الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR- ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR - ANNABA

جامعة باجي مختار- عنابة


**Faculty: Sciences of Engineering** **Year: 2018**
**Department: Electronics**

# THESIS

**Presented in order to obtain the diploma of Ph.D 3rd Cycle**

## Entitled

# Diagnosis of Uncertain Systems using Principal Component Analysis

**Option: Automatic and Signals**

**By: AIT-IZEM Tarek**

**Committee Members:**

| | | |
|---|---|---|
| **President:** | **DOGHMANE Noureddine,** | **Pr. Univ. Annaba** |
| **Supervisor:** | **HARKAT Mohamed-Faouzi** | **Pr. Univ. Annaba** |
| **Reviewers:** | **RAMDANI Messaoud** | **Pr. Univ. Annaba** |
| | **LACHOURI Abderezzak** | **Pr. Univ. Skikda** |
| **Guest:** | **KRATZ Frédéric** | **Pr. Université d'Orléans, INSA-CVL** |

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR- ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR - ANNABA

جامعة باجي مختار- عنابة


**Faculté: Sciences de l'ingéniorat** **Année : 2018**
**Département: Electronique**

# THÈSE

**Présentée en vue de l'obtention du diplôme de Doctorat 3ème Cycle**

## Intitulé

# Diagnostic de Fonctionnement des Systèmes Incertains par Analyse en Composantes Principales

**Option : Automatique et Signaux**

**Par : AIT-IZEM Tarek**

**Devant le Jury :**

| | | |
|---|---|---|
| **Président:** | **DOGHMANE Noureddine,** | **Pr. Univ. Annaba** |
| **Directeur de Thèse :** | **HARKAT Mohamed-Faouzi** | **Pr. Univ. Annaba** |
| **Examinateurs :** | **RAMDANI Messaoud** | **Pr. Univ. Annaba** |
| | **LACHOURI Abderezzak** | **Pr. Univ. Skikda** |
| **Invité :** | **KRATZ Frédéric** | **Pr. Univ. d'Orléans,INSA-CVL** |

*To the memory of my father Madjid Ait-Izem (1951-2006).*

*To my wonderful loving mother.*

*To my brothers, and sister, . . .*

# *Acknowledgements*

This work was done while the author was with the laboratory LASA at Badji-Mokhtar Annaba University, Algeria, and with the laboratory PRISME at INSA-CVL Bourges, France. First and foremost, I would like to give the most sincere thanks to Prof. Dr. Mohamed-Faouzi Harkat and Prof. Dr. Frédéric Kratz, who are respectively my supervisor and co-supervisor. I am grateful for their guidance and support in my research work and the preparation of this dissertation. I would also like to express my deepest appreciation to Prof. Dr. Messaoud Djeghaba, who is also my thesis supervisor in the early years of my thesis, for his insightful discussion about this work and his constructive comments.

I would also thank the professors of both poles, UBMA and INSA-CVL, for their guidance and helpful comments. Many thanks should go to the colleagues in LASA and PRISME, and the friends in Bourges, who have made my time spent at this lovely french city both substantial and pleasurable.

Finally, I would like to give my gratitude to my family members for their persistent and unconditional understanding and supporting.

# ملخص

في هذه الأطروحة، نقوم بالتركيز على كشف وعزل أخطاء أجهزة الاستشعار باستعمال تحليل المكونات الرئيسية. في الحالات العملية، يؤدي الإرتياب في قياسات أجهزة الاستشعار إلى صعوبة كبيرة في عملية صنع القرار عند القيام بعملية الكشف، مما يثير ويزيد من عدد الإنذارات الكاذبة والقرارات غير الدقيقة. في شكلها الإعتيادي، لا تميز طريقة تحليل المكونات الرئيسية بين نقاط البيانات وأخطاء القياس المرتبطة التي تختلف تبعا للظروف التجريبية. وبالتالي، فإننا نحقق في حل حديث ومتين، الذي يتمثل في إستعمال النطاقات البيانية والتي تمثل مختلف ملاحظات قياسات أجهزة الاستشعار للمتغيرات خلال فواصل زمنية.

وبالتالي، فإن نمذجة العمليات تتم على أساس ما يسمى طريقة تحليل المكونات الرئيسية للبيانات ذات النطاق، حيث يتم تفصيل أربع من أكثر الطرق المعروفة، ويتم تقديم تمديد منهجية عملية الرصد القائم على طريقة تحليل المكونات الرئيسية التقليدية في حالة البيانات الجديدة ذات النطاق. ويشمل ذلك ما يلي: تعريف مؤشرات جديدة للرصد والكشف عن الأعطال، وإدخال مبدأ إعادة بناء المتغيرات لعزل الأعطال، وهو ما يستعمل أيضا في استنباط معيار جديد لتحديد عدد المكونات الرئيسية للاحتفاظ بها في نموذج تحليل المكونات الرئيسية للبيانات ذات النطاق. ونوضح كيفية تنفيذ الكشف عن أخطاء أجهزة الاستشعار وعزلها باستعمال الخوارزمية المقترحة باستخدام بيانات اصطناعية، حيث يتم إجراء دراسة معمقة ومقارنة باستخدام محاكاة مونت كارلو لمختلف نماذج المكونات الرئيسية ذات النطاق و مؤشرات الرصد، وتطبيقها على آلة الطحن و جهاز عمود التقطير.

**المفاتيح:** تحليل المكونات الرئيسية، الإرتياب، البيانات ذات النطاق، الكشف عن الأخطاء والعزل، مؤشرات الرصد، مبدأ إعادة بناء المتغيرات.

# *Abstract*

In this manuscript, we emphasis on the detection and isolation of process sensor faults based on Principal Component Analysis (PCA). In real life cases, the sensor data uncertainties cause a significant difficulty in control decision making, which evokes and increases the number of false alarms and imprecise decisions. In its standard form, PCA makes no distinction between data points and the associated measurement errors which vary depending on experimental conditions. Thus, we investigate a recent and robust solution which consists in capturing the variability of the multivariate observations by interval-valued variables.

Process modelling is therefore performed based on the so-called PCA for interval-valued data, where four of the most known methods are detailed , and the extension of the methodology of conventional PCA based statistical process monitoring is presented for the new interval-valued data case. This includes : the introduction of new interval monitoring statistics for the detection of faults, the introduction of an interval reconstruction approach for the isolation of faults, which is also used in deriving a new criterion for the determination of the number of principal components to retain for the interval PCA model.

The implementation of the proposed sensor fault detection and isolation methods are illustrated using synthetic data, where a more in-depth study and comparison is performed using Monte-Carlo simulations for the various PCA methods and monitoring statistics, and applied to milling machine data and distillation column process benchmark.

**Keywords :** Principal Component Analysis, Uncertainties, Interval Data, Fault detection and Isolation, Monitoring statistics, Reconstruction Principle.

# *Résumé*

Les travaux présentés dans ce manuscrit sont axés sur la détection et la localisation de défauts en utilisant l'analyse en composantes principales (ACP). Dans les cas réels, les incertitudes des données de capteurs provoquent des difficultés dans la prise de décision par-rapport à la présence des défauts, ce qui augmente le nombre de fausses alarmes et de décisions imprécises. Sous sa forme standard, l'ACP ne fait aucune distinction entre les mesures collectées et les erreurs de mesure associées, qui varient en fonction des conditions expérimentales. Ainsi, nous étudions une solution récente et robuste qui consiste à capturer la variabilité des observations multivariées par des variables de type intervalle.

La modélisation des processus est donc basée sur ce que l'on appelle l'ACP par intervalles. Dans une première partie, nous detaillons quatre des méthodes les plus connues. Par la suite, nous proposons une extension de la méthodologie de detection et localisation de défauts à base d'ACP conventionelle, au cas intervalle. Cela comprend : l'introduction de nouvelles statistiques de détection des défauts pour le cas intervalle du modèle ACP, l'introduction d'une approche de reconstruction par intervalles pour la localisation des défaut, et qui est également utilisée pour dériver un nouveau critère de détermination du nombre de composantes principales à retenir pour le modèle ACP par intervalle.

Les méthodes de détection et localisation des défauts proposées (à base d'ACP intervalle) sont illustrées par des données synthétiques avec une étude approfondie et une comparaison à l'aide de simulations de Monte-Carlo. Une application réele est presenté à la fin sur les données d'une fraiseuse, et un benchmark de la collonne de distillation

**Keywords :** Analyse en Composantes Principales, Incertitudes, Données intervalles, Detection et Localisation des Défauts, Statistiques de détection, Principe de Reconstruction.

# Contents

# List of Figures

# List of Tables

# Introduction

With increasing demands for efficiency and product quality, and progressing integration of automatic control systems in high-cost and safety-critical processes, the field of monitoring, fault detection and fault diagnosis has gained a very important role. Indeed, fault detection and diagnosis has been very much regarded as an integrated part for many process control systems. High reliability and safety of many industrial processes require the development of an effective fault detection and isolation (FDI) algorithms. In this context, many methods have been developed to improve the reliability and durability of systems, so that the faults related to system operations should be detected, isolated, and corrected in time to avoid severe damages to the systems.

The successful operation of modern complex systems is also largely dependent on the validity of sensor signals providing information for display and control. To enhance safety and improve plant performance, redundant sensors are often installed for measuring critical variables, which can generally be categorized into hardware and analytical redundancies. The general idea of hardware redundancy approaches is to measure one critical variable using two or more sensors and then detect and isolate the faulty sensor. These approaches have been widely used in safety-critical systems for their simplicity and robustness. Without the use of additional sensors, the analytical redundancy approaches identify the functional relations between the measured variables via a mathematical model that can be either developed based on the underlying physics or derived directly from the measurements. Residuals between the sensor measurements and the modeled outputs can then be generated for the detection and isolation of the faulty sensor. Analytical redundancy approaches can be further categorized according to the type of their required a priori knowledge as model-based methods, knowledge-based expert systems, and data-driven methods. This classification was used to present the state of the art methods for process fault detection and diagnosis (Venkatasubramanian et al., 2003a; Venkatasubramanian et al., 2003b; Venkatasubramanian et al., 2003c) as presented in Figure 1.

Fault diagnosis strategies possess both advantages and disadvantages, and the suitable approach should be selected according to the specific applications (Isermann, 2005). In the late seventies, expert systems were largely employed for diagnosis; nonetheless the use of expert knowledge in diagnosis started to show some major limitation. The acquiring and maintenance of the needed expert knowledge became a hard task in the deployment of diagnostic expert systems. As a possible alternative, model based diagnosis approaches have been developed

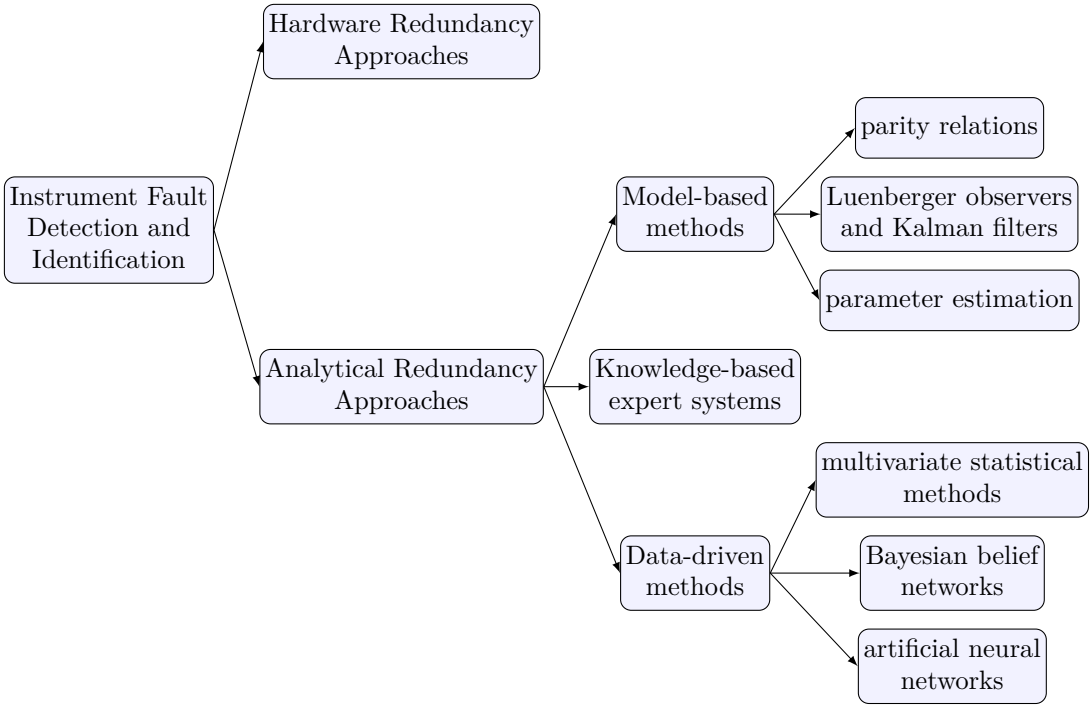



FIGURE 1 – Categorization of Methods for Instrumentation Fault Detection and Identification

by exploiting models of the diagnosed system that had been created while designing it. However, generalized diagnosis-oriented system models are usually difficult to obtain. In this case, data-driven fault diagnosis approaches seem to be more relevant for practical use, especially in large scale complex systems. The data-driven approach makes use of the information from the system's history data. In other words, diagnosis can be done based on different variables that are measured through various sensors of the process. However, most manufacturers still rely on traditional Statistical Process Control (SPC) methods which use univariate statistical analysis such as mean, median, standard deviation, etc. Unfortunately, univariate statistics often miss the underlying patterns in process data. This is where advanced Multivariate Statistical Process Control (MSPC) methods come in handy. Among the most common approaches used in MSPC is Principal Component Analysis (PCA), which is the main subject of this thesis. PCA was originated by (Pearson, 1901) and later developed by (Hotelling, 1933; Jolliffe, 2002). The applications of PCA is discussed in (Cooley and Lohnes, 1971), (Wax and Kailath, 1985), (Rao, 1964), (Jackson, 1991), (Harmon and Duboc, 1995) and many others works. PCA based diagnosis strategies have shown their high performances in the last two decades, and have received much attention for fault detection and isolation applications (Jackson and Mudholkar, 1979), (Piovoso et al., 1992), (MacGregor and Kourti, 1995), (Ku et al., 1995), (Qin, 2003), (Harkat et al., 2006). The principle of PCA based diagnosis can be summarized as follows: A PCA model is first established from history data, which is considered as the training phase. Then, by using the trained model, new data can be tested according to that model, giving feed about the state of the system, corresponding to healthy or faulty state. This is considered as the

testing phase. Thus, fault detection and isolation (FDI) can be achieved with these two phases.

