Semisupervised Bayesian Method for Soft Sensor Modeling with Unlabeled Data Samples

Zhiqiang Ge and Zhihuan Song
State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Zhejiang University, Hangzhou 310027, Zhejiang, P.R. China

DOI 10.1002/aic.12422
Published online October 20, 2010 in Wiley Online Library (wileyonlinelibrary.com).

Most traditional soft sensors are built upon the labeled dataset that contains equal numbers of input and output data samples. However, the output variables that correspond to quality variables and other important controlled variables are always difficult to obtain in chemical processes. Therefore, we may only obtain the output data for a small portion of the whole dataset and have much more input data samples. In this article, a semisupervised method is proposed for soft sensor modeling, which can successfully incorporate the unlabeled data information. To determine the effective dimensionality of the latent space, the Bayesian regularization method is introduced into the semisupervised model structure. Two industrial application case studies are provided to evaluate the feasibility and efficiency of the newly developed probabilistic soft sensor. © 2010 American Institute of Chemical Engineers *AIChE J,* 57: 2109–2119, 2011
Keywords: semisupervised learning, Bayesian regularization, soft sensor modeling, probabilistic

Introduction

In chemical processes, soft sensors have been widely used for estimating product or other important controlled variables. Traditionally, soft sensors are expressed as first principle models, which are based on physical principles such as mass, components, and energy balances of the process. However, they often need complete knowledge of the process, which is difficult to obtain from complex chemical processes. On the other hand, data-based method for soft sensor development has become very popular in recent years. This is because a huge number of process data have been recorded by the distributed control system, which has been widely installed in modern industrial processes.

Along last several decades, data-based soft sensors have been intensively researched both in industry and in academy areas. Among all developed soft sensor models, conventionally used ones include: principal component analysis (PCA)/principal component regression (PCR),[1–3] partial least squares (PLS),[4–8] artificial neural networks,[9–12] and support vector machine.[13–20] According to the recent review on data-driven soft sensor in process industry, the most acceptable technique is PCA/PCR.[21] In our opinion, the main reason is PCA/PCR is very simple to implement in practical application. When the soft sensor acts as a functional module in the process control system, the integration between different parts of the system will become much easier when the simple PCA/PCR method is used as the soft sensor. To our knowledge, up to now, PCA/PCR is still a hot research spot both in the chemical engineering and other related areas.[22–24]

Generally, data-based soft sensors always need complete data samples (both input and output variables) for modeling. In this article, we denote the dataset that contains both input and output data samples as the labeled dataset and denote the one that only consists of input data samples as the unlabeled dataset. Therefore, most traditional data-based soft sensors are modeled upon the labeled dataset. However, as we know, the output variables that correspond to quality variables and other important controlled variables are always

Correspondence concerning this article should be addressed to Z. Ge at zqge@iipc.zju.edu.cn or Z. Song at zhsong@iipc.zju.edu.cn.

difficult to measure. To this end, we may only obtain the output data for a small portion of the whole dataset and have much more input data samples. It is straightforward that we can just use the labeled dataset for soft sensor development and ignore the information of the unlabeled dataset. However, as the data information is not sufficiently used, the performance of the developed soft sensor may not be guaranteed, especially when the portion of the labeled dataset is very small.

Model training with both labeled and unlabeled data samples is termed as semisupervised learning in the statistical machine learning area, which has caught much attention in recent years.[25–37] If we treat the unlabeled data samples as missing data in the probabilistic model, several missing data treatment methods can be incorporated to address the unlabeled data samples, such as the expectation-maximization (EM) algorithm.[25,27] By incorporating the unlabeled data information into the supervised model or directly training the model from both labeled and unlabeled datasets, semisupervised models can always perform better than existing methods. Recent surveys on the semisupervised learning method have been given by Zhu and Chapelle et al.[30,31] in which different semisupervised learning methods have been introduced. Traditional semisupervised learning methods include self-training-based methods, probabilistic generative model-based methods, cotraining methods, graph-based methods, and so on. In this article, however, we only focus on the probabilistic generative model-based semisupervised method. Corresponding to the PCA method, PCR can be considered as its supervised form because the output data information has been incorporated. Recently, a semisupervised form of PCA has been proposed in the machine learning area, depending on which the modeling performance has been greatly improved.[38] However, the method cannot determine the dimensionality of the latent space automatically, which is very important in latent model structure such as PCR. To our best knowledge, the semisupervised form of PCR has not yet been reported in the chemical engineering area, especially for soft sensor modeling.

In this article, we intend to develop the soft sensor based on both labeled and unlabeled datasets. Because of the wide utilization of the PCR method for soft sensor development, a semisupervised learning strategy is incorporated into this model, based on which all available data information can be used. Different from the traditional PCR method, the new semisupervised approach constructs the model through a probabilistic manner and, thus, can model the process noise information simultaneously. However, like other probabilistic statistic methods such as probabilistic principal component analysis and factor analysis, an important issue is how to determine the dimensionality of latent variable space. As the probabilistic model structure itself does not provide any mechanism to determine this important value, other methods should be used. In our opinion, the determination of the latent space dimensionality can be considered as a model complexity problem, which can be treated by the Bayesian method.[39–41] Different from existing dimensionality determination methods such as cross validation and try-and-error, the Bayesian method can automatically select the dimensionality through the modeling process and does not need any additional validation dataset.

Hence, a semisupervised Bayesian PCR (SBPCR) model is proposed. Contributions of this article can be summarized as: (1) A new semisupervised PCR model is proposed for soft sensor development under both labeled and unlabeled datasets; (2) based on the semisupervised PCR model, a further Bayesian regularization step is provided for dimensionality selection of the latent variable space; (3) the developed semisupervised Bayesian method is used for probabilistic soft sensor modeling under both labeled and unlabeled data samples. The rest of this article is organized as follows: In Section "Principal Component Regression," the traditional PCR method is briefly introduced, followed by the detailed description of the proposed SBPCR model for soft sensor development. In Section "Case Studies," two industrial application case studies are provided to evaluate the performance of the proposed method. Finally, some conclusions are made.

Principal Component Regression

Suppose the measurement matrix can be represented as $\mathbf{X} \in R^{n\times m}$, where n is the number of data sample and m is the number of measured variables. The predicted variable matrix can be given as $\mathbf{Y} \in R^{n\times r}$, where n is the number of data sample and r is the number of predicted variables. In this article, for simplicity, it is assumed that the mean values of all data samples have been removed. The aim of PCR is to find a set of principal components that 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 measurement matrix $\mathbf{X}$, and the second step is to calculate the regression matrix between the extracted principal components and the predicted variable matrix $\mathbf{Y}$. The traditional derivation of PCR can be expressed 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 $\mathbf{P} \in R^{m\times k}$ is the loading matrix, $\mathbf{T} \in R^{n\times k}$ is the principal component matrix, k is the selected number of principal components, $\mathbf{C} \in R^{r\times k}$ is the regression matrix, and $\mathbf{E}$ and $\mathbf{F}$ are the residuals matrices with appropriate dimensions. For a new data sample $\mathbf{x}_{\text{new}} \in R^{m\times l}$, the predicted variables can be calculated as follows:

$$\hat{\mathbf{y}}_{\text{new}} = \mathbf{C}\mathbf{P}^T\mathbf{x}_{\text{new}}. \tag{3}$$

SBPCR Model Development

In this section, the proposed semisupervised PCR model is first developed through a probabilistic way, which is followed by its Bayesian regularization method. Then, a new soft sensor will be built based on the SBPCR model.

