Mixture Semisupervised Principal Component Regression Model and Soft Sensor Application

Zhiqiang Ge
State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Dept. of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

Dept. of Chemical and Materials Engineering, University of Alberta, Edmonton AB T6G2G6, Canada

Biao Huang
Dept. of Chemical and Materials Engineering, University of Alberta, Edmonton AB T6G2G6, Canada

Zhihuan Song
State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Dept. of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

DOI 10.1002/aic.14270
Published online November 9, 2013 in Wiley Online Library (wileyonlinelibrary.com)

Traditionally, data-based soft sensors are constructed upon the labeled historical dataset which contains equal numbers of input and output data samples. While it is easy to obtain input variables such as temperature, pressure, and flow rate in the chemical process, the output variables, which correspond to quality/key property variables, are much more difficult to obtain. Therefore, we may only have a small number of output data samples, and have much more input data samples. In this article, a mixture form of the semisupervised probabilistic principal component regression model is proposed for soft sensor application, which can efficiently incorporate the unlabeled data information from different operation modes. Compared to the total supervised method, both modeling efficiency and soft sensing performance are improved with the inclusion of additional unlabeled data samples. Two case studies are provided to evaluate the feasibility and efficiency of the new method. © 2013 American Institute of Chemical Engineers *AIChE J*, 60: 533–545, 2014

Keywords: semisupervised modeling, probabilistic principal component regression, soft sensor, mixture probabilistic modeling

Introduction

Nowadays, soft sensors have been widely used for property estimation of quality and key variables in chemical processes. This is mainly because those important variables are difficult to measure online; instead, their values are usually obtained through analyzers or lab analyses. However, both analyzers and lab analyses are expensive, time consuming, and introduce a significant time delay to the control system. The main advantage of the soft sensor is that it can provide real-time measurements for those important variables, by using other highly correlated but easy-to-measure process variables. Traditional soft sensors are developed through first-principle models, which are subject to process knowledge and experiences of experts. However, the acquisition of both process knowledge and expert experiences is difficult and time consuming, especially for modern complex chemical processes.

On the other hand, without requirements of either detailed process knowledge or expert experiences, data-based soft sensor has become popular in industrial applications. In the past years, a huge volume of process data has been recorded by the distributed control system, which contains important information of the process. Therefore, by constructing relationships between secondary process variables and quality/key variables, a data-based soft sensor can be formulated. Among all developed soft sensor models, conventionally used ones include: principal component analysis/principal component regression (PCR),[1–3] partial least squares (PLS),[4–8] artificial neural networks,[9–12] kernel-based models,[13–18] Bayesian methods[19–22] and so forth.

Typically, both of the input data (from secondary process variables) and output data (from quality/key variables) are required for soft sensor modeling. Here, we represent the dataset which contains both input and output data samples as the labeled dataset, the one which only consists of input data samples is denoted as the unlabeled dataset. Traditional data-based soft sensors are usually built upon the labeled dataset. However, the output data of the soft sensor which correspond to quality/key variables are usually difficult to obtain, for example, through complex lab analysis. Also, compared to the input data, the sampling rate of the output data is often much lower, and as a result, only a small portion of the input data samples have their corresponding output data

Correspondence concerning this article should be addressed to B. Huang at biao.huang@ualberta.ca.

© 2013 American Institute of Chemical Engineers

values; others are unlabeled. Therefore, in practice, we may only have a small number of labeled data samples for soft sensor modeling, and a large number of other unlabeled data samples could have been ignored.

The motivation of the present article is to incorporate both the labeled and unlabeled datasets for soft sensor modeling. Although the unlabeled dataset has no output values, it can contain important process information. From a probabilistic viewpoint, the distribution of the input variables can hardly be captured by a small number of data samples. By utilizing more input data samples, the estimation of the distribution of the input variables could be significantly improved. For a latent probabilistic model, the input and output data are connected by latent variables which can be extracted from input variables. Therefore, if the estimation of the distribution of the input data can be improved by the added unlabeled data samples, the quality of the extracted latent variables can also be improved, which in turn results in an improved relationship between input and output variables.

In our previous article, a probabilistic form of the semisupervised PCR model has been introduced, which can efficiently incorporate the unlabeled data for soft sensor modeling.[23] However, the inherent nature of this single probabilistic model has limited the soft sensor to linear and single mode processes. For those processes which have several different operation modes or the relationship between input and output data is nonlinear, the semisupervised PCR-based soft sensor may not function efficiently. For multimode process modeling, there are already several useful contributions, for example, local modeling approach, external analysis method, Bayesian inference-based strategy and so forth.[24–29] In the present article, a mixture probabilistic form of the semisupervised PCR model is developed, in which several local probabilistic PCR (PPCR) models are formulated through Bayesian inference and posterior estimation. Based on the developed mixture semisupervised PCR model, a new soft sensor is then constructed for estimation of quality/key variables for chemical processes. For online soft sensing of a new data sample, the estimated output of each local PPCR model is combined through a weighted probabilistic coefficient, which is based on the posterior probability of each local model corresponding to the new data sample. Therefore, compared to the single model structure, the mixture probabilistic semisupervised PCR model can provide a soft combination result from different local models; meanwhile, the mode or membership information can be automatically located for each operating region. While the single model is restricted in Gaussian, linear, and single operating mode processes, the mixture model can be used in more general cases, for example, multimode processes, nonlinear processes, non-Gaussian processes, and so forth.

The rest of this article is organized as follows. In section entitled Preliminaries, brief introductions of traditional PCR, probabilistic PCR, and the semisupervised PPCR models are given, followed by the detailed description of the mixture semisupervised PCR model in section entitled Mixture Semisupervised PCR Model Development. In section entitled Online Soft Sensing Based on Mixture Semisupervised PCR Model, a new soft sensor is developed based on the mixture semisupervised PCR model. Case studies of a numerical simulation example and an industrial application are provided in section entitled Case studies. Finally, conclusions are made in section entitled Conclusions.

Preliminaries

PCR

Given the input and output dataset, $\mathbf{X} \in R^{n\times m}$ and $\mathbf{Y} \in R^{n\times r}$, where n is the number of data sample, m and r are numbers of input and output variables. The aim of PCR is to find a set of principal components which span the original measurement variable space. The procedure of PCR can be divided into two steps. The first step is to extract principal components from the input dataset $\mathbf{X}$, and the second step is to calculate the regression matrix between the extracted principal components and the output dataset $\mathbf{Y}$. The PCR model structure is given as follows[2]

$$\mathbf{X}=\mathbf{T}\mathbf{P}^T+\mathbf{E} \tag{1}$$

$$\mathbf{Y}=\mathbf{T}\mathbf{C}^T+\mathbf{F} \tag{2}$$

where $\mathbf{P} \in R^{m\times q}$ is the loading matrix, $\mathbf{T} \in R^{n\times q}$ is the principal component matrix, q is the selected number of principal components, $\mathbf{C} \in R^{r\times q}$ is the regression matrix, and $\mathbf{E}$ and $\mathbf{F}$ are the residuals matrices with appropriate dimensions.

PPCR

Given the data information $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^T \in R^{n\times m}$ and $\mathbf{Y}=[\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]^T \in R^{n\times r}$, different from the PCR model, the probabilistic PCR model is derived through the following generative manner[23]

$$\mathbf{x}=\mathbf{P}\mathbf{t}+\mathbf{e} \tag{3}$$

$$\mathbf{y}=\mathbf{C}\mathbf{t}+\mathbf{f} \tag{4}$$

where $\mathbf{P} \in R^{m\times q}$, $\mathbf{C} \in R^{r\times q}$ are weighting matrices, $\mathbf{t} \in R^{q\times 1}$ is the latent variable vector, $\mathbf{e} \in R^{m\times 1}$ and $\mathbf{f} \in R^{r\times 1}$ are measurement noises of input and output variables. In this probabilistic model, it is assumed that both probability density functions of the latent variable and the measurement noise are Gaussian, that is, $p(\mathbf{t})=N(0, \mathbf{I})$, $p(\mathbf{e})=N(0, \sigma_{\mathbf{x}}^2\mathbf{I})$, and $p(\mathbf{f})=N(0, \sigma_{\mathbf{y}}^2\mathbf{I})$, where $\mathbf{I}$ is an identity matrix, $\sigma_{\mathbf{x}}^2$ and $\sigma_{\mathbf{y}}^2$ are noise variances of input and output variables. Therefore, based on the property of conditional independence, the marginal probability $p(\mathbf{x}, \mathbf{y})$ can be formulated by integrating out the latent variables, which is given as follows

$$p(\mathbf{x}, \mathbf{y}|\mathbf{P}, \mathbf{C}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2)=\int p(\mathbf{x}|\mathbf{t}, \mathbf{P}, \sigma_{\mathbf{x}}^2)p(\mathbf{y}|\mathbf{t}, \mathbf{C}, \sigma_{\mathbf{y}}^2)p(\mathbf{t})d\mathbf{t} \tag{5}$$

The optimal model parameter set $\{\mathbf{P}, \mathbf{C}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2\}$ can be determined by maximizing the following likelihood function

$$L(\mathbf{P}, \mathbf{C}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2)=\ln \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{y}_i|\mathbf{P}, \mathbf{C}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2) \tag{6}$$

