sensors

MDPI

Article

Student's-*t* Mixture Regression-Based Robust Soft Sensor Development for Multimode Industrial Processes

Jingbo Wang, Weiming Shao * and Zhihuan Song *

State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China; wangjingbobo@zju.edu.cn

* Correspondence: shaowm@zju.edu.cn (W.S.); songzhihuan@zju.edu.cn (Z.S.)

Received: 4 September 2018; Accepted: 13 November 2018; Published: 15 November 2018

Abstract: Because of multiple manufacturing phases or operating conditions, a great many industrial processes work with multiple modes. In addition, it is inevitable that some measurements of industrial variables obtained through hardware sensors are incorrectly observed, recorded or imported into databases, resulting in the dataset available for statistic analysis being contaminated by outliers. Unfortunately, these outliers are difficult to recognize and remove completely. These process characteristics and dataset imperfections impose challenges on developing high-accuracy soft sensors. To resolve this problem, the Student's-*t* mixture regression (SMR) is proposed to develop a robust soft sensor for multimode industrial processes. In the SMR, for each mixing component, the Student's-*t* distribution is used instead of the Gaussian distribution to model secondary variables, and the functional relationship between secondary and primary variables is explicitly considered. Based on the model structure of the SMR, a computationally efficient parameter-learning algorithm is also developed for SMR. Results conducted on two cases including a numerical example and a real-life industrial process demonstrate the effectiveness and feasibility of the proposed approach.

Keywords: robust soft sensor; multimode process; Student's-*t* mixture regression; Gaussian mixture model; expectation maximization

1. Introduction

In industrial processes, there is a class of quality-related variables that is very important but difficult to measure, such as melt index in the polypropylene process, catalyst activation in chemical reactions, thickness of strip in the hot rolling process, octane number of gasoline, etc. Measurements of these quality variables are conventionally obtained by expensive online analyzers or time-consuming laboratory analysis, which introduces huge investment cost or large time delay [1]. Soft sensors, which are essentially mathematical models, are capable of predicting these key variables (referred to as "primary variables") online using easy-to-measure process variables (referred to as "secondary variables") such as flow rate, temperature, pressure, etc. Therefore, soft sensors are economical and real-time alternatives to conventional measurement of quality variables, and play an important role in process monitoring, closed-loop control, process optimization and so forth [2–6]. Owing to their advantages, in recent years, soft sensors have been intensively researched and extensively applied to industrial processes [7–11].

The methods for soft sensor modeling can generally be categorized into two groups, which are first-principle methods [12] and data-driven methods [13]. As modern industrial processes grow increasingly complex, it is difficult to obtain first-principle models. By contrast, data-driven models can be easily obtained because a large amount of process data that reflects the true operating conditions is collected in databases via field instruments [10,14]. Thus, data-driven soft sensors have gained increasing attention and popularity in real industrial processes. In the past decade, a variety of modeling algorithms have been developed and applied to construct soft sensor models. Partial least squares [15] and principle component regression [16] which are linear models for describing the relationship between quality variables and secondary variables, have been studied systematically and are widely used in real applications. Aiming at dealing with process non-linearities, soft sensors based on artificial neural networks [17] and support vector machines [18] have also been developed. Extensive reviews for the approaches and applications of soft sensors in real industrial processes can be found in [19].

Due to multiple product-grade requirements, feedstock changes, load variations, seasonal operations, etc., most industrial processes work with multiple operation modes [20]. The multimode characteristics result in process variables that are no longer Gaussian, and the functional relationship between primary and secondary variables being strongly non-linear [2], which increases the difficulty in developing high-accuracy soft sensor models. To deal with these issues, the finite mixture model (FMM) has been widely investigated and applied to real-life industrial processes. The Gaussian mixture model (GMM), which is one of the most widely adopted approaches in the FMM family, possesses the capability of approximating arbitrary unknown random distributions, including those with multiple peaks; meanwhile, GMM provides a simple and computationally efficient maximum-likelihood estimation framework by means of the expectation-maximization (EM) algorithm. Over the past few years, several studies based on GMM have been conducted for soft sensor development [13,21]. Gaussian mixture regression (GMR) treats the input space and output space together to obtain the joint probability density function (PDF) of quality and secondary variables. Then, the conditional PDF of primary variables given secondary variables can be calculated directly from their joint PDF, which can be used to derive the regression relationship between quality and secondary variables.

However, the parameter-learning procedure for the GMM is extremely sensitive to outliers, which may cause the estimated PDF of interested variables to be significantly distorted or excessive components to be required for capturing the tails of the distributions [22–25]. The outliers can be partitioned into two types, conspicuous outliers and indistinctive outliers, according to whether they are beyond the physical meaning or not. Conspicuous outliers can be easily examined and eliminated, while it is difficult to discriminate and address indistinctive outliers.

To tackle this issue, the Student's-*t* mixture model (SMM) has been proposed as an alternative to GMM, which provides stronger robustness against outliers by means of heavier tails [26]. In the Student's-*t* distribution, an additional parameter ν (often called degrees of freedom) compared to Gaussian distribution can be viewed as the robustness-tuning parameter. Recently, the SMM has been applied in signal/image processing applications such as human action recognition [27], medical imaging for segmentation [28], and fall detection [29], through which the SMM has achieved much better performance compared with the GMM. However, up to now, to our best knowledge, no literature has been found reporting the soft sensor based on the SMM for industrial processes. Therefore, the use of SMM for soft sensor application has not been explored. In this paper, the Student's-*t* mixture regression (SMR) structure for the purpose of soft sensor development, which explicitly considers the functional dependency between the primary and secondary variables, is first proposed, followed by an EM algorithm-based parameter-learning algorithm for the SMR.

The rest of this paper is organized as follows. In Section 2, a brief review of the Student's-*t* distribution and SMM are represented, followed by the elaboration of SMR as well as the procedure for parameter-learning and soft sensor development based on SMR in Section 3. In Section 4, the effectiveness and feasibility of the SMR are verified in two case studies including a numerical example and a real-life industrial process. Finally, conclusions and future work are given in Section 5.

2. Preliminaries

2.1. Student's-t Distribution

The PDF of a d-dimensional Student's-t distribution, with mean $\boldsymbol{\mu}$, precision matrix $\boldsymbol{\Lambda}$ and degree of freedom ν, is denoted as

$$St(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\nu/2 + d/2)|\boldsymbol{\Lambda}|^{1/2}}{\Gamma(\nu/2)(\nu\pi)^{d/2}}(1 + \frac{\Delta^2}{\nu})^{-(\nu+d)/2} \tag{1}$$

where $\Gamma(t) = \int_0^\infty z^{t-1}e^{-z}dz$ is the Gamma function, and $\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}$.

The Student's-t distribution can be viewed as an infinite mixture of scaled Gaussian distributions, i.e.,

$$St\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu\right) = \int_0^\infty \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) Gam\left(\eta\middle|\frac{\nu}{2}, \frac{\nu}{2}\right) \mathrm{d}\eta \tag{2}$$

where $\mathcal{N}(\cdot)$ represents the Gaussian distribution, η stands for the intermediate latent variable which is helpful for deriving the analytical solution, and $Gam(\cdot)$ denotes the Gamma distribution.

Figure 1 illustrates the Student's-t distribution with fixed mean vector and covariance matrix but various degrees of freedom. It can be seen that the Student's-t distribution degrades the Gaussian distribution in the limit $\nu \to +\infty$. Moreover, the tail of the Student's-t distribution tends to be heavier when the degree of freedom $\nu \to 0$. Therefore, the Student's-t distribution possesses the potentiality to mitigate the adverse effect of outliers in contrast to the Gaussian distribution.

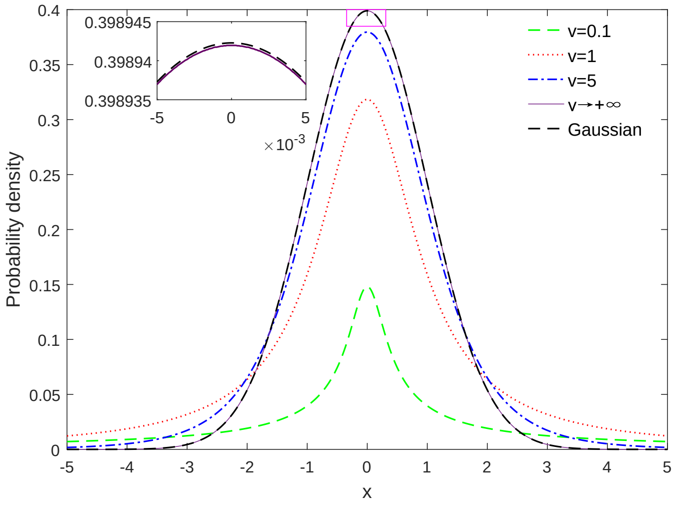


Figure 1. Illustration of Student's-t distribution with various degrees of freedom.

2.2. Student's-t Mixture Model

Assume the secondary variable $\boldsymbol{x}$ follows the mixture distributions with K components as

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k St(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k) \tag{3}$$

where the mixing coefficients $\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots, \pi_K\}$ satisfy $\sum_{k=1}^{K} \pi_k = 1$ together with $0 \leqslant \pi_k \leqslant 1$. In addition, let us introduce a K-dimensional assignment latent variable $\boldsymbol{z} = (z_1, \cdots, z_K)$ associated with $\boldsymbol{x}$, in which z_k for $k = 1, 2, \cdots, K$ are binary variables, i.e., $z_k \in \{0, 1\}$. In addition, only one of the z_k for $k = 1, 2, \cdots, K$ can be assigned with value 1, and the rest ones are all 0. Therefore, we have the constraint $\sum_{k=1} z_k = 1$. If certain $z_k = 1$, it means that the k-th component is responsible for generating the corresponding observed sample.

