Universidade Estadual de Campinas
Faculdade de Engenharia Química
Departamento de Processos Químicos
Área de Concentração: Desenvolvimento de Processos Químicos

Técnicas Estatísticas Multivariadas para o Monitoramento de Processos Industriais Contínuos

(Multivariate Statistical Techniques for the Monitoring of Continuous Industrial Processes)

Tese apresentada à Faculdade de Engenharia Química como parte dos requisitos exigidos para a obtenção do título de Doutor em Engenharia Química

Autora: Celeste María Díaz Cónsul
Orientador: Prof. Dr. Rubens Maciel Filho

Campinas- São Paulo-Brasil
19 de Março de 2002

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

D543t

Díaz Cónsul, Celeste María
Técnicas estatísticas multivariadas para o monitoramento de processos industriais contínuos / Celeste María Díaz Cónsul. --Campinas, SP: [s.n.], 2002.

Orientador: Rubens Maciel Filho.
Tese (doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Química.

1. Controle de qualidade. 2. Análise de componentes principais. 3. Análise discriminante. 4. Engenharia – Métodos estatísticos. 5. Localização de falhas (Engenharia). I. Díaz Cónsul, Celeste María. II. Universidade Estadual de Campinas. Faculdade de Engenharia Química. III. Título.

Tese de Doutorado defendida por Celeste María Díaz Cónsul e aprovada, em 19 de março de 2002 pela banca examinadora constituída pelos doutores:

Prof. Dr. Rubens Maciel Filho – Orientador

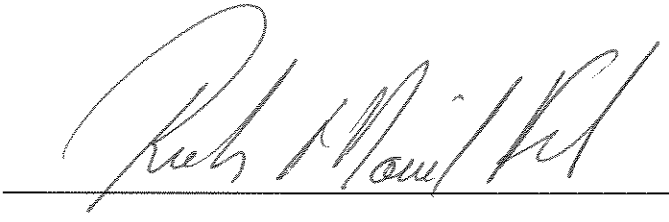
Prof. Dr. Antônio José Gonçalves da Cruz – UFSCAR

Dr. Odair Araujo - Rhodia

Prof. Dr. Antônio Carlos Luz Lisboa – UNICAMP

Dr. Rogério Favinha Martini – Pós-doutoramento/FAPESP

Esta versão corresponde à redação final da Tese de Doutorado defendida por Celeste María Díaz Cónsul e aprovada pela comissão julgadora em 19 de março de 2002.



Prof. Dr. Rubens Maciel Filho

A mis padres Ana María y Genaro
por su inmenso cariño y por ser
ejemplos completos de dedicación
al trabajo y superación profesional.
A mis queridas hermanas y a mis
adorables sobrinos.

Agradecimentos

Quando terminei o mestrado uma coisa da qual não tinha dúvidas era querer usar os conhecimentos do mestrado num doutorado aplicado. Fazer o doutorado em Engenharia Química é algo que tem representado para mim um estágio superior, do ponto de vista profissional, abrindo-me novos caminhos, novas inquietudes e ajudando-me a esclarecer meu futuro profissional.

Muito tem contribuído para isto meu orientador. O Prof. Rubens é para mim um exemplo de professor e pesquisador, além de uma excelente pessoa. Ele se dedica, com esmero e absoluta satisfação, aos seus alunos. Agradeço a ele por ter confiado em mim desde o primeiro momento.

Á Prof. Maria Regina pelo seu dinamismo, espírito positivo e animo infinito. Sua contribuição e muito importante no ambiente de trabalho do nosso laboratório, coisa que acho tem sido muito positiva para a realização desta tese. Aos companheiros do LOPCA e LPDS meu agradecimento.

Agradeço também a CAPES pelo período de suporte financeiro, embora curto serviu de incentivo para produzir mais resultados em menos tempo. Sem suporte financeiro fica difícil manter o ritmo de trabalho, só fazendo um grande esforço consegue-se terminar. Também acredito que o grande interesse que senti pelo assunto desta tese, desde o primeiro momento, fez com que o esforço fosse mais prazeroso.

Ao meu esposo, um agradecimento especial pelo seu amor e por sempre se preocupar e ocupar com minhas coisas, ajudando-me em tudo e me dando força.

Aos meus amigos Tuca, Marcelo, Marcelo Melo, Carol, Dani, Thais e Zé pelas conversas, pelo carinho e pela quantidade de conhecimentos de todo tipo que levarei comigo e me farão lembrar deles cada vez que for

empreender alguma coisa na minha vida. Para vocês um obrigada! Deixo por escrito meu agradecimento, minha a amizade e meu carinho incondicional.

Por último, embora não menos importante, agradeço a minha família, em Cuba e a July e Ruben aqui. Eles me animam e também, porque não dizer?, me mimam mesmo longe fisicamente. Muito obrigada a minha mãe, meu pai, minhas irmãs e meus sobrinhos. Suas lembranças me encorajam e me motivam ao me traçar novas metas e desafios. Gracias!

Abstract

In the industry of chemical processes, a lot of variables are manipulated and monitored at the same time. In these cases, it start to be of extreme importance the stages of data treatment and the development of models for representation of the process. One of the most important goals are detection and identification of faults in the process. Using multivariable statistical techniques, as Principal Components Analysis (PCA) and Fisher's Discriminant Analysis (FDA), is possible to take advantage of the data multivariable nature and it is possible to proceed with the detection of monitoring problems as well as diagnosing which the causes of these behaviors.

In this work is considered as study case a hydrogenation of phenol to cyclohexanol reactor. Historical data, with a great number of variables and observations, were collected during the operation of the process. The general idea of the method of PCA is to explain the covariance structure of the data through some few lineal combinations of the original variables, which try to reflect the dimensions truly important. The acting of the process then can be monitored in the space of the principal components, of smaller dimension. Using the model PCA was possible the evaluation and identification of a group of faults in the process. On the other hand, using a bank of faults, adequately built, FDA got to classify the observations with a good classification tax. A reflection on the importance of the use of these multivariable techniques for detection and fault diagnose is presented with the evaluation of the obtained results.

Resumo

Na indústria de processos químicos, geralmente varias variáveis são manipuladas e monitoradas ao mesmo tempo. Nestes casos passam a ser de extrema importância as etapas de tratamento de dados e o desenvolvimento de modelos de representação do processo para a detecção e identificação de falhas no processo. Usando técnicas estatísticas multivariadas, como Análise de Componentes Principais (PCA) e Análise Discriminante de Fisher (FDA), tira-se proveito da natureza multivariada dos dados e é possível proceder com a detecção de problemas de monitoramento no processo, assim como diagnosticar quais as causas destes comportamentos.

Neste trabalho consideramos como caso estudo um reator de hidrogenação do fenol a ciclo-hexanol. Dados históricos, com um grande número de variáveis e observações, foram coletados durante a operação do processo. A idéia geral do método da PCA é explicar a estrutura de variância e covariância dos dados através de umas poucas combinações lineares das variáveis originais, as quais tentam refletir as dimensões verdadeiramente importantes. O desempenho do processo poderá então ser monitorado no espaço das componentes principais, dimensionalmente menor. Usando o modelo PCA e alguns gráficos auxiliares foi possível a avaliação e identificação de um conjunto de falhas no processo. Por outro lado, usando um banco de falhas apropriadamente construído FDA conseguiu classificar todas as observações amostradas com uma boa taxa de classificação. Uma reflexão sobre a importância do uso destas técnicas multivariadas na detecção de falhas é apresentada junto a avaliação dos resultados obtidos.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Motivation

Large amounts of data are usually available in many chemical processes. The data can be analyzed to determine whether or not a fault has occurred in the process. A fault is defined as abnormal process behavior whether associated with equipment failure, equipment wear, or extreme process disturbances. This is very important in industry for efficiency, security, quality of products and environmental restrictions.

Techniques for analysis of complex data sets, with a great number of measured variables, is inside the field of the multivariate statistics. There are several techniques particularly important in industry of chemical process: Principal Components Analysis, Discriminant Analysis, Factorial Analysis and Cluster are some of them. Multivariate statistical methods have became very useful for their ability to describe major trends in a data set, specially Principal Components Analysis (PCA), which has been widely used for this purpose (Hiden et. al., 1999). This method in many ways forms the basis for multivariate data analysis (Wold et al., 1987). It is worthwhile mentioning, that PCA can also be used to accompany the variations of processes (Wetherill, 1991). In this context it is possible to consider the variability as information (Shunta, 1995).

The general idea of the method of the PCA is to explain the covariance structure of the data through some few lineal combinations of the original variables. The general objectives are: reduction of the data and interpretation (Johnson and Wichern, 1992).

In industry, a great number of variables are usually measured and stored, as databases, in the computer, during the operation of a process. These

variables, in general, are highly correlated and the real dimension of the monitored process is considerably smaller than that represented by the number of variables of the process collected. PCA reduces the dimensionality of the process creating a new group of variables, called principal components, which try to reflect the dimensions truly important. Then, the performance of the process can be monitored in the PCA space, dimensionally smaller. (Zhang, et. al, 1997).

PCA relies on the formation of a statistical model based on historical process data to establish normal operating behavior. New data is then compared with the normal operation model to detect a change in the system. This model can be handled with less variables than the original number, since an important feature of PCA is dimensionality reduction with relatively little loss of information, or at least with a prior knowledge about the desired information.

On the other hand, Fisher's Discriminant Analysis (FDA) is a multivariable technique that works with data that present a group structure, or class, known a *priori*. Exploratory by nature, this technique has as main objectives to find discriminant functions, or new axes, that describe graphically and algebraically the separation among the groups as well as rules that allow to classify a new individual in one of the known groups, minimizing the risk of misclassification. In this case it will be called as group to each type of fault.

1.2 Geometric illustration

To appreciate the behavior of these techniques in the space of the data, some appropriate graphical representations will be used to illustrate it.

1.2.1 Principal Components Axes

The geometrical representation of the principal components is illustrated in Figure 1 for two variables.

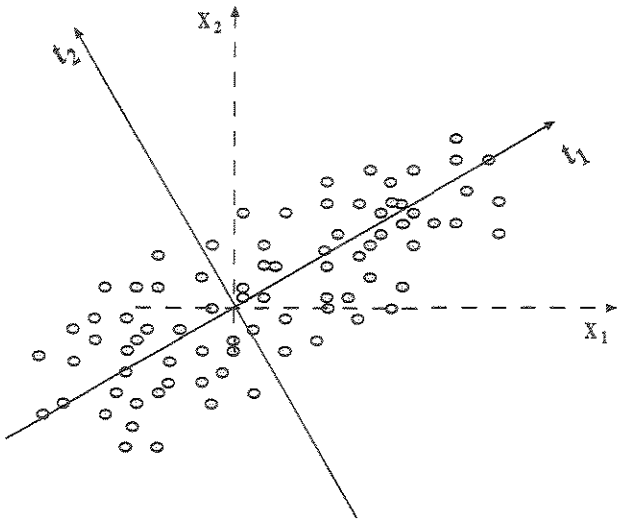
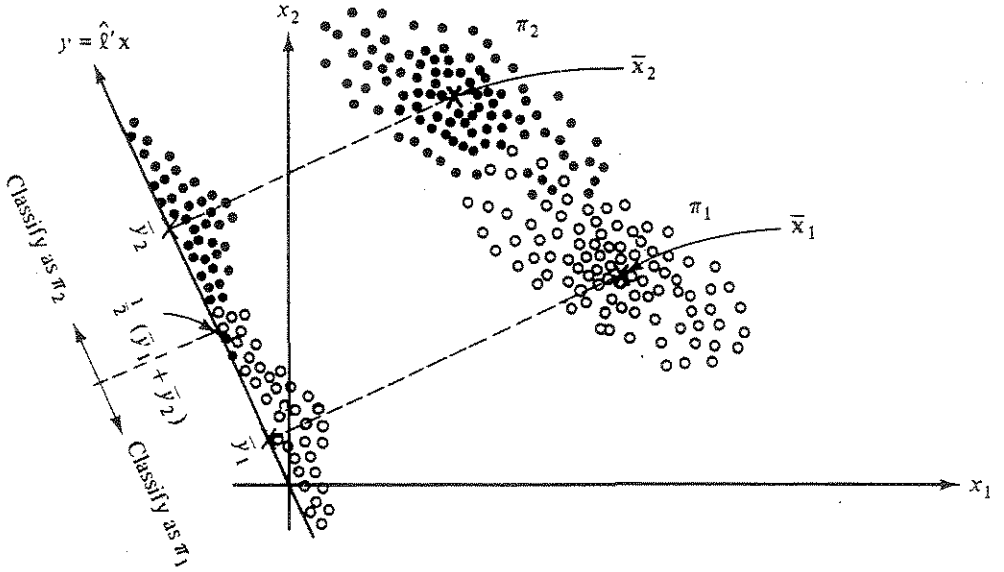


Figure 1. Hypothetical representation of Principal Components for two variables.

1.2.2 Fisher's Discriminant Axes

Fisher's procedure is illustrated in Figure 2, schematically, for two variables. All points in the scatterplots are projected onto a line in the direction $\hat{l}$, and this direction is varied until the samples are maximally separated.



Source: JOHNSON AND WICHERN, 1992.

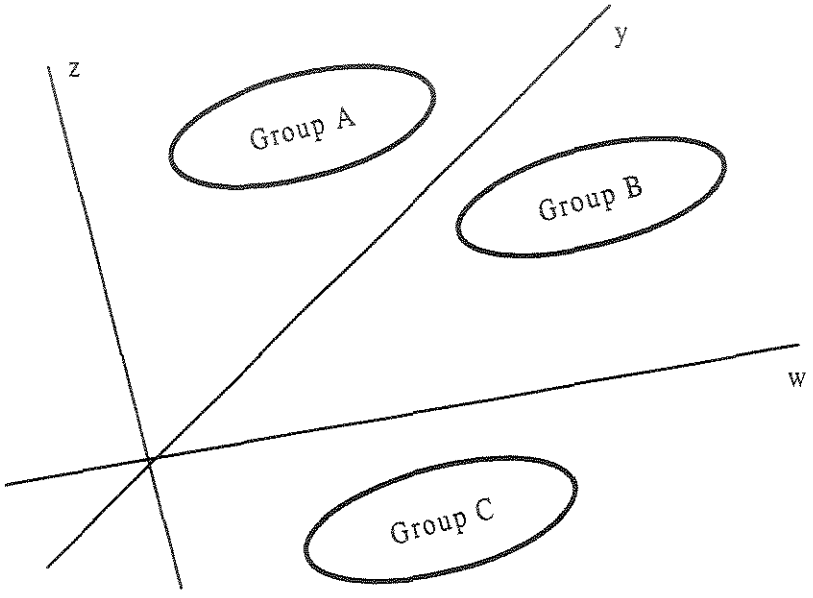
Figure 2. A pictorial representation of Fisher's procedure for two variables and two groups.

1.2.3 Geometric comparison of Principal Components and Fisher's Discriminant Axes

To have a geometric idea of the similarities and differences between PCA and FDA axes, it will be presented an example to proceed.

Let us consider three hypothetical groups of bivariate data whose graphic representations are in Figure 3. The three directions, y, w and z here represented are the following ones:

- When the structure of groups is unknown, and considering the collection of data as a whole, y is the first Principal Component, which describes the direction of maximum variability of the complete data set.



Source: KRZANOWSKI, W.J., 1988.

Figure 3. Hypothetical representation of the approximate directions for the first Principal Component y, when it interests the total variability; the first Principal Component w, when it interests the variability inside of the groups and z, the first Fisher's Discriminant Function, when it interests the variability among the groups.

- When the interest is the variability inside the groups, that is equal for all in this example, this can be seen in the direction of the Principal Component w.
- Finally, when it is the separation between the groups the most important, z, that it is the first Fisher's Discriminant Function, is

convenient to represent the direction where it is seen the better separation of the groups.

1.3 Scope, Goals and Approach

In general, the goal of this research was to determine how PCA and FDA could be used to enhance process monitoring and control. This general goal was broken down into two areas of application: process monitoring and process analysis. In each of these two areas there are specific questions which were attacked. These are outlined below:

Process monitoring: Can PCA and FDA be used as effective process monitoring tools? Earlier studies have shown that PCA appears to model the "normal" process variation and have indicated that PCA may be useful for identifying process upsets and failures. Other recent studies have shown that while PCA development models are based on data collect for each fault class, the FDA approach simultaneously uses all of the data to obtain a single lower dimensional model used to diagnose faults. Can limits be developed around the methods so that they can be used in a straightforward fashion for fault detection? These and other issues are considered in this work.

Process analysis: What can be learned about multivariate processes using PCA and FDA? Often, otherwise unrecognized relationships between variables and samples are made apparent when the data is subjected to PCA. Also, studies from other fields have indicated that FDA is useful as a pattern recognition technique. Does this hold in practice?

These were subjects that served as motivation for this work. It is intended as specific objective to build an itinerary that serves as guide to the use of PCA and FDA for detection and identification of faults in data of a continuous process. Using for this the digital exit of the program in Fortran, derived of the calculations, and also auxiliary graphs that are shown very useful for interpretation ends as well as to take decisions. In short, the objectives of this work are:

- to deepen in the use of the PCA and FDA techniques for detection and identification of faults,
- to implement the use of these techniques in Fortran seeking the subsequent use of this program starting from a group of industrial data,
- to explore the presentation of the results digitally and graphically,
- to build an itinerary with necessary indications of how to use these multivariate statistical techniques in a continuous chemical processes, having as target the personnel of industries, as users,
- to develop a software in Fortran with the facilities required to the users to take decisions based on the PCA and FDA methods.

1.4 Thesis Organization

The thesis is organized as follows.

First is given an introduction of the work, in Chapter 1, with an explanation of the motivation, the geometric meaning of the techniques of interest and the objectives of the work. Then, in Chapter 2, a recount of the Statistical Process Control is done and some of the papers published in this area until recent dates are commented.

A review with the theory about Principal Components Analysis and Fisher's Discriminant Analysis is made in Chapter 3. In Chapter 4 we talk about the proposed procedures and the form used to validate them. It will be explained the construction of the test data and the training data, PCA model and the data set for performing FDA.

The results obtained are showed in Chapter 5. An analysis of the behavior of the PCA model is made for each type of fault studied as well as the results of the application of FDA.

In Chapter 6 we draw an itinerary that has as objectives to help as a guide in the application of these techniques to the multivariable process control. Discussions and conclusions are presented in Chapter 7. Recommendations with proposed future works and other applications are found in Chapter 8.

Chapter 9 is dedicated to bibliographic references and Appendix 1 and Appendix 2 present some theoretical results of interest for PCA and FDA. The flowcharts for Fortran routines are presented in Appendix 3.

Chapter 2: History and Bibliographic Review

2.1 Recounts of the appearance of Statistical Process Control

The Statistical Process Control (SPC) is an important tool in the modern industry. SPC and related techniques of survey inspection were developed in the last century. In May of 1924 Walter A. Shewhart of Bell Telephone Laboratories made the first sketch of a graph of modern control. In 1931 the important paper of the new techniques was presented to the Royal Statistical Society. SPC was used widely in the Second World War in England and in the United States, but it lost some importance when the industries abandoned the warlike production. The Japanese industry applied SPC thoroughly and it proved the benefit of it. Countries as England and the United States are being forced to introduce SPC with the objective of competing with the Japanese (Wetherill and Brown, 1991).

SPC examines if a process is working in the due way or not, evaluating collected data. If abnormalities are detected, the idea is to determine the reasons for this behavior and to eliminate the causes, producing solutions using statistical techniques (Ipek et al., 1999).

Although it usually understands both as the same thing, there exist differences between statistical control of quality and statistical control of processes. With the first, traditionally the product quality properties are charted to determine if the process is in state of "statistical control", traditional multivariate control charts are shown to be very effective for detecting events when the multivariate space is not too or ill-conditioned. However, product quality data may not be available frequently, but only every few hours. However, many process measurements such as the temperature profile down the reactor, for instance, the coolant

temperatures, and the solvent and initiator flowrates are available on a frequent basis. The signature of any special events or faults occurring in the process that will eventually affect the product, should also appear in the process data X. Therefore monitoring the process may be preferable. By looking at the process as well as the quality variables, in fact statistical process control has been considered, as opposed to statistical quality control (SQC), as mentioned by Kourti and MacGregor, 1996.

The Multivariate Statistical Control Process, according to Saibt, et. al, 1996, consists of two basic procedures: the control of the means and the control of the variability of the process. SPC should be seen as a statistical analysis of the variations of the process and their causes. The differences among the decisions taken based on facts and those taken only using intuition can be enormous (Wetherill and Brown, 1991).

In any production process, some variation in the quality of the products is inevitable. The built-in theory in the graphs of control originated from the graphs of Shewhart is that this variation can be divided in two categories: random variations and variations due to special causes. These last ones refer to causes on the ones which some control type is considered, for instance, differences in the quality of the raw material, new workers or no specialized, among others. However, the random variations are the variations of the quality due to many complex causes, each one influencing the process slightly. Little can be done in this case, unless the process is modified in its basic requirements.

In some industries, mainly of manufacture, the control graphs are one of the most effective ways to discover when a process is "out of control" in a cheap and safe way. When the process is working abnormally the "sign" will appear in the graph. In the industry of processes, particularly, the situation is more complicated because it is not always clear what to graph or what to do in the case of signs of "out of control". Frequently there only exists a vague knowledge of the relationships among many of these variables. A SPC in such cases involves much more than control graphs.

At the present time many multivariate statistical techniques are being

used in the control of processes. To proceed, it is presented a summary of the great amount of recent published papers about the use of it.

2.2 Published Papers

A paper that described quality control methods for two variables was discuss by Jackson, 1956. The use of PCA for quality control was first suggested in the early paper by Jackson, 1959. Here PCA is introduced both as a method of characterizing a multivariate process and as a control tool associated with control procedures. Jackson and Mudholkar, 1979, discuss the treatment of residuals associated with PCA.

More details on the use of PCA were later provided by Jackson, 1980 and 1985. Jackson's book, 1991, widely referenced, provides a user's guide to principal components, with a compilation of theory and applications.

This method, PCA, has been used and extended in various applications, some examples are:

Kaufmann, 1993, who constructs a model, using PCA, that detects adulteration of edible oils, i.e., where high-priced commodity oils are mixed with lower-priced substitutes. A plot of the first two principal components showed the spread of the different authentic types of oils in the chromatographic measurement space.

Another approach is the sensor fault identification and reconstruction using PCA. In the paper of Dunia et al, 1996, the PCA model captures the measurements correlations and reconstructs each variable by using iterative substitution and optimization. The effect of different sensor faults on model based residuals is analyzed and a new indicator called SVI is defined to determine the status of each sensor. An example using boiler process data demonstrates the attractive features of this indicator.

In their paper, Kosanovich et al., 1996, discuss a variant of PCA, multiway PCA, used to analyze data taken from an industrial batch process. They show in that work that multiway PCA can be used to identify major sources of variability in the processing steps, improving process

understanding.

Wise and Gallagher, 1996, reviews the chemometric approach, the application of mathematical and statistical methods to the analysis of chemical data, to chemical process monitoring and fault detection. They used PCA and other multivariate statistical techniques to assist their goals.

The paper of Martin et al., 1999, reviews the concept of process performance monitoring through an industrial application to a fluidized bed-reactor and a comprehensive simulation of a batch methyl methacrylate polymerization reactor, using PCA and multiway PCA, respectively.

Valle et al., 1999, comment on how principal component analysis has wide applications in signal processing, chemometrics, and chemical processes data analysis. They propose a method, the variance of the reconstruction error criterion, with the comparison to other methods, to select the number of principal components to be retained.

Ralston et al., 2001, use PCA for process modeling, monitoring, fault detection and diagnosis. An enhancement is made by using confidence limits on the residuals of each variable for fault detection. Their results show that the time required for fault detection, using a MATLAB toolbox, is reduced. They identified ways to more effectively monitor processes and to more promptly detect and diagnose faults when they occur, using PCA. A chemical process is used as case study.

FDA provides an optimal lower dimensional representation in terms of discriminating among classes of data (Duda and Hart, 1973; Hudlet and Johnson, 1977), where for fault diagnosis, each class corresponds to data collected during a specific known fault.

According to Russell and Braatz, 1998, PCA has great properties in terms of detection of faults. However they discuss the advantages, from the theoretical point of view, of FDA on PCA, in the item of isolation of the fault.

Chiang et al., 2000, compare the potentialities of FDA, Discriminant

Parcial Least Square (PLS) and PCA. Although FDA has been widely studied in the pattern classification literature and is only slightly more complex than PCA, its use for analyzing data of processes is not frequently found described in the literature. As Chiang et al., 2000, analyzed, this is interesting, since FDA should outperform PCA when the primary goal is to discriminate among groups. They suspect that part of the reason that FDA has been ignored in the chemical process control literature is that more chemical engineers read the statistic literature (where PCA is dominant) than the pattern classification literature (where FDA is dominant).

Chapter 3: Theory and Basic Knowledge of Principal Components Analysis and Fisher's Discriminant Analysis

3.1 Introduction

In this Chapter the theoretical base of the methods of PCA, section 3.2, and FDA, section 3.3, is presented. First we explain the theory in the usual form, i.e., as it is found in the statistical literature (for instance in Johnson and Wichern, 1992). In a second moment the approach is presented for their application in the detection and diagnosis of faults in continuous processes.

3.2 Principal Components Analysis

3.2.1 Theory overview of PCA

PCA is an optimal dimensionality reduction technique in terms of capturing the variance of the data. For a given data matrix X with n rows (observations) and p columns (measurement variables) the covariance matrix of X is defined as

$$\mathbf{S} = \mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}'\mathbf{X}}{n-1} \tag{3.2.1}$$

The X matrix will be called the original data matrix, in this work. For practical convenience, data matrix X is an 'autoscaled' matrix; i.e, adjusted to a mean zero and unit variance by substracting the column averages and dividing each column of the original process data by its standard deviation.

For this reason the equation (3.2.1) is also the correlation matrix for X.

It is good to observe in this point that the standardization affects the size of the coefficients and therefore the relative importance of the variables in the hour of the interpretation; the explanatory power of each component changes because it also changes the magnitude of the eigenvalues. Therefore, in PCA the standardization should be treated carefully. It is recommended in some cases, being more used in the presence of different variables and scales of different measurement in the original variables. Given that, in our case, it will work with a large group of industry variables monitored, which present scales of different measurement, the standardization is justified[1].

