Chemometrics and Intelligent Laboratory Systems 132 (2014) 52–62

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

# Reference set selection with generalized orthogonal Procrustes analysis for multivariate statistical process monitoring of multiple production processes

R.L.J. Coetzer [a], R.F. Rossouw [a,b,*], N.J. Le Roux [b]

[a] Sasol Technology Research and Development, Sasol, Private Bag 1, Sasolburg 1947, South Africa
[b] Department of Statistics and Actuarial Science, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

ARTICLE INFO

*Article history:*
Received 28 June 2013
Received in revised form 9 December 2013
Accepted 9 January 2014
Available online 17 January 2014

*Keywords:*
Biplots
Multiple processes
Process monitoring
Procrustes analysis

ABSTRACT

Multivariate process monitoring is important in industry to ensure that production processes perform as close as possible to optimal operation. However, the selection of a reference set of optimal or expected performance is required for efficient process monitoring in real time. In this paper we present the method of generalized orthogonal Procrustes analysis to select a reference set for the multivariate monitoring of multiple production processes simultaneously. We combine generalized orthogonal Procrustes analysis with principal component analysis (PCA) and biplots to illustrate the implementation of the method and the interpretation of the results which provide important information on the relationships between many process variables and differences between the production processes. The work is motivated by an industrial problem involving the multivariate monitoring of a coal gasification production facility considering many process variables monitored across multiple reactors.


## 1. Introduction

Process monitoring and more specifically multivariate statistical process control have received much attention in the statistical and engineering literature [1,7]. The fundamental approach of the majority of the multivariate process monitoring procedures is to first specify an historical reference set that is within statistical control. Multivariate analytical techniques such as principal component analysis (PCA) are then employed to project process variables onto a lower dimensional space where they can be jointly monitored given the in-control reference set. Throughout this paper we assume that a reference set is required for monitoring whether the process is within statistical control and to detect out-of-control or deviations from expected performance i.e., the reference set is considered as the target for the process. The use of the PCA biplot as a monitoring graph has been discussed at length by [1], and more specifically the use of the biplot as a dynamic tool, which is updated in real time, has been discussed by [7].

Current literature is mostly focused on multivariate statistical monitoring of many process variables simultaneously for a single production process. The selection of a single reference set that is within statistical control or conforms to some specified accepted performance measure(s) for multiple identical production processes simultaneously has (to our knowledge) not been discussed previously. In this paper we present a methodology for selecting a single reference set for multivariate process monitoring across multiple identical production processes. The work is motivated by an industrial problem involving the multivariate monitoring of a coal gasification production facility considering many process variables monitored across multiple mechanically identical reactors. Therefore this paper deals with the situation where classical multivariate process monitoring is extended to the multivariate monitoring of two or more identical processes across several process variables simultaneously.

Sasol, South Africa, gasifies bituminous coal to synthesis gas, which is converted to fuels and chemicals via Sasol's suite of hydrocarbon processes. Since Sasol's coal-to-liquids facility delivers nearly 29% of the fuel requirements in South Africa, the continuous improvement of the coal gasification plant is of critical importance to the company to ensure a stable supply of high quality synthesis gas to the downstream units. Therefore, in order to maintain optimum product yields and to sustain throughput, an efficient multivariate process monitoring methodology is required for those process variables that govern gasifier performance. The coal gasification plant utilizes 84 Sasol® FBDB. The residual sum of squares is given by gasifiers. The Sasol facility in Secunda, South Africa, is the largest coal to syngas production facility of its kind in the world.

Real time data are captured on more than ten process variables for monitoring the performance of each gasifier, with the main output from the individual gasifiers the amount of raw gas ($km^3$ n/h) produced. The gasifiers are referred to as the production processes. In this paper the production processes are grouped into a number of production trains e.g., ten production processes per train. All the trains receive the

* Corresponding author at: Sasol Technology Research and Development, Sasol, Private Bag 1, Sasolburg 1947, South Africa.
*E-mail addresses:* roelof.coetzer@sasol.com (R.L.J. Coetzer), ruan.rossouw@sasol.com (R.F. Rossouw), njlr@sun.ac.za (N.J. Le Roux).

0169-7439/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2014.01.006

same feedstock, but each train has a different operator. Therefore, differences between trains and production processes may be due to operator, mechanical degradation or other process variable deviations.

An important consideration in this specific production facility is stable operation. The raw gas produced by the gasifiers is the feedstock to the remainder of the factory. This coal-to-liquids petrochemical factory is an extremely complex and integrated system. It is therefore of the utmost importance that the raw gas production is as stable as possible, as any upset will result in a disturbance of the stability in the factory as a whole.

In this paper we consider the problem of selecting one production train as a reference set that allows for efficient multivariate process monitoring of all the production processes. The problem is downscaled however to only considering two production processes per train for five production trains. Considering only two production processes per train contains all the complexities of reference set selection when faced with the multivariate monitoring of multiple processes, so that the methodology developed can be applied to any number of production processes and trains within the actual production facility. The motivation behind the selection of one train is that the whole production facility should ideally be operated similarly, and deviations in performance of the production processes are evaluated relative to the same reference of expected or optimal performance. As the production trains (and therefore the production processes on the trains) receive similar feedstock together with mechanically identical production processes, the major cause of differences in performance over trains must be related to operating protocol. Selecting the optimal train as reference set would therefore ensure that the operators on all the trains target the same optimal operating protocol.