Principal component analysis (PCA) method is among the most popular statistical data-based approaches used for process modeling and monitoring purposes. However, in several situations, the process is uncertain and the available information is formalized in terms of intervals.

These uncertainties have a negative impact on the established model, and thus, on the fault detection and isolation (FDI) performances. The need for interval-valued data may arise in connection with the imprecision of measurement devices, process uncertainties or with the data fluctuations in the case of recorded measures during a specific interval of time. In fact, considering the minimum and maximum recorded values offers a more complete insight about the measured phenomenon than considering the average values. For more precision in representing the real data, this uncertainty can be treated by considering an interval representation (Tulsyan and Barton, 2017a; Tulsyan and Barton, 2017b; Tulsyan and Barton, 2017c), instead of a single-valued representation. In this case, the determination of such model requires using new techniques adapted for the interval-valued data. For this propose, several PCA models for interval-valued data are proposed in literature (Cazes et al., 1997; Gioia and Lauro, 2006; Irpino, 2006; Le-Rademacher and Billard, 2012; Palumbo and Lauro, 2003; D'Urso and Giordani, 2004; Wang et al., 2012).

Thereby, the FDI performances can be improved significantly by taking into account additional information on all measurements to keep the systems in operation. In other words, have a more in depth knowledge on the precision of the measurements. Indeed, a more complete information on measurements can be achieved by taking into account the uncertainty of sensor measurement. A recent solution including such additional information can be addressed by describing a set of measurement uncertainty in terms of interval-valued data. Thus, the new interval-valued measurement encloses all the recorded sensor measurements, during a period of time between minimum and maximum values which correspond to the bounds of the interval. However, FDI techniques, and more specifically those based on statistical methods like PCA, have been developed for the analysis of single-valued variables, and are not adapted for the interval-valued case.

First, let us note that the PCA model is based on the eigen-decomposition of covariance data matrix (Wold et al., 1987). So, considering the new type of interval-valued data, the determination of such model is a quite hard task, given the absence of well-established mathematical methods in the matter (for the time-being). Therefore, many researchers investigated the use of geometrical approaches, giving birth to many well approximated interval-valued PCA methods, such as vertices PCA (VPCA), centers PCA (CPCA) (Cazes et al., 1997), midpoints-radii PCA (MRPCA) (Palumbo and Lauro, 2003), symbolic covariance interval-valued PCA (Le-Rademacher and Billard, 2012) and complete-information PCA (CIPCA) (Wang et al., 2012). Other methods for computing interval-valued PCA models can be found in the literature, as in (D'Urso and

Giordani, 2004), (Gioia and Lauro, 2006), (Irpino, 2006), and many others, which tend to work for narrow intervals only, and are not considered in this work.

Therefore, the main objective of this work is to extend the FDI based PCA approaches to deal with interval valued data. In other words, we have to define the new FDI strategies for interval-valued data. For that, we investigate the use of the four most known interval-valued PCA methods CPCA, VPCA, MRPCA, and CIPCA, and extend various FDI techniques based PCA to the interval-valued PCA case.

## Thesis Contributions

This thesis presents a first step towards a long term goal of developing a well-established theory for FDI based interval-valued PCA. In this case, dealing with static interval-valued PCA. Especially, we are interested in answering the following questions:

— How to apply PCA based modeling to deal with interval-valued data (IVD) (interval-valued PCA)? and how to apply FDI strategies in the case of IVD based PCA?

— What are the differences between the conventional PCA and IVD-PCA in term of performances for FDI?

— What are the methods used to determine the number of principal components (PCs) for the IVD-PCA model?

— Is IVD-PCA's use limited to small amount of data? or can it be used for large scale processes?

— Can IVD-PCA based FDI be extended to the dynamic or non-linear case?

First, we studied the possibility of using the VPCA model for FDI, but covering only the univariate FDI problems. This involved studying the means of generating residuals based on IVD-PCA model, and their use in detection. The isolation of faults was performed based on an extended version of the reconstruction principle for VPCA model (Ait-Izem et al., 2014a). Next, we moved to the second model, which is the CPCA, also for the univariate case and following the same steps (Ait-Izem et al., 2014b). We then tried doing the same with the MRPCA model, but this time by including a comparison in terms of FDI performances for the three models (VPCA, CPCA and MRPCA) (Ait-Izem et al., 2015a).

After covering the univariate case, we made a first attempt on the multivariate one using VPCA and CPCA models. This involved the introduction of improved interval statistics, and isolation of faults using the reconstruction principle, (Ait-Izem et al., 2015c). While studying the multivariate case, we noticed the lack of precision of the so-called interval-valued statistics, and decided to develop

a new statistic based on the interval-valued norm, also, including the use of a new model called the CIPCA model (Ait-Izem et al., 2015b).

To investigate the applicability of the proposed FDI strategies in real processes, we used the NASA prognostic center's milling machine data set. We also lead a comparative studies, using Monte-Carlo simulations, between the performances of the proposed interval index and that of the other statistics, based on the CIPCA model, and also in comparison with the conventional PCA (Ait-Izem et al., 2017b). Finally, a global study of all the interval-valued PCA models and FDI statistics was addressed, using Monte-Carlo simulation, with the purpose of finding the best combination of model and statistics to use for FDI. This work included the introduction of the new interval variance of reconstruction error (IVRE) for the determination of the number of PCs to be kept in the IVD-PCA model. Also, an application on a larger scale process (distillation column) is presented (Ait-Izem et al., 2018).

As an attempt to answer the last question, we developed a new dynamic interval-valued PCA scheme for process monitoring. The strategy is based on a combination between MRPCA model and the generalized hebbian algorithm (GHA) learning, in order to update the parameters of the model on-line. A promising work, which was validated using an application on the Tennessee Eastman Process (Ait-Izem et al., 2017a).

## Scientific Production

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2014a). "Fault Detection and Isolation of Uncertain Process Using Interval Principal Component Analysis". In: *International Conference on Technological Advances in Electrical Engineering (ICTAEE-2014)*. Skikda- Algeria.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2014b). "Fault Detection and Isolation Using Interval Vertices Principal Component Analysis". In: *3rd International Conference on Information Processing and Electrical Engineering (ICIPEE-2014)*. Tebessa- Algeria.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2015a). "Fault Detection and Isolation Using Interval Principal Component Analysis Methods". In: *IFAC-PapersOnLine* 48.21. 9th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS 2015, pp. 1402 –1407. DOI: https://doi.org/10.1016/j.ifacol.2015.09.721.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2015b). "Interval PCA Based Fault Detection and Isolation With New Interval SPE Statistic". In: *International Conference on Automatic control, Telecommunication and Signals (ICATS-2015)*. Annaba- Algeria.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2015c). "Vertices and Centers Principal Component Analysis for Fault Detection and Isolation". In: *2nd International Conference on Automation,Control, Engineering and Computer Science (ACECS-2015)*. Sousse- Tunisia: Proceedings of Engineering and Technology (PET).

Ait-Izem, T., M-F. Harkat, F. Kratz, and M. Djeghaba (2017a). "Approche Neuronale d'ACP par Intervalle Appliquèe au Diagnostic". In: *12 ème Congrès International Pluridisciplinaire en Qualité, Sûreté de fonctionnement et Développement durable, (Qualita-2017)*. Bourges- France.

Ait-Izem, T., M-F. Harkat, M. Djeghaba, and F. Kratz (2017b). "Sensor Fault Detection Based on Principal Component Analysis for Interval-Valued Data". In: *Quality Engineering*. DOI: 10.1080/08982112.2017.1391288.

Ait-Izem, T., M-F. Harkat, M. Djeghaba, and F. Kratz (2018). "On the Application of Interval PCA to Process Monitoring: A Robust Strategy for Sensor FDI with New Efficient Control Statistics". In: *Journal of Process Control* 63, pp. 29 –46. ISSN: 0959-1524. DOI: https://doi.org/10.1016/j.jprocont.2018.01.006.

## Thesis Organization

This thesis is divided in two main parts and is organized in four chapters as follows:

1. **PART I - PCA and Application for FDI**
   **Chapter 1: Principal Component Analysis**
   This chapter covers the basic notion of PCA model. It includes a brief introduction on the principle of linear PCA and the procedure for constructing a PCA model, followed by the necessary rules for the determination of the number of retained components in the PCA model. The presented approaches are illustrated through a simulation example.
   **Chapter 2: PCA based Fault Detection and Isolation**
   This chapter is dedicated to the theoretical knowledge of PCA for FDI. Different detection statistics and isolation approaches for use with PCA are presented. Also, the previously presented simulation example is used for illustration of the PCA based FDI scheme.
2. **PART II - PCA for IVD and Application to FDI**
   **Chapter 3: IVD-PCA for FDI**
   This chapter contains a literature review of the most well known PCA approaches for interval-valued data (VPCA, CPCA, MRPCA, CIPCA), which are then dedicated to fault detection and isolation purposes. All the developed theory on interval-valued PCA for FDI, including: definitions of interval-valued statistics, and introduction of new statistics based on the interval squared norm are presented. Extension of the well-known PCA

based reconstruction principle for the interval-valued data and definition of a new criterion for the selection of the number of retained PCs in the IVD PCA model. A simulation example is used to illustrate the presented FDI strategies.

**Chapter 4: Comparative Studies and Applications**

This chapter covers various applications of the proposed FDI strategies based on PCA for interval-valued data. The first section is dedicated to an in-depth study of the interval-valued FDI strategies (univariate and multivariate) for the different IVD PCA models. The study is lead using Monte-Carlo simulation on synthetic data, where a comparison is made in terms of modeling and FDI accuracies. In the next section, the validation of the developed techniques is done using two real application: milling machine data and distillation column process.

At the end of this thesis, conclusion summarizes the contributions of the thesis, and highlights the directions of future works.

# PART 1

# PCA AND ITS APPLICATION FOR FDI

# 1 | Principal Component Analysis



## 1.1 Introduction

Data-driven methods, also known as process history based methods, only require the availability of sufficient data from the process. Various methods have been developed to establish the knowledge database for the underlying system by extracting characteristic features directly from its past performance data. For large-scale processes, such as chemical plants, the development of model-based fault-detection or the gathering of expert knowledge on the process both require a considerable and eventually a too high effort. Then, data driven analysis methods offer an alternative and efficient way. Especially based on methods of multivariate statistical analysis, but mostly using Principal Component Analysis (PCA) and Projection to Latent Structures (PLS) which have received the most attention.

Principal Component Analysis (PCA) is a non-parametric procedure for orthogonal linear transformation of the input data to a new coordinate system, such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. In other words, PCA first selects a normalized direction in $m$-dimensional space ($m$ is the number of columns in the input data,i.e. variables) along which the variance in input data is maximized – this is referred to as the first principal component. It then repeatedly finds other

directions (principal components) in which the variance is maximized. At every step, PCA restricts the search for only those directions that are perpendicular to all previously selected directions. By doing so, PCA aims to reduce the redundancy among input variables. To understand the notion of redundancy, consider an extreme scenario with a data set comprising of two variables, where the first one denotes some quantity expressed in liters, and the other variable represents the same quantity but in gallons. Both these variables evidently capture redundant information, and hence one of them can be removed. In a general scenario, keeping solely the linear combination of input variables would both express the data more concisely and reduce the number of variables. This is why PCA is often used as a dimensionality reduction technique. Of course if, for instance we start out with 2 independent (thus completely unrelated) variables there is no hope of reducing the dimensionality of the problem, and there is no multi-dimensional phenomenon, then all the information will be contained in the sequence of the 2 one-dimensional analyses of the variables.

In this Chapter, we first present the principle of principal component analysis, and how it is used for dimensionality reduction, illustrated with a simulation example. The determination of the PCA model parameters, i.e. the number of components is a critical point. Thus, different choices for criteria to determine the number of components are presented and compared. The optimal choice of components is determined for the simulated example, and the projections onto different sub-spaces are presented and discussed.

## 1.2 Principle of Linear PCA

Suppose that $X = [X_1, X_2, \ldots, X_m]$ is a set of $m$ random variables of $n$ observations, with mean vector $\boldsymbol{M}$ and variance-covariance matrix $\Sigma$. The original data matrix $X$ is usually of $rank = min(n, m)$. We want to bring that rank down using an approximation of $X$ by a matrix of lower rank, say $\ell$, where $\ell < rank(X)$. In order to loose as little as possible of the information we use a decomposition of $X$ that is equal to the sum of $\ell$ matrices of rank 1, their relative importance is measured by what is called eigenvalues.