Mathematically, PCA relies on an eigenvector decomposition of a symmetric, non singular matrix, such as the covariance matrix of the process variables, S (Johnson and Wichern, 1992). It may be reduced to a diagonal matrix Λ by premultiplying and posmultiplying by a particular orthonormal matrix P, this is $\mathbf{P'SP} = \mathbf{\Lambda}$. The diagonal elements of Λ, λ_1, λ_2, ..., λ_p, are called the characteristic roots, latent roots, or eigenvalues of S. The columns of P, p_1, p_2, ..., p_p, are called characteristic vectors, or eigenvectors of S. These characteristic roots may be obtained from the characteristic equation

$$|\mathbf{S} - \lambda\mathbf{I}| = 0 \qquad (3.2.2)$$

where I is the identity matrix the order *p*. The characteristic vectors may be obtained by solution of the equations

$$[\mathbf{S} - \lambda_i\mathbf{I}]\mathbf{q_i} = 0 \qquad (3.2.3)$$

Independently of the length of vector q_i.. For this reason, to obtain a single solution for the problem above, it is convenient to restrict attention to characteristic vectors of unit length, p_i (Johnson and Wichern, 1992). This normalization would be as in expression (3.2.4).

---

[1] A wider discussion on this aspect can be seen in Consul, 2000, section 3.1.

$$\mathbf{p}_i = \frac{\mathbf{q}_i}{\sqrt{\mathbf{q}_i'\mathbf{q}_i}} \tag{3.2.4}$$

for i=1, 2, ..., p.

Geometrically, the described PCA procedure is nothing more than an axis rotation of the covariance matrix, and the elements of the eigenvectors are the direction cosines of the new axes related to the old ones (Geladi and Kowalski, 1986).

Mathematically, in PCA the eigenvectors are the coefficients of the linear combinations of the original variables, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, that transform *p* correlated variables into *p* new uncorrelated variables $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p$ using the following transformation:

$$\mathbf{T}_{nxp} = \mathbf{X}_{nxp}\mathbf{P}_{pxp} \tag{3.2.5}$$

The $\mathbf{P}_{pxp}$ matrix has *p* columns, which are the eigenvectors, and *p* rows, related to each original variable. The columns of the transformed variables matrix T are called principal components of data matrix X. Generally in PCA, $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p$ are called score vectors and $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_p$ loading vectors.

The *i*th principal component would be written as

$$\mathbf{t}_i = \mathbf{X}\mathbf{p}_i \tag{3.2.6}$$

where

$$\text{var}(\mathbf{t}_i) = \mathbf{p}_i'\mathbf{S}\mathbf{p}_i \quad \text{and} \quad \text{cov}(\mathbf{t}_i, \mathbf{t}_k) = \mathbf{p}_i'\mathbf{S}\mathbf{p}_k, \quad \text{for i=1, 2, ..., p.}$$

Here $\mathbf{t}_i$ and $\mathbf{p}_i$ are *n*x1 and *p*x1 vectors, respectively. In detail, the components of the *i*th principal component vector $\mathbf{t}_i$ are calculated as

$$\begin{aligned} t_{1i} &= x_{11}p_{1i} + x_{12}p_{2i} + \ldots + x_{1p}p_{pi} \\ t_{2i} &= x_{21}p_{1i} + x_{22}p_{2i} + \ldots + x_{2p}p_{pi} \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \\ t_{ni} &= x_{n1}p_{1i} + x_{n2}p_{2i} + \ldots + x_{np}p_{pi} \end{aligned} \tag{3.2.7}$$

In other words, the principal components are those uncorrelated linear combinations $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p$ whose variances are as large as possible. The first

principal components in the linear combination with maximum variance. That is, it maximizes $\text{var}(\mathbf{t}_i) = \mathbf{p}_i' \mathbf{S} \mathbf{p}_i$.

The optimization criterion for PCA, before described, can be written also in the form (Russell and Braatz, 1998):

$$\max_{\mathbf{p} \neq 0} \frac{\mathbf{p}' \mathbf{S} \mathbf{p}}{\mathbf{p}' \mathbf{p}} \tag{3.2.8}$$

It Can be shown that each principal component t_i will have mean zero and variance equal to eigenvalue λ_i, for the result[2] to follow:

Result 1: Let S be the covariance matrix associated with the random vector $\mathbf{X} = [\mathbf{X}_1, \ \mathbf{X}_2, \ \ldots, \ \mathbf{X}_p]$. Let S have the eigenvalue-eigenvector pairs (λ_1, e_1), (λ_2, e_2), ..., (λ_p, e_p), where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. The ith principal component is given by

$$\mathbf{t}_i = \mathbf{X}\mathbf{p}_i = \mathbf{X}\mathbf{p}_{1i} + \mathbf{X}\mathbf{p}_{2i} + \cdots + \mathbf{X}\mathbf{p}_{pi}, \quad i = 1,2,\ldots,p$$

with these choices,

$$\begin{aligned} \text{var}(\mathbf{t}_i) &= \mathbf{p}_i' \mathbf{S} \mathbf{p}_i = \lambda_i \\ \text{cov}(\mathbf{t}_i, \mathbf{t}_k) &= \mathbf{p}_i' \mathbf{S} \mathbf{p}_k = 0 \end{aligned} \quad , \text{ for i=1, 2, ..., p.}$$

From result 1, the principal components are uncorrelated and have variances equal to the eigenvalues of S.

3.2.1.1 *Proportion of explained variance by each principal component*

To calculate the variance proportion explained by each principal component needed first to calculate the total variance. Let us see the following result[3] (Johnson and Wichern, 1992):

Result 2: Let $\mathbf{X} = [\mathbf{X}_1, \ \mathbf{X}_2, \ \ldots, \ \mathbf{X}_p]$ have covariance matrix S, with eigenvalue-eigenvector pairs (λ_1,e_1), (λ_2,e_2), ..., (λ_p,e_p), where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. Let $\mathbf{t}_1 = \mathbf{X}\mathbf{p}_1$, $\mathbf{t}_2 = \mathbf{X}\mathbf{p}_2$, ..., $\mathbf{t}_p = \mathbf{X}\mathbf{p}_p$ be the principal components . Then

---

[2] The complete result and its proof can be seen in Appendix 1, section A1.1.

[3] The proof of Result 2 can be seen in Appendix 1, section A1.2.

$$s_{11} + s_{22} + \ldots + s_{pp} = \sum_{i=1}^{p} \text{var}(\mathbf{X}_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_P = \sum_{i=1}^{p} \text{var}(\mathbf{t}_i)$$

Result 2 says that $\text{Total sample variance} = \sum_{i=1}^{p} s_i = \sum_{i=1}^{p} \lambda_i$ and consequently, the proportion of total variance due to, or explained by, the *k*th principal component is:

$$\left(\begin{array}{c} \text{Proportion of total variance} \\ \text{explained by kth principal components} \end{array} \right) = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}, \quad k = 1,2,\ldots,p \tag{3.2.9}$$

If most of the total variance, for large *p*, can be attributed to the first components, then these components can "replace" the original *p* variables without much loss of information.

3.2.2 Tools to develop a PCA model

One interesting property of principal component is the fact that equation (3.2.5) can be inverted to

$$\mathbf{X} = \mathbf{T}\mathbf{P}' \tag{3.2.10}$$

due to the fact that P is orthonormal so that $\mathbf{P}^{-1} = \mathbf{P}'$. Then, multiplying by $\mathbf{P}'$ in both sides of equation (3.2.5), $\mathbf{T}\mathbf{P}' = (\mathbf{X}\mathbf{P})\mathbf{P}' = \mathbf{X}\left(\mathbf{P}\mathbf{P}^{-1}\right) = \mathbf{X}$. Equation (3.2.10) is called PCA model (Wold et al., 1987).

The *p* principal components reproduce the whole covariance structure of the original data. However, using only a few principal components it is possible to reach a high percent of explained variance that can be enough for the purposes of fault diagnosis and to explain the dynamic behavior. Let us suppose that it is decided to be alone with *k* principal components, $k<p$. Then an approximation of the data matrix X can be written as

$$\hat{\mathbf{X}} = \mathbf{T}_{nxk}\,\mathbf{P}'_{kxp} \tag{3.2.11}$$

Here the residual matrix E appears, representing the percent of variance not explained by the PCA model. In this case, model (3.2.10) becomes:

$$\mathbf{X} = \mathbf{T}_{nxk}\,\mathbf{P}'_{kxp} + \mathbf{E} \qquad (3.2.12)$$

3.2.2.1 *Reduction of dimensionality*

Several methods to decide what is the appropriate number of principal components to be chosen exist. It is possible to use the SCREE test, that is a graphical technique widely used for this goal. It consists on plotting all of the characteristic roots, eigenvalues, of the covariance matrix, the values of the roots themselves being the ordinate and the eigenvalues, the abscissa. If the graphic has one break in it, this procedure is a good and easy way to select the principal components number to be retained. In other cases it could be difficult to reach a conclusion and others methods should be used (see Jackson,1991 and Valle et al., 1999). There is a plethora of methods to calculate the number of PC, for example: cumulative percent variance, scree test on residual percent variance, average eigenvalue, parallel analysis, cross validation, etc. As Valle et al., 1999, analyze, the decision to choose the number of principal components is very subjective. Russell, 1998, after a careful analysis, comments that there appears to be no dominant technique. Ku et al., 1995, recommend the parallel analysis method, because in their experience, it has performed overall the best.

Horn, 1965, had already proposed this method (parallel analysis). He suggested generating a random data set having the same number of variables and observations as the set being analyzed. These variables should be normally distributed but uncorrelated. A SCREE plot of these eigenvalues will generally approach a straight line over the entire range. The intersection of this line and the SCREE plot for the original data should indicate the point separating the retained and unretained principal components. The reason for that is that any eigenvalues for the real data, which are above the line obtained for the random data, represent eigenvalues that are larger than they would be by chance alone.

This procedure, comparing the singular value profile to that obtained by

assuming independent measurement variables, will be used in this work to choose the number of principal components to be retained in the PCA model. The dimension is determined by the point at which the two lines cross. This approach is particularly attractive since it is intuitive and easy to automate (an example of this procedure will be seen in Figure 9).

Once the principal components have been obtained from matrix data X, new data can be referenced against the model. For an entire data set X, where X is the new data matrix that has been scaled to the mean and standard deviation of the model data set, the residual matrix E, representing the percent of variance not explained by the PCA model is calculated as

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{T}\mathbf{P}' = \mathbf{X} - \mathbf{X}\mathbf{P}\mathbf{P}' = \mathbf{X}(\mathbf{I} - \mathbf{P}\mathbf{P}') \tag{3.2.13}$$

where matrix P_{pxk} is the matrix with the k eigenvector selected.

3.2.2.2 *Fault detection*

For any sample, a row of the new X, $\mathbf{x}_i'$, the sum of squared residuals is a scalar value sometimes referred to as a lack of fit statistic, Q. For the ith sample,

$$Q_i = \mathbf{e}_i'\mathbf{e}_i = \mathbf{x}_i'(\mathbf{I} - \mathbf{P}_{pxk}\mathbf{P}_{kxp}')\mathbf{x}_i \tag{3.2.14}$$

The Q statistic provides a way to test whether the process has shifted outside normal operation (Ralston et al., 2001). It is a measure of the amount of variation in each sample not captured by the selected number of principal components retained in the model. Variation of data within a confidence limit established for Q from normal data represents process noise. This confidence limit is calculated as

$$Q_\alpha = \theta_1\left[\frac{c_\alpha\sqrt{2\theta_2 h_o^2}}{\theta_1} + 1 + \frac{\theta_2 h_o(h_o - 1)}{\theta_1^2}\right]^{\frac{1}{h_0}} \tag{3.2.15}$$

where $\theta_i = \sum_{j=k+1}^{n} \lambda_j^i$, for $i = 1,2,3$, and $h_o = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$. c_α is the normal deviate corresponding to the upper (1-α) (Wise, 1991).

When the variation of the data is outside the defined confidence limits, the model has not captured the majority of the variance; therefore, the PCA model does not describe the data adequately. In the latter situation, the data are identified as faulty data (McGregor, 1995).

While the Q statistics offer a way to test if the process data has shifted outside the normal operating space, there is a need for a statistic that provides an indication of unusual variability within the normal subspace. This, the normal subspace, may be provided by Hotelling's T^2 statistic (Wise, 1991; Wise and Gallagher, 1996; Jackson, 1991, 1981, 1979). Kourti and MacGregor, 1996, also show that normal operations can be characterized by employing Hotelling's T^2 statistic. For any new sample $\mathbf{x}_i'$, the T^2 value is defined as

$$\mathbf{T}_i^2 = \mathbf{t}_i \Lambda^* \mathbf{t}_i' = \mathbf{x}_i \mathbf{P} \Lambda^* \mathbf{P}' \mathbf{x}_i \qquad (3.2.16)$$

where t_i in this instance refers to the ith row of T_{nxk}, the matrix of k scores vectors from the PCA model. The matrix Λ^* is a diagonal matrix containing the inverse eigenvalues associated with the k eigenvectors (principal components) retained in the model.

Statistical confidence limits for T^2 can be calculated by means of the F-distribution (Johnson and Wichern, 1992) as follows

$$T_{k,n,\alpha}^2 = \frac{k(m-1)}{m-k} F_{k,n-1,\alpha} \qquad (3.2.17)$$

While the Q limit defines a distance off the space that is considered unusual for normal operating conditions, T^2 limit defines an ellipsoid on the space within which the operating point normally projects, see Figure 4.

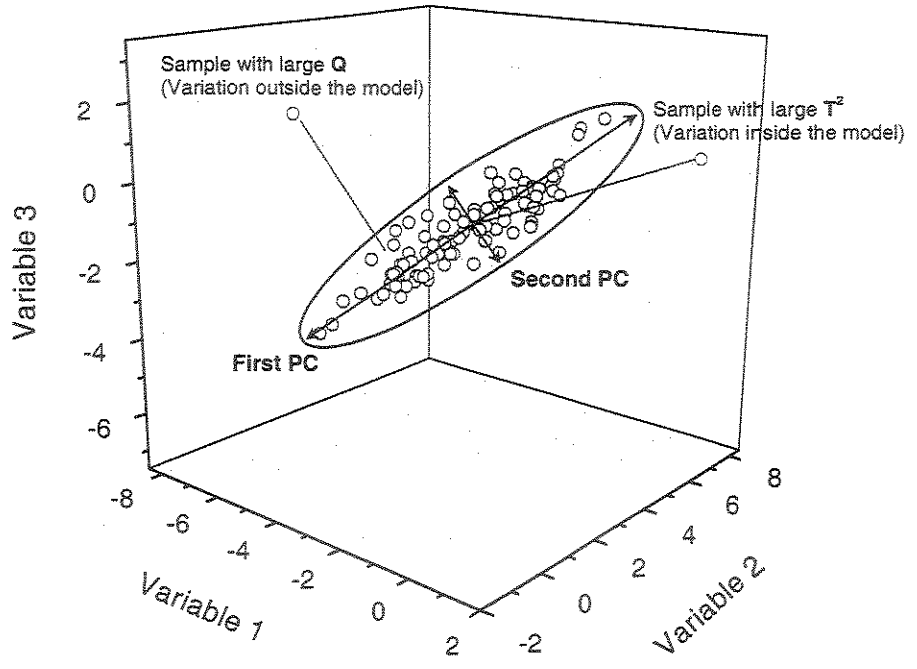


Figure 4. Geometric representation of Q and T^2 statistics.

3.3.2.3 Fault diagnose

Q residual contribution plots provide a way to diagnose a fault. The plot represents the Q residual versus a sample number or a grouping of sample numbers. This gives an approximation to the time a particular fault occurred. These plots are bar graphs of each variables Q contribution. This information is calculated by computing the means of the columns of the E, the residual matrix. As mentioned before, the residual matrix is made up of m samples (row) by n variables (columns). If a certain variable has extremely large residuals for a certain time frame, this contribution plot allows one to narrow down the fault source.

To generate a Q residual contribution plot, two cases need to be considered. The first of which is rather simple and the second is just one step more complicated. The first case deals with finding the Q residual contribution plot for a single sample. To accomplish this, the values listed in the specific sample's row would be plotted.

The second case deals with finding the Mean Q residual contribution plot for a group of samples. This is a little more involved than case one. If Q residual contribution plot for samples n to n+8 is desired, rows n to n+8 are needed. For those rows, each column's mean is calculated. The mean of each column, will give the Q residual contribution for each variable over the specified number of samples from n to n+8.

After having calculated the PCA model, the residual matrix is obtained. A Q residual contribution plot is given showing data for each variable's contribution over samples 26 to 31. To verify what is happening but is not visible to the user, look at the following sample residual matrix (Table 1 below). As always, the rows represent the sample number and the columns represent the variables. For simplicity the matrix is taken to have rows 1-8 and columns 1-8 representing 8 samples on each of the 8 variables respectively.

Table 1. Example of a sample residual matrix.

0.519	0.358	0.058	0.571	0.208	-0.477	1.245	-0.268
1.324	1.170	0.625	-4.049	-0.177	-0.030	0.318	-0.191
0.593	0.383	0.074	0.301	0.408	-0.276	1.114	-0.357
0.400	0.379	-0.220	0.444	0.431	-0.398	1.090	-0.439
-0.166	-0.209	-0.421	0.286	-0.033	-0.082	-1.879	-2.021
-0.014	-0.043	0.626	1.116	-0.410	-0.573	-1.876	-1.972
0.902	1.053	-0.020	0.408	0.080	0.992	-3.432	-2.215
0.332	0.190	-0.046	0.682	-0.277	-0.689	-1.346	-1.925

To create the Q residual contribution plot for sample number five (i.e. row number five in the Table 1), calculate the mean of each variable (column) and plot them. Here since only one sample is being considered the mean will essentially be the value listed in the cell. To create the Q residual contribution plot for samples one to eight (i.e. row number one to eight in the table above), calculate the mean of each variable (column) and plot them. The Q residual contribution plot for samples one to eight is shown below with the calculated variable means.

Table 2. Q residual contribution plot data for data from Table 1.

0.68501	0.41026	0.08432	-0.03013	0.02881	-0.19168	-0.59570	-1.17367

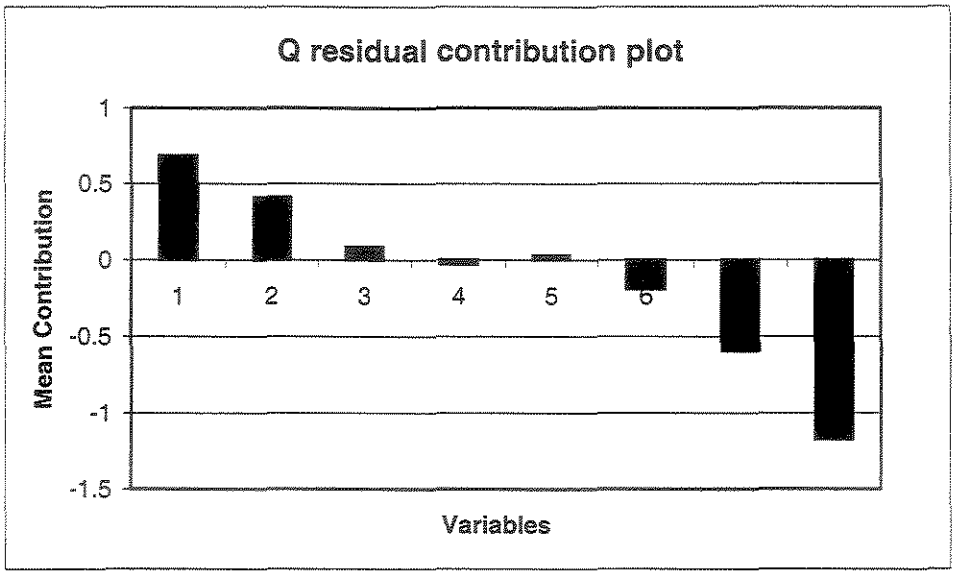


Figure 5. Q residual contribution plot for sample residual matrix in Table 1.

3.3 Fisher's Discriminant Analysis

3.3.1 Theory overview of FDA

The problem of Discriminant Analysis is characterized when is considered *n individuals* or observations, described by a group of *p* quantitative *variables*, $X_1, X_2, \cdots, X_P$, separate in *s groups*, defined a *priori* by an indicative variable. The matrix of data can be written, proceeding the following notation:

$$\mathbf{X}_{nxp} = \begin{bmatrix} X'_{11} \\ \vdots \\ X'_{1n_1} \\ \vdots \\ X'_{s1} \\ \vdots \\ X'_{sn_s} \end{bmatrix} \qquad (3.3.1)$$

where each $X_{ij} = \begin{bmatrix} x_{ij1} \\ \vdots \\ x_{ijp} \end{bmatrix}$ is such that i=1, 2,...., s; j =1, 2,...., n_s. The sub-index i refers to the number of groups and j identifies the number of observations inside of each group. The matrix of data[4] can be written in full detail as:

$$\mathbf{X}_{nxp} = \begin{bmatrix} x_{111} & \cdots & x_{11p} \\ \vdots & & \vdots \\ x_{1n_1 1} & \cdots & x_{1n_1 p} \\ \vdots & & \vdots \\ x_{s11} & \cdots & x_{s1p} \\ \vdots & & \vdots \\ x_{sn_s 1} & \cdots & x_{sn_s p} \end{bmatrix} \qquad (3.3.2)$$

Fisher's idea (Johnson and Witchern, 1992) was to transform the multivariate observations X_{ij}, i=1, 2,...., s; j =1, 2,...., n_s, to univariate observations Y in such way that the Y's derived from one population were separated as much as possible of those derived of the other populations. An important point is that Fisher's approach does not assume that the populations are normal. It does, however, implicitly assume the population covariance matrices are equal[5], $\Sigma_1 = \Sigma_2 = ... = \Sigma_s = \Sigma_W$, because a pooled estimate of the common covariance matrix, $\mathbf{S}_{pooled} = \mathbf{S}_W$, is used as the estimate of covariance matrix within of the populations, Σ_W. The expression of this matrix can be seen to proceed:

---

[4] It is good to emphasize here that this section will be developed without considering the matrix X as being standardized. This because in the case of FDA, the existence of a closed relationship among the coefficients obtained starting from the standardized data and of the data without transforming has been shown (Cónsul, 2000, p.30). Therefore, the development of this section will be made in the most general case.

5 In practice it is common that it doesn't come true the hypothesis S_1=S_2 =.... = S_W, however, the FDA is robust and it can be applied and to work perfectly in these cases (see Gilbert, E.S., 1969).

$$\mathbf{S}_{pooled_{(pxp)}} = \frac{1}{n-s}\sum_{l=1}^{s}\frac{(n_l-1)}{n-s}\mathbf{S}_l = \frac{1}{n-s}\sum_{l=1}^{s}\sum_{t=1}^{n_l}(x_t-\bar{x}_{lt})(x_t-\bar{x}_{lt})' = \frac{1}{n-s}\sum_{l=1}^{s}(X-\bar{X}_l)(X-\bar{X}_l)' \quad (3.3.3)$$

with $i, j = 1,2,...,p$ and vectores X and $\bar{X}$, as defined in section 3.3.2, below.

The Fisher's Discriminant Analysis is a useful technique in practice because it helps to visualize better the separation between populations using some few lineal combinations of the variables, reducing the dimensionality of the problem.

In others hands, FDA provide an optimal lower dimensional representation in terms of discriminating among classes of data (Johnson and Wichern, 1992), where for fault diagnosis, each class corresponds to data collected during a specific known fault.

3.3.2 Fisher's Method for several populations

The interest will be to show the acting of the FDA when is considered each fault studied as a group. As usually with is accessible is a sample[6] of s groups, the notation will be worked in sample level and to proceed will be defined the expressions more used. The vector of means for the total sample will be represented in the following way:

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}_{px1}, \text{ with } \bar{x}_i = \sum_{l=1}^{s}\sum_{t=1}^{n_l}\frac{x_{lti}}{n_t} \quad (3.3.4)$$

with i=1, 2,...., s; j =1, 2,...., n_s. The vector of variables and the mean by group, respectively, as:

---

6 The population treatment for the FDA can be seen, for instance, in Johnson, R.A. and Wichern, D.W., 1992; Mardia et al., J.M., 1979 or Lachenbruch, P.A, 1975.

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}_{px1}, \quad \overline{X}_i = \begin{bmatrix} \overline{x}_{i1} \\ \vdots \\ \overline{x}_{ip} \end{bmatrix}_{px1}, \text{ with } \overline{x}_{ij} = \sum_{t=1}^{n_i} \frac{x_{tij}}{n_t} \tag{3.3.5}$$

On the other hand equation (3.3.6) it is the sum of squares and products crossed between the groups and $SS_W = (n-s)\mathbf{S}_{pooled}$ is the sum of squares and products crossed within the groups.

$$SS_B = \sum_{i=1}^{s} (\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})' \tag{3.3.6}$$

Let consider the lineal combination (3.3.7)

$$Y_{kx1} = \mathbf{C}_{kxp} X_{px1} \tag{3.3.7}$$

The matrix of coefficients C can be written as:

$$\mathbf{C}_{kxp} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & & \vdots \\ e_{i1} & e_{i2} & \cdots & e_{ip} \\ \vdots & \vdots & & \vdots \\ e_{k1} & e_{k2} & \cdots & e_{kp} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1' \\ \mathbf{e}_2' \\ \vdots \\ \mathbf{e}_i' \\ \vdots \\ \mathbf{e}_k' \end{bmatrix} \tag{3.3.8}$$

Explicitly, the lineal combinations are in the following way:

$$\begin{aligned} y_1 &= e_{11}X_1 + e_{12}X_2 + \ldots + e_{1p}X_p \\ y_2 &= e_{21}X_1 + e_{22}X_2 + \ldots + e_{2p}X_p \\ &\vdots \qquad \vdots \qquad \vdots \qquad \vdots \\ y_k &= e_{k1}X_1 + e_{k2}X_2 + \ldots + e_{kp}X_p \end{aligned} \tag{3.3.9}$$

and each one of these lineal combinations can be written, in a more reduced form, as:

$$y_i = \mathbf{e}_i' X \tag{3.3.10}$$

with $i = 1,2,..,k$ and where the vector of coefficients $\mathbf{e}_i'$ is the ith row of the matrix in equation (3.3.8). The lineal combination, in the way seen in equation (3.3.10) and conditioned to the population of interest, they have expectation equal to:

$$E(y_i) = \mathbf{e}_i' E(X \mid \pi_l) = \mathbf{e}_i' \overline{X}_l \qquad (3.3.11)$$

for the combination i and the group $l = 1,2,..,s$, and variance equal to

$$Var(y_i) = \mathbf{e}_i' \, Cov(X) \mathbf{e}_i = \mathbf{e}_i' \mathbf{S}_{pooled} \mathbf{e}_i \qquad (3.3.12)$$

$\mathbf{S}_{pooled}$, as defined in the expression (3.3.3), and X and $\overline{X}$ defined vectors as in (3.3.5) and (3.3.4).