Semisupervised PCR model

Although the traditional PCR method develops the model through a data projection manner, the semisupervised PCR method intends to construct the model through the following generative model structure

$$\mathbf{x}_i = \mathbf{P}\mathbf{t}_i + \mathbf{e}_i$$
$$\mathbf{y}_j = \mathbf{C}\mathbf{t}_j + \mathbf{f}_j, \tag{4}$$

where $i = 1,2,\ldots,n$, $j = 1,2,\ldots,n_1$, n_1 is the size of the labeled dataset, and $n_2 = n - n_1$ is the size of the unlabeled dataset. $\mathbf{P} \in R^{m\times k}$ and $\mathbf{C} \in R^{r\times k}$ are the corresponding loading matrix and regression matrix, respectively, where m is the number of input variables and, r is the number of output variables. $\mathbf{t} \in R^{k\times l}$ is the principal component vector, $\mathbf{e} \in R^{m\times l}$ and $\mathbf{f} \in R^{r\times l}$ are process noises of the input and output variables, respectively. In the probabilistic PCR model structure, it is assumed that both probability density functions of the principal component and the process noise are Gaussian, thus $p(\mathbf{t}) = N(0,\mathbf{I})$, $p(\mathbf{e}) = N(0,\sigma_{\mathrm{x}}^2\mathbf{I})$, and $p(\mathbf{f}) = N(0,\sigma_{\mathrm{y}}^2\mathbf{I})$, where $\mathbf{I}$ is an identity matrix, and σ_{x}^2 and σ_{y}^2 are noise variances of the input and output variables, respectively.

First, the labeled and unlabeled datasets can be denoted as $\mathbf{X}_1 = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}]^T \in R^{n_1\times m}$, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_1}]^T \in R^{n_1\times r}$, and $\mathbf{X}_2 = [\mathbf{x}_{n_1+1}, \mathbf{x}_{n_1+2}, \ldots, \mathbf{x}_{n_1+n_2}]^T \in R^{n_2\times m}$, then the marginal distribution of the labeled and unlabeled data samples can be calculated as follows:

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

$$p(\mathbf{x}_{n_1+i}, |\mathbf{P}, \sigma_{\mathbf{x}}^2) = \int p(\mathbf{x}_{n_1+i}|\mathbf{t}_{n_1+i}, \mathbf{P}, \sigma_{\mathbf{x}}^2)p(\mathbf{t}_{n_1+i})d\mathbf{t}_{n_1+i}, \tag{6}$$

where $j = 1,2,\ldots,n_1$ and $i = 1,2,\ldots,n_2$. It is noted that we have used 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.[38] Following the maximum likelihood framework, the likelihood function can be calculated as

$$p(\mathbf{X}, \mathbf{Y}) = p(\mathbf{X}_1, \mathbf{Y})p(\mathbf{X}_2). \tag{7}$$

For simplicity, the maximum likelihood function can be transformed to the log form. Notice Eqs. 5 and 6, the log-likelihood function can be formulated as

$$L(\mathbf{X}, \mathbf{Y}) = L(\mathbf{X}_1, \mathbf{Y}) + L(\mathbf{X}_2) = \ln\prod_{j=1}^{n_1} p(\mathbf{x}_j, \mathbf{y}_j|\mathbf{P}, \mathbf{C}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2) + \ln\prod_{i=1}^{n_2} p(\mathbf{x}_{n_1+i}|\mathbf{P}, \sigma_{\mathbf{x}}^2). \tag{8}$$

Through maximizing the log-likelihood function, the parameter set of the semisupervised PCR model $\Theta = \{\mathbf{P},\mathbf{C},\sigma_{\mathrm{x}}^2,\sigma_{\mathrm{y}}^2\}$ can be determined. To maximize this log-likelihood function, the well-known EM algorithm can be used, which can guarantee that the log likelihood never decreases when this algorithm is carried out iteratively. In the E-step of the EM algorithm, we are given the parameters Θ_{old} obtained in the previous M-step; the aim of this step is to determine the sufficient statistics of the principal components. In the M-step, it is assumed that the sufficient statistics of the principal components have been obtained, we try to calculate the new parameter set Θ_{new} by maximizing the log-likelihood function.

In the E-step, the posteriori distribution of the principal components under the labeled and unlabeled datasets can be calculated separately, which are given in Eqs. 9 and 10, respectively.

$$p(\mathbf{t}_j|\mathbf{x}_j, \mathbf{y}_j, \mathbf{P}, \mathbf{C}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2) = \frac{p(\mathbf{x}_j|\mathbf{t}_j, \mathbf{P}, \sigma_{\mathbf{x}}^2)p(\mathbf{y}_j|\mathbf{t}_j, \mathbf{C}, \sigma_{\mathbf{y}}^2)p(\mathbf{t}_j)}{p(\mathbf{x}_j, \mathbf{y}_j, \mathbf{P}, \mathbf{C}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{y}}^2)} \tag{9}$$

$$p(\mathbf{t}_{n_1+i}|\mathbf{x}_{n_1+i}, \mathbf{P}, \sigma_{\mathbf{x}}^2) = \frac{p(\mathbf{x}_{n_1+i}|\mathbf{t}_{n_1+i}, \mathbf{P}, \sigma_{\mathbf{x}}^2)p(\mathbf{t}_{n_1+i})}{p(\mathbf{x}_{n_1+i}, \mathbf{P}, \sigma_{\mathbf{x}}^2)}, \tag{10}$$

where $j = 1,2,\ldots,n_1$ and $i = 1,2,\ldots,n_2$. Again, we have used the conditional independence property of the input and output variables given the latent variable. It can be easily inferred that all terms in the right hand of Eqs. 9 and 10 are Gaussian distributed. Therefore, the two calculated posterior probabilities also follow Gaussian distribution; the expected mean and variance values of which can be easily formulated as follows:

$$E(\hat{\mathbf{t}}_j|\mathbf{x}_j, \mathbf{y}_j) = (\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{P} + \sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{C} + \mathbf{I})^{-1}(\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{x}_j + \sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{y}_j)$$
$$\times E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j, \mathbf{y}_j) = (\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{P} + \sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{C} + \mathbf{I})^{-1} + E(\hat{\mathbf{t}}_j|\mathbf{x}_j, \mathbf{y}_j)E^T(\hat{\mathbf{t}}_j|\mathbf{x}_j, \mathbf{y}_j) \tag{11}$$

$$E(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i}) = (\mathbf{P}^T\mathbf{P} + \sigma_{\mathbf{x}}^2\mathbf{I})^{-1}\mathbf{P}^T\mathbf{x}_{n_1+i}$$
$$E(\hat{\mathbf{t}}_i\hat{\mathbf{t}}_i^T|\mathbf{x}_i, \mathbf{y}_i) = \sigma_{\mathbf{x}}^2(\mathbf{P}^T\mathbf{P} + \sigma_{\mathbf{x}}^2\mathbf{I})^{-1} + E(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i}, \mathbf{y}_{n_1+i})E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i}, \mathbf{y}_{n_1+i}). \tag{12}$$

In the M-step, the new parameter set Θ_{new} can be calculated by maximizing the log-likelihood function. This can be done by setting the partial derivatives of $L(\mathbf{X}, \mathbf{Y})$ with respect to each parameter to be zero. The results of updated parameters are given as follows[38]:

$$\frac{\partial[L(\mathbf{X}, \mathbf{Y})]}{\partial\mathbf{P}} = 0 \Rightarrow \mathbf{P}^{\mathrm{new}} = \left[\sum_{j=1}^{n_1}\mathbf{x}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j, \mathbf{y}_j) + \sum_{i=1}^{n_2}\mathbf{x}_{n_1+i}E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})\right] \times \left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j, \mathbf{y}_j) + \sum_{i=1}^{n_2}E(\hat{\mathbf{t}}_{n_1+i}\hat{\mathbf{t}}_{n_1+i}^T|\mathbf{x}_{n_1+i})\right]^{-1} \tag{13}$$

$$\frac{\partial[L(\mathbf{X}, \mathbf{Y})]}{\partial\mathbf{C}} = 0 \Rightarrow \mathbf{C}^{\mathrm{new}} = \left[\sum_{j=1}^{n_1}\mathbf{y}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j, \mathbf{y}_j)\right] \times \left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j, \mathbf{y}_j)\right]^{-1} \tag{14}$$

$$\frac{\partial[L(\mathbf{X},\mathbf{Y})]}{\partial\sigma_{\mathbf{x}}^2}=0 \;\Rightarrow \sigma_{\mathbf{x}}^{2\text{new}}=$$
$$\sum_{j=1}^{n_1}\mathbf{x}_j^T\mathbf{x}_j+\sum_{i=1}^{n_2}\mathbf{x}_{n_1+i}^T\mathbf{x}_{n_1+i}+$$
$$\text{trace}\left\{\mathbf{P}^{\text{new}T}\mathbf{P}^{\text{new}}\left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j)+\sum_{i=1}^{n_2}E(\hat{\mathbf{t}}_{n_1+i}\hat{\mathbf{t}}_{n_1+i}^T|\mathbf{x}_{n_1+i})\right]\right\}$$
$$-2\text{trace}\left\{\mathbf{P}^{\text{new}T}\left[\sum_{j=1}^{n_1}\mathbf{x}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)+\sum_{i=1}^{n_2}\mathbf{x}_{n_1+i}E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})\right]\right\}$$
$$mn \tag{15}$$

$$\frac{\partial[L(\mathbf{X},\mathbf{Y})]}{\partial\sigma_{\mathbf{y}}^2}=0 \;\Rightarrow \sigma_{\mathbf{y}}^{2\text{new}}$$
$$=\frac{\sum_{j=1}^{n_1}\mathbf{y}_j^T\mathbf{y}_j+\text{trace}\{\mathbf{C}^{\text{new}T}\mathbf{C}^{\text{new}}[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j)]\}-2\text{trace}\left\{\mathbf{C}^{\text{new}T}\left[\sum_{j=1}^{n_1}\mathbf{y}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)\right]\right\}}{rn_1}, \tag{16}$$

where trace(·) is a calculator for matrix trace.

