A Data-Driven Statistical Approach for Monitoring and Analysis of Large Industrial Processes[*]

A. Montazeri *, M.H. Ansarizadeh, M.M. Arefi*****

** Engineering Department, Lancaster University, Lancaster, United Kingdom, (e-mail: a.montazeri@lancaster.ac.uk).*
***Faculty of E-Learning, Shiraz University, Shiraz, Iran (e-mail: ansarizadeh@gmail.com)*
**** Department of Power and Control Engineering, Shiraz University, Shiraz, Iran (e-mail: arefi@shirazu.ac.ir)*

Abstract: Monitoring and fault detection of industrial processes is an important area of research in data science, helping effective management of the plant by the remote operator. In this article, a data-driven statistical model of a process is estimated using the principal component analysis (PCA) method and the associated probability density function. The aim is to use the model to monitor and detect the incurred faults in the industrial plant. The experimental data are collected by finding the suitable subsystems of a Recycle Gas in Ethylene Oxide production process, and a subset of nine variables are extracted for further statistical analysis of the system. The performance of the developed model for monitoring purpose is evaluated by using faulty and close to faulty inputs as the new test data. Copy-right © 2019 IFAC

Keywords: Statistical Process Monitoring, Fault detection, Probability Density Function, Principal Component Analysis, Kernel method, Process Control.

1. INTRODUCTION

Statistical process monitoring is a widely used technique for fault diagnosis of chemical processes to improve process quality and productivity. The principal component analysis (PCA) is one of the most popular methods for this purpose. Reduction of faults, improving process the safety, and product quality are the main concerns for the operators in industrial plants. Several methods have been used so far for this purpose, and each of them could be useful under specific circumstances.

One approach in monitoring is using the model based techniques, which requires an exact mechanistic model of the system or process under the study (Venkatasubramanian, et al., 2003), (Montazeri, et al., 2017), (Montazeri & Ekotuyo, July 2016). The great advantage of the model based method is that it provides a more accurate monitoring, and hence decision making process compared to the other techniques. Nevertheless, the disadvantage of this approach is the complexity of the modern industrial processes, and the fact that achieving an accurate model of the system or process requires optimisation and parameter estimation techniques (Montazeri, A. & Poshtan, J., 2009).

Therefore, the second mothed used for monitoring of an industrial system is the so-called knowledge based method (Pokkunuri, 1994). This technique relies on the real time knowledge of the system behavior and experience of remote operator who is expert in working with the system. It is important to note that, this method requires long-term operation with the process to gain deep knowledge and experience by the operator. Despite its cognitive nature, this technique is time consuming, especially under difficult operating conditions of the plant.

It should be noted that both model based and knowledge based methods are still popular, when the model of the system is not too complex, or it is possible to get accurate knowledge about the system operation. However, a more viable approach towards monitoring, diagnosis, and prognosis of the industrial systems is to use the data-driven techniques (MacGregor & Cinar, 2012), (Yin, et al., November 2014). In this category, the information is achieved by processing and analysis of the recorded noisy data. In this way, it is possible to find and model the relationship between different system variables. In these methods, the collected data are used to extract and construct some statistical information for the monitoring purpose. Therefore, these techniques sometimes are referred to as the multivariable statistical process monitoring (MSPM). Amongst various techniques applied for MSPM, Principal Component Analysis (PCA) (Zhou, et al., 2016), (Shlens, April 2009), and Partial Least Squares (PLS) are the most well-known ones.