On the other hand, the general mean for the lineal combinations y_i, $i = 1,2,..,k$, is a constant value equal to $\overline{y} = \sum_{i=1}^{k} \overline{y}_i$, being $\overline{Y} = \begin{bmatrix} \overline{y}_1 \\ \vdots \\ \overline{y}_p \end{bmatrix}_{px1}$ the vector of means for equation (3.3.7). However, since the n observations are divided in s groups, it will be convenient to describe this structure in the new variables y. For this is possible to call of y_{ij} the value of y for the observation j in the group i, with $i = 1,2,..,s$ and $j = 1,2,...,n_i$. In this way the mean of the group i will be $\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ and the general mean $\overline{y} = \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^{s} n_i \overline{y}_i$.

To search if the s groups are well differentiated would have to partition the total sum of squares SS_T, of y_{ij}, in the sum of squares between the groups SS_B, and the sum of squared within the groups SS_W.

$$SS_T = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2 = \mathbf{Y'HY} = \mathbf{e'X'HXe} = \mathbf{e'S}_T\mathbf{e} \qquad (3.3.13)$$

In this $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{11'}$ is the matrix of centralization, whose help makes possible to have convenient matrix representations for the sums of squares and crossed products,

$$SS_B = \sum_{i=1}^{s} n_i (\overline{y}_i - \overline{y})^2 = \sum_{i=1}^{s} n_i \{(\mathbf{e}'\overline{\mathbf{X}}_i - \overline{\mathbf{X}})\}^2 = \mathbf{e'S}_B\mathbf{e} \qquad (3.3.14)$$

$$SQ_W = \sum_{i=1}^{s}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2 = \sum_{i=1}^{s}\mathbf{e}'\mathbf{X}_i'\mathbf{H}_i\mathbf{X}_i\mathbf{e} = \mathbf{e}'\mathbf{S}_W\mathbf{e} \tag{3.3.15}$$

where $\mathbf{H}_i = \mathbf{I}_{n_i x n_i} - \frac{1}{n_i}\mathbf{1}\mathbf{1}'$. As it is known, the relationship among these sums of squares is the following:

$$SS_T = SS_B + SS_W \tag{3.3.16}$$

The Fisher's criterion is particularly attractive because it results of the use of the sum of squares and products crossed between groups and the sum of squared within of the groups. This means that the interest will be to maximized the rate given in the expression (3.3.17), so that the new variable has larger variability between groups relative to the variability within the groups.

$$\frac{SS_B}{SS_W} = \frac{\mathbf{e}'\mathbf{S}_B\mathbf{e}}{\mathbf{e}'\mathbf{S}_W\mathbf{e}} \tag{3.3.17}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are the covariance matrices between and within the groups, respectively. The form of the $\mathbf{S}_B$ matrix can be seen in (3.3.18).

$$\mathbf{S}_B = \begin{bmatrix} s^B_{11} & s^B_{12} & \cdots & s^B_{1p} \\ s^B_{21} & s^B_{22} & \cdots & s^B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s^B_{p1} & s^B_{p2} & \cdots & s^B_{pp} \end{bmatrix}_{pxp} = \frac{1}{s-1}\sum_{l=1}^{s}\left(\bar{X}_l - \bar{X}\right)\left(\bar{X}_l - \bar{X}\right)' \tag{3.3.18}$$

A generic element of this matrix is:

$$s^B_{ii} = \frac{1}{s-1}\sum_{l=1}^{s}\left(\bar{x}_{il} - \bar{x}_i\right)^2 \tag{3.3.19}$$

$$s^B_{ij} = \frac{1}{s-1}\sum_{l=1}^{s}\left(\bar{x}_{il} - \bar{x}_i\right)\left(\bar{x}_{jl} - \bar{x}_i\right)' \tag{3.3.20}$$

with $i, j = 1,2,\ldots, p$.

This means that the interest will be to maximize the rate given in the expression (3.3.17), that is:.

$$\max_{e \neq 0} \frac{\mathbf{e}' \mathbf{S}_B \mathbf{e}}{\mathbf{e}' \mathbf{S}_W \mathbf{e}} \tag{3.3.21}$$

As larger is the rate (3.3.17), will be stronger the indication of larger variability between groups of those within of the groups. One change in the coefficients of the lineal combinations, i.e.,

$$\mathbf{e}_i' = \left(e_{i1}, \quad e_{i2}, \quad \ldots, \quad e_{ip}\right) \tag{3.3.22}$$

with $i = 1,2,\ldots,k$, will change the values of y_{ij} and, therefore, it will produce different values for equation (3.3.17). To maximize the value of this rate will allow to see, in the best possible way, the difference between the groups, and it is for this that the problem of the FDA will be to find the coefficients that maximize (3.3.17), which is an optimization problem, purely mathematical. To solve it is necessary to select $\mathbf{e}$ in such way that equation (3.3.17) be maximum, therefore, the idea here is to work with this expression, presenting the optimization in the form of the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

3.3.2.1 Reduction of dimensionality

The matrix $\mathbf{S}_W^{-1}\mathbf{S}_B$ has k eigenvalues different from zero, where $k = \min(s-1,\ p)$; this constant k determines the number of discriminant functions. Therefore, the space of the new discriminant variables has dimension *k*, smaller than the space of the original variables, of dimension *p*.

The associated eigenvectors, $l_1, l_2, \ldots, l_k$, to the k eigenvalues different from zero, standardized such that $l_i' \mathbf{S}_W l_i = 1$, $\forall i = 1,2,..,k$, will impose a condition that will allow to obtain an unique solution for this problem. It can be proven analytically that the vector of coefficients that maximizes the equation (3.3.17) it is the first eigenvector of $\mathbf{S}_W^{-1}\mathbf{S}_B$. This means that the best lineal combination $y_i = \mathbf{e}_i' X$ of the original variables that exalts the difference among the groups has as coefficients the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

3.3.2.2 *Simplification of the Fisher's discriminant function for the case of two groups*

The objective of Fisher's Discriminant Analysis, namely to maximize the separation between the groups, is reduced, in the case of two groups, to maximize:

$$\phi = \frac{|\bar{y}_1 - \bar{y}_2|}{e' \mathbf{S}_{pooled} e} \tag{3.3.23}$$

The coefficients, e, of the lineal combination will be chosen so that they maximize the reason between the square of the distance between the means in the new variables y and the considered estimate of the covariance matrix within of the groups. The development of the expression (3.3.23) it is:

$$\phi^2 = \frac{(|\bar{y}_1 - \bar{y}_2|)^2}{e' \mathbf{S}^{-1}_{pooled} e} = \frac{(e'\bar{\mathbf{x}}_1 - e'\bar{\mathbf{x}}_2)^2}{e' \mathbf{S}^{-1}_{pooled} e} = \frac{(e'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{e' \mathbf{S}^{-1}_{pooled} e} \tag{3.3.24}$$

Using the Maximization Lemma[7] in quadratic forms[8] arrived, in (3.3.24), to

$$\max_e \frac{(e'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{e' \mathbf{S}^{-1}_{pooled} e} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1}_{pooled} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = D^2,$$

i.e., the maximum is attained when $e = \mathbf{S}^{-1}_{pooled}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, being D^2 the square of the distance between both sample means, measures in units of standard deviation (Mahalanobis distance[9]). The Fisher's Discriminant Function is then algebraically explained as $Y = e' X = (\bar{x}_1 - \bar{x}_2)' \mathbf{S}^{-1}_{pooled} X$, for the case of two variables.

3.3.3 Using Fisher's Discriminants to Classify for several groups

Fisher's discriminants were derived for the purpose of obtaining a low-dimensional representation of the data that separates the populations as much as possible. Although they were derived from separatory

---

7 See the Maximization Lemma in Appendix 2, section A2.1

8 The proof of Maximization Lemma of quadratic forms can be seen in Appendix 2, section A2.1.

considerations, the discriminants also provide the basis for a classification rule.

The Fisher's discriminant procedure can be described (Johnson and Wichern, 1992) as:

Allocated the new observation x_0 to group m if:

$$\sum_{j=1}^{s}\left(\hat{y}_j - \bar{y}_{mj}\right)^2 = \sum_{j=1}^{s}\left[\hat{l}_j'(x_0 - \bar{x}_m)\right]^2 \leq \sum_{j=1}^{s}\left[\hat{l}_j'(x_0 - \bar{x}_i)\right]^2 , \ \forall i \neq m \qquad (3.3.25)$$

where $\hat{l}_j'$ is the jth eigenvector of matrix $\mathbf{S}_W^{-1}\mathbf{S}_B$. In other words, it means to allocate the new observation in the group such that the distance between the coordinates of this observation and the mean of this group be smaller than the distance between this observation and the mean of any other group.

Hypothetical representation of classification procedure

Let us suppose that we have three groups, s=3; identified as A, B and C; and two discriminant axes. The averages of the groups are represented in Figure 6. The classification procedure calculates the distance from the observation, to be classified, to each one of the groups and it puts in the nearest group. In this case X_0 would be classified in group B.

[9] More information about Mahalanobis distance can be seen in Cónsul, 2000, pag.34.

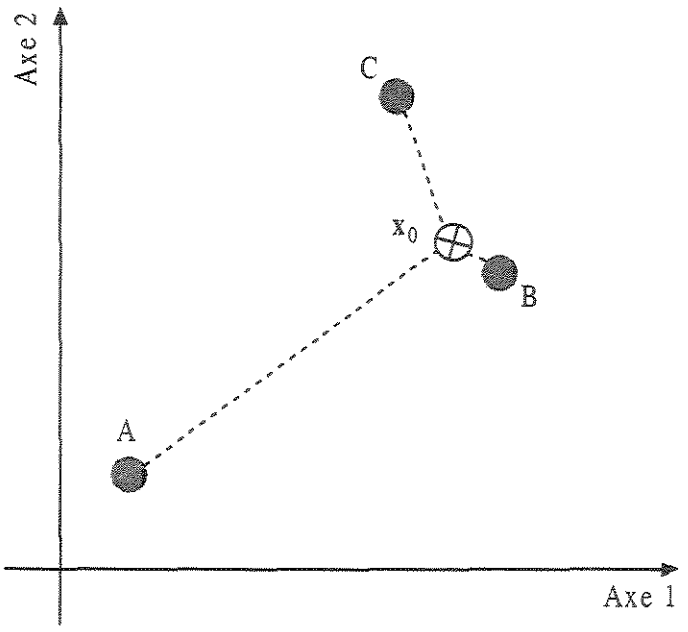


Figure 6. Hypothetical representation of Fisher's classification procedure.

3.3.3.1 *Rates of misclassification by group*

A good classification procedure should result in few misclassifications. In other words, the rates of misclassification should be small. This misclassification rate can be calculated for group i as:

$$\text{rate}(\text{group } i) = \frac{\text{total of misclassificated observations in group } i}{n_i}, \ \forall i = 1,2,...,s \quad (3.3.26)$$

and the percent of misclassification in group i is:

$$\%\ \text{rate}(\text{group } i) = \text{rate}(\text{group } i) \times 100\%, \ \forall i = 1,2,...,s \quad (3.3.27)$$

3.3.3.2 *Mean rate of misclassification*

On the other hand, it is possible to calculate the rate mean of misclassification as:

$$\text{mean rate} = \frac{\sum_{i=1}^{s} \text{rate}(\text{group } i)}{s} \quad (3.3.28)$$

And the mean percentage of misclassification is equal to:

$$\%\ \text{mean rate} = \text{mean rate} \times 100\% \quad (3.3.29)$$

3.3.3.3 *Using Fisher's Discriminants to Classify for two groups*

Fisher's solution to the separation problem can also be used to classify new observations (Johnson and Wichern, 1992). An allocation rule based on Fisher's Discriminant Function is:

Allocate x_0 to group 1 if:

$$y_0 = (\bar{x}_1 - \bar{x}_2)' \mathbf{S}_{pooled}^{-1} x_0 \geq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' \mathbf{S}_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2) \qquad (3.3.30)$$

and allocate x_0 to group 2 if:

$$y_0 = (\bar{x}_1 - \bar{x}_2)' \mathbf{S}_{pooled}^{-1} x_0 < \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' \mathbf{S}_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2) \qquad (3.3.30)$$

The rates of misclassification by group and the mean rate of misclassification can be calculated, for two groups (s=2), using equation (3.3.26) and equation (3.3.28) respectively.

3.4 Conclusions

In this chapter the necessary theoretical base to develop PCA and FDA was shown. In the next chapter we will make the connection between theory and practice to take the analysis ahead with data of an industrial continuous process.

Chapter 4: Proposed procedures

4.1 Introduction

In this chapter we will build data sets adapted to work with PCA and FDA. Also, the procedures necessary for to develop the techniques proposed will be approached step the step. This chapter is explanatory, preparing the conditions for the application of PCA and FDA. The results of the techniques will be seen in Chapter 5.

4.2 Proposed Approach

To develop the methods proposed in the previous Chapter we need to follow some necessary steps. First, it will be specified an appropriate data set, this is, a data set that represents a process under normal conditions of operation. Second, the procedure of autoscaling, before employing the dimensionality reduction techniques. Then, the PCA model is building and the process of generation of faults, and the results for each studied fault is explained. Later, FDA is applied and the behavior of the data in the discriminant space and the classification procedure is developed. Finally a insight about comparation of PCA and FDA is given.

4.3 Validation Procedure for PCA

Historical data was collected from a hydrogenation of phenol to cyclohexanol reactor, shown schematically in Figure 7 (Santana, 1999).

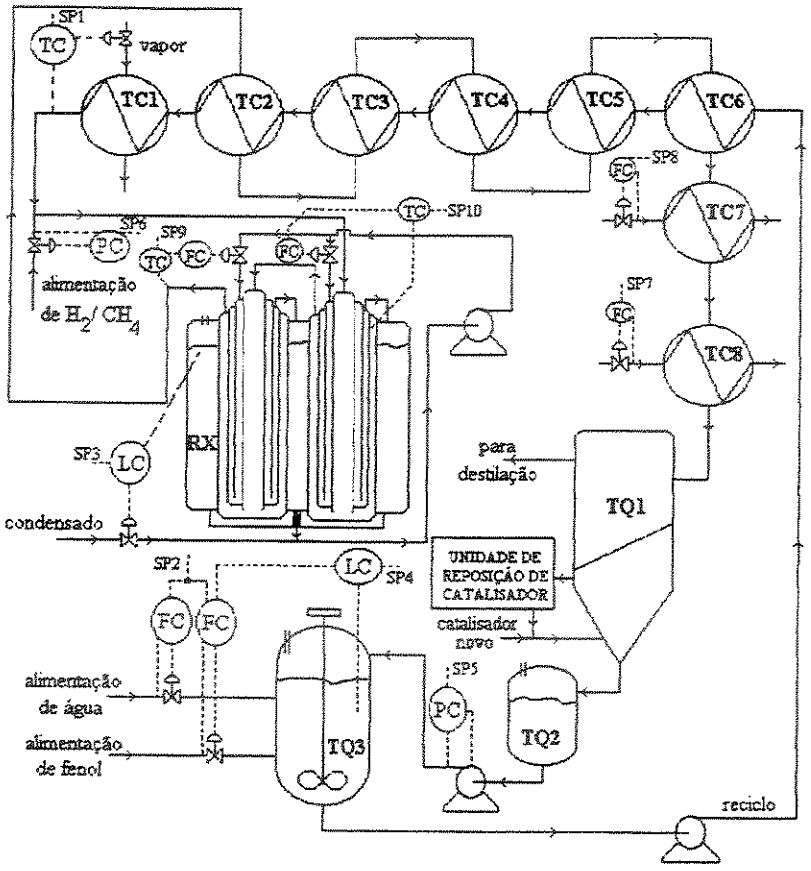


Figure 7. A typical unit of cyclohexanol production.

Figure 7 shows a typical unit of cyclohexanol production. This is formed by storage tanks (TQ2) and mixture of reagents and catalyst (TQ3), by a tank of separation of the products of the reaction of the catalyst (TQ1), by several heat exchanges (TC1 to TC8) and a reactor (RX), which is formed by eight tubular modules immersed in a boiler. The control of the tanks is made controling the operation of the tanks TQ3 and TQ1 basically. In the first level, the proportion of the reagents and the pressure of feeding of the catalyst are maintained under control, while in the last the recycled amount is controlled. There are basically two involved reagents: phenol and hydrogen. Additionally, the reactor is fed by a stream of water and another of recycle with catalyst. The fed water purpose is to move the balance of undesirable reactions and also to improve the thermal change in the reactor. The catalyst is separated from the reaction products in the lung tank TQ1, schematized in Figure 7, and correspondent for

regeneration, being a portion returned to the process. The regenerated catalyst is mixed with a new feed of catalyst. The concentration of residual phenol in tank TQ1 is measured. If this is above the operation specification, the whole current originating from the reactor is recycled, being interrupted the injection of new reagents and the reactor is just used to consume the whole phenol. If there does not exist residual phenol, the recycle of the liquid stream is not made, and just a small portion of this, impregnated in the catalyst, is returned to the process.

The reaction of hidrogenation of phenol is exothermic, and, depending on the temperature of operation of the reactor and of the used catalyst, several products can be formed as acetones or cyclic alcohols, aromatic hydrocarbons and cyclics. The cyclohexanol reactor is constituted by a number of tubular modules immersed in a boiler, being each one of them formed by concentric tubes. In these, there is passage of the reactant mixture as well as of the coolant so that the reaction temperature along the reactor is controlled. Located temperature measurements exist in two different points in each tubular module, and these can suffer problems of incrustation which lead to measurement errors in a significant level. The flow of reactant in the cyclohexanol reactor flows from one tubular module to another, and the first six are similar to each other and they are constituted, each one, of four concentric tubes.

4.3.1 Real Chemical Process Simulation

37 process variables were monitored, from the process described before, with time intervals of 15 minutes for a total of 158 observations. When these data were explored it was possible to note that there was no guarantee that the process was under statistical control. This can be a relatively usual situation when dealing with industrial data analysis since it may be considered that it is not possible to access a data set with a guarantee of coming from a process under "good operating conditions". In this case, it was decided not to use these data directly to construct the

PCA model.

In view of the necessity of the construction of a data set without special-cause variability, a suitable approach was to use random generation data, in an appropriate form, to simulate a real process. The principal problem was the maintenance of a consistent correlation structure inside the new data set, knowing the importance of this aspect for the application of principal components technique. In order to preserve correlation relations between variables, similar to the existing in real industrial process, the idea was to generate the same number of multinormal variables as the monitored process variables (in this case study, 37), using the mean vector and the covariance matrix from the real process data, with the same number of observations (in this case, 158). This data set, X, was used to construct the PCA model.

With the goal to study the potential of the technique, five appropriate data sets, with 37 variables and 100 observations each, called X*new*, were constructed for the fault detection and diagnose step.

Other situations of interest were also simulated. They were situations with two and three simultaneous faults and the case of constant reading in one of the measurement equipments was also simulated and analyzed inserting a constant variable.

4.3.2 Construction of the PCA model

A calculation routine in FORTRAN was developed specifically to achieve the goals of this PCA routine. The implementation in FORTRAN has as the main objective to leave the routines accessible for practical applications in industrial environment with freedom of software interface at lower costs.

4.3.2.1 *Choice of the number of principal components using the parallel analysis method*

The first SCREE plot, for the data set X, is showed in Figure 9 (with the symbol ●). The break point suggests that only six principal components are enough to describe the process.

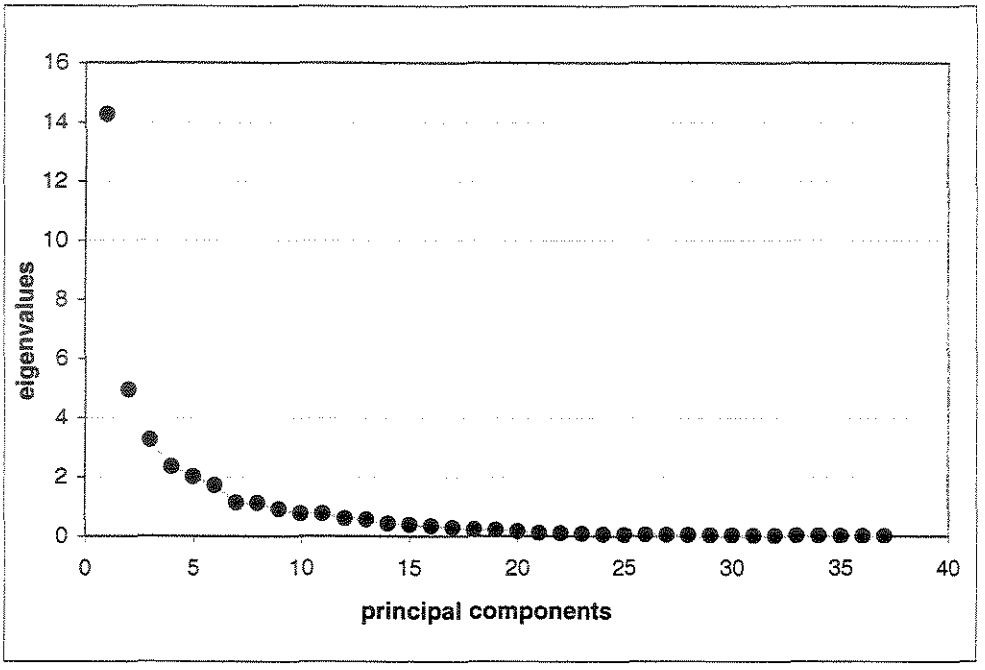


Figure 8. SCREE plot.

At the same time a second SCREE plot for the uncorrelated random data set was plotted, Figure 9 (with the symbol ✖). To make this graph, 37 independent variables with normal distribution were generated randomly, using the mean and standard deviation of the original variables. The eigenvalues of the covariance matrix for these uncorrelated random data set are the values in the second graph. The intersection of this line and the SCREE plot for the original data indicates that the point separating the retained and deleted principal components is also in the principal component six.

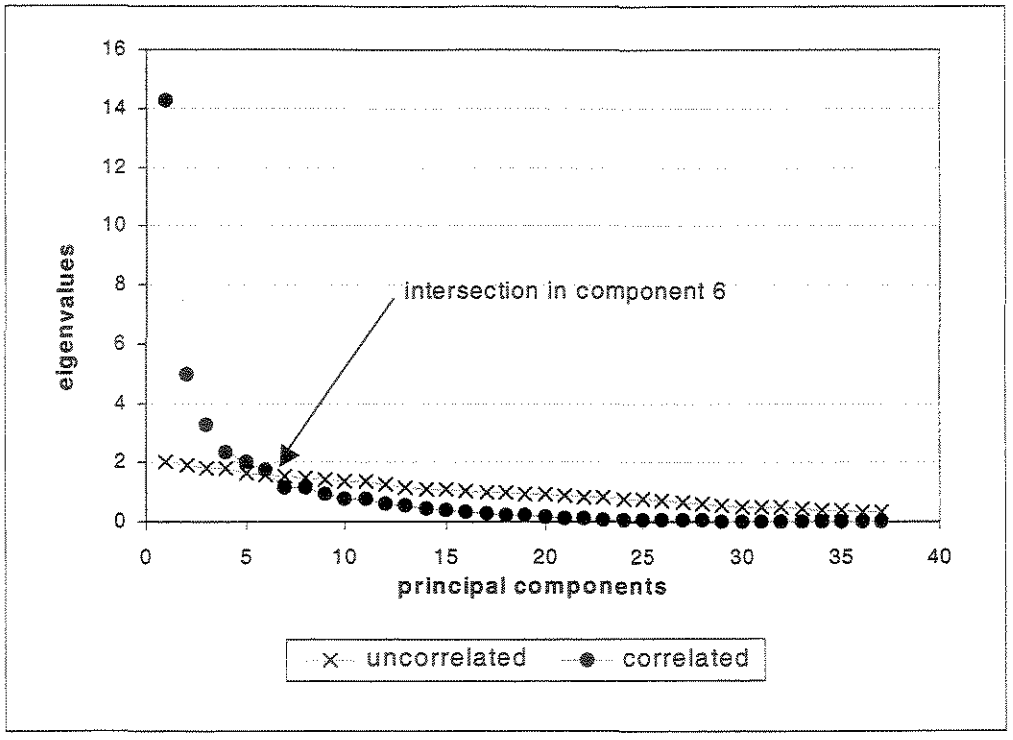


Figure 9. SCREE plot and parallel analysis.

The eigenvalues and percent of variance, for the first ten principal components, are given in Table 3. It may be seen that the first six principal components explain 77.2% of the total variability of matrix X.

Table 3. Eigenvalues and percent of explained variance for the first ten principal components.

Principal component	Eigenvalue of S	% de variance	
		This PC	Cumulative
1	14,27	38,6	38,6
2	4,94	13,3	51,9
3	3,26	8,8	60,7
4	2,35	6,4	67,1
5	2,01	5,4	72,5
6	1,72	4,7	77,2
7	1,13	3,0	80,2
8	1,12	3,1	83,3
9	0,92	2,4	85,7
10	0,78	2,2	87,9

For this model, the calculated values for Q_{crit} and T^2_{crit} limits are 14.96 and

13.42, respectively.

4.4 Validation Procedure FDA

To validate the method of FDA we will focus ourselves in 3 points that we considered fundamental: the construction of the data for the application of the method, the analysis of the behaviors of the data in the discriminant space and the classification of the observations, with the calculations of the misclassification rates.