Semisupervised PCR model

Similar to the probabilistic PCR model, the semisupervised PCR model is also formulated through the generative structure, given as follows[23]

$$\begin{aligned}\mathbf{x}_i&=\mathbf{P}\mathbf{t}_i+\mathbf{e}_i\\ \mathbf{y}_j&=\mathbf{C}\mathbf{t}_j+\mathbf{f}_j\end{aligned}\tag{7}$$

where $i=1,2,\cdots,n$, $j=1,2,\cdots,n_1$, n_1 is the size of the labeled dataset, and $n_2=n-n_1$ is the size of the unlabeled dataset. $\mathbf{P}\in R^{m\times q}$, $\mathbf{C}\in R^{r\times q}$ are weighted matrices, where m is the number of input variables, and r is the number of output variables. $\mathbf{t}\in R^{q\times 1}$ is the latent variable vector, $\mathbf{e}\in R^{m\times 1}$ and $\mathbf{f}\in R^{r\times 1}$ are noises of input and output variables, respectively. Also, it is assumed that both probability density functions of the latent variable and the measurement noise are Gaussian. Given labeled dataset $\mathbf{X}_1=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_{n_1}]^T$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_{n_1}]^T$ and unlabeled dataset $\mathbf{X}_2=[\mathbf{x}_{n_1+1},\mathbf{x}_{n_1+2},\cdots,\mathbf{x}_{n_1+n_2}]^T$, the marginal distribution can be calculated as follows

$$p\left(\mathbf{x}_j,\mathbf{y}_j|\mathbf{P},\mathbf{C},\sigma_{\mathbf{x}}^2,\sigma_{\mathbf{y}}^2\right)=\int p(\mathbf{x}_j|\mathbf{t}_j,\mathbf{P},\sigma_{\mathbf{x}}^2)p(\mathbf{y}_j|\mathbf{t}_j,\mathbf{C},\sigma_{\mathbf{y}}^2)p(\mathbf{t}_j)d\mathbf{t}_j\tag{8}$$

$$p(\mathbf{x}_{n_1+i},|\mathbf{P},\sigma_{\mathbf{x}}^2)=\int p(\mathbf{x}_{n_1+i}|\mathbf{t}_{n_1+i},\mathbf{P},\sigma_{\mathbf{x}}^2)p(\mathbf{t}_{n_1+i})d\mathbf{t}_{n_1+i}\tag{9}$$

where $j=1,2,\cdots,n_1$, $i=1,2,\cdots,n_2$. Following the maximum likelihood framework, the log likelihood function can be derived as

$$\begin{aligned}L(\mathbf{X},\mathbf{Y})=L(\mathbf{X}_1,\mathbf{Y})+L(\mathbf{X}_2)=\ln\prod_{j=1}^{n_1}p(\mathbf{x}_j,\mathbf{y}_j|\mathbf{P},\mathbf{C},\sigma_{\mathbf{x}}^2,\sigma_{\mathbf{y}}^2)\\ +\ln\prod_{i=1}^{n_2}p(\mathbf{x}_{n_1+i}|\mathbf{P},\sigma_{\mathbf{x}}^2)\end{aligned}\tag{10}$$

Through maximizing the Log likelihood function, the parameter set of the semisupervised PCR model $\Theta=\{\mathbf{P},\mathbf{C},\sigma_{\mathbf{x}}^2,\sigma_{\mathbf{y}}^2\}$ can be determined.

Mixture Semisupervised PCR Model Development

In the mixture semisupervised PCR model, we first assume that K individual semisupervised PCR models have been incorporated, and q latent variables are retained in each submodel. The mixture form of the semisupervised PCR model can be formulated as follows

$$\begin{aligned}&\mathbf{x}_{i,k}=\mathbf{P}_k\mathbf{t}_{i,k}+\mathbf{e}_{i,k},k=1,2,\cdots,K\\ &\mathbf{y}_{j,k}=\mathbf{C}_k\mathbf{t}_{j,k}+\mathbf{f}_{j,k},k=1,2,\cdots,K\\ &\mathbf{x}_i=\begin{cases}\sum_{k=1}^{K}p_1(k)\mathbf{x}_{i,k} & 1\le i\le n_1\\ \sum_{k=1}^{K}p_2(k)\mathbf{x}_{i,k} & n_1+1\le i\le n\end{cases}\\ &\mathbf{y}_j=\sum_{k=1}^{K}p_1(k)\mathbf{y}_{j,k}\end{aligned}\tag{11}$$

where $i=1,2,\cdots,n$, $j=1,2,\cdots,n_1$, and we assume that the total number of data samples is n, among which n_1 samples are labeled, and $n_2=n-n_1$ samples are unlabeled. $p_1(k)$ and $p_2(k)$ are mixing proportional values of each individual model for labeled and unlabeled dataset, with constraint $\sum_{k=1}^{K}p_1(k)=1$ and $\sum_{k=1}^{K}p_2(k)=1$. $\mathbf{P}_k$ and $\mathbf{C}_k$ are weighting matrices of the kth individual semisupervised PCR model, $\mathbf{t}_k\in R^{q\times 1}$ is the latent variable vector, $\mathbf{e}_k\in R^{m\times 1}$ and $\mathbf{f}_k\in R^{r\times 1}$ are noise vectors of input and output variables in the corresponding model. It is assumed that both probability density functions of the latent variable and the measurement noise in each individual model are Gaussian; thus $p(\mathbf{t}_k)=N(0,\mathbf{I})$, $p(\mathbf{e}_k)=N(0,\sigma_{\mathbf{x},k}^2\mathbf{I})$, and $p(\mathbf{f}_k)=N(0,\sigma_{\mathbf{y},k}^2\mathbf{I})$. Therefore, the multivariate Gaussian distributions of the input and output data in the kth individual model are given as $p(\mathbf{x}|\mathbf{t}_k)=N(\mathbf{P}_k\mathbf{t}_k+\boldsymbol{\mu}_{\mathbf{x},k},\sigma_{\mathbf{x},k}^2)$ and $p(\mathbf{y}|\mathbf{t}_k)=N(\mathbf{C}_k\mathbf{t}_k+\boldsymbol{\mu}_{\mathbf{y},}\ k,\sigma_{\mathbf{y},k}^2)$. Based on the property of conditional independence of the input and output variables, that is, all input and output variables are conditionally independent to each other given the latent variables, the marginal distribution of the labeled and unlabeled data in each individual model can be determined as follows

$$\begin{aligned}p(\mathbf{x},\mathbf{y}|\mathbf{P}_k,\mathbf{C}_k,\sigma_{\mathbf{x},k}^2,\sigma_{\mathbf{y},k}^2)=\int p(\mathbf{x}|\mathbf{t}_k,\mathbf{P}_k,\sigma_{\mathbf{x},k}^2)\\ p(\mathbf{y}|\mathbf{t}_k,\mathbf{C}_k,\sigma_{\mathbf{y},k}^2)p(\mathbf{t}_k)d\mathbf{t}_k\end{aligned}\tag{12}$$

$$p(\mathbf{x}|\mathbf{P}_k,\sigma_{\mathbf{x},k}^2)=\int p(\mathbf{x}|\mathbf{t}_k,\mathbf{P}_k,\sigma_{\mathbf{x},k}^2)p(\mathbf{t}_k)d\mathbf{t}_k\tag{13}$$

Given the labeled datasets $\mathbf{X}_1=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_{n_1}]^T\in R^{n_1\times m}$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_{n_1}]^T\in R^{n_1\times r}$ and the unlabeled dataset $\mathbf{X}_2=[\mathbf{x}_{n_1+1},\mathbf{x}_{n_1+2},\cdots,\mathbf{x}_{n_1+n_2}]^T\in R^{n_2\times m}$, to obtained the optimal parameter set, the following likelihood function should be maximized

$$p(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})=p(\mathbf{X}_1,\mathbf{Y}|\boldsymbol{\Theta})p(\mathbf{X}_2|\boldsymbol{\Theta})\tag{14}$$

For simplicity, the likelihood function can be transformed to the log likelihood form as follows

$$\begin{aligned}L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})=L(\mathbf{X}_1,\mathbf{Y}|\boldsymbol{\Theta})+L(\mathbf{X}_2|\boldsymbol{\Theta})=\ln\prod_{i=1}^{n_1}p(\mathbf{x}_i,\mathbf{y}_i|\boldsymbol{\Theta})\\ +\ln\prod_{i=n_1+1}^{n}p(\mathbf{x}_i|\boldsymbol{\Theta})=\sum_{i=1}^{n_1}\ln p(\mathbf{x}_i,\mathbf{y}_i|\boldsymbol{\Theta})\\ +\sum_{i=n_1+1}^{n}\ln p(\mathbf{x}_i|\boldsymbol{\Theta})=\sum_{i=1}^{n_1}\ln\sum_{k=1}^{K}p(\mathbf{x}_i,\mathbf{y}_i|k,\boldsymbol{\Theta})p_1(k)\\ +\sum_{i=n_1+1}^{n}\ln\sum_{k=1}^{K}p(\mathbf{x}_i|k,\boldsymbol{\Theta})p_2(k)\end{aligned}\tag{15}$$

where $\boldsymbol{\Theta}=\{\boldsymbol{\Theta}\}_k=\{\mathbf{P}_k,\mathbf{C}_k,\sigma_{\mathbf{x},k}^2,\sigma_{\mathbf{y},k}^2,\boldsymbol{\mu}_{\mathbf{x},k},\boldsymbol{\mu}_{\mathbf{y},k}\}$.

In the Expectation-Maximization (EM) algorithm,[30] instead of maximizing the Log likelihood function directly, the maximization of the expected complete-data Log likelihood function is usually carried out. The derivation of the expected complete-data Log likelihood function is provided in Appendix A. The result is given below. First

$$\begin{aligned}L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})=\sum_{i=1}^{n_1}\sum_{k=1}^{K}\ln p(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_{i,k},k|\boldsymbol{\Theta}))\\ +\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}\ln p(\mathbf{x}_i,\mathbf{t}_{i,k},k|\boldsymbol{\Theta}))\end{aligned}\tag{16}$$

Under the EM algorithm, we treat both of $\mathbf{t}$ and k as hidden variables. Therefore, the expected complete-data Log likelihood function value with respect to joint distribution of $\mathbf{t}$ and k can be derived as

$$
\begin{aligned}
E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]=&\sum_{i=1}^{n_1}\sum_{k=1}^{K}\int p(\mathbf{t}_{i,k},k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})\ln\left[p(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_{i,k},k|\boldsymbol{\Theta})\right]d\mathbf{t}_{i,k}\\
&+\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}\int p(\mathbf{t}_{i,k},k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\ln\left[p(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_{i,k},k|\boldsymbol{\Theta})\right]d\mathbf{t}_{i,k}\\
=&\sum_{i=1}^{n_1}\sum_{k=1}^{K}p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})\int p(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})\ln\left[p(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_{i,k},k|\boldsymbol{\Theta})\right]d\mathbf{t}_{i,k}\\
&+\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\int p(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})\ln\left[p(\mathbf{x}_i,\mathbf{t}_{i,k},k|\boldsymbol{\Theta})\right]d\mathbf{t}_{i,k}\\
=&\sum_{i=1}^{n_1}\sum_{k=1}^{K}p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})\int p(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})\ln\left[p(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_{i,k}|k,\boldsymbol{\Theta})p_1(k)\right]d\mathbf{t}_{i,k}\\
&+\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\int p(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})\ln\left[p(\mathbf{x}_i,\mathbf{t}_{i,k}|k,\boldsymbol{\Theta})p_2(k)\right]d\mathbf{t}_{i,k}\\
=&\sum_{i=1}^{n_1}\sum_{k=1}^{K}p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})\{\ln p_1(k)+\int p(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})\ln\left[p(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_{i,k}|k,\boldsymbol{\Theta})\right]d\mathbf{t}_{i,k}\}\\
&+\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\{\ln p_2(k)+\int p(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})\ln\left[p(\mathbf{x}_i,\mathbf{t}_{i,k}|k,\boldsymbol{\Theta})\right]d\mathbf{t}_{i,k}\}
\end{aligned}
\tag{17}
$$

Then, the second step of the EM algorithm can be used for maximization of the expected complete-data Log likelihood function. In the E-step of the EM algorithm, we are given the parameters $\boldsymbol{\Theta}_{\text{old}}$ obtained in the previous M-step, and the aim of this step is to determine posterior probabilities of hidden variables $\mathbf{t}$ and k and obtain the expected likelihood. In the M-step, we update the new parameter set $\boldsymbol{\Theta}_{\text{new}}$ by maximizing the expected complete-data Log likelihood function $E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]$.

