SPARSE PRINCIPAL COMPONENT ANALYSIS (SPCA) TO FACILITATE KNOWLEDGE DISCOVERY AND PROCESS MONITORING

Ahmet Palazoglu[*,1], and Murat Kulahci[2]
[1]University of California, Davis, Davis, CA 95616, USA
[2]Technical University of Denmark, 2800, Kongens Lyngby, Denmark and Luleå University of Technology, 97187, Luleå, Sweden

Abstract Overview

Process systems possess complex dynamics and multivariate interactions. Data collected from such systems may contain information regarding such interactions and can be used to verify known relationships and/or help facilitate the discovery of new and novel relationships. The traditional use of the principal component analysis (PCA) can be extended to the discovery of novel relationships by imposing sparsity constraints on the principal component loadings. Additional benefit of the sparse PCA (SPCA) emerges in fault diagnosis which is a key component of process monitoring strategies. This paper will introduce methodologies that rely on SPCA in uncovering multivariate relationships and fault signatures.

Keywords

Sparse principal component analysis, process discovery, fault detection and diagnosis.

Introduction

The use of analytics to uncover hidden features and trends in data collected from various platforms has been the focus of an increased number of studies in the last decade. With the recent emphasis on initiatives such as Industry 4.0 and digital transformation of corporate enterprises, the use of data for decision making processes both in the long term and short term gained significant momentum. This is accompanied by the abundance of data collected in various forms from a wide variety of sources. A report published in 2011 by McKinsey Global Institute (McKinsey, 2011) captures the impact of big data on corporate management and operations and outlines the challenges and opportunities for innovation and productivity.

The manufacturing industries are also going through a transformation as big data is fueling the era of optimized smart manufacturing (O'Donovan et al., 2015). A recent perspective article by Reis et al. (2015) articulates the impact of big data and the associated research on the process industries. While the benefits of exploiting big data for decision making is clear, they also caution as to the improper use of data, and overreliance on the results and predictions. A potential pitfall generated by the abundance of data is to overlook the knowledge embedded in process models that capture fundamental physico-chemical relationships that underlie the operation and dynamics of process systems. An important question and a challenge is the incorporation of both data-driven and theory-driven knowledge into the decision making. In this paper, we first explore the use of data-driven techniques not only to verify known relationships among process variables but also to create an environment that can allow discovery of new knowledge in the form of correlated (not necessarily causal) relationships among the key process variables. Such an exercise can further enable the application of real-time process monitoring techniques (Qin, 2015) by providing the

* To whom all correspondence should be addressed

engineers and operators with more focused variable groupings, facilitating root-cause analyses.

We introduce sparse principal component analysis (SPCA) as a technique that can (1) provide the means towards new knowledge discovery, and (2) improve substantially the diagnosis of abnormal/faulty events in process systems. The Tennessee Eastman Process (TEP) is used to illustrate the salient points.

Sparse Principal Component Analysis (SPCA)

Principal component analysis (PCA) is the most commonly used multivariate technique (Joliffe, 2002). In its simplest form, raw data can be measured in two dimensions: number of samples (n) and number of variables (p) where both n and p can be large. PCA extracts the essential information from p variables of the original dataset into k retained principal components (PCs). Often, k is much smaller than p. Nevertheless, all p variables have non-zero loadings on the derived PCs and this, in turn, may confound the interpretation of PCs especially when p is large.

SPCA is a recent technique proposed for producing PCs with sparse loadings via the variance-sparsity trade-off (Trendafilov, 2014, Jolliffe and Cadima, 2016). Zou et al. (2006) proposed a strategy to obtain sparse loadings by reformulating the PCA as a regression problem and imposing LASSO (elastic net) constraints on the L1 norm of the regression coefficients (sparse loadings). SPCA modifies the traditional PCA algorithm whereby the interpretability is improved through limiting the number of non-zero coefficients (loadings). The number of non-zero loadings (NNZL) are referred to as the cardinality or L0 norm of the corresponding component. We have shown that the ability to diagnose a fault is greatly enhanced by identifying the components with most contributions to the fault (Gajjar et al., 2018).

