Forecastable Component Analysis and Partial Least Squares Applied on Process Monitoring

Dan Wang[1, a], Yirong Lu[2, b], Yupu Yang[3, c]

[1,2,3]*Department of Electrical Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China*

[a]*season580@sjtu.edu.cn*, [b]*sx.xc@163.com*, [c]*ypyang@sjtu.edu.cn*

Abstract.

Forecastable Component Analysis (ForeCA) is a new feature extraction method for multivariate time series. ForeCA can find an optimal transformation to dig out the potential forecastable information structure from large amounts of data. This paper combines ForeCA with PLS for industrial process monitoring. This method overcomes the drawback that partial least squares(PLS) rarely use dynamic timing characteristics of system, so it can reflect the dynamic nature of industrial processes better. We use PLS for regression after appropriate forecastable components selected. Finally, we construct CUSUM statistic and SPE statistic for monitoring industrial processes. Simulation results on the Tennessee Eastman (TE) process illustrate the effectiveness of the proposed method for detecting slow drift fault.

Keywords: Forecastable Component Analysis; Partial Least Squares; Fault Detection; TE Process;

Introduction

In recent years, with the ever increasing scales and complexities in modern industries such as chemical industry and metallurgy, more concerns are focused on the safety issues in the industrial process. Methods based on multivariate statistical analysis have become a hotspot in areas of fault detection and diagnosis and have been effectively applied in the industrial process [1].

Partial least squares (PLS) can precisely extract relations between quality variables and process variables within normal working conditions and effectively monitor the process. Moreover, PLS statistical monitoring technique is independent from process mechanism models and it can be trained without fault samples. Therefore, PLS has been widely applied and deeply studied in chemical production's quality control and online monitoring [2]. However, PLS is incapable of reflecting dynamic timing characteristics, which to some extent affects its accuracy in fault detection. Forecastable component analysis (ForeCA) [3], as a brand new statistical signal processing method for multivariate timing signal feature extracting, has overcome the shortcoming. As its capability to predict system's operation tendency through extracting the dynamic characteristics from existing data, ForeCA can essentially describe industrial process through extracted features.

In this paper, ForeCA and PLS are combined for fault detection. By mapping samples onto the predictable subspace and then applying the least squares regression, the predictability of model is further improved. Meanwhile, CUSUM and SPE are constructed to monitor the system so as to better detect faults with mean deviation under two times of standard deviation. This technique has overcome the shortcoming of inability to reflect process timing characteristics by traditional partial least squares. It can predict system's operation tendency and reflect dynamic characteristics. Consequently, the accuracy of fault detection will be improved.

Basic algorithm

The basic idea of ForeCA is to assume a matrix $\mathbf{X} \in R^{n \times m}$, where n is the number of samples, m the number of variables. By linear transformation $\mathbf{W}^{\mathrm{T}} \in R^{k \times n}$, it leads to:

$$S = W^T X \tag{1}$$

where W is a loading matrix while S is a score matrix composed of predictable column vectors,. ForeCA is applied to estimate S and W through observation matrix X.

For a multivariate steady process $\mathbf{X_t}$ of second order, after a linear transformation $\mathbf{y}_t = \mathbf{w}^{\mathrm{T}}\mathbf{X}_t$, where $\mathbf{w} \in R^n$ is the column vector of W in Equation (1), i.e. the predictable component, $\mathbf{y}_t$ can be seen as a univariate stationary process of second order. The optimization problem of ForeCA is given in Literature [3] as:

$$\begin{aligned}
&\max_{\mathbf{w}} \Omega(\mathbf{w}^{\mathrm{T}}\mathbf{X}_t) = \max_{\mathbf{w}}(1+\int_{-\pi}^{\pi} f_y(\lambda)\log_{2\pi} f_y(\lambda)\,\mathrm{d}\lambda) \\
&\quad = \max_{\mathbf{w}}(1+\int_{-\pi}^{\pi} \mathbf{w}^{\mathrm{T}} S_{\mathbf{X}}(\lambda)\mathbf{w}\log_{2\pi}(\mathbf{w}^{\mathrm{T}} S_{\mathbf{X}}(\lambda)\mathbf{w})\,\mathrm{d}\lambda) \\
&\text{s.t.} \quad \mathbf{w}^{\mathrm{T}}\Sigma_{\mathbf{X}}\mathbf{w} = 1
\end{aligned} \tag{2}$$

where $\Omega(\mathbf{w}^{\mathrm{T}}\mathbf{X}_t)$ is forecastability of $\mathbf{y}_t$ when $\mathbf{w}$ is the forecastable component, $S_{\mathbf{X}}(\lambda)$ is the spectral density of multivariate stationary process $\mathbf{X_t}$, i.e. $S_{\mathbf{X}}(\lambda) = \dfrac{1}{2\pi}\sum_{k=-\infty}^{\infty}\Gamma_{\mathbf{X}}(k)\mathrm{e}^{2\pi\mathrm{i}k\lambda},\ \lambda \in [-\pi,\pi]$, $\Gamma_{\mathbf{X}}(k)$ is the ACVF of $\mathbf{X_t}$ and $\Sigma_{\mathbf{X}}$ is the covariance matrix of $\mathbf{X_t}$.

