sensors

MDPI

Article

A KPI-Based Probabilistic Soft Sensor Development Approach that Maximizes the Coefficient of Determination

Yue Zhang [1], Xu Yang [1,*], Yuri A. W. Shardt [2], Jiarui Cui [1] and Chaonan Tong [1]

[1] Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China; s20160638@xs.ustb.edu.cn (Y.Z.); cuijiarui@ustb.edu.cn (J.C.); tcn@ies.ustb.edu.cn (C.T.)

[2] Department of Automation Engineering, Technical University of Ilmenau,
98684 Ilmenau, Thuringia, Germany; yuri.shardt@tu-ilmenau.de

* Correspondence: yangxu@ustb.edu.cn

Received: 28 July 2018; Accepted: 7 September 2018; Published: 12 September 2018

Abstract: Advanced technology for process monitoring and fault diagnosis is widely used in complex industrial processes. An important issue that needs to be considered is the ability to monitor key performance indicators (KPIs), which often cannot be measured sufficiently quickly or accurately. This paper proposes a data-driven approach based on maximizing the coefficient of determination for probabilistic soft sensor development when data are missing. Firstly, the problem of missing data in the training sample set is solved using the expectation maximization (EM) algorithm. Then, by maximizing the coefficient of determination, a probability model between secondary variables and the KPIs is developed. Finally, a Gaussian mixture model (GMM) is used to estimate the joint probability distribution in the probabilistic soft sensor model, whose parameters are estimated using the EM algorithm. An experimental case study on the alumina concentration in the aluminum electrolysis industry is investigated to demonstrate the advantages and the performance of the proposed approach.

Keywords: soft sensor; coefficient of determination maximization strategy; expectation maximization (EM) algorithm; Gaussian mixture model (GMM); alumina concentration

1. Introduction

With the increasing demands placed on industry, requiring a decrease in the defective rate of products, better economic efficiency, and improved safety, there has been a growing demand to develop and implement approaches that can improve the overall control strategy [1]. The first issue that needs to be solved is achieving accurate and real-time estimation of key performance indicators (KPIs) [2]. The difficulty is that these KPIs are usually not easy to measure, or the measurement has significant time delay. Even if some KPIs are measurable, due to the complexity and nonlinearity of modern industrial systems and their complex working conditions, the KPIs may be extremely unreliable [3]. One way to solve the above problems is to develop a soft sensor, which seeks to select a group of easier-to-measure secondary variables that are correlated with the required primary variables (i.e., KPIs in this paper), so that the system is capable of providing process information as often as necessary for control [4,5]. In the development of a successful soft sensor, a good process model is required. The process models can be divided into two major categories: first principles models and data-driven models [6,7]. Although it is desirable to apply mass and energy balances to build a complete first principles model, lack of process knowledge, plant–model mismatch, and nonlinear characteristics limit the applicability of such an approach to the simplest processes. As an alternative, data-driven soft sensors are developed from historical data without necessarily considering any outside process knowledge. Data-driven soft sensors, which solely use available process data to develop a model of the process, have recently attracted considerable attention and have been successfully applied in many fields [8], such as fault detection (FD) and process monitoring, that are important for many industrial processes. Serdio [9] introduced an improved fault detection approach based on residual signals extracted online from system models identified by high-dimensional measurements provided by the multisensor network. The data-driven system identification model can also be combined using multivariate orthogonal space transformations and vectorized time-series models to achieve enhanced residual-based fault detection in condition monitoring systems equipped with a multisensor network [10]. Shardt [11] proposed a data-driven design of a diagnostic-observer-based process monitoring method, which was extended to include the ability to detect changes given infrequent KPI measurements. Yan [12] and Gabrys [13] introduced the most popular data-driven soft sensor modelling techniques, as well as discussing some issues in soft sensor development and maintenance and their possible solutions. Data-driven methods can be divided into three categories: models based on statistical analysis, models based on statistical learning theory [14], and models based on artificial intelligence [15].

Of interest for this paper are models developed using statistical methods to extract the relevant information from the large amounts of industrial data that are produced by the complex processes. Statistical methods have been developed that can handle such large datasets and develop useful models. Common methods include principal component analysis (PCA) [16] and partial least squares (PLS) [17]. PCA is a powerful tool for data compression and information extraction that can simplify the model structure and improve the speed of operations. However, PCA can only deal with the correlations between vectors in the same matrix. To overcome this limitation, PLS was developed as an approach that models the correlation between independent variables and dependent variables. Since PLS only applies to linear systems or weakly nonlinear systems, many nonlinear PLS algorithms have been developed to handle nonlinear systems. The neural-network-based PLS algorithm [18] uses the nonlinear processing capability of a neural network to describe the relationship between variables. However, the determination of the network structure and the selection of network training algorithms are difficult problems. In addition, if there are too many datapoints, the model structure will be very complex and the accuracy will be difficult to guarantee.

On the other hand, considering that data-driven modeling methods use historical data for training, this raises the question of how to handle missing data. Along with issues such as the reliability of sensors and multirate sampling, missing data is common in practical industry process [19,20]. For example, in the aluminum electrolysis process, the alumina concentration is usually obtained manually by laboratory staff. Considering human factors and chemical examination equipment reliability, data loss occurs from time to time. In this case, this type of measurement has different effects on the soft sensor modeling process and state estimation performance. Therefore, in order to make the soft sensor more suitable for practical, complex industrial processes, the missing data problem needs to be taken seriously. Compared with the direct deletion of missing data, the data interpolation method [21] is better able to restore the real situation. Currently, data interpolation methods include the mean substitution method, the regression interpolation method, and the expectation maximization (EM) algorithm. Of these, the mean substitution method can cause biased estimates, and the regression interpolation method is built based on a complete data set, where the linear relationship between the variables with missing values and other variables is necessary, which, in many cases, cannot be satisfied. In fact, the EM algorithm has good practical value as an iterative algorithm for simplifying the maximum likelihood estimation when dealing with missing data in sample sets [22].

Recently, in order to evaluate the accuracy of the model output, the coefficient of determination approach has been considered. The coefficient of determination is the measurement of how well the regression model fits the data [23]. Feng [24] introduced the coefficient of determination as a criterion for comparing the best-wavelength partial least squares regression (PLSR) model with the full-wavelength model. Boyaci [25] used the coefficient of determination to evaluate the adulteration rate of coffee beans, thus ensuring coffee quality. However, these applications only consider the coefficient of determination as an evaluation index without applying it for the modeling process. In general, the coefficient of determination is a criterion that can evaluate the quality of a model and has a concise structure, so it is appropriate to apply it to the soft sensor development process to establish a simpler and more accurate model for complex industrial process.

Therefore, this paper develops a KPI-based soft sensor model with simple structure and high accuracy, using the coefficient of determination method, which also solves the missing data issue using the EM algorithm.

2. Background

2.1. The Gaussian Mixture Model

As a flexible and efficient tool for probabilistic data models, a Gaussian mixture model (GMM) can be used to define any complex probability distribution function and is, therefore widely used in many statistical data modelling applications. In this paper, GMM is used to approximate the joint probability distribution in the soft sensor probability model. The reason for introducing GMM is that, theoretically, any probability distribution can be approximated using the joint weighted Gaussian distribution [26].