Extracting the principal components from the system input/output data leads to T2 Hotelling, and Squared Prediction Error (SPE), which are quite important parameters. PCA and PLS are useful tools to deal with the high dimensional systems, and find the highly correlated system variables for monitoring purpose. Some more recent techniques used for monitoring a dynamic system or process are referred to as probabilistic PCA (Wang, et al., 2002), factor analysis (FA), independent component analysis (ICA) (Lu, et al., 2008), kernel PCA, and support vector data description (SVDD). In this investigation, the kernel method reported in (Ni Zhang, et al., 2014), is utilized to achieve probability density function (PDF) of the collected data from the process, and a sort of kernel PCA technique is applied to reduce the dimensions of the system. It is worth noting that although the traditional PCA will not give a reasonable output when the process data does not have a strictly Gaussian distribution, ICA algorithm can be more useful in such scenarios. In the same line, the conventional MSPM may fail and will not satisfy the requirements to describe the system when it is nonlinear, time-varying, or multiple operation modes are applicable. Some nonlinear modelling approaches, used for process monitoring in these cases are neural network (Zamprogna, 2003), kernel PCA, and linear subspace methods. Moreover, under time-varying conditions, and when multiple operation modes are applicable, adaptive PCA, recursive PLS, Moving Window PCA (Jeng, 2010), multi-model method, local model approach, and Bayesian inference method are suitable machine learning techniques. Since in this study, we are dealing with a nonlinear system, and the multiple operation mode assumption is valid, the effectiveness of PCA with moving window approach is studied. Statistical monitoring methods heavily rely on the data collected from the field using measurement instruments. Although recent advances in measurement technologies and control system designs have made collecting data rather easy, there are huge numbers of measuring variables in major industrial plants, and hence clearance of these variables and their corresponding data is yet an important step. To address this problem in plant-wide or large-scale process monitoring problems, the multi-block PCA and multi-block PLS are used. The system considered for study here is a Mono Ethylene Glycol plant, and the measurement data is gathered from the MORVARID MEG petrochemical company, located in Asaluyeh, Iran.

The article is organized as follows. After this introduction, the mathematical and theoretical backgrounds and concepts are introduced, in section 2. In section 3, the process units and analysis of various subsystems of the process are explained. In section 4, the clearance of data gathered from the process unit and the performance of the proposed monitoring scheme is presented using a simulation study. Finally, conclusions are drawn in Section 5.

2. MATHEMATICAL AND THEORETICAL DEFINITIONS

The probability density function is a basic concept in statistics and probability. By knowing the probability density function of a random variable, it is possible to fully describe the random variable. Basically, two major techniques exists for this purpose: parametric estimation method, and non-parametric estimation method. In parametric estimation technique, it is assumed that the data are belonging to a similar probabilistic distribution, such as normal or Gaussian distribution. In this case, the unknown parameters i.e. mean μ and variance σ^2 should be estimated. In non-parametric estimation technique, the probability function itself is unknown and estimation of the PDF is yet remained to be solved. The random variable x is known to be a normal variable (or Gaussian variable), if its probabilistic density function is defined as

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

For $-\infty < x < +\infty$ and μ and σ are both real values and σ is greater than zero. A special case of this function is when μ=0 and σ=1, and hence it called a standard normal distribution. In a simple estimator, if the random variable x is assumed to have a probabilistic density function of f

$$f(x) = \lim_{h\to 0} \frac{1}{2h} p(x-h<X<x+h) \qquad (2)$$

Where h is the width, and for each h, the function $p(x-h<X<x+h)$ can be estimated by a fraction of the observed samples in the interval $x-h<X<x+h$. By defining the weight function (w)

$$w(x) = \begin{cases} \frac{1}{2} & |x|<1 \\ 0 & \text{other places} \end{cases} \qquad (3)$$

The simple estimator is defined as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x-X_i}{h}\right) \qquad (4)$$

A *simple estimator* is the step function with the mean value of the measurement points, and has a fixed value in a certain interval and drops drastically at the ends. This results in a discontinuous and non-differentiable function. Another estimator is kernel estimator. The *kernel estimator* is generalisation of the simple estimator, explained above aiming to overcome its drawbacks. If the weight function *w* in the simple estimator is replaced with the function *K,* known as the core, the result is called the kernel estimator. For the kernel estimator with core of *K*

$$\int_{-\infty}^{\infty} K(x)dx = 1 \qquad (5)$$

Normally, the core *K* is a symmetric probability density function. Therefore, the kernel estimator with core *K* is defined as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \qquad (6)$$

Where h is the window width or smoothing parameter in (6). For a huge number of sensors used to record and monitor the performance of the system, it is important to reduce the size of the variables to get the minimum principal components. For example, in case that the variables have correlations, or in case that the remote operator has no time or expertise to cluster the data recorded from the sensors.

