Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

Map-Reduce Decentralized PCA for Big Data Monitoring and Diagnosis of Faults in High-Speed Train Bearings ⋆

Qiang Liu * **Dezhi Kong** * **S. Joe Qin** ** **Quan Xu** *

* *State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, Liaoning 110819, China (e-mail: liuq@mail.neu.edu.cn)*
** *Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, CA 90089, USA (e-mail: sqin@usc.edu)*

Abstract: Real-time fault detection and diagnosis of high speed trains is essential for the operation safety. Traditional methods mainly employ rule-based alarms to detect faults when the measured single variable deviates too far from the expected range, with multivariate data correlations ignored. In this paper, a Map-Reduce decentralized PCA algorithm and its dynamic extension are proposed to deal with the large amount of data collected from high speed trains. In addition, the Map-Reduce algorithm is implemented in a Hadoop-based big data platform. The experimental results using real high-speed train operation data demonstrate the advantages and effectiveness of the proposed methods for five faulty cases.

Keywords: Big Data Modeling, Decentralized Principal Component Analysis, Fault Diagnosis, High-Speed Train Operation Safety.

1. INTRODUCTION

The recent successes of big data technology are witnessed in many fields including manufacturing, business, and financial applications. The ultimate power of the big data technology lies in the ability to process enormous amount of data in parallel, which are usually handled by cloud computing. Typically data collection, processing, and analysis happen simultaneously and nearly instantaneously via data streaming, which enables one to merge data from multiple sites and multiple time scales to solve a problem at hand. The enlarged data scope also provides deeper understanding of the problem than using local data alone. Parallel computation and modeling are the choice for implementation.

High-speed trains (HST) are such an example where faults can happen to multiple trains at multiple routes and lead to negative effects on the overall operations, and even fatal disastrous consequences (Jia and Li, 2014). Among all possible faults, the bearing faults, including journal box fault, gear box fault, motor stator fault, motor non-driving end fault, and motor non-driving end fault, happen frequently and are of most important as they are directly related to the operation safety of the trains. In addition, high-speed trains operate as fast as 350km/h with a high traffic density. The faults should be diagnosed and maintained before the negative effects propagate to operation incidences such as an unexpected stop.

At present, bearing faults of high-speed trains are usually detected with rule-based methods, for instance, the automatic diagnosis systems of German ICE train, French TGV, and Japanese Shinkansen (Jia and Li, 2014). In addition, the signal enhancement techniques were studied to monitor the bearing faults using signal synchronization average matrix diagram, inverted spectrum whitening, and linear prediction filter Borghesani et al. (2013). A wavelet and correlation filtering based bearing fault detection method was proposed using the bearing vibration signal (Wang et al., 2011). The Welch technique based mechanical vibration evaluation and gear fault diagnosis method was proposed using the electromagnetic torque signal (Henao et al., 2010). The existing methods are essentially single signal-based modeling and alarm systems that are unable to detect incipient faulty behavior until the effects grow significantly.

On the other hand, multivariate data-driven modeling and fault diagnosis methods developed in recent years are able to detect and diagnose small faults using data correlations (MacGregor et al., 1994; Qin, 2012). The bearings of a high-speed train operate in similar loads and environment that make the temperature of each bearing correlated. The correlations can be conveniently analyzed with multivariate data modeling methods, such as principal component analysis (PCA) and dynamic principal component analysis (DPCA) (Ku et al., 1995).

One challenge in analyzing data from a network of HST operations is the massive amount of data to be handled. As

⋆ Support to this research was provided by the Natural Science Foundation of China (61490704, 61673097, 61573022), the Fundamental Research Funds for the Central Universities (N160804002, N160801001), the Fundamental Disciplinary Research Program of the Shenzhen Committee on Science and Innovation (20160207), and the Texas-Wisconsin-California Control Consortium (TWCCC).

the train operation data come with a high sampling rate, the amount of collected bearing data will be too large to be imported into the regular computation platform and the computational load for bearing data modeling could be too heavy to implement.

In this work, a Map-Reduce based decentralized modeling and monitoring algorithm is developed to implement with parallel computation a block-wise PCA algorithm for modeling large amount of data. The real-time fault monitoring and diagnosis methods are also proposed.