So, our aim is then to define $m$ linear combinations of $X$ that represent the information in $X$ more parsimoniously. Specifically, we need to find $\mathbf{p}_1, \ldots, \mathbf{p}_m$ such that $\mathbf{p}_1^T X, \ldots, \mathbf{p}_m^T X$ gives the same information as $X$.

$$\Sigma = \frac{1}{n-1} X^T X \tag{1.1}$$

The covariance matrix $\Sigma$ is of dimension $m \times m$, real, and symmetric. The diagonal elements are the variances of the individual random variables, while the off-diagonal elements are their covariances. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ be the $m$ roots (eigenvalues) of the matrix $\Sigma$, and let $\mathbf{p}_1, \ldots, \mathbf{p}_m$ be the corresponding eigenvectors. We have to choose an eigenvector $\mathbf{p}_i$ so that $\mathbf{p}_i^T \mathbf{p}_i =$

$1, (i = 1, \ldots, m)$, i.e., a normalized eigenvector. Then, $\mathbf{p}_i^T X$ is the $i^{th}$ principal component of the random variables in $X$. The sample covariance matrix $\Sigma$ of data can then be represented by:

$$\Sigma = [\mathbf{p}_1, \ldots, \mathbf{p}_m] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{bmatrix} \begin{bmatrix} \mathbf{p}_m^T \\ \vdots \\ \mathbf{p}_m^T \end{bmatrix} \tag{1.2}$$

Here are some properties:

1. $Var(\mathbf{p}_i^T X) = \mathbf{p}_i^T \ \Sigma \mathbf{p}_i = \lambda_i$
   We know that $\Sigma \mathbf{p}_i = \lambda_i \mathbf{p}_i$, because $\mathbf{p}_i$ is the eigenvector for $\lambda_i$; thus, $\mathbf{p}_i^T \ \Sigma \mathbf{p}_i = \mathbf{p}_i^T \lambda_i \mathbf{p}_i = \lambda_i$. In other words, the variance of the $i^{th}$ principal component is $\lambda_i$, the $i^{th}$ eigenvalue.
2. $\mathbf{p}_i$ and $\mathbf{p}_j, (j = 1, \ldots, m)$ are orthogonal, i.e., $\mathbf{p}_i^T \mathbf{p}_j = 0$
3. $\text{Cov}(\mathbf{p}_i^T X, \mathbf{p}_j^T X) = \mathbf{p}_i^T \ \Sigma \mathbf{p}_j = \mathbf{p}_i^T \lambda_j \mathbf{p}_j = \lambda_j \mathbf{p}_i^T \mathbf{p}_j = 0$
4. $\text{Tr}(\Sigma) = \lambda_1 + \cdots + \lambda_p =$ the sum of variances for all $m$ principal components, and for the variables $X_1, \ldots, X_m$. Also, the importance of the $i^{th}$ principal component is $\lambda_i / \text{Tr}(\ \Sigma)$ , which is equal to the variance of the $i^{th}$ principal component divided by the total variance in the system of the $m$ random variables, $X_1, \ldots, X_p$; it is the proportion of the total variance explained by the $i^{th}$ component.

To explain why an eigen-decomposition is used for PCA, we recall that PCA is a technique to reduce dimension by:

— Taking linear combinations of the original variables.

— Each linear combination explains the most variance in the data it can.

— Each linear combination is uncorrelated with the others

Or, in mathematical terms:

— For $T_j = \mathbf{p}^T X$ (linear combination for jth component)

— For $i < j, Var(T_i) < Var(T_j)$ (first components explain more variation)

— $\mathbf{p}_i^T \mathbf{p}_j = 0$ (orthogonality)

Finding linear combinations that satisfy these constraints leads us to successive optimizations problems based on Lagrange multipliers, (Jolliffe, 2002) : (maximizing variance) constrained such that $P^T P = 1$ for coefficients $P$ (to prevent the case when variance could be infinite) and constrained to make sure the coefficients are orthogonal. This ultimately leads us at the end to eigenvalues and eigenvectors.

So, Given that $P = [\mathbf{p}_1, \ldots, \mathbf{p}_m]$ is the global eigenvector matrix of data matrix $X$, and $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$ contains the associated eigenvalues positioned diagonally, PCA can be viewed as a linear mapping of the form:

$$T = P^T X \tag{1.3}$$

Where $T \in \mathbb{R}^{n \times m}$ is the principal component matrix, which is the generated new set of variables out of the original variables $X$. The eigenvectors $P$ are the coefficients for the linear transformation. The projection can be reversed back with:

$$X = TP^T \tag{1.4}$$

For a basic geometric interpretation of principal components, suppose we have two variables, $X_1$ and $X_2$, that are centered at their respective means. In Figure 1.1, the ellipse delimits a population of sample points. The first principal component is a line through the widest part of the ellipse and the second component is the line that goes through the less wide part of the ellipse. Or, we take our original frame on the two $X_1$ and $X_2$ axes and perform a transformation around the origin to get a new set of axes according to the computed eigenvectors $P$. The best-fit line (or axis in this case) corresponds to the first principal component which goes through the wide part of points. In PCA, the more correlated the original data, the better this line will explain the actual values of the observed measurements, this line will best explain all the observations with minimum residual error. In other words, the line goes in the direction of maximum variance of the projections.

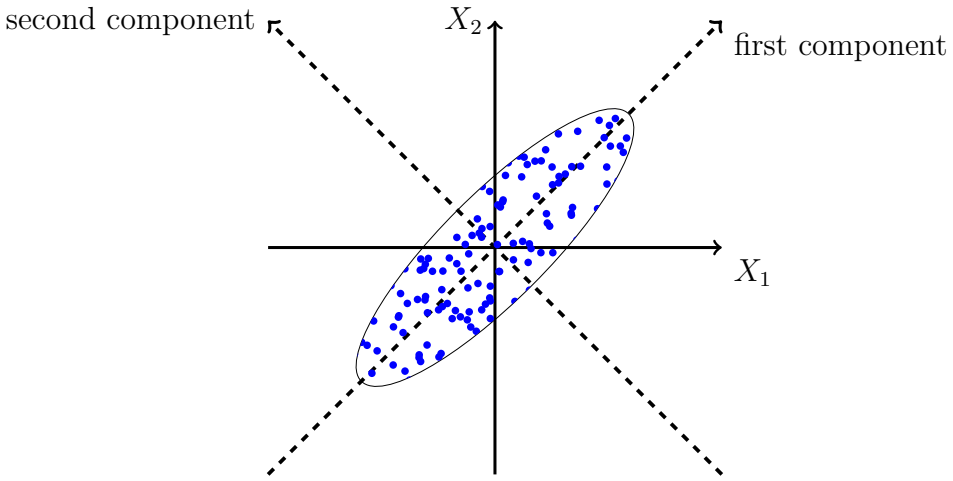



FIGURE 1.1 – Geometrical Representation of Linear PCA

The principal components (PC's) are sorted by descending order of the eigenvalues, which are equal to the variances of the components. Now if the first few $\ell$ PC's account for most of the variation of $X$, then we might interpret these components, obtained via the first $\ell$ eigenvectors $P_\ell = [\mathbf{p}_1, \ldots, \mathbf{p}_\ell]$, as "factors" underlying the whole set $X_1, \ldots, X_m$. This is, in other words, the basis of principal component analysis. The mapping is then performed from $\mathbb{R}^m$ to a lower dimensional space $\mathbb{R}^\ell$.

Thus, further multiplying the PC's by the corresponding principal axes $P_\ell, \ell < m$ yields matrix $\hat{X}$ that has the original $n \times m$ size but is of lower rank (of rank $\ell$). This matrix $\hat{X}$ provides an estimate of the original data from the first $\ell$ PC's and has the lowest possible estimation error. Hence, The transformation matrix $P$ can then be decomposed in two parts, as presented in Figure 1.2. That is, matrix

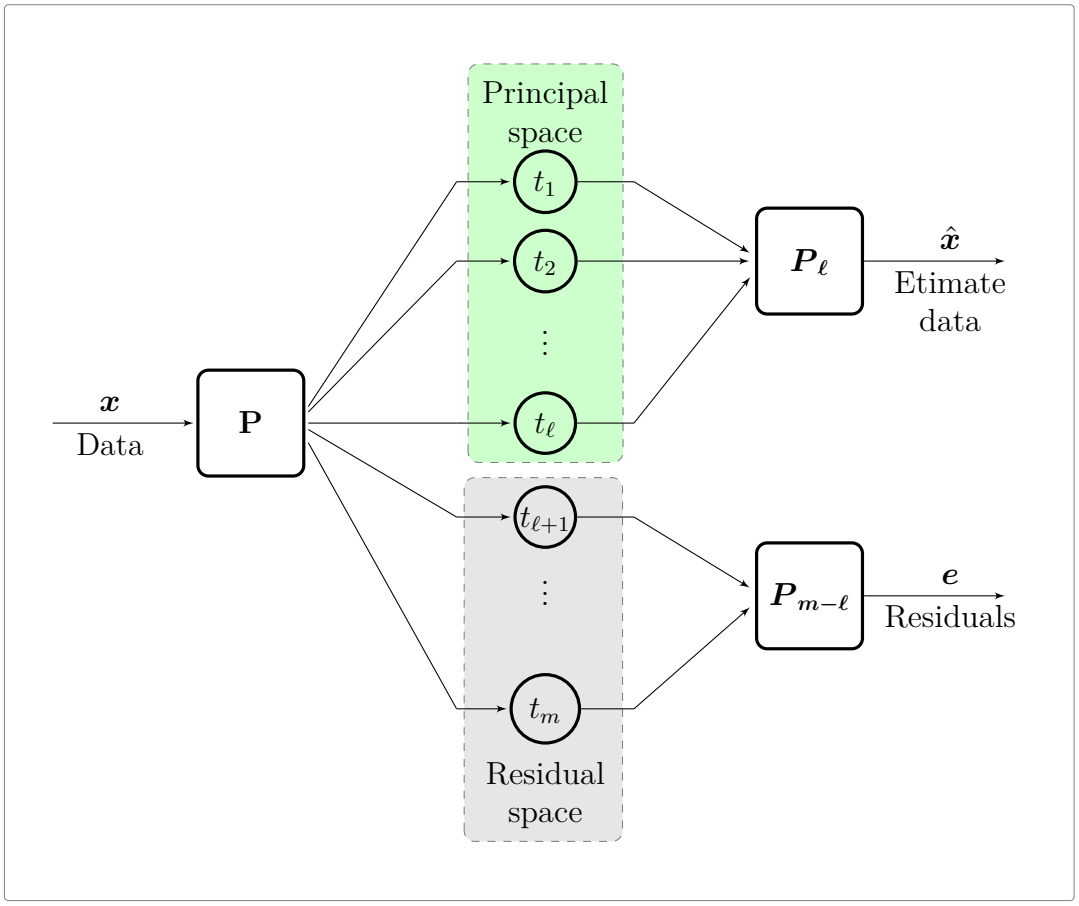



FIGURE 1.2 – Mapping and Inverse Mapping Using Linear PCA

$P_\ell \in \mathbb{R}^{m\times\ell}$ generated by choosing $\ell$ eigenvectors or columns of $P$ corresponding to $\ell$ largest eigenvalues, which transforms the space of the measured variables into the reduced dimension space $\hat{T}$, or principal components, and allows to obtain the estimate reduced rank matrix $\hat{X}$ via inverse mapping. And matrix $P_{m-\ell} \in \mathbb{R}^{m\times(m-\ell)}$ generated by choosing the last $(m-\ell)$ eigenvectors or columns of $P$ allows the projection in the residual sub-space $\tilde{T}$, which allows to obtain the estimation error or residuals $E$, also denoted $\tilde{X}$.

$$\Sigma = \begin{bmatrix} P_\ell & P_{m-\ell} \end{bmatrix} \begin{bmatrix} \Lambda_\ell & 0 \\ 0 & \Lambda_{m-\ell} \end{bmatrix} \begin{bmatrix} P_\ell^T \\ P_{m-\ell}^T \end{bmatrix} \tag{1.5}$$

$$\hat{T} = X\,P_\ell \tag{1.6}$$

$$\tilde{T} = X\,P_{m-\ell} \tag{1.7}$$

$$\hat{X} = X\,P_\ell\,P_\ell^T = X\,C_\ell \tag{1.8}$$

$$E = X - \hat{X} = X\,(I - C_\ell) \tag{1.9}$$

The formulas in 1.6, 1.7, 1.8 and 1.9 are given for the whole population, i.e. for global data matrix $X$, but everything translates more-or-less directly for a sample vector $\mathbf{x}$.
With all its properties, including the powerful capability of analysing correlation between variables, PCA is mostly used as a tool in exploratory data analysis, but also for making predictive models for systems based on their history. In a practical situation, if we consider that the different variables are the sensors of a given process, and the different values of these variables are the corresponding sensor measurements, we can qualify PCA as a modelling approach for the process through it's historical data, i.e. sensor measurements. PCA is considered as one of the most famous methods in the area of latent variable modelling. This statistical model have shown to be a very powerful tool in dealing with complex process data, especially chemical processes. Various uses of such model include process optimization and control, predictive modelling, and process monitoring. This later is the main subject of this thesis.

### 1.2.1 Pre-processing of Data

Let us consider the data matrix $X^g \in \mathbb{R}^{n\times m}$ gathered from a process in normal operating conditions (NOC), that is, under safe operating conditions or when there is no fault in the system. Given that the process is equipped with $j = 1, \ldots, m$ sensors, and the measurement are obtained at each time sample $k$, the collected initial data matrix $X^g$ is given by:

$$X^g = \begin{pmatrix} x_1(1) & x_2(1) & \ldots & x_m(1) \\ x_1(2) & x_2(2) & \ldots & x_m(2) \\ \vdots & \vdots & & \vdots \\ x_1(k) & x_2(k) & \ldots & x_m(k) \\ \vdots & \vdots & & \vdots \\ x_1(n) & x_2(n) & \ldots & x_m(n) \end{pmatrix} \tag{1.10}$$

There have to be no missing data in the used set, and missing values should be excluded or estimated using dedicated algorithms.

It is important to point out that PCA assumes approximate normality of the input data distribution. More precisely, PCA assumes that the distribution of the data to explain can be described by a mean (of zero) and variance alone. This formalism is known as : mean and variance are sufficient statistics. Or, the only zero-mean probability distribution that is fully described by the variance is the Gaussian distribution. In order for this assumption to hold, the probability distribution of input data must be Gaussian, and deviations from a Gaussian could invalidate this assumption. However, PCA may still be able to produce a good low dimensional projection of the data even if the data isn't normally distributed.