If x represents a multidimensional random variable, then the joint probability distribution of the GMM is expressed as

$$p(x\,|\Theta\,) = \sum_{l=1}^{M} \alpha_l p_l(x|\theta_l\,) \tag{1}$$

where α_l is the mixing coefficient, which represents the prior probability of each mixed component; M is the number of mixed components; and $\sum_{l=1}^{M} \alpha_l = 1$. $\Theta = (\theta_1, \theta_2, \cdots, \theta_M)$ is the parameter vector of each mixed component, and each Gaussian probability density function $p_l(x)$ is determined by the parameter $\theta_l = (\mu_l, \Sigma_l)$, where μ_l is the mean and Σ_l is the covariance matrix. The GMM parameters α_l, μ_l, and Σ_l ($l = 1, 2, \ldots, M$) are estimated using the EM algorithm.

2.2. The Expectation Maximization Algorithm

The EM algorithm is a maximum likelihood estimation method for solving model distribution parameters from "incomplete data" and was first introduced in [27]. Each iteration of the algorithm involves two steps, called the expectation step (E-step) and the maximization step (M-step).

2.2.1. E-Step

Given the observation data set X and the current parameters $\Gamma^{(i)}$, the expectation of the log-likelihood function is called the Q-function which can be written as

$$Q(\Gamma, \Gamma^{(i)}) = E\left[\log p(\,X, |\Gamma)\middle| X, \Gamma^{(i)}\right] \tag{2}$$

where γ can represent missing data due to observational conditions and other reasons, and can also refer to hidden variables. Since the direct optimization of the likelihood function is usually very difficult, the relationship between X, Γ, and γ can be established by introducing an additional variable γ to achieve the purpose of simplifying the likelihood function.

2.2.2. M-Step

A new parameter $\Gamma^{(i+1)}$ is calculated by maximizing $Q(\Gamma, \Gamma^{(i)})$ which was obtained from the E-step; that is,

$$\Gamma^{(i+1)} = \underset{\Gamma}{\operatorname{argmax}} Q(\Gamma, \Gamma^{(i)}) \,. \tag{3}$$

The iteration between the E- and M-steps continues until the elements of Γ are less than a given value.

2.3. The Coefficient of Determination

Analysis of variance is an approach for determining the significance and validity of a regression model using variances obtained from the data and model. The coefficient of determination is an analysis of variance approach that seeks to decompose the total variability in the data into various orthogonal components that can then be independently analyzed [23]. For the purposes of analyzing the regression, let the total sum of squares, denoted by TSS, be defined as

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{4}$$

where the real data set is represented as y = <$y_1, y_2, \ldots, y_n$> and $\overline{y}$ refers to the average of y_i. Let the sum of squares due to regression, SSR, be defined as

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \tag{5}$$

where $\hat{y}_i$ denotes the predicted value of the regression model for y_i. The coefficient of determination R^2 represents the ratio of SSR to TSS, that is,

$$R^2 = \frac{SSR}{TSS} \,. \tag{6}$$

Let the sum of squares due to the error, SSE, be defined as

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{7}$$

It can be proved that TSS = SSR + SSE [23,28], so R^2 can also be expressed as

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{\sum\limits_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n} (y_i - \overline{y})^2} \,. \tag{8}$$

3. Development of the Probabilistic Soft Sensor Model

In this section, in order to obtain more accurate KPI estimates, a soft sensor development approach based on maximizing the coefficient of determination is proposed. In addition, the problem of missing data in the training sample set is also considered. In order to more clearly describe the soft sensor development process, Figure 1 shows the modeling flow chart.

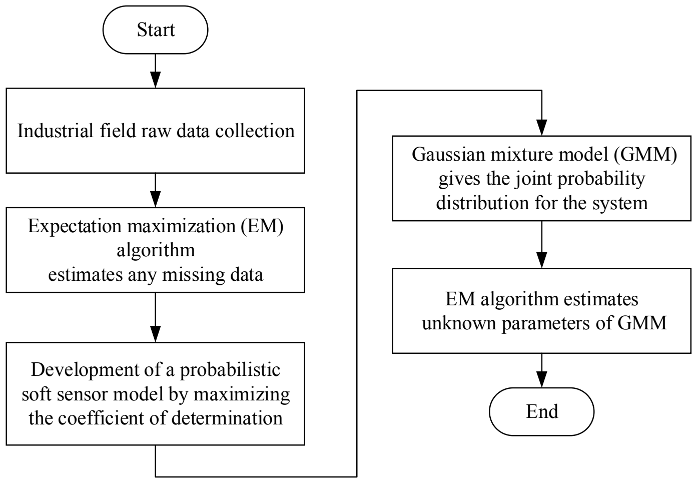

Start
Industrial field raw data collection
Expectation maximization (EM) algorithm estimates any missing data
Development of a probabilistic soft sensor model by maximizing the coefficient of determination
Gaussian mixture model (GMM) gives the joint probability distribution for the system
EM algorithm estimates unknown parameters of GMM
End

Figure 1. The flow chart of soft sensor development process.

3.1. *EM Algorithm Handing Missing Data*

Let $X_1, X_2, \ldots, X_n$ be a random sample from a p-variate normal population, where $X_j = (x_{j1}, x_{j2}, \ldots, x_{j\mathrm{p}})$, $1 \leq j \leq n$, so the training sample set X can be written as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{bmatrix} x_{11}, & x_{12}, & \cdots, & x_{1p} \\ x_{21}, & x_{22}, & \cdots, & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1}, & x_{n2}, & \cdots, & x_{np} \end{bmatrix}. \tag{9}$$

The basic steps for processing missing data using the EM algorithm are given in [29].

3.1.1. E-Step: Prediction

For each sample X_j containing missing values, $X_j = (m_j, a_j)$, where m_j is the missing value and a_j is the available values. Given the population mean and variance, $\overset{\sim}{\mu}^{i}$ and $\overset{\sim}{\Sigma}^{i}$, from the ith iteration and a_j, we use the expectation of the conditional normal distribution of m_j as the estimate of the missing value. The $(i + 1)$th iteration is

$$\widetilde{m}_j^{i+1} = E(m_j|a_j, \widetilde{\mu}^i, \widetilde{\sum}^i) = \widetilde{\mu}_m^i + \widetilde{\sum}_{ma}^i (\widetilde{\sum}_{aa}^i)^{-1} (a_j - \widetilde{\mu}_a^i) \tag{10}$$

where $\overset{\sim}{\mu}^{i}$ is a $p \times 1$ matrix defined as $\widetilde{\mu}^i = \left[\widetilde{\mu}_m^i, \widetilde{\mu}_a^i\right]'$, $\widetilde{\mu}_m^i$ is the mean of the missing part, and $\widetilde{\mu}_a^i$ is the mean of the available part. In addition, $\overset{\sim}{\Sigma}^{i}$ can be written as

$$\widetilde{\sum}^i = \begin{bmatrix} \widetilde{\sum}_{mm}^i \widetilde{\sum}_{ma}^i \\ \widetilde{\sum}_{am}^i \widetilde{\sum}_{aa}^i \end{bmatrix}. \tag{11}$$