It is good to observe that, following the pointed observations in the Note 4, page 24, we opted to continue working with our data matrix standardized.

4.4.1 Constructing the data set for FDA

From the five X_{new} data sets used for PCA, each one containing a different group of disturbed observations, we built a new group of data to be used in FDA. In this matrix, that we will call X_F, the passages containing the faults 1, 2, 3, 4 and 5, were used. The data matrix X_F will be standardized for each variable to have mean zero and variance one.

Starting from the recommendation of Chiang et al. 2000, who said that FDA can be used to detect faults by including the class of data collected during normal process operation, a group without disturbances was included in the matriz X_F as number 6.

To have a better idea we will say that the matrix X_F has this form:

$$X_{F\,(mxp+1)} = [grupo,\quad X_1,\quad \ldots,\quad X_{37}] \qquad (3.2.1.1)$$

Being, in this case, m=117 and p+1=38. This matrix X_F can be described more explicitly as:

$$\mathbf{X}_{F\,(mxp+1)} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n_1,1} & \cdots & x_{n_1,p} \\ 2 & x_{n_1+1,1} & \cdots & x_{n_1+1,p} \\ \vdots & \vdots & \cdots & \vdots \\ 2 & x_{n_1+n_2,1} & \cdots & x_{n_1+n_2,p} \\ \vdots & x_{i1} & \cdots & x_{ip} \\ 6 & x_{\sum_{i=1}^{5} n_i+1,1} & \cdots & x_{\sum_{i=1}^{5} n_i+1,p} \\ \vdots & \vdots & \cdots & \vdots \\ 6 & x_{m1} & \cdots & x_{mp} \end{bmatrix} \qquad (3.2.1.2)$$

where n_1, n_2, ..., n_6 represent the number of observations, or samples, for the groups 1, 2, up to 6, respectively.

4.5 Conclusions

The conditions for the application of the proposed techniques were created. With the proposed tools it is possible to identify abnormal state variables values which may happen either due to sensor faults or by undesired or unexpected operatory conditions.

Chapter 5: Results

5.1 Results for PCA

5.1.1 Disturbing the data with single faults

A data set with 37 variables and 100 samples was generated, using the same random procedure used for generated X. After that, it was standardized, using the same procedure applied to the X matrix. Five variables were disturbed appropriately (see Table 4) in order to allow the performance test of the proposed technique. Each disturbance generates a different X_{new} matrix. Therefore, it was worked with five disturbed X_{new}, each one corresponding to a fault type.

Table 4. Single faults description.

Fault ID	Description of the disturbance
1	The temperature in the boiler (°C) was disturbed (increase) in samples 47 to 69. (Variable #31)
2	The pressure in the boiler (kgf/cm^2) was perturbed (decrease) in samples 81 to 100. (Variable #32)
3	The temperature in the bottom of tube 1 (°C) was perturbed (increase) in samples 1 to 15. (Variable #9)
4	The temperature in the bottom of tube number 3 (°C) was disturbed (increase) in samples 20 to 42. (Variable #13)
5	The temperature in the top of tube number 6 (°C) was disturbed (increase) in samples 25 to 35. (Variable #18)

5.1.2 Results for single faults

The horizontal line in Q residual and T^2 plots represents the 95% confidence limit. Any point above this line is considered evidence of a process fault. The analysis for each fault was as follows:

Fault #1: Variable number 31, it is the temperature in the boiler (°C). The temperature was disturbed in samples 47 to 69.

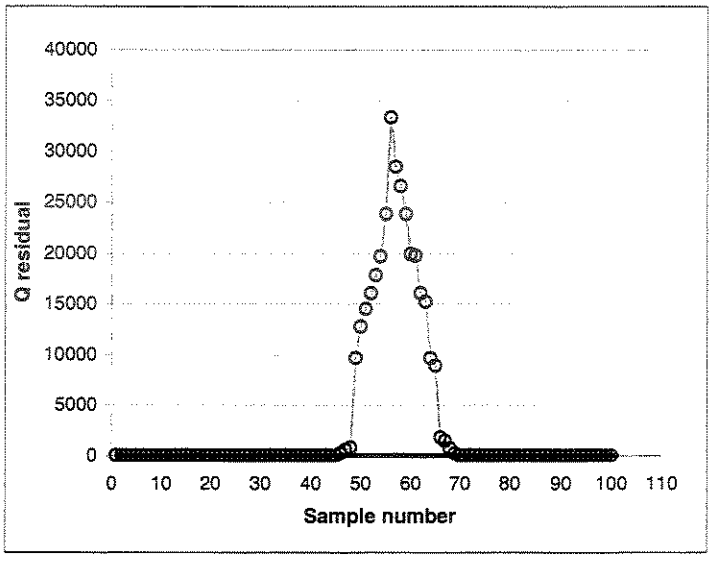


Figure 10. Q residual plot by sample for fault #1.

The disturbance at samples 47-69 is clearly seen in the Q residual plot, see Figure 10. There is an abnormal situation with these observations since they shifted outside the normal operation space, defined by the original X matrix.

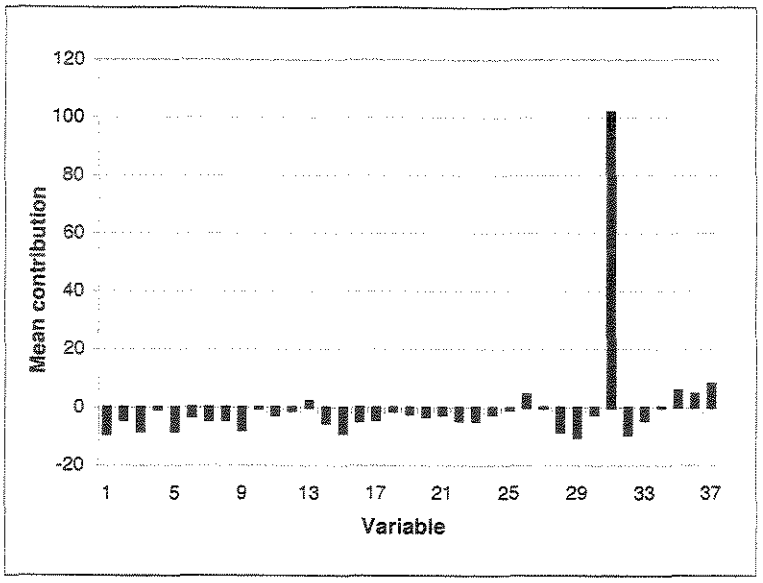


Figure 11. Mean contribution by variable for fault #1.

The mean contribution plot, Figure 11, helps to see who causes the faults. In this case, variable number 31 appears with a larger contribution to the residual matrix E. In fact, this was the perturbed variable.

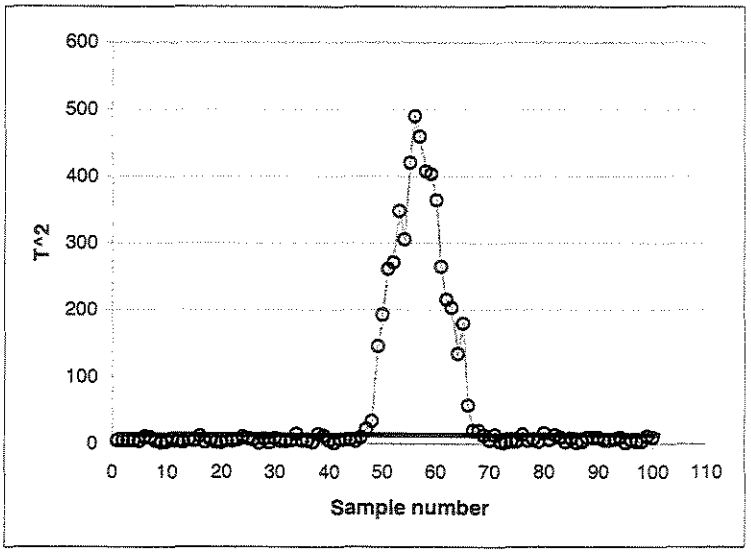


Figure 12. T^2 plot by sample for fault #1.

In the T2 plot, in Figure 12, the same set of samples, 47-69, are outside the limit that define the ellipsoid on the space in which the operating points are normally expected to happen.

The scatter plot, Figure 13, with the scores of principal components 1 and 2, for X(O) and *Xnew* (Δ), also shows that fault points have a different

behavior.

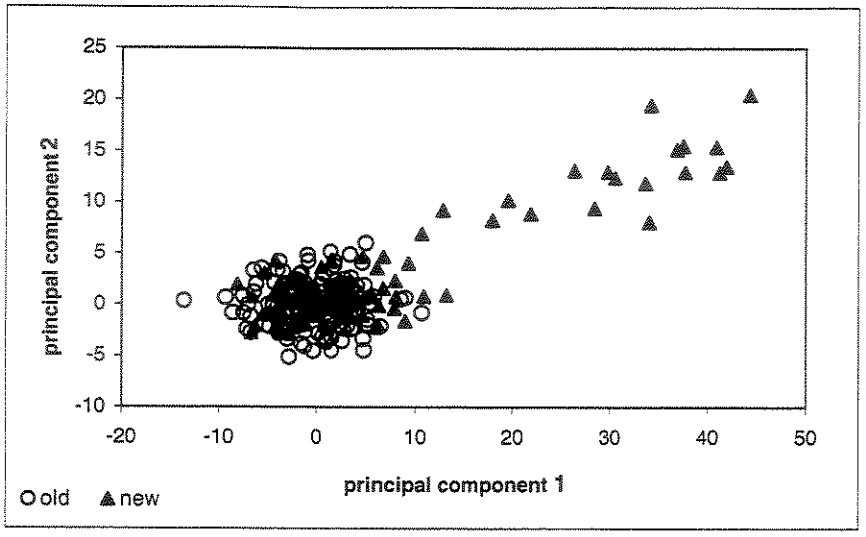


Figure 13. Scores for the two first principal components for fault #1.

Fault #2: Variable number 32, it is the pressure in the boiler (kgf/cm^2). The pressure was perturbed in samples 81 to 100.

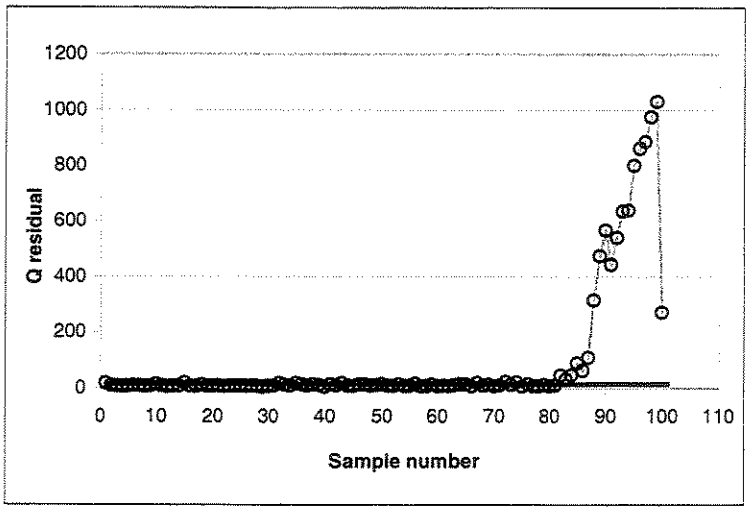


Figure 14. Q residual plot by sample for fault #2.

The disturbance in samples 81-100 is obvious in the residual Q plot, Figure 14. There are unusual variabilities in these samples since they also shifted outside the normal operation space, defined by the original X matrix.

The mean contribution plot, Figure 15, is very useful to see that variable number 32 has the largest contribution to the residual matrix E. Also, this

is a negative contribution since a pressure drop was applied to this variable.

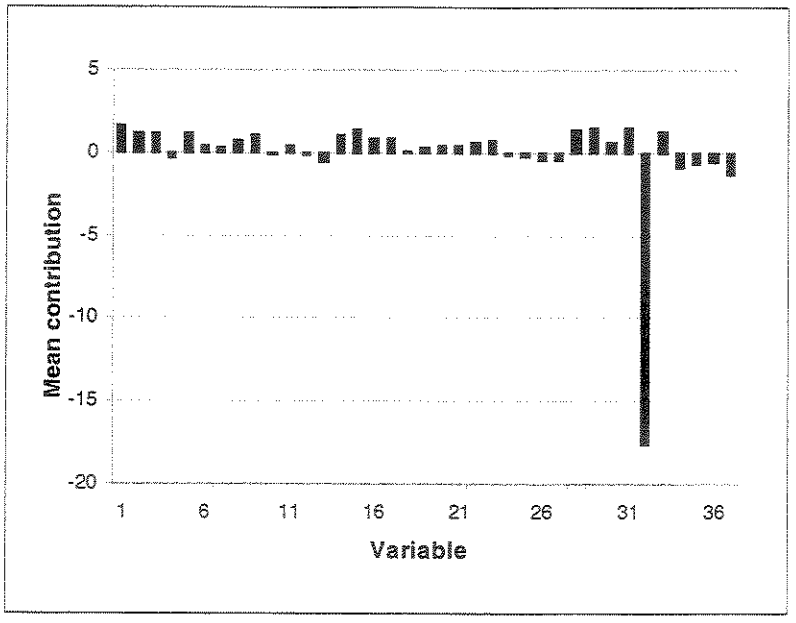


Figure 15. Mean contribution by variable for fault #2.

In the T^2 plot, Figure 16, only a few samples such as 89, 98, 99 are outside the limit 13.42 that defines the ellipsoid on the space in which the operating point normally is expected to happen.

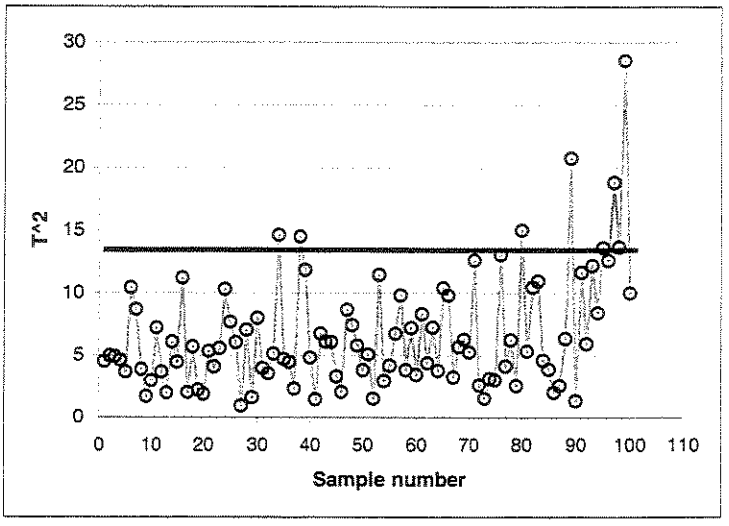


Figure 16. T^2 plot by sample for fault #2.

In the scatter plot with the scores, of principal components 1 and 2, for X(O) and X*new*(Δ), the observations a little outside the set of points are the

ones that had large values for T^2, see Figure 17.

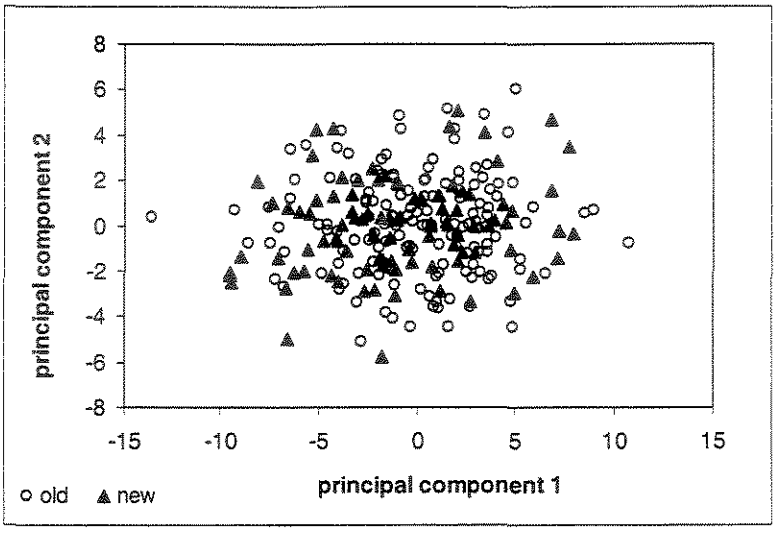


Figure 17. Scores for the two first principal components for fault #2.

It is interesting to point out that the analysis of the scatter plot alone may not be so useful when compared to the Q, mean contributions and T^2 plots. In fact it is proposed in this work the simultaneous use by such statistical representations in order to have full information on the system state.

Fault #3: Variable number 9, it is the temperature in the bottom of tube 1 (°C). The temperature was perturbed in samples 1 to 15.

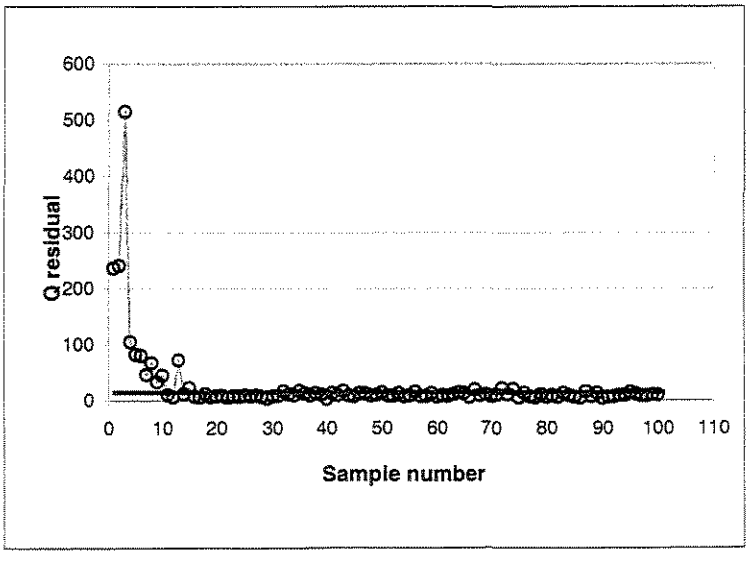


Figure 18. Q residual plot by sample for fault #3.

The disturbances at samples 1 to 15 are seen in the Q residual plot, Figure 18. An abnormal situation is happening with these observations, as it can be observed in the illustration of Figure 19. This plot helps to identify which variable is the cause of such behavior.

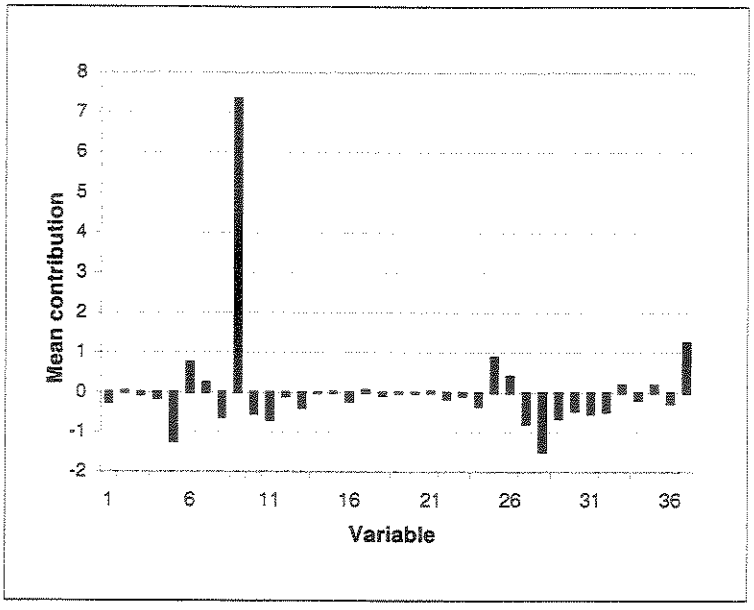


Figure 19. Mean contribution by variable for fault #3.

Variable number 9, just as it was already expected, appears as the largest mean contribution to the data variability, in Figure 19.

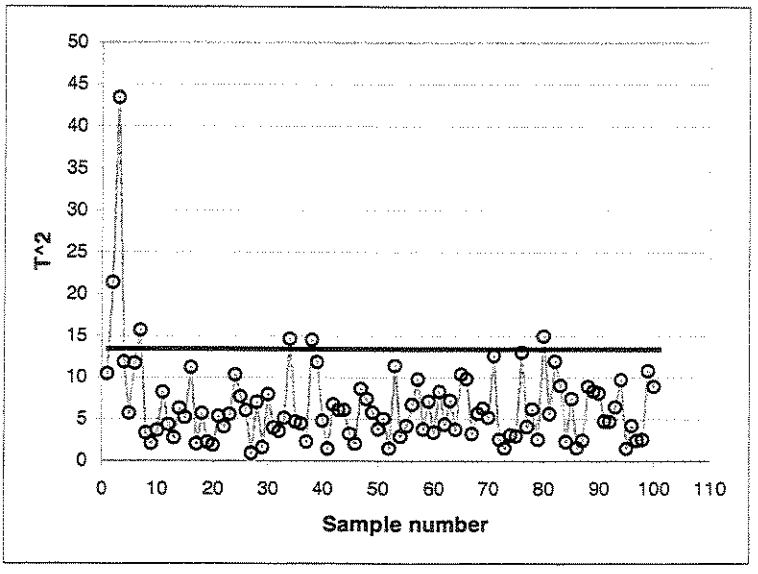


Figure 20. T^2 plot by sample for fault #3.

In the graph of T^2, Figure 20, there does not appear a defined pattern; only some few observations pass the established reliability limit. In this case,

the graph of the scores for the two first principal components, Figure 21, does not allow one to see the occurrence of the fault. This behavior was already expected due to the low values of T^2 in Figure 20.

The scores plot in the first two principal components is completely random, without supplying any information about abnormal behaviors, Figure 21.

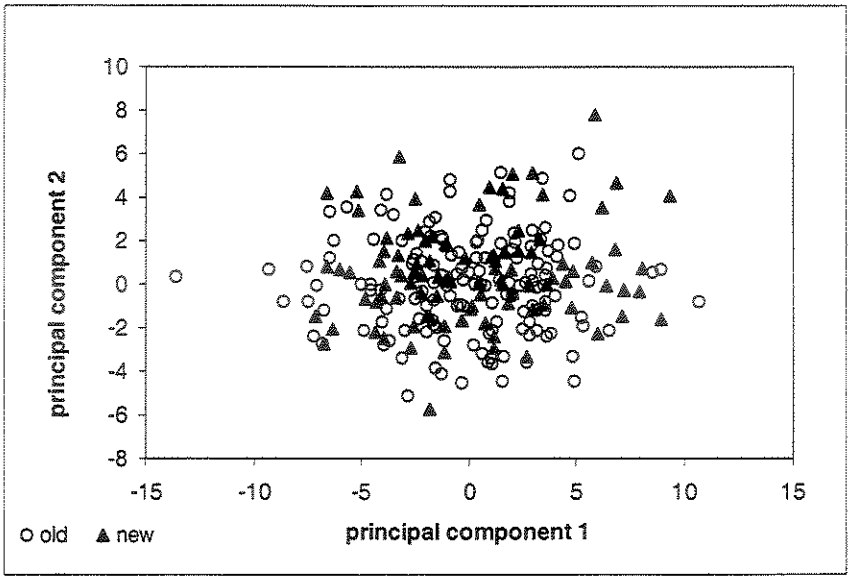


Figure 21. Scores for the two first principal components for fault #3.

Fault #4: Variable number 13, is temperature in the bottom of tube number 3 (°C). The temperature was increased in samples of 20 to 42.

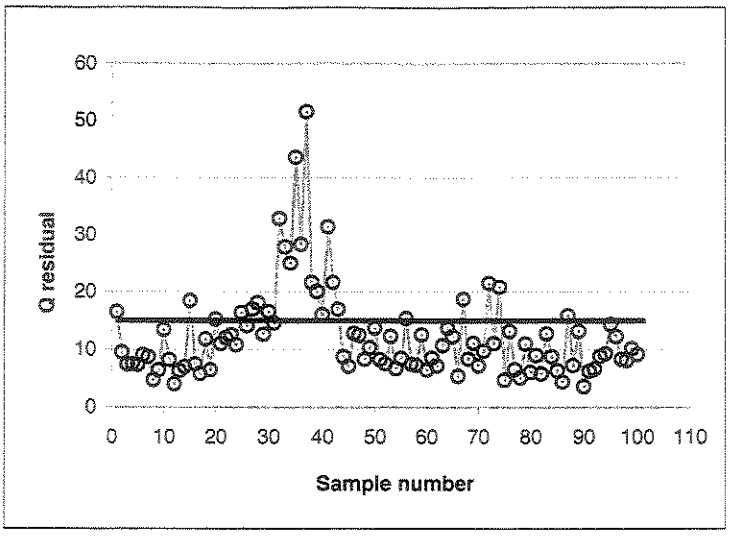


Figure 22. Q residual plot by sample for fault #4.

In the illustration of Figure 22, an abnormal behavior is observed in the samples starting approximately from 25. In fact, the induced disturbance in samples 20 to 25, was not so relevant, but even so it was possible to find out the fault.

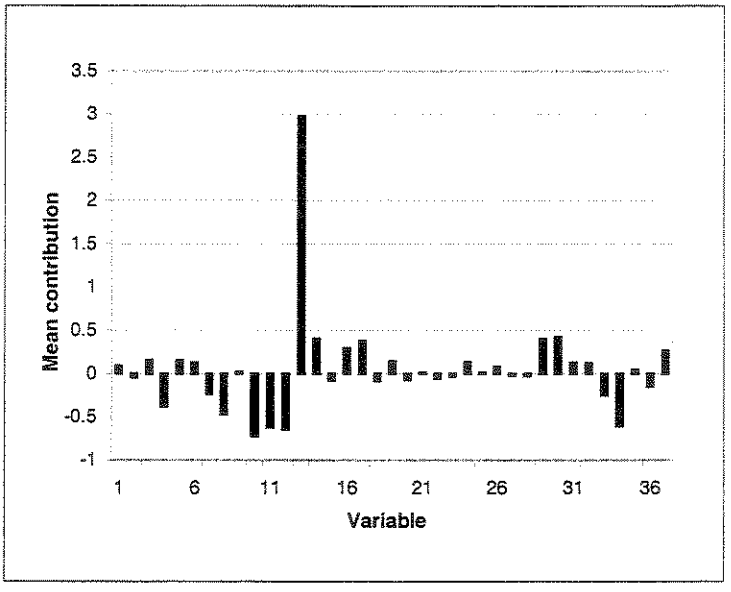


Figure 23. Mean contribution by variable for fault #4.