2. MAP-REDUCE DECENTRALIZED PCA FOR FAULT DIAGNOSIS

2.1 Map-Reduce Decentralized PCA Modeling

Given a mean-centered and appropriately-scaled data matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ consisting of n samples with m variables which can be large, principal component analysis projects $\mathbf{X}$ to a lower-dimensional space as follows:

$$\mathbf{X} = \sum_{i=1}^{A} \mathbf{t}_i \mathbf{p}_i^T + \mathbf{E} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

where $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_A]$ are the set of principal component scores for $\mathbf{X}$, $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_A]$ are the loading vectors for $\mathbf{X}$. $\mathbf{E}$ is the residuals for $\mathbf{X}$. The PCA model, i.e., loading vectors $\mathbf{P}$, and the scores can be derived by two ways. One is singular value decomposition or eigenvalue decomposition. The other one is the nonlinear iterative algorithm that is popular in chemometrics and monitoring applications (Geladi and Kowalski, 1986). Both of the PCA algorithms require the overall modeling data $\mathbf{X}$ be imported into the memory storage. However, for big data applications such as HST bearing data, a single train can generate more than 500MB of historical operation data collected during 20 hours at a sampling rate of one second for 36 bearing variables. The traditional PCA algorithms and regular computers are often incapable to handling the large volume of data.

Fortunately, in the process monitoring literature, block-wise PCA algorithms are available that are developed to make diagnosis easier by decomposing a large number of variables into blocks (Westerhuis et al., 1998; Qin et al., 2001). It is further shown that the multi-block algorithms via partitions of the original data give an equivalent global PCA model for the whole data matrix. Furthermore, Liu et al. (2013) shows that the equivalent PCA model is achieved with hierarchical partitions of the original data matrix. Recently, a distributed PCA in Ge and Song (2013) suggested to build sub-PCA models with each one involving variables related to a principal component. This algorithm, however, can result in a variable being used in many sub-models, which increases the overall computation. In addition, unlike the other multi-block PCA models, it does not yield a model equivalent to performing PCA on the global data. Therefore, in this subsection, a novel Map-Reduce based decentralized PCA is proposed by using multi-block PCA (Qin et al., 2001) for parallel computation and for large volume data.

The collected historical data matrix $\mathbf{X}$ is divided to B blocks

$$\mathbf{X} = [\, \mathbf{X}_1 \;\; \mathbf{X}_2 \;\; \cdots \;\; \mathbf{X}_B \,] \tag{1}$$

and let m_b be the number of variables in $\mathbf{X}_b$.

By block partition in the following algorithm, the block data can be successfully imported into the memory for computation.

1. Scale $\mathbf{X}_b$ to have zero mean and let $\underline{\mathbf{X}}_{b,1} \leftarrow \mathbf{X}_b$ and $i \leftarrow 1$.

2. Choose an initial $\underline{\mathbf{t}}_{T,i}$ and iterate the following equations until convergence of $\underline{\mathbf{t}}_{T,i}$:

$$\text{block loadings: } \underline{\mathbf{p}}_{b,i} \leftarrow \underline{\mathbf{X}}_{b,i}^T \underline{\mathbf{t}}_{T,i} / \left\| \underline{\mathbf{X}}_{b,i}^T \underline{\mathbf{t}}_{T,i} \right\|$$

$$\text{block scores: } \underline{\mathbf{t}}_{b,i} \leftarrow \underline{\mathbf{X}}_{b,i} \underline{\mathbf{p}}_{b,i}$$

$$\underline{\mathbf{T}}_i \leftarrow \left[\underline{\mathbf{t}}_{1,i} \;\; \underline{\mathbf{t}}_{2,i} \;\; \cdots \;\; \underline{\mathbf{t}}_{B,i} \right]$$

$$\text{super loadings: } \underline{\mathbf{p}}_{T,i} \leftarrow \underline{\mathbf{T}}_i^T \underline{\mathbf{t}}_{T,i} / \left\| \underline{\mathbf{T}}_i^T \underline{\mathbf{t}}_{T,i} \right\|$$

$$\text{super scores: } \underline{\mathbf{t}}_{T,i} = \underline{\mathbf{T}}_i \underline{\mathbf{p}}_{T,i}$$

3. Deflate block residuals

$$\underline{\mathbf{X}}_{b,i+1} \leftarrow (\mathbf{I} - \underline{\mathbf{t}}_{T,i} \underline{\mathbf{t}}_{T,i}^T / \underline{\mathbf{t}}_{T,i}^T \underline{\mathbf{t}}_{T,i}) \underline{\mathbf{X}}_{b,i}$$

where $\mathbf{I}$ is an identity matrix of $n \times n$.

