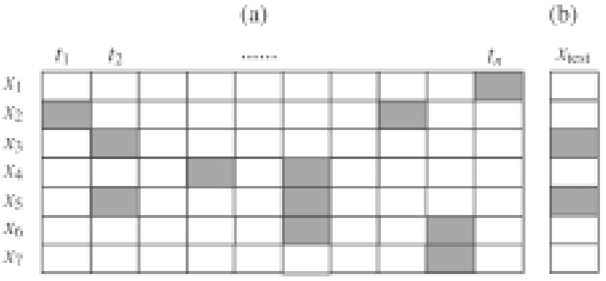

Fig. 1 Examples of missing values at random: (a) a calibration data set, and (b) a test data vector

2.1 Calibration

The PCA model of $\mathbf{X}$ reduced to L (=rank($\mathbf{Z}$) ≤ rank($\mathbf{X}$) = P) dimensional PC space is given by

$$\mathbf{X} = \mathbf{A}\cdot\mathbf{Z} + \mathbf{E} \tag{1}$$

where $\mathbf{A} = \{\boldsymbol{a}_l\}_{l\in L}$ with $\boldsymbol{a}_i^{\mathrm{T}}\cdot\boldsymbol{a}_j \in \{0, 1\}$ for $\{i \neq j, i = j\}$, $\mathbf{Z} = \{\,\underline{\underline{\boldsymbol{z}}}_l\,\}_{l\in L}$ and $\mathbf{E} = \mathbf{X} - \sum_{l\in L}\boldsymbol{a}_l\cdot\underline{\underline{\boldsymbol{z}}}_l$ have been called the loading matrix, the score matrix, and the residual matrix, respectively. All vectors used in this article are column vectors unless specified; we will use a double under bar to denote a row vector, e.g. $\underline{\underline{\boldsymbol{z}}}_l$ is the l-th row of an $L \times N$ matrix $\mathbf{Z}$ while $\boldsymbol{z}_n$ is the n-th column of $\mathbf{Z}$. If $\mathbf{X}^{m} = \{\phi\}$, i.e. there is no missing in $\mathbf{X}$, a least-square sense optimal pair of $\boldsymbol{a}_l$ and $\underline{\underline{\boldsymbol{z}}}_l$ is easily extracted by a nonlinear iterative partial least square (NIPALS) algorithm successively or a singular value decomposition (SVD) technique at once. In fact NIPALS is a method to solve the SVD problem of $\mathbf{X}$, e.g. $[\mathbf{A}, \mathbf{S}, \mathbf{V}] = \mathrm{SVDs}(\mathbf{X}, L)$ then $\mathbf{Z} = \mathbf{S}\cdot\mathbf{V}^{\mathrm{T}}$ and $\mathbf{E} = \mathbf{X} - \mathbf{A}\cdot\mathbf{S}\cdot\mathbf{V}^{\mathrm{T}}$. However, in the case of $\mathbf{X}^{m} \neq \{\phi\}$, the iterative SVD method (Walczak and Massart, 2001) can be applicable. The iSVD method enables us to estimate $\mathbf{X}^{m}$ as well as both $\mathbf{A}$ and $\mathbf{Z}$, e.g.

(1) initial guess of $\underline{\mathbf{X}} = \mathbf{X}^{o} \cup \underline{\mathbf{X}}^{m}_{\rm initial}$
(2) until $\underline{\mathbf{X}}^{m}$ converges, iterate the followings:

$$\begin{aligned}&[\mathbf{A}, \mathbf{S}, \mathbf{V}] = \mathrm{SVDs}(\underline{\mathbf{X}}, L),\\ &\mathbf{X}_{\rm PCA} = \mathbf{A}\cdot\mathbf{S}\cdot\mathbf{V}^{\mathrm{T}},\\ &\underline{\mathbf{X}}^{m} = \mathbf{X}^{m}_{\rm PCA},\ \underline{\mathbf{X}} = \mathbf{X}^{o} \cup \mathbf{X}^{m}\end{aligned} \tag{2}$$

where an 'under bar' denotes the estimate. We may set $L = P$. Then an orthogonal square matrix $\mathbf{A}$, which is to be a new set of basis vectors, transforms the original correlated variable $\boldsymbol{x}$ to a new uncorrelated variable $\boldsymbol{z}$ without any information loss, i.e. $\mathbf{E} = \{\phi\}$. This is the multiple linear regression (MLR) approach. But in general, we set $L < P$ to have robust estimate results. It is the principal component regression (PCR) approach. In essence MLR is a special case of PCR:

$$\mathbf{X}^{m}_{\rm PCR} \leftarrow \mathrm{SVDs}(\underline{\mathbf{X}}, L) \text{ and } \mathbf{X}^{m}_{\rm MLR} \leftarrow \mathrm{SVDs}(\underline{\mathbf{X}}, P) \tag{3}$$

where $L \in \{1, 2, \ldots, \le P\}$. Notice that slight modification of NIPALS algorithm can generate almost all kinds of multivariate projection models, e.g. PLS, continuum regression (CR), cyclic subspace regression (CSR), and so on (Kalivas, 1999; Geladi, 2002). The iterative algorithm can be applied to the intermediate models without serious structural change.

2.2 Prediction

To analyze a new incomplete sample, $\boldsymbol{x}_n = [\boldsymbol{x}^{\mathrm{i}}_n; \boldsymbol{x}^{\mathrm{t}}_n]$, the PCA model can be used, $\boldsymbol{x} = \mathbf{A}\cdot\boldsymbol{z} + \boldsymbol{e}$, as a combined two sub-models: an input model, $\boldsymbol{x}^{\mathrm{i}} = \mathbf{A}^{\mathrm{i}}\cdot\boldsymbol{z} + \boldsymbol{e}^{\mathrm{i}}$, and a target model, $\boldsymbol{x}^{\mathrm{t}} = \mathbf{A}^{\mathrm{t}}\cdot\boldsymbol{z} + \boldsymbol{e}^{\mathrm{t}}$. The combined model form can be expresses as follows:

$$[\boldsymbol{x}^{\mathrm{i}}; \boldsymbol{x}^{\mathrm{t}}] = [\mathbf{A}^{\mathrm{i}}; \mathbf{A}^{\mathrm{t}}]\cdot\boldsymbol{z} + [\boldsymbol{e}^{\mathrm{i}}; \boldsymbol{e}^{\mathrm{t}}] \tag{4}$$

where $\mathbf{A}^{\mathrm{i}}$ and $\mathbf{A}^{\mathrm{t}}$ are submatrices of $\mathbf{A}$ corresponding to $\dim(\boldsymbol{x}^{\mathrm{i}})$ and $\dim(\boldsymbol{x}^{\mathrm{t}})$, respectively. Suppose $\boldsymbol{x} = [x_1; x_2; x_3]$, $\boldsymbol{z} = [z_1; z_2]$, $\boldsymbol{e} = [e_1; e_2; e_3]$ and $\mathbf{A} = [a_{11}, a_{12}; a_{21}, a_{22}; a_{31}, a_{32}]$. If $\boldsymbol{x}^{\mathrm{i}} = [x_1; x_3]$, $\boldsymbol{x}^{\mathrm{t}} = [x_2]$, then $\mathbf{A}^{\mathrm{i}} = [a_{11}, a_{12}; a_{31}, a_{32}]$ and $\mathbf{A}^{\mathrm{t}} = [a_{21}, a_{22}]$; $\boldsymbol{e}^{\mathrm{i}} = [e_1; e_3]$, $\boldsymbol{e}^{\mathrm{t}} = [e_2]$. If $\dim(\boldsymbol{x}^{\mathrm{i}}_n) \neq 0$, the corresponding target vector, $\boldsymbol{x}^{\mathrm{t}}_n$, is predicted via two steps: first least-square sense optimal PC scores using the input model, then to predict the target scores using the target model:

$$\underline{\boldsymbol{z}}_n = \mathbf{A}^{\mathrm{i}\,+}_n\cdot\boldsymbol{x}^{\mathrm{i}}_n \tag{5}$$

$$\underline{\boldsymbol{x}}^{\mathrm{t}}_n = \mathbf{A}^{\mathrm{t}}_n\cdot\underline{\boldsymbol{z}}_n \tag{6}$$

where left pseudo-inverse of $\mathbf{A}^{\mathrm{i}}_n$ is defined by $\mathbf{A}^{\mathrm{i}\,+}_n \equiv (\mathbf{A}^{\mathrm{i}\,\mathrm{T}}_n\cdot\mathbf{A}^{\mathrm{i}}_n)^{-1}\cdot\mathbf{A}^{\mathrm{i}\,\mathrm{T}}_n \neq \mathbf{A}^{\mathrm{i}\,\mathrm{T}}_n$. Notice that since $\dim(\boldsymbol{x}^{\mathrm{i}}_j) \neq \dim(\boldsymbol{x}^{\mathrm{i}}_k)$ for $j \neq k$ was permitted, the number of rows of $\mathbf{A}^{\mathrm{i}}_n$ and $\mathbf{A}^{\mathrm{t}}_n$ are varied according to the sample index n; however, $\mathrm{row}(\mathbf{A}^{\mathrm{i}}_n) + \mathrm{row}(\mathbf{A}^{\mathrm{t}}_n) = \dim(\boldsymbol{x})$, and $\mathrm{column}(\mathbf{A}^{\mathrm{i}}_n) = \mathrm{column}(\mathbf{A}^{\mathrm{t}}_n) = \dim(\boldsymbol{z})$ are preserved $\forall n$.