To solve Equation (2), firstly weighted overlaps average spectrum density estimation algorithm is performed to estimate spectrum density and then EM-Like algorithm is applied to solve for the predictable components. A set of predictable components ordered by descending forecastability can be obtained through detailed steps shown in Literature [3] and the forecastable component matrix $\mathbf{W}^{\mathrm{T}}$ in Equation (1) can hence be solved.

Given input matrix $\mathrm{X} \in \mathrm{R}^{n\times N}$ with n samples, each with N process variables, and output matrix $Y \in \mathrm{R}^{n\times M}$ with n samples, each with M quality variables, PLS is used to model the relations between two sets of data through latent variables, which decomposes the $n\times N$ zero-mean valued matrix X and $n\times M$ zero-mean valued matrix Y into the following forms:

$$\mathrm{X}=\hat{\mathrm{X}}+\mathrm{E}_k=\sum_{i=1}^{k} t_i p_i^T+\mathrm{E}_k=\mathrm{TP}^{\mathrm{T}}+\mathrm{E}_k \tag{3}$$

$$\mathrm{Y}=\hat{\mathrm{Y}}+F_k=\sum_{i=1}^{k} b_i t_i q_i^T+F_k=\mathrm{TBQ}^{\mathrm{T}}+F_k \tag{4}$$

where $\hat{X}$ and $\hat{Y}$ are fitting matrixes, E_k and F_k are fitting error matrixes,

$T=[t_1,...,t_k]$ is the score matrix, P is loading matrix of X, and Q is loading matrix of Y. In PLS model, the loading matrix and score matrix explain the information of itself through maximization, respectively. Meanwhile, the degree of correlation between X and Y is maximized in the process. The most common algorithm to compute PLS model is Nipals algorithm, through which, the forecast regression function of Y is computed as follows:

$$\hat{Y} = XB_{PLS} = XM(P^T M)^{-1} BQ^T \tag{5}$$

where, B_{PLS} is the coefficient matrix, and weighting matrix M is defined by Nipals algorithm as $T = XM$ .In complex multivariate system, PLS algorithm regards independent variable $X \in R^{n\times N}$ and dependent variable $Y \in R^{n\times M}$ as data matrixes with linear relations. This algorithm recombines information of manifest variable systems by utilizing ideas of information decomposition.

Design process monitoring model based on ForePLS

In the industrial process, there exists plenty of slow shift faults. To detect those subtle changes, CUSUM statistic is used for monitoring. In CUSUM statistic, two statistics are defined to detect upward shifts and downward shifts of samples' mean values. They are:

$$S_H(i) = \max[0, x_i - (\mu_0 + K) + S_H(i-1)],\ \ S_H(0) = 0$$

(6)

$$S_L(i) = \max[0, (\mu_0 - K) - x_i + S_L(i-1)],\ \ S_L(0) = 0$$

(7)

$$S_i = \sum_{j=1}^{i} (x_j - \mu_0)$$

(8)

where μ_0 is the actual mean value of sample, and x_j is the jth sample value, which is the mean value of the trained samples. K is called the reference value, and is set as 0.5Δ, where Δ is deviation to be detected and its range falls into [0.5σ, 2σ] and the control limit is five times of the standard variance.

First select a set of observation data $\mathrm{X} \in \mathrm{R}^{n \times N}$ acquired under normal working conditions, where n is the number of variables, N the sample number. ForeCA algorithm is performed on X to derive the following forecastable component matrix: $\mathbf{W}^{\mathrm{T}} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n]^{\mathrm{T}} \in R^{N \times N}$

By combining Equation (3) and (6), $\hat{\mathrm{X}}$ can also expressed as

$$\hat{X} = \mathrm{T}P^T = W^{\mathrm{T}}\mathrm{X}M(P^T M)^{-1}P^T = W^{\mathrm{T}}XM_p \in \hat{S} \tag{9}$$

and the process residual variance is denoted

$$\tilde{X} = W^{\mathrm{T}}X - \hat{X} = W^{\mathrm{T}}X - W^{\mathrm{T}}XM_p = W^{\mathrm{T}}X(\mathrm{I} - M_p) \in \tilde{S} \tag{10}$$

Then the n-dimension process data space is decomposed into two orthogonal complementary subspaces: latent variable space $\hat{S}$ and residual variance space $\tilde{S}$. Similar to PCA defining SPE statistic [4], SPE is defined in the residual variance space $\tilde{S}$, which represents changes in data that have not been explained by the principle component model. The changes are the degree of measured values deviating from principle component model.

$$\mathrm{SPE(i)} = \tilde{x}_i \cdot \tilde{x}_i^T = \left\| W^{\mathrm{T}}X\left(i\right)(\mathrm{I} - M_p) \right\| \tag{11}$$

The control limit of SPE statistic is fixed by kernel density estimation. Refer to Literature [5] for details.

Fault detecting based on ForePLS can be divided into two stages: offline training stage and online detecting stage.