Up to now, we have developed the detailed semisupervised PCR model. An important assumption of this model is that the number of retained principal components is known beforehand. However, how to determine this important number is not an easy task, especially under the semisupervised PCR model structure. As mentioned above, the number selection of the retained principal components can be considered as a model complexity problem. In this case, the Bayesian method can be incorporated, which has been widely used for model selection in different areas. In this method, a hyperparameter is introduced upon the prior distribution of the model parameter, based on which the complexity of the model can be successfully controlled. In the following subsection, the Bayesian regularization method of the semisupervised PCR model will be demonstrated.

Bayesian regularization of semisupervised PCR model

First, a Gaussian prior distribution over each of the two loading matrices {**P**,**C**} of the semisupervised PCR model defined in Eq. 4 can be defined, which are given as follows[39]:

$$p(\mathbf{P}|\alpha)=\prod_{i=1}^{d}\left(\frac{\alpha_i}{2\pi}\right)^{m/2}\exp\left\{-\frac{1}{2}\alpha_i\|\mathbf{p}_i\|^2\right\} \tag{17}$$

$$p(\mathbf{C}|\beta)=\prod_{j=1}^{d}\left(\frac{\beta_j}{2\pi}\right)^{r/2}\exp\left\{-\frac{1}{2}\beta_j\|\mathbf{c}_j\|^2\right\}, \tag{18}$$

where $\mathbf{p}_i$ is the ith column of the loading matrix $\mathbf{P}$, and $\mathbf{c}_j$ is the jth column of the loading matrix $\mathbf{C}$. The two hyperparameter vectors $\alpha=\{\alpha_1,\alpha_2,\ldots,\alpha_d\}$ and $\beta=\{\beta_1,\beta_2,\ldots,\beta_d\}$ are introduced to control the dimensionality of the loading matrices $\mathbf{P}$ and $\mathbf{C}$. If α_i or β_j has a large value, the corresponding $\mathbf{p}_i$ or $\mathbf{c}_j$ will tend to be very small and so would be effectively removed from the loading matrix. Therefore, the effective dimensionality of the principal components can be determined as follows:

$$k=\min[\dim(\mathbf{P}),\dim(\mathbf{C})], \tag{19}$$

where min[·] represents to extract the minimum value and dim(·) represents the dimensionality of the corresponding loading matrix.

To optimize the parameter set of the new SBPCR model, we can maximize the log posterior distribution function, which can be calculated by the Bayesian rule, given as

$$\ln p(\mathbf{P},\mathbf{C}|\mathbf{X},\mathbf{Y})=\ln\frac{p(\mathbf{X},\mathbf{Y}|\mathbf{P},\mathbf{C})p(\mathbf{P})p(\mathbf{C})}{\int\int p(\mathbf{X},\mathbf{Y},\mathbf{P},\mathbf{C})d\mathbf{P}d\mathbf{C}}$$
$$=\ln\frac{p(\mathbf{X}_1,\mathbf{Y}|\mathbf{P},\mathbf{C})p(\mathbf{X}_2|\mathbf{P})p(\mathbf{P})p(\mathbf{C})}{\int\int p(\mathbf{X}_1,\mathbf{Y},\mathbf{P},\mathbf{C})p(\mathbf{X}_2,\mathbf{P})d\mathbf{P}d\mathbf{C}}$$
$$=L(\mathbf{X},\mathbf{Y})-\tfrac{1}{2}\sum_{i=1}^{m-1}\alpha_i\|\mathbf{p}_i\|^2-\tfrac{1}{2}\sum_{j=1}^{m-1}\beta_j\|\mathbf{c}_j\|^2+\text{const}, \tag{20}$$

where we have initially set the dimensionality of the principal component as its maximum value $d=\text{s}\ m-1$. To estimate the optimal values of $\mathbf{P}$, $\mathbf{C}$, and σ_{x}^2, σ_{y}^2, the iterative EM algorithm can again be used. Thus, by maximizing $\mathbf{P}$, $\mathbf{C}$, and σ_{x}^2, σ_{y}^2 in the posterior distribution function given in Eq. 20, the EM algorithm can be constructed as follows.

In the E-step of the EM algorithm, the expected sufficient statistics of the principal components under labeled and unlabeled datasets can be evaluated as follows:

$$E(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)=(\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{P}+\sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{C}+\mathbf{I})^{-1}(\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{x}_j+\sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{y}_j)$$
$$E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j)=(\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{P}+\sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{C}+\mathbf{I})^{-1}+E(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)E^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j) \tag{21}$$

$$E(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})=(\mathbf{P}^T\mathbf{P}+\sigma_{\mathbf{x}}^2\mathbf{I})^{-1}\mathbf{P}^T\mathbf{x}_{n_1+i}$$
$$E(\hat{\mathbf{t}}_{n_1+i}\hat{\mathbf{t}}_{n_1+i}^T|\mathbf{x}_{n_1+i})=\sigma_{\mathbf{x}}^2(\mathbf{P}^T\mathbf{P}+\sigma_{\mathbf{x}}^2\mathbf{I})^{-1}+E(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i}) \tag{22}$$

where $j=1,2,\ldots,n_1$, and $i=1,2,\ldots,n_2$.

Based on the expected sufficient statistics obtained in the E-step, the M-step could update the model parameters $\{\mathbf{P},\mathbf{C},\sigma_{\text{x}}^2,\sigma_{\text{y}}^2,\alpha,\beta\}$ by maximizing the log posterior distribution function given in Eq. 20; the updated results are detailedly demonstrated as follows:

$$\hat{\alpha}_i=\frac{m}{\|\mathbf{p}_i\|^2} \tag{23}$$

$$\hat{\beta}_j=\frac{m}{\|\mathbf{c}_j\|^2} \tag{24}$$

$$\mathbf{P}^{\text{new}}=\left[\sum_{j=1}^{n_1}\mathbf{x}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)+\sum_{i=1}^{n_2}\mathbf{x}_{n_1+i}E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})\right]\times\left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j)+\sum_{i=1}^{n_2}E(\hat{\mathbf{t}}_{n_1+i}\hat{\mathbf{t}}_{n_1+i}^T|\mathbf{x}_{n_1+i})+\sigma_{\mathbf{x}}^2\mathbf{A}\right]^{-1} \tag{25}$$

$$\mathbf{C}^{\text{new}}=\left[\sum_{j=1}^{n_1}\mathbf{y}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)\right]\left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j)+\sigma_{\mathbf{y}}^2\mathbf{B}\right]^{-1} \tag{26}$$