2.3 PCA model of scale variant

The PCA model is scale variant. To ease elucidation, let us define a scaling matrix $\mathbf{B} \equiv \mathrm{diag}(\{b_p\}_{p\in P})$ that changes the scale of $\mathbf{X}$ to $\mathbf{B}\cdot\mathbf{X}$ while it does not affect the correlations among variables. However, the PCA model of $\mathbf{X}$ is totally different from that of $\mathbf{B}\cdot\mathbf{X}$. It means that $\mathbf{X} \to \mathbf{B}\cdot\mathbf{X}$ does not imply neither $\mathbf{A}\cdot\mathbf{Z} \to \mathbf{B}\cdot\mathbf{A}\cdot\mathbf{Z}$ nor $\mathbf{E} \to \mathbf{B}\cdot\mathbf{E}$. The PCA model is also shift variant. Let us introduce a mean-shift vector $\boldsymbol{m} \in \mathbb{R}^{\mathrm{P}}$ that moves $\mathbf{X}$ to $\mathbf{X}' = \{\boldsymbol{x}_n + \boldsymbol{m}\}_{n\in N}$. But the movement does not imply neither $\mathbf{A}\cdot\mathbf{Z} \to \mathbf{A}\cdot\mathbf{Z}'$ nor $\mathbf{E} \to \mathbf{E}'$, where $\mathbf{Z}' = \{\boldsymbol{z}_n + \alpha\cdot\boldsymbol{m}\}_{n\in N}$ and $\mathbf{E}' = \{\boldsymbol{e}_n + (1 - \alpha)\cdot\boldsymbol{m}\}_{n\in N}$, for any $0 \le \alpha \le 1$.

Therefore some pretreatments on a calibration set to have equal importance among variables are a prerequisite for the PCA model through mean-centering ($\boldsymbol{x} \to \boldsymbol{x}^{\mathrm{mc}}$), auto-scaling ($\boldsymbol{x} \to \boldsymbol{x}^{\mathrm{auto}}$), or whitening ($\boldsymbol{x} \to \boldsymbol{x}^{\mathrm{white}}$), where $\boldsymbol{x}^{\mathrm{mc}} = \boldsymbol{x} - \mathrm{mean}(\boldsymbol{x})$, $x_p^{\mathrm{auto}} = \mathrm{var}(x_p)^{-0.5}\cdot(x_p - \mathrm{mean}(x_p))$ $\forall\, p$, $\boldsymbol{x}^{\mathrm{white}} = \mathrm{cov}(\boldsymbol{x})^{-0.5}\cdot(\boldsymbol{x} - \mathrm{mean}(\boldsymbol{x}))$. The pretreatments are also needed to analyze the test sample since the sample should be transformed to have identical statistics of the calibration set. Without any missing in both $\boldsymbol{x}_n$ and $\mathbf{X}$, the pretreatments are simple. But if there are any missing values in $\mathbf{X}$, they are cumbersome because we should estimate the means and the (co)variances of the variables only from $\mathbf{X}^{\mathrm{o}}$.

3. Factor Analysis (FA) from Incomplete Data

The factor analysis (FA) model (Anderson, 1984) has the form of

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

where $\mathbf{A} \in \mathbb{R}^{\mathrm{P\times L}}$ denotes the loading matrix, $\boldsymbol{\mu} \in \mathbb{R}^{\mathrm{P}}$ symbolizes the mean vector of the measurement variable $\boldsymbol{x} \in \mathbb{R}^{\mathrm{P}}$; $\boldsymbol{z} \in \mathbb{R}^{\mathrm{L}}$ represents the factor variable, and $\boldsymbol{e} \in \mathbb{R}^{\mathrm{P}}$ signifies the noise variable. Learning task in the model is obtaining the likelihood-sense optimal parameters set $\Theta \equiv \{\mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ from a (in)complete calibration data set $\mathbf{X} = \mathbf{X}^{\mathrm{o}} \cup \mathbf{X}^{\mathrm{m}}$. To do this, we assume that probability density function (PDF) of the factor is $\mathcal{N}(\boldsymbol{z}\colon \boldsymbol{0}, \mathbf{I})$, where a random variable $\boldsymbol{z}$ can be completely characterized by a Gaussian parametric function with its mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. And the noise is $\mathcal{N}(\boldsymbol{e}\colon \boldsymbol{0}, \boldsymbol{\Lambda} = \mathrm{diag}(\{\lambda_p\}_{p\in P})$, where each noise element, e_p, can have a different variance, λ_p. Furthermore, factors and noises are supposed to be statistically independent of each other, i.e. $p(\boldsymbol{e} \mid \boldsymbol{z}) = p(\boldsymbol{e})$. It means that there is no information in $\boldsymbol{z}$ about $\boldsymbol{e}$.

3.1 Calibration

We consider Eq. (7) as a combined form of two sub-models: an observed model and a missing model,

$$[\boldsymbol{x}^{\mathrm{o}}; \boldsymbol{x}^{\mathrm{m}}] = [\mathbf{A}^{\mathrm{o}}; \mathbf{A}^{\mathrm{m}}]\cdot\boldsymbol{z} + [\boldsymbol{\mu}^{\mathrm{o}}; \boldsymbol{\mu}^{\mathrm{m}}] + [\boldsymbol{e}^{\mathrm{o}}; \boldsymbol{e}^{\mathrm{m}}] \tag{8}$$

It is obvious that Eq. (7) is a special case of Eq. (8) when $\boldsymbol{x}^{\mathrm{m}} = \{\phi\}$. Since the Gaussian PDF is closed to linear operations, and both $\boldsymbol{z}$ and $\boldsymbol{e}$ were assumed to Gaussians; the followings are clear: $\mathcal{N}(\boldsymbol{x}^{\mathrm{m}}\colon \boldsymbol{\mu}^{\mathrm{m}}, \boldsymbol{\Sigma}^{\mathrm{mm}})$, $\mathcal{N}(\boldsymbol{x}^{\mathrm{o}}\colon \boldsymbol{\mu}^{\mathrm{o}}, \boldsymbol{\Sigma}^{\mathrm{oo}})$, $\mathcal{N}(\boldsymbol{x}^{\mathrm{m}} \mid \boldsymbol{z}\colon \boldsymbol{\mu}^{\mathrm{m|z}}, \boldsymbol{\Sigma}^{\mathrm{m|z}})$, and $\mathcal{N}(\boldsymbol{z}|\boldsymbol{x}^{\mathrm{o}}\colon \boldsymbol{\mu}^{\mathrm{z|o}}, \boldsymbol{\Sigma}^{\mathrm{z|o}})$. Here, the mean and the covariance in each PDF are easily derived as follows:

$$\boldsymbol{\Sigma}^{\mathrm{mm}} = \mathbf{A}^{\mathrm{m}}\cdot\mathbf{A}^{\mathrm{mT}} + \boldsymbol{\Lambda}^{\mathrm{m}} \tag{9}$$

$$\boldsymbol{\Sigma}^{\mathrm{oo}} = \mathbf{A}^{\mathrm{o}}\cdot\mathbf{A}^{\mathrm{oT}} + \boldsymbol{\Lambda}^{\mathrm{o}} \tag{10}$$

$$\boldsymbol{\mu}^{\mathrm{m|z}} = \mathbf{A}^{\mathrm{m}}\cdot \mathrm{z} + \boldsymbol{\mu}^{\mathrm{m}} \tag{11}$$

$$\boldsymbol{\Sigma}^{\mathrm{m|z}} = \boldsymbol{\Lambda}^{\mathrm{m}} \tag{12}$$

$$\boldsymbol{\mu}^{\mathrm{z|o}} = \mathbf{A}^{\mathrm{oT}}\cdot(\boldsymbol{\Sigma}^{\mathrm{oo}})^{-1}\cdot(\boldsymbol{x}^{\mathrm{o}} - \boldsymbol{\mu}^{\mathrm{o}}) \tag{13}$$

$$\boldsymbol{\Sigma}^{\mathrm{z|o}} = \mathbf{I} - \mathbf{A}^{\mathrm{oT}}\cdot(\boldsymbol{\Sigma}^{\mathrm{oo}})^{-1}\cdot\mathbf{A}^{\mathrm{o}} \tag{14}$$

Notice that if $\dim(\boldsymbol{z}) < \dim(\boldsymbol{x}^{\mathrm{o}})$, say, $\mathbf{A}^{\mathrm{o}}$ is a thin matrix, inversion of $\boldsymbol{\Sigma}^{\mathrm{oo}}$ is often intractable. For Eqs.

(13) and (14), let us denote $\mathcal{N}([\boldsymbol{x}^{\mathrm{o}}; \boldsymbol{z}]: [\boldsymbol{\mu}^{\mathrm{o}}; \boldsymbol{0}], [\boldsymbol{\Sigma}^{\mathrm{oo}}, \boldsymbol{\Sigma}^{\mathrm{oz}}; \boldsymbol{\Sigma}^{\mathrm{zo}}, \boldsymbol{\Sigma}^{\mathrm{zz}}])$ then $\mathcal{N}(\mathbf{z}|\boldsymbol{x}^{\mathrm{o}}: \boldsymbol{\Sigma}^{\mathrm{zo}}\cdot\boldsymbol{\Sigma}^{\mathrm{oo}-1}\cdot(\boldsymbol{x}^{\mathrm{o}} - \boldsymbol{\mu}^{\mathrm{o}}), \boldsymbol{\Sigma}^{\mathrm{zz}} - \boldsymbol{\Sigma}^{\mathrm{zo}}\cdot\boldsymbol{\Sigma}^{\mathrm{oo}-1}\cdot\boldsymbol{\Sigma}^{\mathrm{oz}})$; where $\boldsymbol{\Sigma}^{\mathrm{oo}} = \mathbf{A}^{\mathrm{o}}\cdot\mathbf{A}^{\mathrm{oT}} + \boldsymbol{\Lambda}^{\mathrm{o}}$, $\boldsymbol{\Sigma}^{\mathrm{oz}} = \mathbf{A}^{\mathrm{o}}$, $\boldsymbol{\Sigma}^{\mathrm{zo}} = \mathbf{A}^{\mathrm{oT}}$, and $\boldsymbol{\Sigma}^{\mathrm{zz}} = \mathrm{I}$. However, there is an efficient way to the type of inversion using the matrix inversion lemma,

$$\begin{aligned}&(\mathbf{A}^{\mathrm{o}}\cdot\mathbf{A}^{\mathrm{oT}} + \boldsymbol{\Lambda}^{\mathrm{o}})^{-1}\\ &= \boldsymbol{\Lambda}^{\mathrm{o}-1} - \boldsymbol{\Lambda}^{\mathrm{o}-1}\cdot\mathbf{A}^{\mathrm{o}}\cdot(\mathbf{I} + \mathbf{A}^{\mathrm{oT}}\cdot\boldsymbol{\Lambda}^{\mathrm{o}-1}\cdot\mathbf{A}^{\mathrm{o}})^{-1}\cdot\mathbf{A}^{\mathrm{oT}}\cdot\boldsymbol{\Lambda}^{\mathrm{o}-1} \qquad (15)\end{aligned}$$

The lemma is essential to calculate both Eqs. (13) and (14), where a more detailed proof is shown in Appendix B. Without the lemma, numerical problems may be inevitable when the FA model is calibrated.