Another important matter to discuss before going any further is the fact that the data matrix $X^b$ must be centred and scaled to zero mean and unit variance prior to using PCA. The reason is that variables may all be on completely different scales, and thus the comparison of their relative importance is then impossible, as well as the creation of linear combinations between them. This problem is solved by doing what we call "Normalization" of data, i.e. centring and standardization (scaling) of the data before the actual analysis takes place, in order to avoid unwanted differences between the different variables. Thus, each variable $X_j$ of the new normalized matrix $X$ is given by:

$$X_j = \frac{X_j^b - M_j}{\sigma_j} \tag{1.11}$$

Where $X_j^b$ is the $j$-th column of matrix $X^b$, and $M_j$ is its mean defined as:

$$M_j = \frac{1}{n}\sum_{k=1}^{n} x_j^b(k) \tag{1.12}$$

and $\sigma_j$ is its estimate variance calculated as:

$$\sigma_j = \frac{1}{n}\sum_{k=1}^{n}(x_j^b(k) - M_j)^2 \tag{1.13}$$

Normalized matrix $X$ is then given by:

$$X = [X_1, X_2, \ldots, X_m] \tag{1.14}$$

### 1.2.2 Simulation Example

We focus in this section on interpretation of the PCA model that we obtain from an eigenvalue decomposition. For that, let us consider the following example constituted of 6 variables, which are defined at different time samples $k$ by the following:

**Example 1 :**

$$\begin{cases} x_1(k) = 0.5v_1(k) - 1.3\sin(k/N)\cos(k/3) + \varepsilon_1(k),\ v_1(k) \sim N(0,1) \\ x_2(k) = 0.9v_2(k) - 1.2\cos(k/4)^3 e^{(-k/2N)} + \varepsilon_2(k),\ v_2(k) \sim N(0,1) \\ x_3(k) = x_1(k) + x_2(k) + \varepsilon_3(k) \\ x_4(k) = 2x_1(k) + x_3(k) + \varepsilon_4(k) \\ x_5(k) = x_2(k) + x_3(k) + \varepsilon_5(k) \\ x_6(k) = 2x_1(k) + x_2(k) + \varepsilon_6(k) \end{cases} \tag{1.15}$$

Where $v_1(k)$ and $v_2(k)$ are realizations randomly generated from a normal distribution with zero mean and unit variance, i.e. $\sim \mathcal{N}(0,1)$, and $\varepsilon_j(k), j = 1, \ldots, 6$

represents Gaussian noise in the range $[-0.2, 0.2]$ added to the measurements. The rest of variables is constituted by different linear analytic redundancy relations. The raw data variables are presented in Figure 1.3.

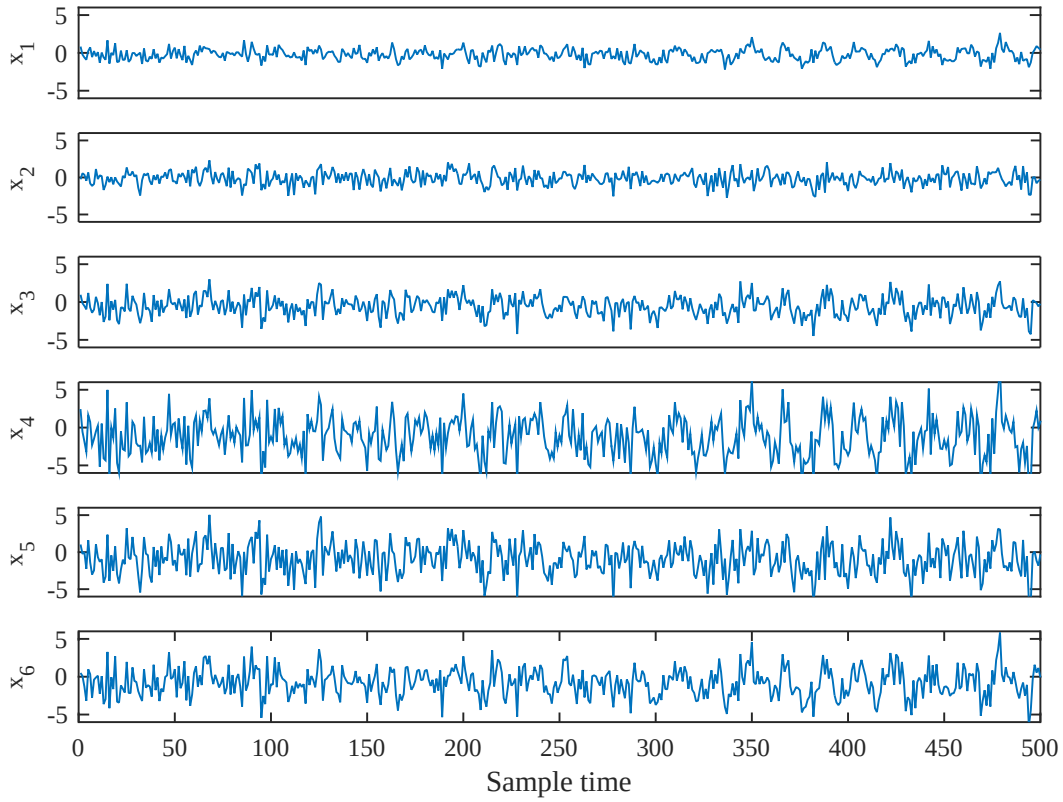



FIGURE 1.3 – Different variables of raw data simulated in example 1

The first step is to center and standardize the data. The box plots in Figure 1.4 show the tendency of raw data and normalized data. We notice that normalizing removes any differences between the different variables of the data.

The covariance matrix of data is then computed, and is given by

$$\Sigma = \begin{pmatrix} 1.0000 & 0.0246 & 0.6207 & 0.9096 & 0.3961 & 0.8380 \\ 0.0246 & 1.0000 & 0.7340 & 0.3798 & 0.8852 & 0.5279 \\ 0.6207 & 0.7340 & 1.0000 & 0.8698 & 0.9376 & 0.8828 \\ 0.9096 & 0.3798 & 0.8698 & 1.0000 & 0.7033 & 0.9432 \\ 0.3961 & 0.8852 & 0.9376 & 0.7033 & 1.0000 & 0.7726 \\ 0.8380 & 0.5279 & 0.8828 & 0.9432 & 0.7726 & 1.0000 \end{pmatrix} \tag{1.16}$$

Next, we perform an eigen-decomposition of $\Sigma$, thus obtaining the eigenvalues and their corresponding eigenvectors. Matrices $\Lambda$ and $P$ are sorted in decreasing

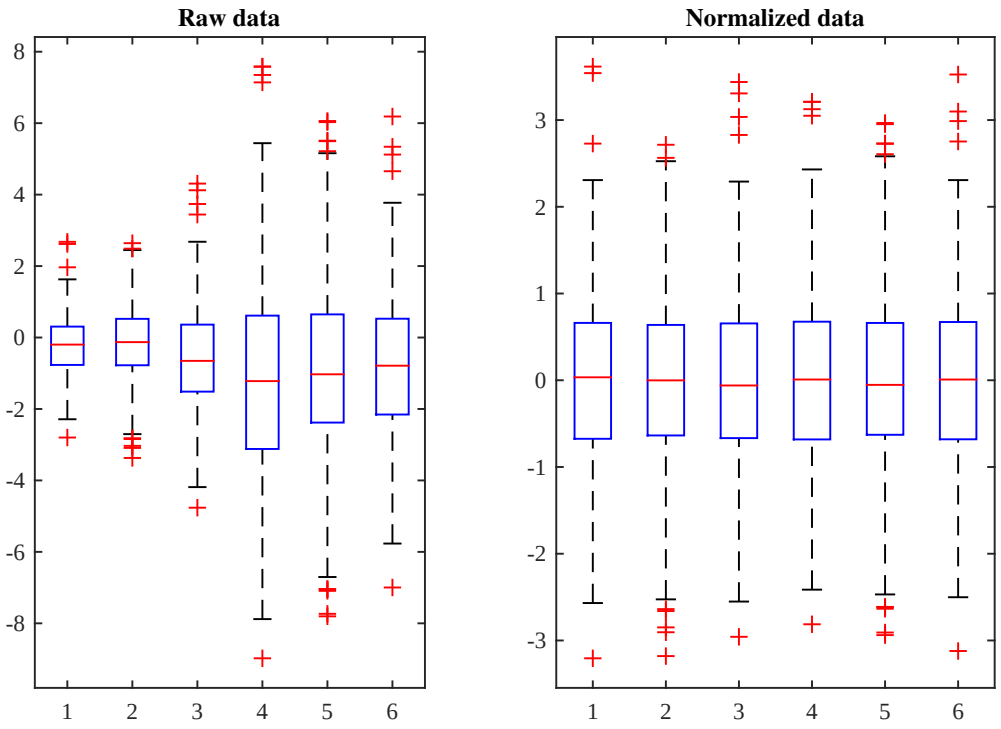



FIGURE 1.4 – Box plots of different variables of data before and after normalization

order, and are given by:

$$\Lambda = \begin{pmatrix} 4.5525 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.3011 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0898 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0250 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0204 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0112 \end{pmatrix} \tag{1.17}$$

$$P = \begin{pmatrix} 0.3471 & 0.5830 & 0.1230 & 0.2316 & -0.3793 & -0.5718 \\ 0.3155 & -0.6346 & 0.4290 & -0.3065 & -0.4196 & -0.2091 \\ 0.4561 & -0.1202 & -0.5242 & -0.3609 & 0.4551 & -0.4066 \\ 0.4373 & 0.2957 & -0.2299 & -0.3830 & -0.3799 & 0.6144 \\ 0.4221 & -0.3579 & -0.2847 & 0.7558 & -0.0781 & 0.1881 \\ 0.4500 & 0.1654 & 0.6262 & 0.0653 & 0.5679 & 0.2262 \end{pmatrix} \tag{1.18}$$

#### 1.2.2.a Interpretation of principal components and residuals

According to PCA, the principal components are given by the following linear relations in terms of eigenvectors :

$$\begin{cases} \mathbf{t}_1 = 0.34\mathbf{x}_1 + 0.31\mathbf{x}_2 + 0.45\mathbf{x}_3 + 0.43\mathbf{x}_4 + 0.42\mathbf{x}_5 + 0.45\mathbf{x}_6 \\ \mathbf{t}_2 = 0.58\mathbf{x}_1 - 0.63\mathbf{x}_2 - 0.12\mathbf{x}_3 + 0.29\mathbf{x}_4 - 0.35\mathbf{x}_5 + 0.0118\mathbf{x}_6 \\ \mathbf{t}_3 = 0.12\mathbf{x}_1 + 0.42\mathbf{x}_2 - 0.52\mathbf{x}_3 - 0.22\mathbf{x}_4 - 0.28\mathbf{x}_5 + 0.62\mathbf{x}_6 \\ \mathbf{t}_4 = 0.23\mathbf{x}_1 - 0.30\mathbf{x}_2 - 0.36\mathbf{x}_3 - 0.38\mathbf{x}_4 + 0.75\mathbf{x}_5 + 0.062\mathbf{x}_6 \\ \mathbf{t}_5 = -0.37\mathbf{x}_1 - 0.41\mathbf{x}_2 + 0.45\mathbf{x}_3 - 0.37\mathbf{x}_4 - 0.07\mathbf{x}_5 + 0.56\mathbf{x}_6 \\ \mathbf{t}_6 = -0.57\mathbf{x}_1 - 0.20\mathbf{x}_2 - 0.40\mathbf{x}_3 + 0.61\mathbf{x}_4 + 0.18\mathbf{x}_5 + 0.22\mathbf{x}_6 \end{cases} \tag{1.19}$$

In order to visualize the impact of the PCA projection, i.e. principal components presented in Figure 1.5, in redefining the data in a new set of axis, let us compute the coefficients of correlation between each variable $(\mathbf{x_1}, \ldots, \mathbf{x_6})$ and each component $(\mathbf{t_1}, \ldots, \mathbf{t_6})$, which are presented in table 1.1. The correlation coefficient of two variables $A$ and $B$ is a measure of their linear dependence. So, if each variable has $n$ observations (samples), then the correlation coefficient is defined as:

$$r(A,B) = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{A_i - \mu_A}{\sigma_B} \right) \left( \frac{B_i - \mu_B}{\sigma_B} \right) \tag{1.20}$$

where $\mu_A$ and $\sigma_A$ are the mean and standard deviation of variable A, respectively, and $\mu_B$ and $\sigma_B$ are the mean and standard deviation of variable B. Alternatively, the correlation coefficient can be defined in terms of the covariance matrix of A and B, as:

$$r(A,B) = \frac{Cov(A,B)}{\sigma_A \sigma_B} \tag{1.21}$$

The values of the coefficient $r$ can range from -1 to 1, with -1 representing a direct negative correlation, 0 representing no correlation, and 1 representing a direct positive correlation.