Variable 13 is clearly the cause of the problem as can be observed in the mean contributions plot, Figure 23.

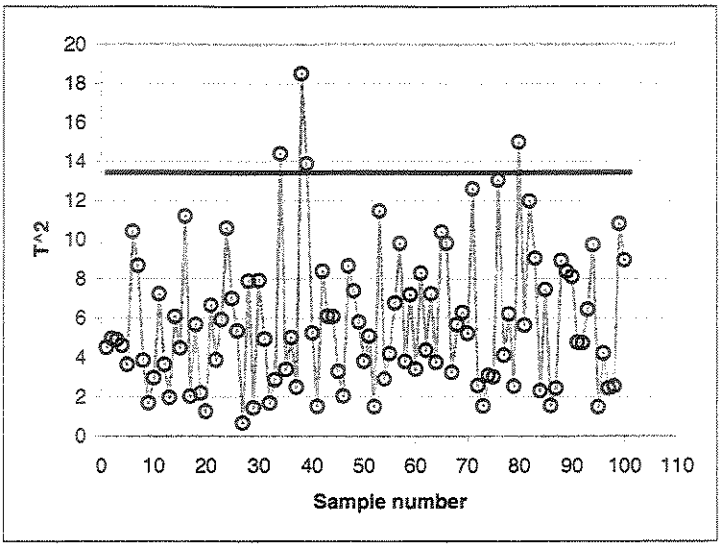


Figure 24. T^2 plot by sample for fault #4.

In the graph of T^2, Figure 24, there does not appear a defined pattern either, with some few observations passing of the established limit. In this case, in Figure 25, it is also seen the occurrence the fault. This behavior is again related with the values of T^2 in the Figure 24.

The scores plot in the first two principal components, Figure 25, is completely random, without showing any evidence of abnormal behavior either. The other plots show the same behavior, and they can be seen in Appendix 1, section A1.3 (Figure 47 and Figure 48).

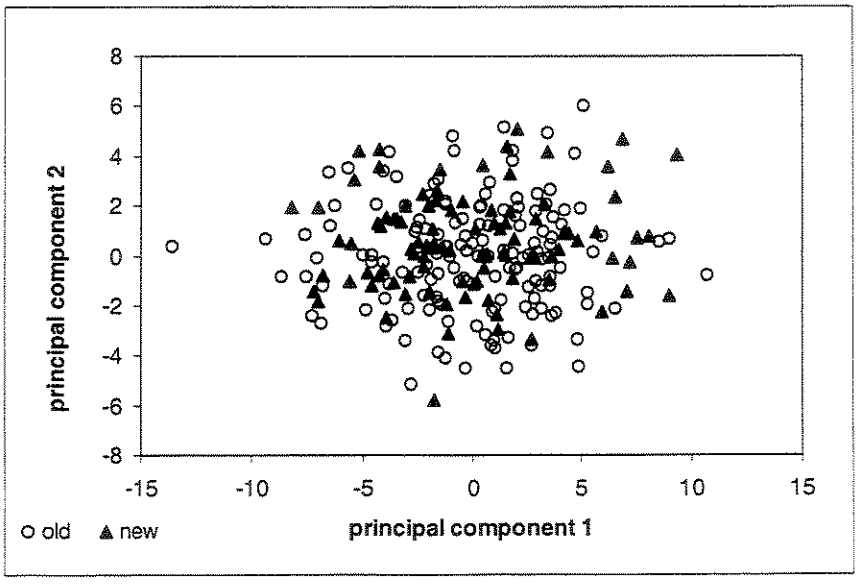


Figure 25. Scores for the two first principal components for fault #4.

Fault #5: Variable number 18, it is the temperature in the top of tube number 6 (°C). The temperature was disturbed in samples 25 to 35.

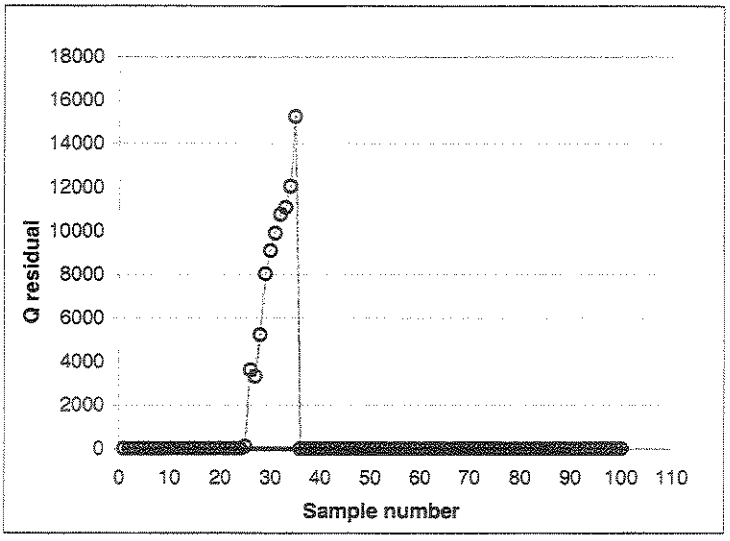


Figure 26. Q residual plot by sample for fault #5.

In this case, Figure 26, the presence of a differentiated pattern of abnormal behavior, in samples 25 to 35, is observed in residual Q plot. When the mean contribution is analyzed by variable for fault #5, Figure 27, it was noted that this behavior is due precisely to the variable 18 that is the one which causes the largest influence in the residual matrix.

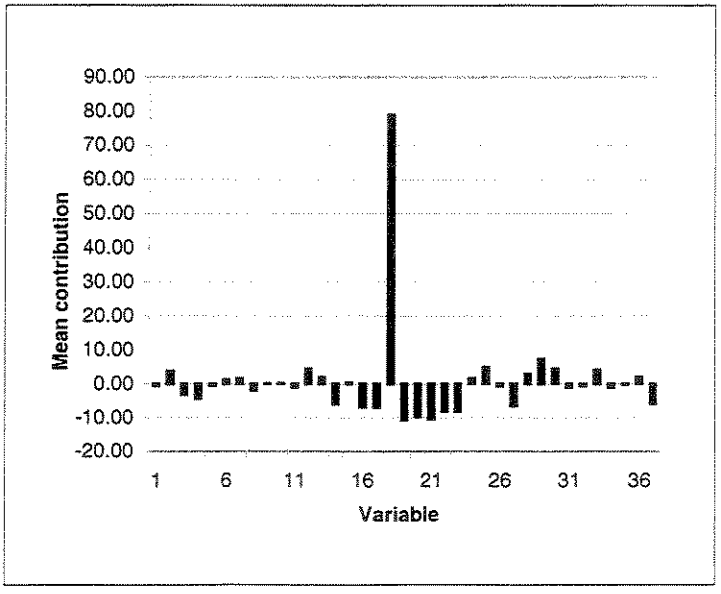


Figure 27. Mean contribution by variable for fault #5.

In the graph of T^2, Figure 28, an unusual behavior can be observed as well, similar to the one observed in the graph of Q. Then, similarly to case of the fault #1, it can be seen the presence of a disturbance in the process that produces large values as much of Q as of T^2. In this case of fault #5, idem to fault #1, this behavior can be identified in the graphs for these two statistics.

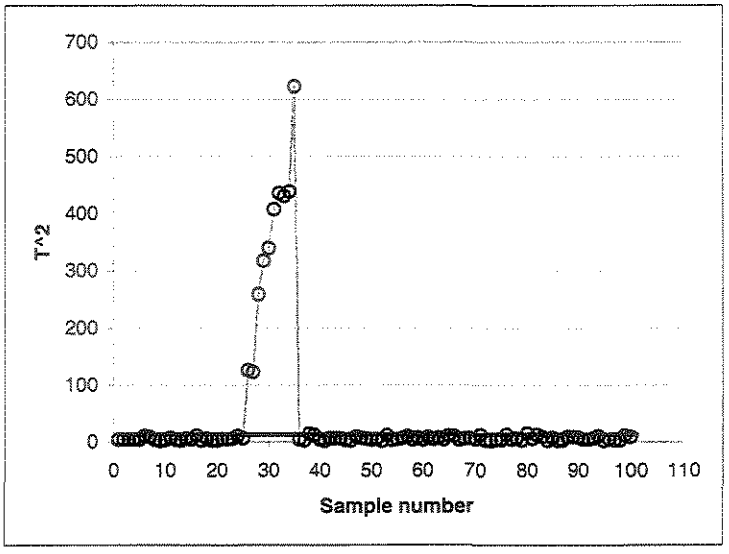


Figure 28. T^2 plot by sample for fault #5.

When this type of fault happens, it can be seen, also in a clear way, in the graph of the scores of the first two principal components, Figure 29. Here it is possible to observe the samples that suffered sharp disturbance.

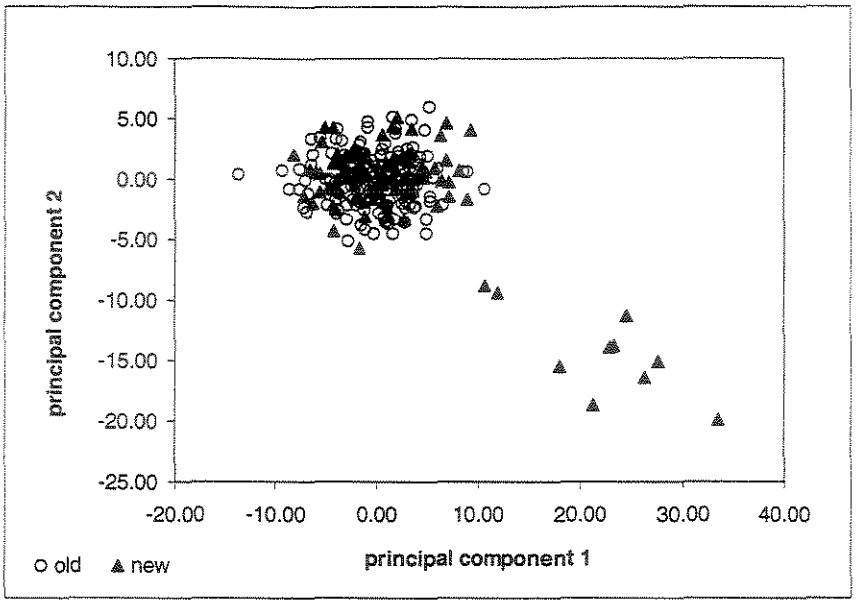


Figure 29. Scores for the two first principal components for fault #5.

5.1.3 Disturbing the data with two simultaneous faults

The description for the simulation of two simultaneous faults is shown in Table 5.

Table 5. Description of the two simultaneous faults.

Fault ID	Description of the disturbance
1 e 2	The temperature in the boiler (°C) was disturbed (increase) in samples 47 to 69 (Variable #31) and the pressure in the boiler (kgf/cm^2) was perturbed (decrease) in samples 1 to 20 (Variable #32)

The objective here is to see if it is possible to detect this couple faults with the use of statistics Q, T^2 and the other auxiliary plots.

5.1.4 Results for two simultaneous faults

In the graph of the Q residues, Figure 30, it can be seen that both fault #1 and fault #2 cross the limit $Q\alpha$, indicating that approximately in samples 47 to 69 and 1 to 20 there appears a disturbance in the process.

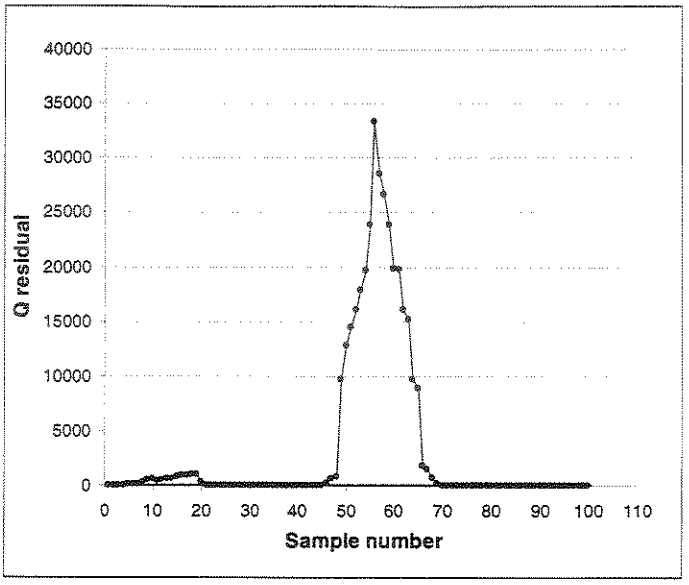


Figure 30. Q residual plot by sample for fault #1 and fault #2.

In the case of statistics T^2, it only captures the fault #1, happening in the samples from 47 to 69.

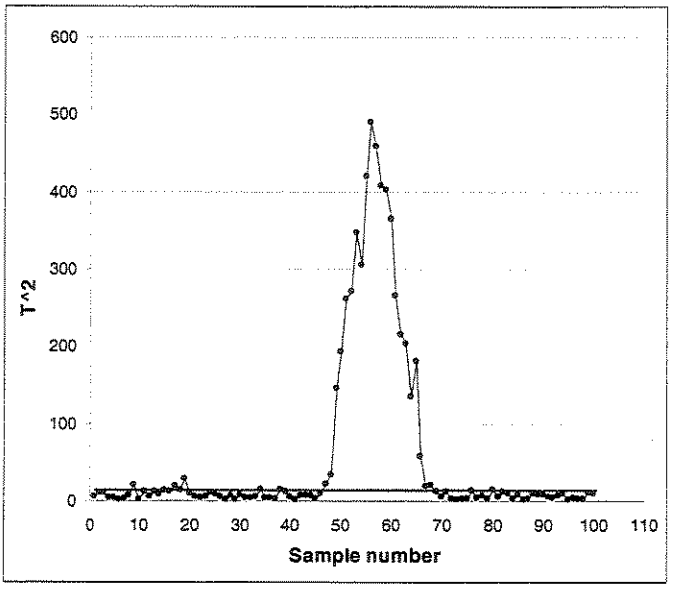


Figure 31. T^2 plot by sample for fault #1 and fault #2.

The mean contributions plot for this double fault is depicted in Figure 32. It may be observed that the two variables with larger contributions are variables 31 and 32, in the same magnitude order and direction in the which they were simulated.

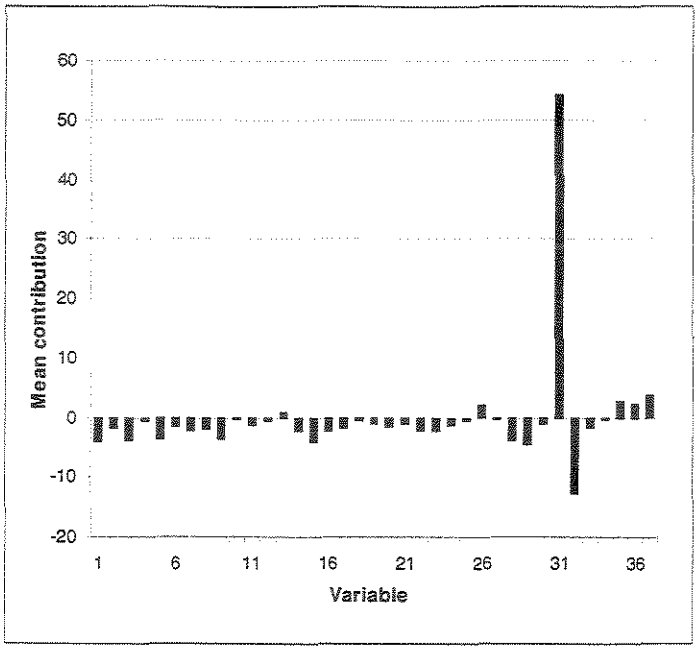


Figure 32. Mean contribution by variable for fault #1 and fault #2.

The behavior of the observations, in the first two principal components, can be seen in Figure 33. In this case, the plot informs about the existence of a strange behavior in the data.

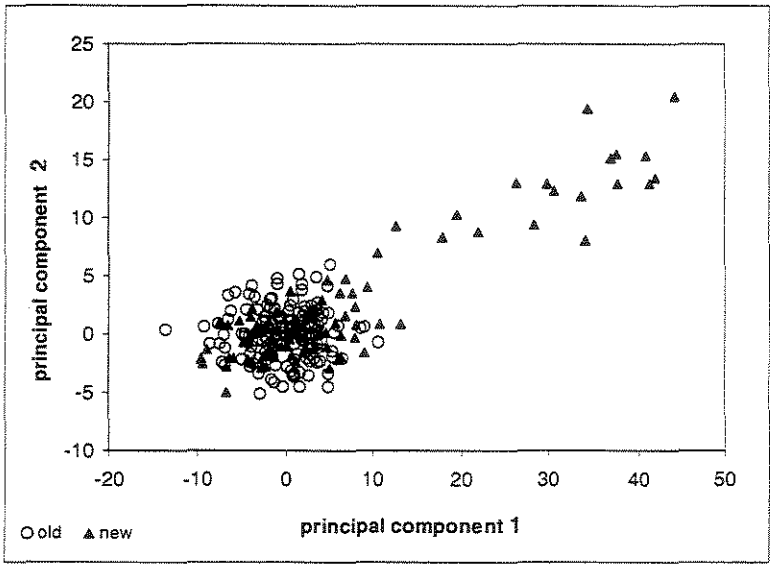


Figure 33. Scores for the two first principal components for fault #1 and fault #2.

5.1.5 Disturbing the data with three simultaneous faults

A case more elaborated, with three simultaneous faults was built to analyze the potentialities of the technique of PCA.

Table 6. Description of the three simultaneous faults.

Fault ID	Description of the disturbance
1, 2 e 3	The temperature in the boiler (°C) was disturbed (increase) in samples 16 to 38 (Variable #31), the pressure in the boiler (kgf/cm²) was perturbed (decrease) in samples 59 to 78 (Variable #32) and the temperature in the bottom of tube 1 (°C) was perturbed (increase) in samples 104 to 118. (Variable #9).

For this, the first three simple faults, views in the Table 4[10], were joined to build a new situation. The description of this triple faults is shown in Table 6.

5.1.6 Results for three faults

Q residual plot against samples shows clearly the three studied faults. Here it is possible to see the disturbances in the samples 16 to 38, 59 to 78 and 104 to 118.

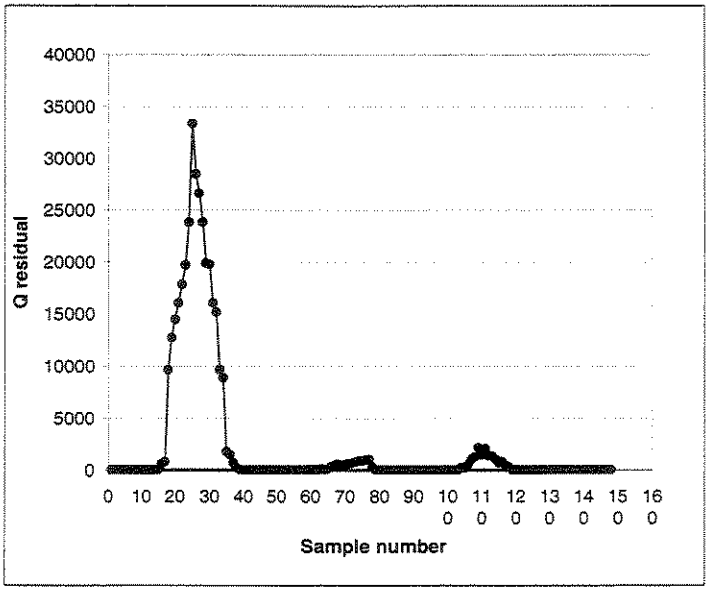


Figure 34. Q residual plot by sample for fault #1, fault #2 and fault #3.

The T^2 plot, Figure 35, only gets to capture the disturbances in faults 1

[10] Except the fault #3, because a more intense disturbance was made in the variable 9 (in the same samples 1 to 15) to get an intermediate situation between the fault #1 and the fault #2. This to get a more interesting situation aiming of evaluation of the technique of PCA

and 3. Fault #2 can not be noticed here.

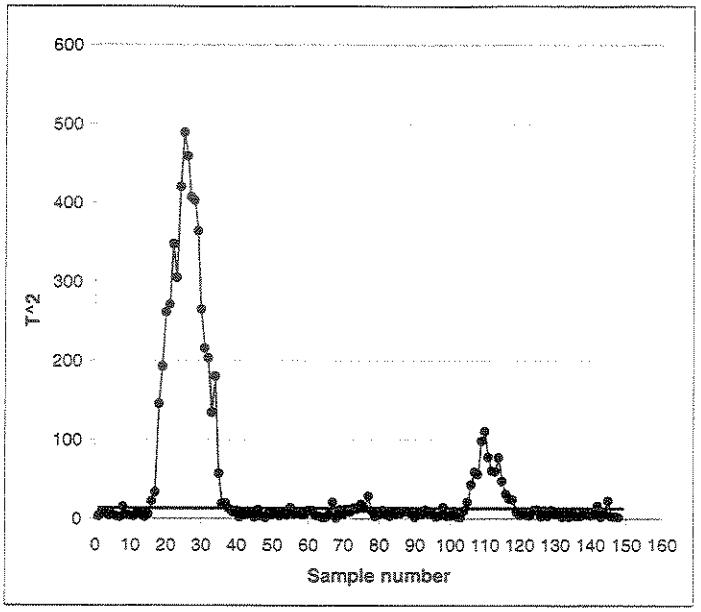


Figure 35. T^2 plot by sample for fault #1, fault #2 and fault #3.

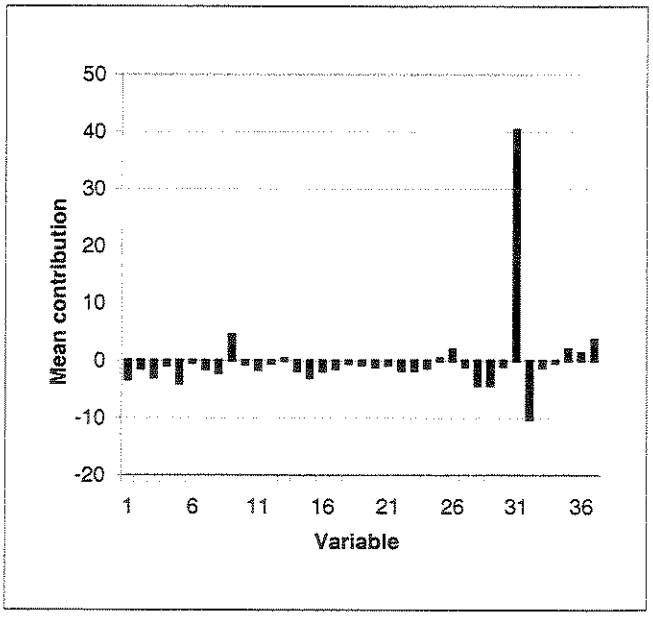


Figure 36. Mean contribution by variable for fault #1, fault #2 and fault #3.

The magnitudes and directions of the medium contributions for the variables 31, 32 and 9 appear as expected. The variable 31 appears with the largest contribution, positive, followed by variable 32, with negative contribution and last by the variable 9 with the smallest contribution among them three, also positive. The rest of the non disturbed variables

have smaller contributions.

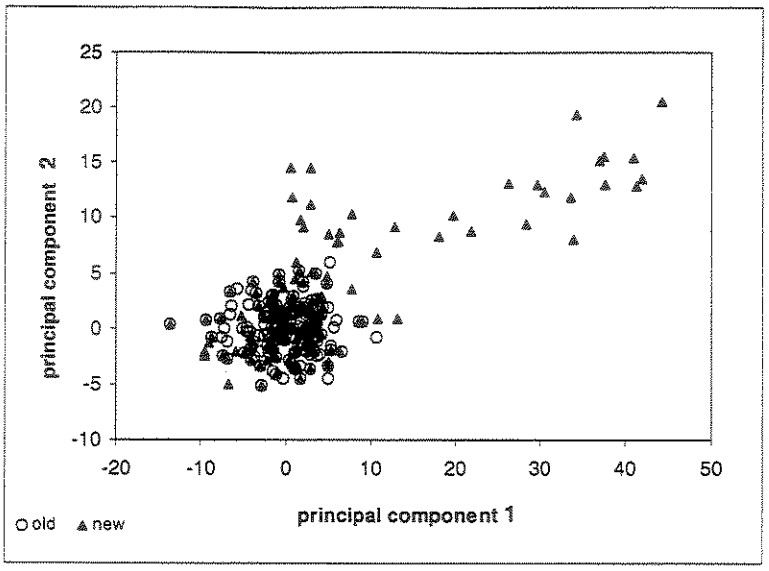


Figure 37. Scores for the first two principal components for fault #1, fault #2 and fault #3.

Given that T^2 detected abnormal behaviors in some samples they are shown in the graph of the scores in the first two principal components[11].

5.1.7 Disturbing the data with a constant reading in one of the measurement equipments

The case of constant reading in one of the measurement equipments was analyzed, too. The intention is to simulate a situation which could occur in practice as a result of errors in the measurement probe.

It was also simulated by the insertion of a constant variable in the data set. To do this, variable number 31 was modified in order to represent this kind of problem.

5.1.8 Results for the simulation of constant reading in one of the measurement equipments

When the results were analyzed, it was seen that this situation is not captured by the technique of PCA. It can be noted in the Q and T^2 plots, the mean contributions plot and the graph of the scores of the

[11] The graphs of the scores of the observations in the principal components 3 - 4 and 5 - 6 can be seen in the Appendix A1, section A1.3. (Figure 49 andFigure 50)

observations in the principal components, Figure 38, Figure 39, Figure 40 and Figure 41, respectively, that no useful information can be obtained.