The prior distribution over $\boldsymbol{z}$ is specified in accordance with the mixing coefficients π_k as

$$p(z_k = 1) = \pi_k \tag{4}$$

Using the *1-of-K* coding scheme, the prior distribution over $\boldsymbol{z}$ can also be written in the form

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{5}$$

Similarly, the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{z}$ is a Student's-*t* distribution

$$p(\boldsymbol{x}|z_k = 1) = St(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k) \tag{6}$$

which can also be written as

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} St(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k)^{z_k} \tag{7}$$

3. Methodology

In practical applications, data collected from industrial processes are very likely to be contaminated by outliers, and it is usually non-trivial to completely remove all outliers. It has been demonstrated that the performance of GMM might be rather disappointing with the presence of outliers because the tails of the Gaussian distribution in many applications are shorter than required [22,30]. To this end, we propose the Student's-*t* distribution mixture regression (SMR) which is detailed in this subsection.

3.1. Student's-t Mixture Regression

Let us denote $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}^{\mathrm{T}} \in R^{N\times d}$ and $\boldsymbol{Y} = \{y_1, \cdots, y_N\}^{\mathrm{T}} \in R^{N\times 1}$ as the input and output space of samples data, and the input variable $\boldsymbol{x}$ is assumed to be generated from Student's-*t* distribution mixture models with K components as Equation (3).

The SMR is illustrated in Figure 2 in the form of a probabilistic graphical model.

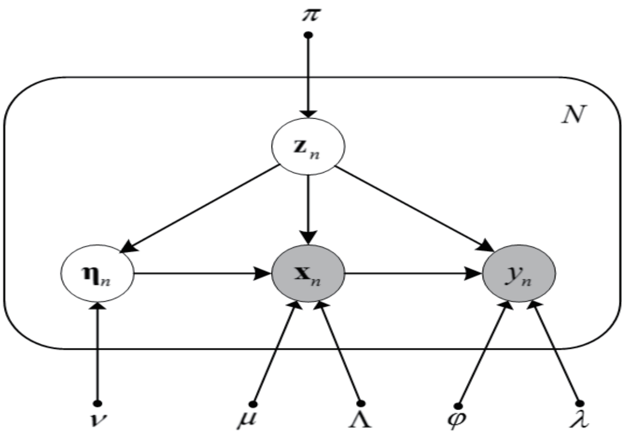


Figure 2. Probabilistic graphical model representation for the Student's-*t* mixture regression model given a set of N independent identically distributed data points $\{\boldsymbol{x}_n, y_n\}$, with corresponding latent variables $\{\boldsymbol{z}_n, \boldsymbol{\eta}_n\}$,where $n = 1, 2, \cdots, N$.

For the convenience of mathematical derivation, let us define

$$p(\eta_{nk}) = p(\boldsymbol{\eta}_n|z_{nk} = 1) = Gam\left(\eta_{nk}\middle|\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \tag{8}$$

where $\boldsymbol{\eta}_n$ means the intermediate latent variable associated with the n-th sample of secondary variables (i.e., $\boldsymbol{x}_n$). Consequently, we have

$$p(\boldsymbol{\eta_n}|\boldsymbol{z}_n) = \prod_{k=1}^{K} Gam\left(\eta_{nk}\middle|\frac{\nu_k}{2}, \frac{\nu_k}{2}\right)^{z_{nk}} \tag{9}$$

The probability distribution over $\boldsymbol{x}_n$ conditioned on two latent variables $\boldsymbol{z}_n = (z_{n1}, \cdots, z_{nK})$ and $\boldsymbol{\eta_n}$ can be obtained as

$$p(\boldsymbol{x}_n|\boldsymbol{\eta}_n, z_{nk} = 1) = \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, (\eta_{nk}\boldsymbol{\Lambda}_k)^{-1}\right) \tag{10}$$

which can also be written as

$$p\left(\boldsymbol{x}_n|\boldsymbol{\eta_n}, \boldsymbol{z}_n\right) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, (\eta_{nk}\boldsymbol{\Lambda}_k)^{-1}\right)^{z_{nk}} \tag{11}$$

For each component, linear dependence of y_n on $\boldsymbol{x}_n$ is introduced. Taking the single-output case for example, for $k = 1, \cdots, K$, we have

$$y_n = \tilde{\boldsymbol{x}}_n^{\mathrm{T}}\boldsymbol{\varphi}_k + \varepsilon_k \tag{12}$$

where $\boldsymbol{\varphi}_k$ represents the regression coefficient vector, ε_k means zero-mean Gaussian-distributed noise variable with covariance λ_k^{-1}, and $\tilde{\boldsymbol{x}}_n = [\boldsymbol{x}_n^{\mathrm{T}}, 1]^{\mathrm{T}}$.

According to Equation (12), for the k-th component, the conditional PDF of y_n given $\boldsymbol{x}_n$ can be obtained as

$$p(y_n|\boldsymbol{x}_n, z_{nk} = 1) = \mathcal{N}(y_n|\tilde{\boldsymbol{x}}_n^{\mathrm{T}}\boldsymbol{\varphi}_k, \lambda_k^{-1}) \tag{13}$$

According to Equation (13), we have

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{z}_n) = \prod_{k=1}^{K} \mathcal{N}(y_n|\tilde{\boldsymbol{x}}_n^{\mathrm{T}}\boldsymbol{\varphi}_k, \lambda_k^{-1})^{z_{nk}} \tag{14}$$

3.2. Parameters Learning for the SMR

The parameters for the SMR that need to be learnt are denoted as $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k, \boldsymbol{\varphi}_k, \lambda_k\}_{k=1}^{K}$. The EM algorithm, consisting of the expectation step (E-step) and maximization step (M-step), is an ideal approach to addressing the issues of missing values [31] (corresponding to the latent variables appeared in the SMR). Therefore, we adopt the EM to perform the parameter-learning task for the SMR.

In the E step, the posterior distribution over latent variables $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N$, which are collectively denoted as $\boldsymbol{Z} = (\boldsymbol{z}_n)_{n=1}^{N}$, associated with the training dataset $(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{x}_n, y_n)_{n=1}^{N}$ can be calculated as

$$\begin{aligned} p(z_{nk} = 1|\boldsymbol{x}_n, y_n) &= \frac{p(z_{nk} = 1)p(y_n|\boldsymbol{x}_n, z_{nk} = 1)p(\boldsymbol{x}_n|z_{nk} = 1)}{\sum_{k=1}^{K} p(z_{nk} = 1)p(y_n|\boldsymbol{x}_n, z_{nk} = 1)p(\boldsymbol{x}_n|z_{nk} = 1)} \\ &= \frac{\pi_k\mathcal{N}(y_n|\tilde{\boldsymbol{x}}_n^{\mathrm{T}}\boldsymbol{\varphi}_k, \lambda_k^{-1})St(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k)}{\sum_{k=1}^{K} \pi_k\mathcal{N}(y_n|\tilde{\boldsymbol{x}}_n^{\mathrm{T}}\boldsymbol{\varphi}_k, \lambda_k^{-1})St(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k)} \end{aligned} \tag{15}$$

Therefore, the expectation of z_{nk} based on the posterior distribution can be calculated as

$$\begin{aligned} \langle z_{nk}\rangle &= p(z_{nk} = 1|\boldsymbol{x}_n, y_n) \\ &= \frac{\pi_k\mathcal{N}(y_n|\tilde{\boldsymbol{x}}_n^{\mathrm{T}}\varphi_k, \lambda_k^{-1})St(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k)}{\sum_{k=1}^{K} \pi_k\mathcal{N}(y_n|\tilde{\boldsymbol{x}}_n^{\mathrm{T}}\varphi_k, \lambda_k^{-1})St(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k)} \end{aligned} \tag{16}$$

Given the latent variable z_n and observed variable $\boldsymbol{x}_n$, the posterior distribution over $\boldsymbol{\eta}_n$ can be calculated as

$$\begin{aligned} & p(\boldsymbol{\eta}_n|\boldsymbol{x}_n, z_{nk}=1) \propto p(\boldsymbol{x}_n|\boldsymbol{\eta}_n, z_{nk}=1)p(\boldsymbol{\eta}_n|z_{nk}=1) \\ & \propto \mathcal{N}\Big(\boldsymbol{x}_n\Big|\boldsymbol{\mu}_k, (\boldsymbol{\eta}_n\boldsymbol{\Lambda}_k)^{-1}\Big)Gam\left(\boldsymbol{\eta}_n\middle|\frac{v_k}{2}, \frac{v_k}{2}\right) \\ & \propto \boldsymbol{\eta}_n^{\frac{d+v_k}{2}-1}\exp\left\{-\frac{(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k)+v_k}{2}\boldsymbol{\eta}_n\right\} \end{aligned} \tag{17}$$

Comparing the definition of the Gamma distribution, we have

$$\begin{aligned} p(\eta_{nk}|\boldsymbol{x}_n) &= p(\eta_n|\boldsymbol{x}_n, z_{nk}=1) \\ &= Gam\Big(\boldsymbol{\eta}_{nk}\Big|\frac{\nu_k+d}{2}, \frac{\nu_k}{2}+\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\Big) \end{aligned} \tag{18}$$

Thus, we can obtain the expectations

$$\langle\eta_{nk}\rangle = \frac{\nu_k+d}{\nu_k+(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k)} \tag{19}$$

$$\langle\ln(\eta_{nk})\rangle = \psi(\frac{\nu_k+d}{2}) - \ln\Big(\frac{\nu_k}{2}+\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\Big) \tag{20}$$

where $\psi(\cdot)$ is the digamma function defined as $\psi(x) = \mathrm{d}\big(\Gamma(x)\big)/\mathrm{d}x$.