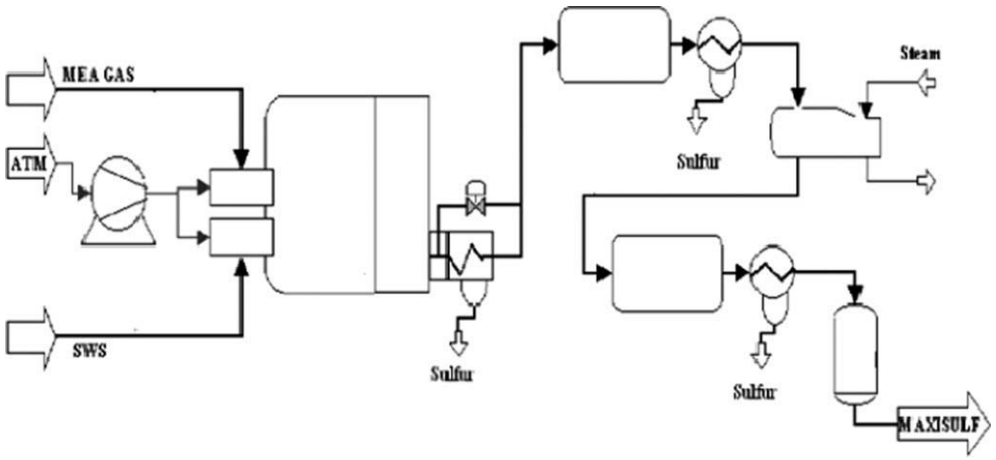


Figure 1. A simplified scheme of the SRU process.[2]

$$\sigma_{\mathbf{x}}^{2\text{new}} = \frac{\sum_{j=1}^{n_1}\mathbf{x}_j^T\mathbf{x}_j + \sum_{i=1}^{n_2}\mathbf{x}_{n_1+i}^T\mathbf{x}_{n_1+i} + \text{trace}\left\{\mathbf{P}^{\text{new}T}\mathbf{P}^{\text{new}}[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j) + \sum_{i=1}^{n_2}E(\hat{\mathbf{t}}_{n_1+i}\hat{\mathbf{t}}_{n_1+i}^T|\mathbf{x}_{n_1+i})]\right\} - 2\text{trace}\left\{\mathbf{P}^{\text{new}T}[\sum_{j=1}^{n_1}\mathbf{x}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j) + \sum_{i=1}^{n_2}\mathbf{x}_{n_1+i}E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})]\right\}}{mn} \tag{27}$$

$$\sigma_{\mathbf{y}}^{2\text{new}} = \frac{\sum_{j=1}^{n_1}\mathbf{y}_j^T\mathbf{y}_j + \text{trace}\{\mathbf{C}^{\text{new}T}\mathbf{C}^{\text{new}}[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j)]\} - 2\text{trace}\left\{\mathbf{C}^{\text{new}T}\left[\sum_{j=1}^{n_1}\mathbf{y}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)\right]\right\}}{rn_1}, \tag{28}$$

where $\mathbf{A} = \text{diag}\,(\alpha_i, i = 1,2,\ldots,d)$ and $\mathbf{B} = \text{diag}\,(\beta_j, j = 1,2,\ldots,d)$ are both diagonal matrices. Therefore, until all of the parameters are converged, the optimal values of $\{\mathbf{P}, \mathbf{C}, \text{and } \sigma_x^2, \sigma_y^2, \alpha, \beta\}$ can be determined by recursively calculating E-step and M-step given above. Detailed derivation of the EM algorithm for the SBPCR model is provided in Appendix A.

Soft sensor model and prediction

After the SBPCR model has been developed, the corresponding soft sensor can be constructed. Suppose the retained number of principal components is selected as k. When the new measurement input data sample $\mathbf{x}_{\text{new}}$ is available, the principal component $\mathbf{t}_{\text{new}}$ can be first calculated as

$$\hat{\mathbf{t}}_{\text{new}} = [\sigma_{\mathbf{x}}^{-2}\mathbf{I} + \mathbf{P}(:,1:k)^T + \mathbf{P}(:,1:k)]^{-1}\mathbf{P}(:,1:k)^T\mathbf{x}_{\text{new}} \tag{29}$$

where $\mathbf{P}(:,1{:}k)$ means to extract the first k columns of the $\mathbf{P}$ loading matrix. Then, the estimated predicted variables can be calculated as follows:

$$\hat{\mathbf{y}}_{\text{new}} = \mathbf{C}(:,1:k)\hat{\mathbf{t}}_{\text{new}} = \mathbf{C}(:,1:k) \times (\sigma_{\mathbf{x}}^{-2}\mathbf{I} + \mathbf{P}(:,1:k)^T\mathbf{P}(:,1:k))^{-1}\mathbf{P}(:,1:k)^T\mathbf{x}_{\text{new}} \tag{30}$$

Similarly, $\mathbf{C}(:,1{:}k)$ represents the first k columns of the $\mathbf{C}$ loading matrix. Therefore, the estimation error of the probabilistic soft sensor is given as follows:

$$\mathbf{er}_{\text{new}} = \mathbf{y}_{\text{new}} - \hat{\mathbf{y}}_{\text{new}}, \tag{31}$$

where $\mathbf{y}_{\text{new}}$ is the real value of the predicted variables.

To evaluate the performance of the developed probabilistic soft sensor, the root mean square error (RMSE) criterion can be used, which is defined as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{L}\left\|\mathbf{y}_j - \hat{\mathbf{y}}_j\right\|^2}{L}}, \tag{32}$$

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

Case Studies

Two industrial application case studies are given in this section for performance evaluation of the proposed method: the sulfur recovery unit (SRU) process and the debutanizer distillation process.

SRU process

The SRU can remove environmental pollutants from acid gas streams before they are released into the atmosphere. Two kinds of acid gases are taken as inputs of this process, which are MEA gas (rich in H_2S) and the sour water stripping (SWS) gas that comes from the SWS plant and is rich in H_2S and NH_3. In this process, MEA and SWS gases are first burnt in reactors, through which H_2S can be transformed into pure sulfur. Then, the products are cooled to generate the liquid sulfur. Finally, the liquid sulfur is passed through high-temperature converters, and the sulfur can be produced.[42] A simplified scheme of the SRU process can be described in Figure 1.

To monitor the conversion process and improve the sulfur extraction rate, soft sensors are necessary, which can be used for online measuring the concentrations of both H_2S and

Table 1. Input Variables of the Soft Sensor in SRU Process

Input Variables Number	Variable Descriptions
x_1	MEA gas flow
x_2	First air flow
x_3	Second air flow
x_4	Gas flow in SWS zone
x_5	Air flow in SWS zone

SO_2 in the tail gas of each sulfur line. In this special SRU process, five process variables are used as the input variables, which are listed in Table 1, and the concentrations of H_2S and SO_2 are considered as the two output variables. For model construction, a total of 1680 data samples have been collected. Similarly, we partition the dataset into two parts: the labeled dataset and the unlabeled dataset. Different proportion values of the labeled dataset have been studied, which are between 50 and 5% of the total data samples.

First, the number of principal components should be determined. Through the same Bayesian regularization method in the numerical example, the number of principal components is selected as 3. Different from the numerical example, only the last dimension of the loading matrix **C** has been switched off. The α or β values are presented in Table 2, through which we can easily find that only the last value of β has become infinity. Correspondingly, the three-dimensional exhibition of the squared element value of the loading matrix **C** is shown in Figure 2. Therefore, the number of the principal components should be determined as 3.

After the number of principal components has been determined, we are in the position to evaluate the performance of the semisupervised PCR soft sensor. For comparison, we use the following three different soft sensors: (1) the traditional PLS-based soft sensor, (2) the soft sensor based on self-training method and PCR, and (3) the soft sensor based on the missing data estimation and PCR. The detailed prediction results of all four soft sensors under different proportion values of labeled data are provided in Table 3. Through this table, we can find that with the increase number of labeled data samples for model construction, the prediction performances of all methods will be improved. However, the semisupervised PCR soft sensor performs much better than that of the PLS soft sensor, especially when the proportion of the labeled data samples is small. It is noted the other two soft sensors also perform much better than the PLS-based soft sensor, this is because both of them have used additional unlabeled data samples, which is the same case of the proposed method. The detailed prediction results of semisupervised PCR- and PLS-based soft sensors under the lowest proportion value are presented in Figure 3. By comparing the two results, it can be found that the accuracy of the semisupervised PCR method is much higher than that of the PLS method. Particularly, the prediction performance has been greatly improved in the data samples

Table 2. α and β Values of Different Principal Component Directions in SRU Process

Principal Component Directions	1	2	3	4
α values	2.90	2.93	7.07	37.89
β values	1.07	1.10	1.56	Inf

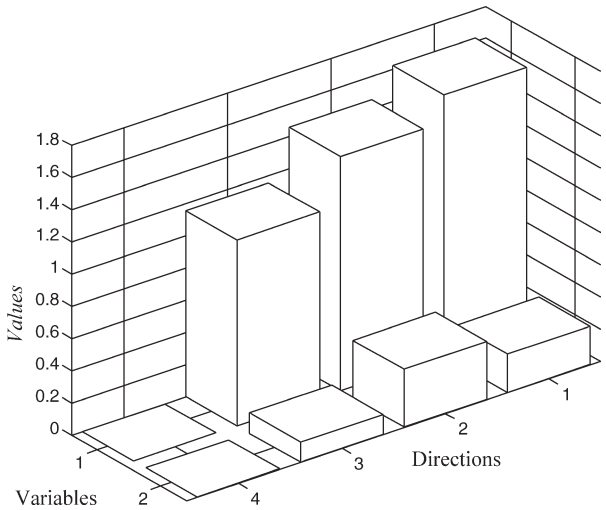


Figure 2. Three-dimensional exhibitions of the squared element values in the loading matrix C.