4. $i \leftarrow i + 1$

5. Iterate i until all desired components are computed.

The Map-Reduce based computation scheme can be described as follows. The block loadings $\underline{\mathbf{p}}_{b,i}$ and block scores $\underline{\mathbf{t}}_{b,i}$ are computed by Map, with the block score matrix $\underline{\mathbf{T}}_i$, super loadings $\underline{\mathbf{p}}_{T,i}$, and super scores $\underline{\mathbf{t}}_{T,i}$ computed by Reduce. After that, the loadings of PCA can be obtained from the block loadings and super loadings of multiblock PCA based on the work of Qin et al. (2001). Normally the computation load and time cost will decrease for each Map when the number of Maps grows.

2.2 PCA-based Fault Monitoring and Diagnosis

Three statistical indices, i.e., squared prediction errors (SPE), T-square, and combined index, can be defined and compared with the corresponding control limits to monitor the fault in a similar way to the work of Cherry and Qin (2006). After that, contribution plot or reconstruction based contribution can be used to pinpoint the faulty variable (Alcala and Qin, 2009). Details are omitted in this work due to limited space. The interested readers are encouraged to have more information from the work of Qin (2012).

It is noted the monitoring and diagnosis results can be obtained in real-time by multiplying the real-time sample vector with the loading computed in the modeling stage.

2.3 Dynamic PCA with Map-Reduce

Dynamic extension is necessary and an easy way is to include lagged value of data in the augmented data matrix Ku et al. (1995). The dynamic PCA builds dynamic relations of the variables according to the PCA decomposition of the augmented data with a the time orders.

Decentralized DPCA for large amount of data can also be achieved similar to the decentralized PCA algorithm. Faults can be detected thereafter according to the unexpected deviation of the dynamic relations. The SPE statistic is used for this work.

3. IMPLEMENTATION AND EXPERIMENTS FOR HST FAULT DIAGNOSIS

The proposed Map-Reduce decentralized PCA for modeling and fault diagnosis is implemented in the Hadoop platform of big data analytics. The overall system is composed of two parts, i.e., off-line modeling system and on-board fault diagnosis system. The sensor data generated on-board are first transferred to store using the Hadoop distributed file system (HDFS). The Map-Reduce decentralized PCA modeling algorithm is achieved in the cloud computing cluster with distributed computers. The decentralized PCA model is then used for on-board real-time monitoring and diagnosis.

The platform uses Hadoop as its basic framework and combines HDFS and HBase databases as the data storage framework (Gudmundsson et al., 2012). A lossless data compression method is presented to reduce the data storage space and improve storage efficiency.

The proposed methods are implemented using the platform to model the bearing temperature data collected from 36 variables, including 4 bearings each with 7 temperature sensors, and additional 8 bearing box. The data are collected at a sampling of 1 second and more than 60,000 samples of almost 20 hours are used for modeling. The historical operation data is too big to be analyzed with regular computers. Map-Reduce decentralized PCA that divides the original data into the memory of cloud computing cluster is implemented. The computation speed of Map-Reduce PCA with 8 nodes is almost 7 times faster than the one with single nodes.

Five faulty cases, including Case 1 (journal box fault), Case 2 (gear box fault), Case 3 (motor stator fault), Case 4 (motor non-driving end fault), and Case 5 (motor non-driving end fault) are used to demonstrate the proposed decentralized PCA and dynamic extension of decentralized PCA with the traditional rule based methods.

The fault monitoring results are shown in Fig. 3-Fig. 7 while the red line indicates where the true fault occurred. The faults are successfully detected by both of decentralized PCA and dynamic PCA with Map-Reduce. For example, faults in Case 1 are detected by combined index of decentralized PCA around Time 3280(second) in Fig. 3(a) and SPE index of dynamic extension of decentralized PCA around Time 3015(second) in Fig. 3(b). In addition, PCA based contribution plots for the five faulty cases are provided as shown in Fig. 8(a)-(e), respectively. The faulty variables are successfully identified. As shown in Fig. 8(a), the third variable, i.e., journal box fault, is identified as the faulty bearing, which is also verified by the practitioners.

