Chemical Engineering Science 122 (2015) 573–584

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Robust supervised probabilistic principal component analysis model for soft sensing of key process variables

Jinlin Zhu, Zhiqiang Ge*, Zhihuan Song

State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027, PR China

HIGHLIGHTS

- A robust probabilistic method is proposed to deal with the soft sensing problem.
- Mixture of robust probabilistic principal component analysis with Student *t*-distribution is used for modeling.
- The expectation–maximization algorithm is employed for parameter learning.
- Bayes soft alignment method of local predictions is developed for online soft sensing.
- The superiority of the developed method is tested on the debutanizer column.

ARTICLE INFO

Article history:
Received 28 March 2014
Received in revised form
8 September 2014
Accepted 15 October 2014
Available online 23 October 2014

Keywords:
Soft sensor
Robust modeling
Supervised probabilistic model
Process data modeling
Outliers

ABSTRACT

In this paper, a robust and mixture form of supervised probabilistic principal component analysis model is proposed to deal with the soft sensing problem, particularly for those processes with multiple operating conditions and the collected datasets may contain outliers. Under the framework of latent variable models, the commonly adopted multivariate Gaussian distribution assumption is replaced by the multivariate student *t*-distribution so as to tolerate the notorious outliers by using the adjusted heavy tail. After the construction of robust probabilistic model, the iterative expectation–maximization algorithm is derived to perform the parameter estimation for both single and mixture models. For online soft sensing application, the Bayes rule is introduced for soft alignment of local prediction results. Two case studies are provided for performance evaluation of the proposed method, both in comparison with the conventional supervised model. Results indicate that the new model is much more reliable under outlier contaminated and multimode conditions.

1. Introduction

Nowadays, industrial plants are usually equipped with a large number of hardware sensors in order to collect process data for monitoring and control (Kadlec et al., 2009; Ge et al., 2013a, 2013b). In most cases, the key variables or quality variables which indicate the production state are hardly available by hardware sensors and have to undergo lab analyses. However, lab analyses are expensive and time-consuming thus may cause time delay which is contradictive to the real-time requirement for process monitoring and control. To overcome the deficiency of hardware instruments, predictive models called soft sensors are usually constructed to perform quality variables predictions with the available easy-to-measure process variables (Facco et al., 2009; Fujiwara et al., 2012; Yan et al., 2004).

Roughly speaking, there are two main types of soft sensors, namely model-based soft sensors and data-based soft sensors. Model-based techniques require the explicit model dynamical evolution presentations. However, a detailed and complete state-space description for complex industrial systems can be expensive, time consuming and sometimes technically unavailable (Lin et al., 2007). On the contrary, the data-based soft sensors can be derived directly based on the data measurements which demand the least model knowledge. Meanwhile, with the development of the distributed control systems and the popularization of large capacity database techniques, a great volume of process data could be collected and recorded, which provides a great convenience for statistical data analysis and diagnosis. As a result, the data-based sensors have gained continuous and increasing focus in the field over the last few years (Khatibisepehr et al., 2013). Conventionally, most data-based soft sensors resort to statistical regression

* Corresponding author. Tel.:+86 87951442.
E-mail address: gezhiqiang@zju.edu.cn (Z. Ge).

http://dx.doi.org/10.1016/j.ces.2014.10.029
0009-2509/© 2014 Elsevier Ltd. All rights reserved.

methods such as multivariate regression, principal component regression (PCR) and partial least regression (PLS) (Ahmed et al., 2009; Kaneko et al., 2009). There are also other intelligent soft sensors based on machine learning techniques like artificial neural networks (ANN), support vector regression (SVR) and their corresponding improvements (Chen and Wang, 1998; Yu, 2012). Although models like PCR have been widely used, a main drawback for these methods is the lack of a proper probabilistic definition of underlying uncertainty introduced by data noise. To get over this issue, the probabilistic version of PCR (PPCR) based on EM algorithm has been proposed under the framework of latent variable models (Zhou et al., 2013). By introducing the random distribution property for each variable, the PPCR is more extendable and elegant in model selection and parameter estimation with the Bayesian inference mechanism. Recently, the PPCR has also been extended into the mixture form and has been experimentally proved effective in dealing with the multi-mode industrial cases (Ge et al., 2011).

Despite the noisy uncertainties, another notorious problem for soft sensor modeling is the appearance of outliers (Hodge and Austin, 2004; Khatibisepehr and Huang, 2008; Khatibisepehr et al., 2013). The outliers are often considered as those measurements deviate from the typical ranges of collected process data. It should be noted that most soft sensors are constructed under the ideal condition that no outliers are mixed within the modeling cases (De La Torre and Black, 2003). Unfortunately, it is known that most industrial datasets contain outliers due to incorrectly observed, recorded or copied process measurements. Generally speaking, outliers can be treated as obvious ones and non-obvious ones according to whether or not the values are beyond the meaningful ranges (Kadlec et al., 2009). In some conditions, certain meaningless outliers can be easily excluded given the upper and lower limits for each sensor. However, many outliers include the obvious ones may not simply exceed the given limits. Hence, manual evaluation and discard can be arduous and inefficient. Moreover, simply discard of outliers could be another drawback since the information loss can lead to a biased estimation for parameters (Fang and Jeong, 2008). As alternatives, some methods turn to design robust estimations for the original regression models such as robust PCR, robust PLS and robust partial least squares regression (PLSR) etc (Gil and Romera, 1998; Hubert and Branden, 2003; Hubert and Verboven, 2003). However, none of these gives the general probabilistic explanation for process data uncertainties.