3.1.2. M-Step: Estimation

We compute the maximum likelihood estimates as follows:

$$\widetilde{\mu}^{i+1} = \overline{X}^{i+1} \tag{12}$$

$$\widetilde{\sum}^{i+1} = \frac{(n-1)S^{i+1}}{n} \tag{13}$$

where $\overline{X}^{i+1}$ is the mean of the samples and S^{i+1} is the sample standard deviation, and they are all sufficient statistics. For a normal population, the importance of sufficient statistics is that the total information about μ and Σ in the data matrix X is contained in $\overline{X}$ and S, regardless of the sample size n. By transforming $\overline{X}$ and S, two new sufficient statistics T_1 and T_2 [29], given by

$$T_1 = n\overline{X} \tag{14}$$

$$T_2 = (n-1)S + n\overline{X}\overline{X}' \tag{15}$$

are obtained. Combining Equations (14) and (15) with Equations (12) and (13) gives

$$\widetilde{\mu}^{i+1} = \frac{T_1^{i+1}}{n} \tag{16}$$

$$\widetilde{\sum}^{i+1} = \frac{1}{n}{T_2}^{i+1} - \widetilde{\mu}^{i+1}(\widetilde{\mu}^{i+1})' \tag{17}$$

where

$$\widetilde{m_j m_j{}'}^{i+1} = E(m_j m_j{}' | a_j, \widetilde{\mu}^i, \widetilde{\sum}^i) = \widetilde{\sum}^i{}_{mm} - \widetilde{\sum}^i{}_{ma} (\widetilde{\sum}^i{}_{aa})^{-1} \widetilde{\sum}^i{}_{am} + \widetilde{m}_j^{i+1} \left(\widetilde{m}_j^{i+1}\right)' \tag{18}$$

$$\widetilde{m_j a_j{}'}^{i+1} = E(m_j a_j{}' | a_j, \widetilde{\mu}^i, \widetilde{\sum}^i) = \widetilde{m}_j^{i+1} \left(a_j\right)'. \tag{19}$$

The iteration between the E- and M-steps continues until the elements of $\widetilde{\mu}$ and $\widetilde{\Sigma}$ are less than a given value. Therefore, the iteration result $\widetilde{m}$ is the optimal substitution for the missing values, resulting in a complete training sample set X.

3.2. Soft Sensor Development Approach Based on the Coefficient of Determination Maximization Strategy

For the complete training sample set X obtained from Section 3.1, which can be written as

$$X = \begin{bmatrix} x_{11}, & x_{12}, & \cdots, & x_{1p} \\ x_{21}, & x_{22}, & \cdots, & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1}, & x_{n2}, & \cdots, & x_{np} \end{bmatrix} \tag{20}$$

let $\left(x_1, x_2, \cdots x_{p-1}\right)$ denote the secondary variables, and x_p denote the KPI. Our objective is to estimate x_p from $\left(x_1, x_2, \cdots x_{p-1}\right)$.

R^2 measures the fraction of the total variance in the model explained by the regression with the given variables [23]. The range of R^2 is [0,1]. Let x_p be the y mentioned in Section 2.3. Then, the coefficient of determination is

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n} \left(x_{ip} - \hat{x}_{ip}\right)^2}{\sum\limits_{i=1}^{n} \left(x_{ip} - \overline{x}_p\right)^2}. \tag{21}$$

If the secondary variables in the soft sensor model do not account for the variance of x_p, the estimate of x_{ip}, denoted $\hat{x}_{ip}$, is exactly equal to the sample mean of x_{ip}, denoted $\overline{x}_{ip}$. In this case, SSR is 0 and SSE equal to TSS, so R^2 = 0. On the other hand, if $\left(x_{i1}, x_{i2}, \cdots x_{i(p-1)}\right)$ fully explains the variance of x_{ip}, for i = 1, 2, … , n, it follows that $x_{ip} = \overline{x}_{ip}$, i.e., each error is zero and SSR = TSS, so R^2 = 1. In general, R^2 does not take the extreme values 0 or 1, but instead takes a certain value between the two [28]. For the case where the number of variables, p, is much smaller than the sample number n, the closer R^2 is to 1, the better the model. Therefore, when the model for the KPI maximizes R^2, it becomes the best estimate of the KPI, that is,

$$1-\frac{\sum\limits_{i=1}^{n}\left(x_{ip}-\widetilde{x}_{ip}\right)^2}{\sum\limits_{i=1}^{n}\left(x_{ip}-\overline{x}_p\right)^2}=\max\left[1-\frac{\sum\limits_{i=1}^{n}\left(x_{ip}-K_i\right)^2}{\sum\limits_{i=1}^{n}\left(x_{ip}-\overline{x}_p\right)^2}\right] \tag{22}$$

where $\widetilde{x}_{ip}$ is the best estimate of x_{ip}, and K_i represents all possible estimates of x_{ip}. Simplifying the above equation gives

$$\frac{\sum\limits_{i=1}^{n}\left(x_{ip}-\widetilde{x}_{ip}\right)^2}{\sum\limits_{i=1}^{n}\left(x_{ip}-\overline{x}_p\right)^2}=\min\left[\frac{\sum\limits_{i=1}^{n}\left(x_{ip}-K_i\right)^2}{\sum\limits_{i=1}^{n}\left(x_{ip}-\overline{x}_p\right)^2}\right] \tag{23}$$

where x_{ip} and $\overline{x}_p$ are both computed values. Equation (23) can then be written as

$$\sum_{i=1}^{n}\left(x_{ip}-\widetilde{x}_{ip}\right)^2=\min\left[\sum_{i=1}^{n}\left(x_{ip}-K_i\right)^2\right]. \tag{24}$$

Multiplying Equation (24) on both sides by n^{-1} gives

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_{ip}-\widetilde{x}_{ip}\right)^2=\min\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_{ip}-K_i\right)^2\right]. \tag{25}$$

Considering that the mathematical expectation of a discrete random variable is

$$E(x)=\sum_i x_i p_i \tag{26}$$

where x_i represents the ith value of the random variable x and p_i represents its probability, Equation (26) can be expressed as

$$E\left\{\|x_p-\widetilde{x}_p\|^2\right\}=\min E\left\{\|x_p-K\|^2\right\} \tag{27}$$

where K denotes all possible estimates of the KPI x_p, and $\widetilde{x}_p$ represents the best estimate of the KPI when the coefficient of determination R^2 is maximized. Since x_p is derived from the soft sensor models and secondary variables, the above equation can be written as

$$\widetilde{x}_p=\underset{K}{\operatorname{argmin}}E\left[\|x_p-K\|^2\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right]. \tag{28}$$

In order to establish a more direct connection between $\widetilde{x}_p$ and $(x_{i1}, x_{i2}, \ldots, x_{i(p-1)})$, the left-hand side of Equation (28) will be simplified further. Firstly, it can be noted that K does not have an impact on the simplification, that is,

$$\begin{aligned}
&E\left[\|x_p-K\|^2\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right]\\
=\;&E\left[\|x_p-E\left(x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right)+E\left(x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right)-K\|^2\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right]\\
=\;&E\left[\|x_p-E\left(x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right)\|^2\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right]+E\left[\|E\left(x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right)-K\|^2\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right]\\
&+E\left[\left[x_p-E\left(x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right)\right]^T\left[E\left(x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right)-K\right]\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right]\\
&+E\left[\left[E\left(x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right)-K\right]^T\left[x_p-E\left(x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right)\right]\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right]
\end{aligned} \tag{29}$$