3.1.1 MLE for a complete data set The maximum likelihood estimator (MLE) is a well-known technique to calibrate the FA model. Let's define an augmented data set $\mathbf{D} \equiv \{\boldsymbol{d}_n\}_{n\in N} \in \mathbb{R}^{(\mathrm{P+L})\times \mathrm{N}}$ with $\boldsymbol{d}_n = [\boldsymbol{x}^{\mathrm{o}}_n; \boldsymbol{x}^{\mathrm{m}}_n; \boldsymbol{z}_n] \in \mathbb{R}^{(\mathrm{P+L})}$, where each data vector in $\mathbf{D}$ consists of an observed vector, a missing vector, and a factor vectors. Notice that the factor vector is regarded as a permanent missing vector with the fixed dimension L while the missing vector can have arbitrary dimension from zero to $(\dim(\boldsymbol{x}) - 1)$ according to its sample index. Given both $\boldsymbol{x}^{\mathrm{m}}$ and $\boldsymbol{z}$ in $\boldsymbol{d}$, that is, $\boldsymbol{d}$ is a complete vector, probability density of $\boldsymbol{d}$, $p(\boldsymbol{d})$, and log likelihood of $\boldsymbol{d}$, $\ell(\boldsymbol{d})$, are given by

$$\begin{aligned}p(\boldsymbol{d}: \Theta) &= p(\boldsymbol{z})\cdot p(\boldsymbol{x} \mid \boldsymbol{z}: \Theta)\\ &= \mathcal{N}(\mathbf{z}: \boldsymbol{0}, \mathbf{I})\cdot\mathcal{N}(\boldsymbol{x} \mid \mathbf{z}: \mathbf{A}\cdot\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Lambda}) \qquad (16)\end{aligned}$$

$$\begin{aligned}\ell(\Theta: \boldsymbol{d}) &\equiv \log p(\boldsymbol{d}: \Theta)\\ &= 0.5\cdot\log\det(\boldsymbol{\Lambda}^{-1})\\ &\quad - 0.5\cdot(\boldsymbol{x} - \mathbf{A}'\cdot\boldsymbol{z}')^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot(\boldsymbol{x} - \mathbf{A}'\cdot\boldsymbol{z}') + \delta \qquad (17)\end{aligned}$$

where $\mathbf{A}' \equiv [\mathbf{A}, \boldsymbol{\mu}] \in \mathbb{R}^{\mathrm{P}\times(\mathrm{L}+1)}$, $\boldsymbol{z}' \equiv [\boldsymbol{z}; 1] \in \mathbb{R}^{\mathrm{L}+1}$, and δ represents Θ independent terms in the log likelihood function. The Gaussian PDF of $\boldsymbol{x}$ is $\mathcal{N}(\boldsymbol{x}: \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-0.5\cdot\dim(\boldsymbol{x})}\cdot\det(\boldsymbol{\Sigma}^{-1})^{0.5}\cdot\exp(-0.5\cdot(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\cdot(\boldsymbol{x} - \boldsymbol{\mu}))$. Let us suppose an independent and identically distributed (*iid*) condition on $\mathbf{D}$. Mathematically, *iid* on $\mathbf{D}$ implies that $p(\boldsymbol{d}_i \mid \mathbf{D} \setminus \boldsymbol{d}_i) = p(\boldsymbol{d}_i) = p(\boldsymbol{d}_j)$, where the first equality indicates an independent condition and the second identity represents an identically distributed condition. Given both $\boldsymbol{x}^{m}_n$ and $\boldsymbol{z}_n$ $\forall$ n under *iid* on $\mathbf{D}$, the log likelihood of $\mathbf{D}$, $\mathcal{L}(\mathbf{D})$, is simply the sum of $\ell(\Theta: \boldsymbol{d}_n)$ $\forall$ n.

$$\begin{aligned}&\mathcal{L}(\Theta: \mathbf{D})\\ &\equiv \textstyle\sum_{n\in N} \ell(\Theta: \boldsymbol{x}_n)\\ &= \textstyle\sum_{n\in N}[0.5\cdot\log\det(\boldsymbol{\Lambda}^{-1})\\ &\qquad - 0.5\cdot(\boldsymbol{x}_n - \mathbf{A}'\cdot\boldsymbol{z}'_n)^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot(\boldsymbol{x}_n - \mathbf{A}'\cdot\boldsymbol{z}'_n)] + \delta\\ &\qquad (18)\end{aligned}$$

Therefore the likelihood-sense optimal Θ that maximizes $\mathcal{L}(\Theta: \mathbf{D})$ is found by

$$\begin{aligned}&(\partial/\partial\mathbf{A}')[\mathcal{L}(\Theta: \mathbf{D})] = \mathbf{0}\\ &\Rightarrow \mathbf{A}' = (\textstyle\sum_{n\in N}\boldsymbol{x}_n\cdot\boldsymbol{z}_n'^{\mathrm{T}})\cdot(\textstyle\sum_{n\in N}\boldsymbol{z}_n'\cdot\boldsymbol{z}_n'^{\mathrm{T}})^{-1} \qquad (19)\end{aligned}$$

$$\begin{aligned}&(\partial/\partial\boldsymbol{\Lambda}^{-1})[\mathcal{L}(\Theta: \mathbf{D})] = \mathbf{0}\\ &\Rightarrow \boldsymbol{\Lambda} = N^{-1}\cdot\mathrm{diag}\textstyle\sum_{n\in N}(\boldsymbol{x}_n\cdot\boldsymbol{x}_n^{\mathrm{T}} - \mathbf{A}'\cdot\mathbf{z}_n'\cdot\boldsymbol{x}_n^{\mathrm{T}}) \qquad (20)\end{aligned}$$

where $(\partial/\partial\mathbf{A})\cdot(\boldsymbol{x} - \mathbf{A}\cdot\mathbf{z})^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot(\boldsymbol{x} - \mathbf{A}\cdot\mathbf{z}) = -2\cdot\boldsymbol{\Lambda}^{-1}\cdot(\boldsymbol{x} - \mathbf{A}\cdot\mathbf{z})\cdot\mathbf{z}^{\mathrm{T}}$ for symmetric $\boldsymbol{\Lambda}$, $(\partial/\partial\boldsymbol{\Lambda}^{-1})\cdot[\log\det(\boldsymbol{\Lambda}^{-1})] = \boldsymbol{\Lambda}^{\mathrm{T}}$, and $(\partial/\partial\boldsymbol{\Lambda}^{-1})\cdot[(\boldsymbol{x} - \mathbf{A}\cdot\mathbf{z})^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot(\boldsymbol{x} - \mathbf{A}\cdot\mathbf{z})] = (\boldsymbol{x} - \mathbf{A}\cdot\mathbf{z})\cdot(\boldsymbol{x} - \mathbf{A}\cdot\mathbf{z})^{\mathrm{T}}$ and $\mathrm{diag}(\cdot)$ denotes set off-diagonal elements in the parenthesis to zeros. Note that $\mathbf{A}'$ in Eq. (20) should be the resultant of Eq. (19). If a mean-centered $\mathbf{X}$ is used to these equations, Eq. (19) indicates the least-square solution, $\mathbf{A} = \mathbf{X}\cdot\mathbf{Z}^{+}$, and Eq. (20) implies slightly biased, $N\cdot(N - 1)^{-1}$, the sample error variance matrix whose diagonals are the variances of the estimation errors, i.e. $\boldsymbol{\Lambda} = N^{-1}\cdot\mathrm{diag}(\mathbf{X}\cdot(\mathbf{I} - \mathbf{Z}^{+}\cdot\mathbf{Z})\cdot\mathbf{X}^{\mathrm{T}})$.

3.1.2 EM for incomplete data Only given $\mathbf{X}^{\mathrm{o}}$ in $\mathbf{D} = \mathbf{X}^{\mathrm{o}} \cup \mathbf{X}^{\mathrm{m}} \cup \mathbf{Z}$, direct maximization of $\mathcal{L}(\Theta: \mathbf{D})$ is impossible through analytic ways. An expected complete data log likelihood function can be assumed, as it has been called an energy function, $q(\cdot)$ for $\ell(\cdot)$ first.

$$\begin{aligned}&q(\Theta|\Theta_t)\\ &\equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}:\Theta t}[\ell(\boldsymbol{d}: \Theta)]\\ &= \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}:\Theta t}[0.5\cdot\log\det(\boldsymbol{\Lambda}^{-1})\\ &\qquad - 0.5\cdot(\boldsymbol{x} - \mathbf{A}'\cdot\boldsymbol{z}')^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot(\boldsymbol{x} - \mathbf{A}'\cdot\boldsymbol{z}')] + \delta\\ &\qquad (21)\end{aligned}$$

where $\mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}:\Theta t}[\ell(\boldsymbol{d}: \Theta)] \equiv \iint \ell(\boldsymbol{x}^{\mathrm{o}}, \boldsymbol{x}^{\mathrm{m}}, \boldsymbol{z}: \Theta)\cdot p(\boldsymbol{z}, \boldsymbol{x}^{\mathrm{m}}|\boldsymbol{x}^{\mathrm{o}}: \Theta_t)\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x}^{\mathrm{m}}$, and Θ_t denotes an estimate of Θ at iteration time t. If $\boldsymbol{x}^{\mathrm{m}}$ were given, then $q(\Theta|\Theta_t) = \mathcal{E}_{\mathbf{z}|\mathbf{x}:\Theta t}[\ell(\boldsymbol{d}: \Theta)]$ which results in the standard FA. Since *iid* on $\mathbf{D}$ was assumed, an energy function $\Theta(\cdot)$ for $\mathcal{L}(\cdot)$ is just the summation of $q_n(\Theta|\Theta_t)$ $\forall$ n.

$$\begin{aligned}&\mathcal{Q}(\Theta|\Theta_t)\\ &\equiv \textstyle\sum_{n\in N} q_n(\Theta|\Theta_t)\\ &= \textstyle\sum_{n\in N}\mathcal{E}_{\boldsymbol{z}n,\boldsymbol{x}\mathrm{m}n|\boldsymbol{x}\mathrm{o}n:\Theta t}[0.5\cdot\log\det(\boldsymbol{\Lambda}^{-1})\\ &\qquad - 0.5\cdot(\boldsymbol{x}_n - \mathbf{A}'\cdot\mathbf{z}_n')^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{-1}\cdot(\boldsymbol{x}_n - \mathbf{A}'\cdot\mathbf{z}_n')]\\ &\quad + \delta \qquad (22)\end{aligned}$$

Instead of intractable $\mathcal{L}(\cdot)$, maximize Eq. (22) via two steps: an expectation step (E-step) and a maximization step (M-step). The E-step evaluates $\mathcal{Q}$ given from previous M-step resultant Θ, and the maximization step (M-step) finds maximizing Θ of the E-step resultant $\mathcal{Q}$. The iterative algorithm altering the E-step and the M-step has been named an EM algorithm, and the algorithm can never decrease $\mathcal{L}(\cdot)$ as iteration proceeds (Dempster *et al.*, 1977), in other words, Θ_t converges to likelihood-sense (local) optimal Θ as t increases. There might be local optimal since there is no way to find a global optimum solution for a general nonlinear functions. The local region of the solution candidates of the function can be extended via a genetic algorithm, a Boltzmann machine, and so on.