Recently, some researchers have tried to introduce the student t-distribution into the probabilistic modeling framework so as to handle the outliers in a more natural and friendly manner (Jylänki et al., 2011; Luttinen et al., 2012; Wei and Li, 2012). Compared with the Gaussian distribution which is commonly appeared for constructing PPCR and PPCA, the student t-distribution shows more stability and compatibility due to its heavy tail. The heavy tail is usually adjusted by the parameter called degree of freedom which is learned and adapted from the training procedure. Therefore, the student t-distribution is more robust to outliers than the Gaussian one (Zhu et al., 2014). Notice that different from the robust PCR/PLS that takes the hard strategy by detecting and rejecting outliers, the Student t model softly explains the outliers with probabilistic framework which in terms makes the robust model more elegant and extensible for modeling as well as parameter learning. Due to the desirable and elegant interpret ability for outliers, more and more studentized models can be found which have been developed upon the original Gaussian ones such as the student t-mixture model, the robust PPCA and its mixture form etc (Archambeau et al., 2006, 2008; Chatzis et al., 2009; Svensén and Bishop, 2005). However, as far as we know, no studies have been investigated to develop the robust soft sensors and make the industrial applications from the student t aspect.

The motivation of this article is to propose a novel soft sensing model called robust supervised probabilistic principal component analyzer. First, the conventional SPPCA is modified into the student t-version to conduct the robust modeling phase within potential outliers. Then, the obtained model is extended into the mixture version with the EM algorithm so as to cope with the multimode data characteristics. During the online soft sensing procedure, for each new coming measurement, instead of hard assignment, the estimated value from each local model is softly aligned with the corresponding weight, and the global estimation is considered as the current time production quality. Notice that the local weight is computed with the posterior of the measurement with respect to each local model, which can be realized with the Bayes rule.

The rest of this paper is organized as follows. In Section 2, the conventional supervised PPCA are revisited. Then the robust supervised PPCA is proposed and the corresponding mixture model is further developed with the EM algorithm in Sections 3 and 4, respectively. Followed by the algorithm, the online soft sensing mechanism is developed based on the proposed model. After that, two case studies are used to validate the proposed method in Section 6. Finally, the conclusions are made.

2. Preliminaries

In this section, the traditional SPPCA is briefly revisited. As a first step and introduction, we first come to the PPCA method.

2.1. PPCA

Given data set $\{\mathbf{x}_n|\mathbf{x}_n \in R^D\}_{n=1}^N$, where N is the sample number, D is the number of dimension for data space, the PPCA tries to find a linear projection from the original measurements to a lower dimensional latent vectors $\{\mathbf{t}_n|\mathbf{t}_n \in R^d\}_{n=1}^N$, the generative model can be described as follows (Tipping and Bishop, 1999):

$$\mathbf{x}_n = \mathbf{P}\mathbf{t}_n + \boldsymbol{\mu} + \mathbf{e}_n \tag{1}$$

where $\mathbf{e} \in R^{D\times 1}$ denotes the noise, $\mathbf{P} \in R^{D\times d}$ is the orthogonal projection matrix, $\mathbf{t} \in R^{d\times 1}$ is the latent variable vector, $\boldsymbol{\mu} \in R^D$ denotes the offset, $d < D$. Notice that here the latent dimension number is assumed to be given and we will discuss the choice for latent dimensionality later on. In PPCA, the probability distributions for latent variable and the noise are both assumed as Gaussian ones; therefore, we have $p(\mathbf{t}_n) = N(\mathbf{0}, \mathbf{I}_d)$ $p(\mathbf{e}_n) = N(\mathbf{0}, \tau\mathbf{I}_D)$ and $p(\mathbf{x}|\mathbf{t}_n) = N(\mathbf{P}\mathbf{t}_n + \boldsymbol{\mu}, \tau\mathbf{I}_D)$. Here, $\mathbf{I}_d$ denotes the d-dimensional identity matrix for covariance, τ represents the magnitude and equals to the square of standard deviation. The parameters for PPCA is $\boldsymbol{\Theta} = \{\mathbf{P}, \boldsymbol{\mu}, \tau\}$ can be obtained by maximum likelihood approach or the EM algorithm, for details, one can refer to Archambeau et al. (2008), Kim and Lee (2003).

2.2. SPPCA

The SPPCA extends the PPCA by incorporating the label information into the projection phase. The aim of extension is to build predictive models so as to deal with the soft sensing problems. Different from PLS which simply consider the inter covariance, SPPCA finds the projection based on both the inter covariance and the intra covariance between the input and the output (Yu et al., 2006). For comparison, the probabilistic graphic models for PPCA and SPPCA are depicted in Fig. 1(a) and (b).

Given input (process variable related) dataset $\{\mathbf{x}_n|\mathbf{x}_n \in R^{D_x}\}_{n=1}^N$ and output (quality relevant related) dataset $\{\mathbf{y}_n|\mathbf{y}_n \in R^{D_y}\}_{n=1}^N$, the generative model for SPPCA is given as (Ge et al., 2011)

$$\mathbf{x}_n = \mathbf{P}\mathbf{t}_n + \boldsymbol{\mu}_x + \mathbf{e}_n \tag{2}$$