Subsequently, in the M step, with the assumption that the samples are independent and identically distributed, the expectation of complete data log-likelihood function is first formulated as

$$\begin{aligned} L(\Theta) &= \langle\ln p(\boldsymbol{X}, \mathbf{Y}, \boldsymbol{Z}, \boldsymbol{\eta})\rangle \\ &= \langle\ln p(\boldsymbol{Y}|\mathbf{X}, \boldsymbol{Z})\rangle + \langle\ln p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\eta})\rangle + \langle\ln p(\boldsymbol{\eta}|\mathbf{Z})\rangle + \langle\ln p(\boldsymbol{Z})\rangle \end{aligned} \tag{21}$$

where

$$\langle\ln p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{nk}\rangle\Big\{-\frac{1}{2}\ln(2\pi)+\frac{1}{2}\ln(\lambda_k)-\frac{1}{2}\lambda_k(y_n-\tilde{\boldsymbol{x}}_n^{\mathrm{T}}\varphi_k)^2\Big\} \tag{22}$$

$$\langle\ln p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\eta})\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{nk}\rangle\Big\{-\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(|\boldsymbol{\Lambda}_k|)+\frac{d}{2}\langle\ln(\eta_{nk})\rangle-\frac{\langle\eta_{nk}\rangle}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\Big\} \tag{23}$$

$$\langle\ln p(\boldsymbol{\eta}|\mathbf{Z})\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{nk}\rangle\Big\{-\ln\big(\Gamma(\frac{\nu_k}{2})\big)+\frac{\nu_k}{2}\ln(\frac{\nu_k}{2})+(\frac{\nu_k}{2}-1)\langle\ln(\eta_{nk})\rangle-\frac{\nu_k}{2}\langle\eta_{nk}\rangle\Big\} \tag{24}$$

$$\langle\ln p(\boldsymbol{Z})\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{nk}\rangle\ln(\pi_k) \tag{25}$$

and $\boldsymbol{\eta} = (\boldsymbol{\eta}_n)_{n=1}^{N}$.

Setting the derivatives of Equation (21) with respect to $\boldsymbol{\mu}_k$ to zero leads to

$$\begin{aligned} \frac{\partial L(\Theta)}{\partial\boldsymbol{\mu}_k} &= \sum_{n=1}^{N}\langle z_{nk}\rangle\langle\eta_{nk}\rangle\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k) = 0 \\ &\Rightarrow \boldsymbol{\mu}_k = \sum_{n=1}^{N}\langle z_{nk}\rangle\langle\eta_{nk}\rangle\boldsymbol{x}_n\Big/\sum_{n=1}^{N}\langle z_{nk}\rangle\langle\eta_{nk}\rangle \end{aligned} \tag{26}$$

Similarly, we have

$$\begin{aligned}\frac{\partial L(\Theta)}{\partial \boldsymbol{\Lambda}_k} &= \sum_{n=1}^{N}\langle z_{nk}\rangle\Big\{\boldsymbol{\Lambda}_k^{-1} - \langle \eta_{nk}\rangle(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\Big\} = 0 \\ \Rightarrow \boldsymbol{\Lambda}_k^{-1} &= \sum_{n=1}^{N}\langle z_{nk}\rangle\langle \eta_{nk}\rangle(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\Big/\sum_{n=1}^{N}\langle z_{nk}\rangle\end{aligned} \tag{27}$$

$$\begin{aligned}\frac{\partial L(\Theta)}{\partial \lambda_k} &= \sum_{n=1}^{N}\langle z_{nk}\rangle\Big(\lambda_k^{-1} - (y_n - \tilde{\boldsymbol{x}}_n^{\mathrm{T}}\boldsymbol{\varphi}_k)^2\Big) = 0 \\ \Rightarrow \lambda_k &= \left\{\sum_{n=1}^{N}\langle z_{nk}\rangle(y_n - \tilde{\boldsymbol{x}}_n^{\mathrm{T}}\boldsymbol{\varphi}_k)^2\Big/\sum_{n=1}^{N}\langle z_{nk}\rangle\right\}^{-1}\end{aligned} \tag{28}$$

Setting the derivatives of Equation (21) with respect to $\boldsymbol{\varphi}_k$ to zero leads to

$$\begin{aligned}\frac{\partial L(\Theta)}{\partial \boldsymbol{\varphi}_k} &= \widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{R}_k\widetilde{\boldsymbol{X}}\boldsymbol{\varphi}_k - \widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{R}_k\boldsymbol{Y} = 0 \\ \Rightarrow \boldsymbol{\varphi}_k &= (\widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{R}_k\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{R}_k\boldsymbol{Y}\end{aligned} \tag{29}$$

where $\boldsymbol{R}_k = diag(\langle z_{1k}\rangle, \langle z_{1k}\rangle, \cdots, \langle z_{nk}\rangle)$, $\widetilde{\boldsymbol{X}} = [\boldsymbol{X}, \mathbf{1}]$, $\mathbf{1}$ is the column with all element 1.

The parameter ν_k can be obtained by solving the non-linear equation as follows.

$$\frac{\partial L(\Theta)}{\partial \nu_k} = \sum_{n=1}^{N}\langle z_{nk}\rangle\Big\{-\psi(\frac{\nu_k}{2}) + \ln(\frac{\nu_k}{2}) + 1 + \langle \ln(\eta_{nk})\rangle - \langle \eta_{nk}\rangle\Big\} = 0 \tag{30}$$

$$\Rightarrow -\psi(\frac{\nu_k}{2}) + \ln(\frac{\nu_k}{2}) + 1 + \frac{\sum_{n=1}^{N}\langle z_{nk}\rangle\left(\langle \ln(\eta_{nk})\rangle - \langle \eta_{nk}\rangle\right)}{\sum_{n=1}^{N}\langle z_{nk}\rangle} = 0 \tag{31}$$

Please note that it has been proved that the left-hand side of Equation (31) strictly decreases from $+\infty$ to a minus value as ν_k increases in $(0,+\infty)$ [32]. Therefore, solving Equation (31) for ν_k is not difficult by the means of many one-dimensional search methods, such as the dichotomy method.

Using the constraint $\sum_{k=1}^{K}\pi_k = 1$ and introducing the Lagrange multiplier γ, we can obtain

$$\begin{cases}\frac{\partial \widetilde{L}(\Theta)}{\partial \pi_k} = \sum_{n=1}^{N}\langle z_{nk}\rangle\Big/\pi_k + \gamma = 0 \\ \sum_{k=1}^{K}\pi_k = 1\end{cases} \Rightarrow \pi_k = \sum_{n=1}^{N}\langle z_{nk}\rangle/N \tag{32}$$

where $\widetilde{L}(\Theta) = L(\Theta) + \gamma(\sum_{k=1}^{K}\pi_k - 1)$.

In the light of the updated equations such as the derivation above, the robustness of SMR compared with GMR can be clearly seen with the use of degrees of freedom ν. As the degrees of freedom parameter ν is introduced, the outliers with large Mahalanobis distance have small value of the expectation of η_{nk} as can be drawn from Equation (19), resulting in the outliers being down-weighted and the influence of outliers on parameters estimation being significantly reduced. Taking the precision matrix of each component, for example, based on GMR the updated equation will be converted into $\boldsymbol{\Lambda}_k^{-1} = \sum_{n=1}^{N}\langle z_{nk}\rangle(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\Big/\sum_{n=1}^{N}\langle z_{nk}\rangle$, which means that the data's outliers will highly influence the estimates. However, taking this example to the extreme, the outliers which are extremely different to the majority of dataset are down-weighted to zero because in the SMR the $\langle \eta_{nk}\rangle$ associated with the outliers will be zero, resulting in the influence of outliers on precision matrix estimates being removed.

As the model above-mentioned parameters are updated by iterative learning, the iterative process terminates when $L(\Theta)$ converges, and the convergence criterion can be defined as

$$\left|\frac{L(\Theta_t) - L(\Theta_{t-1})}{L(\Theta_{t-1})}\right| < \varepsilon \tag{33}$$

where $L(\Theta_t)$ denotes the value of $L(\Theta)$ at the tth iteration and ε represents the threshold value, which is specified by the user.

Up to now, we can summarize the detailed procedure for training the SMR in Algorithm 1.

Algorithm 1 Pseudocode for training SMR.

```
Given $K$, initialize $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k, \boldsymbol{\varphi}_k, \lambda_k\}_{k=1}^{K}$, and the maximum iteration times;
Set $t = 0$;
while $t <$ maximum iteration times do
    Set $t = t + 1$;
    for $k = 1, \cdots, K$; $n = 1, \cdots, N$ do
        Calculate $\langle z_{nk} \rangle$ using Equation (16);
        Calculate $\langle \eta_{nk} \rangle$ and $\langle \ln \eta_{nk} \rangle$ using Equation (19) and Equation (20), respectively;
    end for
    for $k = 1, \cdots, K$ do
        Update $\boldsymbol{\mu}_k$, $\boldsymbol{\Lambda}_k$, $\lambda_k$ and $\boldsymbol{\varphi}_k$, $\pi_k$ with Equation (26), Equation (27), Equation (28), Equation (29)
        and Equation (32) respectively;
        Solve Equation (31) for $\nu_k$;
    end for
    Calculate $L(\Theta)$ using Equation (21).
    if the convergence criterion in Equation (33) is satisfied then
        Terminate while;
    end if
end while
```

3.3. Soft Sensor Development Based on SMR

Based on the SMR, a soft sensor model can be easily developed for predicting the quality variable y_q when a sample $\mathbf{x}_q$ of process variables is available.