250–450 and 1450–1600, which are highlighted by two ellipses. To be clear, the detailed results of these two special periods are enlarged and are shown in Figures 4a, b, respectively. It is straightforward that performance of the soft senor will be improved when new labeled samples are included, because the additional data information has been incorporated. However, compared with PLS, the semisupervised PCR method can incorporate the labeled and unlabeled data information simultaneously. In our opinion, semisupervised PCR is more efficient in extracting data features than PLS when the training dataset consists of unlabeled data samples.

Debutanizer distillation process

Traditionally, the debutanizer distillation process 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 very 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 gray circles represent the used process variables in this case study for soft sensor development.[43]

For prediction of the butane content in this process, seven input variables have been selected, which are listed in Table 4. A total of 1400 data samples have been collected in the normal process. The dataset can be partitioned into two parts: the modeling database (1000 samples) for training and the testing dataset (1000 samples) for testing. Similarly, we partition the training dataset into two parts: the labeled dataset and the unlabeled dataset. Different proportion values of the labeled dataset have been studied, which are between 50 and 10% of the total data samples.

Previously, the number of principal components should be determined for the prediction model. Through the Bayesian regularization method, a total of three principal components have been selected. Similarly, the values of α or β are tabulated in Table 5, through which we can easily find that the

Table 3. Prediction Results (RMSE) of Two Soft Sensors under Different Proportions of Labeled Data

Soft Sensors/Proportion		5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Semisupervised PCR	y_1	0.1266	0.0780	0.0591	0.0595	0.0572	0.0571	0.0549	0.0548	0.0545	0.0538
	y_2	0.0637	0.0592	0.0586	0.0572	0.0572	0.0570	0.0570	0.0568	0.0567	0.0567
PLS	y_1	0.2012	0.1775	0.1576	0.0639	0.0587	0.0585	0.0577	0.0564	0.0554	0.0545
	y_2	0.0774	0.0775	0.0731	0.0577	0.0569	0.0574	0.0569	0.0572	0.0571	0.0569
Self-training	y_1	0.1312	0.1054	0.0624	0.0601	0.0574	0.0573	0.0561	0.0555	0.0545	0.0541
	y_2	0.0654	0.0601	0.0588	0.0572	0.0571	0.0572	0.0571	0.0569	0.0568	0.0567
Missing data estimation	y_1	0.1712	0.1313	0.1176	0.0599	0.0572	0.0573	0.0553	0.0551	0.0547	0.0544
	y_2	0.0689	0.0631	0.0617	0.0575	0.0573	0.0572	0.0574	0.0571	0.0569	0.0569

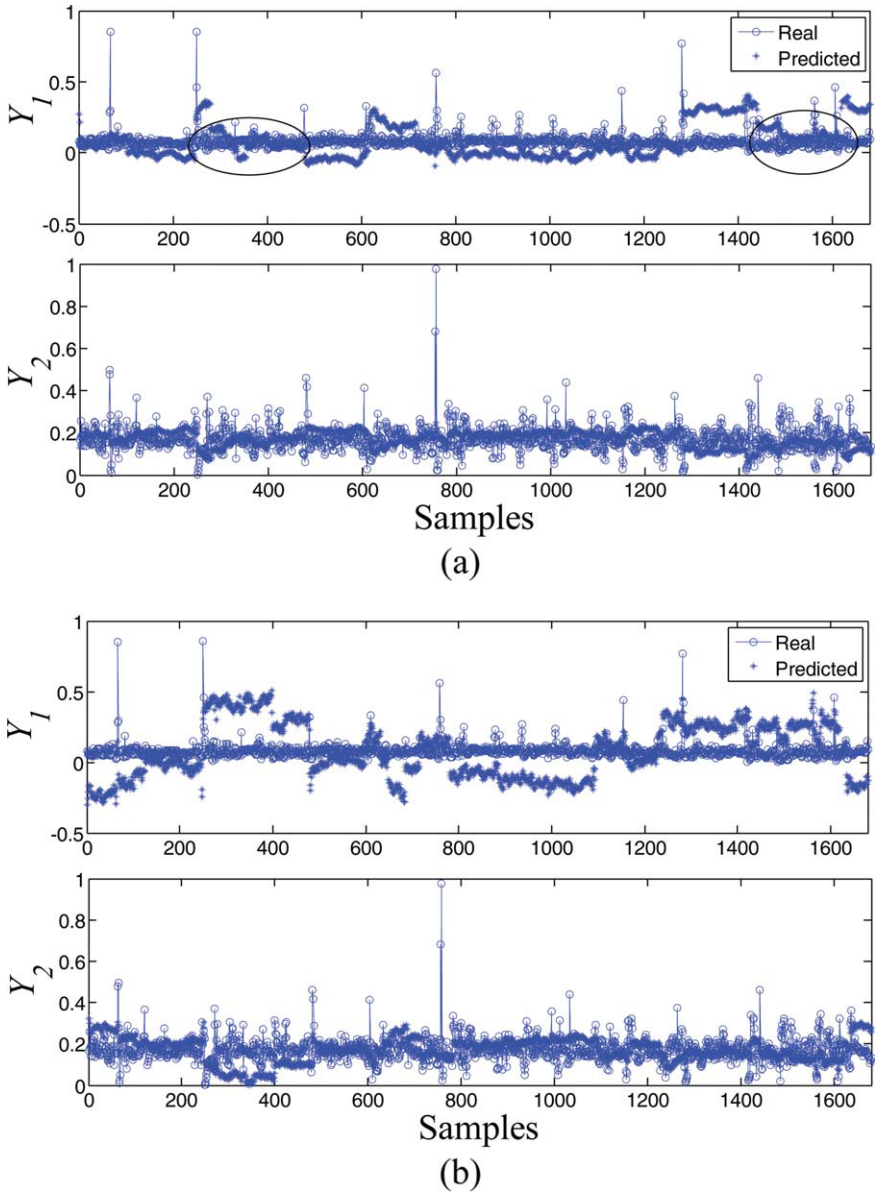
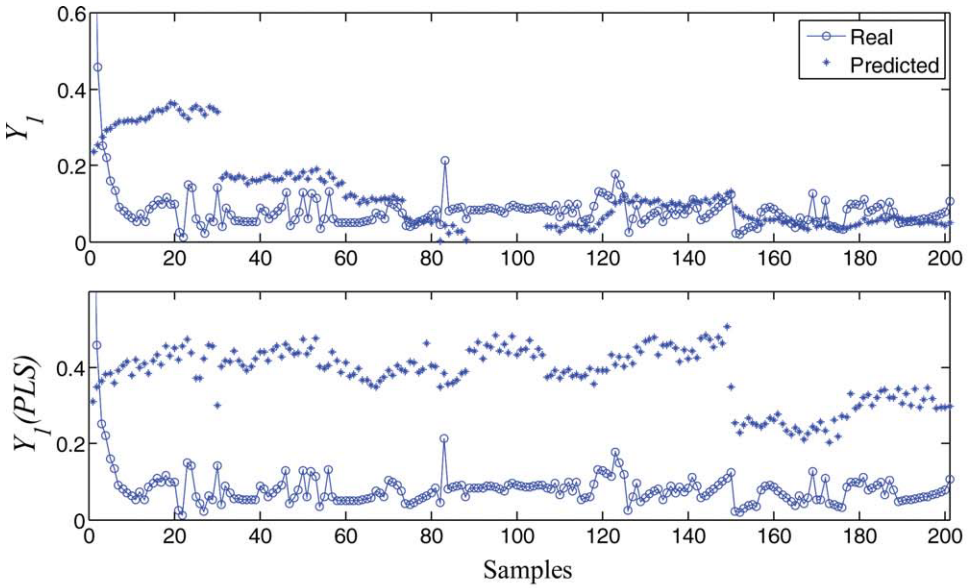


Figure 3. Prediction results under the proportion value of 5%.