Let us now examine the magnitudes of the coefficients, the larger the absolute value of the coefficient, the more important the corresponding variable is in calculating the component, i.e. the more of information from that variable is contained in that component. The sign of the coefficients reveals the direction of the projection.

| | Principal components | | | | | |
|---|---|---|---|---|---|---|
| Variables | $\mathbf{t_1}$ | $\mathbf{t_2}$ | $\mathbf{t_3}$ | $\mathbf{t_4}$ | $\mathbf{t_5}$ | $\mathbf{t_6}$ |
| $\mathbf{x_1}$ | 0.7062 | 0.6563 | 0.1076 | 0.0514 | 0.0062 | -0.0818 |
| $\mathbf{x_2}$ | 0.6879 | -0.7526 | 0.1333 | -0.0780 | -0.0980 | 0.0349 |
| $\mathbf{x_3}$ | 0.9726 | -0.2018 | -0.1039 | -0.0807 | 0.0685 | -0.0007 |
| $\mathbf{x_4}$ | 0.9293 | 0.2868 | -0.0093 | -0.0659 | -0.0218 | 0.0805 |
| $\mathbf{x_5}$ | 0.9005 | -0.4686 | -0.0353 | 0.0693 | -0.0166 | 0.0749 |
| $\mathbf{x_6}$ | 0.9562 | 0.1471 | 0.2516 | 0.0255 | 0.1217 | 0.0408 |

TABLE 1.1 – Correlation levels between variables and principal components

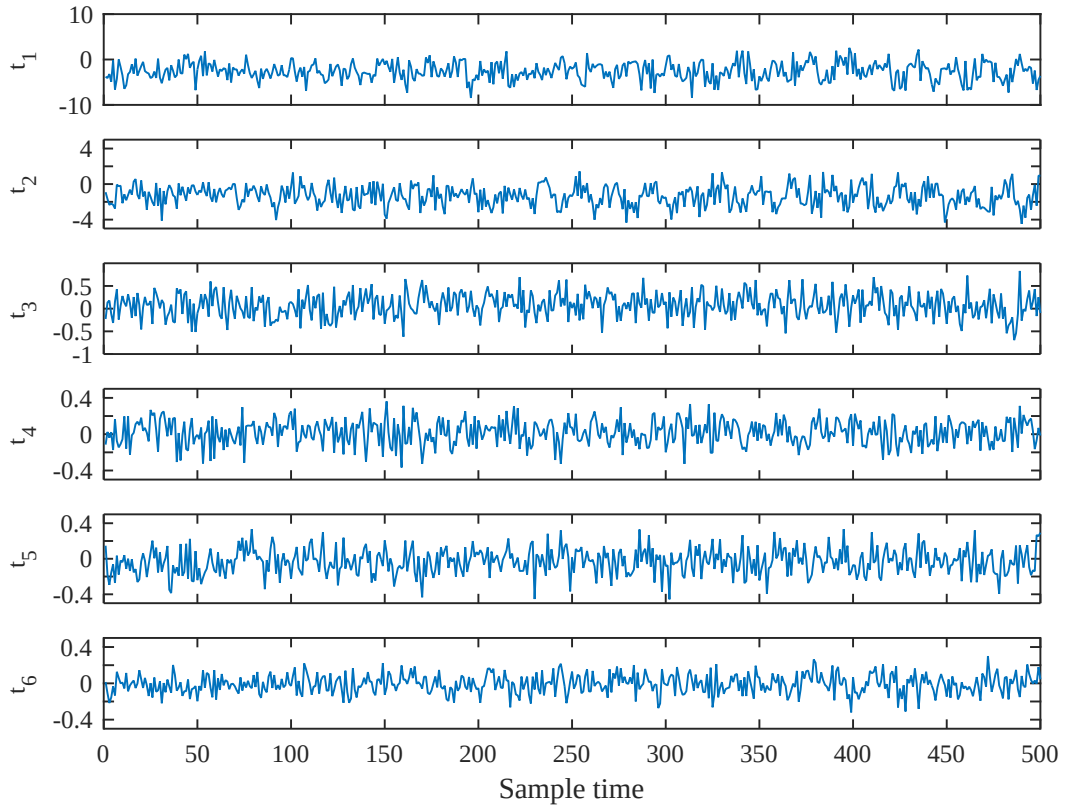



FIGURE 1.5 – Time evolution of principal components

We can clearly notice that the first two components $\mathbf{t_1}$ and $\mathbf{t_2}$ account for most of the information contained in the original covariance matrix $X$, this is further shown in Figure 1.5 which represents the evolution of principal components in time. In the other hand, the four remaining components contain very few information on the data, and are considered useless. Or, these components are not much needed to estimate the data. The loss of information between the original data matrix $X$ and its estimate from the first chosen components is defined as the distance between them (difference) as in 1.9, and is known as residual distance, estimation error, or simply residuals. The question that logically comes after is what are the method to compute the number of components (or axes, or factors, or dimensions) to retain in order to have the most accurate model ?

## 1.3 Determining the Number of Useful Components

While conducting PCA on a set of variables, the number of extracted components is equal to the number of variables being analysed, necessitating that we decide how many of these components are truly meaningful and worthy of being kept. The expectation is that the first few components will account for meaningful information (high variance), and that the last components will tend to account for less meaningful

information (low variance). Therefore, The next step of the analysis is to determine how many useful components should be retained for the PCA model. However, The most common problem in the use of PCA is specifically the determination of this number of components to retain. Several comparisons have been lead in literature that focus on the effectiveness of various criteria (Zwick and Velicer, 1986)(Valle et al., 1999), such as: parallel analysis (Horn, 1965), minimum average partial (MAP) (Velicer, 1976), scree test (Cattell, 1966), Chi-square test (Bartlett, 1950), eigenvalue greater than one rule (Kaiser, 1960), cross-validation (Wold, 1978), the cumulative percentage of variance, and the variance of reconstruction error (VRE) (Dunia and Joe Qin, 1998). All methods give fairly good results, and have each advantages and drawbacks, it is then advisable to carefully choose the criterion that is most suitable for the desired application. We describe in what follows some of the most common criteria that may be used in choosing the number of components.

### 1.3.1 The Eigenvalue Greater Than One Rule

In principal component analysis, one of the most commonly used criteria for solving the number of components problem is the eigenvalue-one criterion, also known as the Kaiser rule (Kaiser, 1960). This approach simply retains the components corresponding to an eigenvalue greater than 1 ($\lambda_i > 1$). The logic behind this criterion is straightforward, the more variables that load (or project) onto a particular component (i.e., have a high correlation with the component), the more important the component is in summarizing the data. In other words, the eigenvalue is an index indicating how good a component is in summarizing the data. Recalling that data matrix is of dimension $n \times m$ , and that we have $m$ eigenvalues when diagonalizing its covariance, we can be sure that an eigenvalue greater than 1.0 means that the corresponding component contains at least the same amount of information as a single variable. Thus, components with eigenvalues less than 1.0 are viewed as trivial, and are not retained. This rule has been often criticized that it sometimes results in the selection of too many components. The general advice is to use this method as a starting point, try another method, and then select a number of components such that the resulting components seem valid.

### 1.3.2 Cumulative Percentage of Variance

PCA is all about variance of data, high variance means signal or meaningful information, and low variance means noise. This means that PCA filters data by estimating the data with only high variance components. The optimality criterion for component determination used in this case is an information criterion. Recalling that principal components are ordered according to decreasing variance, we have to retain enough PC's so that their cumulated variance (CV) approximates the total variance of the original variables. In practice, we consider the following information ratio:

$$CPV(\ell) = 100 \left( \frac{\sum_{j=1}^{\ell} \lambda_j}{\sum_{j=1}^{m} \lambda_j} \right) \% \tag{1.22}$$

Therefore, we have to choose $\ell$ so that the cumulative percentage of variance $CPV(\ell)$ is sufficiently high. Generally, values in the range 70-80% are considered satisfactory, but for more accurate models higher CPV ratio is often required (over 90%).

### 1.3.3 Cross-Validation Criterion

Another mean of choosing the number of principal components is to fit the model to only part of the available data (training set), and to measure how well models with different numbers of extracted components fit the other part of the data (validation set). In other words, perform a cross-validation. The number of principal components to be retained is then usually one that optimizes some criterion or selection rule. Various choices exist in the litterature, but here we only present the criterion known as the predicted residual sum of squares (PRESS) described by (Wold, 1978). This criterion is given by,

$$PRESS(\ell) = \frac{1}{Nm} \sum_{k=1}^{N} \sum_{i=1}^{m} \left( \hat{x}_i^{\ell}(k) - x_i(k) \right)^2 \tag{1.23}$$

where $N$ is the size of the used validation set. By calculating $PRESS(\ell)$, (for $\ell = 1, \ldots, m$), we pick the first value of $\ell$ that minimizes $PRESS(\ell)$ and consider it as the desired number of components to keep in the model. Drawbacks of this method are high computational cost, as we construct various PCA models for $(\ell = 1, \ldots, m)$. Also, it leads in some cases to extracting too many components, and thus having an overfit model, i.e. a model that matches the training data too well, inducing the loss of the model's predictive ability.

### 1.3.4 Variance of the Reconstruction Error

Perhaps the most suitable and accurate approach for the determination of the PC's for practicel purpose, is the variance of reconstruction error (VRE) introduced in Dunia and Joe Qin, 1998. It's principle consists in estimating a variable from other process variables using the PCA model. The reconstruction accuracy is thus related to the capacity of the PCA model to reveal the redundancy relations among the variables. First of all, we begin by defining the reconstruction principle then we show how to use this principle for determining the accurate number of components for the PCA model.

Let $\mathbf{x}(k)$ be a sample vector of the data matrix $X$ in 1.14, $\mathbf{x}(k)$ is given by:

$$\mathbf{x}(k) = [x_1(k), \ldots, x_m(k)]^T \tag{1.24}$$

We know that an estimation, or reconstruction, of vector $\mathbf{x}(k)$ can be performed via matrix $C_\ell = P_\ell P_\ell^T$. That is one of the basic properties of a PCA model. Now let's choose from $\mathbf{x}(k)$ a sample corresponding to variable $i$, we thus write our vector $\mathbf{x}(k)$ as:

$$\mathbf{x}(k) = [x_1(k) \ldots x_{i-1}(k) \; x_i(k) \; x_{i+1}(k) \ldots x_m(k)]^T \tag{1.25}$$

The highlighted element $x_i(k)$ is the sample $k$ of variable $i$ that we would like to reconstruct. The estimate $\hat{x}_i(k)$ from PCA model can be used as a reconstruction of

$x_i(k)$. However, we need this reconstruction to be based on the linear PCA model established from the rest of variables, meaning all process variables except the $i$-th one, in order to eliminate its influence. The drawback of conventional estimation approach is that the $x_i(k)$ contained in $\mathbf{x}(k)$ is used in the estimate. Therefore, the estimate is somewhat contaminated or affected by $x_i(k)$. To eliminate the effect of the $i^{th}$ sensor or variable, we feedback the estimation of the $i^{th}$ variable $\hat{x}_i(k)$ to the input to replace $x_i(k)$ and iterate until it converges to a value $z_i(k)$ as shown in Figure 1.6. Every iteration through the PCA model is an orthogonal projection to the principal component subspace as explained in (Harkat, 2002). The iteration is defined by (Dunia and Joe Qin, 1998), and can be represented by the following expression:

$$z_i\,(k) = \frac{[c_{-i}^T \quad 0 \quad c_{+i}^T]}{1 - c_{ii}}\mathbf{x}(k) \tag{1.26}$$

Where $z_i(k)$ is the reconstructed value of measurement $x_i(k)$ and $c_{-i}$, $c_{+i}$ denote, respectively, the first $(i-1)$ and last $(m-i)$ elements of the $i$-th column of matrix $C_\ell$.

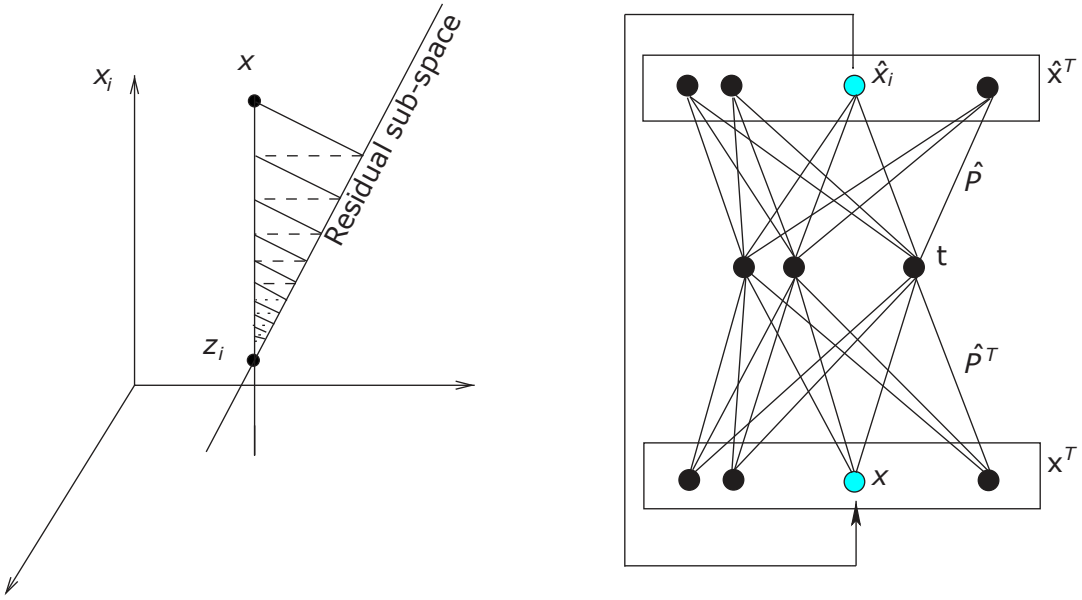



FIGURE 1.6 – Iterative approach for reconstruction

Thus, the new reconstructed vector $\mathbf{x}^{(i)}(k)$ denotes the vector that contains all the elements of $\mathbf{x}(k)$ except the $i$-th element which is replaced by its reconstruction $z_i(k)$ calculated as in 1.26. $\mathbf{x}^{(i)}(k)$ is then given by:

$$\mathbf{x}^{(i)}(k) = [x_1(k) \ldots x_{i-1}(k) \;\; z_i(k) \;\; x_{i+1}(k) \ldots x_m(k)]^T \tag{1.27}$$

**Remark.** *PCA has the ability to "Reconstruct" the original variables from the principal component, or at least from the $\ell$ first ones, using an inverse mapping. However, we'd rather like to refer to this operation as "Estimation" and not reconstruction in order to avoid ambiguity with the reconstruction explained here.*

A more generalized form of the reconstructed vector $\mathbf{x}^{(i)}(k)$ can be defined as:

$$\mathbf{x}^{(i)}(k) = G^{(i)}\mathbf{x}(k) \tag{1.28}$$

The projection matrix $G^{(i)}$ is given by:

$$G^{(i)} = \begin{bmatrix} \xi_1 & \dots & \mathrm{g}_i & \dots & \xi_m \end{bmatrix}^T, \ \mathrm{g}_i^T = \frac{\begin{bmatrix} c_{-i}^T & 0 & c_{+i}^T \end{bmatrix}}{1 - c_{ii}} \tag{1.29}$$

Or, $G^{(i)}$ is expressed in terms of $C_\ell$ as:

$$G^{(i)} = I_m - \xi_j (\xi_j^T C_\ell \xi_j)^{-1} \xi_j^T C_\ell) \tag{1.30}$$

Where $\xi_i = [0 \dots 1 \dots 0]$ is the vector of reconstruction direction with all elements equal to 0 except the $i$-th which value is 1, and $I_m$ is the identity matrix of size $m$.