In order to minimize the above equation, the following should hold:

$$K=E\left[x_p\middle|\left(x_1,x_2,\cdots x_{p-1}\right)\right] \tag{30}$$

which can be rewritten as

$$\widetilde{x}_p = E\left[x_p\middle|\left(x_1, x_2, \cdots x_{p-1}\right)\right] . \tag{31}$$

Furthermore, $E\left[x_p\middle|\left(x_1, x_2, \cdots x_{p-1}\right)\right]$ can be expanded according to the definition of expectation, giving

$$\begin{aligned}\widetilde{x}_p &= E\left[x_p\middle|\left(x_1, x_2, \cdots x_{p-1}\right)\right] \\ &= \int x_p p\left[x_p\middle|\left(x_1, x_2, \cdots x_{p-1}\right)\right] dx_p . \\ &= \int x_p \frac{p(x_1,x_2,\cdots x_{p-1},x_p)}{p(x_1,x_2,\cdots x_{p-1})} dx_p\end{aligned} \tag{32}$$

Thus, this establishes the basic framework of the probabilistic soft sensor model with KPI optimal estimation.

The next part is to solve the joint probability distribution in the model.

In this paper, GMM is used to approximate the joint probability distribution. Let $p(x_e) = p(x_1, x_2, \cdots x_{p-1})$; that is,

$$p(x_e) = \sum_{j=1}^{M} \alpha_j p\left(x_{je}\middle|\theta_j\right) \tag{33}$$

$$p\left(x_e, x_p\right) = \sum_{l=1}^{M} \alpha_l p\left(x_{le}, x_{lp}\middle|\theta_l\right) . \tag{34}$$

In order to deduce the specific representation of the KPI optimal estimation $\widetilde{x}_p$ under the proposed probabilistic soft sensor model, we first introduce Lemma 1.

Lemma 1. [30] *Let G(x; μ, Σ) be a multidimensional normal density function with mean μ and covariance matrix Σ. Let* $x^T = \left(x_1^T, x_2^T\right)$, $\mu^T = \left(\mu_1^T, \mu_2^T\right)$, *and* $\Sigma = \begin{bmatrix} \Sigma_{11}\Sigma_{12} \\ \Sigma_{21}\Sigma_{22} \end{bmatrix}$*; then, the joint probability density is*

$$p(x) = G(x_1; \mu_1, \Sigma_{11}) G\left(x_2; \mu_{x_2|x_1}, \Sigma_{x_2|x_1}\right) \tag{35}$$

where

$$\mu_{x_2|x_1} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(\mu_1 - x_1) \tag{36}$$

$$\Sigma_{x_2|x_1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} . \tag{37}$$

Proof. The details of the proof can be found in [30].

Using Lemma 1, it follows that

$$\begin{aligned} & p\left(x_{le}, x_{lp}\right) \\ = \; & G(x_l; \mu_l, \Sigma_l) \\ = \; & G(x_{le}; \mu_{le}, \Sigma_{lee}) G\left(x_{lp}; \mu_{lp|e}, \Sigma_{lp|e}\right)\end{aligned} \tag{38}$$

where $\mu_l = \left(\mu_{le}^T, \mu_{lp}^T\right)$ and $\Sigma_l = \begin{bmatrix} \Sigma_{lee}\Sigma_{lep} \\ \Sigma_{lpe}\Sigma_{lpp} \end{bmatrix}$. Therefore, Equations (33) and (34) can be written as

$$p(x_e) = \sum_{j=1}^{M} \alpha_j G\left(x_{je}; \mu_{je}, \Sigma_{jee}\right) \tag{39}$$

$$p\left(x_e, x_p\right) = \sum_{l=1}^{M} \alpha_l G(x_{le}; \mu_{le}, \Sigma_{lee}) G\left(x_{lp}; \mu_{lp|e}, \Sigma_{lp|e}\right) . \tag{40}$$

Substituting Equations (39) and (40) into Equation (32) gives

$$\begin{aligned}\widetilde{x}_p &= \int x_p \frac{p(x_e,x_p)}{p(x_e)} dx_p \\ &= \int x_p \frac{\sum\limits_{l=1}^{M} \alpha_l G(x_{le};\mu_{le},\Sigma_{lee}) G\left(x_{lp};\mu_{lp|e},\Sigma_{lp|e}\right)}{\sum\limits_{j=1}^{M} \alpha_j G\left(x_{je};\mu_{je},\Sigma_{jee}\right)} dx_p\ .\end{aligned} \tag{41}$$

Extracting the sum in the numerator to outside the integral gives

$$\widetilde{x}_p = \sum_{l=1}^{M} \int x_p \frac{\alpha_l G(x_{le};\mu_{le},\Sigma_{lee}) G\left(x_{lp};\mu_{lp|e},\Sigma_{lp|e}\right)}{\sum\limits_{j=1}^{M} \alpha_j G\left(x_{je};\mu_{je},\Sigma_{jee}\right)} dx_p\ . \tag{42}$$

In order to make the derivation more concise, the positions of some factors in the integral are changed as follows:

$$\begin{aligned}\widetilde{x}_p &= \sum_{l=1}^{M} \int \frac{\alpha_l G(x_{le};\mu_{le},\Sigma_{lee})}{\sum\limits_{j=1}^{M} \alpha_j G\left(x_{je};\mu_{je},\Sigma_{jee}\right)} x_p G\left(x_{lp};\mu_{lp|e},\Sigma_{lp|e}\right) dx_p \\ &= \sum_{l=1}^{M} \frac{\alpha_l G(x_{le};\mu_{le},\Sigma_{lee})}{\sum\limits_{j=1}^{M} \alpha_j G\left(x_{je};\mu_{je},\Sigma_{jee}\right)} \int x_p G\left(x_{lp};\mu_{lp|e},\Sigma_{lp|e}\right) dx_p\ .\end{aligned} \tag{43}$$

When the integral part is the conditional expectation, the above equation can be simplified to

$$\widetilde{x}_p = \sum_{l=1}^{M} \frac{\alpha_l G(x_{le};\mu_{le},\Sigma_{lee})}{\sum\limits_{j=1}^{M} \alpha_j G\left(x_{je};\mu_{je},\Sigma_{jee}\right)} \mu_{lp|e}\ . \tag{44}$$

Therefore, the detailed soft sensor model expression of the KPI optimal estimation is obtained.

In this paper, unknown parameters in the model are estimated using the EM algorithm. The iterative equations of the EM algorithm for estimating the GMM parameters are [31]

$$\mu_l^{(i+1)} = \frac{\sum\limits_{j=1}^{n} \gamma_{jl}^{(i+1)} X_j}{\sum\limits_{j=1}^{n} \gamma_{jl}^{(i+1)}},\ \Sigma_l^{(i+1)} = \frac{\sum\limits_{j=1}^{n} \gamma_{jl}^{(i+1)} \left(X_j - \mu^{(i)}\right)^2}{\sum\limits_{j=1}^{n} \gamma_{jl}^{(i+1)}},\ \alpha_l^{(i+1)} = \frac{\sum\limits_{j=1}^{n} \gamma_{jl}^{(i+1)}}{n} \tag{45}$$

where γ_{jl} represents the responsivity of the mixed component l on the training sample data X_j. It can be written as

$$\gamma_{jl}^{(i+1)} = \frac{\alpha_l p(X_j|\theta_l)}{\sum\limits_{l=1}^{M} \alpha_l p(X_j|\theta_l)}\ . \tag{46}$$

Consequently, the above steps give the GMM parameters, and the KPI optimal estimate $\widetilde{x}_p$ follows.