E-Step: To evaluate $\mathcal{Q}(\Theta|\Theta_t)$ let us approximate $p(\boldsymbol{x}^{\mathrm{m}}|\mathbf{z}, \boldsymbol{x}^{\mathrm{o}}) \approx p(\boldsymbol{x}^{\mathrm{m}}|\underline{\mathbf{z}})$ where $\underline{\mathbf{z}} \equiv \mathcal{E}[\mathbf{z}|\boldsymbol{x}^{\mathrm{o}}]$. It implies that the observed vector is only used to infer the factor scores which are responsible for all generated signals from the concerned process. In generative latent variable models, e.g. FA, probabilistic PCA, hidden markov model, mixture modeling, measurement scores, no matter how many they are, are treated just as shadows of latent variables' combinations. With the approximation, the conditional joint density over two unknowns, given a known is factorized to

$$p(\boldsymbol{z}, \boldsymbol{x}^{\mathrm{m}}|\boldsymbol{x}^{\mathrm{o}}) \approx p(\boldsymbol{z}|\boldsymbol{x}^{\mathrm{o}})\cdot p(\boldsymbol{x}^{\mathrm{m}}|\underline{\boldsymbol{z}}) \tag{23}$$

Since we knows $p(\boldsymbol{z}|\boldsymbol{x}^{\mathrm{o}}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{x}^{\mathrm{o}}\!: \boldsymbol{\mu}^{\mathrm{z|o}}, \boldsymbol{\Sigma}^{\mathrm{z|o}})$ and $p(\boldsymbol{x}^{\mathrm{m}}|\underline{\boldsymbol{z}}) = \mathcal{N}(\boldsymbol{x}^{\mathrm{m}}|\underline{\boldsymbol{z}}\!: \boldsymbol{\mu}^{\mathrm{m|\underline{z}}}, \boldsymbol{\Sigma}^{\mathrm{m|\underline{z}}})$, Eq. (23) is evaluated using Eqs. (11)–(14). Furthermore, using this factorization, the expected elements which implicitly existed in the $\mathcal{Q}$ function are easily derived as follows:

$$\underline{\boldsymbol{z}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{z}] \approx \boldsymbol{\mu}^{\mathrm{z|o}} \tag{24}$$

$$\underline{\boldsymbol{zz}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{z}\cdot\boldsymbol{z}^{\mathrm{T}}] \approx \boldsymbol{\Sigma}^{\mathrm{z|o}} + \boldsymbol{\mu}^{\mathrm{z|o}}\cdot\boldsymbol{\mu}^{\mathrm{z|oT}} \tag{25}$$

$$\underline{\boldsymbol{x}}^{\mathrm{m}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{x}^{\mathrm{m}}] \approx \boldsymbol{\mu}^{\mathrm{m|\underline{z}}} \tag{26}$$

$$\underline{\boldsymbol{xx}}^{\mathrm{m}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{x}^{\mathrm{m}}\cdot\boldsymbol{x}^{\mathrm{mT}}] \approx \boldsymbol{\Sigma}^{\mathrm{m|\underline{z}}} + \boldsymbol{\mu}^{\mathrm{m|\underline{z}}}\cdot\boldsymbol{\mu}^{\mathrm{m|\underline{z}T}} \tag{27}$$

$$\underline{\boldsymbol{x}}^{\mathrm{m}}\underline{\boldsymbol{z}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{x}^{\mathrm{m}}\cdot\boldsymbol{z}^{\mathrm{T}}] \approx \boldsymbol{\mu}^{\mathrm{m|\underline{z}}}\cdot\boldsymbol{\mu}^{\mathrm{z|oT}} \tag{28}$$

$$\underline{\boldsymbol{zx}}^{\mathrm{m}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{z}\cdot\boldsymbol{x}^{\mathrm{mT}}] \approx \boldsymbol{\mu}^{\mathrm{z|o}}\cdot\boldsymbol{\mu}^{\mathrm{m|\underline{z}T}} \tag{29}$$

They are also calculated by Eqs. (11)–(14). Augmented forms of the expectation elements, which explicitly exist in the $\mathcal{Q}$ function are evaluated as follows using Eqs. (24)–(29):

$$\underline{\boldsymbol{z}}' \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{z}'] = [\underline{\boldsymbol{z}};\ 1] \tag{30}$$

$$\underline{\boldsymbol{zz}}' \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{z}'\cdot\boldsymbol{z}'^{\mathrm{T}}] = [\underline{\boldsymbol{zz}},\ \underline{\boldsymbol{z}};\ \underline{\boldsymbol{z}}^{\mathrm{T}},\ 1] \tag{31}$$

$$\underline{\boldsymbol{x}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{x}] = [\boldsymbol{x}^{\mathrm{o}};\ \underline{\boldsymbol{x}}^{\mathrm{m}}] \tag{32}$$

$$\underline{\boldsymbol{xx}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{x}\cdot\boldsymbol{x}^{\mathrm{T}}] = [\boldsymbol{x}^{\mathrm{o}}\cdot\boldsymbol{x}^{\mathrm{oT}},\ \boldsymbol{x}^{\mathrm{o}}\cdot\underline{\boldsymbol{x}}^{\mathrm{mT}};\ \underline{\boldsymbol{x}}^{\mathrm{m}}\cdot\boldsymbol{x}^{\mathrm{oT}},\ \underline{\boldsymbol{xx}}^{\mathrm{m}}] \tag{33}$$

$$\underline{\boldsymbol{xz}}' \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{x}\cdot\boldsymbol{z}'^{\mathrm{T}}] = \underline{\boldsymbol{x}}\cdot\underline{\boldsymbol{z}}'^{\mathrm{T}} \tag{34}$$

$$\underline{\boldsymbol{z}'\mathrm{x}} \equiv \mathcal{E}_{\boldsymbol{z},\boldsymbol{x}\mathrm{m}|\boldsymbol{x}\mathrm{o}}[\boldsymbol{z}'\cdot\boldsymbol{x}^{\mathrm{T}}] = \underline{\boldsymbol{z}}'\cdot\underline{\boldsymbol{x}}^{\mathrm{T}} \tag{35}$$

Notice that the most probable inferences on all unobserved states, i.e. factors and missings, and their correlations are evaluated and used in the EM algorithm of the FA model calibration while the least square of the PCA model simply adapts the orthogonal projection of $\boldsymbol{x}$ to the latent space as the best inference of the states.

M-step: Similar to Eqs. (19) and (20), likelihood-sense optimal Θ that maximizes $\mathcal{Q}(\Theta|\Theta_t)$ is obtained by

$$(\partial/\partial\mathbf{A}')[\mathcal{Q}(\Theta|\Theta_t)] = \mathbf{0}$$
$$\Rightarrow \mathbf{A}' = (\textstyle\sum_{n\in N}\underline{\boldsymbol{xz}}'_n{}^{\mathrm{T}})\cdot(\textstyle\sum_{n\in N}\underline{\boldsymbol{zz}}'_n)^{-1} \tag{36}$$

$$(\partial/\partial\boldsymbol{\Lambda}^{-1})[\mathcal{Q}(\Theta|\Theta_t)] = \mathbf{0}$$
$$\Rightarrow \boldsymbol{\Lambda} = N^{-1}\cdot\mathrm{diag}\textstyle\sum_{n\in N}(\underline{\boldsymbol{xx}}_n - \mathbf{A}'\cdot\underline{\boldsymbol{z}'\boldsymbol{x}}_n) \tag{37}$$

As before, $\mathbf{A}'$ in Eq. (37) should be the resultant of Eq. (36). Notice that $\underline{\boldsymbol{xx}}_n \neq \underline{\boldsymbol{x}}_n\cdot(\underline{\boldsymbol{x}}_n)^{\mathrm{T}}$ and $\underline{\boldsymbol{zz}}'_n \neq \underline{\boldsymbol{z}}'_n\cdot(\underline{\boldsymbol{z}}'_n)^{\mathrm{T}}$ but $\underline{\boldsymbol{xz}}'_n = \underline{\boldsymbol{x}}_n\cdot(\underline{\boldsymbol{z}}'_n)^{\mathrm{T}}$ and $\underline{\boldsymbol{z}}'\underline{\boldsymbol{x}}_n = \underline{\boldsymbol{z}}'_n\cdot(\underline{\boldsymbol{x}}_n)^{\mathrm{T}}$ as we expressed in Eqs. (30)–(35).

EM algorithm to calibrate the FA model from incomplete data set is summarized as follows: unobserved states and their correlations are the likelihood-sense optimal inferred by E-step using Eqs. (24)–(29) given Θ. Then we can update Θ by the M-step using Eqs. (30) and (31), given the inference results which have the forms represented in Eqs. (24)–(26). A lot of iterations altering the E-step and the M-step converge to a likelihood-sense (local) optimal Θ definitely. Appendix A is the summary of these procedures; it may be helpful to realize the method in computer programs, e.g. Matlab.