The variance of reconstruction error quantifies the information lost during the reconstruction process. In fact, there is always a part of the measurement variation that cannot be reconstructed. This part is called the reconstruction error, and it can be computed in the reconstruction direction $\xi_j$ by the difference between the original and reconstructed measures by $\xi_j^T(\mathbf{x}(k) - \mathbf{x}^{(i)}(k))$.

In the same way that the estimation error using PCA model is calculated in 1.9 based on matrix $C_{m-\ell} = I - C_\ell$, the reconstruction error is calculated as:

$$\xi_j^T(\mathbf{x}(k) - \mathbf{x}^{(i)}(k)) = \frac{\tilde{\xi}_j^T \mathbf{x}(k)}{\tilde{\xi}_j^T \tilde{\xi}_j} \tag{1.31}$$

Where $\tilde{\xi}_j = (I - C_\ell)\xi_j$, and $\tilde{\xi}_j^T \tilde{\xi}_j = (1 - c_{ii})$. The variance of reconstruction error is then equal to the variance of the difference between the original and reconstructed variable, and is given by:

$$\rho_i(\ell) = Var\left\{\xi_j^T(\mathbf{x}(k) - \mathbf{x}^{(i)}(k))\right\} = \frac{\tilde{\xi}_j^T \Sigma \tilde{\xi}_j}{(\tilde{\xi}_j^T \tilde{\xi}_j)^2} \tag{1.32}$$

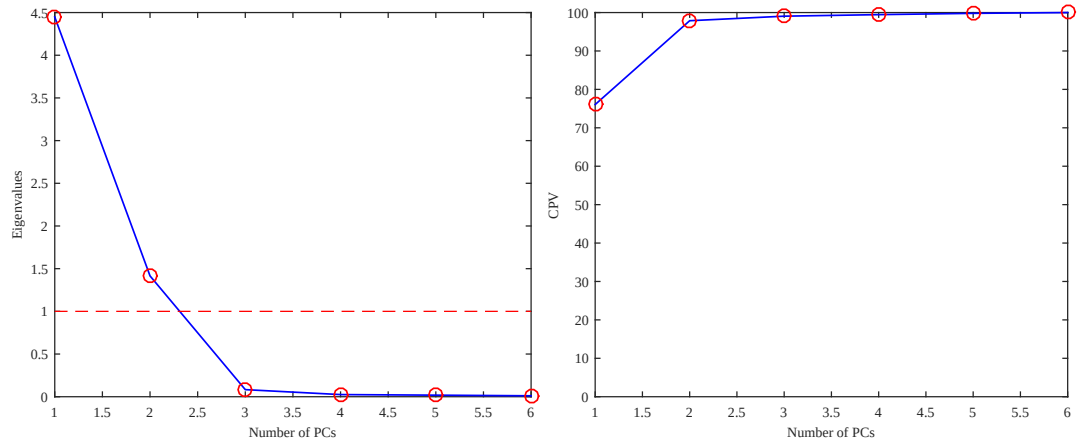
Thus, the number of components to be retained according to the the variance or reconstruction error criterion (VRE), is the one that minimizes the global reconstruction error given by:

$$VRE(\ell) = \sum_{j=1}^{m} \frac{\rho_i(\ell)}{\xi_j^T \Sigma \xi_j} \tag{1.33}$$

In other words, the criterion defined in 3.48 must ensure that the chosen number of components is the optimal $\ell_{opt}$ that have the best reconstruction possible .i.e the lowest reconstruction error.

$$\ell_{opt} = \arg\min_{\ell} \{VRE(\ell)\} \tag{1.34}$$

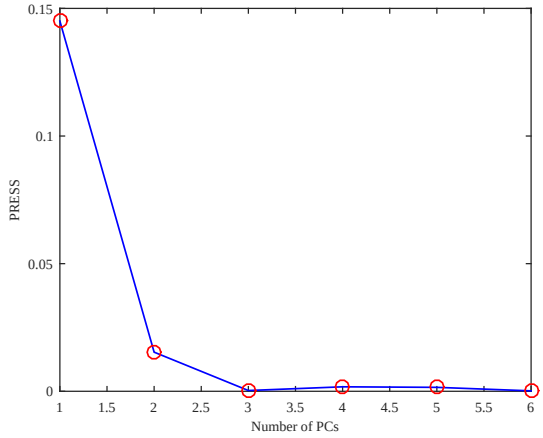
With its ability to eliminate the effects of a variable in a desired direction, the reconstruction of variables principle can serve for various purposes including the use for filling missing values in the data. Applied in process monitoring, and in the case of a sensor problem/fault, the reconstruction technique can be applied to the process to restore in-control operations, i.e. determining optimal replacement values of faulty values in data. The reconstruction of variables is also a powerful method for isolation of detected faults.



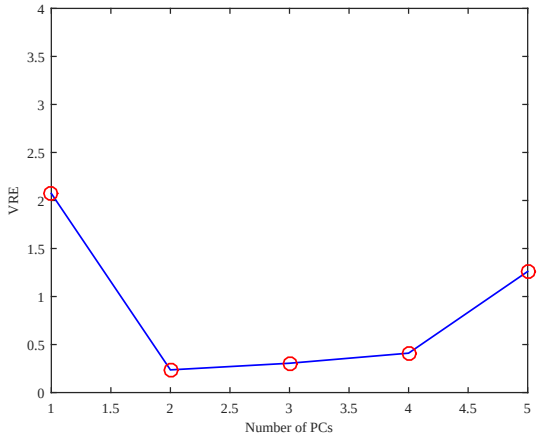

(A) Eigenvalues




(B) CPV




(C) PRESS




(D) VRE

FIGURE 1.7 – Evolution of the different presented criteria in terms of principal components

In order to compare between the different criteria presented in this section, let us discuss the results for the determination of the number of components for the simulation example in 1, which are presented in Figure 4.6. The eigenvalue rule in 1.7a and the variance of reconstruction error in 1.7d give the same number of components to retain, which corresponds to two ($\ell = 2$) components. This resulting number of PC's perfectly matches the results of the correlation study given in table 1.1. In the other hand, the PRESS criterion in 1.7c yields three components, where the cumulative percentage of variance in 1.7b depends on the required variance to obtain (1 component for more than 80% and 2 components for 98%). The VRE criterion is by far the most accurate criterion for best results in a practical situation, that is because it minimizes the error in variable reconstruction, and even though the eigenvalue rule gives the same results, it tends to perform poorly in practical situations.

So, if we consider that the number of components retained for the example 1 is $\ell = 2$, we can estimate the reduced rank matrix $\hat{X}$ from the first two components by reverse mapping as in 1.8. The estimated variables are presented in Figure 1.8, we can clearly see that the PCA model yields a good estimation of data based on only the first two principal components.

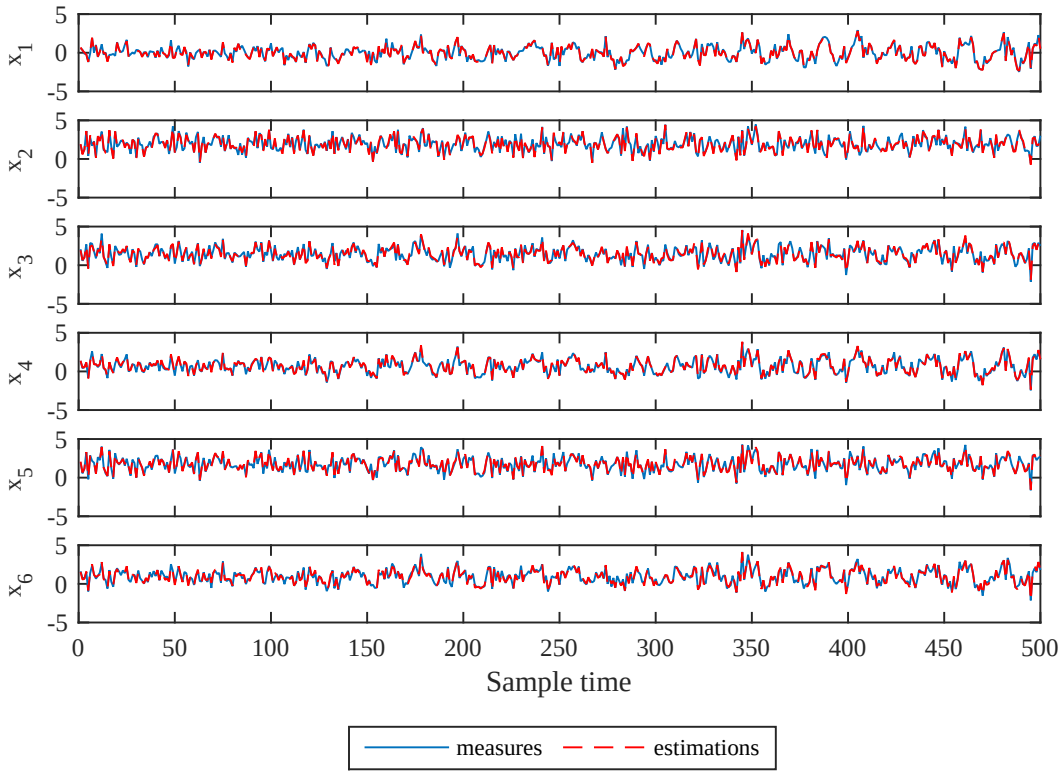



FIGURE 1.8 – Comparison between measures and estimates of example 1 using PCA model for $\ell = 2$ components

## 1.4 Conclusion

In this chapter, we presented the basic principle of process modeling using linear principal component analysis approach. The different steps to follow in order to construct an adequate PCA model are presented. We also presented interpretations of the different projection onto sub-spaces, i.e. principal components and residuals, and their relation to the initial data. The identification of the PCA model requires the determination of the number of principal components (PCs). Several approaches are used in the literature to identify the optimal number of components to retain for the PCA model, we presented four most known criteria: the eigenvalue one rule, the cumulative percentage of variance, the cross-validation criterion, and the variance of reconstruction error criterion. This last criterion is very interesting for diagnostic purposes, because it allows the exploitation of the redundancies that exist between the different variables, and ensures a minimum reconstruction error by the model. A simulation example is used to illustrate the use and accuracy of these critera for the determination of the PCA model's structure.

# 2 | PCA based Fault Detection and Isolation



## 2.1 Introduction

The classical approaches to supervision is to check the limits of single variables based on univariate statistical techniques and alarm the operators in case of fault occurrence. However, these approaches determine limits for each observation variable without using any information from the other variables. Because process data are in general serially correlated, and especially in chemical complex processes, classical univariate approaches lack sensitivity to many faults occurring in such processes. The need to handle these correlations between variables has led to the development and employment of process monitoring statistics based on multivariate approaches for monitoring complex processes.

Multivariate statistical methods, such as Principal Component Analysis (PCA), have been widely applied in the process industry. Due to their capability of dimension reduction and the non-necessity of an explicit input-output model, these methods have been used for performance monitoring in high-dimensional complex systems that have correlated inputs/outputs. In (Kresta et al., 1991), a multivariate monitoring procedure analogous to the univariate Shewart Chart was proposed, in which methods are employed to compress available measurements into a low-dimension space while

retaining most of the information. A considerable number of researchers have centred their interest in the application of multivariate statistical methods for process modelling and fault detection in various applications, and especially using PCA (MacGregor and Kourti, 1995), (Doymaz et al., 2001), (Qin, 2003), (Tien et al., 2004), (Wang and Xiao, 2004), (Harkat et al., 2006), (Sharmin et al., 2008), (Harrou et al., 2013), (Nasri et al., 2015).

In general, the procedure begins by establishing a principal component plan or model under normal operations, and then an index is calculated to evaluate the process performance. Among the commonly used indices are the Squared Prediction Error (SPE) that calculates the perpendicular distance between a new observation and the established estimate using principal components, and the Hotelling's $T^2$ statistic that represents the variability in the principal component subspace (Qin, 2003). Other indices include the Squared Weighted Error (SWE) and the combined index. Based on the calculated index, a proper threshold can be established for fault detection, where exceeding values with respect to that threshold are considered as faulty entries. The fault isolation field using PCA has also been explored, and many solution were introduced, including the contribution chart by (MacGregor et al., 1994), the multi-block method (Chen and McAvoy, 1998), the partial PCA (Huang et al., 1999), the reconstruction based isolation (Dunia et al., 1996a), (Dunia et al., 1996b), and many others. The contribution chart determines the contribution from each process variable to the prediction errors, while the multi-block method groups the process variables into several blocks with each corresponding to a specific section of the monitored process. The partial PCA consists in constructing several partial models, where the reconstruction based approach relies on the reconstruction of variables in several directions, both method eliminate the effect of the fault by eliminating its effect, one by constructing partial model, and the other by performing a fault free reconstruction. All the proposed methods demonstrated their capabilities to identify the variables that causes the deviation of the process performance from its normal conditions.

A limitation of PCA is that most real industrial processes operate at a number of different conditions and modes. So, applying classical static PCA approach to such a process can produce excessive number of false alarms or alternatively, missed detection of process faults, which significantly compromises the reliability of the monitoring strategy. Hence, extensive research has been carried out to address this limitation of PCA, and many solutions are available which can be resumed in three classes (De Ketelaere et al., 2015), (Rato et al., 2016), and these are Dynamic PCA (DPCA) (Ku et al., 1995), Recursive PCA (RPCA) (Li et al., 2000), and Moving Window PCA (MWPCA) (Wang et al., 2005). DPCA was developed to handle autocorrelation, whereas RPCA and MWPCA are able to handle non-stationary data. Other adaptations of PCA have been proposed to cope with non-linearities in processes such as the auto-associative neural networks (Kramer, 1991), the principal curves and neural networks approach (Dong and McAvoy, 1996), the method based on input-training neural network (Jia et al., 2000) and the Kernel PCA (Scholkopf et al., 1998). However, despite these existing research efforts, the detection and isolation of sensor faults in real system remains always a challenging problem.

