Journal of Process Control xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Process Control

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

Comparative study on monitoring schemes for non-Gaussian distributed processes

Gang Li [a,*], S. Joe Qin [b,1]

[a] Department of Inertia Technology & Navigation Guidance Instruments, Beihang University, Beijing 100191, China
[b] Chinese University of Hong Kong, Shenzhen 518172, China

ARTICLE INFO

Article history:
Received 23 February 2016
Received in revised form 30 June 2016
Accepted 25 August 2016
Available online xxx

Keywords:
Non-Gaussian distribution
Independent component analysis
Kernel density estimation
Gaussian mixture model
Support vector data description
Statistical pattern analysis
Neyman Pearson lemma

ABSTRACT

Traditional multivariate statistical process monitoring techniques usually assume measurements follow a multivariate Gaussian distribution so that T^2 can be used for monitoring. The assumption usually does not hold in practice. Many efforts have been spent on redefining a proper boundary of control region for non-Gaussian distributed processes. These efforts lead to new models such as independent component analysis (ICA), statistical pattern analysis (SPA), and new techniques such as kernel density estimation (KDE), support vector data description (SVDD). However, it has not been stated clearly how a latent structure will affect monitoring performance. In this paper, most of main stream methods for non-Gaussian process monitoring are recalled and categorized. The essential problem formulation of process monitoring is summarized from a general case and then explained in both Gaussian and non-Gaussian distribution, respectively. According to this formulation, KDE and SVDD methods are effective but time-consuming to extract proper control region of non-Gaussian distributed processes. Dimension reduction models are more beneficial to overcome the curse of dimensionality, rather than extracting non-Gaussian data structure. Besides, the monitoring of non-Gaussian processes can be converted into the monitoring of Gaussian processes according to central limitation theorem.

1. Introduction

Multivariate statistical process monitoring (MSPM) technology has received a great success in fault detection and diagnosis for many manufacturing processes, which relies on latent structure models such as principal component analysis (PCA) and partial least squares (PLS) [1–7]. MSPM can be viewed as the application of data science to manufactural industries, which digs valuable information from huge routine process data and improves the reliability and efficiency greatly [8].

A typical MSPM data model usually starts with a dimension reduction procedure, which provides a latent structure of data. Taking PCA model as an example, principal subspace and residual subspace are generated after a PCA model is built. Then, corresponding statistics are available for process monitoring in each subspace. If measured variables follow a multivariate normal distribution, the optimal control region in principal subspace is a hyper ellipsoid, while the control region in residual space is usually defined as a hyper sphere. However, it is very common that some measured variables do not follow a Gaussian distribution. As a result, PCA and PLS models do not function effectively and lead to inaccurate detection result.

In order to deal with such challenges, many methods have been well developed, which can be roughly divided into three types. The first type is to modify conventional latent structure of multivariate data. Independent component analysis (ICA) is typical modification of PCA regarding non-Gaussian distribution, which searches a linear combination of variables with the highest non-Gaussianity [9]. While principal components in PCA are independent under multivariate Gaussian distribution, components in ICA are claimed to be independent under non-Gaussian distribution. Generally, ICA model recovers essential signals efficiently for non-Gaussian distributed data. After ICA was firstly introduced to process monitoring by Kano et al. [9], related research work has been reported in many occasions [10,11].

Although independent components in ICA are more informative, they usually do not follow a multivariate normal distribution. Consequently, kernel density estimation (KDE) is adopted to determine the boundary of few retained independent components. Kernel density estimation is widely used to estimate the control region of

* Corresponding author.
E-mail address: gangli@buaa.edu.cn (G. Li).
[1] On leave from the Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA.

http://dx.doi.org/10.1016/j.jprocont.2016.08.007
0959-1524/

latent variables in many models [12,13]. However, it is not the only way. For the mixture of multiple Gaussian distributions, Gaussian mixture model (GMM) was proposed to describe the whole distribution with fewer Gaussian distributions based on Bayesian criterion [14]. As GMM does not reduce the dimensionality of measurements, a mixture of factor analysis model was proposed for process monitoring inspired by GMM [15]. Recently, another technique called support vector data description (SVDD) has been studied in process monitoring [16–19]. SVDD attributes the boundary of normal data to a hyper sphere and learns sphere radius directly from data in a similar way to support vector machine. With the help of kernel function, the original data which may not be distributed as a sphere can be mapped into a high-dimensional feature space. The advantage of SVDD over KDE is high computational efficiency. Similar to SVDD, Ge and song employed a one-class support vector machine to determine control region with a hyper plane instead of hyper sphere which is also learned directly with unsupervised data [20].

Instead of searching the control region, there is another type of solution, which generates features that approximately follow Gaussian distribution from original measurements. The local approach, which constructs statistics based on parameters of latent structure, has been proposed to perform the monitoring of industrial processes [21,22]. The parameters in their models are eigenvalues of covariance matrix. Although original data may not follow Gaussian distribution, the featured statistics from a window of samples may approximately follow a Gaussian distribution according to central limit theorem (CLT). Similarly, a new framework called statistic pattern analysis is proposed to perform fault detection of both continuous and batch processes [23,24]. SPA framework considers not only mean and covariance of samples, but also high order statistics like skewness and kurtosis. Similar to local approach, the feature vector of SPA approximately follow a multivariate Gaussian distribution. This approach works well regardless of original distribution and is more sensitive to tiny faults because noisy level falls down when statistics are used for monitoring directly. However, it usually causes non-ignorable delay to fault detection depending on window size.

