Neurocomputing 82 (2012) 167–178

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

On-line principal component analysis with application to process modeling

Jian Tang [a,*], Wen Yu [b], Tianyou Chai [a,c], Lijie Zhao [a,d]

[a] State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110004, China
[b] Departamento de Control Automatico, CINVESTAV-IPN, Av.IPN 2508, México D.F. 07360, Mexico
[c] Research Center of Automation, Northeastern University, Shenyang 110004, China
[d] College of Information Engineering, Shenyang Institute of Chemical Technology, Shenyang 110142, China

ARTICLE INFO

Article history:
Received 4 April 2011
Received in revised form
21 October 2011
Accepted 24 October 2011
Communicated by S. Fiori
Available online 29 December 2011

Keywords:
Principal component analysis
Recursive algorithm
On-line modeling
Process modeling

ABSTRACT

Principal component analysis (PCA) has been widely applied in process monitoring and modeling. The time-varying property of industrial processes requires the adaptive ability of the PCA. This paper introduces a novel PCA algorithm, named on-line PCA (OLPCA). It updates the PCA model according to the process status. The approximate linear dependence (ALD) condition is used to check each new sample. A recursive algorithm is proposed to reconstruct the PCA model with selected samples. Three types of experiments, a synthetic data, a benchmark problem, and a ball mill load experimental data, are used to illustrate our modeling method. The results show that the proposed OLPCA is computationally faster, and the modeling accuracy is higher than conventional moving window PCA (MWPCA) and recursive PCA (RPCA) for time-varying process modeling.

1. Introduction

Many important process parameters in industrial processes, such as product quality and some key process parameters, cannot be measured by hardware sensors. Nowadays, these variables mainly depend on the manual timed sampling and titration in a chemical laboratory. The manual titration has a high precision, but the sampling interval is too long, and some parameters have to be judged by the experts' experience. They cannot be used for real-time process monitoring and control effectively. Soft sensing models are alternative methods for immeasurable variables [1].

The data in the industrial processes are strongly coupled. The industrial environment is often called "data rich but information poor". How to select useful variables is a key job to construct a soft sensing model [2]. Including all variables in a model not only increases the complexity of the model, but also produces a negative effect in the model training phase. It has been shown that the principal component analysis (PCA) can successfully monitor the changes of industrial processes [3], including chemical [4] and microelectronics manufacturing processes [5]. Without losing generality of the system, PCA transforms the original input variables into the independent variables in a new reduced space. It is one of the most popular modeling techniques applied to data-driven soft sensors [2]. In practice, PCA is a popular multivariate statistical technique for data compression, which is usually applied as the pre-processing step followed by the actual computational learning method [2]. It has been combined with many computational learning methods such as artificial neural networks [6], neuro-fuzzy systems [7], support vector machines [8], fuzzy C-means [9], and multi-model technology [10]. PCA has also been applied to the power spectrum of the acoustic emission of mechanical equipments, which shows that the extracted spectral principal components (PCs) give a significant improvement in fault detection and fault diagnosis [11,12]. There is a direct relation between PCA and neural networks. The use of a particular Hebbian learning rule (Oja's learning rule [13]) results in a network that performs PCA [14].

Partial least square (PLS) captures the maximal covariance between two data blocks. It has been widely applied in chemometrics, steady-state process modeling, dynamic modeling and process monitoring [15]. However, PLS is inappropriate to capture the nonlinear characteristics between the predictor and response variables. The over-fitting problem of the PLS regression requires more latent variables than they need [16]. Genetic algorithm shows effectiveness for PLS variable selection [17]. Some nonlinear approaches, such as quadratic PLS [18], neural network PLS [19], fuzzy PLS [20], and kernel PLS [21], are also useful to overcome the above problems of PLS.