To begin with, the posterior distribution of the associated latent variable $z_q = (z_{q1}, \cdots, z_{qK})$ is calculated as

$$\begin{aligned} p(z_{qk} = 1 | \boldsymbol{x}_q) &= \frac{p(\boldsymbol{x}_q | z_{qk} = 1) p(z_{qk} = 1)}{\sum_{k=1}^{K} p(\boldsymbol{x}_q | z_{qk} = 1) p(z_{qk} = 1)} \\ &= \frac{\pi_k St(\boldsymbol{x}_q | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k)}{\sum_{k=1}^{K} \pi_k St(\boldsymbol{x}_q | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k)} \triangleq R_{qk} \end{aligned} \tag{34}$$

Subsequently, the probability distribution y_q conditioned on $\boldsymbol{x}_q$ can be obtained as

$$\begin{aligned} p(y_q | \boldsymbol{x}_q) &= \sum_{k=1}^{K} p(y_q | \boldsymbol{x}_q, z_{qk} = 1) p(z_{qk} = 1 | \boldsymbol{x}_q) \\ &= \sum_{k=1}^{K} R_{qk} N(y_q | \tilde{\boldsymbol{x}}_q^{\mathrm{T}} \boldsymbol{\varphi}_k, \lambda_k^{-1}) \end{aligned} \tag{35}$$

Finally, the prediction of y_q can be obtained as

$$\widehat{y}_q = \sum_{k=1}^{K} R_{qk} \tilde{\boldsymbol{x}}_q^{\mathrm{T}} \boldsymbol{\varphi}_k \tag{36}$$

4. Case Studies

In this section, the proposed method is first evaluated using a numerical example and then applied to develop soft sensors for an industrial primary reformer in an ammonia synthesis plant [33]. For comparison purposes, the performance of multiple dynamic PLS (Multi-DPLS) [34,35] and GMR are also provided as benchmarks. Please note that the Multi-DPLS is realized by first referring to the work in [34], where the GMM is used for data clustering, followed by constructing a sub-PLS model for each data cluster. Then, we extend the PLS model to the DPLS model by augmenting the input vector according to [35].

The root mean squares error (RMSE) is used to evaluate the prediction accuracies of various methods, which is defined as

$$\text{RMSE} = \sqrt{\sum_{n=1}^{N_t} (y_n - \widehat{y}_n)^2 \Big/ N_t} \tag{37}$$

where y_n and $\widehat{y}_n$ are the true value and predicted value of quality variable, respectively, and N_t is the size of the testing dataset.

To deal with the influence of randomness of initial parameters, a total of 100 simulations are carried out for both the GMR and SMR, and their final parameters are selected as those that can minimize the RMSE on the validating dataset, while the generalization performance of various methods are evaluated on the testing dataset. The configurations of the used computer are given as follows: CPU: Core i5-4570 (3.2 GHz × 2), RAM: 8 GB, OS: Windows 10, and Software: MATLAB (R2016b). The CPU time (CPT) spent in offline model training (CPT_{trn}, in seconds) and in online predicting (CPT_{tst}, in seconds) are employed to assess the computational efficiency for different methods. In both case studies, the threshold values for diagnosing the convergence for the SMR and GMR are set as 10^{-6}.

4.1. Numerical Example

We assume a 2-dimensional input variables $\boldsymbol{x} = (x_1, x_2)^{\text{T}}$ and a scalar output y are generated from a mixture of three Student's-t distributions based on Equations (3) and (12), in which the configurations of each component are listed in Table 1. Please note that as the non-diagonal elements for the precision matrices $\boldsymbol{\Lambda}_k$ are not zero, the correlations among the input variables are taken into consideration, which can be captured by the proposed model using Equation (3). In addition, in our model setting, the vector $\boldsymbol{x} = (x_1, x_2)^{\text{T}}$ is assumed to obey a mixture of multivariate Student's-t distributions, and we do not need to build one SMM for each of variable. Figure 3 illustrates the data distributions from the input space, which clearly shows the multimode characteristics.

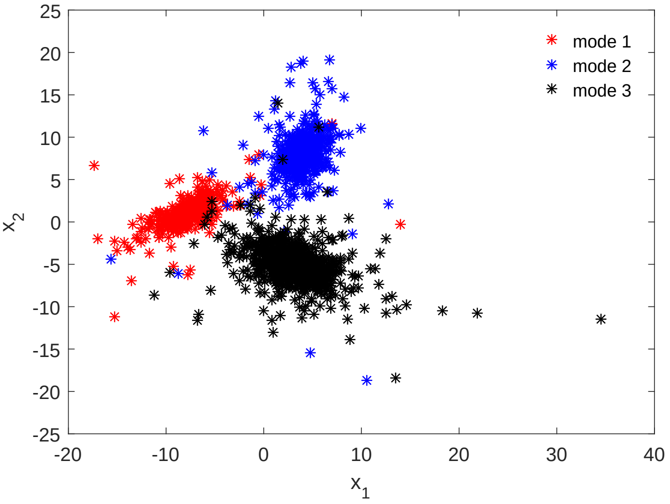


Figure 3. Visualization of the data distribution in the input space.

Table 1. Configuration of three Student component.

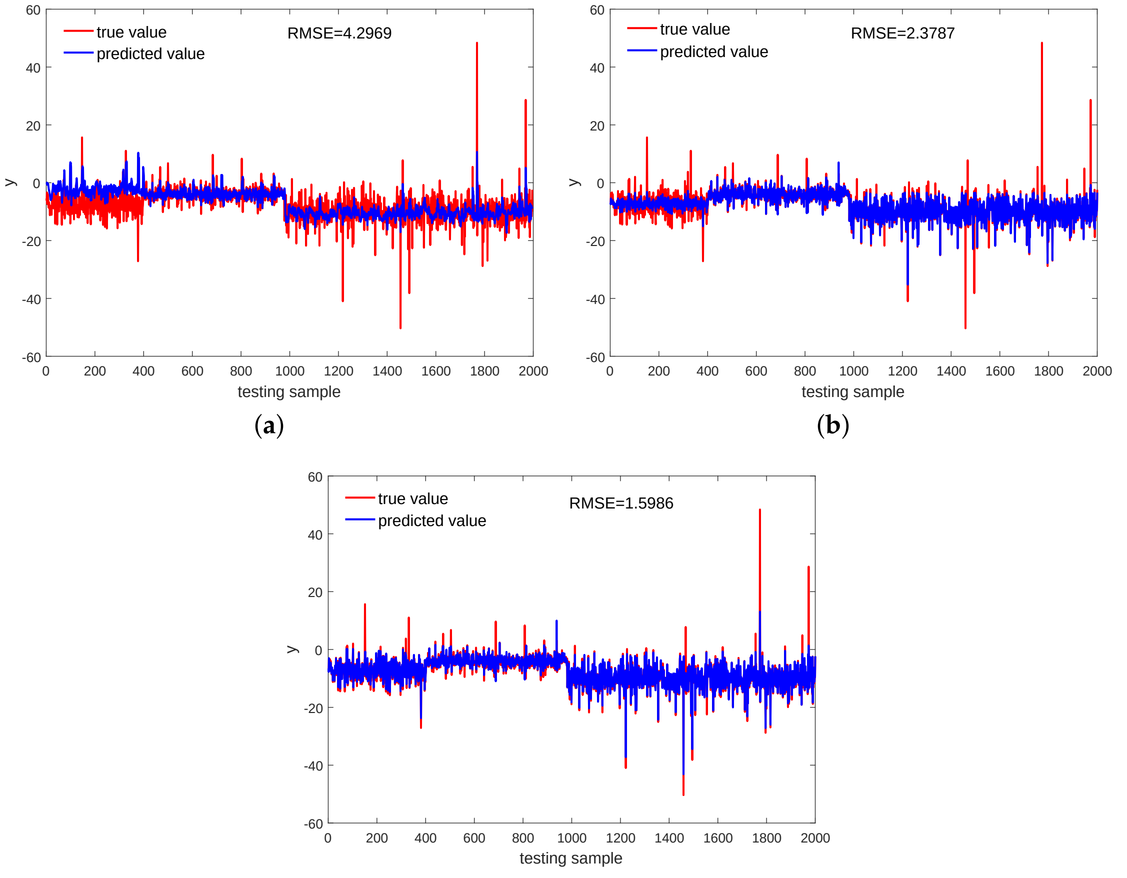
	$k=1$	$k=2$	$k=3$
π_k	0.2	0.3	0.5
$\boldsymbol{\mu}_k$	$\begin{bmatrix}-8 & 1\end{bmatrix}$	$\begin{bmatrix}4 & 8\end{bmatrix}$	$\begin{bmatrix}3 & -5\end{bmatrix}$
$\boldsymbol{\Lambda}_k$	$\begin{bmatrix}2.0 & 1.0\\1.0 & 1.0\end{bmatrix}$	$\begin{bmatrix}1.0 & 0.5\\0.5 & 2.0\end{bmatrix}$	$\begin{bmatrix}3.0 & -1.0\\-1.0 & 1.5\end{bmatrix}$
ν_k	3	3	3
$\boldsymbol{\varphi}_k$	$\begin{bmatrix}1 & 1 & 0\end{bmatrix}^{\mathrm{T}}$	$\begin{bmatrix}1 & -1 & 0\end{bmatrix}^{\mathrm{T}}$	$\begin{bmatrix}-1 & 1 & 2\end{bmatrix}^{\mathrm{T}}$
λ_k	0.25	0.25	0.25

In the simulation, three datasets, namely the training dataset, validating dataset, and testing dataset, each of which consists of 2000 samples, were generated. The training dataset is used for parameter learning, while the validating dataset is used for determining the initialized values of model parameters for the Multi-DPLS, GMR, and SMR models. In this example, the number of mixing components for Multi-DPLS, GMR, and SMR models were set as 3 in advance; in addition, the dimensionality of the latent space for each sub-PLS model in the Multi-DPLS was set as 2. The performance of various methods are evaluated on the testing dataset, which is unseen at the training stage. Moreover, 1%, 3% and 5% outliers are randomly added into the input data samples, respectively.