The results are summarized in Table 2 while the alarm limits for single variable based fault diagnosis are listed in Table 1. As the false alarm rate and missed alarm rate are trade-off, the thresholds for single variables are selected in Table 1 according to priori knowledge in practice and receiver operating characteristic curve (ROC) to eliminate false alarms. From Table 2, all of the five faulty cases are successfully detected by the three methods. The detection time of decentralized PCA and dynamic extension of decentralized PCA is earlier (800 seconds to 8,000 seconds) than that of rule based method. Compared to decentralized PCA, dynamic PCA with Map-Reduce can detect the fault much earlier which demonstrates an advantage.

Table 1. Thresholds for Rule Based Method.

Sensor Location	Warning (°C)	Alarms (°C)
Journal box	100	120
Gear box	110	130
Motor stator	160	180
Motor non-driving end	90	110
Motor driving end	110	130

4. CONCLUSIONS

In this paper, multivariate latent variable modeling method is proposed for bearing fault diagnosis of high-speed trains: i) the proposed decentralized PCA algorithm is suitable for Map-Reduce based implementation of PCA for large amount of data; ii) the decentralized PCA and dynamic PCA based methods demonstrate earlier fault detection compared to the rule based alarms, while dynamic PCA outperforms PCA; iii) the contribution plot of decentralized PCA successfully pinpoints the faulty bearing.

REFERENCES

Alcala, C. and Qin, S. (2009). Reconstruction-based contribution for process monitoring. *Automatica*, 45, 1593–1600.

Borghesani, P., Pennacchi, P., Randall, R.B., Sawalhi, N., and Ricci, R. (2013). Application of cepstrum pre-whitening for the diagnosis of bearing faults under variable speed conditions. *Mechanical Systems and Signal Processing*, 36, 370–384.

Cherry, G. and Qin, S. (2006). Multiblock principal component analysis based on a combined index for semiconductor fault detection and diagnosis. *IEEE Transactions on Semiconductor Manufacturing*, 19, 159–172.

Ge, Z. and Song, Z. (2013). Distributed PCA model for plant-wide process monitoring. *Industrial and Engineering Chemistry Research*, 52, 1947–1957.

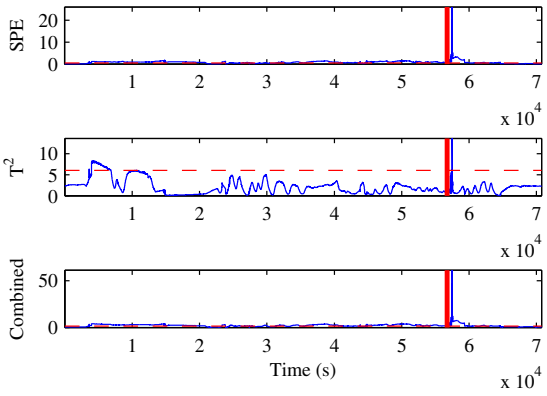
Geladi, P. and Kowalski, B. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185, 1–17.

Gudmundsson, G.P., Amsaleg, L., and Jonsson, B.P. (2012). Distributed high-dimensional index creation using Hadoop, HDFS and C++. In *CBMI*, 1–6.

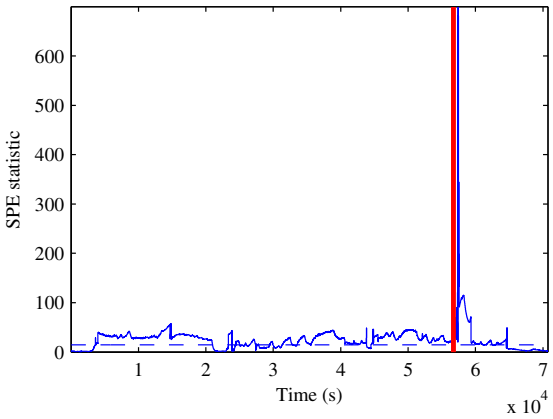
Henao, H., Kia, S., and Capolino, G. (2010). Torsional-vibration assessment and gear-fault diagnosis in railway traction system. *IEEE Transactions on Industrial Electronics*, 58(5), 1707–1717.

