ISA Transactions 53 (2014) 1516–1527

Contents lists available at ScienceDirect

ISA Transactions

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

Plant-wide process monitoring based on mutual information–multiblock principal component analysis

Qingchao Jiang, Xuefeng Yan *

Key Laboratory of Advanced Control and Optimization for Chemical Processes of Ministry of Education, East China University of Science and Technology, P.O. Box 293, MeiLong Road No. 130, Shanghai 200237, PR China

ARTICLE INFO

Article history:
Received 11 June 2013
Received in revised form
29 May 2014
Accepted 30 May 2014
Available online 20 June 2014
This paper was recommended for publication by Prof. A.B. Rad.

Keywords:
Plant-wide process monitoring
Mutual information
Multiblock PCA
Support vector data description

ABSTRACT

Multiblock principal component analysis (MBPCA) methods are gaining increasing attentions in monitoring plant-wide processes. Generally, MBPCA assumes that some process knowledge is incorporated for block division; however, process knowledge is not always available. A new totally data-driven MBPCA method, which employs mutual information (MI) to divide the blocks automatically, has been proposed. By constructing sub-blocks using MI, the division not only considers linear correlations between variables, but also takes into account non-linear relations thereby involving more statistical information. The PCA models in sub-blocks reflect more local behaviors of process, and the results in all blocks are combined together by support vector data description. The proposed method is implemented on a numerical process and the Tennessee Eastman process. Monitoring results demonstrate the feasibility and efficiency.

1. Introduction

With the rapid development of new sensors and data gathering equipment, multivariate statistical process monitoring (MSPM) methods, have progressed quickly in recent decades [1–7]. Among these MSPM methods, principal component analysis (PCA) usually serves as the most fundamental one and has been researched a lot [8–15]. As a dimensionality reduction technique, PCA can effectively handle high dimensional and correlated data by projecting the data into two lower-dimensional spaces: the systematic space containing most variation and the residual space with least variances. For process monitoring purpose, two statistics represented by Mahalanobis and Euclidean distances are constructed to detect the changes in the two spaces, respectively [1,16]. Successful applications of PCA monitoring have been reported in both academic and industrial areas [17–20]. However, for the plant-wide process which always has many different parts and operation units, PCA monitoring may not function well [21,22].

Given the importance in process industries, plant-wide process monitoring has gained increasing attentions in recent years. To capture the relations between complex process variables and reflect local behaviors of a process, hierarchical and multiblock statistical methods have been proposed and extensively extended [21–28]. MacGregor et al. developed the multiblock projection method that allows one to establish monitoring charts for each sub-block as well as the entire process [21]. Westerhuis et al. compared several multiblock and hierarchical PCA/partial least square (PLS) methods from a algorithmic viewpoint [22]. Qin et al. provided further analysis of the multiblock methods discussed by Westerhuis and defined block and variable contributions upon T^2 and Q statistics for decentralized monitoring [29]. Choi and Lee further analyzed the statistics and blocks suggested by Qin et al., and defined the block or variable contributions to the statistics in multiblock PLS monitoring [24]. Kohonen et al. demonstrated the efficiency and difference of multiblock PLS method in relation to normal PLS through an industrial continuous process [27]. Zhang and Ma proposed a multiblock kernel independent component analysis method which considered the non-Gaussianity for fault detection and diagnosis [30]. Ge and Song proposed a two level multiblock independent component analysis and principal component analysis method for monitoring performance and fault interpretation improvements [25].

The block division is an initial and a key step in multiblock monitoring, and the methods mentioned above usually divide the blocks under the assumption that some process knowledge is known. However, in a complex plant-wide process, the process knowledge is not always available or only a part of them is available. Meanwhile, limited by the cognition or experience, people may have different viewpoints on the division. In this case, the block division step should be carried out automatically and the totally data-driven methods are of particular interest. More

* Corresponding author. Tel.: +86 21 64251036.
E-mail addresses: jiang_qch@hotmail.com (Q. Jiang), xfyan@ecust.edu.cn (X. Yan).

http://dx.doi.org/10.1016/j.isatra.2014.05.031
0019-0578/

recently, Ge and Song proposed a distributed PCA (DPCA) for plant wide process monitoring [31]. The DPCA constructed sub-blocks through different directions of PCA principal components and then the original feature space can be divided into several sub-feature spaces automatically. This totally data-driven method has significantly promoted the development of multiblock monitoring, however, there are still some issues that need to be discussed. First, as is well known, the PCA can only model linear correlations between variables but performs poor for the non-linear relations. Second, PCA would only consider the mean and variance information and ignore the statistical distribution knowledge. These shortcomings of PCA would affect the division performance and limit its practical application.

In this study, a new totally data-driven multiblock PCA method based on mutual information (MI) (MI–MBPCA) is proposed to monitor large-scale processes. MI is a relatively new statistical analysis technique, which can quantitatively measure the statistical dependency between two random variables estimated from entropy. In comparation with PCA, it not only considers the linear correlation, but also takes the nonlinear relation into account. The divisions would be more proper. The monitoring results in all blocks are combined together using support vector data description (SVDD), which tries to find a sphere with minimum volume containing all normal observations and rejecting the abnormal ones.