In such situations, the principal components analysis (PCA) is a method to extract the main and most important components from the huge number of existing variables. In fact, PCA will reduce the dimension of variables. Principal component analysis works on covariance or correlation matrix of the recorded data, if the data are numeric and normalised. The principal component is a linear combination of the main predictors in the dataset. The main component of the dataset can be written based on main predictors as

$$z_1 = \varphi_{11}x_1 + \varphi_{21}x_2 + \cdots + \varphi_{p1}x_p \qquad (7)$$

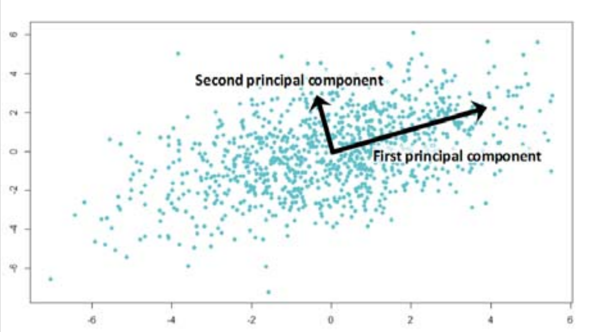


Fig. 1. The orthogonality of the vectors of the first and second principal components.

Here z_1 is the first principal component and φ_{pi} is the load vector of the ith component. The number of load vectors is limited by the sum of squares equal to one. The reason for this choice is that a large amount of loads may lead to a very large variance. This value also defines the direction of the main component in z_1 to data which is the most diverse. Also x_1 to x_p are normalized predictions. The average of the normalized predictions is zero and their standard deviation is equal to one. Therefore, the first principal component is a linear combination of the main predictions, which considers the largest variance in the dataset. In another word, this component determines the range of change in the first component, which contains most of the information. The range of change on other components is less than the first one.

The second component can be written in a similar way. The second principal component z_2 is a linear combination of the main predictors that preserves the remaining variance in the dataset and is uncorrelated with, z_1 therefore

$$z_2 = \varphi_{12}x_1 + \varphi_{22}x_2 + \cdots + \varphi_{p2}x_p \tag{8}$$

If the first two principal components are uncorrelated, their directions should be orthogonal. This is illustrated in Fig. 1 for a sample dataset. All other principal components will follow a concept similar to what is stated here. In the other word; they maintain the amount of the remained variance, without correlating with the previous components. Generally, with a dataset of n×p dimension, *min (n-1, p)* of the components are constructed.

3. PROCESS OF MEG

The data use in this study are achieved from a model of MEG[1] plant as a petrochemical process from MORVARID petrochemical company. The MEG plant has the most critical petrochemical process. The reactions inside the process can be quite critical, if the incurred faults are not recognized at the early stages.

Any large-scale petrochemical plant is made of several subsystems and the data analysis algorithm should be applied to all subsystems. MEG petrochemical plant has five main subsystems or units, which are numbered from 100 to 500 in order: these can be referred to as EO Reaction unit, CO_2 Removal and EO Recovery unit, Light Ends Removal unit, Glycol Reaction and Recovery unit and Glycol Purification Section unit (Shell Global Solutions, 2016). It is important to choose units that are critical in terms of fault detection and prediction for the remote operator. An important subsystem contained as a part of units 100 and 200 of the plant is RYG loop. This is a closed loop system in which the gas inside is mostly Ethylene Oxide. EO consists of the Ethylene gas and Oxygen gas, generated, as a result of the main reactions inside the MEG plant. Because of the hazards resulting from the reaction, detecting and predicting the faults is quite important. This loop is started by merging ethylene and oxygen (without reaction) as the main feed gases inside the oxygen mixing nozzle. Before reaction is taking place, the gas should be preheated by a heat exchanger called feed/product exchanger. The required heat for this exchanger is generated by the reactor products. After pre heating of gas, the reaction can take place. Production of EO gas is the result of two continues tubular reactors, working in parallel. This is an exothermic reaction. The product gas is too hot and needs to be cooled down. This process is carried out in two steps; first with the primary product cooler, and then with the feed/product exchanger use to preheat the reactors feed. The heat produced from the primary product cooler will be used to generate steam.

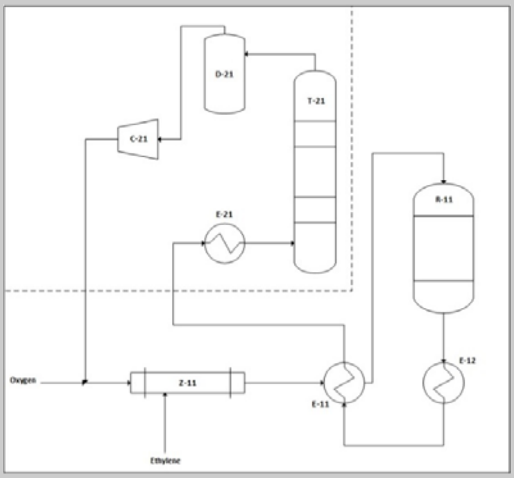


Fig. 2. Schematic of RYG unit working as a loop.