In the E-step of the EM algorithm, we are given the parameters $\boldsymbol{\Theta}_{\text{old}}$ obtained in the previous M-step, and to determine two posterior probabilities: $p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})$ and $p(\mathbf{t}_i|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})$ for the labeled dataset, $p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})$ and $p(\mathbf{t}_i|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})$ for the unlabeled dataset. According to the Bayesian rule, $p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})$ and $p(k|\mathbf{x}_i,\Theta_{\text{old}})$ can be determined as

$$
p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})=\frac{p(\mathbf{x}_i,\mathbf{y}_i,|k,\boldsymbol{\Theta}_{\text{old}})p_1(k|\boldsymbol{\Theta}_{\text{old}})}{p(\mathbf{x}_i,\mathbf{y}_i|\boldsymbol{\Theta}_{\text{old}})}\tag{18}
$$

$$
p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})=\frac{p(\mathbf{x}_i,|k,\boldsymbol{\Theta}_{\text{old}})p_2(k|\boldsymbol{\Theta}_{\text{old}})}{p(\mathbf{x}_i|\boldsymbol{\Theta}_{\text{old}})}\tag{19}
$$

where $p_1(k|\Theta_{\text{old}})$ and $p_2(k|\Theta_{\text{old}})$ are proportion values calculated in the previous M-step, and $\sum_{k=1}^{K}p_1(k|\Theta_{\text{old}})=1$ and $\sum_{k=1}^{K}p_2(k|\Theta_{\text{old}})=1$. $p(\mathbf{x}_i,\mathbf{y}_i,|k,\boldsymbol{\Theta}_{\text{old}})$ and $p(\mathbf{x}_i,|k,\boldsymbol{\Theta}_{\text{old}})$ are multivariate Gaussian distributions which can be easily formulated, and the two denominators are normalizing constants which need not be evaluated. Similarly, $p(\mathbf{t}_i|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})$ and $p(\mathbf{t}_i|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})$ can be determined as

$$
p(\mathbf{t}_i|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})=\frac{p(\mathbf{x}_i|\mathbf{t}_i,k,\boldsymbol{\Theta}_{\text{old}})p(\mathbf{y}_i|\mathbf{t}_i,k,\boldsymbol{\Theta}_{\text{old}})p(\mathbf{t}_i|k,\boldsymbol{\Theta}_{\text{old}})}{p(\mathbf{x}_i,\mathbf{y}_i|k,\boldsymbol{\Theta}_{\text{old}})}\tag{20}
$$

$$
p(\mathbf{t}_i|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})=\frac{p(\mathbf{x}_i|\mathbf{t}_i,k,\boldsymbol{\Theta}_{\text{old}})p(\mathbf{t}_i|k,\boldsymbol{\Theta}_{\text{old}})}{p(\mathbf{x}_i|k,\boldsymbol{\Theta}_{\text{old}})}\tag{21}
$$

As all terms are Gaussian distributed, $p(\mathbf{t}_i|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})$ and $p(\mathbf{t}_i|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})$ are also Gaussian, with their expected means and variances given as follows[23,31]

$$
\begin{aligned}
E(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})=&(\sigma_{\mathbf{x},k}^{-2}\mathbf{P}_k^T\mathbf{P}_k+\sigma_{\mathbf{y},k}^{-2}\mathbf{C}_k^T\mathbf{C}_k+\mathbf{I})^{-1}\\
&[\sigma_{\mathbf{x},k}^{-2}\mathbf{P}_k^T(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})+\sigma_{\mathbf{y},k}^{-2}\mathbf{C}_k^T(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k})]
\end{aligned}\tag{22}
$$

$$
\begin{aligned}
E(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})=&(\sigma_{\mathbf{x},k}^{-2}\mathbf{P}_k^T\mathbf{P}_k+\sigma_{\mathbf{y},k}^{-2}\mathbf{C}_k^T\mathbf{C}_k+\mathbf{I})^{-1}\\
&+E(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})E^T(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})
\end{aligned}\tag{23}
$$

where $i=1,2,\cdots,n_1$, and

$$
E(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})=(\sigma_{\mathbf{x},k}^{2}\mathbf{I}+\mathbf{P}_k^T\mathbf{P}_k)^{-1}\mathbf{P}_k^T(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})\tag{24}
$$

$$
\begin{aligned}
E(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})=&\sigma_{\mathbf{x},k}^{2}(\mathbf{P}_k^T\mathbf{P}_k+\sigma_{\mathbf{x},k}^{2}\mathbf{I})^{-1}\\
&+E(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})E^T(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})
\end{aligned}\tag{25}
$$

where $i=n_1+1,n_1+2,\cdots,n$.

In the M-step of the EM algorithm, the new parameters are updated by maximizing the expected complete-data Log likelihood function $E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]$. The results of updated parameters are given as

$$p_1(k)=\frac{1}{n_1}\sum_{i=1}^{n_1} p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}) \tag{26}$$

$$p_2(k)=\frac{1}{n_2}\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}}) \tag{27}$$

$$p(k)=\frac{1}{n}\left\{\sum_{i=1}^{n_1} p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})+\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\right\} \tag{28}$$

$$\boldsymbol{\mu}_{\mathbf{x},k}^{\text{new}}=\frac{\begin{array}{c}\sum_{i=1}^{n_1} p(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}})[\mathbf{x}_i-\mathbf{P}_k E(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})]\\ +\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\Theta_{\text{old}})[\mathbf{x}_i-\mathbf{P}_k E(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})]\end{array}}{\sum_{i=1}^{n_1} p(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}})+\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\Theta_{\text{old}})} \tag{29}$$

$$\boldsymbol{\mu}_{\mathbf{y},k}^{\text{new}}=\frac{\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right)\left[\mathbf{y}_i-\mathbf{C}_k E\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]}{\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right)} \tag{30}$$

$$\begin{aligned}\mathbf{P}_k^{\text{new}}=&\left[\sum_{i=1}^{n_1} p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})E^T(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})\right.\\ &\left.+\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})E^T(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})\right]\\ &\times\left[\sum_{i=1}^{n_1} p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right.\\ &\left.+\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]^{-1}\end{aligned} \tag{31}$$

$$\begin{aligned}\mathbf{C}_k^{\text{new}}=&\left[\sum_{i=1}^{n_1} p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})\left(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k}\right)E^T(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})\right]\\ &\times\left[\sum_{i=1}^{n_1} p(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}})E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]^{-1}\end{aligned} \tag{32}$$

$$\sigma_{\mathbf{x},k}^{2\text{new}}=\frac{\begin{array}{c}\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left\{(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})^T(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})-2E^T(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})\mathbf{P}_k^{\text{new}\,T}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})\right.\\ \left.+trace\left[\mathbf{P}_k^{\text{new}\,T}\mathbf{P}_k^{\text{new}}E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]\right\}+\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\left\{(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})^T(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})\right.\\ \left.-2E^T(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}})\mathbf{P}_k^{\text{new}\,T}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k})+trace\left[\mathbf{P}_k^{\text{new}\,T}\mathbf{P}_k^{\text{new}}E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]\right\}\end{array}}{m\left\{\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)+\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\right\}} \tag{33}$$

$$\sigma_{\mathbf{y},k}^{2\text{new}}=\frac{\begin{array}{c}\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left\{\left(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k}\right)^T\left(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k}\right)\right.\\ -2E^T(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}})\mathbf{C}_k^{\text{new}\,T}\left(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k}\right)\\ \left.+trace[\mathbf{C}_k^{\text{new}\,T}\mathbf{C}_k^{\text{new}}\left(E\left(\mathbf{t}_i\mathbf{t}_i^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]\right\}\end{array}}{r\left\{\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\right\}} \tag{34}$$

where $p(k)$ is the overall proportional value of the input dataset $\mathbf{X}=\{\mathbf{X}_1,\mathbf{X}_2\}$. Detailed derivation of the M-step for the mixture semisupervised model is provided in Appendix B. By updating and recalculating the E-step and the M-step of the EM algorithm until convergence, the optimal parameter set can be obtained.