When new data comes, the nonlinear models based on PCA/PLS have to be rebuilt using all data. To track the changing characteristics of the industrial process, attention should be paid more on the recent data, i.e., the soft sensing models should be

* Corresponding author.
E-mail address: freeflytang@gmail.com (J. Tang).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.10.026

updated [22], and PCA/PLS models are adaptive. Moving window and recursive technique are popular methods to adjust PCA/PLS models with respect to the time-varying data. For example, the moving window PCA (MWPCA) [23], exponentially weighted moving average (EWMA) [21], sample-wise/block-wise recursive PLS (RPLS) [15], recursive PCA (RPCA) [24–26]. In [24] the correlation matrix was applied in RPCA for the industry process monitoring. In [25] the neural network method was used to obtain the PCs of the input matrix recursively. In [26], rank-one perturbation based method was used to update the PCs of the covariance matrix for input variables. MWPCA generates a new process model by including the newest sample and excluding the oldest one by moving a time-window. However, the window has to cover a large number of data points in order to include sufficient process variation for modeling and monitoring purposes, and small windows may adapt to process changes so quickly that abnormal behavior remains undetected. Even the fast MWPCA cannot solve the window size problem essentially. Recursive techniques, on the other hand, update the model for an ever increasing dataset that includes new samples without discarding old ones. The drawback is the influence of the old data that can last for a very long time. Furthermore, the criterion of PLS minimizes the empirical risk, which may cause the over fitting problem.

In this paper, using an on-line adaptive technique, the standard RPCA is modified into an on-line version for process modeling. It is called on-line PCA (OLPCA). This paper shows how to simplify the computing procedure in the OLPCA, and how to modify the OLPCA for process modeling, and how to deal with the trade-off problem of the prediction accuracy and learning time in OLPCA. The off-line training samples were used to construct the PCA process model. When a new sample comes, the on-line prediction based on the initial training samples decides if this new sample would be used to update the PCA model. Here the approximate linear dependence (ALD) condition is used to check the linear independence of the new sample, such that the new model is not represented by a linear combination of its previously admitted samples. The ALD condition and the score matrix are calculated recursively.

This paper is organized as follows. A brief description of the PCA method for nonlinear system modeling is introduced in Section 2. Then, OLPCA is proposed in Section 3, including the increase of the dictionary, calculation of the score matrix for the process model, and the reconstruction of the process model. Experimental comparisons modeling the synthetic, benchmark data and experiment-based mill load data with OLPCA are given in Section 4. At last, the conclusions are derived in Section 5.

2. Principal component analysis for process modeling

The following discrete-time nonlinear system will be modeled by a PCA model:

$$y_l = f(\mathbf{z}_l), \quad l = 1,2,\ldots,k \tag{1}$$

where $\mathbf{z}_l = [y(l-1), y(l-2), \cdots, u(l-d), u(l-d-1), \cdots]$, $f(\cdot)$ is an unknown time-varying nonlinear function, representing the plant dynamics, $u(l)$ and $y(l)$ are the measurable input and the output of the nonlinear plant, $d$is the time delay. $\mathbf{z}(l)$ is the input to the nonlinear function $f(\cdot)$. The mechanism-based or data-driver based approach will be used to estimate $f(\cdot)$, such that the estimate output $\hat{y}_l$ closes to the real y_l.

The key idea of PCA model is to reduce the number of variables by building their linear combinations [3]. The new set of variables is called PCs, which are uncorrelated and in order. The first few variables give the most contributions. For a conventional PCA, the raw data $\mathbf{X}_k^0 \in R^{k\times p}$ is made of k samples (rows) and p variables (columns). $\mathbf{X}_k^0$ is first normalized to zero mean and unit variance. The normalized $\mathbf{X}_k$ is decomposed as follows:

$$\mathbf{X}_k = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_h\mathbf{p}_h^T + \mathbf{t}_{h+1}\mathbf{p}_{h+1}^T + \cdots + \mathbf{t}_p\mathbf{p}_p^T \tag{2}$$

where $\mathbf{t}_i$ and $\mathbf{p}_i (i = 1,\ldots,p)$ are the PC score and loading, respectively. $\mathbf{p}_i$ is actually the ith eigenvector of the correlation matrix $\mathbf{R}_k \in \Re^{p\times p}$ according to