The coal gasification process is a continuous process and the performance of the reactors is monitored in real time. However, the performance of the production processes may change from week to week due to planned or unexpected changes in the feed. In particular, the feedstock blending schedule is updated weekly according to feed availability, and other drivers. The feedstock could therefore potentially change on a weekly basis. The implication is that a base time unit for reference set selection should be a week of data. Therefore, we are also interested in the optimal combination of a specified number of weeks (with a minimum of two weeks) to employ for the selection of the optimal train as the reference set. The motivation behind the selection of the optimal number of weeks is that deviations in performance of the production processes can be evaluated relative to the same reference of expected or optimal feed. A minimum of two weeks of data is required for obtaining the most accurate results from the multivariate PCA and biplot analysis.

We present a methodology using generalized orthogonal Procrustes analysis (GOPA) [6] for selecting the optimal reference set for the multivariate monitoring of the multiple identical production processes. More specifically, GOPA is applied to select the optimal combination of a specified number of weeks as well as the optimal production train. We illustrate the use of the optimal reference set on a PCA biplot for multivariate process monitoring. Procrustes analysis and the more generalized GOPA have not been applied previously for reference set selection within the context of multivariate process monitoring. Although it is not the case here, we note that this procedure allows for different variables to be measured at the various production processes.

The paper is outlined as follows. In Section 2 we discuss the industrial problem under investigation and the difficulties with selecting a reference set for multiple production processes. In Section 3 we provide a brief introduction to the mathematics underlying a Procrustes analysis. In Sections 4.1 to 4.3 we present the results and discuss the implementation of the proposed method for the multivariate monitoring of multiple processes in a production facility. We conclude in Section 5 by proposing a procedure for reference set selection and its implementation on a commercial production facility.

## 2. Problem setting

Consider Fig. 1, which represents a flow sheet of the production facility under study. Specifically, the facility under investigation consists of ten identical production processes (reactors) grouped into five production trains of two processes each. In the current study, time weighted hourly average data for eleven process variables were captured over a consecutive ten week period for all ten processes. The process variables include measurements on production volume, utility consumption such as oxygen and steam consumption and other stability measures on the reactors. Table 1 provides a description of the types of variables used in the study. Note that the variables are indicated by neutral labels due to confidentiality restrictions imposed by the company under consideration.

Canonical variate analysis (CVA) biplots are employed to groups of data to maximize the between group relative to within group variance [5]. Therefore, to illustrate the complexity of the industrial problem, a CVA biplot was constructed to visualize the differences between the different production processes for some selected weeks. Fig. 2 depicts the CVA biplot for weeks 1, 3, 6 and 10. Ninety percent (90%) of alpha bags are added to the display to quantify the within group variability for each group [4]. From Fig. 2(a), it is observed that the performance of reactors five (RX5) and ten (RX10) are different compared to the other reactors across the trains for week 1. Specifically, readings of variables S2, S3 and S5 are much higher for reactor ten compared to the other reactors. Since these variables all indicate stability it can be concluded that reactor ten was more unstable than the other reactors during week one. However, reactor ten and reactor nine (RX9) are on train five, but they have a very different performance in terms of stability. Reactor five and reactor six (RX6) are on train three, also with a very different performance in stability, and different from the reactors on train five. The differences between the trains may be due to different operators, but the differences between the reactors on the same train may be due to the specific process variable deviations or mechanical wear and tear.

From Fig. 2(b), huge variability is observed in the performance of reactor nine (RX9) compared to the other reactors across the trains for week three. The variability in performance of reactor nine is observed for the production, utility and stability variables. Compared to Fig. 2(a) for week one, it highlights the problem of changes in performance from week to week for one or more of the production processes across the different trains. Fig. 2(c) and (d) illustrates the performance of the reactors for weeks six and ten respectively. Specifically, more stable production periods are observed across the production processes (reactors).

The CVA biplots in Fig. 2 illustrate the differences in performance between the production processes across different weeks, and although the feed is not expected to change much from week to week it illustrates the differences that may occur due to the blending, planning and scheduling of the feedstock. Therefore, the selection of the optimal combination of weeks will provide the ability to monitor deviations in performance of the production processes relative to the same reference of expected or optimal feed.

Furthermore, the CVA biplots in Fig. 2 display the differences in performance between the production processes across the trains, and the selection of the one production train as a reference set will provide the ability to monitor deviations in performance of the production processes relative to the same reference of expected or optimal performance. The problem reflects the difficulties with the implementation of a monitoring methodology on the actual production facility, where a reference set needs to be selected for the simultaneous monitoring of all the trains or processes which may consist of up to ten production processes per train. We apply generalized orthogonal Procrustes analysis (GOPA) [6] for selecting the reference set of the optimal combination of weeks and the production train.

In the current problem where all the production processes have the same process variables, it might be argued that it is more appropriate to