3.1.3 Variance explanation ratio A covariance matrix of $\boldsymbol{x}$ has been estimated by the sum of its systematic part $\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}}$, and noise part $\boldsymbol{\Lambda}$, i.e. $\underline{\boldsymbol{\Sigma}}^{\mathrm{x}} = \mathbf{A}\cdot\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Lambda}$, in the model. Denoting the systematic part of the p-th variable in $\boldsymbol{x}$ as x^{s}_p, then the variance explanation ratio of the p-th variable in $\boldsymbol{x}$, x_p, is given by $r_p = \mathrm{var}(x^{\mathrm{s}}_p)\cdot\mathrm{var}(x_p)^{-1}$. Hence, the average variance explanation ratio, r_{avg}, can be defined by the mean of all the ratios.

$$r_p = \mathrm{diag}(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})_p\cdot[\mathrm{diag}(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})_p + \mathrm{diag}(\boldsymbol{\Lambda})_p]^{-1} \tag{38}$$

$$r_{\mathrm{avg}} = P^{-1}\cdot\textstyle\sum_{p\in P} r_p \tag{39}$$

where $\mathrm{diag}(\cdot)_p \in \mathbb{R}^1$ and $\mathrm{diag}(\cdot)_p^{-1} \in \mathbb{R}^1$ denote the p-th diagonal element in the parenthesis and its inverse, respectively. Like the score plot of the PCA model, $\dim(\boldsymbol{z})$ of the FA model can be decided by plotting $r_{\mathrm{avg}}(l)$ for $l = 1, 2, \cdots, \dim(\boldsymbol{x})$. Notice that total variance explanation ratio in the PCA model is obtained from

$$\begin{aligned} r^{\mathrm{PCA}} &= [\textstyle\sum_{p\in P}\mathrm{var}(x^{\mathrm{s}}_p)]\cdot[\textstyle\sum_{p\in P}\mathrm{var}(x_p)]^{-1} \\ &= [\textstyle\sum_{p\in P}\mathrm{diag}(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})_p]\cdot[\textstyle\sum_{p\in P}\mathrm{diag}(\mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})_p \\ &\qquad + \mathrm{diag}(\boldsymbol{\Lambda})_p]^{-1} \\ &= \|\mathbf{A}\|_{\mathrm{F}}^2\cdot\mathrm{tr}(\underline{\boldsymbol{\Sigma}}^{\mathrm{x}})^{-1} \end{aligned} \tag{40}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes a Frobenius norm. Thus an important difference to find an appropriate number of latent variables between FA and PCA is that FA decides the number by the summation of all the elements' ratios while PCA determine it by the ratio of the two summations:

$$r^{\mathrm{FA}} = P^{-1}\cdot\sum_{p\in P}[\mathrm{var}(x^{\mathrm{s}}{}_p)\cdot\mathrm{var}(x_p)^{-1}] \text{ vs.}$$
$$r^{\mathrm{PCA}} = [\sum_{p\in P}\mathrm{var}(x^{\mathrm{s}}{}_p)]\cdot[\sum_{p\in P}\mathrm{var}(x_p)]^{-1} \tag{41}$$

They also indicate that PCA depends on variances of variables but FA does not.

3.2 Prediction

For a newly measured sample $\boldsymbol{x}_n = [\boldsymbol{x}^{\mathrm{i}}{}_n; \boldsymbol{x}^{\mathrm{t}}{}_n]$, similar to PCA, a target vector is predicted via two steps: infer factor scores first, and then predict target score using the inference result.

$$\underline{\mathbf{z}}_n \equiv \mathcal{E}[\mathbf{z}|\boldsymbol{x}^{\mathrm{i}}{}_n] = \beta_n\cdot(\boldsymbol{x}^{\mathrm{i}}{}_n - \boldsymbol{\mu}^{\mathrm{i}}{}_n) \tag{42}$$

$$\underline{\boldsymbol{x}}^{\mathrm{t}}{}_n \equiv \mathcal{E}[\boldsymbol{x}^{\mathrm{t}}{}_n|\underline{\mathbf{z}}_n] = \mathbf{A}^{\mathrm{t}}{}_n\cdot\underline{\mathbf{z}}_n + \boldsymbol{\mu}^{\mathrm{t}}{}_n \tag{43}$$

where $\beta_n \equiv \mathbf{A}^{\mathrm{i}}{}_n{}^{\mathrm{T}}\cdot\boldsymbol{\Sigma}^{\mathrm{ii}}{}_n{}^{-1}$; $\boldsymbol{\Sigma}^{\mathrm{ii}}{}_n{}^{-1} = (\mathbf{A}^{\mathrm{i}}{}_n\cdot\mathbf{A}^{\mathrm{i}}{}_n{}^{\mathrm{T}} + \boldsymbol{\Lambda}^{\mathrm{i}}{}_n)^{-1}$ should be calculated regarding to $\dim(\boldsymbol{x}^{\mathrm{i}}{}_n)$ using Eq. (15). Notice that PCA simply used orthogonal projection to infer the latent score and predict the target score as expressed in Eqs. (5) and (6); but FA utilizes conditional densities for these tasks. In fact Eq. (5) is the special case of Eq. (42) when using a mean centered calibration set, $\boldsymbol{\mu}^{\mathrm{i}}{}_n = 0$, and restrict the variances of the noise variables to have infinitesimal, i.e. $\lim_{\lambda p\to 0} \mathbf{A}^{\mathrm{i}}{}_n{}^{+} = \beta_n \;\forall\; p$. And Eq. (6) is identical to Eq. (43) for the mean centered data.

3.3 FA model of scale invariant

Contrast to the PCA model, the FA model is invariant to both scale-change and mean-shift. For rescaled x by scaling matrix $\mathbf{B} \equiv \mathrm{diag}(\{b_p\}_{p\in P})$, i.e. $\boldsymbol{x} \to \mathbf{B}\cdot\boldsymbol{x}$, the rescaling does not affect the FA model structure, it only influences model parameters: $\mathbf{A} \to \mathbf{B}\cdot\mathbf{A}$, $\boldsymbol{\mu} \to \mathbf{B}\cdot\boldsymbol{\mu}$, and $\boldsymbol{\Lambda} \to \mathbf{B}\cdot\boldsymbol{\Lambda}\cdot\mathbf{B}$. For mean shifted $\boldsymbol{x}$ by $\boldsymbol{m}$, i.e. $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{m}$, it indicates simply $\boldsymbol{\mu} \to \boldsymbol{m}\cdot\boldsymbol{\mu}$. When developing the FA model, we are free from the pretreatments both on the calibration set and the tested sample, which are essential in projection models.

3.4 Statistical tests

An estimation error of the input part in a new sample is given by $\underline{\boldsymbol{e}}^{\mathrm{i}}{}_n = \boldsymbol{x}^{\mathrm{i}}{}_n - \underline{\boldsymbol{x}}^{\mathrm{i}}{}_n$ where $\underline{\boldsymbol{x}}^{\mathrm{i}}{}_n = \mathbf{A}^{\mathrm{i}}\cdot\underline{\mathbf{z}}_n + \boldsymbol{\mu}^{\mathrm{i}}$. If the developed FA model is still valid to the sample, Mahalanobis squared norms of input error (MSNE), $\underline{\boldsymbol{e}}^{\mathrm{i}}{}_n{}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{\mathrm{i}}{}_n{}^{-1}\cdot\underline{\boldsymbol{e}}^{\mathrm{i}}{}_n$, should follow the chi-square distribution with $\dim(\boldsymbol{x}^{\mathrm{i}}{}_n)$ degrees of freedom since the FA model assumed $\mathcal{N}(\boldsymbol{e}^{\mathrm{i}}: \boldsymbol{0}, \boldsymbol{\Lambda}^{\mathrm{i}})$. Thus if Eq. (38) is satisfied for the sample, the developed FA model would be adequate to the sample with the α level of significance (LoS). Under the condition, both types of the variance explanation ratios, elements based ratios in Eq. (38) and the average ratio in Eq. (39), are also expected to the sample. Furthermore, the process which generates $\boldsymbol{x}^{\mathrm{i}}{}_n$ is considered as in-control with α LoS if Eq. (45) is held since $\mathcal{N}(\boldsymbol{e}^{\mathrm{i}}: \boldsymbol{0}, \mathbf{I})$.

$$\underline{\boldsymbol{e}}^{\mathrm{i}}{}_n{}^{\mathrm{T}}\cdot\boldsymbol{\Lambda}^{\mathrm{i}-1}\cdot\underline{\boldsymbol{e}}^{\mathrm{i}}{}_n \le \chi^{-2}{}_{(1-\alpha;\,\dim(\boldsymbol{x}n\mathrm{i}))} \tag{44}$$

$$\underline{\mathbf{z}}_n{}^{\mathrm{T}}\cdot\underline{\mathbf{z}}_n \le \chi^{-2}{}_{(1-\alpha,\,L)} \tag{45}$$

where χ^{-2} denotes the inverse of cumulative chi-square PDF. Equations (44) and (45) are closely related to the Q test (Jackson and Mudholkar, 1979), and Hotelling's $\mathcal{T}^2$ test (Hotelling, 1947), respectively; they are popular tests in statistical process control (SPC) and fault detection (MacGregor and Kourti, 1995; Wise and Gallagher, 1996). Note here that Q and $\mathcal{T}^2$ tests are given by

$$\begin{aligned}\mathrm{Q}: \boldsymbol{x}^{\mathrm{T}}\cdot(\mathbf{I} - \mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})\cdot\boldsymbol{x} &\\ \le \theta_1\cdot[(\mathcal{N}_{\mathrm{s}}{}^{-1}{}_{(1-\alpha)}\cdot(2\cdot\theta_2\cdot h_0{}^2)^{0.5}\cdot\theta_1{}^{-1}) &\\ + (\theta_2\cdot h_0\cdot(h_0 - 1)\cdot\theta_1{}^{-2}) + 1]^{1/h0}&\end{aligned} \tag{46}$$

$$\mathcal{T}^2: \|\mathbf{S}_{\mathrm{S}}{}^{-0.5}\cdot\boldsymbol{x}\|^2 \le L\cdot(N^2 - 1)\cdot(N^2 - N\cdot L)^{-1}\cdot\mathcal{F}^{-1}{}_{(1-\alpha;\,L,N-L)} \tag{47}$$

where $\mathcal{N}_{\mathrm{s}}{}^{-1}$ denotes the inverse of cumulative $\mathcal{N}(0, 1)$; using the following SVD results of $\mathbf{X}$: $[\mathbf{A} = \{\boldsymbol{a}_p\}_{p\in P}$, $\mathbf{S} = \mathrm{diag}(\{s_p\}_{p\in P})$, $\mathbf{V}] = \mathrm{SVDs}(\mathbf{X}, P)$, $\theta_j = \sum_{i=L+1,\ldots,P} s_i{}^{2\cdot j}$ for $j \in \{1, 2, 3\}$, $h_0 = 1 - (2\cdot\theta_1\cdot\theta_3)\cdot(3\cdot\theta_2{}^2)^{-1}$, $(\mathbf{S}_{\mathrm{S}}{}^{-0.5})^{\mathrm{T}} = \{s_p{}^{-1}\cdot\boldsymbol{a}_p\}_{p\in L}$.

