Journal of Process Control 35 (2015) 30–40

Contents lists available at ScienceDirect

Journal of Process Control

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

Multivariate fault isolation via variable selection in discriminant analysis

Te-Hui Kuang[a], Zhengbing Yan[b], Yuan Yao[a,*]

[a] Department of Chemical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan
[b] College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou 325035, China

ARTICLE INFO

Article history:
Received 8 February 2015
Received in revised form 14 July 2015
Accepted 21 August 2015
Available online 14 September 2015

Keywords:
Fault isolation
Multivariate statistical process monitoring
Variable selection
LASSO
Elastic net

ABSTRACT

In multivariate statistical process monitoring (MSPM), isolation of faulty variables is a critical step that provides information for analyzing causes of process abnormalities. Although statistical fault detection has received considerable attention in academic research, studies on multivariate fault isolation are relatively fewer, because of the difficulty in analyzing the influences of multiple variables on monitoring indices. The commonly used tools for fault isolation, such as contribution plots, reconstruction-based methods, etc., have several shortcomings limiting their implementation. To solve the problems of the existing methods, this paper reveals the relationship between the problems of multivariate fault isolation and variable selection for discriminant analysis. Furthermore, by revealing the equivalence between discriminant analysis and regression analysis, the problem of multivariate fault isolation is further formulated in a form of penalized regression which can be solved efficiently using state-of-the-art algorithms. Instead of offering a single suggested set of faulty variables, the proposed method provides more information on the relevance of process variables to the detected fault, facilitating the subsequent root-cause diagnosis step after isolation. The benchmark Tennessee Eastman (TE) process is used as a case study to illustrate the effectiveness of the proposed method.

1. Introduction

For ensuring safe and efficient operation of manufacturing and chemical processes, multivariate statistical process monitoring (MSPM), which extracts process information from historical operating data, plays an important role [1]. As discussed in [2], a complete MSPM procedure can be divided into four steps, including fault detection, fault isolation, fault diagnosis, and process recovery. The target of fault detection is to identify process abnormality as soon as possible after its occurrence, while the purpose of fault isolation is to recognize the process variables contributing most to the detected fault. In the research field of MSPM, different types of fault detection methods have been intensively investigated. In comparison, studies on fault isolation are relatively fewer, due to the difficulty in analyzing the influences of multiple variables on monitoring indices.

For isolating faulty variables, contribution plots [3] have been widely used as standard tools in MSPM, which judge the status of each process variable by comparing its contribution to the monitoring index and a predetermined control limit. Although easy to use, contribution plots have been criticized by many researchers (e.g. [4]) for suffering from the "smearing" effect, i.e. the influence of faulty variables on the contributions of non-faulty variables. Due to such effect, the contributions of non-faulty variables may fall outside the corresponding control limits as well, resulting in misleading isolation results. As discussed in [5], although the contributions of process variables follow a certain distribution in the normal operating condition, the contributions of non-faulty variables usually do not follow the same distribution when a fault occurs. Therefore, it is improper to offer a suggested set of faulty variables based on the control limits derived from normal operating data.

Besides contribution plots, reconstruction-based methods are also widely used, which identify faulty variables by minimizing the reconstructed monitoring index along certain "fault directions". In the early version of this type of methods, all candidate fault directions are assumed to be known [6]. Such assumption is usually too strong in practice. An approach was developed in [7] for extracting the fault directions from historical fault data, which relaxes the requirement for reconstruction. However, sufficient historical fault data are rarely available in real industry. To solve these problems, reconstruction-based contribution (RBC) [8] was proposed, where the amount of reconstruction of the monitoring

* Corresponding author. Tel.: +886 3 5713690; fax: +886 3 5715408.
E-mail address: yyao@mx.nthu.edu.tw (Y. Yao).

http://dx.doi.org/10.1016/j.jprocont.2015.08.011
0959-1524/

index along a variable direction is regarded as that variable's contribution. Although RBC is capable of dealing with both known and unknown faults, the smearing effect can still be observed in the RBC contribution plot, as noted by the authors themselves.

To overcome the drawbacks of the conventional reconstruction-based methods, the branch and bound (BAB) technique was combined with missing data analysis, which addresses the fault isolation problem by solving a combinatorial optimization problem [9,10]. Although BAB is more efficient than exhaustive search, the computational burden of this approach is still heavy for a process with a massive number of variables. Therefore, such method is not suitable for online implementation.

In addition to the above-mentioned methods, pattern classification techniques, such as Fisher's discriminant analysis (FDA), support vector machine (SVM), etc., have also been utilized to classify the detected faults [2,11–13]. Instead of isolating faulty variables, fault classification determines fault types directly by calculating the similarities between the fault data and historical process data corresponding to different types of known events. To achieve accurate classification, sufficient historical fault data are required for model training. However, in practice it is unusually to have a complete dataset containing all types of faults. Therefore, unknown faults may not be well treated. Recently, fault classification was combined with BAB to deal with both known and unknown faults [14,15]. Of course, the shortcoming of BAB is inherited by such methods.