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

last three values of β have become infinity, which means their corresponding dimensions should be switched off from the model structure. Correspondingly, the squared element values of the loading vector **C** are shown in Figure 6, through which we can also find that the last three dimensions of the loading vector have become zero. Next, the testing dataset is used to evaluate the performance of the four different soft sensors. Detailed prediction results of all soft sensors under



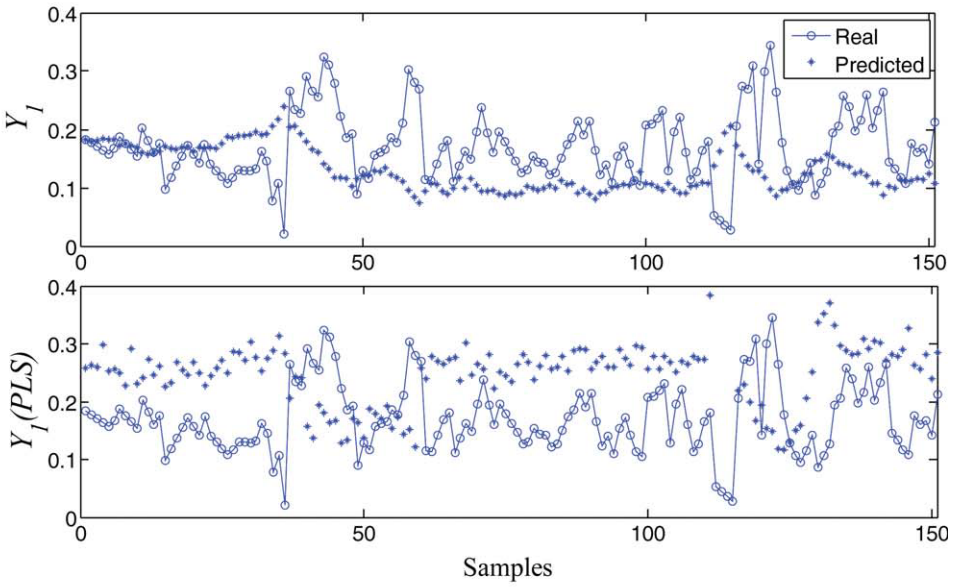


Figure 4. Detailed prediction results of two particular time periods.

(a) 250–450 and (b) 1450–1600. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

different proportion values of labeled data are provided in Table 6. Similarly, the semisupervised PCR-based soft sensor has better prediction performance than other three soft

Table 4. 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

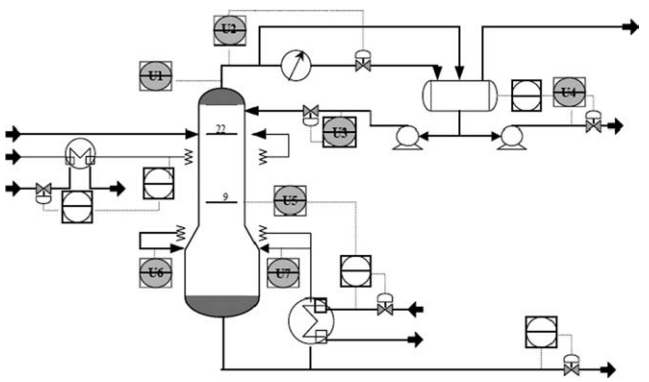


Figure 5. The flowchart of the debutanizer column.[2]

Table 5. α and β Values of Different Principal Component Directions in Debutanizer Column

Principal Component Directions	1	2	3	4	5	6
α values	3.06	6.17	6.66	7.03	10.33	10.35
β values	4.99	5.48	6.58	Inf	Inf	Inf

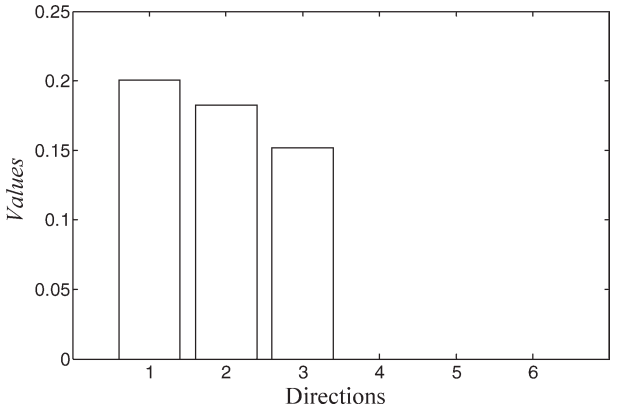


Figure 6. Squared element values in the loading vector C.

sensors in most cases. Compared with other soft sensors, the PLS-based soft sensor has poor performance, because it has only used the labeled dataset for modeling and totally ignored the information of unlabeled data samples. It is noted that all three semisupervised manner soft sensors show significant improvements to the traditional PLS-based soft sensor. Therefore, based on industrial application case studies of this example and the SRU process, we think it is necessary to include the unlabeled data sample for soft sensor development. Compared with the traditional supervised soft sensor, the semisupervised soft sensor is more efficient for online estimation and prediction. The detailed prediction results of semisupervised PCR- and PLS-based soft sensors under the proportion value of 10% are presented in Figure 7.

Conclusions

In this article, a semisupervised Bayesian method has been proposed for soft sensor modeling. Different from traditional soft sensor models, the unlabeled data information can also be incorporated in the new model structure, based on which the prediction performance of the soft sensor can be further improved. To determine the effective dimensionality of the principal components, the Bayesian regularization method is used into the semisupervised PCR model. As a result, the dimensionality of the principal components can be automatically determined, which is controlled by two hyper-parameters in the Bayesian model. To evaluate the efficiency of the newly developed soft sensor, two case studies have been studied, both of which gave satisfactory results.

Table 6. Prediction Results (RMSE) of the Soft Sensor under Different Proportions of Labeled Data

Soft Sensors/Proportion	10%	20%	30%	40%	50%
Semisupervised PCR	0.1810	0.1785	0.1683	0.1530	0.1505
PLS	0.1948	0.1862	0.1765	0.1579	0.1568
Self-training	0.1845	0.1795	0.1689	0.1529	0.1507
Missing data estimation	0.1889	0.1805	0.1715	0.1541	0.1516

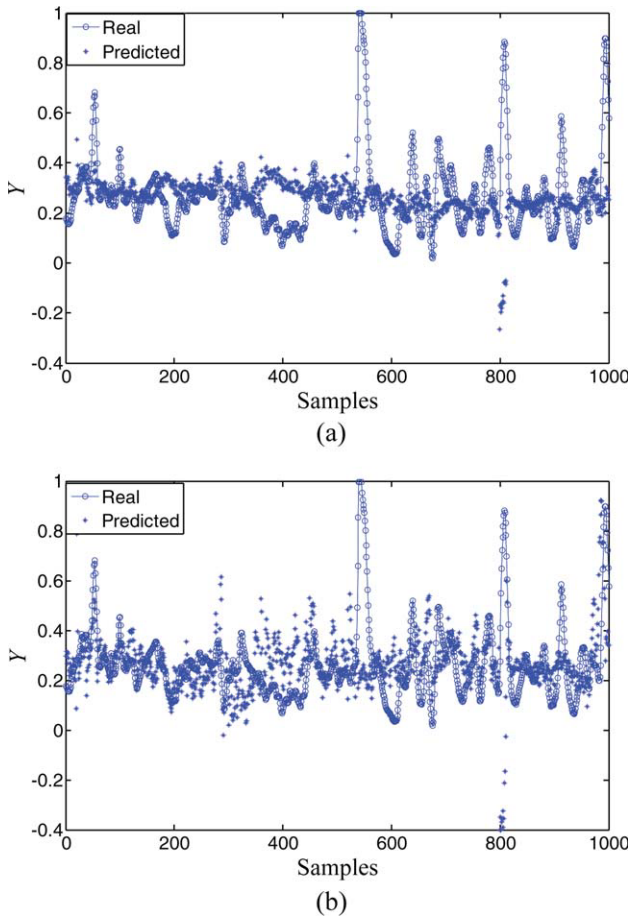


Figure 7. Prediction results under the proportion value of 10%.