There are several advantages on the proposed tests over the conventional tests:

(1) They provide simpler forms than the existing tests.

(2) They work with the same measuring unit, i.e. squared Mahalanobis norm. All statistical inference should be based on the Mahalanobis measuring unit; however, the left-hand term in the Q test indicates the squared Euclidian norm of the estimation errors, i.e. $\|\underline{\boldsymbol{e}}\|^2$, since $\underline{\boldsymbol{e}} = \boldsymbol{x} - \mathbf{A}\cdot\underline{\mathbf{z}} = (\mathbf{I} - \mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})\cdot\boldsymbol{x}$, and $(\mathbf{I} - \mathbf{A}\cdot\mathbf{A}^{\mathrm{T}})$ is an idempotent matrix. It is the reason why the right-hand term in the test is so complicated. But the left-hand term in the $\mathcal{T}^2$ test represents the squared Mahalanobis norm of principal components' scores with the normalizing constants $\{s_p{}^{-1}\}_{p\in L}$. Note here that the limiting condition of $L\cdot(N^2 - 1)\cdot(N^2 - N\cdot L)^{-1}\cdot\mathcal{F}^{-1}{}_{(1-\alpha;\,L,\,N-L)}$ when $N \approx \infty$ is $\chi^{-2}{}_{(1-\alpha;\,L)}$; however, we do not set the condition for the upper control limit of Q.

(3) They can work on partial available elements among all the measurements; while Q and $\mathcal{T}^2$ tests should wait for all measurements to be gathered. This point may be the decisive merit of the tests when applying the tests to an on-line monitored process with some malfunctioning sensors or missing values.

(4) They provide statistical normality tests for all individuals both in $\underline{\mathbf{z}}_n = \{\underline{z}_l\}_{l\in L}$ and $\underline{\boldsymbol{e}}^{\mathrm{i}}{}_n = \{\underline{e}_p\}_{p\in\dim(\boldsymbol{x}n\mathrm{i})}$. Notice that $\mathcal{N}(\underline{\mathbf{z}}_n: \boldsymbol{0}, \mathbf{I})$ and $\mathcal{N}(\underline{\boldsymbol{e}}^{\mathrm{i}}{}_n: \boldsymbol{0}, \mathrm{diag}\{\lambda_p\}_{p\in\dim(\boldsymbol{x}n\mathrm{i})})$ indicate $\{\mathcal{N}(\underline{z}_l: 0, 1)\}_{l\in L}$ and $\{\mathcal{N}(\lambda_p{}^{-0.5}\cdot\underline{e}_p: 0, 1)\}_{p\in P}$, respectively, where $p(\underline{z}_i|\underline{z}_j) = p(\underline{z}_i)$ and $p(\underline{e}_i|\underline{e}_j) = p(\underline{e}_i)$ for $i \ne j$. Therefore if the p-th sensor in $\boldsymbol{x}^{\mathrm{i}}{}_n$ is fault then Eq. (48) will not be satisfied with α LoS. And if Eq. (49) is not hold for the extracted l-th factor score in $\underline{\mathbf{z}}_n$, $\underline{z}_l$, then the factor is considered responsible for the process' out-of-control condition with α LoS.

$$\underline{e}_p \in [\lambda_p{}^{0.5}\cdot\mathcal{N}_{\mathrm{s}}{}^{-1}{}_{(0.5\cdot\alpha)}, \lambda_p{}^{0.5}\cdot\mathcal{N}_{\mathrm{s}}{}^{-1}{}_{(1-0.5\cdot\alpha)}] \tag{48}$$

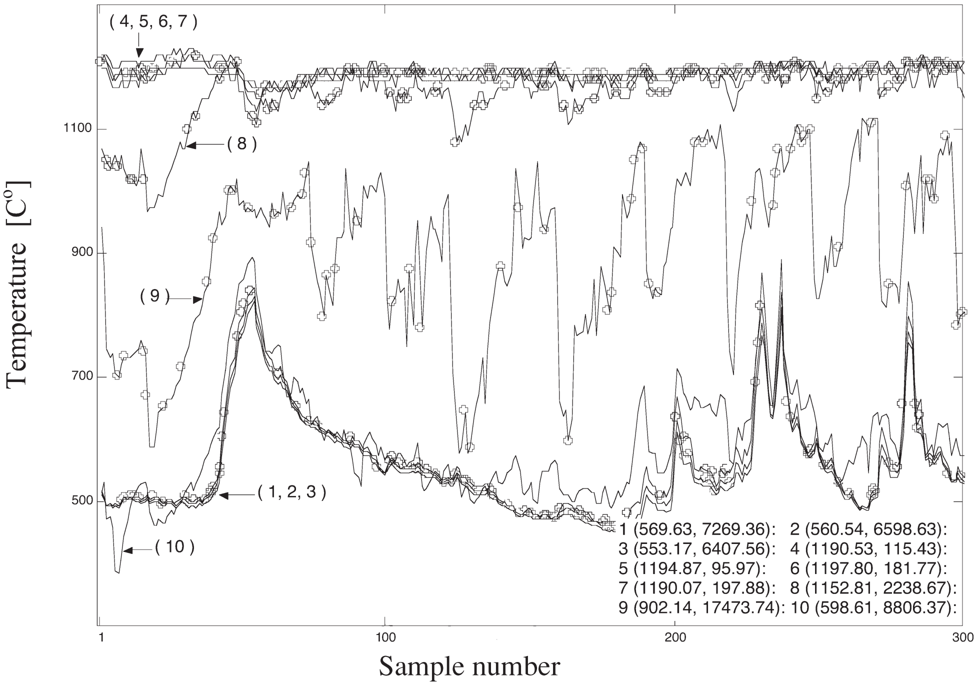

Fig. 2 Liquid-fed ceramic melter (LFCM) data set made to have artificial 20% missing values at random

$$\underline{z}_l \in [\mathcal{N}_{\text{s}}^{-1}{}_{(0.5\cdot\alpha)}, \mathcal{N}_{\text{s}}^{-1}{}_{(1-0.5\cdot\alpha)}] \tag{49}$$

These tests are comparable to the contribution plot used in conventional SPC.

4. Simulation and Results

Multivariate process data from liquid-fed ceramic melter (LFCM) data were employed to test the performance of the proposed method, where the data set are taken from the PLS_toolbox ver. 3.1 (Eigenvetor Research, Inc.). A slurry of nuclear waste and glass-forming chemicals were fed into the LFCM, and glass was periodically poured from the melter, resulting in time-dependent variations in the glass level. The LFCM was monitored with 10 thermocouples located in two thermo-wells within the glass pool (Stork *et al.*, 1997). Raw data set was treated to have 20% missing at random. The data pattern and missing values are shown in **Figure 2**, where the missing values are symbolized by '○'. In the figure, $p(m, v)$ indicates that the p-th variable in the raw data set, without any missing value, has mean m and variance v. As shown in the figure, every 10 variables show the strong correlation among the measured variables and also have remarkably different statistics of means and variances. If there were negligible missing data, the pretreatment step for PCA would be a simple task. However, 20% missing may not be a negligible quantity at all.

To develop a multivariate model from incomplete data set and predict the missing values in the set, we partitioned the overall incomplete set $\mathbf{X} = \{\boldsymbol{x}_j\}_{j\in 300}$ into

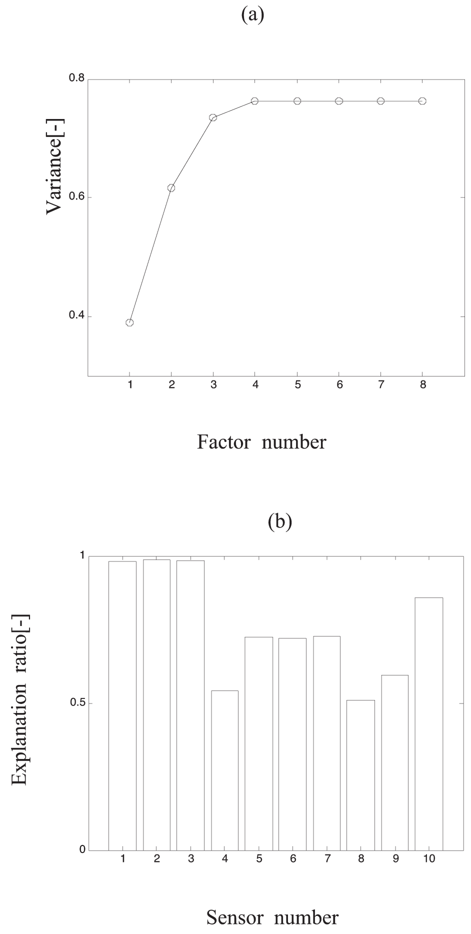

Fig. 3 Plots of variance explanation ratios: (a) averaged variance explanation ratio, and (b) individual elements' variance explanation ratios