Online Soft Sensing Based on Mixture Semisupervised PCR Model

Based on the developed mixture semisupervised PCR model, a soft sensor can be constructed for online prediction of key variables in the process. After we obtain the new data sample $\mathbf{x}_{new}$, the first step is to calculate its posterior probability in each local operating region, given as

$$p(k|\mathbf{x}_{\text{new}},\boldsymbol{\Theta})=\frac{p(\mathbf{x}_{\text{new}}|k,\boldsymbol{\Theta})p(k|\boldsymbol{\Theta})}{p(\mathbf{x}_{\text{new}}|\boldsymbol{\Theta})} \tag{35}$$

The latent variables in each local model $\mathbf{t}_{k,new}$ can be estimated as

$$\hat{\mathbf{t}}_{k,\text{new}}=\left(\sigma_{\mathbf{x},k}^2\mathbf{I}+\mathbf{P}_k^T\mathbf{P}_k\right)^{-1}\mathbf{P}_k^T(\mathbf{x}_{\text{new}}-\boldsymbol{\mu}_{\mathbf{x},k}) \tag{36}$$

Then the prediction in the corresponding local region can be calculated as

$$\hat{\mathbf{y}}_{k,\text{new}}=\mathbf{C}_k\hat{\mathbf{t}}_{k,\text{new}}=\mathbf{C}_k\left(\sigma_{\mathbf{x},k}^2\mathbf{I}+\mathbf{P}_k^T\mathbf{P}_k\right)^{-1}\mathbf{P}_k^T(\mathbf{x}_{\text{new}}-\boldsymbol{\mu}_{\mathbf{x},k}) \tag{37}$$

Under the mixture probabilistic model structure, the final prediction is provided as the following weighted form

$$\hat{\mathbf{y}}_{\text{new}}=\sum_{k=1}^{K} p(k|\mathbf{x}_{\text{new}},\boldsymbol{\Theta})\hat{\mathbf{y}}_{k,\text{new}} \tag{38}$$

with prediction error of the soft sensor given as

$$\mathbf{er}_{\text{new}}=\mathbf{y}_{\text{new}}-\hat{\mathbf{y}}_{\text{new}} \tag{39}$$

where $\mathbf{y}_{\text{new}}$ is the real value of the quality/key variables. To evaluate the performance of the soft sensor, the root mean square error (RMSE) criterion is typically used, which is defined as follows

$$RMSE=\sqrt{\frac{\sum_{j=1}^{L}\|\mathbf{y}_j-\hat{\mathbf{y}}_j\|^2}{L}} \tag{40}$$

where $j=1,2,\cdots,L$, $\mathbf{y}_j$ and $\hat{\mathbf{y}}_j$ are real and predicted values, respectively, and L is the total number of test data samples.

Case Studies

Numerical example

The numerical example consists of five variables in each operation mode, which is constructed as follows

$$\begin{aligned}\mathbf{x}_1&=\mathbf{P}_1\mathbf{t}_1+\mathbf{e}_1, \mathbf{y}_1=\mathbf{C}_1\mathbf{t}_1+\mathbf{f}_1\\ \mathbf{x}_2&=\mathbf{P}_2\mathbf{t}_2+\mathbf{e}_2, \mathbf{y}_2=\mathbf{C}_2\mathbf{t}_2+\mathbf{f}_2\\ \mathbf{x}_3&=\mathbf{P}_3\mathbf{t}_3+\mathbf{e}_3, \mathbf{y}_3=\mathbf{C}_3\mathbf{t}_3+\mathbf{f}_3\end{aligned} \tag{41}$$

where $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$ are three random 5×3 matrices, $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ are three random 1×3 matrices, $\{\mathbf{t}_1,\mathbf{t}_2,\mathbf{t}_3\}\sim N(0,\mathbf{I})$, $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}\sim N(0,0.05^2\mathbf{I})$, and $\{\mathbf{f}_1,\mathbf{f}_2,\mathbf{f}_3\}\sim N(0,0.05^2\mathbf{I})$ are measurement noises of input and output data. Here, we have assumed that the process data are formed by three different operation modes, and the noise levels in those three operation modes are assumed to be same.

To build the mixture semisupervised PCR model, 1000 data samples are generated in each operation modes, among which 100 samples are labeled, other 900 samples only contain the input measurements **x**. Therefore, a total of 3000 samples are available for model constructions, with 300 labeled samples and 2700 unlabeled samples. The prior probabilities of the three operation modes are 1/3 for both of the labeled and unlabeled datasets. The first two dimensions of the labeled dataset and the whole dataset are plotted in Figures 1a, b, respectively. When only labeled data are used, it can be seen in Figure 1a that the three different clusters are difficult to characterize. With the inclusion of the unlabeled dataset, we can clearly see three data clusters in Figure 1b. Under the Gaussian assumption, all of the three data clusters are distributed in an ellipse-like shape. For this multimode dataset, if we use a single Gaussian model, the modeling result will be too conservative, that is, a large ellipse encloses all data samples in different modes. Therefore, compared to the single Gaussian model, the dataset can be modeled more accurately by using the mixture model structure.

Based on both labeled and unlabeled datasets, the mixture semisupervised PCR model is developed. The estimated prior

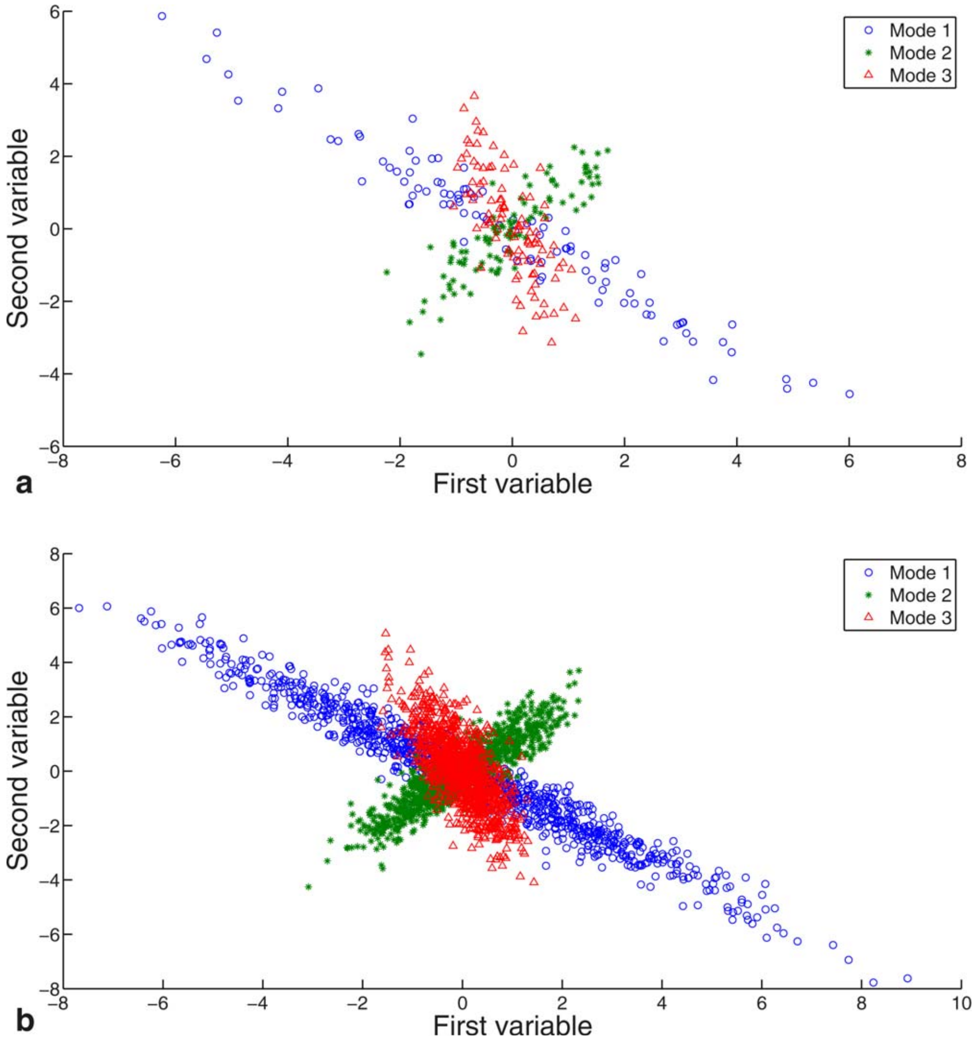

Figure 1. Data characteristics of labeled dataset and the whole dataset.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

probabilities of the three data clusters are 0.3331, 0.3335, and 0.3334 for the labeled dataset and 0.3333, 0.3332, and 0.3335 for the unlabeled dataset, which are very close to the real values. Detailed posterior probabilities of the labeled and unlabeled datasets under the three local regions are illustrated in Figures 2 and 3. It can be seen in both of the two figures that different data samples have been correctly assigned to their corresponding operation modes. The estimated noise variances are 0.0024, 0.0026, and 0.0025 for the input data and 0.0021, 0.0027, and 0.0022 for the output data, which are the averaged values based on 50 Monte-Carlo simulations. Compared to the output data, the estimation accuracy of noise variances for the input data is higher, and this is because more input data samples have been incorporated for modeling.

Besides, we can also evaluate the distribution of the extracted latent variables by the mixture semisupervised PCR model. For comparison, the supervised form of the mixture PCR model is also developed, which refers to modeling by ignoring the unsupervised data. With the inclusion of more unlabeled data samples, it is expected that the extracted latent variables are more likely to follow the Gaussian distribution than those extracted by the mixture supervised PCR model in which only a small number of labeled data samples have been used. Detailed probability distributions of the two extracted latent variables by both semisupervised and supervised models are shown in Figure 4, in which the first column corresponds to the semisupervised method and the second column corresponds to the supervised method. It can be clearly seen from this figure that the two components in the first column are more like Gaussian distribution than those in the second column. More precisely, if we carry out Jarque-Bera test which is a goodness-of-normality-fit test of a component, it can be found that the values of two extracted latent variables by the semisupervised method are closer to the critical cut-off value than those extracted by the supervised method. Therefore, compared to the supervised method, the extracted latent variables of the semisupervised method is more accurate.