The EO gas in RYG unit will go to EO absorber column. The main job of this column is to absorb the EO gas and CO_2, but there will also be some byproducts such as O_2, C_2H_4, CO. Some EO is also leaving from the top of the tower and return to RYG loop.

It should be noted that the propulsion force for the gas is a compressor and the compressor also exists in the loop. The recycle gas compressor works similar to the heart of the loop and the gas will not flow without this component.

A schematic block diagram of the process is shown in Fig. 2. Having the basic understanding of the system, it is possible to analyse the variables shown in the loop in Fig. 2. There are several types of measurement instrument such as flow, pressure and temperature as shown in this figure. Before clearing the variables and the collected data, an investigating for the specific circumstances of this process should be done. As it's mentioned before, the main reaction of this system is EO reaction inside the reactors. These reactors work with

[1] Mono Ethylene Glycol

catalyst. Two types of catalysts are used in the MEG plants for this purpose; High Activity (HA) catalyst and High Selectivity (HS) catalyst. Each of these catalysts has their own circumstances and can affect other variables. Therefore, if the statistical model is specified according to one of the catalysts, it will not be valid if the operator changes the catalyst types and a new model should be constructed.

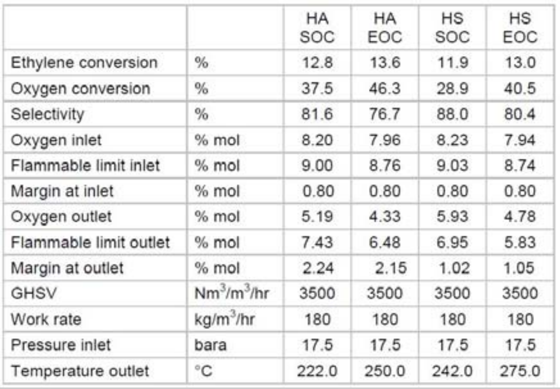
It should be noted that from the start of cycle (SOC) to the end of the cycle (EOC), the process has different behaviors, according to the catalyst life-time. In another word, when the catalyst is getting rich and aged, the process conditions are changed. These new conditions are considered as the multiple operation modes. Table 1 shows different operation modes of the process when the catalyst is changing. To address this issue, the moving window method is used in the numerical study reported in the next section.

4. PREPARATION AND IMPLEMENTATION OF THE ALGORITHMS

4.1 Clearing Variables and Process Data

Since many variables and instruments exists in the process loop shown in Fig. 2, the first step is to clear and process the data. The first thing to address is to identify the difference between various types of instruments, their functions, and the purpose of their use. Signals acquired from the sensors can have several purposes. For example, some of them are used only as the indicator, which are less important. Some others have also the alarm points, some are used as interlocks, and some are used for controlling purposes. Considering the importance of the functionality of the signals, some of the variables could be eliminated.

Table 1. Difference circumstances affect process conditions.



		HA SOC	HA EOC	HS SOC	HS EOC
Ethylene conversion	%	12.8	13.6	11.9	13.0
Oxygen conversion	%	37.5	46.3	28.9	40.5
Selectivity	%	81.6	76.7	88.0	80.4
Oxygen inlet	% mol	8.20	7.96	8.23	7.94
Flammable limit inlet	% mol	9.00	8.76	9.03	8.74
Margin at inlet	% mol	0.80	0.80	0.80	0.80
Oxygen outlet	% mol	5.19	4.33	5.93	4.78
Flammable limit outlet	% mol	7.43	6.48	6.95	5.83
Margin at outlet	% mol	2.24	2.15	1.02	1.05
GHSV	$Nm^3/m^3/hr$	3500	3500	3500	3500
Work rate	$kg/m^3/hr$	180	180	180	180
Pressure inlet	bara	17.5	17.5	17.5	17.5
Temperature outlet	°C	222.0	250.0	242.0	275.0

Moreover, it should be noted that in the path of the process loop, some instruments are placed repeatedly. For example, after feed/product exchanger, and before reactor in Fig. 2, two temperature instruments are installed, and hence one of them can be eliminated.