According to the proportion of the sample number of each mode, the outliers are generated by transforming a certain coordinate of some sample data randomly selected to the value far away from its center. For example, 3% rate outliers are added to the training dataset containing 2000 samples, which is to say there are 12, 18, and 30 outliers added to each mode, respectively [36].

By using trial and error, the order of the Multi-DPLS is determined as 4, i.e., the values of input variables in the past four moments are also used to estimate the value of the current output. Recall that in this case, data samples were generated independently with each other without dynamics. The reason the Multi-DPLS with the order of 4 achieves the best performance can be explained as follows. The augmented input vector is helpful at improving the classification accuracy for the GMM, because the samples at some augmented sampling instances may be located at non-overlapped areas among the three modes; meanwhile, the PLS can deal with the data-collinearity. That is why the performance of the Multi-DPLS gets enhanced when the order increases. However, as the order further increases, the dimensionality of the input vector significantly increases, too, which leads to inaccurate estimations of the probability density functions. That is why performance of the Multi-DPLS deteriorates when the order is greater than 4.

Predictions for y by the models based on the Multi-DPLS, GMR, and SMR with the outlier rate set as 3% are visualized in Figure 4, from which for the Multi-DPLS large deviations existing in the first mode and third mode can be clearly found. This is because the information of output space in the mode identification step is ignored, and then the performance of clustering the high-dimensional data is rather unsatisfactory, leading to a PLS model built into each mode that cannot explain the true functional dependency between the output and input variables well. In contrast, the GMR and SMR-based models, which treat the input space and output space together, are more powerful at modeling the multimode process. However, intuitively, we can recognize that the SMR performs better compared with the GMR in terms of predicting samples from the first mode.



(**a**) (**b**)

(**c**)

Figure 4. With 3% rate outliers, predictions for the output variable achieved by: (**a**) Multi-DPLS, (**b**) GMR, (**c**) SMR.

For more in-depth analyses, predictive accuracies of three methods on the validating dataset and testing dataset are quantified in Table 2. As can be seen, the performance of the Multi-DPLS model is rather disappointing, while the predictive accuracies of the GMR and SMR models are much higher. In addition, one can see that as the number of outliers increases, the performances of both the GMR and SMR-based models deteriorate. However, the deteriorations for the SMR-based model are much slighter compared with those for the GMR-based model. To be specific, as the outlier rate rises from 1% to 3% and 5%, the generalization RMSE for the GMR-based model is increased by 41.8% and 63.8%, respectively; in contrast, the increment of generalization RMSE for the SMR-based model is only 5.1% and 7.7%, respectively, which demonstrate that the SMR-based model is much more robust against outliers compared with the GMR-based model.

Table 2. RMSE of various methods on the validating and testing datasets.

Outliers	Dataset	Multi-DPLS	GMR	SMR
1%	validating	3.9414	1.9097	1.5939
	testing	4.1216	1.6776	1.5208
3%	validating	4.0450	2.0692	1.6398
	testing	4.2969	2.3787	1.5986
5%	validating	4.1307	2.2223	1.7352
	testing	4.3127	2.7476	1.6388

For probabilistic methods such as the GMR and SMR, correctly estimating the PDFs of process variables is a prerequisite to high predictive accuracy. In this synthetic case, the estimations of PDFs of x_1 and x_2 with different amounts of outliers are illustrated in Figures 5–7. One can readily recognize that due to the long tails of data distributions, the PDFs of x_1 and x_2 estimated by GMR have been significantly skewed compared with the data histograms and true PDFs. In addition, such distortion becomes more severe as the number of outliers increase. In particular, the GMR basically fails to capture the middle peak from the x_2 direction. By contrast, the PDFs estimated by the SMR fit the data histograms well, and are barely affected by the increase of outliers, which is the reason that the SMR-based model can provide satisfactory performance with various numbers of outliers.

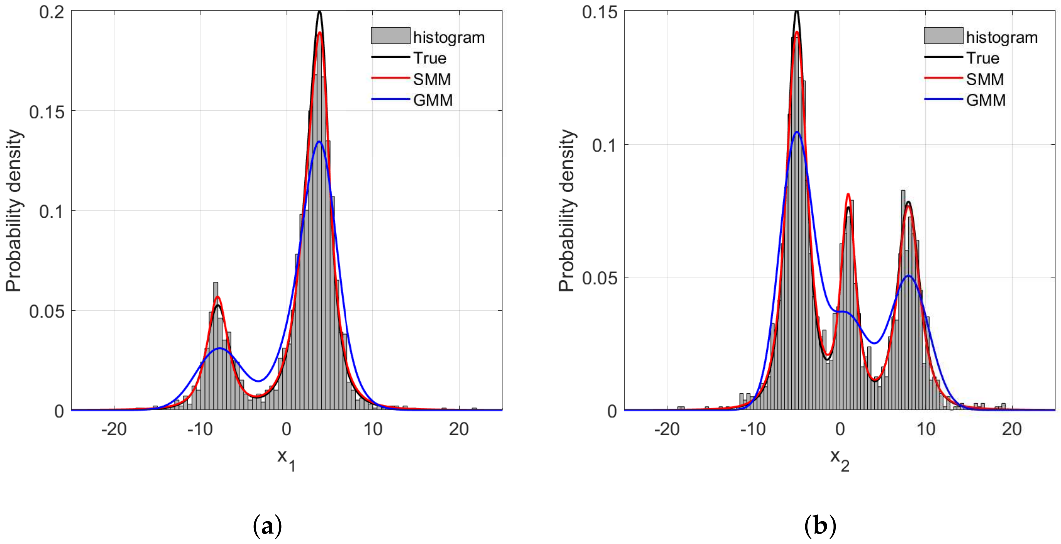


Figure 5. The frequency histogram and probability density curve with 1% rate outliers: (**a**) x_1 direction; (**b**) x_2 direction.

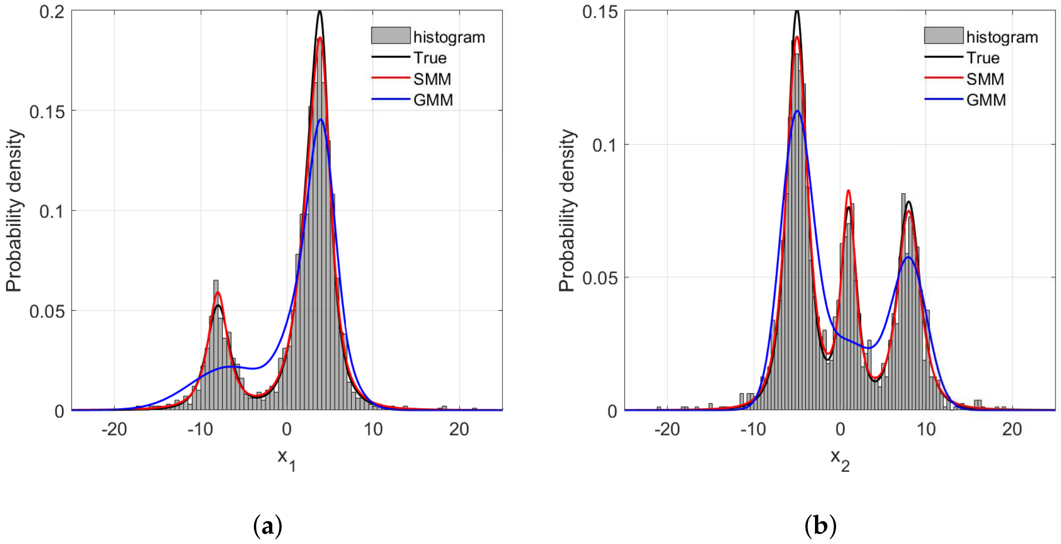
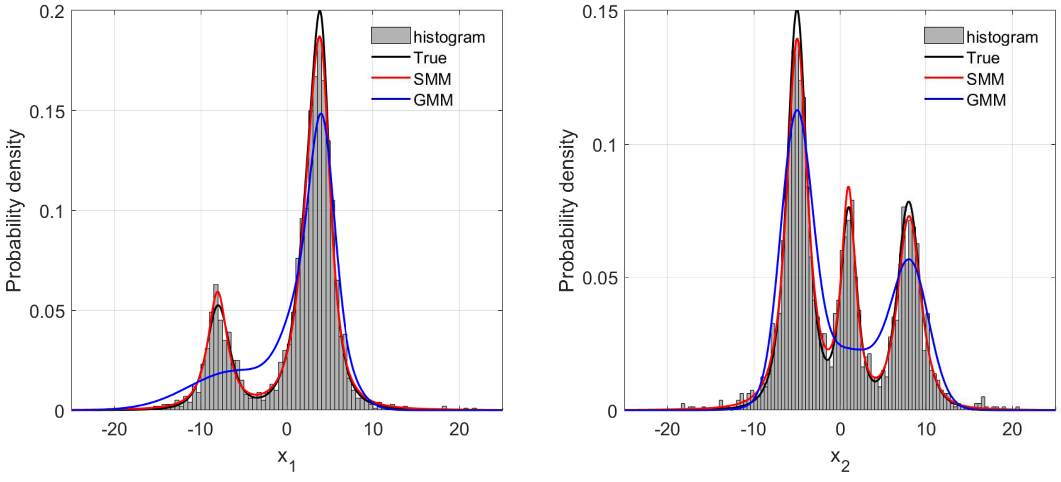


Figure 6. The frequency histogram and probability density curve with 3% rate outliers: (**a**) x_1 direction; (**b**) x_2 direction.