Although there are so many models and methods to deal with non-Gaussian data, there is a lack of summary that points out the essential formulation in non-Gaussian process monitoring. This paper firstly recalls main techniques used for the monitoring of non-Gaussian distributed processes and then tries to understand basic problems in process monitoring. The function of latent structure in process monitoring will be discussed. Different techniques are compared through theoretical analysis and several case studies. Conclusions are given in the last section.

2. Control region for multivariate statistical process monitoring

2.1. Hypothesis test based multivariate statistics

Suppose $\mathbf{x} \in \mathbf{R}^m$ represents a measurement vector for process monitoring which consists of m variables. When a process operates under a steady condition, $\mathbf{x}$ can be viewed as a random vector with probability density function (PDF) $p(\mathbf{x})$. The problem of multivariate statistical process monitoring can be stated as a hypothesis test problem with null hypothesis $H_0 : \mathbf{x} \sim p_N(\mathbf{x})$ and alternative hypothesis $H_1 : \mathbf{x} \sim p_F(\mathbf{x})$, where $p_N(\mathbf{x})$, $p_F(\mathbf{x})$ represent PDF for normal data and faulty data, respectively. Generally speaking, $p(\mathbf{x})$ is unknown for both normal and faulty situations, however there are usually abundant normal data and few faulty data. For the case that normal and faulty data both follow a relatively stable pattern, which means p_N and p_f are both approximatively fixed, the problem become easier and reduces to a simple test. Furthermore, if both normal and faulty data are sufficient, the problem is simplified to a binomial classification problem for single fault detection or a multinomial classification problem for multiple faults detection. This is a typical supervised learning problem in the area of machine learning. There are many efficient discriminators available such as logistic regression, support vector machine, decision tree and so on.

Table 1
Confusion matrix.

		Decision (detection)	
		Accept H_0 (not detected)	Reject H_0 (detected)
Truth	H_0 (Normal)	Correct decision	Type I error (false alarm)
	H_1 (Fault)	Type II error (missing alarm)	Correct decision

For the case that faulty pattern are arbitrary and very few faulty data is available, the hypothesis test are composite test, which means there is not a single pdf for faulty data. In such a case, a closed boundary indicating normal region is preferred other than a separating boundary between normal and faulty data.

Consider the most simplest scenario that both p_N and p_F are known and there is a sample to be classified. The above problem of hypothesis test can be solved by partitioning the whole space into two regions, namely control region S_N and rejection region S_F, where $S_N \bigcap S_F = \phi$, $S_N \bigcup S_F = \mathbf{R}^m$. The sample is decided as normal if and only if $\mathbf{x} \in S_N$, otherwise a fault is detected. There are two types of error, which are indicated in confusion matrix (Table 1). Denote $P_N(S_F)$ as the probability of type I error (i.e. probability of false alarm), and $P_F(S_F)$ as 1- probability of type II error (i.e. probability of fault detection), then a model with low $P_N(S_F)$ and high $P_F(S_F)$ is preferred. Given a detection logic with a certain threshold, it is always possible to adjust $P_N(S_F)$ and $P_F(S_F)$ by tuning the threshold. In general, when $P_N(S_F)$ goes down, $P_F(S_F)$ will go down too. It is hence reasonable to search a hypothesis test, represented by a partition S_N and S_F that maximizes the $P_F(S_F)$ with a given upper limit of $P_N(S_F)$. The problem can be formulated as the following optimization over the set S_F:

$$\begin{aligned} \max_{S_F} \quad & P_F(S_F) = \int_{\mathbf{x} \in S_F} p_F(\mathbf{x}) d\mathbf{x} \\ s.t. \quad & P_N(S_F) = \int_{\mathbf{x} \in S_F} p_N(\mathbf{x}) d\mathbf{x} \le \alpha \end{aligned} \tag{1}$$

In general, the solution of the above optimization is difficult to figure out. Fortunately, with the development of statistics, Neyman and Pearson had solved this problem perfectly.

Lemma 1 (Neyman-Pearson Lemma [25]). *In all the test function ϕ of simple hypothesis test $H_0 : \mathbf{x} \sim p_N(\mathbf{x})$ and $H_1 : \mathbf{x} \sim p_F(\mathbf{x})$, the most powerful test ϕ^* with level α is to accept H_0 when $\mathbf{x} \in S_N^* = \{x : p_F(\mathbf{x}) \le k p_N(\mathbf{x})\}$ and reject H_0 when $x \in S_F^* = \{x : p_F(x) > k p_N(x)\}$, where k is selected so that $P_N(S_F^*) = \alpha$.*

The most powerful test with level α means $P_F(S_F^*) \ge P_F(S_F)$ for arbitrary test ϕ with $P_N(S_F) \le \alpha$. This lemma provides a guidance on how to construct the most efficient detection index and how to determine the control limit. However, in most cases of process monitoring, only normal data is available for estimating the distribution of p_N and there is no faulty data for p_F. According to principle of maximum entropy, it is reasonable to assume p_F as a uniform distribution on a fixed range where measured data can reach, which indicates

$$S_N^* = \{x : p_N(x) \ge p_\alpha\} \tag{2}$$

with p_α satisfying $P_N(S_F^*) = \alpha$. Therefore, either given the family of PDF with unknown parameters or only data, once the parameters

Please cite this article in press as: G. Li, S.J. Qin, Comparative study on monitoring schemes for non-Gaussian distributed processes, J. Process Control (2016), http://dx.doi.org/10.1016/j.jprocont.2016.08.007