4. Case Study

In this section, the effectiveness and feasibility of the proposed soft sensor model approach based on maximizing the coefficient of determination are evaluated through an industrial aluminum electrolytic production process. To show the advantages of the probabilistic soft sensor framework, the estimations are compared with the real values. For performance evaluation, the root-mean-squared error (RMSE) index is used.

4.1. Soft Sensor Development for Industrial Aluminum Electrolytic Process

Aluminum is widely used in construction and electrical industries [32]. The main method currently chosen for smelting aluminum plants is the cryolite–alumina molten salt electrolysis process, in which the electrochemical reaction process takes place in an electrolytic cell. Figure 2 shows the internal structure of the electrolytic cell.

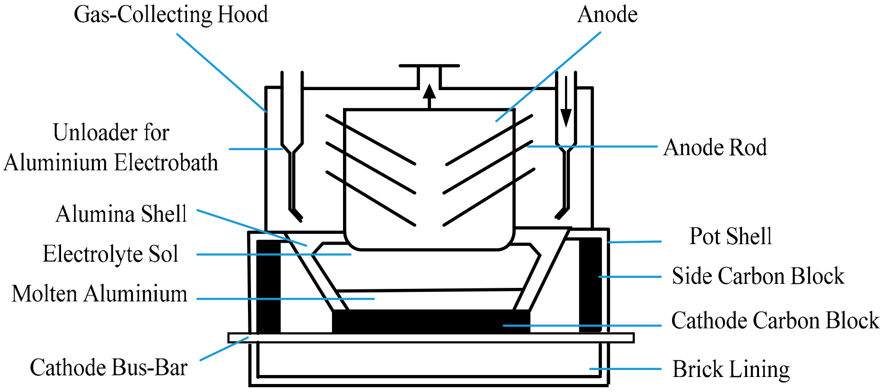


Figure 2. The internal structure of the aluminum electrolytic cell.

Molten cryolite is a solvent in which aluminum oxide is dissolved as a solute, forming a melt with good electrical conductivity. Carbon materials are used as cathodes and anodes, and a direct current is passed through them. The thermal energy of the direct current is used to melt the cryolite and maintain a constant electrolysis temperature. Furthermore, the electrochemical reaction occurs between the two electrodes, where the product at the cathode is aluminum liquid, and carbon dioxide and other gases are generated at the anode. The chemical reaction of the electrolytic process is

$$2Al_2O_3 + 3C \rightarrow 4Al + 3CO_{2\uparrow}.$$

The chemical reaction can produce gases other than carbon dioxide and carbon monoxide, as well as fluorocarbon gases. The gas purifying device uses alumina and fluorine generated in the mixed gas to produce fluorinated alumina, and the fluorinated alumina is then recycled to the electrolytic cell for chemical reaction. Figure 3 shows the process flow diagram of the aluminum electrolysis process.

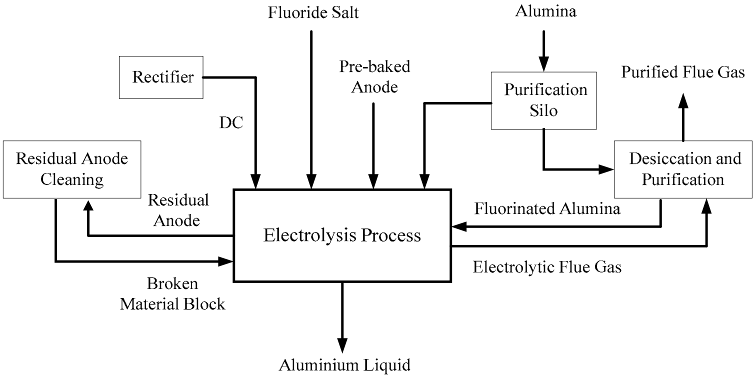


Figure 3. The process flow diagram of the aluminum electrolysis process.

The main control goal of the aluminum electrolysis process is to keep the alumina concentration in the electrolysis cell stable within a certain range, preferably between 1.5% and 3.5% [33]. The control of alumina concentration relates to energy consumption and economic benefits of the aluminum electrolytic production process. On one hand, when the alumina concentration is too low, an additional chemical reaction occurs at the anode, which can easily cause a sudden rise in the cell voltage and the energy balance of the cell is destroyed. On the other hand, when the concentration reaches saturation, if the feeder continues to add alumina at the time, the raw material will be deposited at the bottom of the cell, so that the resistance increases and the current efficiency becomes low. Therefore, it is necessary to keep the alumina concentration in the proper range.

In soft sensor development for the aluminum electrolytic process, the measurable variables, the voltage x_1 between the two electrodes obtained by the first voltage measuring instrument; the anode conductor current x_2; the voltage x_3 between the two electrodes obtained by the second voltage measuring instrument; and the alumina concentration x_4 provided by an electrochemical analyzer, were selected as the secondary variables. The interelectrode voltage refers to the voltage between the anode guide and the corresponding cathode steel bar. The alumina concentration y provided by the laboratory is the primary variable for the model. Figure 4 shows a diagram of the process measurement system.

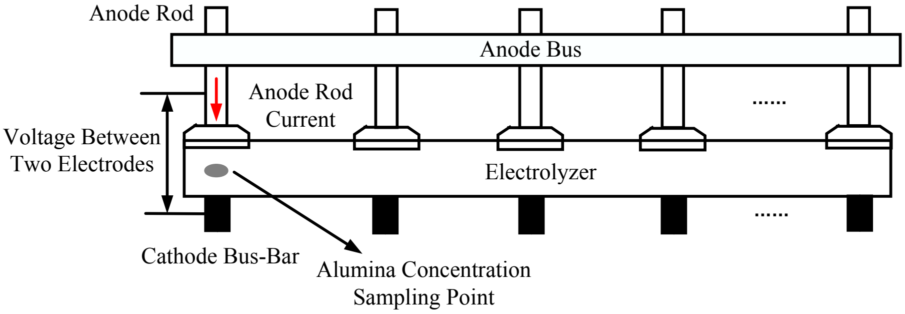


Figure 4. Schematic diagram of the variable collection system.

The variables $x_1(k)$, $x_2(k)$, $x_3(k)$, $x_4(k)$, and $y(k)$ form the joint probability distribution

$$p(x(k)) = p(x_1(k), x_2(k), x_3(k), x_4(k), y(k))\ . \tag{47}$$

The soft sensor was then developed according to the process described in Section 3 of this paper. It is assumed that $M = 2$.

4.2. Experimental Results

4.2.1. EM Algorithm and Missing Values

We took 600 complete data groups from the training sample set, and deleted 10%, 20%, or 30% of the alumina concentration variable data. Then, the mean substitution method, the regression interpolation method, and the EM algorithm were used to process the sample set with missing values. Tables 1–3 show the mean and RMSE of the alumina concentration sample set for the three method simulations for missing ratios of 10%, 20%, and 30%.

Table 1. Comparison of three data interpolation methods for a 10% missing rate.