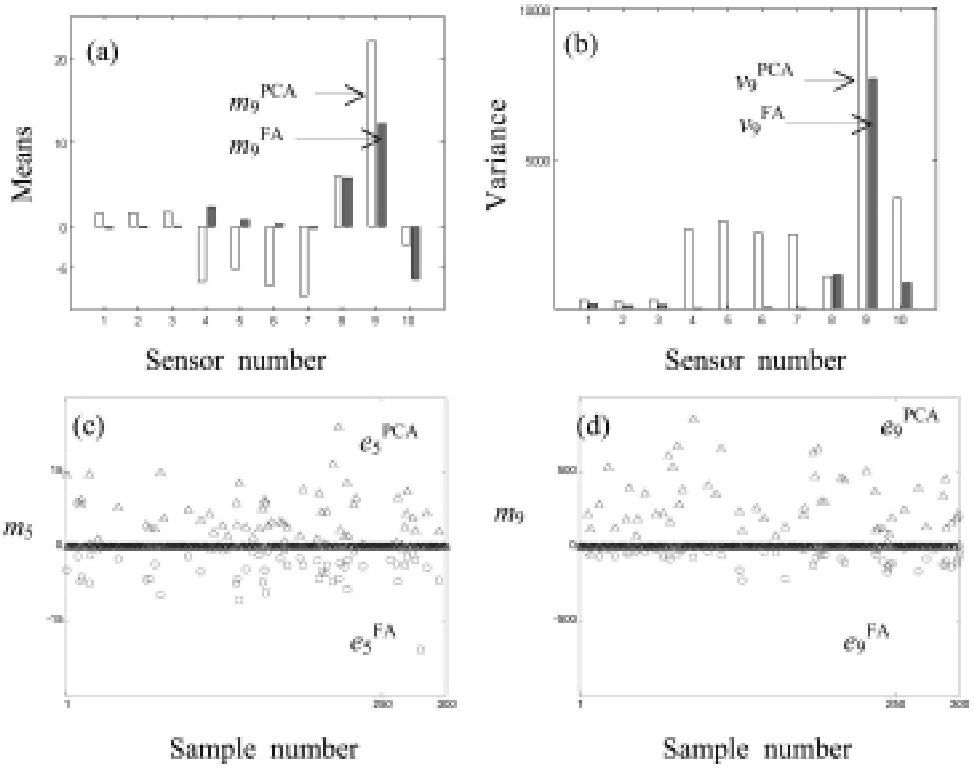

Fig. 4 Comparisons of the prediction performance between the PCA based method and the FA based method. (a) Means of prediction errors of missing values by PCA and FA, (b) variances of prediction errors by PCA and FA, (c) minimal variance variable, x_5, and (d) maximal variance variable, x_9

two subsets: a calibration set $\mathbf{X}_{\mathrm{cali}} = \{\boldsymbol{x}_n\}_{n\in 250}$ and a test set $\mathbf{X}_{\mathrm{test}} = \{\boldsymbol{x}_n\}_{n=251,\ldots,300}$. Multivariate models based on both PCA and FA are calibrated from the observed part in $\mathbf{X}_{\mathrm{cali}}$, i.e. $\{\boldsymbol{x}^{\mathrm{o}}_n\}_{n\in 250}$. Using the developed models, the missing values in the n-th calibration sample $\boldsymbol{x}^{\mathrm{m}}_n$ by $\boldsymbol{x}^{\mathrm{o}}_n$ are estimated and the target values in the n-th test sample $\boldsymbol{x}^{\mathrm{t}}_n$ by $\boldsymbol{x}^{\mathrm{i}}_n$, are predicted.

Figure 3 shows the variance explanation ratios with an increased factor number and each sensor's variance explanation ratio. The appropriate number of factors is found to be four by plotting averaged variance explanation ratios as increasing the factor number from 1 to 8 in Figure 3(a). Under the four factor model structure, Figure 3(b) depicts each sensor's variance explanation ratios and 78% variances of individual elements are explained by the FA model. Notice that the ratios are arranged by $r_2 > r_3 \ldots > r_4 > r_8$; however, the order of the ratios does not depend on the variances of the variables but on the correlations between measurements and factors.

Prediction results for $\{\boldsymbol{x}^{\mathrm{m}}_n\}_{n\in 250}$ and $\{\boldsymbol{x}^{\mathrm{t}}_n\}_{n=251,\ldots,300}$ using the PCA and the FA models are presented in **Figure 4**. Here, the means and the variances of the prediction errors estimated by PCA and FA for the missing values are shown in hollow bars and in gray bars, respectively. The mean and the variance of the p-th variable are shown as m_p and v_p, respectively. These were calculated by (a) $m_p = N_{\mathrm{pm}}^{-1}\cdot\sum_{n\in N\mathrm{pm}}(x_{p\,n}^{\,\mathrm{m}} - \underline{x}_{p\,n}^{\,\mathrm{m}})$, and (b) $v_p = (N_{\mathrm{pm}} - 1)^{-1}\cdot\sum_{n\in N\mathrm{pm}}[(x_{p\,n}^{\,\mathrm{m}} - \underline{x}_{p\,n}^{\,\mathrm{m}}) - m_p]^2$, where N_{pm} represents the number of missing elements in the p-th variable, and under bar denotes the estimate. Among all the variables, two extreme cases which are (c) minimal variance variable, x_5, and (d) maximal variance variable, x_9, are chosen for the detailed comparison. In order to simplify the comparison, ± absolute values are plotted in both (c) and (d), e.g. $e_{\mathrm{PCA}} = \mathrm{abs}(x^{\mathrm{m}} - \underline{x}^{\mathrm{m}}_{\mathrm{PCA}})$ with a triangle symbol, $e_{\mathrm{FA}} = -\mathrm{abs}(x^{\mathrm{m}} - \underline{x}^{\mathrm{m}}_{\mathrm{FA}})$ with a circle mark. As shown in the figure, The FA model gives better results for both biases and variances of the prediction errors than the PCA one. The results are straightforward because FA is devised to consider the probability density information of all the variables and its measuring unit is the Mahalanobis distance, while PCA is governed by Euclidian distance measure.

There are a calibrated multivariate model of the process and an incomplete new sample which is generated from the concerned process. Then there are two important questions which should be answered to monitor the process status when analyzing a new sample: the first question is whether the developed model is still valid to the sample or not, and the second question is if the model is proper to the sample, how we can decide the process condition from the sample, i.e.,

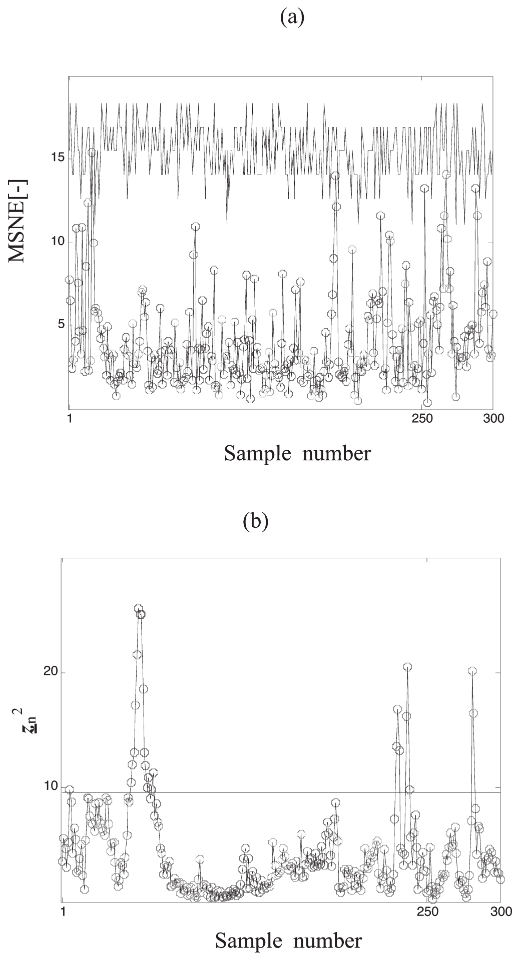

Fig. 5 Multivariate control charts of the FA model, (a) model error test, (b) process' in-control test

in-control or out-of-control. If there were no missing value in the tested sample, Q and $\mathcal{T}^2$ tests of the PCA model might be the answers to these questions. When the sample has any missing elements, these questions should be answered only by the available part in the sample, i.e. $\boldsymbol{x}^{\mathrm{i}}_n$.

When only partial information of the sample is given, **Figure 5** can be the answers for the first question and the second question, respectively. Figure 5 shows multivariate control charts of the FA model. Figure 5(a) shows the model error test of the tested samples using Mahalanobis squared norms of input error scores (MSNE), and Figure 5(b) shows in-control test using Mahalanobis squared norms of extracted factor scores. Chi-square upper control limits with solid lines were calculated regarding the 0.05% level of significance in both the charts, that is, 95% upper control limits. Notice that the 95% upper control limits in Figure 5(a) are varied according to the number of available measurements in the sample. However, the limit to check the in-control in Figure 5(b) is invariant to all the samples because of dim($\mathbf{z}_j$) = 4 $\forall$ j. As shown in Figure 5(a), both the available parts in calibration samples, $\{\boldsymbol{x}^{\mathrm{o}}_n\}_{n=1,\ldots,250}$, and test samples, $\{\boldsymbol{x}^{\mathrm{i}}_n\}_{n=251,\ldots,300}$, indicate that all samples are harmonized with the developed FA model with the 0.05% level of significance. Extracted factor scores from the FA model are enough to decide the current process condition. If there were some violations in the figure, we can also test which elements in the available part, $\boldsymbol{x}^{\mathrm{o}}_n$ or $\boldsymbol{x}^{\mathrm{i}}_n$, are responsible for these violations using Eq. (48). Therefore, the following general decision can come out that there is an out-of-control sample by Figure 5(b) and the propriety of the decision is confirmed by Figure 5(a).

Conclusions

A new calibration method of the FA model from an incomplete calibration set and a prediction method for the missing values in the calibration set and a newly measured sample are proposed. The proposed method gives better estimation results for the missing values than the well-known PCA based method. The results come from an underlying fundamental difference in each method. While PCA seeks to find a least-square sense optimal solution and a Euclidian distance is used as the measuring unit of similarities among data, FA aims to get a maximum likelihood optimal solution and the Mahalanobis distance is to be the unit. On the other hand, when they are applied to auto-scaled data, the PCA based method results similar to the FA based method since the Euclidian distance is converted to the Mahalanobis distance. However, the scaling is often impossible when the data has some missing values.