Merola (2015) proposed an algorithm where the sparse loadings are computed by a reformulation of PCA as a regression with backward elimination. Known as least squares sparse principal component analysis (LS SPCA), it has a number of advantages by preserving the uncorrelated nature of PCs, allowing one to control the cardinality and variance captured by the PCs. This work will use Zou's algorithm in knowledge discovery and the Merola's algorithm for fault detection and diagnosis

Forward SPCA

Once k and the loading matrix are obtained for a given dataset, the optimum sparse loadings are sought by considering the trade-off between sparsity and cumulative percent variance (CPV) explained (Gao et al, 2018). In this process, the determination of the NNZL on each SPC is critical. As a heuristic for process systems, one can consider the basic pairwise causality between a manipulated variable and its corresponding controlled variable. In most cases this would be the minimal expected relationship among process variables. The forward SPCA approach is initialized by this heuristic rule to find the optimum NNZL for each SPC sequentially: i.e., the optimal NNZL is found for the first SPC with the maximum CPV by fixing that of the other (k-1) SPCs. After this, the optimal NNZL is found for the second SPC with the maximum CPV by fixing that of the other (k-1) SPCs. These steps are repeated until the difference between the new CPV and the old CPV reaches a pre-defined limit, thus revealing no further improvement in the variance explained by decreasing sparsity.

The goal of forward SPCA is to determine if significant insight can be gained by systematically sacrificing sparsity. With this forward SPCA method, only the first few SPCs contain more non-zero loadings, which achieving the desired sparsity and insight. Furthermore, one can readily distinguish and grasp the dominant information patterns captured by each SPC through the change of the loadings in each search step. Once the optimum sparse loadings of SPCs are obtained, one can extract valuable process knowledge by attempting to interpret them. The dominant process variables having relatively high loadings on one SPC are most likely to be strongly correlated due to their inherent operational characteristic. On the other hand, the process variables having relatively small loadings on one SPC would have weak correlations with the other variables.

Case Study

TEP (Downs and Vogel, 1993) has five major unit operations: an exothermic reactor, a product condenser, a vapor-liquid separator, a recycle compressor and a reboiled product stripper. A total of 33 variables that consist of 22 continuous process measurements and 11 manipulated variables are selected in this study. 960 normal samples with sampling rate of 3 min are used to build the SPCA model (http://web.mit.edu/braatzgroup/links.html).

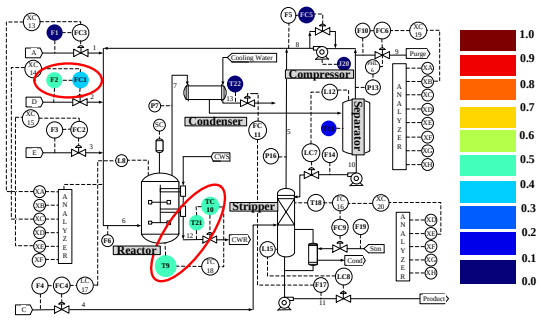

Figure 1. NNZL for SPC 3. Color code indicates the magnitude of loadings

Figure 1 shows the NNZL for SPC 3 where the highlighted measured variables indicate a strong correlation between the reactor temperature control loops and the Feed

D flow controllers. This is an unexpected relationship and demonstrates the potential for knowledge discovery as to the influence of component D on the reactor operation.

Parallel Coordinates

Recently, Gajjar et al. (2016) demonstrated the advantages of parallel coordinates and PCA for real-time process monitoring. With parallel coordinates, the perception barrier of 3-dimensional visualization can be broken facilitating the visualization of multidimensional problems and studying trends in multivariate datasets. Parallel coordinate visualization not only aids our pattern recognition ability without having to visualize the data in a combinatorial fashion but also enables us to extract insights from the dataset. In this case, the resulting SPCs are arranged in order of decreasing variance captured, and since SPCs are uncorrelated, the order becomes irrelevant.