(a) Semisupervised PCR and (b) PLS. [Color figure can be viewed in the online issue, which is available at wiley onlinelibrary.com.]

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China. (61004134, 60974056), the National 863 High Technology Research and Development Program of China (2009AA04Z154), and China Postdoctoral Science Foundation (20090461370).

Notation

$\mathbf{X} \in R^{n\times m}$ = process variable matrix
$\mathbf{X}_1 \in R^{n_1\times m}$ = labeled process variable matrix
$\mathbf{X}_2 \in R^{n_2\times m}$ = unlabeled process variable matrix
$\mathbf{Y} \in R^{n\times r}$ = predicted variable matrix
n = number of data samples
n_1 = number of labeled data samples
n_2 = number of unlabeled data samples
m = number of process variables
r = number of predicted variables
k = selected number of principal components
$\mathbf{P} \in R^{m\times k}$ = loading matrix of the PCR model
$\mathbf{T} \in R^{n\times k}$ = principal component matrix of the PCR model

$\mathbf{C} \in R^{r\times k}$ = regression matrix of the PCR model
$\mathbf{x}_{new} \in R^{m\times l}$ = new data sample
$\mathbf{e} \in R^{m\times l}$ = noise vector of input variables
$\mathbf{f} \in R^{r\times l}$ = noise vector of output variables
σ_x^2 = noise variance of the input variable
σ_y^2 = noise variance of the output variable
Θ = parameter set of the semisupervised PCR model
trace(·) = a calculator for matrix trace
α = hyperparameter set corresponding to the loading matrix $\mathbf{P}$
β = hyperparameter set corresponding to the loading matrix $\mathbf{C}$
min[·] = extract the minimum value
dim(·) = dimensionality of the corresponding loading matrix
const = constant value
d = maximum dimensionality value of the principal component
$\mathbf{A}$ = diagonal matrix of the hyperparameter set α
$\mathbf{B}$ = diagonal matrix of the hyperparameter set β
$\mathbf{P}(:,1{:}k)$ = first k columns of the $\mathbf{P}$ loading matrix
$\mathbf{C}(:,1{:}k)$ = first k columns of the $\mathbf{C}$ loading matrix
RMSE = root mean square error

Literature Cited

1. Jolliffe IT. *Principal Component Analysis*. Aberdeen, UK: Springer Verlag, 2002.
2. Fortuna L, Graziani S, Rizzo A, Xibilia MG. *Soft Sensors for Monitoring and Control of Industrial Processes*. London: Springer, 2007.
3. Kano M, Nakagawa Y. Data-based process monitoring, process control and quality improvement: recent developments and applications in steel industry. *Comput Chem Eng*. 2008;32:12–24.
4. Qin SJ. Recursive PLS algorithm for adaptive data modeling. *Comput Chem Eng*. 1998;22:503–514.
5. Kruger U, Chen Q, Sandoz DJ, McFarlane RC. Extended PLS approach for enhanced condition monitoring of industrial processes. *AIChE J*. 2001;47:2076–2091.
6. Li CF, Ye H, Wang GZ, Zhang J. A recursive nonlinear PLS algorithm for adaptive nonlinear process modeling. *Chem Eng Technol*. 2005;28:141–152.
7. Zhao CH, Wang FL, Mao ZZ, Lu NY, Jia MX. Quality prediction based on phase-specific average trajectory for batch processes. *AIChE J*. 2008;54:693–705.
8. Zhang YW, Zhang Y. Complex process monitoring using modified partial least squares method of independent component regression. *Chem Intell Lab Syst*. 2009;98:143–148.
9. Mandic DP, Chambers JA. *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability*. Chichester, UK: Wiley, 2001.
10. 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.
11. Himmelblau DM. Accounts of experience in the application of artificial neural networks in chemical engineering. *Ind Eng Chem Res*. 2008;47:5782–5796.
12. 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.
13. Vapnik VN. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
14. Scholkopf B, Smola AJ. *Learning with Kernels: Support Vector Machine, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.
15. Agrawal M, Jade AM, Jayaraman VK, Kulkarni BD. Support vector machine: a useful tool for process engineering applications. *Chem Eng Prog*. 2003;98:57–62.
16. Taylor JS, Crisianini N. *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press, 2004.
17. Yan WW, Shao HH, Wang XF. Soft sensing modeling based on support vector machine and Bayesian model selection. *Comput Chem Eng*. 2004;28:1489–1498.
18. Laskov P, Gehl C, Kruger S, Muller KR. Incremental support vector learning: analysis, implementation and application. *J Mach Learn Res*. 2006;7:1909–1936.
19. Jain P, Rahman I, Kulkarni BD. Development of a soft sensor for a batch distillation column using support vector regression techniques. *Chem Eng Res Des*. 2007;85:283–287.
20. Zhang YW. Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM. *Chem Eng Sci*. 2009;64:801–811.
21. Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Comput Chem Eng*. 2009;33:795–814.
22. Xiao XS, Mukkamala R, Cohen RJ. A weighted-principal component regression method for the identification of physiologic systems. *IEEE Trans Biomed Eng*. 2006;53:1521–1530.
23. Barros AS, Pinto R, Delgadillo I, Rutledge DN. Segmented principal component transform-partial least squares regression. *Chem Intell Lab Syst*. 2007;89:59–68.
24. Keithley RB, Heien ML, Wightman RM. Multivariate concentration determination using principal component regression with residual analysis. *Trends Anal Chem*. 2009;28:1127–1136.
25. Dempster A, Laird N, Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J R Stat Soc Ser B*. 1977;39:1–38.
26. Pedrycz W, Waletzky J. Fuzzy clustering with partial supervision. *IEEE Trans Syst Man Cyber B*. 1997;27:787–795.
27. Bennett K, Demiriz A. Semi-supervised support vector machines. *Adv Neural Inf Process Syst*. 1999;11:368–374.
28. Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. *Proceedings of the 18th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 2001: 19–26.
29. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7:147–177.
30. Zhu X. Semi-supervised learning in literature survey. Technical report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.
31. Chapelle O, Zien A, Scholkopf, B. *Semi-Supervised Learning*. London: MIT Press, 2006.
32. Cardoso-Cachopo A, Oliveira AL. Semi-supervised single-label text categorization using centroid-based classifiers. *Appl Comput*. 2007;1:844–851.
33. Hein M, Audibert JY, Luxburg UV. Graph laplacians and their convergence on random neighborhood graphs. *J Mach Learn Res*. 2007;8:1325–1368.
34. King BR, Guda C. Semi-supervised learning for classification of protein sequence data. *Sci Program*. 2008;16:5–29.
35. Gornitz N, Kloft M, Brefeld U. Active and semi-supervised data domain description. *Lect Notes Artif Intell*. 2009;5781:407–422.
36. Zhang Y, Yeung DY. Semi-supervised multi-task regression. *Lect Notes Artif Intell*. 2009;5782:617–631.
37. Prakash S, Robles-Kelly A. A semi-supervised approach to space carving. *Pattern Recognit*. 2010;43:506–518.
38. Yu SP, Yu K, Tresp V, Kriege HP, Wu MR. Supervised probabilistic principal component analysis. *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, 2006:464–473.
39. Bishop CM, Bayesian PCA. *Adv Neural Inf Proc Syst*. 1999;11:382–388.
40. Bishop CM. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
41. Nounou MN, Bakshi BR, Goel PK, Shen XT. Bayesian principal component analysis. *J Chemom*. 2002;16:576–595.
42. Fortuna L, Rizzo A, Sinatra M, Xibilia MG. Soft analyzers for a sulfur recovery unit. *Control Eng Pract*. 2003;11:1491–1500.
43. Fortuna L, Graziani S, Xibilia MG. Soft sensors for product quality monitoring in debutanizer distillation column. *Control Eng Pract*. 2005;13:499–508.