An additional method to eliminate some instruments from the scope of work is to consider the destination of the signals generated by the instruments. For example, some signals are pointed to DCS[2] and some are referring to ESD[3] subsystems. Normally, next to each ESD instrument, there is a DCS instrument measuring the same process variable. The main purpose of the ESD instrument is for shut down, and it is used to show that the alarm stage is passed and the fault has already happened. Moreover, collecting data for a long time from ESD is difficult, and hence it is not suitable to extract the statistical information from. In this way, it is possible to eliminate the instruments, which are pointed to ESD system.

Another way to clear unnecessary variables is to compare different variables with the same origin. For example, here the statistical behavior of flow and pressure instruments, placed in the same process are examined. Figure 3 top shows the flow data recorded from the mixing nozzle's input, and the resulting pressure at the output is plotted at the bottom. Careful examination of Fig. 3 shows that, the behavior of these two variables is very similar and one of them can be eliminated.

Following the clearance procedure explained just now, the final variable list of the RYG process loop can be reduced to nine variables. All variables are listed in the Table 2. Data from these variables are gathered from January 1st, 2016 to October 3rd, 2017 from the MORVARID MEG plant. This equals to 632 days and one hour, or 15169 hours of unit operation in total. During this period, there are times that the unit was shut down, and therefore the corresponding data should be eliminated. In total and without considering the data cleaned during the unit shutdown times, 3200 samples of nine plant variables under different operating conditions are recorded. The number of recorded samples after cleaning procedure is reduced to 1500.

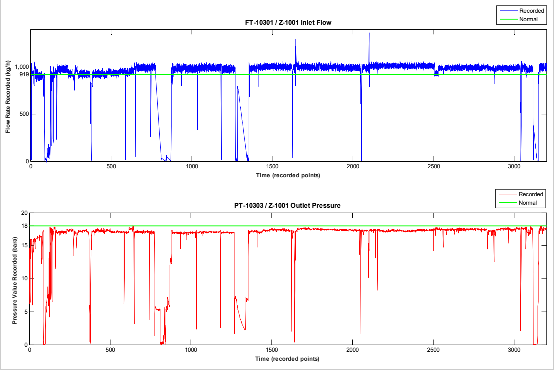


Fig. 3. Input flow and output pressure of mixing nozzle.

Table 2. The final list of variables after clearing.

Input flow to mixing nozzle
Input flow to feed / product exchanger
Pressure input to reactors
Temperature input to reactors
Temperature input to primary product cooler
Temperature output from reactors
Pressure output from reactors
Pressure output from compressor
Temperature output from compressor

[2] Distributed Control System
[3] Emergency Shut Down

4.2 Statistical Monitoring of the Process

In this section, the analysis results on the recorded experimental data is presented. Since the data are collected from different process variables at different scales, the first step towards analyzing the experimental data and converting them to the process information is using the standardization and normalization technique. The collected data for nine process variables listed in Table 2 is plotted in Fig. 4 after normalization. The PCA algorithm is applied on the standardized data to find its principal components. The total variance of the model after adding each principal component is plotted in a bar chart in Fig. 5.

The percentages on the *y*-axis at the right hand side of Fig. 5, is a measure showing how many principal components of the data are able to represent the full data collected from the process. As can be seen from Fig. 5, 85% of the full data can be reconstructed with two variables while this will rise to around 95% for three variables. It is also possible to infer from Fig. 5 that with only five variables rather than nine variables, it is possible to get very close to 100% representation of the collected data. The blue curve in Fig. 5 shows that by increasing the number of principal components it is possible to reach rapidly to 100% reconstruction of the complete data. The variables of the system using the first two and three principal components of the data are plotted in Figs. 6 and 7, respectively. The red points in these two figures are distributed statistical data and the green ones are new test data used for evaluation of the proposed technique.

To estimate the probability density function of the statistical data, a simple kernel algorithm is used with the first two principal components calculated using the PCA method. The estimated probability density function is shown in Fig. 8. In order to evaluate the accuracy and smoothness of the estimated PDF, different confidence intervals, i.e. 90%, 93%, 95% and 97% are compared. It should be noted that the intervals must be continuous without any insulated area.

To find the best confidence interval, the performance of the statistical model against new test data is evaluated for different confidence intervals. By considering the faulty or near faulty data amongst the new test data, the success rate of the statistical model for different confidence intervals are calculated and the best model is selected. The results show that the confidence interval 93% with two principal components extracted from the PCA algorithm and the probability density function estimated using the simple kernel method shows the best result.