$$\begin{cases} \mathbf{R}_k \approx \frac{1}{k-1}\mathbf{X}_k^T \cdot \mathbf{X}_x \\ (\mathbf{R}_k - \lambda_k)\mathbf{P}_k = 0 \end{cases} \tag{3}$$

where λ_k are the eigenvalues of $\mathbf{R}_k$, $\mathbf{P}_k \in \Re^{p\times p}$ are the eigenvectors of $\mathbf{R}_k$. As the $\mathbf{T}_k \in \Re^{k\times p}$ are the orthogonal projections of $\mathbf{X}_k$ in the new basis $\mathbf{P}_k$, the scores $\mathbf{T}_k$ are calculated as

$$\mathbf{T}_k = \mathbf{X}_k\mathbf{P}_k. \tag{4}$$

The dimension reduction is obtained by decomposing $\mathbf{X}_k$ as

$$\mathbf{X}_k = \hat{\mathbf{X}}_k + \tilde{\mathbf{X}}_k = \hat{\mathbf{T}}_k\hat{\mathbf{P}}_k^T + \tilde{\mathbf{T}}_k\tilde{\mathbf{P}}_k^T \tag{5}$$

where $\hat{\mathbf{X}}_k$ and $\tilde{\mathbf{X}}_k$ are the modeled and residual components, respectively; $\hat{\mathbf{P}}_k \in \Re^{p\times h}$ contains the first h eigenvectors of $\mathbf{R}_k$, called loading matrix and PCA models, whose range space is called the principal component subspace (PCS); $\hat{\mathbf{T}}_k \in \Re^{n\times h}$ is the projection of $\mathbf{X}_k$ onto $\hat{\mathbf{P}}_k$, called score matrix; $\tilde{\mathbf{P}}_k^T \in \Re^{p\times(p-h)}$ is the residual loading, and its range space is called residual subspace (RS); $\tilde{\mathbf{T}}_k \in \Re^{n\times(p-h)}$ is the residual score. $\hat{\mathbf{T}}_k$ and $\tilde{\mathbf{T}}_k$ can be rewritten as

$$\hat{\mathbf{T}}_k = \mathbf{X}_k\hat{\mathbf{P}}_k \tag{6}$$

$$\tilde{\mathbf{T}}_k = \mathbf{X}_k\tilde{\mathbf{P}}_k \tag{7}$$

Since the industrial process is multi-variable, with strong coupling and uncertainty, the mechanism based modeling approach requires a lot of process expert knowledge. The data-driver modeling is a black-box method. It only depends on the empirical observations. Directly throwing away some input variables, which have a lower contribution for modeling, causes the information loss. Since the correlation among the process variables is an important issue, an effective method is to transform the original process variables into independent variables, and keep the original information as much as possible. In this case, the new independent variables are used to construct the nonlinear system model. The correlation among the process variables is removed, and the complex modeling becomes easier.

PCA has been used to address multiple co-linearity problems, and nonlinear industry process modeling [2]. The pre-processed data, i.e., the score matrix $\hat{\mathbf{T}}_k$ in (6) is used as the new variable to construct nonlinear model. The input $\mathbf{z}_l$ of the nonlinear model in (1) is replaced by $\hat{\mathbf{t}}_l$. Most of the industrial processes are time-varyingly slow. At the time m, if the nonlinear model is changed as

$$y_m = f'(\mathbf{z}_m), \quad m = k+1, k+2, \ldots \tag{8}$$

where $f'(\cdot)$ is the new nonlinear model; $\mathbf{z}_m$ is the input variable at the time m. When PCA is used to construct the nonlinear model on-line, at least three steps should be carried out:

(1) The PCA model is updated effectively;
(2) the input $\mathbf{z}_m$ is updated;
(3) a new nonlinear model for $f'(\cdot)$ is reconstructed.

3. On-line principal component analysis for process modeling

In this section, a new PCA method for nonlinear systems on-line modeling is proposed. It consists of four parts: (1) performance