Jia, H. and Li, L. (2014). Thinking on improving the manufacturing level and operation quality of high speed train in China. *Chinese Railways*, 1, 30–33.

Ku, W., Storer, R., and Georgaki, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 30, 179–196.

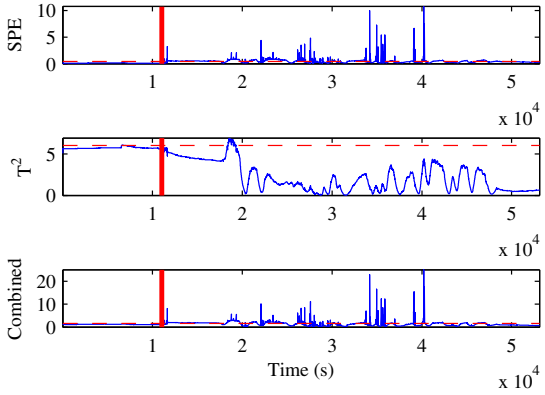


(a) Decentralized PCA

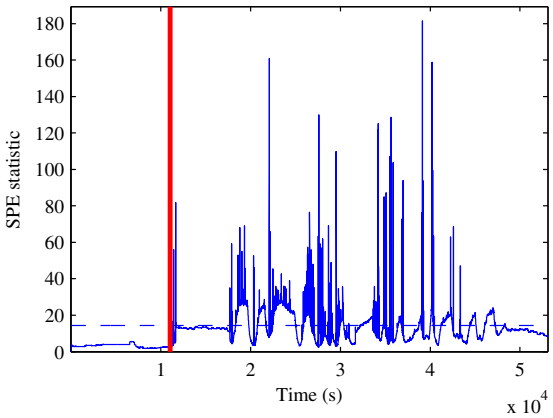


(b) Dynamic Extension of Decentralized PCA

Fig. 1. Monitoring result for faulty case 1.

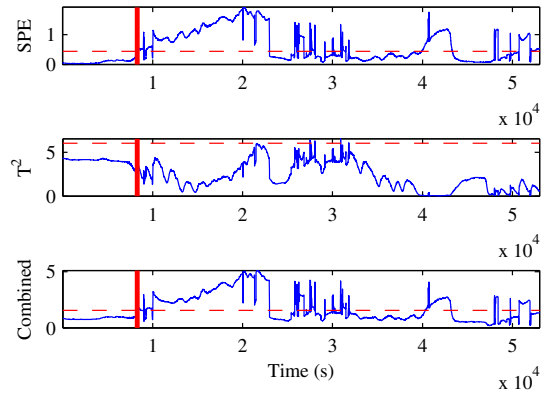


(a) Decentralized PCA

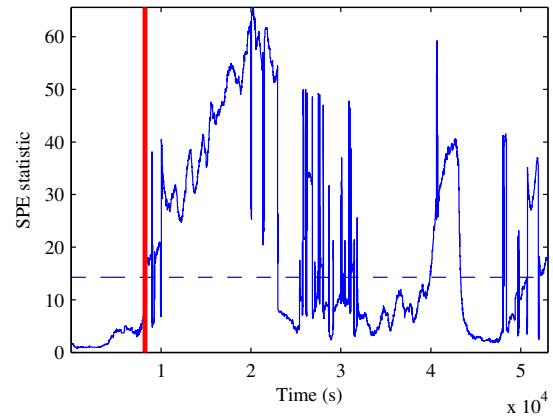


(b) Dynamic Extension of Decentralized PCA

Fig. 2. Monitoring result for faulty case 2.