In this paper, the multivariate fault isolation problem is re-investigated from another point of view and transformed into a variable selection problem in discriminant analysis. By revealing the equivalence between discriminant analysis and regression analysis, the problem can be further formulated in a form of penalized regression, where L_1 regularization is introduced into a standard multiple regression model to achieve variable selection. In the field of statistics, such a model is called the least absolute shrinkage and selection operators (LASSO) [16], which can be solved efficiently using state-of-the-art algorithms. Therefore, the computational burden of fault isolation is significantly reduced, especially when the number of process variables is large. In addition, instead of giving a single suggested set of faulty variables based on improper control limits as conventional contribution plots do, the proposed method provides more information about the relative importance of each process variable on the detected fault. Such information facilitates the subsequent root-cause diagnosis step after isolation, while the smearing effect is avoided. For better dealing with highly correlated faulty variables, the LASSO-based isolation method can be further revised by adding an L_2 penalty term into the objective function of least squares regression. By doing so, the multivariate fault isolation problem is described by an elastic net (EN) [17] model.

The rest of this paper is organized as follows. In Section 2, the motivations of this paper are briefly introduced. Then, the LASSO- and EN-based fault isolation methods are proposed in Section 3. In Section 4, the capability of the proposed methods is illustrated using the case studies on the benchmark Tennessee Eastman (TE) process [18]. Finally, conclusions are drawn in Section 5 to summarize the paper.

2. Motivations

The basic idea of this paper is as follows. As mentioned in previous, the goal of fault isolation is to identify variables responsible for the detected process abnormality. In other words, the faulty variables to be isolated are those discriminating normal process measurements and fault samples. Therefore, in a sense, the task of fault isolation is identical to identifying the discriminating variables in a two-class problem, where the historical normal operating data are regarded as belonging to one class and the data associated with the detected fault are assigned to the other class. Consequently, the multivariate fault isolation problem is equivalent to a variable selection problem in discriminant analysis.

FDA, named for its inventor Fisher [19], is one of the most famous techniques in discriminant analysis, which is used for fault isolation in the following of this paper. In previous research, it has been well established that an FDA model is identical to a least squares regression model with predictors and response variable chosen properly [20]. Suppose that a set of d-dimensional samples $\Xi = \Xi_1 \cup \Xi_2 = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ contains samples from two different classes $\Xi_1 = \{\mathbf{x}_1^1, \ldots, \mathbf{x}_{n_1}^1\}$ and $\Xi_2 = \{\mathbf{x}_1^2, \ldots, \mathbf{x}_{n_2}^2\}$, where $\Xi = \Xi_1 \cup \Xi_2$, n, n_1 and n_2 are the numbers of observations in Ξ, Ξ_1 and Ξ_2, respectively, $n = n_1 + n_2$. FDA aims to finding a projection direction that maximizes the separation of class means and minimizes the within-class variance. Usually, this is achieved by solving an optimization problem. In order to link least squares regression to FDA, a predictor matrix is defined as

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_1 & \mathbf{X}_1 \\ -\mathbf{1}_2 & -\mathbf{X}_2 \end{bmatrix}, \tag{1}$$

where $\mathbf{1}_i$ is a column vector containing n_i ones, and $\mathbf{X}_i$ is an n_i-by-d matrix whose rows are the samples belonging to Ξ_i. Define a response vector $\mathbf{y}$ as

$$\mathbf{y} = \begin{bmatrix} \dfrac{n}{n_1}\mathbf{1}_1 \\ \dfrac{n}{n_2}\mathbf{1}_2 \end{bmatrix}, \tag{2}$$

where the constants n/n_i compensate the effect of the unbalanced sample sizes. For a least squares regression problem $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$, the regression coefficient vector $\boldsymbol{\beta}$ can be obtained by minimizing the residual sum of squares (RSS). The objective function is formulated as

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3}$$

Denote $\boldsymbol{\beta} = \begin{bmatrix} W_0 \\ \boldsymbol{\omega} \end{bmatrix}$, where W_0 is the coefficient corresponding to the first column in $\mathbf{X}$ and the vector $\boldsymbol{\omega}$ contains the remaining coefficients. In [20], it is be proved that $\boldsymbol{\omega}$ is the direction vector that FDA searches for. In other words, the solution of the least squares regression problem is identical to that of FDA.

Accordingly, the multivariate fault isolation problem can be further transformed to a variable selection problem in regression analysis. Recent research in the field of statistics shows that penalized regression, such as LASSO and EN, is a promising technique in variable selection for regression models [16,17], which minimizes the residual sum of squares under certain constraint(s). By doing so, penalized regression tends to produce a number of zero coefficients enforcing sparsity in the solution, and thereby conducts parameter estimation and variable selection simultaneously. In addition, both LASSO and EN can be solved efficiently using elegant algorithms based on least-angle regression (LARS) [21].

The above findings motivate this paper.

3. Penalized regression for multivariate fault isolation

3.1. LASSO-based fault isolation

The variable selection problem in least squares regression can be described in an optimization form as

$$\min_{\boldsymbol{\beta}} ((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \left\|\boldsymbol{\beta}\right\|_0), \tag{4}$$