	Mean Substitution Method	Regression Interpolation Method	EM Algorithm	Real Value
Mean	2.4133	2.4225	2.4225	2.4259
RMSE	0.0867	0.4209	0.0698	0

Table 2. Comparison of three data interpolation methods for a 20% missing rate.

	Mean Substitution Method	Regression Interpolation Method	EM Algorithm	Real Value
Mean	2.4139	2.4217	2.4215	2.4259
RMSE	0.1451	0.4075	0.1361	0

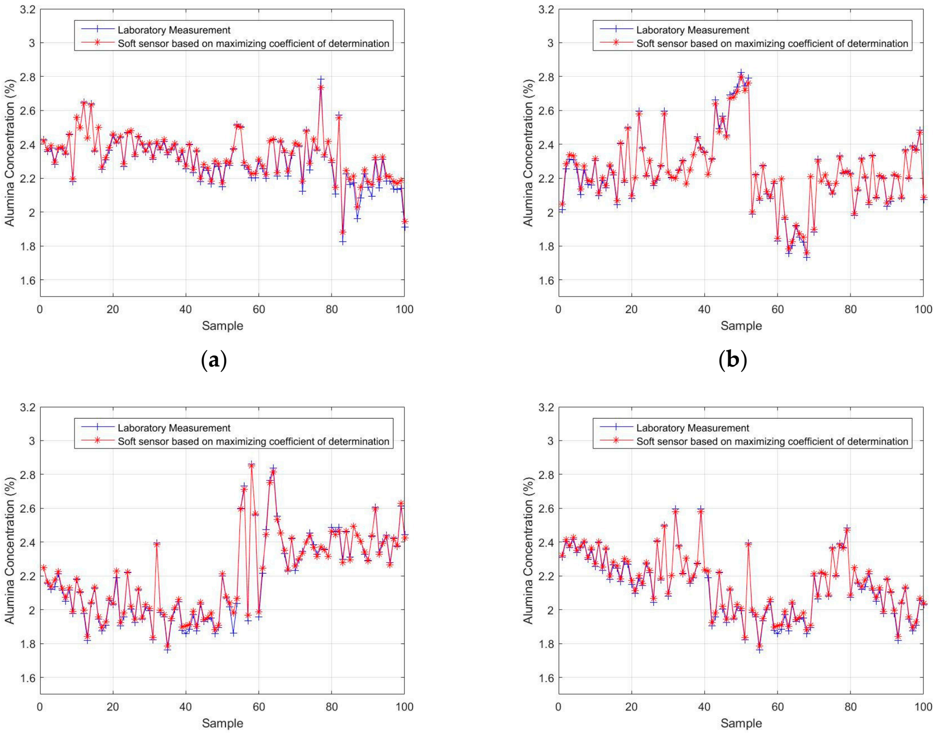
Table 3. Comparison of three data interpolation methods for a 30% missing rate.

	Mean Substitution Method	Regression Interpolation Method	EM Algorithm	Real Value
Mean	2.4140	2.4204	2.41198	2.4259
RMSE	0.1700	0.4068		0

First, comparing the mean value, we can see from the above tables that the means of the regression interpolation method and the EM data interpolation method are closer to the mean of the real value set, and the mean substitution method is less effective. Obviously, the RMSE of the EM data interpolation method is much smaller than that of the regression interpolation method. Therefore, the accuracy and effectiveness of the EM data interpolation method in processing missing values is verified. Further, if there is a problem with missing values in the practical industrial process, the EM algorithm can be selected for data interpolation.

4.2.2. Experimental Results of the Soft Sensor Model Based on Maximizing the Coefficient of Determination

In order to verify the feasibility of the proposed approach, a test sample set was used to validate the designed soft sensor model. The test sample set was divided into four subsets of 100 samples. The actual alumina concentration measurement obtained from the laboratory was compared with the output of the soft sensor model to acquire an estimated performance evaluation of the model. The results are shown in Figure 5. Figure 5a–d show the estimated alumina concentrations based on the first, second, third, and fourth test subsets, respectively. Table 4 shows the root-mean-square errors (RMSE) of the four test subsets. It can be seen that, overall, the soft sensor model based on maximizing the coefficient of determination can accurately track the overall trends in the process. The alumina concentration output by the model is approximately the same as the actual laboratory measurement.



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

(**c**) (**d**)

Figure 5. The soft-sensor-estimated alumina concentrations, based on maximizing the coefficient of determination, compared with the actual laboratory measurement using (**a**) the first test subset, (**b**) the second test subset, (**c**) the third test subset, and (**d**) the fourth test subset.

Table 4. The RMSE values of the four test subsets.

Test Subset	RMSE
First	0.0231
Second	0.0145
Third	0.0209
Fourth	0.0155

4.2.3. Comparison with BP and LSSVM

The backpropagation (BP) neural network and the least-squares, support vector machine (LSSVM) model were applied to the test sample set, and the first test subset was used for performance comparison. The parameters of the comparison algorithms were determined as follows: The number of hidden layer nodes in the BP neural network model was 100 and the activation function of the hidden layer was a sigmoid [34]. The kernel function of the LSSVM model was the radial basis function (RBF), and the kernel parameter and regular parameter were 1 and 20, respectively [34]. For each model, the number of secondary variables was 4, and the number of primary variables was 1. It could be seen that the two comparison models need different parameters in order to achieve an accurate estimation performance, while this is not necessary for the soft sensor model based on maximizing the coefficient of determination. The estimated results are shown in Figures 6 and 7. Figure 6 shows the estimated values of the soft sensor based on the BP neural network for the first test subset, and Figure 7 shows the estimated values of the soft sensor based on the LSSVM for the first test subset. It can be seen from Figure 6 that the soft sensor based on a BP neural network can roughly follow the trend of the laboratory measurements, but the error is still large at many points. It can be seen from Figure 7 that the overall performance of the soft sensor based on LSSVM is better than that based on a BP neural network, but compared with Figure 5a, it is obvious that the estimation of some extreme points is not as accurate as that given by the soft sensor based on maximizing the coefficient of determination.

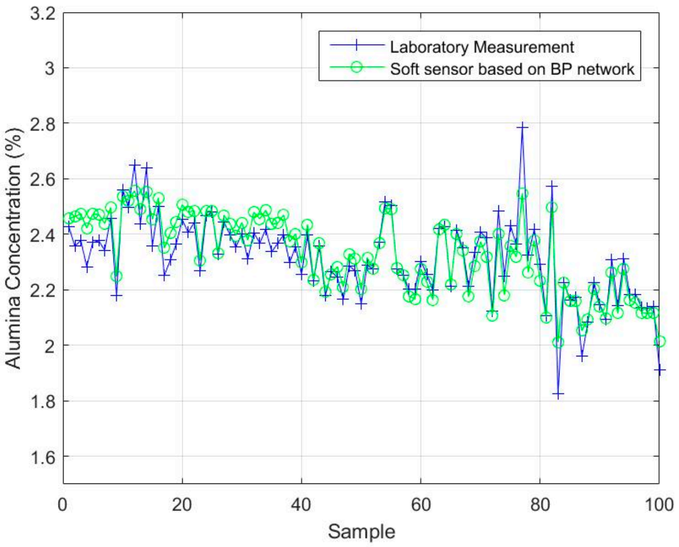


Figure 6. The estimated values of the soft sensor based on a backpropagation (BP) network compared with actual laboratory measurements.