The probabilistic model such as FA is more suitable than the other projection models when a statistical decision should be made. All kinds of statistical decisions are based on statistics; more generally, it is the matter of the probability densities of variables within the model. But the model building and the decision making are completely separated in the projection model, e.g. PCA based process monitoring. In the probabilistic modeling approach, building the model and making the decision are done in one fold. Another good merit, which was not discussed in this article, of the probabilistic approaches is that it can be extended to mixture modeling, e.g. a mixture of Gaussians (Duda *et al*., 2001), a mixture of factor analyzers (Ghahramani and Hinton, 1996), a mixture of probabilistic PCA (Tipping and Bishop, 1997), etc. In fact, there is no guarantee of assuming that all probabilistic density functions of the FA model would be Gaussians. But there is a good approximation technique of non-Gaussian density through the summations of several Gaussian densities, so called a mixture of Gaussian modeling (MOG). For a nonlinear function, there are two types of modeling techniques: (1) to model it by appropriate nonlinear function directly, or (2) to develop several local linear models for a nonlinear function, then approximate the original to the linear combinations of the local models. Future work on the both topics, called a generic nonlinear modeling and a mixture modeling of FA, will be carried out.

Acknowledgment

This work was supported by a Korea Research Foundation Grant (KRF-2004-041-D00151) and also by a grant (No. R01-2002-000-00007-0) from Korea Science & Engineering Foundation.

Appendix A: FA Model Calibration from Incomplete Data Set

FA model calibration from an incomplete data set is trained by the EM algorithm.

Set initial Θ then

$$\text{E-step: } Q(\Theta|\Theta_t) = \mathcal{E}[\mathcal{L}(\mathbf{D}\text{: } \Theta)|\mathbf{X}^{\mathrm{o}}\text{: } \Theta_t] \tag{A1}$$

$$\text{M-step: } \Theta_{t+1} = \arg_{\mathrm{Q}} \max\text{: } Q(\Theta|\Theta_t) \tag{A2}$$

Until Θ converges, iterate the E-step and the M-step:

The idea is realized as follows:

Set initial $\Theta = \{\mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ and let $\mathbf{A}' = [\mathbf{A}, \boldsymbol{\mu}]$.

Until Θ converge, iterate the followings:

For n =1, 2, …, N,

if $\dim(\boldsymbol{x}^{\mathrm{o}}{}_n) \le \dim(\boldsymbol{z})$, then $\boldsymbol{\Sigma}^{\mathrm{oo}}{}_n{}^{-1} = (\mathbf{A}^{\mathrm{o}}{}_n \cdot \mathbf{A}^{\mathrm{o}}{}_n{}^{\mathrm{T}} + \boldsymbol{\Lambda}^{\mathrm{o}}{}_n)^{-1}$

if $\dim(\boldsymbol{x}^{o}{}_n) > \dim(\boldsymbol{z})$,
then $\boldsymbol{\Sigma}^{\mathrm{oo}}{}_n{}^{-1} = \boldsymbol{\Lambda}^{\mathrm{o}}{}_n{}^{-1} - \boldsymbol{\Lambda}^{\mathrm{o}}{}_n{}^{-1} \cdot \mathbf{A}^{\mathrm{o}}{}_n \cdot (\mathbf{I} + \mathbf{A}^{\mathrm{o}}{}_n{}^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{\mathrm{o}}{}_n{}^{-1} \cdot \mathbf{A}^{\mathrm{o}}{}_n)^{-1} \cdot \mathbf{A}^{\mathrm{o}}{}_n{}^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{\mathrm{o}}{}_n{}^{-1}$

$$\beta_n = \mathbf{A}^{\mathrm{o}}{}_n{}^{\mathrm{T}} \cdot \boldsymbol{\Sigma}^{\mathrm{oo}}{}_n{}^{-1} \tag{A3}$$

$$\underline{\boldsymbol{z}}_n = \beta_n \cdot (\boldsymbol{x}^{\mathrm{o}}{}_n - \boldsymbol{\mu}^{\mathrm{o}}{}_n)\text{: } \boldsymbol{\Sigma}^{\mathrm{z|o}}{}_n = \mathbf{I} - \beta_n \cdot \mathbf{A}^{\mathrm{o}}{}_n\text{: } \underline{\boldsymbol{zz}}_n = \boldsymbol{\Sigma}^{\mathrm{z|o}}{}_n + \underline{\boldsymbol{z}}_n \cdot \underline{\boldsymbol{z}}_n{}^{\mathrm{T}} \tag{A4}$$

$$\underline{\boldsymbol{z}}'{}_n = [\underline{\boldsymbol{z}}_n; 1]\text{: } \underline{\boldsymbol{zz}}'{}_n = [\underline{\boldsymbol{zz}}_n, \underline{\boldsymbol{z}}_n; \underline{\boldsymbol{z}}_n{}^{\mathrm{T}}, 1] \tag{A5}$$

$$\underline{\boldsymbol{x}}^{m}{}_n = \mathbf{A}^{\mathrm{m}}{}_n \cdot \underline{\boldsymbol{z}}_n + \boldsymbol{\mu}^{\mathrm{m}}{}_n\text{: } \boldsymbol{\Sigma}^{\mathrm{m|z}}{}_n = \boldsymbol{\Lambda}^{\mathrm{m}}{}_n\text{: } \underline{\boldsymbol{xx}}^{\mathrm{m}}{}_n = \boldsymbol{\Sigma}^{\mathrm{m|z}}{}_n + \underline{\boldsymbol{x}}^{\mathrm{m}}{}_n \cdot \underline{\boldsymbol{x}}^{\mathrm{m}}{}_n{}^{\mathrm{T}} \tag{A6}$$

$$\underline{\boldsymbol{x}}_n = [\boldsymbol{x}^{\mathrm{o}}{}_n; \underline{\boldsymbol{x}}^{\mathrm{m}}{}_n]\text{: } \underline{\boldsymbol{xx}}_n = [\boldsymbol{x}^{\mathrm{o}}{}_n \cdot \boldsymbol{x}^{\mathrm{o}}{}_n{}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{o}}{}_n \cdot \underline{\boldsymbol{x}}^{\mathrm{m}}{}_n{}^{\mathrm{T}}; \underline{\boldsymbol{x}}^{\mathrm{m}}{}_n \cdot \boldsymbol{x}^{\mathrm{o}}{}_n{}^{\mathrm{T}}, \underline{\boldsymbol{xx}}^{\mathrm{m}}{}_n] \tag{A7}$$

End for

$$\mathbf{A}' = (\textstyle\sum_{n\in N} \underline{\boldsymbol{x}}_n \cdot \underline{\boldsymbol{z}}'{}_n{}^{\mathrm{T}}) \cdot (\textstyle\sum_{n\in N} \underline{\boldsymbol{zz}}'{}_n)^{-1} \tag{A8}$$

$$\boldsymbol{\Lambda} = N^{-1} \cdot \mathrm{diag} \textstyle\sum_{n\in N} (\underline{\boldsymbol{xx}}_n - \mathbf{A}' \underline{\boldsymbol{z}}'{}_n \cdot \underline{\boldsymbol{x}}_n{}^{\mathrm{T}}) \tag{A9}$$

$$\mathrm{A} = \mathbf{A}'{}_{[:,\, 1:L]}\text{: } \boldsymbol{\mu} = \mathbf{A}'{}_{[:,\, L+1]} \tag{A10}$$

End until

$$\text{Final } \boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\} \tag{A11}$$

Appendix B: Proof of Eq. (15) Using a Matrix Inversion Lemma

For any invertible **U**, in this case **U** = **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}} \\
&\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}} \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{A}^{\mathrm{T}} \\
&\qquad\qquad - (\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}} \\
&\qquad\qquad - (\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}$$

Literature Cited

Anderson, T. W.; An Introduction to Multivariate Statistical Analysis, 2nd ed., Wiley-Interscience, New York, U.S.A. (1984)

Dempster, A., N. Laird and D. Rubin; "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Stat. Soc. B*, **39**, 1–38 (1977)

Duda, R. O., P. E. Hart and D. G. Stork; Pattern Classification, 2nd ed., Wiley-Interscience, New York, U.S.A. (2001)

Geladi, P.; "Some Recent Trends in the Calibration Literature," *Chemom. Intell. Lab. Syst.*, **60**, 211–224 (2002)

Ghahramani, Z. and G. E. Hinton; The EM Algorithm for Mixtures of Factor Analyzers, Technical Report CRG-TR-96-1, University of Toronto, Canada (1996)

Grung, B. and R. Manne; "Missing Values in Principal Component Analysis," *Chemom. Intell. Lab. Syst.*, **42**, 125–139 (1998)

Hotelling, H.; Multivariate Quality Control, Illustrated by the Air Testing of Sample Bombsights Techniques of Statistical Analysis, C. Eisenhart, M. W. Hastay and W.A. Wallis eds., McGraw-Hill, New York, U.S.A. (1947)

Jackson, J. and G. Mudholkar; "Control Procedures for Residuals Associated with Principal Component Analysis," *Techometrics*, **21**, 341–349 (1979)

Kalivas, J. H.; "Interrelationships of Multivariate Regression Methods Using Eigenvector Basis Sets," *J. Chemom.*, **13**, 111–132 (1999)

Kim, D. S. and I. Lee; "Process Monitoring Based on Probabilistic PCA," *Chem. Intell. Lab. Syst.*, **67**, 109–123 (2003)

MacGregor, J. F. and T. Kourti; "Statistical Process Control of Multivariate Processes," *Contr. Eng. Pract.*, **3**, 403–414 (1995)

Stork, C. L., D. J. Veltkamp and B. R. Kowalski; "Identification of Multiple Sensor Disturbances during Process Monitoring," *Anal. Chem.*, **69**, 5031–5036 (1997)

Tipping, M. E. and C. M. Bishop; Mixtures of Probabilistic Principal Component Analyzers, Technical Report NCRG/97/003, pp. 1–29, Aston University, Birmingham, U.K. (1997a)

Tipping, M. E. and C. M. Bishop; "Probabilistic Principal Component Analysis," *J. Roy. Stat. Soc. B*, **61**(Part 3), 611–622 (1997b)

Walczak, B. and D. L. Massart; "Dealing with Missing Data Part 1," *Chemom. Intell. Lab. Syst.*, **58**, 15–27 (2001)

Wise, B. and N. Gallagher; "The Process Chemometrics Approach to Process Monitoring and Fault Detection," *J. Process Contr.*, **6**, 329–348 (1996)