(**a**) (**b**)

Figure 7. The frequency histogram and probability density curve with 5% rate outliers: (**a**) x_1 direction; (**b**) x_2 direction.

For the numerical example the time consumed by these three methods are listed in Table 3.

Table 3. Average CPT (in second) consumed by various methods for the numerical example.

Outliers	CPT_{trn}			CPT_{tst}		
	Multi-DPLS	**GMR**	**SMR**	**Multi-DPLS**	**GMR**	**SMR**
1%	0.0283	0.0095	0.0951	0.0013	0.001	0.00072
3%	0.0148	0.0135	0.1087	0.0012	0.0012	0.000846
5%	0.0193	0.0164	0.1099	0.0011	0.001	0.000777

It is easily seen from Table 3 that the differences between CPT_{tst} for these three methods can be negligible. The CPT_{trn} for the Multi-DPLS and GMR are comparable. Please note that in the SMR the parameter ν_k is estimated by solving a non-linear equation with the help of the dichotomy method, which results in more time for iterative learning. However, the computational efficiency for a soft sensor based on SMR is still acceptable.

4.2. Primary Reformer

The primary reformer is an important part of hydrogen-manufacturing units in the ammonia synthesis process for producing NH_3, which is the main material in the urea synthesis process. The flowchart of the primary reformer is illustrated in Figure 8.

Main transformation reactions set off in the primary reformer are

$$\begin{aligned} & C_n H_{2n+2} + n\,H_2O \overset{\triangle}{\longleftrightarrow} n\,CO + (2n+1)\,H_2 \\ & CH_4 + H_2O \overset{\triangle}{\longleftrightarrow} CO + 3\,H_2 \\ & CO + H_2O \overset{\triangle}{\longleftrightarrow} CO_2 + H_2 \end{aligned} \tag{38}$$

According to the reaction mechanism, the temperature in the furnace plays a significant role in the purity of hydrogen; thus, the temperature should be strictly monitored and controlled, which is realized by manipulating the burning conditions at the dense burner. One of the effective approaches to stabilizing the burning condition is to control the oxygen concentration in the furnace at the specified interval. However, the measurement of oxygen concentration (i.e., the quality-related variable for the primary reformer) in practice is expensive, due to an exorbitant mass spectrometer, or time-consuming, due to offline laboratory analysis, both of which fail to satisfy the requirement of real-time control and production. To cope with this issue, a soft sensor based on a historical dataset is desirable for online estimation of the oxygen concentration, which is illustrated with a dark green block in Figure 8.

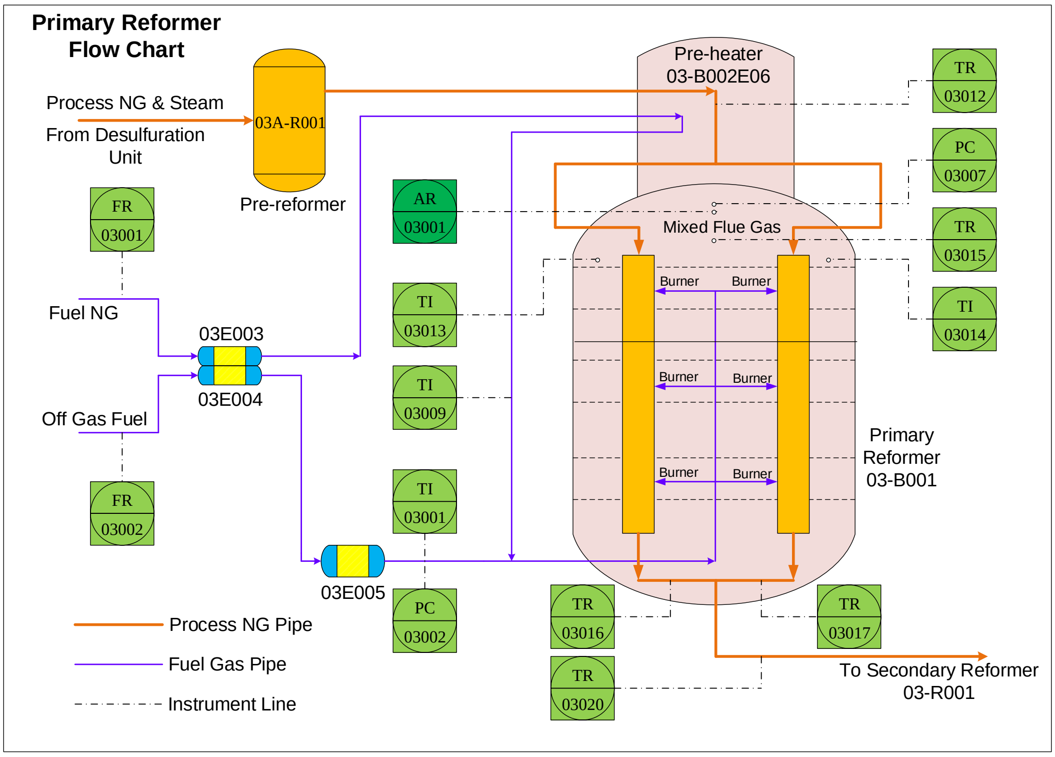


Figure 8. Flowchart of the primary reformer.

Based on expert knowledge of process mechanisms and experiences from engineers, 13 process variables, including pressures and temperatures, are selected as secondary variables for soft sensor modeling, which are illustrated with light-green blocks in Figure 8. Detailed descriptions of these secondary variables are presented in Table 4.

Table 4. Descriptions of process variables in the primary reformer.

Tags	Descriptions
FR03001.PV	Flow rate of fuel NG into 03B001
FR03002.PV	Flow rate of fuel off gas into 03B001
PC03002.PV	Pressure of fuel off gas at 03E005's exit
PC03007.PV	Pressure of furnace flue gas at 03B001's exit
TI03001.PV	Temperature of fuel off gas at 03E005's exit
TI03009.PV	Temperature of fuel NG at 03B002E06's exit
TR03012.PV	Temperature of process gas at 03B001's entrance
TI03013.PV	Temperature of furnace flue gas at 03B001's top left
TI03014.PV	Temperature of furnace flue gas at 03B001's top right
TR03015.PV	Temperature of mixed furnace flue gas at 03B001's top
TR03016.PV	Temperature of transformed gas at 03B001's left exit
TR03017.PV	Temperature of transformed gas at 03B001's right exit
TR03020.PV	Temperature of transformed gas at 03B001's exit

A total of 7000 samples recorded from January 2015 to July 2015 were collected from the database of distributed control systems of a real-world primary. The collected samples are evenly partitioned into three parts, i.e., 2000 samples serve as the training dataset, 2000 samples are used as the validating dataset for model selection, and the remaining 3000 samples constitute the testing dataset for evaluating the generalization performance of various soft sensors. By taking the testing samples, for example, it is obvious that the process basically involves five large operating conditions, as shown by the dash-dot blue line in Figure 9, which indicates that the primary reformer is characterized by multiple modes.

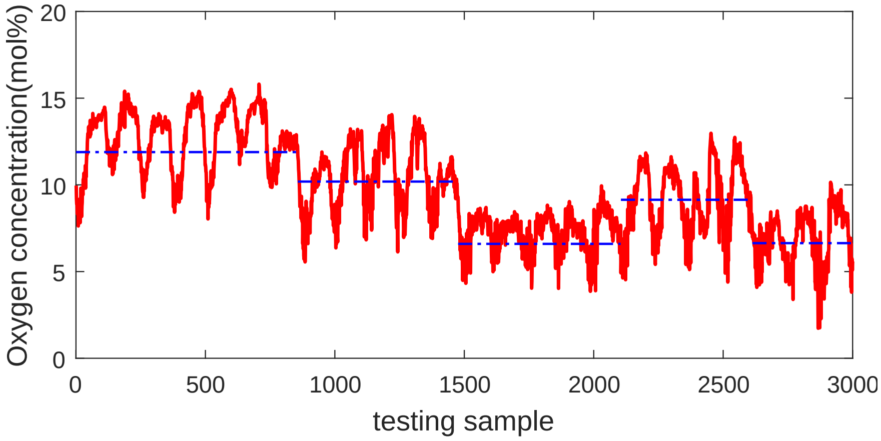


Figure 9. Visualization of multimode characteristics of the primary reformer.

As with the numerical example, the order of the Multi-DPLS is determined as 3, and Figure 10 shows that the number of components and the dimensionality of latent space are determined as 12 and 8, respectively, in which the Multi-DPLS has the minimum RMSE on the validating dataset. In addition, the initial values of model parameters for the GMR and SMR-based soft sensors, as well as the model selections for them are also completed on the validating dataset. In particular, the best performances for the GMR and SMR-based soft sensors with various numbers of mixing components (i.e., K) are visualized in Figure 11a, which indicates that both the predictive RMSEs of GMR and SMR-based soft sensors reach the minimum at $K = 18$. However, for the SMR-based soft sensor, we see that as $K \geq 15$, the validating RMSE almost stabilize at 0.88. Considering the fact that the larger the K, the higher the model complexity, we determine the optimal K for the SMR-based soft sensor as 15. Based on the same consideration, the optimal K for the GMR-based soft sensor is selected as 18. Meanwhile, for the GMR and SMR-based soft sensors, the generalization performances on the testing dataset are compared in Figure 11b.