Appendix: Derivation of the EM Algorithm for the SBPCR Model

In the E-step of the EM algorithm, the posteriori distribution of the principal components under labeled and unlabeled data samples can be calculated as follows:

$$p(\mathbf{t}|\mathbf{x},\mathbf{y},\mathbf{P},\mathbf{C},\sigma_{\mathbf{x}}^2,\sigma_{\mathbf{y}}^2) = \frac{p(\mathbf{x}|\mathbf{t},\mathbf{P},\sigma_{\mathbf{x}}^2)p(\mathbf{y}|\mathbf{t},\mathbf{C},\sigma_{\mathbf{y}}^2)p(\mathbf{t})}{p(\mathbf{x},\mathbf{y},\mathbf{P},\mathbf{C},\sigma_{\mathbf{x}}^2,\sigma_{\mathbf{y}}^2)} \quad \text{(A1)}$$

$$p(\mathbf{t}|\mathbf{x},\mathbf{P},\sigma_{\mathbf{x}}^2) = \frac{p(\mathbf{x}|\mathbf{t},\mathbf{P},\sigma_{\mathbf{x}}^2)p(\mathbf{t})}{p(\mathbf{x},\mathbf{P},\sigma_{\mathbf{x}}^2)}. \quad \text{(A2)}$$

It is noted that all terms in the right side of Eqs. 35 and 36 are Gaussian distributed. Therefore, the posteriori distributions of the principal components under both types of data samples are also Gaussian. Thus, the expected sufficient statistics of the corresponding principal components can be easily obtained as follows:

$$E(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j) = (\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{P} + \sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{C} + \mathbf{I})^{-1}(\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{x}_j + \sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{y}_j)$$

$$E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j) = (\sigma_{\mathbf{x}}^{-2}\mathbf{P}^T\mathbf{P} + \sigma_{\mathbf{y}}^{-2}\mathbf{C}^T\mathbf{C} + \mathbf{I})^{-1} + E(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)E^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j) \quad \text{(A3)}$$

$$E(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i}) = (\mathbf{P}^T\mathbf{P} + \sigma_{\mathbf{x}}^2\mathbf{I})^{-1}\mathbf{P}^T\mathbf{x}_{n_1+i}$$

$$E(\hat{\mathbf{t}}_{n_1+i}\hat{\mathbf{t}}_{n_1+i}^T|\mathbf{x}_{n_1+i}) = \sigma_{\mathbf{x}}^2(\mathbf{P}^T\mathbf{P} + \sigma_{\mathbf{x}}^2\mathbf{I})^{-1} + E(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i}) \quad \text{(A4)}$$

where $j = 1,2,\ldots,n_1$ and $i = 1,2,\ldots,n_2$. In the M-step, by maximizing the log posterior distribution function given in Eq. 20 with respect to $\mathbf{P}$, $\mathbf{C}$, σ_{x}^2, and σ_{y}^2 and set them to zero, yields the following results

$$\frac{\partial[\ln p(\mathbf{P},\mathbf{C}|\mathbf{X},\mathbf{Y})]}{\partial\mathbf{P}} = 0 \Rightarrow$$

$$\left\{\sum_{j=1}^{n_1}[\sigma_{\mathbf{x}}^{-2}(\mathbf{x}_j - \hat{\mathbf{P}}E(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j))E^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)] + \sum_{i=1}^{n_2}[\sigma_{\mathbf{x}}^{-2}(\mathbf{x}_{n_1+i} - \hat{\mathbf{P}}\mathrm{E}(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i}))E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})]\right\} - \hat{\mathbf{P}}\mathrm{A} = 0$$

$$\Rightarrow (\sigma_{\mathbf{x}}^{-2}\mathbf{I})^T\hat{\mathbf{P}}\mathrm{A} - \left\{\sum_{j=1}^{n_1}[\sigma_{\mathbf{x}}^{-2}(\mathbf{x}_j - \hat{\mathbf{P}}\mathrm{E}(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j))E^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)] + \sum_{i=1}^{n_2}[\sigma_{\mathbf{x}}^{-2}(\mathbf{x}_{n_1+i} - \hat{\mathbf{P}}\mathrm{E}(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})]\right\} = 0$$

$$\Rightarrow \mathbf{P}^{\mathrm{new}} = \left[\sum_{j=1}^{n_1}\mathbf{x}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j) + \sum_{i=1}^{n_2}\mathbf{x}_{n_1+i}E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})\right] \times \left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j) + \sum_{i=1}^{n_2}E(\hat{\mathbf{t}}_{n_1+i}\hat{\mathbf{t}}_{n_1+i}^T|\mathbf{x}_{n_1+i}) + \sigma_{\mathbf{x}}^2\mathbf{A}\right]^{-1} \quad \text{(A5)}$$

$$\frac{\partial[\ln p(\mathbf{P},\mathbf{C}|\mathbf{X},\mathbf{Y})]}{\partial\mathbf{C}} = 0 \Rightarrow$$

$$\sum_{j=1}^{n_1}\left[\sigma_{\mathbf{y}}^{-2}(\mathbf{y}_j - \hat{\mathbf{C}}E(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j))E^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)\right] - \hat{\mathbf{C}}\mathrm{B} = 0$$

$$\Rightarrow (\sigma_{\mathbf{y}}^{-2}\mathbf{I})^T\hat{\mathbf{C}}\mathbf{B} - \sum_{i=1}^{n}[(\mathbf{y}_i - \hat{\mathbf{C}}E(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j))E^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)] = 0$$

$$\Rightarrow \mathbf{C}^{\mathrm{new}} = \left[\sum_{j=1}^{n_1}\mathbf{y}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)\right]\left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j) + \sigma_{\mathbf{y}}^2\mathbf{B}\right]^{-1} \quad \text{(A6)}$$

$$\frac{\partial[\ln p(\mathbf{P},\mathbf{C}|\mathbf{X},\mathbf{Y})]}{\partial\sigma_{\mathbf{x}}^2} = 0 \Rightarrow$$

$$\sigma_{\mathbf{x}}^{2\mathrm{new}} = \frac{\sum_{j=1}^{n_1}\mathbf{x}_j^T\mathbf{x}_j + \sum_{i=1}^{n_2}\mathbf{x}_{n_1+i}^T\mathbf{x}_{n_1+i} + \mathrm{trace}\left\{\mathbf{P}^{\mathrm{new}T}\mathbf{P}^{\mathrm{new}}\left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j) + \sum_{i=1}^{n_2}E(\hat{\mathbf{t}}_{n_1+i}\hat{\mathbf{t}}_{n_1+i}^T|\mathbf{x}_{n_1+i})\right]\right\} - 2\mathrm{trace}\left\{\mathbf{P}^{\mathrm{new}}\left[\sum_{j=1}^{n_1}E^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)\mathbf{x}_j + \sum_{i=1}^{n_2}E^T(\hat{\mathbf{t}}_{n_1+i}|\mathbf{x}_{n_1+i})\mathbf{x}_{n_1+i}\right]\right\}}{mn} \quad \text{(A7)}$$

$$\frac{\partial[\ln p(\mathbf{P},\mathbf{C}|\mathbf{X},\mathbf{Y})]}{\partial\sigma_{\mathbf{y}}^2} = 0 \Rightarrow$$

$$\sigma_{\mathbf{y}}^{2\mathrm{new}} = \frac{\sum_{j=1}^{n_1}\mathbf{y}_j^T\mathbf{y}_j + \mathrm{trace}\left\{\mathbf{C}^{\mathrm{new}T}\mathbf{C}^{\mathrm{new}}\left[\sum_{j=1}^{n_1}E(\hat{\mathbf{t}}_j\hat{\mathbf{t}}_j^T|\mathbf{x}_j,\mathbf{y}_j)\right]\right\} - 2\mathrm{trace}\left\{\mathbf{C}^{\mathrm{new}T}\left[\sum_{j=1}^{n_1}\mathbf{y}_jE^T(\hat{\mathbf{t}}_j|\mathbf{x}_j,\mathbf{y}_j)\right]\right\}}{rn_1}. \quad \text{(A8)}$$

Manuscript received Feb. 6, 2010, revision received Apr. 23, 2010, and final revision received Aug. 27, 2010.