5. CONCLUSIONS

This article is proposing a statistical monitoring technique by analyzing the relevant data and variables collected from the process under the study. An important aspect of statistical process monitoring is to respect the conditions of the process, such as multiple operation modes, and choose the right subsystem when the whole unit is too big for estimation of the statistical model. Having this priori knowledge the next step is clearing the system variables and their data, especially for a system with huge number of variables. To reduce the dimensions of the system, PCA technique is used. The results show that two and three principal components are able to give a good approximation of the whole system. The probability density function of data is estimated using a kernel based technique. The process variables and associated data are cleared by looking at the function of measurement instruments and destination of the signals. To address the multiple operation mode condition of the plant, a moving window approach is applied. To evaluate the effectiveness of the estimated model the new test data are imported and the results are compared for different confidence intervals. As the next step, it is aimed to develop a user-friendly software working based on the method proposed in this study.

6. ACKNOWLEDGMENT

The authors would like to acknowledge Engineering and Physical Sciences Research Council (EPSRC), grant number EP/R02572X/1, and National Centre for Nuclear Robotics.

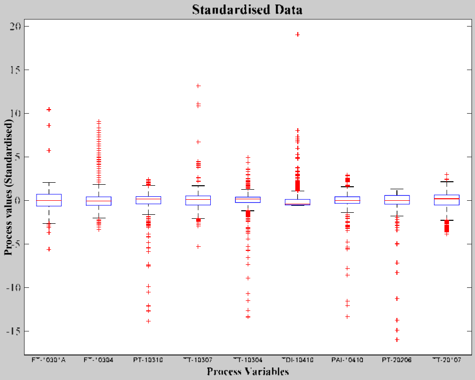


Fig. 4. Normalized data.

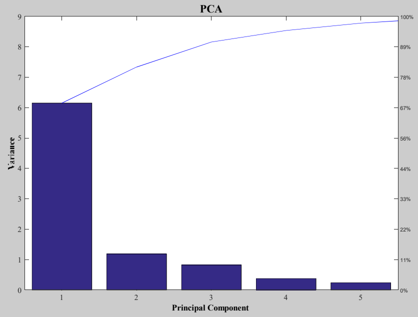


Fig. 5. Cumulative chart of main components variance.

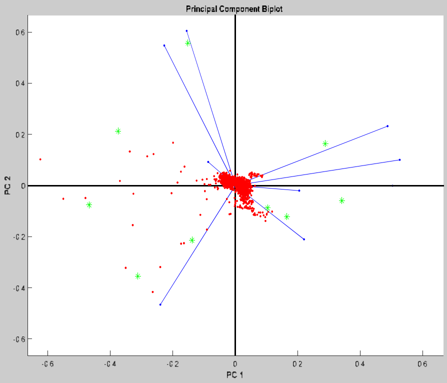


Fig. 6. Estimated model with the first two principal components.

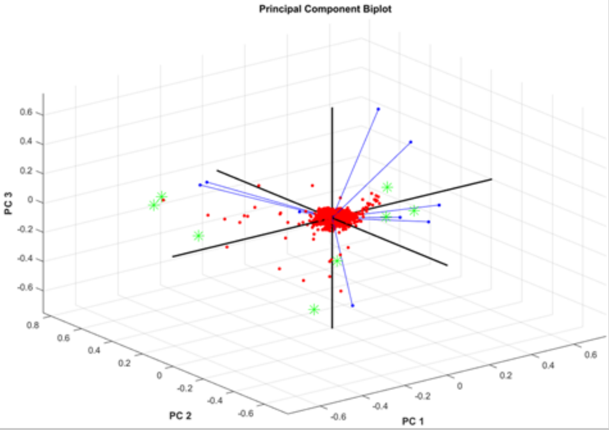


Fig. 7. Estimated model with the first three principal components.

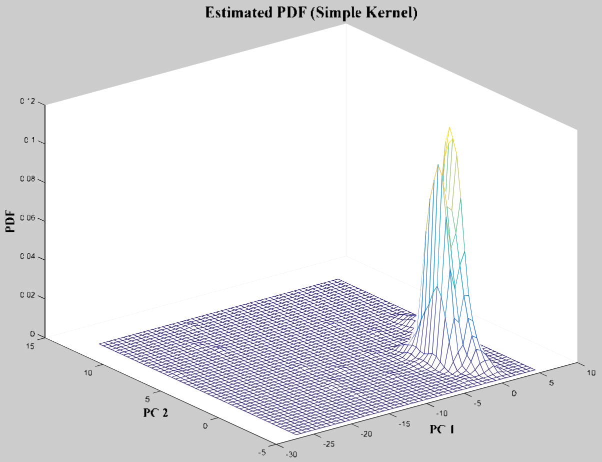


Fig. 8. Probability density function - Simple kernel.