(a) Decentralized PCA

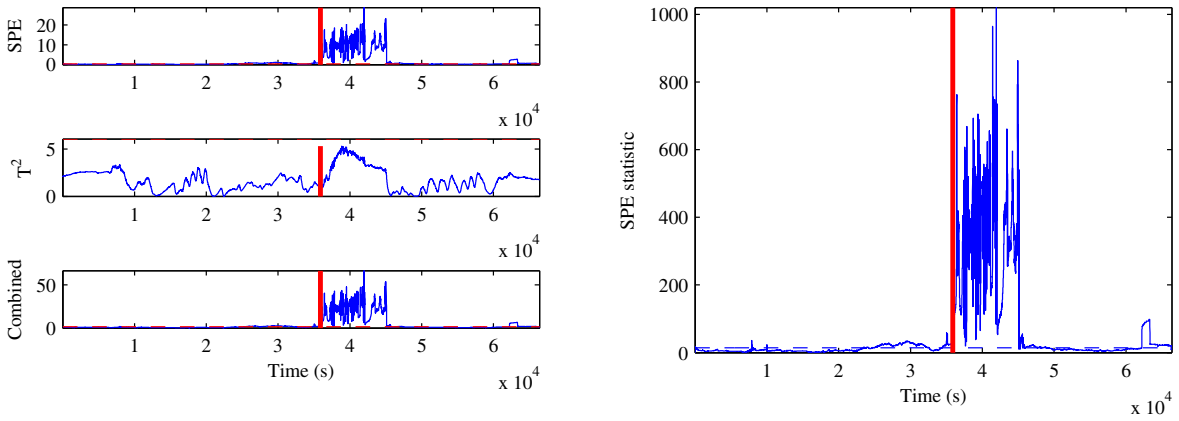


(b) Dynamic Extension of Decentralized PCA

Fig. 3. Monitoring result for faulty case 3.

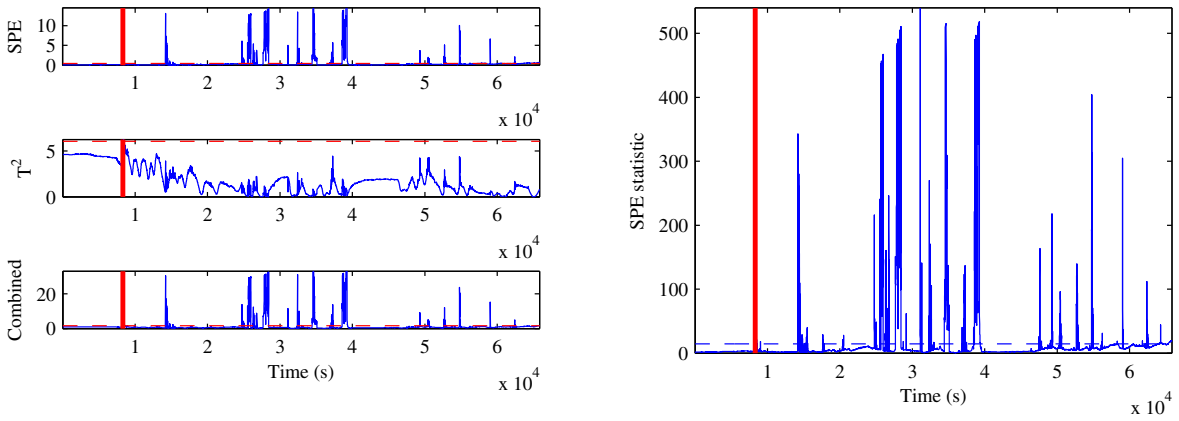
Table 2. Comparison Results of Rule-Based Method, PCA, and DPCA.

Faulty Cases	Faulty Variable No.	Detected Time		
		Rule-Based	PCA	DPCA
Journal box fault	3	57420	56830	56760
Gear box fault	26	19340	11090	11060
Motor stator fault	13	9000	8240	8233
Motor non-driving end fault	29	35930	35030	34515
Motor driving end fault	28	14200	8340	8323



(a) Decentralized PCA (b) Dynamic Extension of Decentralized PCA

Fig. 4. Monitoring result for faulty case 4.



(a) Decentralized PCA (b) Dynamic Extension of Decentralized PCA

Fig. 5. Monitoring result for faulty case 5.

Liu, Q., Qin, S., and Chai, T. (2013). Decentralized fault diagnosis of continuous annealing processes based on multi-level PCA. *IEEE Transactions on Automation Science and Engineering*, 10, 687–698.