The rest of the paper is organized as follows. In Section 2, the PCA monitoring, the MI and the SVDD techniques are briefly reviewed. In Section 3, the MI–MBPCA for both fault detection and diagnosis are introduced in detail. The efficiency of the proposed method is demonstrated by case studies on a numerical process and the TE process in Section 4. Finally, in Section 5, conclusions are made and some discussions are provided.

2. Preliminaries

In this section, the PCA process monitoring, the MI and the SVDD are briefly reviewed.

2.1. Principal component analysis

PCA involves the decomposition of a data matrix $\boldsymbol{X} \in R^{n\times m}$, with n observations and m measured variables, into a lower-dimension space. This subspace is spanned by the eigenvectors of the covariance of correlation matrix associated with $\boldsymbol{X}$, and contains the most variation information of the process data. The decomposition can be expressed as follows [32]:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{1}$$

where $\boldsymbol{P}$ is the loading matrix which projects $\boldsymbol{X}$ onto the principal component (PC) space; $\boldsymbol{T}$ is the PC scores, which are uncorrelated with each other. Usually, only the first few dominant PCs are selected in $\boldsymbol{P}$, and then the residual matrix $\boldsymbol{E}$ is obtained. The loading matrix $\boldsymbol{P}$ could be got through singular value decomposition (SVD) and the number of PCs could be determined by the cumulative percent variance (CPV) rule [1,32].

For process monitoring purpose, T^2 and Q statistics are constructed to monitor the changes in the dominant subspace and the residual subspace, respectively. Given an observation vector $\boldsymbol{x} \in R^{m\times 1}$, the T^2 statistic of the first k PCs can be calculated as [1,16]

$$T^2 = \boldsymbol{x}^T \boldsymbol{P}(\boldsymbol{\Lambda})^{-1} \boldsymbol{P}^T \boldsymbol{x} \tag{2}$$

where $\boldsymbol{\Lambda} \in R^{k\times k}$ is a diagonal matrix denoting the estimated covariance matrix of principal component scores. The Q statistic

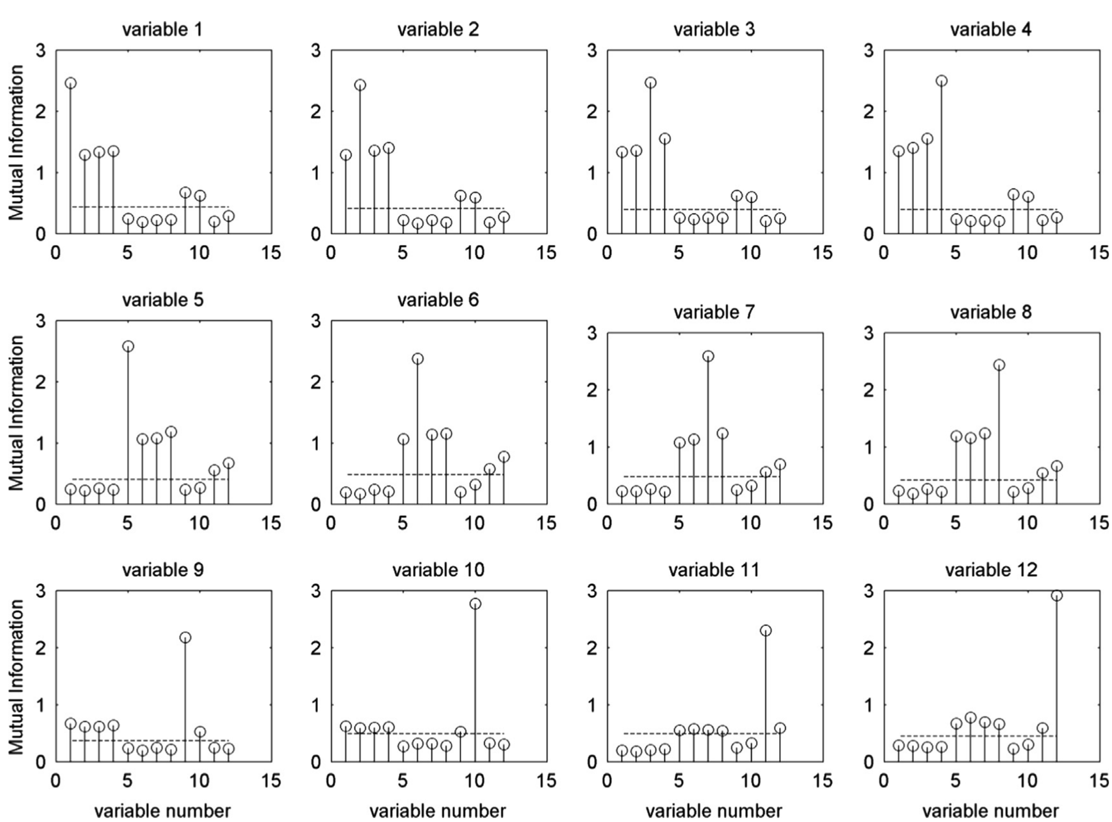

Fig. 1. The mutual information between each two variables.