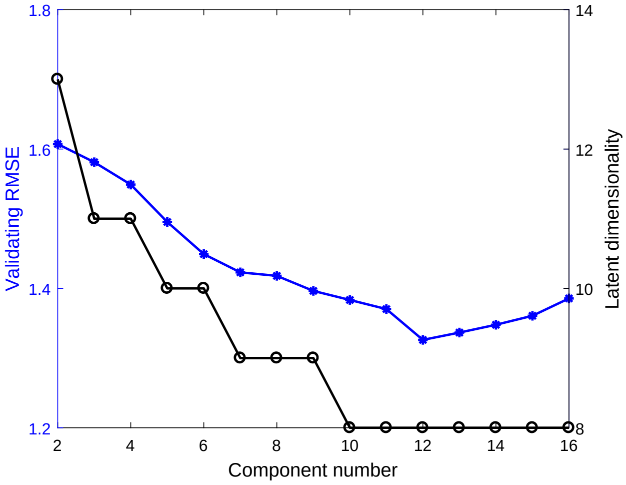
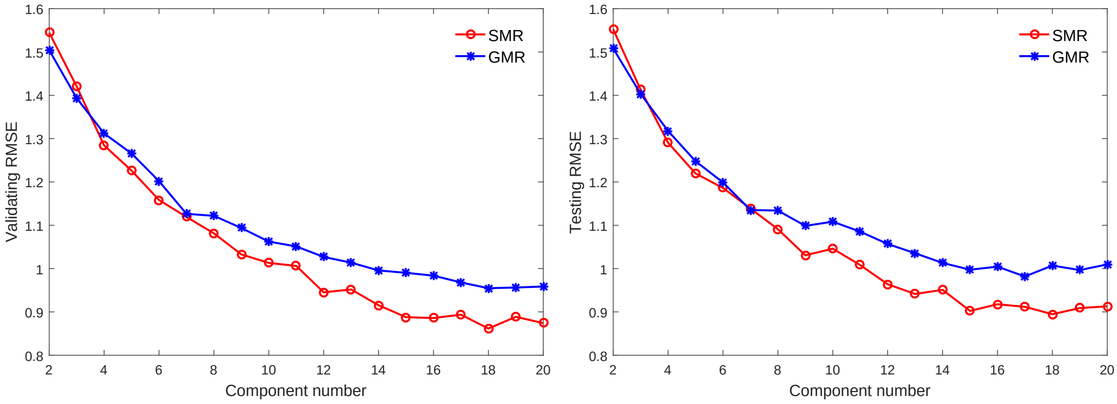


Figure 10. The validating RMSE and latent dimensionality based on Multi-DPLS.



(**a**) (**b**)

Figure 11. The RMSE on: (**a**) the validating datasets, (**b**) the testing datasets.

From Figure 11a,b we can recognize that: (1) the selected optimal values of K and initialized model parameters upon the validating dataset can basically embody the true generalization performance upon the testing dataset for the GMR and SMR-based soft sensors; (2) upon both the validating and testing dataset, although with small values of K (≤ 7), the performances of the two soft sensors are comparable, and the SMR-based soft sensor starts to show apparent predictive advantage over the GMR-based one as $K \geq 8$; (3) the number of components is much larger than the number of operating conditions, because each operating condition may consist of several modes. The underlying reason for this phenomenon is that for complex processes, one Student's-t distribution still does not model one operating condition well, and more Student's-t distributions are required for one operating condition; and (4) the number of components is mainly determined through the division of the spatial pattern of input and output variables rather than the number of input variables, so there is no relationship between the number of components and the number of input variables in the mixture models.

Predictions of the O_2 concentration by soft sensors based on the Multi-DPLS, GMR, and SMR are visualized in Figure 12, where their generalization abilities are also presented in terms of RMSE. As can be seen, the Multi-DPLS model has worst performance. Except for the ignorance of output information in the mode identification, the other reason is that the augmented input vector has high dimensionality (52 dimensions in the primary reformer), resulting in an exponentially increasing number of samples being required to acquire the correct estimations of probability distribution of each mode. In contrast, both the GMR and SMR-based soft sensors, which employ mixture component models, can significantly improve the prediction performance. Scatter plot comparisons among the Multi-DPLS, GMR and SMR presented in Figure 13 could provide more insights. It can be clearly seen that the predictions obtained by the Multi-DPLS are more scattered. However, predictions of soft sensors based on GMR and SMR lean much closer to the black diagonal line, indicating higher predictive accuracy. Moreover, since the SMR takes the robustness against outliers into consideration, the predictions obtained by SMR have tighter scatters around the black diagonal line, which demonstrates the advantages of SMR compared with GMR. The predictive RMSE on the testing dataset also demonstrate that the SMR-based soft sensor has stronger generalization ability than the GMR-based one. For further quantitative analyses, the determination coefficients for the Multi-DPLS, GMR, and SMR are also calculated as 0.7729, 0.8655, and 0.9233, respectively, from which the same conclusion can be drawn.

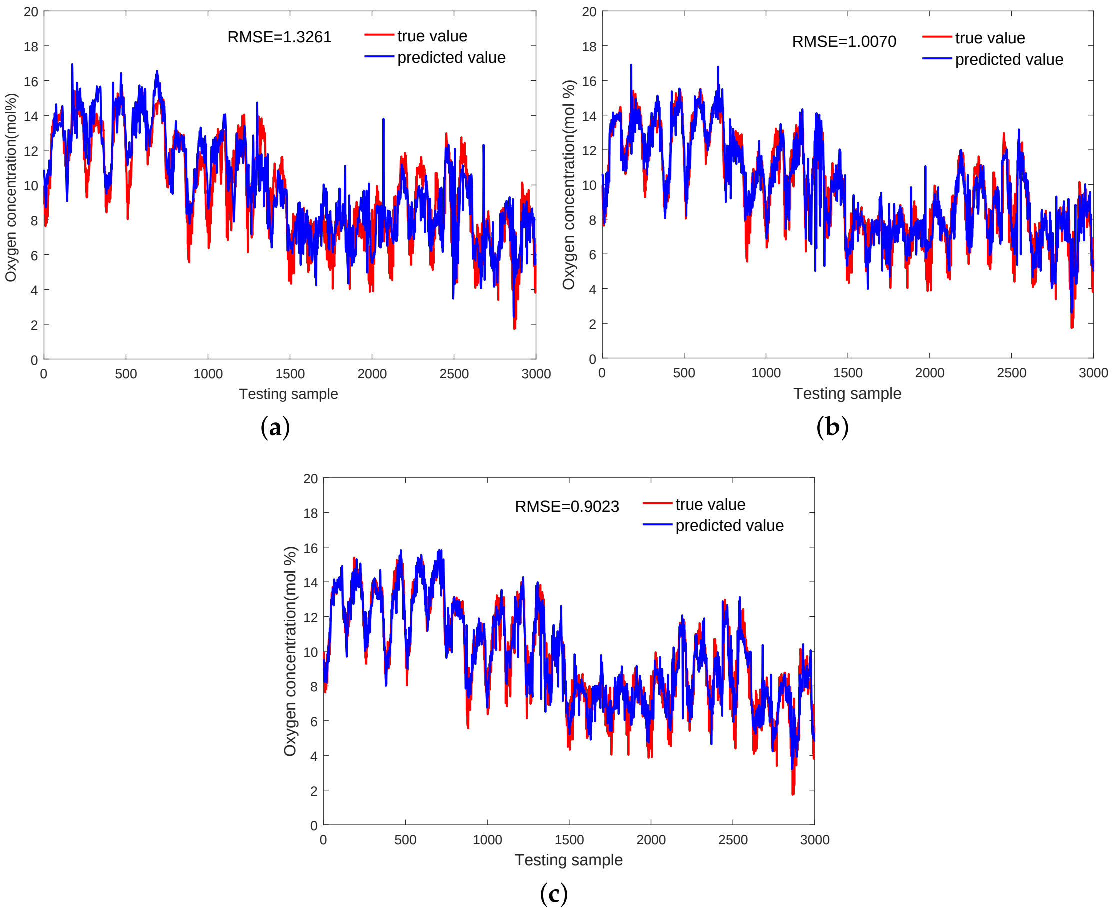


Figure 12. Predictions of the oxygen concentration achieved by: (**a**) Multi-DPLS, (**b**) GMR, (**c**) SMR.

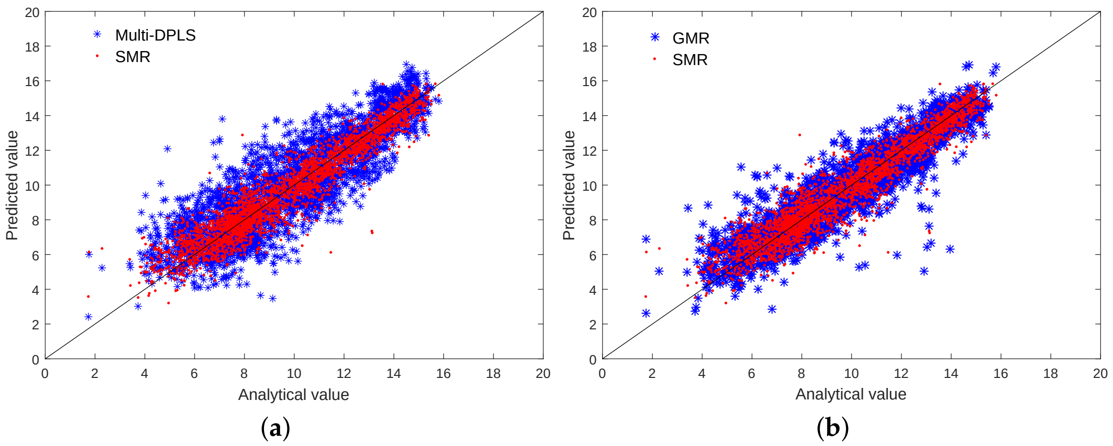


Figure 13. Scatter plot comparisons for estimating the concentration of O_2: (**a**) Multi-DPLS and SMR; (**b**) GMR and SMR.