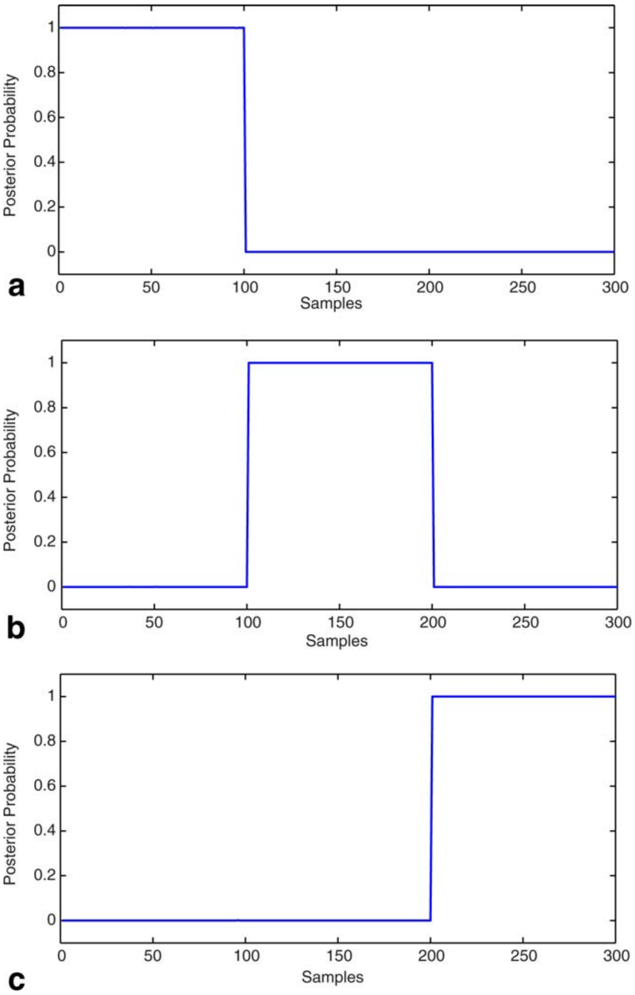

Figure 2. Posterior probabilities of the labeled data samples in different modes.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

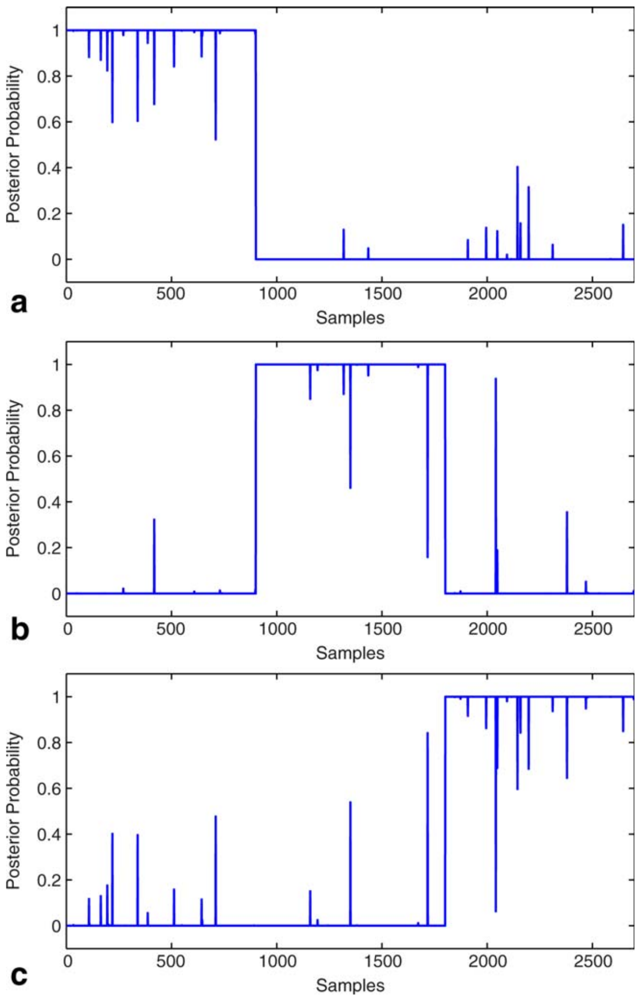

Figure 3. Posterior probabilities of the unlabeled data samples in different modes.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

To examine the modeling effort of the mixture semisupervised PCR model, the model training time is compared among the single semisupervised PCR model, mixture PCR model, and the mixture semisupervised PCR model. Under the same running environment: Windows 7; Matlab 7.5, Dual Core 3.2GHz, 8.0Gb RAM, the CPU running time of those three models are 31.6548s, 37.9557s, 113.0719s. It can

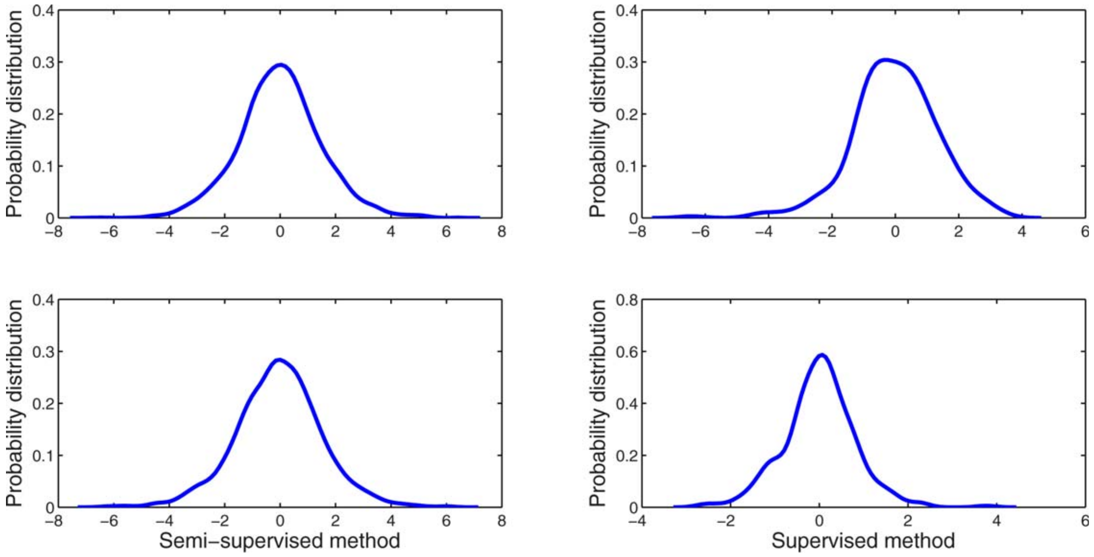

Figure 4. Probability distributions of two extracted latent variables by both semisupervised and supervised methods.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

be seen that the modeling effort of the mixture semisupervised PCR model is much harder than both single PCR and mixture PCR model. While the single PCR model incorporates a large number of additional unlabeled data samples, the mixture PCR model only uses the part of labeled data samples. That is why the running time of the single PCR model and the mixture PCR model is not quite different from each other.

Industrial application

Traditionally, the debutanizer column is a part of the desulfuring and naphtha splitter plant. In the naphtha stream, propane and butane are removed as overheads. To improve the control quality of the debutanizer column, real-time estimation of the butane content is important. A number of sensors are installed on the plant for product quality monitoring. The detailed description of the debutanizer column is shown in Figure 5, in which grey circles represent the used process variables in this case study for soft sensor development.[32]

For prediction of the butane content in this process, seven input variables have been selected, which are listed in Table 1. A total of 2000 data samples have been collected under the normal operating condition, which are provided by Fortuna et al.[33] The dataset is partitioned into two parts: the modeling dataset (1000 samples) and the testing dataset (1000 samples). For comparison, both of the mixture semisupervised PCR and mixture supervised PCR models are developed. In each individual model of the two mixture models, three latent variables are selected. To be fair, the number of individual models in both supervised and semisupervised mixture models is selected as 3. Different numbers of labeled datasets are used for examination of the soft sensing performance of the two mixture models, which are between 5% and 50% of the whole training dataset.

Detailed soft sensing results of the testing dataset are provided in Table 2. It can be seen that with the increase of the number of labeled data samples, both of the two methods can provide more and more accurate estimation results. However, with the use of unlabeled data samples, the mixture semisupervised PCR model-based soft sensor has obtained more accurate results than the mixture supervised PCR model-based soft sensor under the same number of labeled data samples. Therefore, compared to the total supervised regression model, the semisupervised regression model improves the soft sensing performance, especially when the ratio between the number of labeled data samples and the number of unlabeled data samples is low.

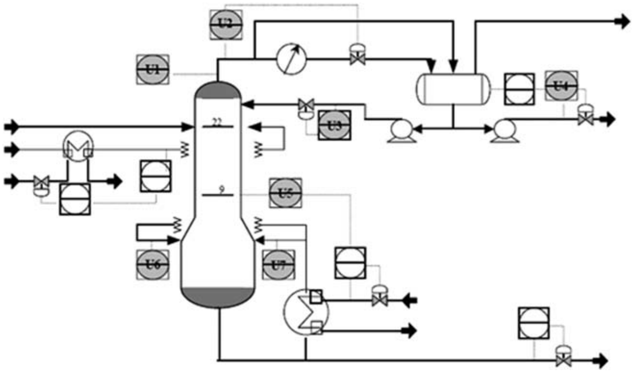

Figure 5. The flowchart of the debutanizer column.[32]

Table 1. Input Variables in the Debutanizer Column

Input variables	Description
u_1	Top temperature
u_2	Top pressure
u_3	Reflux flow
u_4	Flow to next process
u_5	6th tray temperature
u_6	Bottom temperature
u_7	Bottom temperature

Table 2. RMSE Values of the Two Methods Under Different Numbers of Labeled Samples

Methods/Portions	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
MSSPCR	0.2266	0.1857	0.1755	0.1736	0.1701	0.1641	0.1633	0.1605	0.1583	0.1548
MSPCR	0.2838	0.2414	0.1969	0.1885	0.1806	0.1723	0.1698	0.1671	0.1655	0.1641

Particularly, detailed soft sensing results of the 10% labeled dataset case are illustrated in Figure 6 for the two methods. The estimated values of the prior probabilities of the three individual models for the labeled and unlabeled datasets are tabulated in Table 3. Correspondingly, the estimated posterior probabilities of data samples in labeled and unlabeled datasets under different individual models are shown in Figures 7 and 8, respectively. It can be seen that there are only a small portion of data samples that belong to the third operation mode, which is consistent with the results of estimated prior probabilities for different individual models. The estimated noise variances of the input and output data for different individual models are given in Table 4. For the testing dataset, detailed information of the estimated posterior probabilities is provided in Figure 9, which is quite similar to those of the unlabeled data samples in the training dataset. Most of the first 500 data samples have been assigned to the first two operation modes, only a part of the last 500 data samples are assigned to the third operation mode. This is because the data characteristics of the training dataset and the testing dataset are similar to each other.