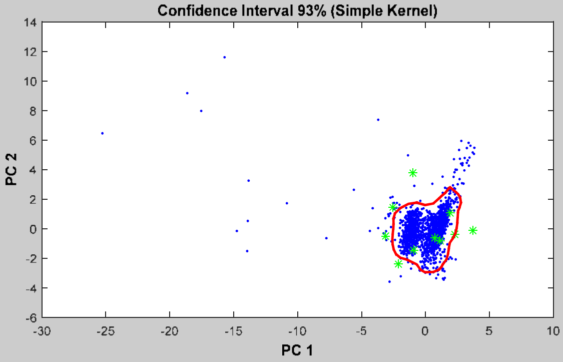


Fig. 9. The estimated PDF for 93% of confidence interval.

7. REFERENCES

Jeng, J.-C., 2010. Adaptive Process Monitoring Using Efficient Recursive PCA And Moving Window PCA Algorithms. *Journal of the Taiwan Institute of Chemical Engineers, 41,* p. 475–481.

Lu, C.-J., Lee, T.-S. & Chiu, a. C.-C., 2008. Statistical Process Monitoring Using Independent Component Analysis Based Disturbance Separation Scheme. *IEEE, IJCNN2008,* pp. 232-237.

MacGregor, J. & Cinar, A., 2012. Monitoring, Fault Diagnosis, Fault-Tolerant Control and Optimization: Data Driven Methods. *Computers and Chemical Engineering 47,* pp. 111-120.

Montazeri, A. & Ekotuyo, J., July 2016. *Development of dynamic model of a 7DOF hydraulically actuated tele-operated robot for decommissioning applications.* Boston, s.n., pp. 1209-1214.

Montazeri, A., West, C., Monk, S. & Taylor, C., 2017. Dynamic modelling and parameter estimation of a hydraulic robot manipulator using a multi-objective genetic algorithm. *International Journal of Control, 90 (4), 661-683.*

Montazeri, A. & Poshtan, J., 2009. GA-based optimization of a MIMO ANC system considering coupling of secondary sources in a telephone kiosk. *Applied Acoustics, 70(7), 945-953.*

Montazeri, A. & Poshtan, J., 2009. Optimizing a multi-channel ANC system for broadband noise cancellation in a telephone kiosk using genetic algorithms. *Shock and Vibration, 16(3), 241-260.*

Ni Zhang, X. T., Cai, L. & Deng, X., 2014. Process Fault Detection Based On Dynamic Kernel Slow Feature Analysis. *Computers and Electrical Engineering, 41,* pp. 9-17.

Pokkunuri, B., 1994. Knowledge Based Simulation for Process Monitoring And Regulatory Control. *Intelligent Systems Engineering.*

Shell Global Solutions, 2016. Enhancements in Ethylene Oxide/Ethylene Glycol Manufacturing Technology. *Shell Global Solutions International BV (WHITE PAPER).*

Shlens, J., April 2009. A Tutorial On Principal Component Analysis. *Center for Neural Science, New York University, Version 3.01,* p. 12.

Venkatasubramanian, V., Rengaswamy, R., Yin, K. & Kavuri, a. S. N., 2003. A Review of Process Fault Detection and Diagnosis Part I: Quantitative Model based Methods. *Computers and Chem. Eng. 27, 293—311.*

Wang, H., Song, Z. & Wang, H., 2002. Statistical Process Monitoring Using Improved PCA With Optimized Sensor Locations. *Journal of Process Control, 12,* p. 735–744.

Yin, S., Ding, S. X., Xie, X. & Luo, a. H., November 2014. A Review on Basic Data-Driven Approaches for Industrial Process Monitoring. *IEEE Transactions On Industrial Electronics, VOL. 61, NO. 11,* pp. 6418-6428.

Zamprogna, E., 2003. Process Monitoring and Control Using Artificial Neural Networks and Other Advanced Techniques.

Zhou, B., HaoYe, HaifengZhang & MingliangLi, 2016. Process Monitoring of Iron-Making Process In A Blast Furnace With PCA-Based Methods. *Control EngineeringPractice47.*