In this chapter, we present different approaches for sensor fault detection and isolation based on static version of PCA. The first section is dedicated to the different statistics, or control charts that are common with all applications of PCA based FDI strategies. That is, the squared prediction error (SPE), the Hotteling's $T^2$, and the squared weighted error (SWE). Also, the threshold calculation for each index is presented.

Next, we present the fault detection strategy based on static PCA, and the procedures for the isolation of faults with particular focus on the reconstruction based approach. Then we illustrate the presented strategy with a simulation example.

## 2.2 Fault detection

The PCA based monitoring scheme, as any SPC scheme, is carried out in two phases. Phase l is dedicated to elaborating a model of the process, where monitoring charts are built according to a set of historical in-control data. Once the performances of the process has been understood and modelled, and the assumptions of its behaviour and process stability are checked, the next phase is engaged. Phase II is the exploitation of the model, the charts are used to monitor the process using new available data from new runs on the process.

One of the most powerful tools in process quality control is the statistical control chart first developed in the 1920's by Walter Shewhart. Multivariate control charts however were introduced in 1947 by Harold Hotelling (Hotelling, 1947). They allow to aggregate information concerning a few process variables on one control chart using some statistic. This statistic is a measure of distance between the values of these variables while taking into account the structure of correlations between variables in the form of covariance matrix. In the case of PCA, it can serve as model established from collected data on a process, where new observation measurements in the testing set can be projected into the lower dimensional space. This new data can then be decomposed into a principal component part and a residual part. Each of these projections can be monitored separately or jointly using control charts, where several Shewhart type statistics can be used for this purpose.

### 2.2.1 Squared Prediction Error (SPE)

A typical statistic to detect abnormal situations in data is the squared prediction error ($SPE$), also known as the $Q$ statistic. This statistic measures the total sum of variations in the residual space, i.e. the portion of the observation space corresponding to the $m-\ell$ smallest eigenvalues. It is defined as the residual distance and is computed as the squared 2-norm of this residuals, or, as its sum of squares. For an observation vector $\mathbf{x}(k)$, the $SPE$ is given by,

$$SPE(k) = \|\mathbf{e}(k)\| = \sum_{i=1}^{m} (e_i(k))^2 \tag{2.1}$$

given that :

$$\mathbf{e}(k) = \mathbf{x}(k) - \hat{\mathbf{x}}(k) = (1 - C_\ell)\mathbf{x}(k) \tag{2.2}$$

A threshold can be applied to define the normal variations of noise for the SPE statistic, and a violation of this threshold would indicate that the noise has significantly changed. i.e. indicating the presence of abnormal situation or fault in the data. The control

threshold for the SPE statistic ($\delta_\alpha^2$), with a significance level $\alpha$, has been approximated by (Jackson and Mudholkar, 1979), as:

$$\delta_\alpha^2 = \theta_1 \left[ \frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{2.3}$$

Where $\theta_i = \sum_{j=\ell+1}^{m} \lambda_j^i$, given that $\lambda_j$ is the $j$-th eigenvalue of covariance matrix $\Sigma$, and $i = 1, 2, 3$. $h_0 = \frac{2\theta_1\theta_3}{3\theta_2^2}$ . $c_\alpha$ is $(1-\alpha)\times 100$ percentile for a standard normal distribution.

Later on, (Nomikos and MacGregor, 1995) demonstrated that the control limit for the SPE can also be calculated from the approximate solutions based on quadratic forms developed in (Box, 1954). The distribution of $SPE$ is approximated in (Box, 1954) by a non-central chi-square ($\chi^2$) distribution where g is the scale parameter and h the freedom degree.

$$\delta_\alpha^2 = g\chi_{h,\alpha}^2 \tag{2.4}$$

(Nomikos and MacGregor, 1995) used this approach and proceeded to estimate the parameters of the non-central chi-square ($\chi^2$), which are computed in terms of the mean and the variance of the $SPE$ calculated under NOC. Thus, $g$ and $h$ for the $SPE$ threshold in 2.4 are given by:

$$g = \frac{b}{2a} \tag{2.5}$$

$$h = \frac{2a^2}{b} \tag{2.6}$$

where $a$ is the estimate mean of $SPE$ and $b$ is its estimate variance.

### 2.2.2 Hottelling $T^2$ statistic

Another common statistic is the Hotteling $T^2$ statistic. This statistic provides an indication of unusual variability within the principal subspace. It is calculated as the scaled squared 2-norm of first principal components vector $\hat{\mathbf{t}}(k)$. Or, the value of one sample is equal to the sum of squares of the adjusted (unit variance) projections on each of the principal components, and is given by:

$$T^2(k) = \hat{\mathbf{t}}(k)^T \Lambda_\ell^{-1} \hat{\mathbf{t}}(k) = \sum_{i=1}^{\ell} \frac{t_i^2(k)}{\lambda_i} \tag{2.7}$$

where $\Lambda_\ell$ is the diagonal matrix consisted of the $\ell$ largest eigenvalues $\lambda_i = [\lambda_1, \ldots, \lambda_\ell]$ of $\Sigma$, this is acting as a scaling in respect of the variances for the $\ell$ first components. The process is considered normal for a given significance level $\alpha$ if $T^2 < \tau_\alpha^2$. The control threshold $\tau_\alpha^2$ can be obtained by different approaches, but it is frequent to use the following expression (Tracy et al., 1992):

$$\tau_\alpha^2 = \frac{\ell(N-1)^2}{N(N-\ell)} F_{\ell,(N-\ell),\alpha} \tag{2.8}$$

where $F_{\ell,(N-\ell),\alpha}$ is the $(1-\alpha)\times 100$ percentile of the Fisher distribution with $\ell$ and $(N-\ell)$ degrees of freedom. The $T^2$ index can also be well approximated by a Chi-square ($\chi^2$) distribution with $\ell$ degrees of freedom (Qin, 2003),

$$\tau_\alpha^2 \sim \chi_{\ell,\alpha}^2 \tag{2.9}$$

### 2.2.3 Squared Weighted Error (SWE)

Another control statistic is the squared weighted error ($SWE$) which is a symmetric implementation of the $T^2$ statistic in the residual sub-space, it is also know as the Hawkins $T_H^2$ statistic, and is defined as:

$$SWE(k) = \tilde{\mathbf{t}}(k)^T \Lambda_{m-\ell}^{-1} \tilde{\mathbf{t}}(k) \tag{2.10}$$

The process is considered faulty if it exceeds its detection threshold $\epsilon_\alpha^2$ for a given significance level $\alpha$, given by

$$\epsilon_\alpha^2 = \frac{(m-\ell)(N^2-1)}{N(N-m-\ell)} F_{m-\ell,N-m-\ell} \tag{2.11}$$

where $F_{m-\ell,N-m-\ell}$ is the $(1-\alpha)\times 100$ percentile of the Fisher distribution with $m-\ell$ and $(N-m-\ell)$ degrees of freedom. Similarly, the control threshold for the $SWE$ statistic can be approximated by a Chi-square $\chi^2$ with $m-\ell$ degrees of freedom.

$$\epsilon_\alpha^2 = \chi_{(m-\ell),\alpha}^2 \tag{2.12}$$

As in the case of the Hotelling $T^2$ statistic, the $SWE$ statistic is scaled with respect to the $(m-\ell)$ eigenvalues/variances.

### 2.2.4 Fault Detection Scheme

A typical process monitoring scheme contains one or more measures, that are collected by process sensors. These measures in some way represent the state or behavior of the process. The idea is to convert the data collected from the process into a few meaningful measures, and thereby assist the operators in determining the status of the process, and if necessary in diagnosing the faults. So, Assuming that we have collected a large number $n$ of measurements during a training period for a given process, the first step in fault detection scheme is building the model. The establishment of an accurate PCA model is one of the key issues in the design of a PCA fault detection scheme. This model, constructed under NOC, is used for the residual evaluation process. Considering that the process will be monitored by one/several statistics, these need to be computed along with their corresponding control thresholds. The next phase is the exploitation of the constructed model, i.e. fault detection phase. Now if we consider new data available from the process, fault detection is defined as the practice of determining whether these new observations from the process are in control or not. This is done by calculating the new monitoring charts, and controlling them using the computed thresholds during the previous phase. A faulty condition is declared if the computed statistic exceeds its detection threshold.

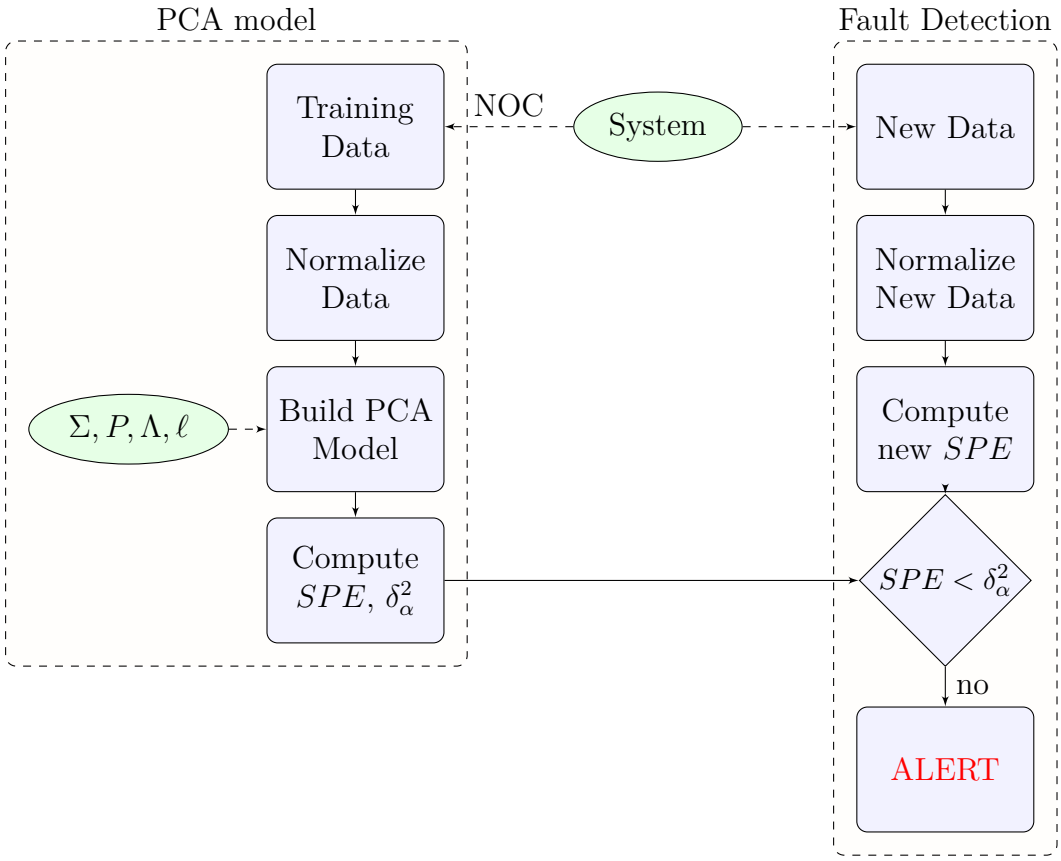



FIGURE 2.1 – Fault Detection Scheme Based on PCA Model

Process fault detection with PCA will proceed as described in Figure 2.1. This is the case where the $SPE$ statistic is used to control the process. In the fault detection phase, first the SPE statistic for the new observation is computed, then the SPE is checked against the corresponding threshold $\delta^2_\alpha$. If the $SPE$ statistic exceeds the threshold ($SPE > \delta^2_\alpha$) then a fault is detected and we proceed to the isolation of the fault. In this case the detected fault has broken the correlation structure of the model. On the contrary, if the SPE is in-control ($SPE < \delta^2_\alpha$) then the process is considered in control. This type of detection scheme is mostly adapted for stationary processes as it is based on a static PCA model. The trained model in this case remains static as new observations are obtained. Therefore, it will not adjust as underlying parameter values change, i.e. changes in mean and variance, and does not attempt to model relationships between observations at different time samples, i.e. changes in autocorrelation structure.

#### 2.2.4.a Generalized indices and detectability conditions

For a generalized fault detection scheme, let us consider that $\Upsilon$ and $\Gamma^2$ are respectively a quadratic statistic and its corresponding threshold, where $\Upsilon$ could be any of the detection statistics presented above. In fact, each of these detection indices is a quadratic distance, and is defined as a squared euclidean norm of a projection of the

sample vector $\mathbf{x}(k)$, given by:

$$\Upsilon(k) = \left\| H^{\frac{1}{2}}\mathbf{x}(k) \right\|^2 = \mathbf{x}(k)^T H^{\frac{1}{2}}\mathbf{x}(k) \tag{2.13}$$

Where $H^{\frac{1}{2}}$ is a positive semi-definite matrix, which is defined according to table 2.1 for different detection statistics, (Mnassri, 2012).

| Detection index $\Upsilon$ | Characteristic Matrix $H^{\frac{1}{2}}$ | Detection threshold $\Gamma^2$ |
|---|---|---|
| $SPE$ | $P_{m-\ell}P_{m-\ell}^T$ | $\delta^2$ |
| $T^2$ | $P_\ell \Lambda_\ell^{\frac{1}{2}} P_\ell^T$ | $\tau^2$ |
| $SWE$ | $P_{m-\ell}\Lambda_{m-\ell}^{\frac{1}{2}}P_{m-\ell}^T$ | $\epsilon^2$ |

TABLE 2.1 – Detection statistics

A generalized control limit $\Gamma^2$ for a significance level $(1-\alpha)$ can be computed based on quadratic form approximations (Box, 1954), for the index $\Upsilon$, as:

$$\Gamma_\alpha^2 = g_\Upsilon \chi^2_{(h_\Upsilon,\alpha)} \tag{2.14}$$

Where $g_\Upsilon$ and $h_\Upsilon$ are the scale parameter and the degrees of freedom computed as follows:

$$g_\Upsilon = \frac{tr\left[(\Sigma H)^2\right]}{tr\left[\Sigma H\right]} \tag{2.15}$$

$$h_\Upsilon = \frac{\left(tr\left[\Sigma H\right]\right)^2}{tr\left[(\Sigma H)^2\right]} \tag{2.16}$$

Given that $\Sigma$ is the covariance matrix of data, and $tr[.]$ is the trace of square matrix. Let us recall that in a faulty case the data are corrupted with faults $\mathbf{f}(k)$ at sample time $k$, in different projection directions $\Xi_j$. The sample vector $\mathbf{x}(k)$ index can be rewritten as:

$$\mathbf{x}(k) = \mathbf{x}^*(k) + \Xi_j \mathbf{f}(k) \tag{2.17}$$

where $\Xi_j$ is the matrix which columns are those of an identity matrix corresponding to fault directions. The necessary condition of detectability for the fault $\Xi_j\mathbf{f}(k)$ using an index $\Upsilon$ is then given by the following inequality:

$$\left\| H^{\frac{1}{2}}\Xi_j\mathbf{f}(k) \right\|^2 > 2\Gamma^2 \tag{2.18}$$

The condition in 2.18 is valid for all quadratic indices defined in table 2.1.