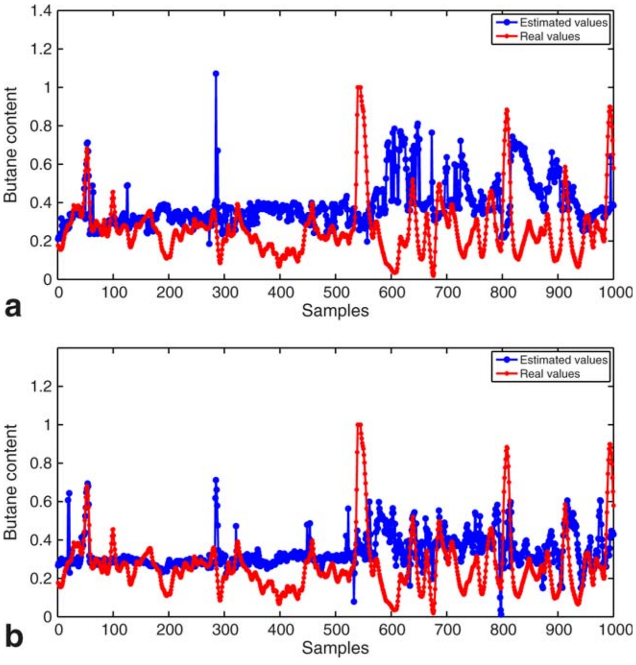

Figure 6. Quality estimation results.

(a) Mixture supervised PCR and (b) mixture semisupervised PCR. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 3. Prior Probabilities of Different Individual Models

Individual Models	#1	#2	#3
Prior probability of labeled dataset	0.5992	0.3287	0.0720
Prior probability of unlabeled dataset	0.3986	0.4876	0.1139

Conclusions

In this article, a mixture probabilistic semisupervised PCR model has been developed, based on which an efficient soft sensor development approach was constructed for online estimation of key variables in the process. Different from the single semisupervised PCR model, the mixture model was able to incorporate unlabeled data samples from different operating modes; thus it can be used in more general situation, such as multimode processes, nonlinear process, and so forth. With the inclusion of additional unlabeled data

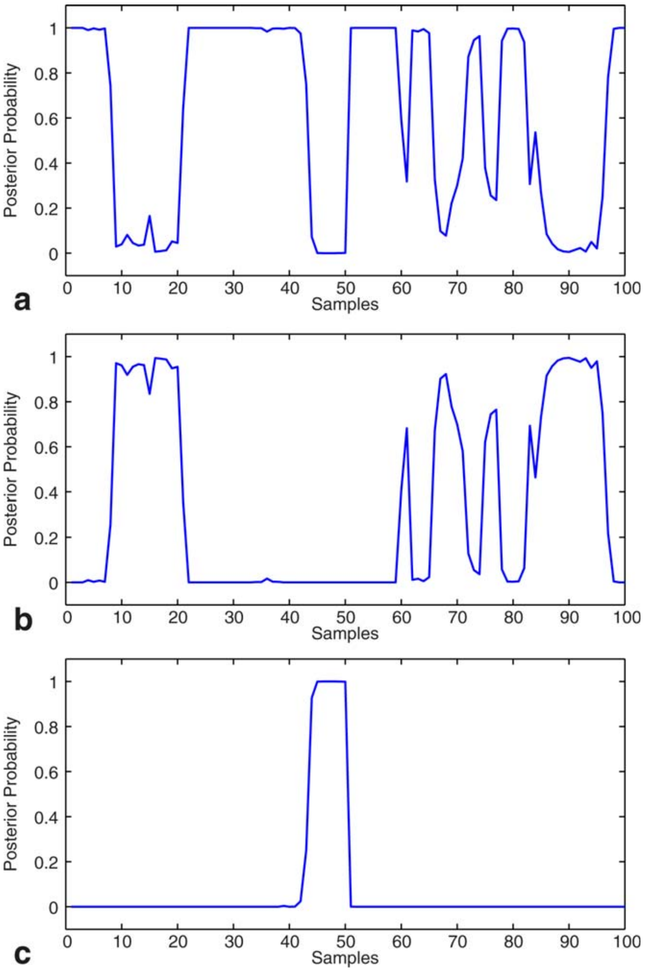

Figure 7. Posterior probability for the labeled data samples in the training dataset.

(a) First local region; (b) second local region; and (c) third local region. [Color figure can be viewed in the online issue, which is available at wileyonline library.com.]

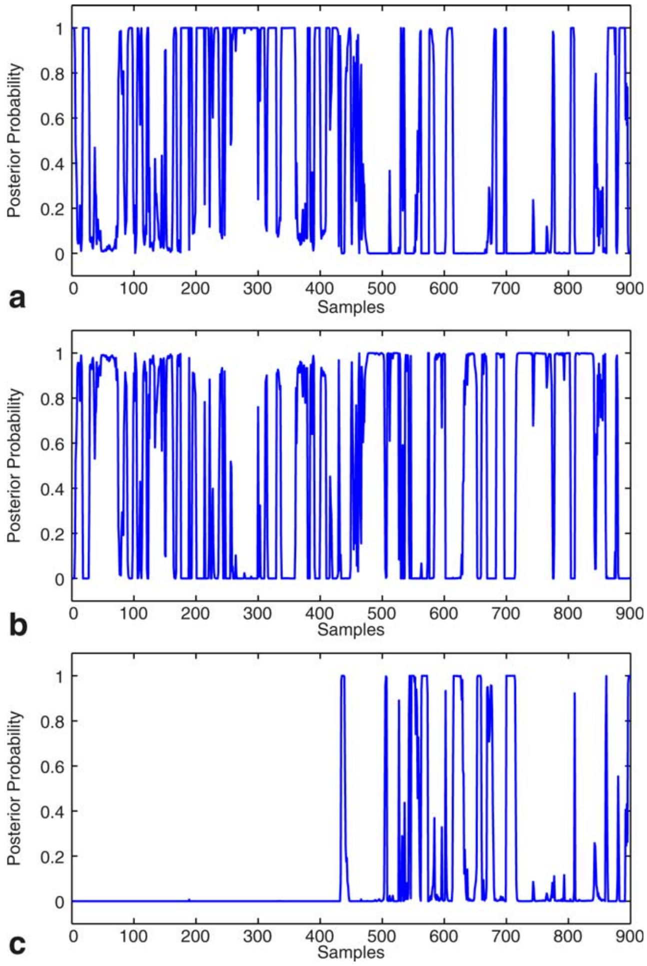

Figure 8. Posterior probability for the unlabeled data samples in the training dataset.

(a) First local region; (b) second local region; and (c) third local region. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

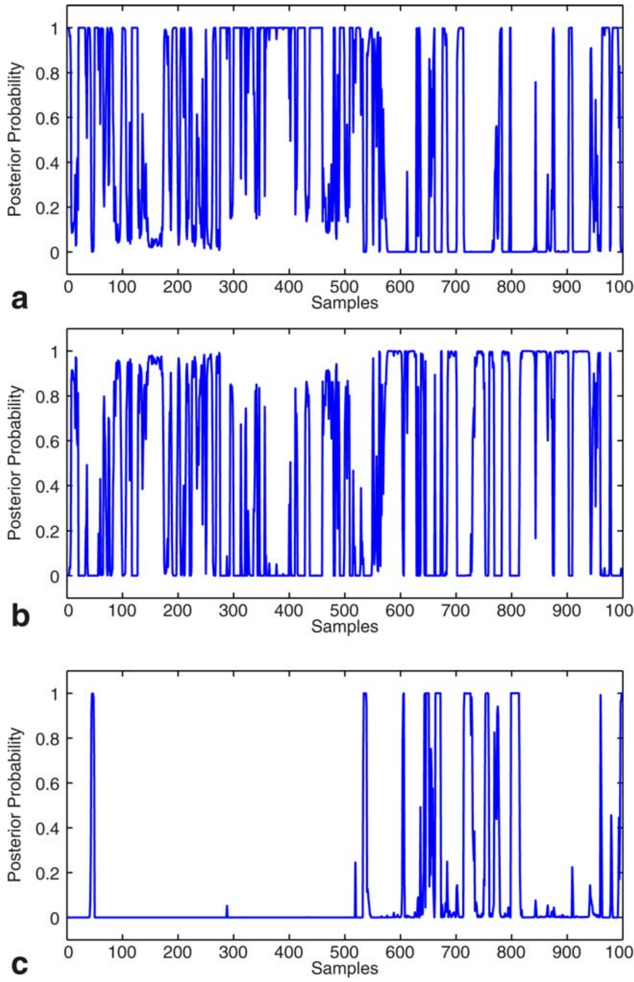

Figure 9. Posterior probability for the testing dataset.

(a) First local region; (b) second local region; and (c) third local region. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

samples, the modeling performance of the supervised model has been improved. Compared to the mixture supervised PCR model, the mixture semisupervised PCR model-based soft sensor can provide more accurate estimation results, particularly when the number of available labeled data samples is small in the process. Both of the numerical example and the industrial data case study have demonstrated the feasibility of the developed model. Although we have applied the traditional PCR method in the present article, the semisupervised modeling idea can be easily extended to other approaches, such as PLS models. Similar to the PCR method, PLS can also model the relationships between the secondary process variables and the quality variables by maximizing the covariance between these two types of variables.

Table 4. Estimated Noise Variances of the Input and Output Data

Individual Models	Input Data	Output Data
#1	0.0020	0.0098
#2	0.0006	0.0013
#3	0.0028	0.0001

Acknowledgments

This article was supported in part by the National Science Engineering Research Council of Canada (NSERC), Alberta Innovates Technology Futures (AITF), the National Natural Science Foundation of China (NSFC) (61370029), National Project 973 (2012CB720500), and the Fundamental Research Funds for the Central Universities (2013QNA5016).