The consumed time by these three methods in the primary reformer process are tabulated in Table 5, from which one can readily find that the Multi-DPLS requires more time to train the model because the dimensionality of the augmented input vector is very high. Although the SMR-based soft sensor is also a time-consuming method due to the dichotomy method, the prediction accuracy is much higher than Multi-DPLS.

Table 5. Average CPT (in second) consumed by various methods for the primary reformer process.

Time/Method	Multi-DPLS	GMR	SMR
CPT_{trn}	4.4908	0.847	3.002
CPT_{tst}	0.0731	0.009	0.005

As for the computational burden based on SMR, we can note that: (1) in the numerical example the input variables are two-dimensional, where the CPT_{trn} is much less than the primary reformer process of which the input variables are 13-dimensional. This is because as more variables are considered, the larger the size of the precision matrices (whose inversions are involved); (2) the computational burden depends on the number of mixing components, and the more mixing components, the more parameters needing to be learnt, which results in more time for model training; and (3) if the input variables are correlated, the non-diagonal elements of covariance are not equal to zero, leading to more time consumed in inverting the covariance matrix.

5. Conclusions

In this paper, with the aim of dealing with outliers when developing soft sensors for multimode industrial processes, we have proposed a robust modeling approach referred to as the Student's-*t* mixture model (SMR). Our novel contribution is twofold. First, a regressive model structure with finite mixture of Student's-*t* distributions has been designed, and the corresponding parameter-learning algorithm based on the EM algorithm has also been developed. Second, case studies have been conducted on both numerical and real-word industrial datasets to evaluate the performance of SMR. The results have demonstrated that SMR can handle multimode characteristics well and is more robust against outliers compared to some state-of-the-art methods.

In our future work, two challenging issues are taken into consideration: (1) how to complete the model-selection and parameter-learning tasks without traversing all candidate numbers of mixing components, and without the validating dataset; and (2) how to deal with the performance degradation of the soft sensor caused by time-variation factors. Our solution is to formulate an adaptive Bayesian SMR (BSMR), which randomizes model parameters (including the number of mixing components K) and updates the BSMR in a recursive fashion online.

Author Contributions: Software, J.W. and W.S.; Supervision, Z.S.; Writing—original draft, J.W.; Writing—review & editing, W.S. and Z.S.

Funding: This work was supported by the Natural Science Foundation of China (61703367), China Postdoctoral Science Foundation, and State Key Laboratory of Industrial Control Technology Open Project Funding (ICT1800384).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shao, W.; Tian, X. Semi-supervised Selective Ensemble Learning Based On Distance to Model for Nonlinear Soft Sensor Development. *Neurocomputing* **2017**, *222*, 91–104. [CrossRef]
2. Ge, Z.; Song, Z.; Gao, F. Review of Recent Research on Data-Based Process Monitoring. *Ind. Eng. Chem. Res.* **2013**, *10*, 3543–3562. [CrossRef]
3. Kruger, U.; Xie, L. Statistical Monitoring of Complex Multivariate Processes: With Applications in Industrial Process Control. *J. Qual. Technol.* **2012**, *45*, 118–120.
4. Ge, Z.; Song, Z.; Ding, S.; Huang, B. Data mining and analytics in the process industry: The role of machine learning. *IEEE Access* **2017**, *5*, 20590–20616. [CrossRef]
5. Ge, Z. Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemom. Intell. Lab. Syst.* **2017**, *171*, 16–25. [CrossRef]
6. Fortuna, L.; Graziani, S.; Xibilia, M. Soft sensors for product quality monitoring in debutanizer distillation columns. *Control Eng. Pract.* **2005**, *13*, 499–508. [CrossRef]

7. Sharmin, R.; Sundararaj, U.; Shah, S.; Vande Griend, L.; Sun, Y.-J. Inferential sensors for estimation of polymer quality parameters: Industrial application of a PLS-based soft sensor for a LDPE plant. *Chem. Eng. Sci.* **2006**, *61*, 6372–6384. [CrossRef]
8. Fortuna, L.; Graziani, S.; Rizzo, A.; Xibilia, M.G. *Soft Sensors for Monitoring and Control of Industrial Processes*; Spring Science & Business Media: Berlin/Heidelberg, Germany, 2007.
9. Kano, M.; Fujiwara, K. Virtual sensing technology in process industries: Trends and challenges revealed by recent industrial applications. *J. Chem. Eng. Jpn.* **2013**, *46*, 1–17. [CrossRef]
10. Kadlec, P.; Gabrys, B.; Strandt, S. Data-driven soft sensors in the process industry. *Comput. Chem. Eng.* **2009**, *33*, 795–814. [CrossRef]
11. Kadlec, P.; Grbić, R.; Gabrys, B. Review of adaptation mechanisms for data-driven soft sensors. *Comput. Chem. Eng.* **2011**, *35*, 1–24. [CrossRef]
12. Prasad, V.; Schley, M.; Russo, L.P.; Bequette, B.W. Product property and production rate control of styrene polymerization. *J. Process Control* **2002**, *12*, 353–372. [CrossRef]
13. Yuan, X.; Ge, Z.; Song, Z. Soft sensor model development in multiphase/multimode processes based on Gaussian mixture regression. *Chemom. Intell. Lab. Syst.* **2014**, *138*, 97–109. [CrossRef]
14. 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. [CrossRef]
15. Shao, W.; Tian, X. Adaptive soft sensor for quality prediction of chemical processes based on selective ensemble of local partial least squares models. *Chem. Eng. Res. Des.* **2015**, *95*, 113–132. [CrossRef]
16. Ge, Z.; Song, Z. Semisupervised Bayesian method for soft sensor modeling with unlabeled data samples. *AIChE J.* **2011**, *57*, 2109–2119. [CrossRef]
17. Gonzaga, J.C.B.; Meleiro, L.A.C.; Kiang, C.; Maciel Filho, R. ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process. *Comput. Chem. Eng.* **2009**, *33*, 43–49. [CrossRef]
18. Kaneko, H.; Funatsu, K. Database monitoring index for adaptive soft sensors and the application to industrial process. *AIChE J.* **2014**, *60*, 160–169. [CrossRef]
19. Kano, M.; Ogawa, M. The state of the art in chemical process control in Japan: Good practice and questionnaire survey. *J. Process Control* **2010**, *20*, 969–982. [CrossRef]
20. Souza, F.A.A.; Araújo, R. Mixture of partial least squares experts and application in prediction settings with multiple operating modes. *Chemom. Intell. Lab. Syst.* **2014**, *130*, 192–202. [CrossRef]
21. Zhu, J.; Ge, Z.; Song, Z. Variational Bayesian Gaussian mixture regression for soft sensing key variables in non-Gaussian industrial processes. *IEEE Trans. Control Syst. Technol.* **2017**, *25*, 1092–1099. [CrossRef]
22. Peel, D.; McLachlan, G.J. Robust mixture modeling using the *t* distribution. *Stat. Comput.* **2000**, *10*, 339–348. [CrossRef]
23. Chatzis, S.; Varvarigou, T. Robust fuzzy clustering using mixtures of Student's-*t* distributions. *Pattern Recognit. Lett.* **2008**, *29*, 1901–1905. [CrossRef]
24. Svensén, M.; Bishop, C.M. Robust Bayesian mixture modelling. *Neurocomputing* **2005**, *64*, 235–252. [CrossRef]
25. Gerogiannis, D.; Nikou, C.; Likas, A. The mixtures of Student's *t*-distributions as a robust framework for rigid registration. *Image Vis. Comput.* **2009**, *27*, 1285–1294. [CrossRef]
26. Zhang, H.; Wu, Q.M.J.; Nguyen, T.M. Image segmentation by a new weighted Student's *t*-mixture model. *IET Image Process.* **2013**, *7*, 240–251. [CrossRef]
27. Moghaddam, Z.; Piccardi, M. Robust density modelling using the student's t-distribution for human action recognition. In Proceedings of the 2011 18th IEEE International Conference on Image Processing (ICIP), Brussels, Belgium, 11–14 September 2011; pp. 3261–3264.
28. Nguyen, T.M.; Wu, Q.M.J. Robust student's-*t* mixture model with spatial constraints and its application in medical image segmentation. *IEEE Trans. Med. Imaging* **2012**, *31*, 103–116. [CrossRef] [PubMed]
29. Makantasis, K.; Doulamis, A.; Matsatsinis, N.F. Student-*t* background modeling for persons'fall detection through visual cues. In Proceedings of the 2012 13th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Dublin, Ireland, 23–25 May 2012.
30. Nguyen, T.; Wu, Q.M.J.; Zhang, H. Asymmetric mixture model with simultaneous feature selection and model detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 400–408. [CrossRef] [PubMed]
31. Bishop, C. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, NY, USA, 2006.

32. Liu, C.; Rubin, D.B. ML estimation of the *t* distribution using EM and its extensions, ECM and ECME. *Stat. Sin.* **1995**, *5*, 19–39.
33. Yao, L.; Ge, Z. Moving window adaptive soft sensor for state shifting process based on weighted supervised latent factor analysis. *Control Eng. Pract.* **2017**, *61*, 72–80. [CrossRef]
34. Peng, K.; Zhang, K.; You, B.; Dong, J. Quality-related prediction and monitoring of multi-mode processes using multiple PLS with application to an industrial hot strip mill. *Neurocomputing* **2015**, *168*, 1094–1103. [CrossRef]
35. Shao, W.; Tian, X.; Wang, P. Supervised local and non-local structure preserving projections with application to just-in-time learning for adaptive soft sensor. *Chin. J. Chem. Eng.* **2015**, *23*, 1925–1934. [CrossRef]
36. Zhu, J.; Ge, Z.; Song, Z. Multimode process data modeling: A Dirichlet process mixture model based Bayesian robust factor analyzer approach. *Chemom. Intell. Lab. Syst.* **2015**, *142*, 231–244. [CrossRef]