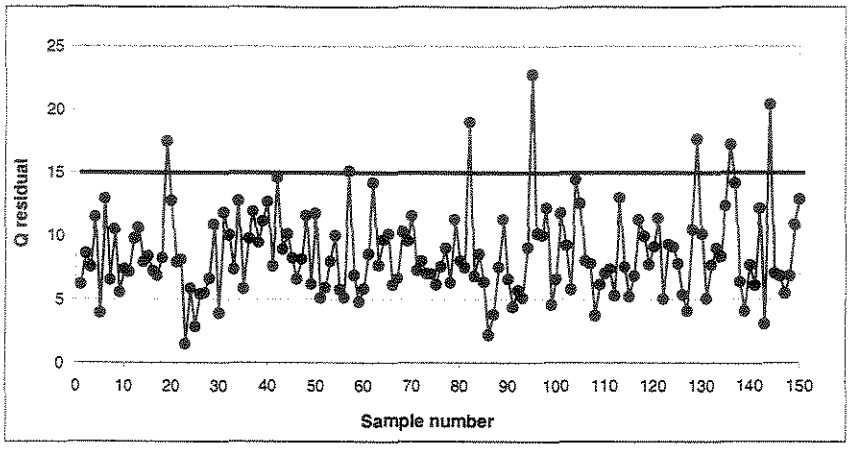


Figure 38. Q residual plot by sample for a constant reading fault.

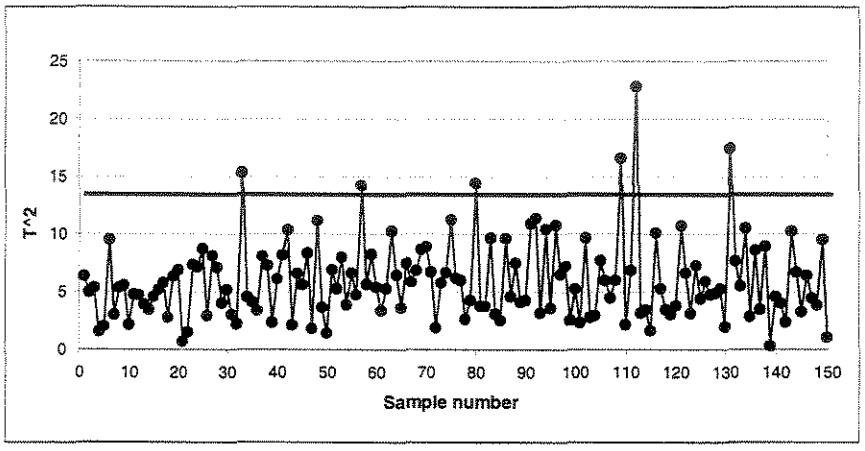


Figure 39. T^2 plot by sample for a constant reading fault.

In the case of the mean contributions plot, Figure 40, it is easy to notice that the contributions were so small that the scale had to be reduced many times to at least get to draw the graph.

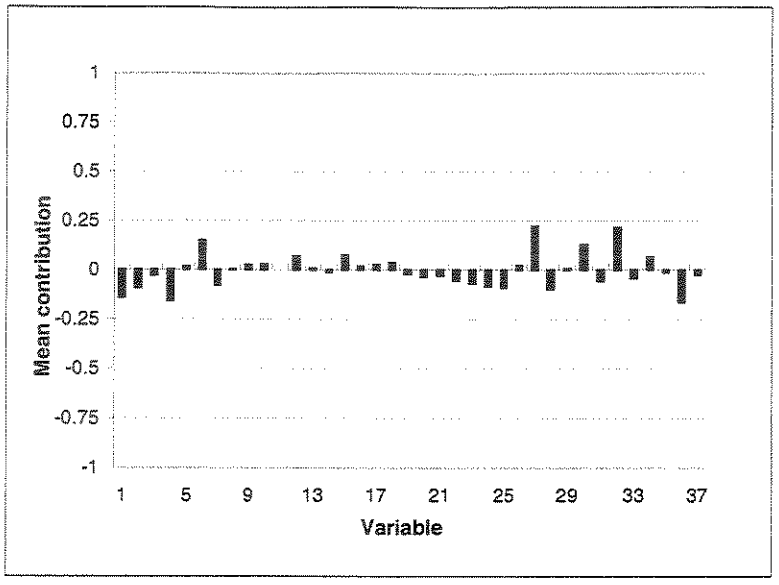


Figure 40. Mean contribution by variable for a constant reading fault.

The graph of the coordinates of the observations in the first two main components shows that still in the directions of maximum variability of the new data (▲), where the equipment fault with constant reading was included, the variabilities produced by this behavior do not cross the normal variabilities of the process, captured in the model PCA (O).

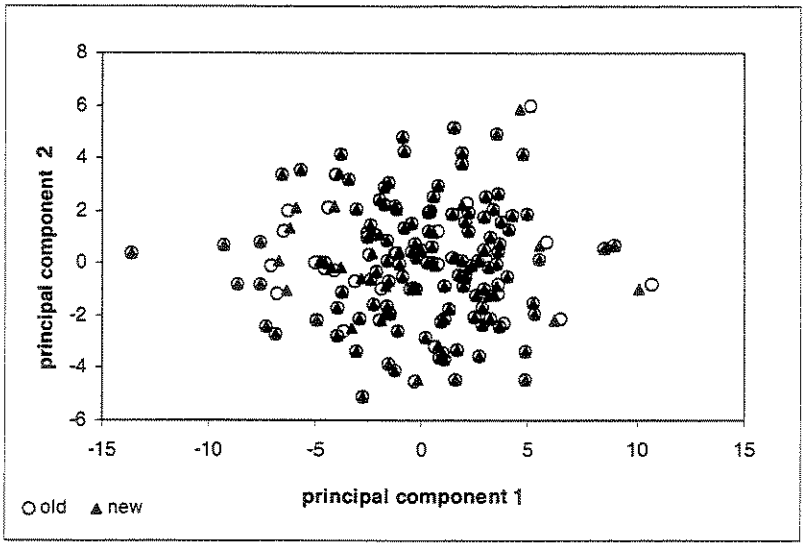


Figure 41. Scores for the first two principal components for a constant reading fault.

In the other main components this behavior is also "masked" inside the normal behavior.

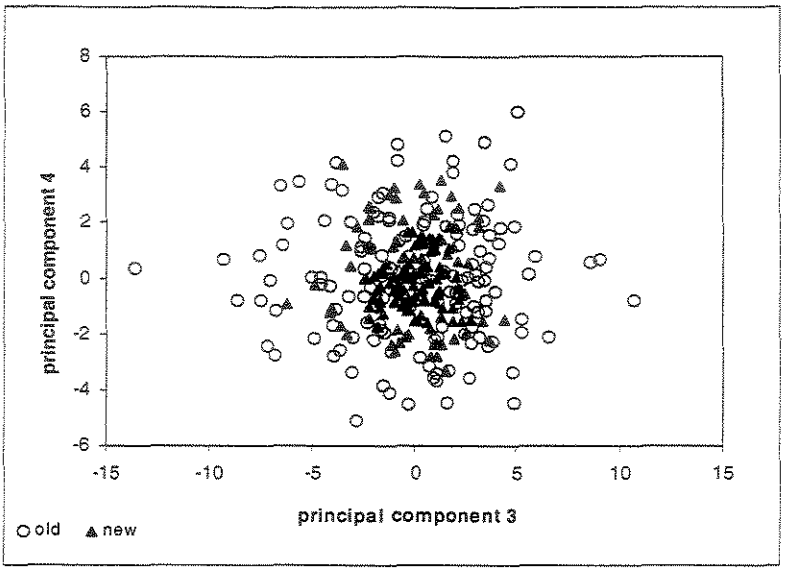


Figure 42. Scores for the principal components 3 and 4, for a constant reading fault.

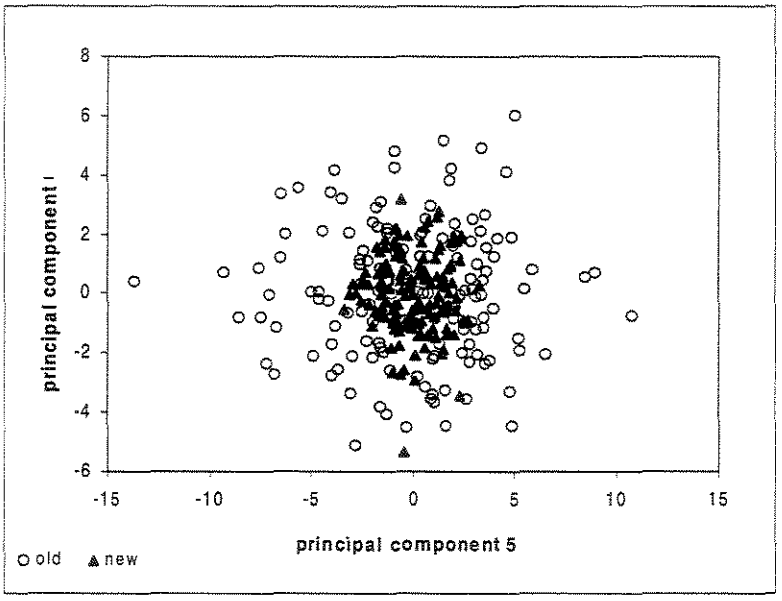


Figure 43. Scores for the principal components 5 and 6, for a constant reading fault.

Proposal to detect this type of situations

As it was seen, PCA is shown to be unable to identify faults of constant reading of a measurement equipment in a group of samples. Therefore, the detection task and identification of faults is incomplete in cases like this.

To identify this situation a good option is to build control charts for all the variables involved in the process, before applying PCA, that is a more refined analysis.

Following, the Xbar chart will be shown to see the behavior of the disturbed variable 32. An Xbar chart is a control chart of means. It is possible to use Xbar charts to track the process level and detect the presence of special causes.

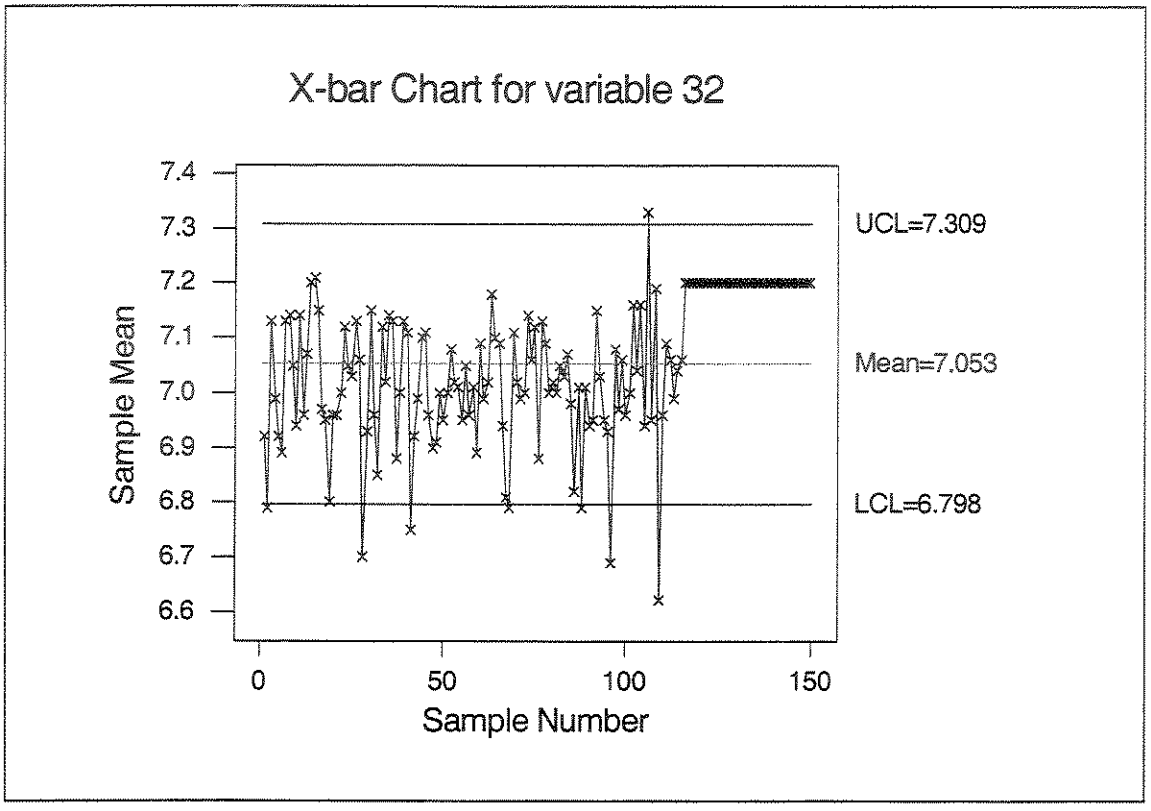


Figure 44. Control chart for variable 32.

Indeed, the Xbar chart shows the constant value of 7.20 kgf/cm^2, in the variable 32, starting from the sample 116 and going up to 150.

5.2 Results for FDA

A data set X_F with 38 variables, including the group variable that concerns the single fault type to which belongs each observation, and 117 samples was constructed. After that, it was standardized, to have mean zero and variance one.

If one observes the eigenvalues of the matrix $\mathbf{S}_W^{-1}\mathbf{S}_B$, in Table 7, it is possible to see that the first accumulates the 47.44% of the variance

between the groups of faults, the second 24.93% and the third 14.67%. Adding the first third, accumulate 87.04% of the variance, in other words, almost all the variance of the data of the process can be explained by the first three discriminant variables. In terms of the simulated industrial process, the case study of this work, the discriminant variables will be investigated to know the information that will carry each one.

Table 7. Eigenvalues of the matrix $S_W^{-1}S_B$

No.	Eigenvalue	Proportion	Cumulative	%
1	23.61	0.47	0.47	47.44
2	12.41	0.25	0.72	72.37
3	7.30	0.15	0.87	87.04
4	3.89	0.08	0.95	94.86
5	2.56	0.05	1.00	100.00

To see what is happening in the discriminant space, it is necessary to build-up the graphs of the discriminant plans, which are made starting from the coefficients, (3.3.22), of the Fisher's discriminant functions[12] (3.3.9).

5.2.1 Behavior of the data in the discriminant space

Now, the graph of the observations in the discriminant plane will be presented. In Figure 45 it is shown the individual scores, for each observation of the matrix X_F, in the first two discriminant axes.

It is reminded that each individual's coordinates in each discriminant axis are calculated using (3.3.9) with the values of the coefficients from Table 15 (in Appendix 2, section A2.2).

[12] The values of the coefficients, in this case, are in Appendix 2, A2.3 (Table 15)

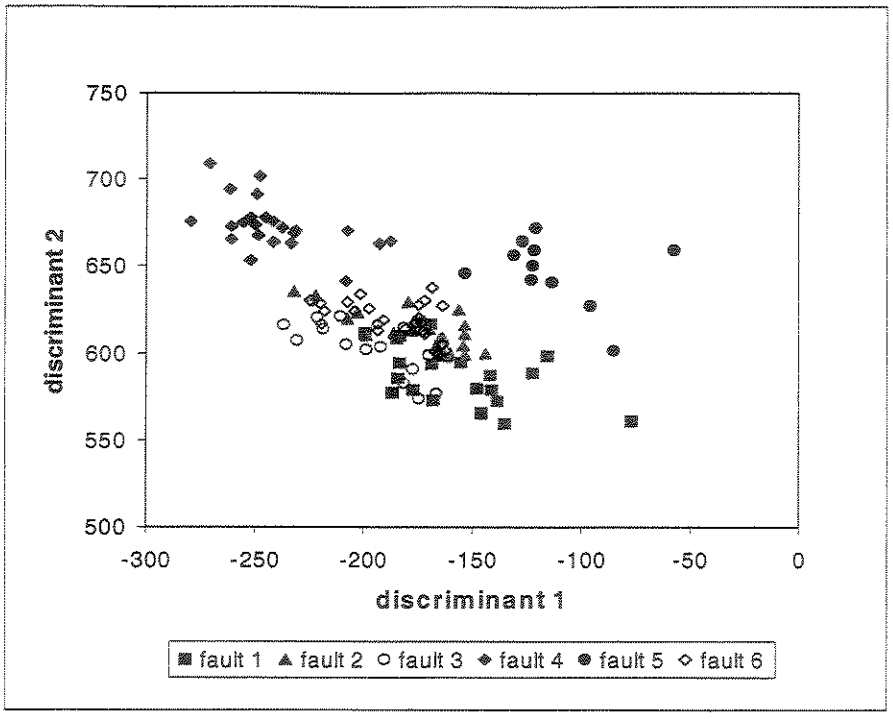


Figure 45. Graph of the scores in the first two discriminant functions[13].

It can be observed that, in Figure 45, faults 4 and 5 are more moved away in the direction of maximum separation of the groups, direction of discriminant 1. The other faults appear very close some to the other, and close to the group 6 that indicates good operating conditions.

This can indicate that inside this bank of faults, only created above these five simple faults, the group number 4 and 5 would be better differentiated.

5.2.1.1 Relative positions between the means of the groups

The means of the six groups in the new discriminant axes are given for the coordinates in Table 8:

Table 8. Group means by discriminant variables

groups	Disc.1	Disc.2	Disc.3	Disc.4	Disc.5
1	0.93	-2.53	-4.78	-0.57	-0.53
2	-0.72	-2.39	3.15	-3.03	-1.01

[13] The other plots for discriminant 1 and 3 and discriminant 2 and 3 can be seen in Appendix 2, section A2.3. (Figure 51 and Figure 52)

groups	Disc.1	Disc.2	Disc.3	Disc.4	Disc.5
3	-0.43	-2.18	1.89	3.88	-2.09
4	-5.94	5.38	-0.52	-0.10	-0.15
5	12.71	5.25	0.66	-0.12	-0.19
6	-0.15	-1.71	0.93	0.76	2.77

This group means can be seen graphically in the discriminant planes, in Figure 46

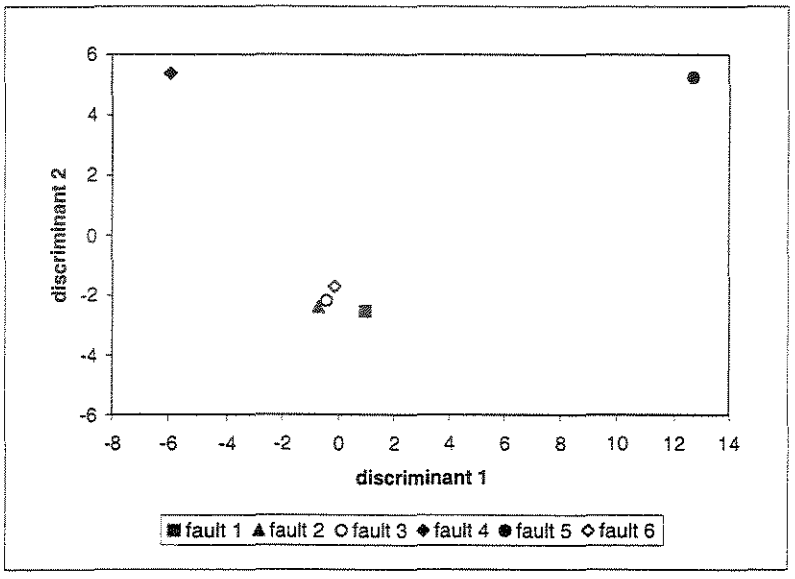


Figure 46. Graph of the relative positions between the means of the groups in the first two discriminant functions.

Indeed the averages of groups 1, 2 and 3 are in the neighborhood of the average of group 6. To observe this behavior, in other discriminant plans, see the plots in Figure 53 and Figure 54 (Appendix 2, section A2.3).

The distances between the means of the groups can be seen in Table 9. As it was seen in Figure 46, the most distant mean groups are 4 and the 5; followed by 2 and 5, 5 and 6 and 1 and 5. On the other hand, the nearest mean groups are 3 and 6, followed by 2 and 6, and 2 and 3.

Table 9: The distance between group means.

Groups	1	2	3	4	5	6
1	0.00	14.03	11.39	26.95	31.15	10.30
2	14.03	0.00	7.91	21.58	32.93	6.96
3	11.39	7.91	0.00	18.71	28.85	5.90

Groups	1	2	3	4	5	6
4	26.95	21.58	18.71	0.00	47.43	20.81
5	31.15	32.93	28.85	47.43	0.00	31.16
6	10.30	6.96	5.90	20.81	31.16	0.00

5.2.2 Classification

Ignoring the group structure, known a *priori*, the observations were put back in the nearest group, following the criterion of Fisher given in equation 3.3.25. For this it was calculated the distances initially between each point and the average of each group, in the discriminant space, choosing the nearest group to the point of interest.

The results of the classification analysis, with the allocated group and the calculated minimum distances by observation, are in the Table 14 in Appendix 2, section A2.3. More details can be obtained with all the calculated distances in the Table 13, Appendix 2, section A2.3.

The Table 10 contains a summary of the classification. It is seen that all observations are classified correctly. The exceptions are two observations of the group 2 that were allocated in group 6.

Table 10: Classification table by groups.

groups	1	2	3	4	5	6
1	23	0	0	0	0	0
2	0	18	0	0	0	2
3	0	0	15	0	0	0
4	0	0	0	23	0	0
5	0	0	0	0	11	0
6	0	2	0	0	0	25

5.2.3 Misclassification rates

The misclassification rates and the percentages by fault group are in the Tables 11 and 12. It can be seen that the rates of misclassification are very good.

Table 11: Misclassification rates by fault groups.

1	2	3	4	5	6
0.0	0.1	0.0	0.0	0.0	0.0

Table 12: Percentage of missclassification by fault groups, %.

1	2	3	4	5	6
0.0	10.0	0.0	0.0	0.0	0.0

The rate average and the percentage average of misclassification are 0.0167 and 1.67% respectively.

Conclusions

In the case of the results for PCA it can be said that all the faults studied, with the exception of the constant fault, were identified with the aid of statistics Q and T^2, diagnosing their causes with the mean contributions for each variable. It was always possible to see the faults with the Q statistic, only in a few cases it was also possible with T^2. When the abnormal behavior was observed also with T^2, it was visible in the graph of the observations in the discriminant planes, otherwise it was not.

To detect constant faults, univariate control charts should be built in the initial stage of the descriptive statistics analysis. This is because this type of behavior will not come out with PCA.

FDA shows the best possible separation among groups of faults and it made an excellent classification of the observations, resulting in very low rates of misclassification.

Chapter 6: Guide of application of PCA and FDA for detection and diagnosis of faults

6.1 Introduction

The development of the technique of principal components is not trivial, specially when applied to the monitoring of processes. It is because the amount of aspects to be checked to accomplish a correct analysis. Considering this, we found necessary, and it was one of the objectives of this work, to facilitate to the users of the industry an itinerary to develop this technique, seeking a way of taking decisions about the behavior of the process. On the other hand, we also developed an itinerary to help the application of the Fisher's discriminant analysis.

Application's Guide of Principal Components Analysis to the Monitoring of a Continuous Industrial Process

6.2.1 Model building

6.2.1.1 *Preparing the data set*

1. The first step is to collect data of the continuous process. It must be made sure that this matrix of initial data, with n lines (observations) and p columns (variables) represents the process under normal conditions of operation.
2. To standardize the initial matrix X.

6.2.1.2 *To choose the number of principal components to be retained*

1. To calculate the covariace matrix, S, of X (see equation 3.2.1).

2. To calculate the eigenvalues and eigenvectors of S (see equation 3.2.2)
3. To calculate the variance percentages explained by each principal component (see equation 3.2.9). To observe the values and to decide until which component, *k*, the values of explained variance contribute with differentiated information of the rest, in this case to keep only these components (to see example of the section 4.3.2).

Note: This decision can be aided drawing the SCREE plot with the eigenvalues of S (to see Figure 8). In case of doubt use the tool of the parallel analysis (it Figure 9).

6.2.1.3 To build the residual matrix E

1. To build the P_k matrix, i.e, only with the first *k* eigenvalues of S.
2. To calculate the residual matrix E, using the matrix expression 3.2.13.

6.2.2 Fault detection

6.2.2.1 Preparation of the data

1. To take a new reading of data (to take care so that this matrix X_{new} contains measurements of the same variables and in the same order of the original, X, of the model PCA previously built). Therefore it will have variable *p* and *m* observations (*m*>*p*).
2. To standardize Xnew using the same means and the same standard deviations of the matrix X.

6.2.2.2 To calculate the value of Q for each observation of Xnew

1. To calculate the value of $Q(i)$, i=1,2,..., m; for each observation of Xnew, using the matrix expression in 3.2.14.
2. To calculate the limit Q_α by the equation 3.2.15.
3. Rule of decision: (1) the observation *i* is considered fault suspicion if $Q(i)>Q_\alpha$. (2) the observation is considered a fault if there are a set of neighboring observations that also present an abnormal pattern.

Note: This decision can be aided using the graph for Q (see example in

Figure 10).

6.2.2.3 *To calculate the value of T^2 for each observation of Xnew*

1. To calculate the value of $T^2(i)$, $i = 1,2,\ldots, m$; for each observation of X_{new}, using the matrix expression in 3.2.16.
2. To calculate the limit $T^2_{k,n,\alpha}$ by the equation 3.2.17.
3. Rule of decision: (1) the observation i is considered fault suspicion if $T^2(i) > T^2_{k,n,\alpha}$. (2) the observation i is considered a fault if there were a set of neighboring observations that also present an abnormal pattern.

Note: This decision can be aided with the graph for T^2 (see example in Figure 12).

6.2.2.4 *Scores of the observations in the principal components.*

1. To calculate the coordinates of each observation in the k principal component chosen (to see formulates in equation 3.2.7).

   Note (1): If the analysis of T^2, step 6.2.2.3 of this itinerary, does not show the existence of faults, in general the individuals' coordinates will not have any behavior abnormal to show either.

   Note (2): To build the graph of the coordinates of the observations in the principal components (see example in Figure 13); it is very useful to see better the behavior of the data.

6.2.3 Fault Diagnosis

6.2.3.1 *To calculate the contributions measured for variable Q_{mean}*

1. To identify the observations with faults resulting of the step 6.2.2.2, item 3 above.
2. To locate the residues of those observations in the residual matrix E.
3. To calculate the mean by variable, columns of matrix E, only for the residues of those observations with fault.
4. Rule of decision: The contribution of the variables to the fault will be

measured by the magnitude of the average calculated in the previous step.

Note: To create the Q mean contribution plot will help to diagnose the variables causing the faults.