Figures 8–10 show the soft sensor estimates based on different modelling methods as a function of the laboratory measurements. The green circles show the BP neural network model; the purple circles the LSSVM model; and the red circles the proposed coefficient of determination maximization model. In the ideal case, the circles should lie on the blue $y = x$ line. In practice, deviations from this behavior can provide information about the accuracy of the models. The BP neural network soft sensor produces a soft sensor system that has a consistent bias, since the values are consistently located above the $y = x$ line. Furthermore, the bias in the LSSVM soft sensor model is smaller, but there also seems to be a calibration issue, since the data does not lie parallel to the $y = x$ line. Finally, the proposed model has the smallest deviations and the most ideal performance.

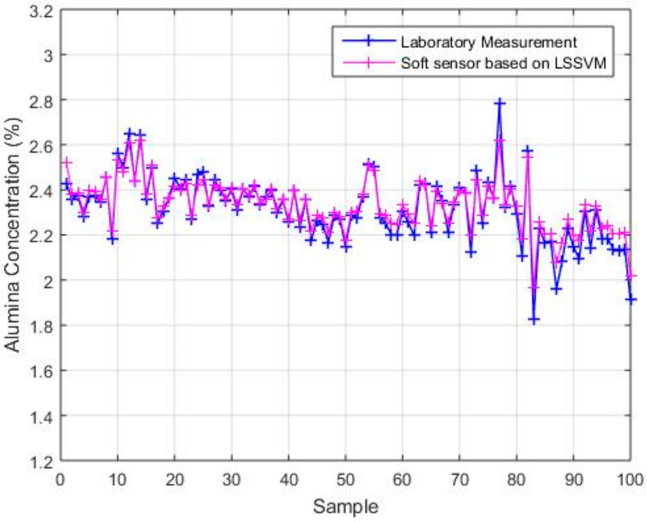


Figure 7. The estimated values of the soft sensor based on LSSVM compared with actual laboratory measurements.

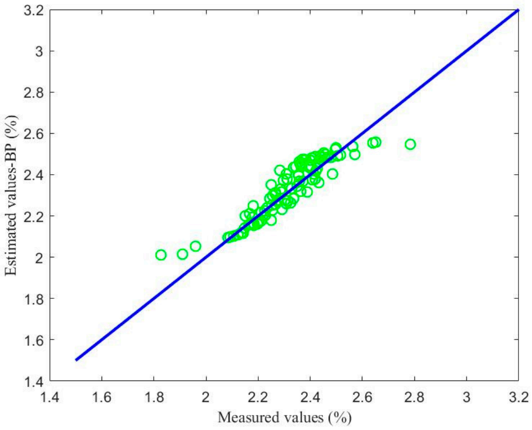


Figure 8. Comparison between the soft sensor based on a BP neural network and laboratory measurements.

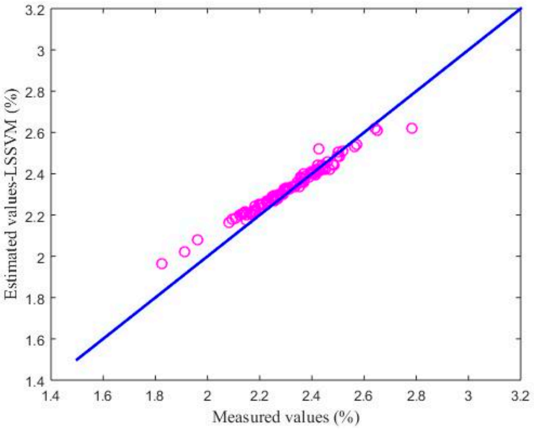


Figure 9. Comparison between the soft sensor based on LSSVM and laboratory measurements.

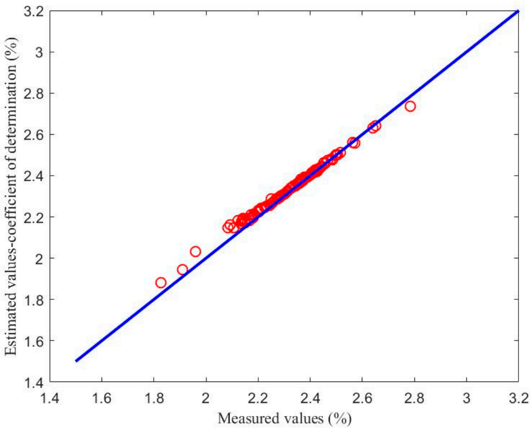


Figure 10. Comparison between the soft sensor based on maximizing the coefficient of determination and laboratory measurements.

To better illustrate the performance of the proposed soft sensor model, Table 5 shows the RMSE values for the different methods. As can be seen from Table 5, the RMSE of the proposed method is smallest, which means that the estimation effect of the proposed model is better than those of the BP neural network model and the LSSVM model.

Table 5. The comparison of the RMSE between the three modelling methods.

Method	RMSE
BP neural network	0.0616
LSSVM	0.0431
Maximizing the Coefficient of Determination	0.0231

5. Conclusions

In this paper, a new KPI estimation method for probabilistic soft sensor development is proposed based on maximizing the coefficient of determination. The joint probability distribution in the probability model is approximated using GMM, while the EM algorithm is used to estimate the GMM parameters. In addition to providing accurate, real-time estimates of the KPIs, this paper also considers the missing values that training sample sets often face and uses the EM algorithm for processing. The resulting soft sensor design method was tested on a case study of the alumina extraction process, which shows that the proposed method can provide alumina concentration estimations that are consistent with the actual measurements obtained from laboratory tests. Future work will focus on applying the proposed soft sensor development approach to solving various problems such as dealing with dynamic, non-Gaussian, or batch processes.

Author Contributions: Y.Z. and X.Y. conceived the idea, while Y.A.W.S. and X.Y. provided assistance with the development and implementation of the methods. J.C. and C.T. provided the industrial data and the experimental set-up for case study, respectively. Y.Z. performed the simulations and analysed the data, with assistance from X.Y. and Y.A.W.S. Y.Z. wrote the paper with editorial assistance from Y.A.W.S. and X.Y.

Funding: This research was funded by the National Natural Science Foundation of China, grant number [#61673053 and #61603034]; the Beijing Natural Science Foundation, grant number [#4162041 and #3182027]; and the National Key R&D Program of China, grant number [2017YFB0306403].