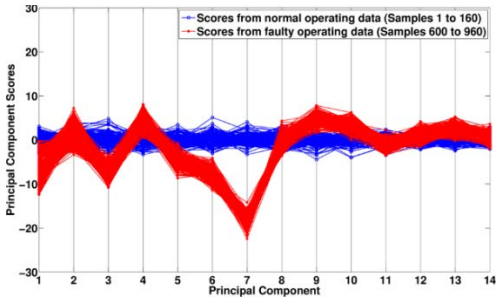

Figure 2. Visualizing 14 PCs in parallel coordinates for fault #1 with 85% CPV PCA.

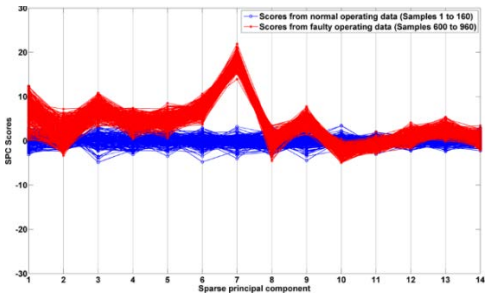

Figure 3. Visualizing 14 SPCs in parallel coordinates for fault #1 with 85% CPV PCA.

Fault Diagnosis with SPCA and Parallel Coordinates

Here LS SPCA is used to develop a methodology for detection and diagnosis of faults. In the off-line step, data is acquired under the normal operating regime and standardized. This is followed by the definition of the cumulative percent variation threshold target and the application of LS SPCA to generate the loading matrix. Next, the control limits for the SPCs are established and the Random Forest algorithm (Gajjar et al., 2018) is trained on scores and residuals obtained from faulty data. The on-line step involves the acquisition of new data, its projection to get the new scores and checking to see if the projection is within control limits. If not, a fault is declared and the diagnosis step is initiated using Random Forest.

Case Study

An LS SPCA model is developed for the TEP data to explain 85% variance with 14 SPCs. Figures 2 and 3 compare the resulting PCs and SPCs in parallel coordinates where the scores from faulty process operation can be easily distinguished from the normal operating region for fault #1. As expected, the fault signatures are visually different for the same fault in LS SPCA vs. PCA. This greatly facilitates the fault diagnosis performance of the proposed method.

Conclusions

Sparse PCA is introduced as a means to discover known and/or hidden relationships among process variables and to diagnose faulty events in process operation. More discussion of results will be available at the conference.

References

Downs, J. J., Vogel, E. F. (1993). A plant-wide industrial process control problem. *Comp. Chem. Eng. 17*, 245.

Gajjar, S.; Palazoglu, A. (2016). A data-driven multidimensional visualization technique for process fault detection and diagnosis. *Chemom. Intel. Lab. Systems, 154*, 122.

Gajjar, S., Kulahci, M., Palazoglu, A. (2018). Real-time Fault Detection and Diagnosis using Sparse Principal Component Analysis. *J. Proc. Control. 67*, 112.

Gao, H., Gajjar, S., Kulahci, M., Zhu, Q., Palazoglu, A. (2016). Process Knowledge Discovery Using Sparse Principal Component Analysis. *Ind. Eng. Chem. Res. 55*, 12046.

Jolliffe, I.T., Principal Component Analysis. (2002) *Springer*. New York, NY.

Jolliffe, I.T., Cadima, J. (2016). Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A.* 374, 20150202.

McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey & Company.

Merola, G. M. (2015). Least squares sparse principal component analysis: a backward elimination approach to attain large loadings. *Australian & New Zealand J of Statistics*. *57*, 391.

O'Donovan, P., Leahy, K., Bruton, K., O'Sullivan, D.T.J. (2015). Big data in manufacturing: a systematic mapping study. *J Big Data*. 2:20.

Qin, S.J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Ann. Rev. in Control*. *36*, 220.

Reis, M.S., Braatz, R. D., Chiang, L.H. (2016). Big Data: Challenges and Future Research Directions. *Chem. Eng. Prog*. March, 46.

Trendafilov, N.T. (2014). From simple structure to sparse components: a review. *Comp. Statistics*. *29*, 431.

Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse Principal Component Analysis. *J Comp. Graph. Statistics*. *15*, 265.