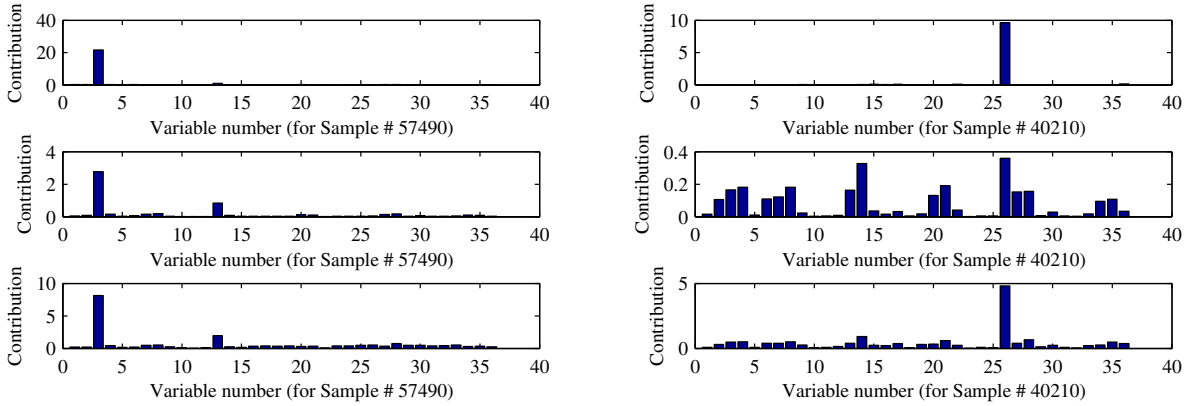
MacGregor, J., Jaeckle, C., Kiparissides, C., and Koutoudi, M. (1994). Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal*, 40(5), 826–838.

Qin, S.J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, 36(2), 220–234.

Qin, S., Valle, S., and Piovoso, M. (2001). On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics*, 15, 715–742.

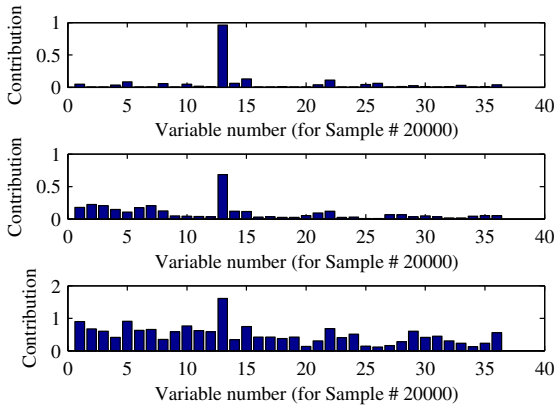
Wang, S., Huang, W., and Zhu, Z. (2011). Transient modeling and parameter identification based on wavelet and correlation filtering for rotating machine fault diagnosis. *Mechanical Systems and Signal Processing*, 25, 1299–1320.

Westerhuis, J.A., Kourti, T., and MacGregor, J.F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12, 301–321.

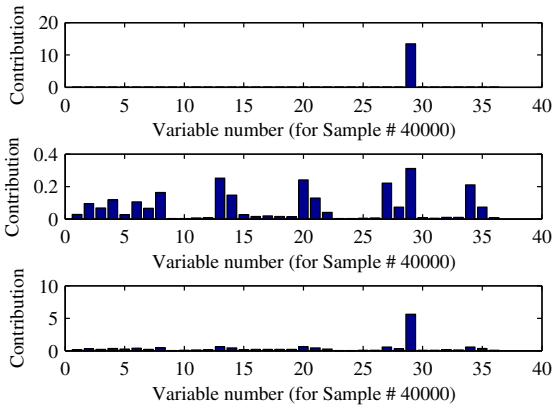


(a) Case 1

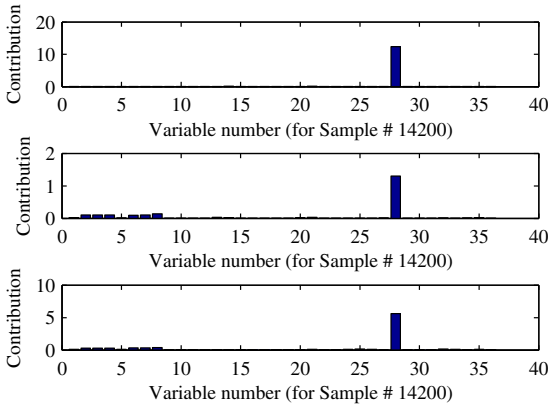
(b) Case 2



(c) Case 3



(d) Case 4



(e) Case 5

Fig. 6. Faulty variable diagnosis results for PCA.