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

References

1. Shardt, Y.A.W.; Mehrkanoon, S.; Zhang, K.; Yang, X.; Suykens, J.; Ding, X.S.; Peng, K.X. Modelling the Strip Thickness in Hot Steel Rolling Mills Using Least-squares Support Vector Machines. *Can. J. Chem. Eng.* **2018**, *96*, 171–178. [CrossRef]

2. Zhang, K.; Shardt, Y.A.W.; Chen, Z.W.; Yang, X.; Ding, S.X.; Peng, K.X. A KPI-Based Process Monitoring and Fault Detection Framework for Large-Scale Processes. *ISA Trans.* **2017**, *68*, 276–286. [CrossRef] [PubMed]
3. Stanojevic, P.; Orlic, B.; Misita, M.; Tatalovic, N.; Lenkey, G.B. Online Monitoring and Assessment of Emerging Risk in Conventional Industrial Plants: Possible Way to Implement Integrated Risk Management Approach and KPI's. *J. Risk Res.* **2013**, *16*, 501–512. [CrossRef]
4. Paulsson, D.; Gustavsson, R.; Mandenius, C.F. A Soft Sensor for Bioprocess Control Based on Sequential Filtering of Metabolic Heat Signals. *Sensors* **2014**, *14*, 17864–17882. [CrossRef] [PubMed]
5. Abeykoon, C. A novel soft sensor for real-time monitoring of the die melt temperature profile in polymer extrusion. *IEEE Trans. Ind. Electron.* **2014**, *61*, 7113–7123. [CrossRef]
6. Yuan, X.F.; Ge, Z.Q.; Huang, B.; Song, Z.H. A Probabilistic Just-in-Time Learning Framework for Soft Sensor Development with Missing Data. *IEEE Trans. Control Syst. Technol.* **2017**, *25*, 1124–1132. [CrossRef]
7. Chen, K.; Liang, Y.; Gao, Z.L. Just-in-Time Correntropy Soft Sensor with Noisy Data for Industrial Silicon Content Prediction. *Sensors* **2017**, *17*, 1830. [CrossRef] [PubMed]
8. Khatibisepehr, S.; Huang, B.; Khare, S. Design of inferential sensors in the process industry: A review of Bayesian methods. *J. Process Control* **2013**, *23*, 1575–1596. [CrossRef]
9. Serdio, F.; Lughofer, E.; Zavoianu, A.C.; Pichler, K.; Pichler, M.; Buchegger, T.; Efendic, H. Improved fault detection employing hybrid memetic fuzzy modeling and adaptive filters. *Appl. Soft. Comput.* **2017**, *51*, 60–82. [CrossRef]
10. Serdio, F.; Lughofer, E.; Pichler, K.; Buchegger, T.; Pichler, M.; Efendic, H. Fault detection in multi-sensor networks based on multivariate time-series models and orthogonal transformations. *Inf. Fusion* **2014**, *20*, 272–291. [CrossRef]
11. Shardt, Y.A.W.; Hao, H.Y.; Ding, S.X. A New Soft-Sensor-Based Process Monitoring Scheme Incorporating Infrequent KPI Measurements. *IEEE Trans. Ind. Electron.* **2015**, *62*, 3843–3851. [CrossRef]
12. Yan, W.W.; Shao, H.H.; Wang, X.F. Soft sensing modeling based on support vector machine and Bayesian model selection. *Comput. Chem. Eng.* **2004**, *28*, 1489–1498. [CrossRef]
13. Kadlec, P.; Gabrys, B.; Strandt, S. Data-driven Soft Sensors in the process industry. *Comput. Chem. Eng.* **2009**, *33*, 795–814. [CrossRef]
14. Shang, C.; Gao, X.Q.; Yang, F. Novel Bayesian Framework for Dynamic Soft Sensor Based on Support Vector Machine with Finite Impulse Response. *IEEE Trans. Control Syst. Technol.* **2014**, *22*, 1550–1557.
15. Fujiwara, K.; Kano, M.; Hasebe, S. Development of correlation-based pattern recognition algorithm and adaptive soft-sensor design. In Proceedings of the IFAC Symposium on Advanced Control of Chemical Processes (ADCHEM), Istanbul, Turkey, 12–15 July 2009.
16. Yuan, X.; Ye, L.; Bao, L.; Ge, Z.; Song, Z. Nonlinear feature extraction for soft sensor modeling based on weighted probabilistic PCA. *Chemom. Intell. Lab. Syst.* **2015**, *147*, 167–175. [CrossRef]
17. Geladi, P. Notes on the history and nature of partial least squares (PLS) modelling. *J. Chemom.* **1988**, *2*, 231–246. [CrossRef]
18. Qin, S.J.; McAvoy, T.J. Nonlinear PLS modeling using neural networks. *Comput. Chem. Eng.* **1992**, *16*, 379–391. [CrossRef]
19. Khatisbisepehr, S.; Huang, B. Dealing with Irregular Data in Soft Sensors: Bayesian Method and Comparative Study. *Ind. Eng. Chem. Res.* **2008**, *47*, 8713–8723. [CrossRef]
20. Qi, F.; Huang, B.; Tamayo, E.C. A Bayesian Approach for Control Loop Diagnosis with Missing Data. *AICHE J.* **2010**, *56*, 179–195. [CrossRef]
21. Newman, D.A. Missing Data: Five Practical Guidelines. *Organ. Res. Methods* **2014**, *17*, 372–411. [CrossRef]
22. Zhang, K.K.; Gonzalez, R.; Huang, B.; Ji, G.L. Expectation-Maximization Approach to Fault Diagnosis with Missing Data. *IEEE Trans. Ind. Electron.* **2015**, *62*, 1231–1240. [CrossRef]
23. Shardt, Y.A.W. *Statistics for Chemical and Process Engineers: A Modern Approach*; Springer International Publishing: Cham, Switzerland, 2015; ISBN 978-3-319-21508-2.
24. Feng, C.H.; Makino, Y.; Yoshimura, M.; Rodriguez, F.J. Estimation of adenosine triphosphate content in ready-to-eat sausages with different storage days, using hyperspectral imaging coupled with R statistics. *Food Chem.* **2018**, *264*, 419–426. [CrossRef] [PubMed]
25. Sezer, B.; Apaydin, H.; Bilge, G.; Boyaci, I.H. Coffee arabica adulteration: Detection of wheat, corn and chickpea. *Food Chem.* **2018**, *264*, 142–148. [CrossRef] [PubMed]

26. Sun, S.L.; Zhang, C.S.; Yu, G.Q. A Bayesian network approach to traffic flow forecasting. *IEEE Trans. Intell. Transp. Syst.* **2006**, *7*, 124–132. [CrossRef]
27. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **1977**, *39*, 1–38.
28. Stock, J.H.; Watson, M.W. *Introduction to Econometrics*, 3rd ed.; Addison-Wesley: Bosten, MA, USA, 2010; ISBN 978-0-13-800900-7.
29. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*, 6th ed.; Pearson: London, UK, 2007; ISBN 978-0-13-187715-3.
30. Rao, C.R. *Linear Statistical Inference and Its Applications*; Wiley: New York, NY, USA, 1973; ISBN 978-0-47-031643-6.
31. Bilmes, J.A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov model. *Int. Comput. Sci. Inst.* **1998**, *4*, 126.
32. Mouedhen, G.; Feki, M.; Wery, M.D.P.; Ayedi, H.F. Behavior of aluminum electrodes in electrocoagulation process. *J. Hazard. Mater.* **2008**, *150*, 124–135. [CrossRef] [PubMed]
33. Yao, Y.C.; Cheung, C.Y.; Bao, J.; Skyllas-Kazacos, M.; Welch, B.; Akhmetov, S. Estimation of spatial alumina concentration in an aluminium reduction cell using a multilevel state observer. *AICHE J.* **2017**, *63*, 2806–2818. [CrossRef]
34. Zhang, S.; Zhang, T.; Yin, Y.X.; Xiao, W.D. Alumina concentration detection based of the kernel extreme learning machine. *Sensors* **2017**, *17*, 2002. [CrossRef] [PubMed]