6.3 Application's Guide of Fisher's Discriminant Analysis to the Monitoring of a Continuous Industrial Process.

6.3.1 Classification

6.3.1.1 *Construction of a Bank of Faults*

1. The first step will be to group a set of observations monitored previously where it is known (for the application of a previous PCA or built or complemented with the people's participation with important experience and wide knowledge of the process in study) the existence of some specific fault types. In other words, to the set of monitored variables it will be added a group variable, informing the fault types that characterize each row of the matrix (to see the form of the matrix in equation 3.3.2, or 3.2.1.2 for more details).

   Note (1): if the fault is well differentiated it will avoid confusion with the others, minimizing like this the rate of misclassification of new observations to be tested.

2. A group should have been included that represents the behavior of the process under normal conditions of operation.

   Note (2): it can be worked initially with few known faults but an important point to take into consideration is that to increase the Bank of Faults will increase the possibilities to classify new observations of the process in the correct fault type.

   Note (3): Always observe that this method may not be sensitive to faults not contained in the training dates.

6.3.1.2 *Calculating the means by groups in the new discriminant axis*

1. To evaluate the means of each group in the equation 3.3.10 to calculate their coordinates in each discriminant axis. The result will be a matrix with the number of rows equal to that of groups and the number of columns equal to the number of discriminant functions.

6.3.1.3 *Classification Rule*

1. First calculate the distances of each observation to each group. The result will be a matrix with the number of rows idem to the number of observations and the number of columns same to the number of discriminant functions.
2. Form the matrix of distances of the previous step the distance measured for observation (row) is chosen and that observation is classified in that group. We do this for each observation.

6.3.1.4 *Calculating the rates of misclassification by faults*

1. To calculate the rate of misclassification by group use equation 3.3.26.
2. The percentage is calculated using equation 3.3.27.

6.3.1.5 *Mean rate of missclasification*

1. The average rate of misclassification is calculated using equation 3.3.28.
2. The average percentage of misclassification is calculated using equation 3.3.29.

6.3.2 Visualization and differentiation of the faults

6.3.2.1 *Calculating the scores for each observation in discriminant axes*

1. To calculate the coordinates of each observation in the discriminant axes (see equation 3.3.10). This will generate a matrix with number of rows equal to the number of observations and, as columns, the number of discriminant functions.

Note (1): to build the graph of the coordinates of the observations in the principal components (see example in Figure 13) it is very useful to see better the behavior of the data.

Note (2): When the analysis of T^2, step 6.2.2.3, don't show the existence of faults, in general the individuals' coordinates will not have any behavior abnormal to show either.

Chapter 7: Discussions and Conclusions

7.1 Discussions

Different kinds of faults can arise in everyday industrial practice. In this work some examples of situations that could appear were analyzed using Principal Components Analysis and Fisher's Discriminant Analysis.

According to PCA:

- Using both Q and T^2 statistics in the analysis of results from principal components technique it was possible to detect the occurrence of faults, for most of the simulated cases, and the Q and T^2 confidence limits proved to have a very good discriminatory feature for this detection.
- Not all of the fault types are able to be detected. Among these cases it is worthwhile to mention the faults produced not by abrupt changes in the variability of the process but rather by the contrary effect, variability null or almost null. An example of this situation is the fault that produces the freezing of the reading of one of the measurement equipments, being in a fixed value. The results showed that PCA does not allow one to see this pattern of behavior which is understandable if it is thought that the method is driven precisely to the detection of large variabilities.

  Taking into consideration that these situations can appear with certain frequency, the proposal is to build control graphs (e.g. Xbar chart) for each variable, before entering in the application of PCA.
- The mean contribution plot behaves as a useful tool to diagnose the cause of the problem. This graph provides information of the magnitude of the contribution of each variable to the total variability of process as

well as the direction, positive or negative, of this magnitude.

- The graph of scores for the principal components is a useful discriminating tool when the fault produces large values of T^2; in these cases the abnormal samples can be seen clearly. On the other hand, when the fault produces large values of Q, but low or non-important values of T^2, the graphic of scores alone does not help to see the fault behavior.

According to FDA:

- The classification procedure was shown satisfactory with a percentage average of misclassification of 1.67%.
- It is understandable that the more complete is the bank of faults the more possibilities there will be to apply FDA to the data of a continuous industrial process with effective results.
- The graph of the scores of the observations/samples allow one to visualize the behavior of the groups in the discriminant space. The quality of the representation will depend upon the differentiation degree between the different types of studied faults.

The implementation of a software in Fortran appears to be a useful contribution for the application of such methods in the industrial practice, since they are more flexible than the existing commercial packages which, generally, do not allow the construction of the PCA model with the proposed procedure in a straightforward and cheaper manner.

7.2 Conclusions

Many of the faults induced were accurately detected, in such a way that it was possible to identify with exactness the samples where the faults took place and which were the responsible variables.

A number of statistics tools were described which show a great potential for identification of fault diagnosis and abnormal operations, using PCA.

Some of these, such as T^2, are special cases of general multivariate control situations and may be employed either with or without the use of principal components. The Q statistic, on the other hand, are developed precisely to deal with residuals related to PCA.

It was noticed the need to maintain the initial exploratory analysis making univariate control charts seeking faults of the type "constant reading" in measurement equipments, the ones which, in general, cannot be detected using PCA.

On the other hand, FDA allows one to classify other observations of the process with a bank of faults built with the known faults. In the studied cases it allowed to classify the observations with a low rate of misclassification. FDA also allows one to have an idea of the space distribution of the different kinds of faults.

The representation of a chemical misclassification process could be developed with efficiency using empirical models based in historical data, like PCA. Different modeling approaches could be established using multivariate statistical techniques. They are very useful, allowing the acquisition of valuable information for the purpose of efficient control of the process.

7.3 Papers and publications developed during the elaboration of this work

During the elaboration of this work the following papers were submitted and/or presented in events and national and international Congresses.

- Díaz-Cónsul, C.M. and Maciel-Filho, R., "Multivariate Statistical Techniques for the Monitoring of Continuous Industrial Processes", FOCAPO 2003, Coral Springs, Florida, January 12-15 (2003).
- Díaz-Cónsul, C.M. and Maciel-Filho, R., "Controle Estatístico Multivariado para um Processo Contínuo", COBEQ – 2002, Natal, August (2002).
- Díaz-Cónsul, C.M. and Maciel-Filho, R. "Implementation of a Software in Fortran for Statistical Process Control of a Continuous Chemical

Process", SCI 2002, Orlando, Florida, July 14-18 (2002).

- Díaz-Cónsul, C.M. and Maciel-Filho, "Multivariate Statistical Control for a Continuous Process", AIChE Annual Meeting 2001, Nevada (approved) (2001).
- Díaz-Cónsul, C.M. and Maciel-Filho, R. "Control of a Continuous Process using Multivariate Statistical Methods", EPFEQ-II, Campinas, São Paulo, Brazil, September (2001).
- Díaz-Cónsul, C.M. and Maciel-Filho, R. "Controle Estatístico Multivariado para um Processo Contínuo", Brazilian Journal of Chemical Engineering, (Submit for publication in December, 2001).
- Díaz-Cónsul, C.M. and Maciel-Filho, R. "Control of a Continuous Process using Multivariate Statistical Methods", 51st Canadian Chemical Engineering Conference, Halifax, October (2001).
- Díaz-Cónsul, C.M. and Maciel-Filho, R. "Process Control of a Continuous Process using Multivariate Statistical Methods", Canadian Journal of Chemical Engineering, (Submit for publication in December, 2001).

Chapter 8: Recommendations for Future Works

8.1 Future works

Anywhere where large amounts of monitoring data are available, specially in the industry of chemical processes, it will be necessary the use of all the possible tools to extract conclusions about the behavior of the process. The application of multivariable statistical techniques is one of the ways to increase the knowledge of the process.

In the section 8.2 it will be made a revision of the recent application of other multivariable statistical techniques (or variations of techniques already known) for the application of detection and diagnosis of faults. These cases can serve as inspiration to seek other applications in industry and to develop future works in this field.

8.2 Other multivariable methods applied to statistical process monitoring

Below we comment on several other modern applications of PCA, found in the literature, as well as other multivariate methods that can be worked to study their potentialities.

8.2.1 PCA Multi-way

Martin, 1999, reviews the concepts of monitoring the acting of the process through an industrial application in a reactor of fluidized bed and of a simulation of a polymerization reactor (batch methyl methacrylate polymerization reactor). The author introduces the use of *Multi-way* Principal Components Analysis for the case of the processes with batches.

This refers to the inclusion of the *time* in the analysis, so that the problem has to be analyzed in three directions: the different measured variables, the lots or batches, and the different intervals of time inside which are made the measurements.

LOUWERSE, 2000, discuss the multivariate statistical control of batch processes (MSPC) in models three-way, with the following purposes: (1) to show how the models of Principal Components can be used for data of batch processes and how a new batch is projected in each model, (2) the theory of batch graphs of MSPC is described and gotten better, and (3) a method to treat with batches "no concluded" is introduced for the on-line monitoring. LOUWERSE presents and compares the monitoring for three worked models.

To monitor the acting of the process in real time, CHEN et.al, 1998b, propose the technique of *Multi-way* Principal Components Analysis, as an alternative cheaper than the traditional analytical instruments. The method goes beyond the system in stationary state and it supplies the approximate monitoring in real time for continuous processes.

This monitoring can detect faults more quickly, compared with others approximate monitoring. Several important subjects for the proposed approach are discussed by CHEN et.al, 1998b; some of them are: the structure of entrance of the model, pre-treatment of the data and the reach of the predictable horizons. An extension multi-block of the basic methodology is also treated to facilitate the isolation of the fault. A process of Tennessee Eastman is used to demonstrate the power of the new approximate monitoring proposed.

Proposal of Work with Multi-way PCA

A possible research topic to be developed is to use the *Multi-way* PCA to deal with time varying process, which should be coupled with the tools developed in this work considering the system in pseudo steady-state. A hierarchical approach could be developed with the *Multi-way* PCA on the top of the sequence.

8.2.2 No linear PCA

For processes highly not linear, the lineal monitoring form sometimes is inefficient due to the dimensionality of the process not always represented by a small number of lineal principal components. The variables of the process, correlated not lineally, can be reduced to a group of principal components no linear, with the application of a proposal of No linear Principal Components Analysis. A more efficient monitoring of the process, now can be implemented, in the space of No linear Principal Components Analysis, in few dimensions (ZHANG, et. al, 1997).

In parallel with the conventional multivariable graphs, ZHANG, 1997, uses the graph of accumulated scores, which provides a significant exit in the separation of different conditions and/or operation faults, leading to robust preventions, of badly potential operations of the plant. Besides, ZHANG, et. al, 1997, demonstrate the effectiveness of the No linear Principal Components Analysis, and the other proposals in their article, in the monitoring conditions of a polymerization reactor.

HIDEN, et. al. (1999), review briefly the attempts to extend the lineal Principal Component Analysis to no linear and it proposes a technique "symbolic oriented" to No linear Principal Components based on genetic programming. The applicability of this proposal is shown using two simple systems no linear and collected data of a column of industrial distillation.

Proposal of Work with No linear PCA and comments

This technique is appearing as a new approach, therefore the theoretical base should be revised carefully, before any application. However the results published up to now are encouraging and this technique could be incorporated to the analysis of data of chemical processes industry, to test their advantages.

8.2.3 PLS

CHEN, et.al, 1998, developed a controller using multivariable statistical

models, presented to reach the objective of producing products of high quality, inside a continuous process, working with the difficulty that quality measurements many times are not available on-line or they are not available to be produced. A model of Principal Components, which incorporates "late" variables in the time, it is used and the control objectives are expressed in the space of the scores of the Principal Components. A controller is projected in the structure of the model of predictive control (MPC) and it is used to control equivalent representation of the process in the space of the scores of the Principal Components. The predictive model, for the algorithm MPC, is built using Partial Least Squares (PLS). The proposed controller is tested in two cases of study, which involve a column of binary distillation and a process in Tennessee Eastman.

Proposal of Work with PLS

New controller structures or algorithms could be used to replace the PCA approach in order to take full advantage of multivariable and non-linear controller.

8.2.4 Discriminant PLS

On the other hand, CHIANG, et. al., 2000, propose the use of the Fisher's Discriminant Analysis (FDA), and Discriminant Partial Least Square (DPLS) as an alternative for the diagnosis of faults, according to him better than the Principal Components Analysis. It shows the use of these techniques applied to collected data of the simulator of chemical plant Tennessee Eastman.

Proposal of Work with Discriminant PLS

This approach could be incorporated in the developed software to increase to robustness.

8.2.5 Online SPC

Guh, et.al. 1999, developed an intelligent tool to do an economical on-line SPC. The article includes a detailed revision of the moderated progresses reported in the literature in the field of the automation of SPC until the moment of their publication. Taking this into consideration the idea is to use the developed software to build-up an online-SPC.

8.2.6 Dynamic PCA

Ku et al., 1995, proposed the disturbance detection and isolation for dynamic systems using an extension of PCA to monitor dynamic chemical process. This dynamic approach constructs a time series model from the data and is referred to as *Dynamic PCA* or *DPCA*, which could be incorporated in the tools developed in this work.

Chapter 9: Bibliographic References

- Chen, G., McAvoy, T.J., Piosovo, M.J. "A Multivariate Statistical Controler for on-line Quality Improvement". Journal in Process Control, 8, 2, 139-149 (1998).
- Chiang, L.H., Russell, E.L and Braatz, R.D., "Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares and principal component analysis". Chemometrics and Intelligent Laboratory Systems, 50, 2, p. 243-252 (2000).
- Cónsul, C.M.D., "Auxílios à Interpretação na Análise Discriminante de Fisher". Dissertação de Mestrado, Estatística. IMECC. UNICAMP (2000).
- Duda, R. O. and Hart, P. E. "Pattern Classification and Scene Analysis" Wiley, New York, (1973).
- Geladi, P and Kowalski, B.R, "Partial least squares regression (PLS): a tutorial", Analytica Chimica Acta 185, 1-17 (1986).
- Gilbert, E. "The Effect of Unequal Variance-Covariance Matrices on Fisher´s Linear Discriminant Function". Biometrics, 25, 3, September (1969).
- Guh, R.S., Tannock, J.D.T., O`Brien, C. "IntelliSPC: A Hybrid Intelligent Tool for On-Line Economical Statistical Process Control. Expert Systems with Applications". 17, 195-212 (1999).
- Hiden, H.G., "Non-linear principal components analysis using genetic programming", Computers and Chemical Engineering, 23, 413-425 (1999).
- Horn, J.L., "A rationale and test for the number of factors in factor analysis", Psychometrika, 30, 179-185 (1965).
- Hudlet, R. and Johnson, R. "Classification and Clustering" Academic Press, New York, pp 371-394, (1977).

- Ipek, H., Ankara, H. and Ozdag, H. "Technical Note the Application of Statistical Process Control". Minerals Engineering. 12, 7, p. 827-835. (1999).
- Jackson, J.E., "Quality Control Methods for Two Related Variables", Industrial Quality Control, Vol. XII, No. 7, p. 2-6, (1956).
- Jackson, J.E., "Quality Control Methods for Several Related Variables", Technometrics, 1, 4, 359-377 (1959).
- Jackson, J.E. and Mudholkar, G.S. "Control Procedures for Residuals Associated with Principal Components Analysis", Technometrics, 21, 3, 341-349 (1979)..
- Jackson, J.E., "Principal Components and Factor Analysis I: Principal components", Journal. of Quality Technology, 12, 4, 201-213 (1980).
- Jackson, J.E., "Multivariate Quality Control", Communications in Statistic: Part A, Theory and Methods, 14, 11, 201-213 (1985).
- Jackson, J.E., "A user's guide to principal components", John Wiley & sons, Inc. Series in probability and mathematical statistics: Applied probability and statistics (1991).
- Johnson, R.A and Wichern, D.W., "Applied multivariate statistical analysis", Prentice-Hall, Inc., Englewood Cliffs, N.J. Third Edition (1992).
- Kosanovich, K.A., Dahl, K.S. and Piosovo M.J., "Improved process understanding using multiway principal components analysis". Ind. Eng. Chem. Res., 35, 138-146 (1996).
- Kourti, T. and MacGregor, J.F., "Multivariate SPC Methods for Process Control and Product Monitoring", Journal of Quality Technology, 28, 4, 409-428 (1996).
- Krzanowski, W.J., "Principles of Multivariate Analysis: a users perspective", New York: Clarendon Press Oxford, Oxford Statistical Science Series, 563 p., (1988).
- Ku, W., Storer, R.H. and Georgakis, C., "Disturbance detection and isolation by dynamic principal component analysis", Chemometrics and Intelligent Laboratory Systems, 30, 179-196 (1995).

- Lachenbruch, P.A. "Discriminant Analysis". Collier Macmillan Canada Ltd. (1975).
- Mardia, K.V., Kent, J.T. and Bibby, J.M. "Multivariate Analysis". London: Academic Press, 521 p., (1979).
- Martin, E.B., Morris, A.J. and Kiparissides, C., "Manufacturing performance enhancement through multivariate statistical process control". Anuals Reviews in Control, 23, 35-44 (1999).
- McGregor, J.F, Kourti, T., "Statistical process control of multivariate processes", Control Engineering Practice, 3, 3, 403-414 (1995).
- Ralston, P., DePuy G. and Graham, J.H., "Computer-based monitoring and fault diagnosis: a chemical process case study", ISA Transactions, 40, 85-98 (2001).
- Russell, E.L and Braatz, R.D., "Fault isolation in industrial processes using Fisher's Discriminant Analysis", in "Proc. of the Conf.", held at Snowbird, Utah, July 5-10, 1998. AIChe Symposium Series No. 320, Vol. 94 (1998), pp.380-385.
- Santana, P.L.; Tvrzská, M.; Maciel, R.; Dechechi, E.C. "Real time optimization of a multiphase reactor: Modeling the evaporation effect". Chem. Eng. Science /no prelo/ 1999.
- Saibt, E.F., Barchet, V.M.F. and Radharamanan, R. "Use of multivariate analysis in controlling a soft drink fabrication process". Computers Ind. Engng, v.31, No. 1/2, 261-264 (1996).
- Shunta, J.P., "Achieving World Class manufacturing Through Process Control". Prentice-Hall, Inc (1995).
- Valle, S., Li, W. and Qin, S.J., "Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods", Industrial Engineering Chemical Research, 38, 4389-4401 (1999).
- Wetherill, G.B. and Brown, D.W. "Statistical Process Control: Theory and practice". Chapman & Hall. (1991)
- Wise, B.M., "Adapting multivariate analysis for monitoring and modeling of dynamic systems". PhD Thesis, University of Washington (1991).

- Wise, B.M. and Gallagher, N.B., "The process chemometrics approach to process monitoring and fault detection", Journal of Process Control, 6, 6, 329-348 (1996).
- Wold, S., Esbensen, K. and Geladi P., "Principal components analysis: a tutorial", Chemometrics and Intelligent Laboratory Systems, 2, 37-52 (1987).
- Zhang, J., Martin, E.B. and Morris A.J., "Process monitoring using non-linear statistical techniques", Chemical Engineering Journal, 67, 181-189 (1997).

Appendix 1: Principal Components Analysis

A1.1 Result 1

Let S be the covariance matrix associated with the random vector $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]$. Let S have the eigenvalue-eigenvector pairs (λ_1, e_1), (λ_2, e_2), ..., (λ_p, e_p), where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. The *i*th principal component is given by

$$\mathbf{t}_i = \mathbf{X}\mathbf{p}_i = \mathbf{X}\mathbf{p}_{1i} + \mathbf{X}\mathbf{p}_{2i} + \cdots + \mathbf{X}\mathbf{p}_{pi}, \quad i = 1,2,\ldots,p$$

with these choices,

$$\begin{aligned} \text{var}(\mathbf{t}_i) &= \mathbf{p}_i'\mathbf{S}\mathbf{p}_i = \lambda_i \\ \text{cov}(\mathbf{t}_i, \mathbf{t}_k) &= \mathbf{p}_i'\mathbf{S}\mathbf{p}_k = 0 \end{aligned} \quad , \text{ for i=1, 2, ..., p.}$$

Proof: can be seen in Johnson and Wichern, 1992, p. 358.

A1.2 Result 2

Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]$ have covariance matrix S, with eigenvalue-eigenvector pairs (λ_1, e_1), (λ_2, e_2), ..., (λ_p, e_p), where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. Let $\mathbf{Y}_1 = \mathbf{e}_1'\mathbf{X}, \mathbf{Y}_2 = \mathbf{e}_2'\mathbf{X}, \ldots, \mathbf{Y}_p = \mathbf{e}_p'\mathbf{X}$ be the principal components. Then

$$s_{11} + s_{22} + \cdots + s_{pp} = \sum_{i=1}^{p} \text{var}(\mathbf{X}_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{var}(\mathbf{Y}_i)$$

Proof: can be seen in Johnson and Wichern, 1992, p. 359.

A1.3 Other plots of the coordinates, scores, of the observations in the principal components

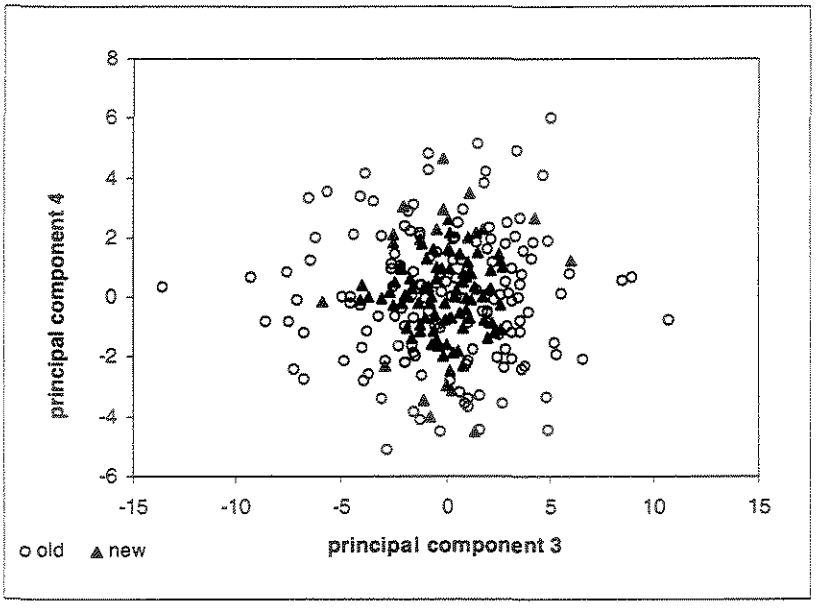


Figure 47. Scores for principal components 3-4 for fault #4.

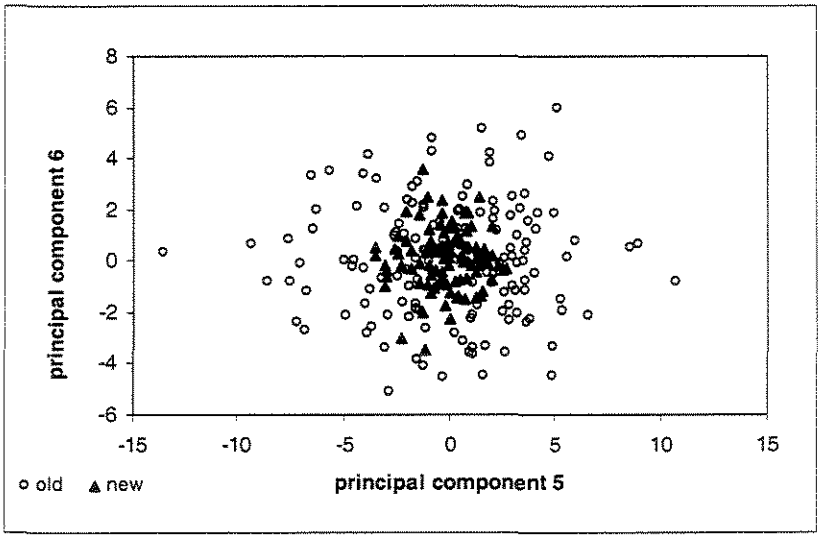


Figure 48. Scores for principal components 5-6 for fault #4.

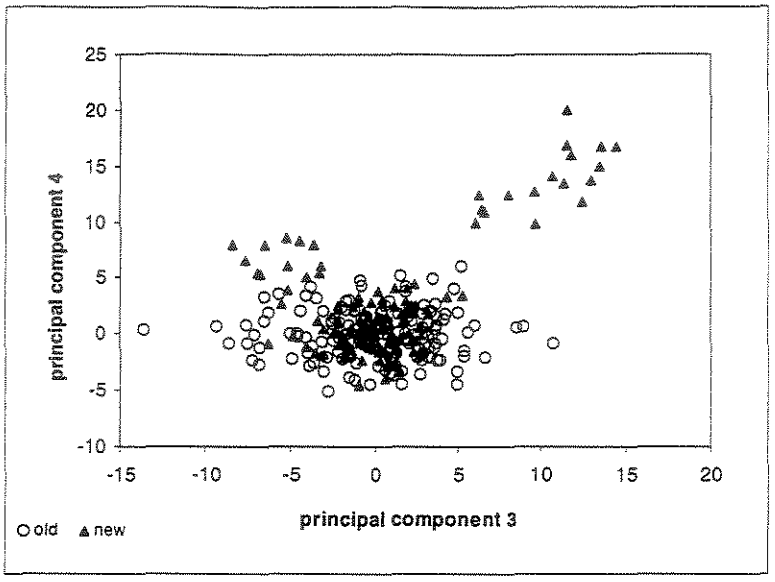


Figure 49. Scores for the principal components 3-4 for fault #1, fault #2 and fault #3.