- Offline training stage: First collect training data X under normal working condition. After pre-processing, ForeCA algorithm is applied to extract forecastable component matrix W. Later PLS regression is conducted in the forecastable subspaces. Then training CUSUM and SPE statistics in the forecastable subspace. Finally, the control limits of the two statistics are obtained: H and SPE_α.
- Online detecting stage: First, according to real-time collected data under unknown status, the forecastable model is applied on the online data to compute CUSUM and SPE statistics for each sample data. Then compare the two statistics with the corresponding control limits and it is thus determined whether faults occurred. If the statistics below the limits, it demonstrates the system works within the change range predicted by the forecastable model, which means the system normal. Otherwise, the system

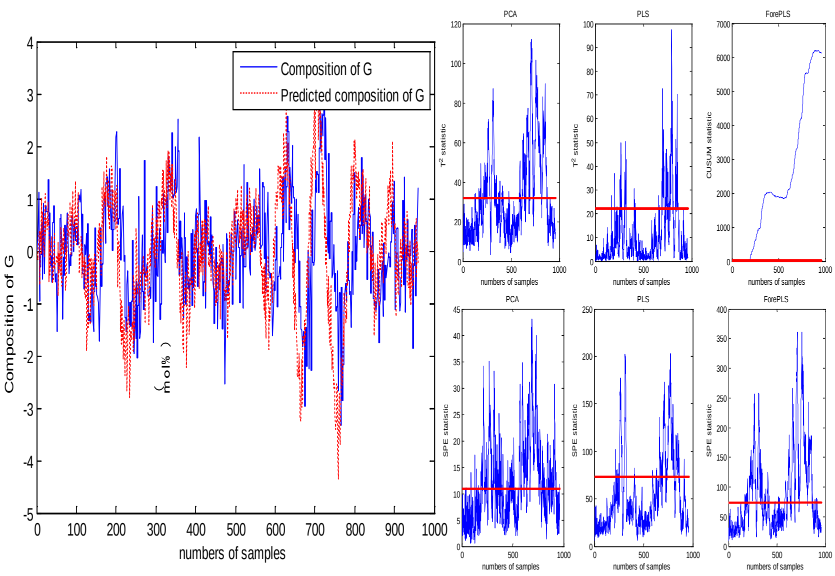
is out of the predictable range and indicates the system out of order.

•

Test results

In TE model[6], the sample data for training consists of 500 vectors of 52 dimensions and the sample data for detecting consists of 960 vectors of 52 dimensions. Faults are introduced since the 161st sample. In this article, G and H in the selection process, i.e. MEAS35 and MEAS36 are regarded as ForePLS's quality variable Y. 22 process variables: MEAS1~22 and 11 operational variables: MV1~11 are set as X. Fig.1 (a) shows that ForePLS has a better forecastability than PLS to predict the amount of reactant G in TE process.

Below is analysis of an example of a typical fault IDV(10). When fault IDV (10) occurs, the temperature of feeding C changes randomly. To verify the effectiveness ForePLS, PCA and PLS are compared. In the experiment, the latent variable number of ForePLS is 6, PLS is 9, principal component number of PCA is 15. Fig.1 shows the results of PCA, PLS and ForePLS detecting Fault IDV(10). It can be seen that the accuracies of T^2 and SPE of PCA are 45.6% and 53.9%, CUSUM and SPE statistics of ForePLS are 96.5% and 52.9%, respectively. This demonstrates that the fault detecting method based on ForePLS introduced in this article is more accurate than PCA and PLS for detecting random changes.



(a) Amount Changing Curve of G in product(b) PCA, PLS and ForePLS methods

Fig.1. The experimental results when fault 10 occur

Conclusion

A process monitoring method based on forecastable component analysis and least square regression is introduced. This method overcomes the shortage of traditional least square algorithm's inability to reflect process timing characteristics. Furthermore, it can effectively predict the tendency of system, reflect system's dynamic characteristics. Through monitoring CUSUM and SPE statistics in predictable space, this method is able to detect delicate faults like slow shift and random changing faults. The simulation results of TEP model shows that the method is more precise and effective than traditional methods PCA and PLS.

References

[1] Venkatasubramanian V, Rengaswamy R, Kavuri S N, et al. A review of process fault detection and diagnosis: Part III: Process history based methods[J]. Computers & chemical engineering, 2003, 27(3): 327-346.

[2] Joe Qin S. Statistical Process Monitoring: Basics and Beyond[J]. Journal of Chemometrics, 2003, 17(8 - 9): 480-502.

[3] Goerg G. Forecastable Component Analysis[C]//Proceedings of the 30th International Conference on Machine Learning (ICML-13). 2013: 64-72.

[4] Venkatasubramanian V, Rengaswamy R, Kavuri S N, et al. A Review of Process Fault Detection and Diagnosis: Part III: Process History Based Methods[J]. Computers & Chemical Engineering, 2003, 27(3): 327-346.

[5] Lee J M, Yoo C K, Lee I B. Statistical Process Monitoring with Independent Component Analysis[J]. Journal of Process Control, 2004, 14(5): 467-485.

[6] Downs J J, Vogel E F. A Plant-wide Industrial Process Control Problem[J]. Computers & Chemical Engineering, 1993, 17(3): 245-255.