Literature Cited

1. Keithley RB, Heien ML, Wightman RM. Multivariate concentration determination using principal component regression with residual analysis. *Trends Anal Chem.* 2009;28:1127–1136.
2. Ge ZQ, Gao FR, Song ZH. Mixture probabilistic PCR model for soft sensing of multimode processes. *Chemom Intell Lab Syst.* 2011;105:91–105.
3. Huang SM, Yang JF. Improved principal component regression for face recognition under illumination variations. *IEEE Signal Process Lett.* 2012;19:179–182.
4. Kruger U, Chen Q, Sandoz DJ. Multivariate statistical process monitors. *US Patent,* 2006;7:062417.
5. Zhang YW, Zhang Y. Complex process monitoring using modified partial least squares method of independent component regression. *Chemom Intell Lab Syst.* 2009;98:143–148.
6. Yu J. Multiway Gaussian mixture model based adaptive kernel partial least squares regression method for soft sensor estimation and

reliable quality prediction of nonlinear multiphase batch processes. *Ind Eng Chem Res.* 2012;51:13227–13237.

7. Galicia HJ, He QP, Wang J. Comparison of the performance of a reduced-order dynamic PLS soft sensor with different updating schemes for digester control. *Control Eng Pract.* 2012;20:747–760.
8. Ni WD, Tan SK, Ng WJ, Brown SD. Localized adaptive recursive partial least squares regression for dynamic system modeling. *Ind Eng Chem Res.* 2012;51:8025–8039.
9. Lee MW, Joung JY, Lee DS, Park JM, Woo SH. Application of a moving-window-adaptive neural network to the modeling of a full-scale anaerobic filter process. *Ind Eng Chem Res.* 2005;44:3973–3982.
10. Gonzaga JCB, Meleiro LAC, Kiang C, Filho RM. ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process. *Comput Chem Eng.* 2009;33:43–49.
11. Yan XF. Hybrid artificial neural network based on BP-PLSR and its application in development of soft sensors. *Chemom Intell Lab Syst.* 2010;103:152–159.
12. Bhattacharya S, Pal K, Pal SK. Multi-sensor based prediction of metal deposition in pulsed gas metal arc welding using various soft computing models. *Appl Soft Comput.* 2012;12:498–505.
13. Yoo CK. Nonlinear monitoring and prediction model in an industrial environmental process. *J Chem Eng Jpn.* 2008;41:32–42.
14. Liu Y, Hu NP, Wang HQ, Li P. Soft chemical analyzer development using adaptive least-squares support vector regression with selective pruning and variable moving window size. *Ind Eng Chem Res.* 2009; 48:5731–5741.
15. Ge ZQ, Song ZH. Nonlinear soft sensor development based on relevance vector machine. *Ind Eng Chem Res.* 2010;49:8685–8693.
16. Kaneko H, Funatsu K. Development of soft sensor models based on time difference of process variables with accounting for nonlinear relationship. *Ind Eng Chem Res.* 2011;50:10643–10651.
17. Yu J. Online quality prediction of nonlinear and non-Gaussian chemical processes with shifting dynamics using finite mixture model based Gaussian process regression approach. *Chem Eng Sci.* 2012; 82:22–30.
18. Wibowo A, Desa MI. Kernel based regression and genetic algorithms for estimating cutting conditions of surface roughness in end milling machining process. *Exp Syst Appl.* 2012;39:11634–11641.
19. Khatibisepehr S, Huang B. Dealing with irregular data in soft sensor: Bayesian method and comparative study. *Ind Eng Chem Res.* 2008; 47:8713–8723.
20. Jin X, Wang SY, Huang B, Forbes F. Multiple model based LPV soft sensor development with irregular/missing process output measurement. *Control Eng Pract.* 2012;20:165–172.
21. Yu J. A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. *Comput Chem Eng.* 2012;41:134–144.
22. Khatibisepehr S, Huang B, Xu FW, Espejo A. A Bayesian approach to design of adaptive multi-model inferential sensors with application in oil sand industry. *J Process Control.* 2012;22:1913–1929.
23. Ge ZQ, Song ZH. Semisupervised Bayesian method for soft sensor modeling with unlabeled data samples. *AIChE J.* 2011;57:2109–2119.
24. Ge ZQ, Song ZH. Online monitoring of nonlinear multiple mode processes based on adaptive local model approach. *Control Eng Pract.* 2008;16:1427–1437.
25. Ge ZQ, Yang CJ, Song ZH, Wang HQ. Robust online monitoring for multimode processes based on nonlinear external analysis. *Ind Eng Chem Res.* 2008;47:4775–4783.
26. Yu J, Qin SJ. Multimode process monitoring with Bayesian inference based finite Gaussian mixture models. *AIChE J.* 2008;54:1811–1829.
27. Ge ZQ, Song ZH. Multimode process monitoring based on Bayesian method. *J Chemom.* 2009;23:636–650.
28. Ge ZQ, Song ZH. Mixture Bayesian regularization method of PPCA for multimode process monitoring. *AIChE J.* 2010;56:2838–2849.
29. Feital T, Kruger U, Dutra J, Pinto JC, Lima EL. Modeling and performance monitoring of multivariate multimodal processes. *AIChE J.* 2013;59:1557–1569.
30. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B.* 1977;39: 1–39.
31. Yu SP, Yu K, Tresp V, Kriege HP, Wu MR. Supervised probabilistic principal component analysis. In: *12th ACM International Conference on Knowledge Discovery and Data Mining.* 2006:464–473.
32. Fortuna L, Graziani S, Xibilia MG. Soft sensors for product quality monitoring in debutanizer distillation column. *Control Eng Pract.* 2005;13:499–508.
33. Fortuna L, Graziani S, Rizzo A, Xibilia MG. *Soft Sensors for Monitoring and Control of Industrial Processes.* London: Springer, 2007.

Appendix A

According to likelihood function

$$\begin{aligned}L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})=&\sum_{i=1}^{n_1}\ln\sum_{k=1}^{K}p\left(\mathbf{x}_i,\mathbf{y}_i|k,\boldsymbol{\Theta}\right)p_1(k)\\&+\sum_{i=n_1+1}^{n}\ln\sum_{k=1}^{K}p(\mathbf{x}_i|k,\boldsymbol{\Theta})p_2(k)\\=&\sum_{i=1}^{n_1}\ln\sum_{k=1}^{K}\int p\left(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_k,k|\boldsymbol{\Theta}\right)d\mathbf{t}_k\\&+\sum_{i=n_1+1}^{n}\ln\sum_{k=1}^{K}\int p(\mathbf{x}_i,\mathbf{t}_k,k|\boldsymbol{\Theta})d\mathbf{t}_k\end{aligned}\tag{A1}$$

Based on Jensen's inequality: $f\left(\sum_{k=1}^{K}\alpha_k\mathbf{x}_k\right)\geq\sum_{k=1}^{K}\alpha_k f(\mathbf{x}_k)$ where $f(\bullet)$ is an arbitrary convex function, and α_k is the weight, the likelihood function can be written as follows

$$\begin{aligned}L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})=&\sum_{i=1}^{n_1}\ln\sum_{k=1}^{K}\int p\left(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_k,k|\boldsymbol{\Theta}\right)d\mathbf{t}_k\\&+\sum_{i=n_1+1}^{n}\ln\sum_{k=1}^{K}\int p(\mathbf{x}_i,\mathbf{t}_k,k|\boldsymbol{\Theta})d\mathbf{t}_k\\=&\sum_{i=1}^{n_1}\ln\sum_{k=1}^{K}\int p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\mathrm{old}}\right)\frac{p\left(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_k,k|\boldsymbol{\Theta}\right)}{p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{old}\right)}d\mathbf{t}_k\\&+\sum_{i=n_1+1}^{n}\ln\sum_{k=1}^{K}\int p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{\mathrm{old}})\frac{p(\mathbf{x}_i,\mathbf{t}_k,k|\boldsymbol{\Theta})}{p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{old})}d\mathbf{t}_k\\\geq&\sum_{i=1}^{n_1}\sum_{k=1}^{K}\int p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\mathrm{old}}\right)\ln\frac{p\left(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_k,k|\boldsymbol{\Theta}\right)}{p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{old}\right)}d\mathbf{t}_k\\&+\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}\int p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{\mathrm{old}})\ln\frac{p(\mathbf{x}_i,\mathbf{t}_k,k|\boldsymbol{\Theta})}{p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{old})}d\mathbf{t}_k\\=&\left\{\sum_{i=1}^{n_1}\sum_{k=1}^{K}\int p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\mathrm{old}}\right)\ln p\left(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_k,k|\boldsymbol{\Theta}\right)d\mathbf{t}_k\right.\\&\left.+\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}\int p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{\mathrm{old}})\ln p(\mathbf{x}_i,\mathbf{t}_k,k|\boldsymbol{\Theta})d\mathbf{t}_k\right\}\\&-\left\{\sum_{i=1}^{n_1}\sum_{k=1}^{K}\int p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{old}\right)\right.\\&\ln p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{old}\right)d\mathbf{t}_k\\&\left.+\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}\int p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{old})\ln p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{old})d\mathbf{t}_k\right\}\end{aligned}\tag{A2}$$

where

$$\begin{aligned}C=&\sum_{i=1}^{n_1}\sum_{k=1}^{K}\int p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\mathrm{old}}\right)\ln p\left(\mathbf{t}_k,k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\mathrm{old}}\right)d\mathbf{t}_k\\&+\sum_{i=n_1+1}^{n}\sum_{k=1}^{K}\int p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{\mathrm{old}})\ln p(\mathbf{t}_k,k|\mathbf{x}_i,\boldsymbol{\Theta}_{\mathrm{old}})d\mathbf{t}_k\end{aligned}$$

is a constant term which is only associated with the old model parameter $\boldsymbol{\Theta}_{\mathrm{old}}$. As a result, the rest two terms in Eq. (A2) correspond to the expectation of the complete-data log likelihood function, with respect to the joint distribution of the hidden variables $p(\mathbf{t}_k, k|\mathbf{x}_i, \boldsymbol{\Theta}_{\text{old}})$. Therefore, the expected complete-data log likelihood is actually a lower bound of the original log likelihood. In the EM algorithm, due to the computational simplicity, the expected complete-data log likelihood is alternatively maximized instead of the original log likelihood.