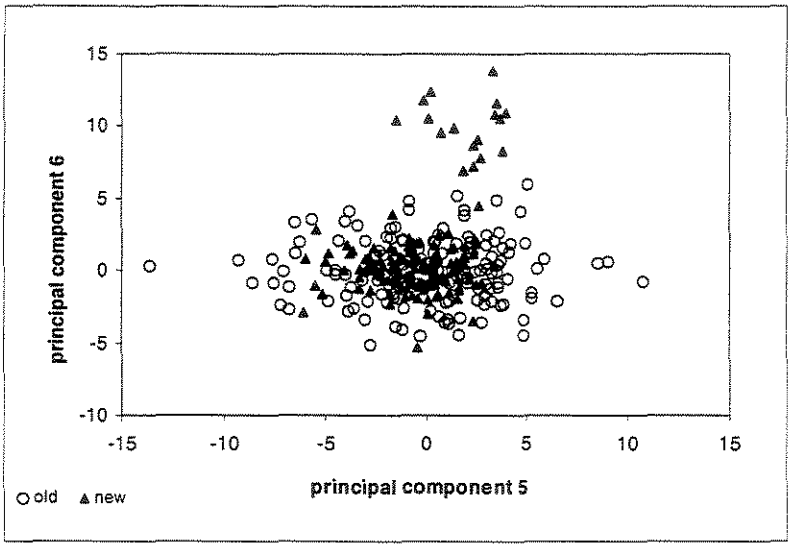


Figure 50. Scores for principal components 5-6 for fault #1, fault #2 and fault #3.

Appendix 2: Fisher Discriminant Analysis

A2.1 Maximization Lemma of quadratic forms

Maximization Lemma: Let $\mathbf{B}_{pxp}$ be positive definite and $\mathbf{d}_{px1}$ be a given vector. Then for an arbitrary nonzero vector $\mathbf{x}_{px1}$, $\max_{\mathbf{x}\neq 0}\frac{(\mathbf{x}'\mathbf{d})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{d}'\mathbf{B}^{-1}\mathbf{d}$, with the maximum attained when $\mathbf{x}_{px1} = c\left(\mathbf{B}^{-1}\right)_{pxp}\mathbf{d}_{px1}$, $\forall c \neq o$.

Proof: can be seen in Johnson and Wichern, 1992, p. 66.

A2.2 Classification

Table 13: Distances between each observation and each single fault mean.

obs.	grupo 1	grupo 2	grupo 3	grupo 4	grupo 5	grupo 6
1	4.0	46.7	54.7	108.2	199.4	28.3
2	9.0	46.7	41.4	101.2	182.6	18.8
3	0.7	82.8	74.9	139.5	231.5	49.9
4	1.1	85.0	75.0	129.3	223.7	48.8
5	0.8	68.8	77.3	137.3	233.3	49.8
6	7.2	96.3	108.4	135.3	232.7	80.9
7	3.7	79.8	94.1	157.5	219.2	65.5
8	1.0	72.8	75.0	138.7	253.8	49.6
9	4.8	82.1	89.0	133.5	282.9	70.5
10	4.9	80.9	76.5	155.1	278.0	68.4
11	12.9	136.3	118.7	188.1	265.6	107.5
12	4.1	103.8	100.5	165.7	243.1	67.9
13	3.3	92.4	76.4	163.2	235.5	66.7

obs.	grupo 1	grupo 2	grupo 3	grupo 4	grupo 5	grupo 6
14	4.0	80.3	86.8	127.5	244.3	43.6
15	6.1	95.1	81.6	133.8	266.3	78.1
16	0.8	75.7	62.5	133.3	217.2	44.3
17	0.6	79.5	80.1	136.0	228.0	56.2
18	2.4	82.6	73.7	130.0	230.5	41.6
19	3.8	102.1	94.9	132.4	258.7	69.6
20	8.8	44.1	37.0	111.7	205.5	16.8
21	5.1	51.0	45.1	89.9	226.2	35.4
22	15.2	31.8	22.4	97.7	225.9	21.8
23	9.3	49.4	47.6	120.9	203.6	16.7
24	46.4	12.2	30.6	101.0	214.7	6.3
25	44.4	4.1	33.3	103.0	242.1	23.6
26	50.5	12.5	29.4	74.5	251.2	8.3
27	64.3	5.6	33.3	91.9	210.2	17.8
28	56.4	5.1	30.3	107.6	206.3	28.2
29	67.9	4.6	38.3	92.4	267.2	17.7
30	50.0	2.7	36.4	97.2	243.7	20.5
31	71.6	0.2	47.2	105.7	257.9	31.3
32	82.3	1.5	60.0	100.9	281.8	38.5
33	100.4	4.1	79.8	139.2	266.4	53.5
34	92.4	2.7	52.0	111.8	298.9	41.0
35	65.2	1.8	61.4	114.9	252.5	43.4
36	77.9	1.3	63.7	116.3	264.0	47.3
37	96.4	3.8	66.3	143.8	240.9	45.7
38	90.6	3.2	71.8	130.0	282.8	57.9
39	98.7	8.8	73.1	153.6	264.4	73.8
40	105.3	5.9	85.2	139.3	283.7	66.9
41	121.6	8.4	87.4	155.2	312.8	70.1
42	100.0	7.3	92.8	148.9	280.2	67.7
43	53.8	3.4	38.0	82.4	259.4	25.8
44	119.7	97.0	12.1	173.1	252.5	79.3
45	94.8	76.2	5.0	124.9	312.2	58.4
46	147.9	130.2	26.8	192.3	342.8	120.9
47	79.8	71.8	2.3	138.7	251.6	47.9
48	57.7	39.2	1.1	99.1	248.1	28.7

obs.	grupo 1	grupo 2	grupo 3	grupo 4	grupo 5	grupo 6
49	68.2	47.5	0.5	119.1	249.6	38.7
50	74.8	39.0	1.4	106.6	255.3	33.3
51	58.2	45.7	1.1	110.4	220.6	27.5
52	38.2	44.1	4.9	97.1	232.5	21.0
53	72.5	46.6	1.0	105.8	234.2	28.5
54	72.3	27.2	7.7	79.3	249.3	21.4
55	32.3	41.1	7.0	99.2	221.9	24.1
56	79.8	57.4	1.7	107.6	261.5	31.8
57	60.4	42.4	7.0	106.1	256.1	13.3
58	55.9	34.9	2.3	116.0	228.6	24.3
59	104.5	86.9	99.7	6.1	347.8	92.5
60	103.0	68.8	92.4	6.5	302.9	63.4
61	88.3	74.7	76.3	5.4	305.0	55.6
62	84.5	66.7	71.3	7.9	274.8	50.4
63	105.2	85.8	89.5	3.2	321.0	82.7
64	97.3	88.1	72.0	6.8	345.1	66.5
65	121.7	107.6	105.2	0.2	344.7	91.2
66	90.4	90.7	94.2	4.9	314.3	64.6
67	105.3	92.8	102.9	4.5	342.1	68.8
68	132.0	102.0	100.8	1.1	348.2	90.9
69	114.7	88.3	94.6	3.9	369.7	75.3
70	105.6	92.7	92.1	1.4	315.6	77.0
71	151.3	134.7	130.1	1.5	361.5	114.2
72	195.2	170.0	166.4	8.4	389.2	149.0
73	168.6	153.8	160.6	9.9	358.4	151.9
74	211.6	195.3	213.0	16.1	447.3	173.2
75	154.4	148.2	146.3	4.7	388.9	137.1
76	228.3	206.4	208.3	19.8	467.1	201.5
77	146.7	113.7	119.6	1.0	367.7	102.2
78	139.8	130.5	115.9	2.2	374.7	102.1
79	120.2	106.9	99.8	2.0	313.0	92.1
80	153.3	123.9	135.6	2.6	362.5	104.7
81	154.7	137.6	134.1	4.2	390.9	105.6
82	288.1	322.8	302.5	407.6	5.5	289.0
83	183.4	205.7	197.6	280.7	3.9	177.7

obs.	grupo 1	grupo 2	grupo 3	grupo 4	grupo 5	grupo 6
84	215.7	254.5	252.6	349.2	1.6	216.6
85	316.7	350.4	355.1	455.2	9.7	313.0
86	311.6	326.5	319.7	423.1	7.1	302.1
87	358.7	382.3	383.4	496.6	15.8	355.2
88	235.9	261.5	255.9	342.1	1.1	234.3
89	231.8	256.0	244.4	341.3	0.4	222.1
90	197.2	211.3	198.7	305.8	4.1	175.9
91	106.8	122.9	120.5	228.6	28.3	108.2
92	165.4	190.1	197.5	298.5	11.3	152.4
93	47.5	31.4	28.5	88.2	222.6	0.4
94	47.3	31.5	32.3	90.9	228.2	0.1
95	59.7	42.4	40.0	95.2	225.9	1.1
96	60.0	24.4	32.1	81.1	231.4	2.5
97	58.2	41.8	38.8	111.6	251.6	1.8
98	30.7	38.6	33.6	91.4	198.5	2.5
99	35.5	34.9	30.9	120.6	214.0	3.1
100	52.3	30.4	38.9	96.1	226.7	0.5
101	36.8	35.5	33.2	95.9	220.1	0.7
102	53.6	52.3	31.0	127.7	210.5	4.5
103	48.4	41.2	46.1	101.3	229.9	1.0
104	42.2	28.3	36.7	98.2	192.7	1.7
105	55.5	32.1	31.2	99.3	270.0	2.8
106	49.2	32.4	43.2	87.4	250.2	1.5
107	48.2	37.5	39.9	111.1	192.7	1.6
108	39.3	38.0	36.8	108.8	228.2	1.1
109	43.2	28.7	38.3	98.9	232.9	0.8
110	60.0	45.2	55.2	92.4	228.1	3.0
111	54.2	49.7	31.6	101.9	224.2	2.0
112	50.9	51.7	35.7	97.9	228.9	1.9
113	36.0	35.0	43.6	71.5	232.3	3.5
114	54.2	28.7	19.4	74.0	220.8	4.5
115	64.9	24.9	32.0	93.2	215.2	3.1
116	47.3	32.3	40.3	88.9	213.8	0.6
117	49.7	36.0	40.0	102.5	242.8	0.7

Table 14: Classification table by observations, with the calculated minimum distances.

obs.	Group original	Group allocated	Calculated minimum distance
1	1	1	4.0
2	1	1	9.0
3	1	1	0.7
4	1	1	1.1
5	1	1	0.8
6	1	1	7.2
7	1	1	3.7
8	1	1	1.0
9	1	1	4.8
10	1	1	4.9
11	1	1	12.9
12	1	1	4.1
13	1	1	3.3
14	1	1	4.0
15	1	1	6.1
16	1	1	0.8
17	1	1	0.6
18	1	1	2.4
19	1	1	3.8
20	1	1	8.8
21	1	1	5.1
22	1	1	15.2
23	1	1	9.3
24	2	6	6.3
25	2	2	4.1
26	2	6	8.3
27	2	2	5.6
28	2	2	5.1
29	2	2	4.6
30	2	2	2.7
31	2	2	0.2
32	2	2	1.5
33	2	2	4.1
34	2	2	2.7

obs.	Group original	Group allocated	Calculated minimum distance
35	2	2	1.8
36	2	2	1.3
37	2	2	3.8
38	2	2	3.2
39	2	2	8.8
40	2	2	5.9
41	2	2	8.4
42	2	2	7.3
43	2	2	3.4
44	3	3	12.1
45	3	3	5.0
46	3	3	26.8
47	3	3	2.3
48	3	3	1.1
49	3	3	0.5
50	3	3	1.4
51	3	3	1.1
52	3	3	4.9
53	3	3	1.0
54	3	3	7.7
55	3	3	7.0
56	3	3	1.7
57	3	3	7.0
58	3	3	2.3
59	4	4	6.1
60	4	4	6.5
61	4	4	5.4
62	4	4	7.9
63	4	4	3.2
64	4	4	6.8
65	4	4	0.2
66	4	4	4.9
67	4	4	4.5
68	4	4	1.1
69	4	4	3.9

obs.	Group original	Group allocated	Calculated minimum distance
70	4	4	1.4
71	4	4	1.5
72	4	4	8.4
73	4	4	9.9
74	4	4	16.1
75	4	4	4.7
76	4	4	19.8
77	4	4	1.0
78	4	4	2.2
79	4	4	2.0
80	4	4	2.6
81	4	4	4.2
82	5	5	5.5
83	5	5	3.9
84	5	5	1.6
85	5	5	9.7
86	5	5	7.1
87	5	5	15.8
88	5	5	1.1
89	5	5	0.4
90	5	5	4.1
91	5	5	28.3
92	5	5	11.3
93	6	6	0.4
94	6	6	0.1
95	6	6	1.1
96	6	6	2.5
97	6	6	1.8
98	6	6	2.5
99	6	6	3.1
100	6	6	0.5
101	6	6	0.7
102	6	6	4.5
103	6	6	1.0
104	6	6	1.7

obs.	Group original	Group allocated	Calculated minimum distance
105	6	6	2.8
106	6	6	1.5
107	6	6	1.6
108	6	6	1.1
109	6	6	0.8
110	6	6	3.0
111	6	6	2.0
112	6	6	1.9
113	6	6	3.5
114	6	6	4.5
115	6	6	3.1
116	6	6	0.6
117	6	6	0.7

Table 15. Fisher's discriminant coefficients

var	Disc.1	Disc.2	Disc.3	Disc.4	Disc.5
1	0.920	0.649	-0.843	0.490	0.805
2	-0.089	0.278	-0.074	-0.042	-0.054
3	1.464	-0.662	-1.294	0.333	0.965
4	-0.310	-0.167	0.252	-0.113	0.079
5	0.558	0.367	0.414	-0.377	-0.252
6	0.262	0.111	0.183	0.130	0.291
7	-0.110	-0.037	0.077	-0.357	0.034
8	-2.051	1.317	2.276	-1.000	-3.880
9	-0.005	-0.272	0.320	1.129	-1.247
10	1.281	-2.209	-0.060	0.115	-1.243
11	1.091	-0.349	-1.508	0.786	4.290
12	-0.883	0.573	0.636	0.884	-0.975
13	-2.978	3.050	-0.432	-0.357	-0.458
14	-2.946	-0.956	5.063	1.896	-0.819
15	0.497	-0.284	-0.312	-0.806	0.565
16	-3.268	0.519	-1.082	1.412	2.054
17	1.628	2.241	-2.738	-1.517	-2.368
18	3.967	2.048	0.110	-0.280	-0.383

var	Disc.1	Disc.2	Disc.3	Disc.4	Disc.5
19	0.737	-0.547	-0.192	-1.076	0.329
20	-0.748	-1.607	-0.688	2.179	1.806
21	1.292	-0.468	-0.592	-0.484	-3.528
22	0.177	-0.030	-1.733	-1.008	-1.080
23	-1.724	0.764	3.506	-1.009	2.402
24	-0.241	0.490	0.126	-0.149	-0.205
25	-0.529	-0.044	0.369	0.129	0.871
26	-0.231	0.129	0.356	-0.140	0.528
27	0.024	0.138	-0.346	-0.104	-0.356
28	-0.274	-0.059	-0.140	-0.234	2.291
29	0.589	-0.438	-0.068	0.004	-2.359
30	0.096	-0.284	-0.370	-0.190	-0.398
31	0.340	-0.386	-2.172	-0.586	-0.627
32	0.390	0.265	-1.129	1.446	0.792
33	0.052	-0.555	-0.105	-0.037	-0.787
34	-0.699	0.063	0.300	-0.022	-0.115
35	0.075	0.087	-0.599	-0.164	-0.299
36	0.056	0.649	0.455	0.362	-0.610
37	-0.280	0.278	0.455	0.259	0.022

A2.3 Other plots of discriminant planes.

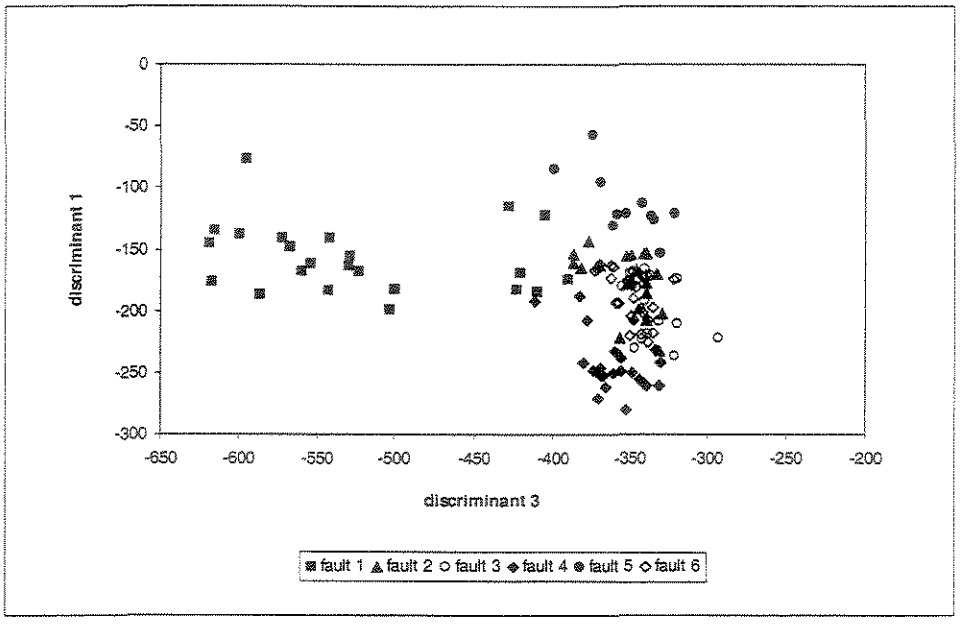


Figure 51. Graph of the scores in the first and third discriminant functions.

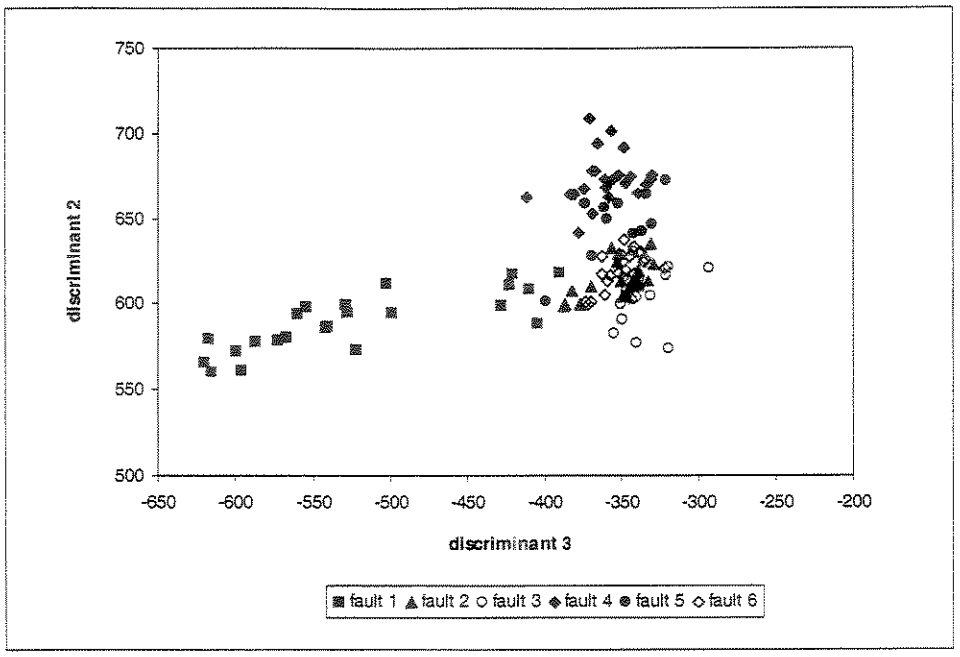


Figure 52. Graph of the scores in the second and third discriminant functions.

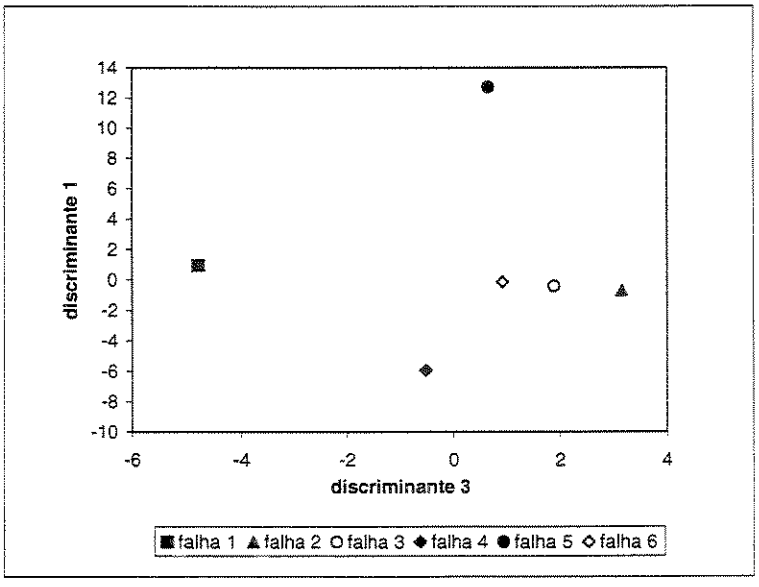


Figure 53. Graph of the relative positions between the means of the groups in the first and third discriminant functions.

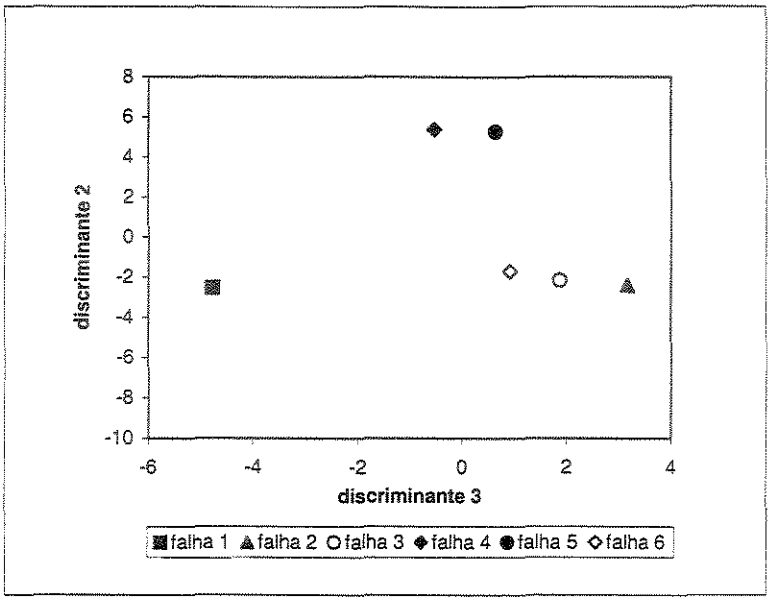
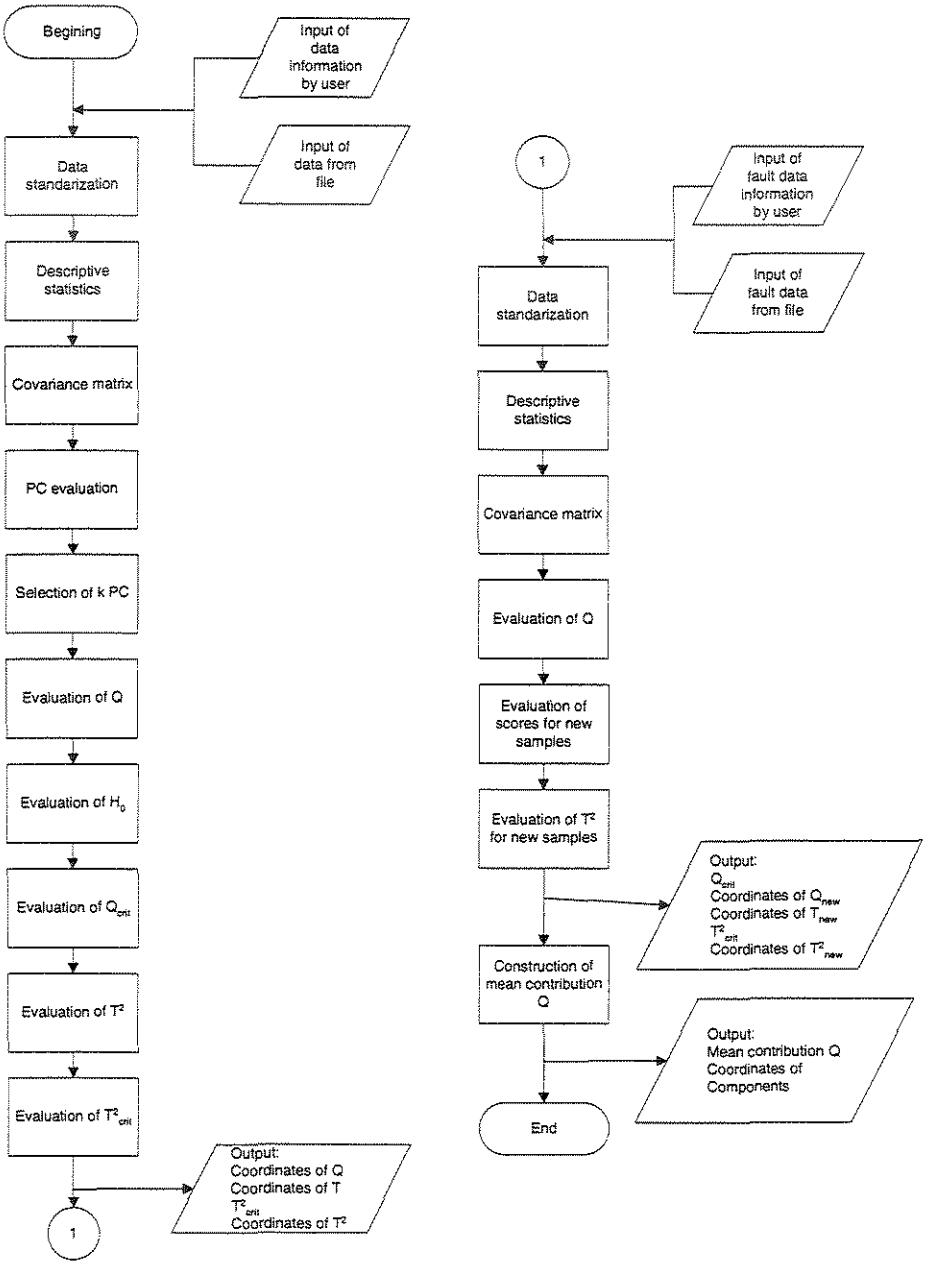


Figure 54. Graph of the relative positions between the means of the groups in the second and third discriminant functions.

Appendix 3: Flowchart of Fortran Programs

PCA program



FDA program