Appendix B

In the M-step, by maximizing the expected Log likelihood function of the complete-data with respect to each one of the parameter set, including $p_1(k)$, $p_2(k)$, $\mathbf{P}_k, \mathbf{C}_k, \sigma^2_{\mathbf{x},k}, \sigma^2_{\mathbf{y},k}, \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\mu}_{\mathbf{y},k}$, the parameter values can be updated for calculation of the next E-step. First, revisit the expected value of the complete-data Log likelihood function, and rewrite it as follows

$$\begin{aligned} E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})] = \sum_{i=1}^{n_1}\sum_{k=1}^{K} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\{\ln p_1(k) \\ + \int p\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\ln\left[p\left(\mathbf{x}_i,\mathbf{y}_i,\mathbf{t}_{i,k}|k,\boldsymbol{\Theta}\right)\right]d\mathbf{t}_{i,k}\Big\} \\ + \sum_{i=n_1+1}^{n}\sum_{k=1}^{K} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\{\ln p_2(k) \\ + \int p\left(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\ln\left[p\left(\mathbf{x}_i,\mathbf{t}_{i,k}|k,\boldsymbol{\Theta}\right)\right]d\mathbf{t}_{i,k}\Big\} \end{aligned} \tag{B1}$$

In the above equation, the related terms with the proportional value $p_1(k)$ and $p_2(k)$ can be separated as follows

$$f_1(k) = \sum_{i=1}^{n_1}\sum_{k=1}^{K} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\ln p_1(k) \tag{B2}$$

$$f_2(k) = \sum_{i=n_1+1}^{n}\sum_{k=1}^{K} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\ln p_2(k) \tag{B3}$$

Introducing a Lagrange multiplier λ into each of $f_1(k)$ and $f_2(k)$, and noted the proportion value follows the constraint $\sum_{k=1}^{K} p_1(k)=1$ and $\sum_{k=1}^{K} p_2(k)=1$, the updated value of $p_1(k|\Theta)$ and $p_2(k|\Theta)$ can be obtained by maximizing

$$g_1(k) = f_1(k) + \lambda_1\left(\sum_{k=1}^{K} p_1(k) - 1\right) \tag{B4}$$

$$g_2(k) = f_2(k) + \lambda_2\left(\sum_{k=1}^{K} p_2(k) - 1\right) \tag{B5}$$

Setting their derivatives to zero with respect to $p_1(k)$ and $p_2(k)$, $k=1,2,\cdots,K$ we can get

$$\begin{aligned} \sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right) + \lambda_1 p_1(k) = 0 \\ \Rightarrow p_1(k) = -\frac{\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)}{\lambda_1} \end{aligned} \tag{B6}$$

$$\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\Theta_{\text{old}}) + \lambda_2 p_2(k) = 0 \Rightarrow p_2(k) = -\frac{\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})}{\lambda_2} \tag{B7}$$

Summing both sides over k, the above equations become

$$\left.\begin{aligned} \sum_{k=1}^{K} p_1(k) = -\frac{\sum_{i=1}^{n_1}\sum_{k=1}^{K} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)}{\lambda} \\ \sum_{k=1}^{K} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right) = 1 \\ \sum_{k=1}^{K} p_1(k) = 1 \end{aligned}\right\} \Rightarrow \lambda_1 = -n_1 \tag{B8}$$

$$\left.\begin{aligned} \sum_{k=1}^{K} p_2(k) = -\frac{\sum_{i=n_1+1}^{n}\sum_{k=1}^{K} p(k|\mathbf{x}_i,\Theta_{\text{old}})}{\lambda_2} \\ \sum_{k=1}^{K} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right) = 1 \\ \sum_{k=1}^{K} p_2(k) = 1 \end{aligned}\right\} \Rightarrow \lambda_2 = -(n-n_1) = -n_2 \tag{B9}$$

Substitute them into Eqs. (B6) and (B7), the updated values of the proportion become as

$$p_1(k) = \frac{1}{n_1}\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right) \tag{B10}$$

$$p_2(k) = \frac{1}{n_2}\sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}}) \tag{B11}$$

Based on the this derivation, the overall proportional value of the input dataset $\mathbf{X}=\{\mathbf{X}_1,\mathbf{X}_2\}$ can be determined as

$$p(k) = \frac{1}{n}\left\{\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right) + \sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\right\} \tag{B12}$$

On the other hand, the updated values of $\mathbf{P}_k^{\text{new}}$, $\mathbf{C}_k^{\text{new}}$, $\sigma^{2\text{new}}_{\mathbf{x},k}$, $\sigma^{2\text{new}}_{\mathbf{y},k}$, $\boldsymbol{\mu}^{\text{new}}_{\mathbf{x},k}$ and $\boldsymbol{\mu}^{\text{new}}_{\mathbf{y},k}$ can also be determined by maximizing $E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]$. Therefore, setting the derivative of $E(L_2)$ respect to $\mathbf{P}_k$ to zero $\frac{\partial E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]}{\partial \mathbf{P}_k}=0$, the value of $\mathbf{P}_k^{\text{new}}$ can be calculated as follows

$$\begin{aligned} &\frac{\partial E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]}{\partial \mathbf{P}_k} = 0 \Rightarrow \\ &\left\{\sum_{i=1}^{n_1}\left[p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left(\mathbf{x}_i - \mathbf{P}_k E\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i\right) - \boldsymbol{\mu}_{\mathbf{x},k}\right)\right.\right. \\ &\left.\left. E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i\right)\right]\right\} + \left\{\sum_{i=n_1+1}^{n}\left[p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\left(\mathbf{x}_i - \mathbf{P}_k E\left(\mathbf{t}_{i,k}|\mathbf{x}_i\right)\right.\right.\right. \\ &\left.\left.\left. - \boldsymbol{\mu}_{\mathbf{x},k}\right)E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i\right)\right]\right\} = 0 \\ &\Rightarrow \mathbf{P}_k^{new} = \left[\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x},k}\right)E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right. \\ &\quad\left. + \sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})\left(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x},k}\right)E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right] \\ &\quad\times\left[\sum_{i=1}^{n_1} p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right. \\ &\quad\left. + \sum_{i=n_1+1}^{n} p(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}})E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]^{-1} \end{aligned} \tag{B13}$$

Similarly, by setting the derivative of $E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]$ with respect to other parameters to zero, their updated values can be calculated as

$$
\begin{aligned}
&\frac{\partial E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]}{\partial \mathbf{C}_k}=0 \Rightarrow \\
&\left\{\sum_{i=1}^{n_1}\left[p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left(\mathbf{y}_i-\mathbf{C}_k E\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i\right)-\boldsymbol{\mu}_{\mathbf{y},k}\right)\right.\right. \\
&\quad \left.\left. E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i\right)\right]\right\}=0 \\
&\Rightarrow \mathbf{C}_k^{\text{new}}=\left[\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k}\right)E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\right] \\
&\quad \times\left[\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]^{-1}
\end{aligned}
\tag{B14}
$$

$$
\begin{aligned}
&\frac{\partial E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]}{\partial \boldsymbol{\mu}_{\mathbf{x},k}}=0 \Rightarrow \\
&\boldsymbol{\mu}_{\mathbf{x},k}^{\text{new}}=\frac{\displaystyle\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right)\left[\mathbf{x}_i-\mathbf{P}_k E\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]+\sum_{i=n_1+1}^{n}p\left(k|\mathbf{x}_i,\Theta_{\text{old}}\right)\left[\mathbf{x}_i-\mathbf{P}_k E\left(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]}{\displaystyle\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right)+\sum_{i=n_1+1}^{n}p\left(k|\mathbf{x}_i,\Theta_{\text{old}}\right)}
\end{aligned}
\tag{B15}
$$

$$
\frac{\partial E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]}{\partial \boldsymbol{\mu}_{\mathbf{y},k}}=0 \Rightarrow \boldsymbol{\mu}_{\mathbf{y},k}^{\text{new}}=\frac{\displaystyle\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right)\left[\mathbf{y}_i-\mathbf{C}_k E\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]}{\displaystyle\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\Theta_{\text{old}}\right)}
\tag{B16}
$$

$$
\begin{aligned}
&\frac{\partial E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]}{\partial \sigma_{\mathbf{x},k}^2}=0 \Rightarrow \\
&\sigma_{\mathbf{x},k}^{2\text{new}}=\frac{\begin{aligned}&\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left\{\left(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k}\right)^T\left(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k}\right)-2E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\mathbf{P}_k^{\text{new}\,T}\left(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k}\right)\right.\\&\left.+trace\left[\mathbf{P}_k^{\text{new}\,T}\mathbf{P}_k^{\text{new}}E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]\right\}+\sum_{i=n_1+1}^{n}p\left(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left\{\left(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k}\right)^T\left(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k}\right)\right.\\&\left.-2E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i,k,\Theta_{\text{old}}\right)\mathbf{P}_k^{\text{new}\,T}\left(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x},k}\right)+trace\left[\mathbf{P}_k^{\text{new}\,T}\mathbf{P}_k^{\text{new}}E\left(\mathbf{t}_{i,k}\mathbf{t}_{i,k}^T|\mathbf{x}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]\right\}\end{aligned}}{m\left\{\displaystyle\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)+\sum_{i=n_1+1}^{n}p\left(k|\mathbf{x}_i,\boldsymbol{\Theta}_{\text{old}}\right)\right\}}
\end{aligned}
\tag{B17}
$$

$$
\begin{aligned}
&\frac{\partial E[L(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})]}{\partial \sigma_{\mathbf{y},k}^2}=0 \Rightarrow \\
&\sigma_{\mathbf{y},k}^{2\text{new}}=\frac{\begin{aligned}&\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\left\{\left(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k}\right)^T\left(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k}\right)\right.\\&-2E^T\left(\mathbf{t}_{i,k}|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\mathbf{C}_k^{\text{new}\,T}\left(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y},k}\right)\\&\left.+trace\left[\mathbf{C}_k^{\text{new}\,T}\mathbf{C}_k^{\text{new}}\left(E\left(\mathbf{t}_i\mathbf{t}_i^T|\mathbf{x}_i,\mathbf{y}_i,k,\boldsymbol{\Theta}_{\text{old}}\right)\right]\right\}\end{aligned}}{r\left\{\displaystyle\sum_{i=1}^{n_1}p\left(k|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\Theta}_{\text{old}}\right)\right\}}
\end{aligned}
\tag{B18}
$$

Manuscript received May 19, 2013, and revision received Oct. 5, 2013.