### 2.2.5 EWMA Filtering

For further enhancement of the performances of detection statistics, (Wold, 1994) successfully combined EWMA filters in conjunction with PCA and PLS. The Exponontial Weighted Moving Average (EWMA) filter can be used on the generated residuals to reduce noise's impact on the data, thus reducing false alarms. The EWMA filter is much more efficient in the case when the treated data are not normally distributed.

Consequently, the filtered residuals are closer to normal distribution than the unfiltered residuals (Qin et al., 1997). The expression of EWMA filter applied on a residual vector $\mathbf{e}(k)$ is given by:

$$\mathbf{e}_{(f)}(k) = (1-\beta)\mathbf{e}_{(f)}(k-1) + \beta\mathbf{e}(k) \tag{2.19}$$

Where $\mathbf{e}_{(f)}(k)$ is the filtered residuals, and $\beta$ is a diagonal matrix which elements are the forgetting factors. If we note $\beta = \gamma I$, then $0 < \gamma < 1$ is adjusted according to the faults to detect, i.e. $\gamma$ close to zero to favour abrupt changes detection (mean changes), and close to one to favour slow changes detection (variance changes).
The filtered $SPE$, denoted $SPE_{(f)}$, is then computed from the filtered residuals $\mathbf{e}_{(f)}(k)$ as:

$$SPE_{(f)}(k) = \left\|\mathbf{e}_{(f)}(k)\right\| \tag{2.20}$$

However, control threshold $\delta^2_\alpha$ for the $SPE$ in 2.3 and 2.4 cannot be used for detection changes in the filtered version $SPE_{(f)}$, as the filtered $SPE$ defines a smaller region than the $SPE$. (Qin et al., 1997) defined the control threshold for the $SPE_{(f)}$ in terms of the conventional $SPE$'s limit ($\delta^2_\alpha$) and the forgetting factor $\gamma$, as:

$$\delta^2_{\alpha(f)} = \frac{\gamma}{2-\gamma}\delta^2_\alpha \tag{2.21}$$

This EWMA scheme for filtering residuals and $SPE$ can also, more or less, be applied for the $T^2$ and $SWE$ statistic. However, it is necessary to point out that these statistics measure different situations of the process, and have different behaviors as of detecting abnormal situations. The SPE and SWE statistics measure variability in the residual space, i.e. variability that is not included in the normal process correlation, which often indicates an aberrant information, while the $T^2$ index measures the distance from the origin in the principal component subspace. As the normal region defined by the control limit for $T^2$ is usually much larger than that of the SPE, it usually takes a much larger fault magnitude to exceed the $T^2$ control limit, while the SPE can detect any small to moderate faults. In the other hand, the use of the SWE index may be more effective in some cases, because taking into account the residual variances can enable to detect faults that can not be detected using the SPE index. However, as the $SWE$ statistic is scaled with respect to the last $(m-\ell)$ eigenvectors, this can result in some ill-conditioning results when the last $(m-\ell)$ eigenvalues are very close to zero, and can also be not defined in some cases.

A similar EWMA filtering scheme can be applied directly on the desired statistic, without having to filter residuals. Thus, for a given statistic $\Upsilon$, the general expression of EWMA filtered index $\Upsilon_{(f)}$ is given by:

$$\Upsilon_{(f)}(k) = (1-\gamma)\Upsilon_{(f)}(k-1) + \gamma\Upsilon \tag{2.22}$$

The corresponding control threshold $\Gamma^2_{(f)}$ then has to be computed from the approximate distribution of the filtered index $\Upsilon_{(f)}(k)$ using equations 2.4, 2.9 and 2.12 , for $SPE_{(f)}$, $T^2_{(f)}$ and $SWE_{(f)}$, respectively.

## 2.3 Fault Isolation

Once a fault has been detected, the next step is to determine the cause of the out-of-control status, i.e. the isolation of faults. Isolating or identifying the fault in processes can be a challenging task, especially in chemical processes that are highly integrated and complex. Practically, many variables of the process go out-of-control simultaneously and in a short time period when a fault occurs. The aim of fault identification is then to determine which variables are most relevant to diagnosing the fault, or in other words which variables are responsible of the fault. In the literature, many researches treated the topic of fault isolation, and several approaches have been proposed for isolation of faults in MSPC based on PCA models.

The most popular approach to fault isolation is the contribution plot approach (MacGregor et al., 1994), (Miller et al., 1998). This approach requires no prior knowledge except for a normal PCA model, and is classified as an unsupervised method for isolation. The principle of contribution based isolation is to determine the effects of the fault, on the observed vector of measurements, by quantifying the contribution of each process variable to the individual scores of the PCA representation. Other methods, so-called supervised methods, include Partial PCA, reconstruction-based approach, and many others. Partial PCA, introduced by (Gertler et al., 1999) was inspired by ideas borrowed from parity relations. By performing PCA on subsets of variables, a set of structured residuals can be obtained in the same way as structured parity relations. The structured residuals are utilized in composing an isolation scheme for sensor and actuator faults, according to a properly designed incidence matrix. If prior knowledge or historical data of the faults are available, the reconstruction-based approach, (Dunia et al., 1996a), (Dunia et al., 1996b) and (Dunia and Qin, 1998), can lead to more conclusive results.

Different other methodologies based on fault reconstruction on fault signature extraction (Yoon and MacGregor, 2001) in addition to different classification techniques based on the use of partial least squares discriminant analysis (Sjöstrjöm et al., 1986) were successfully applied to fault diagnosis. A good review on isolation methods and applications developed during the last two decades can be found in (Qin, 2012) and (Russell et al., 2012).

### 2.3.1 Contribution Plots

This method was proposed by (MacGregor et al., 1994) and (Miller et al., 1998) and is a widespread solution for fault identification when there is no *a priori* knowledge about the different types of fault in the process.

In the case of the SPE, for a new observation $\mathbf{x}(k)$, and in a given direction $j$, the contribution of the variable $x_j(k)$ to the $SPE$ is given by the following expression:

$$Cont_j(k) = \left(\xi_j^T(\mathbf{x}(k) - \hat{\mathbf{x}}(k))\right)^2 = \left(\xi_j^T C_{(m-\ell)}\mathbf{x}(k)\right)^2 \tag{2.23}$$

Where $\xi_j$ is the direction vector with all elements equal to 0 except the $j$-the which value is 1. The contributions calculated are based on the residuals of each sensor/variable

at every sample. Thus, the sensor with the largest error is considered faulty, since it has a major contribution to the $SPE$ index used for fault detection.

Using the definition of (Alcala and Qin, 2009), the contributions of variables in the genralized case for the quadratic index $\Upsilon$ is given by :

$$Cont_j^{\Upsilon}(k) = \left(\xi_j^T H^{\frac{1}{2}} \mathbf{x}(k)\right)^2 \tag{2.24}$$

where $H^{\frac{1}{2}}$ is defined in table 2.1 for the $SPE$, the Hotteling's $T^2$ index and the Hawkins $SWE$ statistic.

The $j$-the variable is considered faulty if:

$$\frac{Cont_j^{\Upsilon}(k)}{\Gamma^2} > 1 \tag{2.25}$$

where $Cont_j^{\Upsilon}(k)$ is the contribution calculated for the index $\Upsilon$, and $\Gamma^2$ is the corresponding control threshold. Using the previous formulas for the corresponding index, the total contribution of variable $X_j$ is calculated as:

$$CONT_j^{\Upsilon} = \sum_{k=1}^{n} Cont_j^{\Upsilon}(k) \tag{2.26}$$

This contributions arranged in the corresponding bar charts for all variables are known as contribution plots and are excellent tools for quickly identifying the observation variables that are related to the detected fault.

### 2.3.2 Partial PCA

Partial PCA, introduced in (Gertler et al., 1999), is a PCA performed on a reduced rank matrix $X^{(i)}$, where some variables in the original data matrix $X$ are missing. Therefore, the residuals will only be sensitive to faults in the variables which are present in the reduced model. In other words, structured residuals are designed so that each residual is sensitive to a particular subset of fault. Thus, the faults associated with the variables $i$ eliminated from the partial PCA model will yield a fault free residuals. The structured residuals are characterized by an incidence matrix, where the rows of this matrix correspond to residuals and its columns to faults. Table 2.2 is an example of a highly isolating structure for 3 variables, also know as theoretical signature matrix. The intersections between lines and columns in matrix of table 2.2 indicate the relation of the residual $(e_1, e_2, e_3)$ to the faults $(f_1, f_2, f_3)$, where a "1" indicates that the residual is sensitive to the fault, where a "0" indicates that the residual is not sensitive to the fault.

A more compact solution for the isolation of faults based on partial PCA models is then performed using the desired fault detection index calculated from the structured residuals. The global procedure is done in two phases and is resumed in the following steps, as explained in (Harkat, 2002)

*Phase I: PCA sub-models*

| | $f_1$ | $f_2$ | $f_3$ |
|---|---|---|---|
| $e_1$ | 0 | 1 | 1 |
| $e_2$ | 1 | 0 | 1 |
| $e_3$ | 1 | 1 | 0 |

TABLE 2.2 – Highly isolating incidence matrix

1. Conduct a PCA on the data matrix.
2. Construct highly isolating incidence matrix.
3. Construct a set of PCA sub-models, i.e. Partial PCA's, corresponding to the theoretical structure predefined in step 2.
4. Determine the thresholds of the desired statistic for fault detection.

*Phase II: Isolation of faults*

1. Acquire a new set of testing data.
2. Compute the new fault detection index for each partial PCA model.
3. Compare the index to the corresponding threshold and construct the experimental signature faults, i.e. put "1" in case of presence of fault, and "0" in case of its absence.
4. Compare the experimental signature of the faults with the columns of the incidence matrix to determine the faulty variable.

### 2.3.3 Reconstruction-based Approach

(Dunia et al., 1996a) introduced an approach to process sensor fault detection and isolation using PCA, based on the reconstruction principle. The obtained PCA considers the best model for fault reconstruction using the variance of reconstruction error criterion, see section 1.3.4. The reconstruction of process faults consists of estimating the sample vector $\mathbf{x}(k)$ by eliminating the effect of the fault $\mathbf{f}(k)$ in direction $\xi_i$ (Dunia et al., 1996a), (Dunia et al., 1996b), (Dunia and Qin, 1998), thus :

$$\mathbf{x}^{(i)}(k) = \mathbf{x}(k) - \xi_i \mathbf{f}(k) \tag{2.27}$$

The reconstruction of the corrupted observation vector $\mathbf{x}(k)$ is conducted in direction $i$, which replaces the observation in this direction by its fault-free reconstruction using the following expression.

$$\mathbf{x}^{(i)}(k) = G^{(i)}\mathbf{x}(k) \tag{2.28}$$

The projection matrix $G^{(i)}$ is given by:

$$G^{(i)} = \begin{bmatrix} \xi_1 & \dots & \mathrm{g}_i & \dots & \xi_m \end{bmatrix}^T, \ \mathrm{g}_i^T = \frac{\begin{bmatrix} c_{-i}^T & 0 & c_{+i}^T \end{bmatrix}}{1 - c_{ii}} \tag{2.29}$$

Or, $G^{(i)}$ is expressed in terms of $C_\ell$ as:

$$G^{(i)} = I_m - \xi_j(\xi_j^T C_\ell \xi_j)^{-1}\xi_j^T C_\ell) \tag{2.30}$$

Where $\xi_i = [0 \dots 1 \dots 0]$ is the vector of reconstruction direction with all elements equal to 0 except the $i$-th which value is 1, and $I_m$ is the identity matrix of size $m$. This method for isolation then suspects that each variable is faulty and suggests to reconstruct it using the PCA model from the remaining variables. Subsequently, the residual vector is defined by:

$$\mathbf{e}^{(i)}(k) = (I - C_\ell)G^{(i)}\mathbf{x}(k) \tag{2.31}$$

So, if we note by $SPE^{(i)}$ the expression of the SPE index computed after reconstructing the $i$-th variable, this particular index will not be affected with faults in this direction. In a more generalized way, the fault reconstruction methodology is explained in terms of various detection statistics. Let us consider the case of an unknown fault $\xi_j\mathbf{f}(k)$. The detection statistic $\Upsilon^{(i)}$ is the calculated index after reconstruction of the $i$-th variable, i.e. in direction $\xi_i$, and is given by:

$$\Upsilon^{(i)}(k) = \left\| H^{\frac{1}{2}}\mathbf{x}^{(i)}(k) \right\|^2 \tag{2.32}$$

Given that $\Gamma^{(i)2}$ is the corresponding control threshold for the index $\Upsilon^{(i)}(k)$ , and according to the reconstruction principle, the reconstructed index $\Upsilon^{(i)}(k)$ will be in control (not faulty) in the case when the direction of the real fault $\xi_j\mathbf{f}(k)$ is the same as the reconstruction direction $\xi_i$, that is, when $\xi_i = \xi_j$. The isolation can then be performed based on a comparison between a predetermined theoretical signatures matrix (incidence matrix) as in 2.2, and the experimental signatures matrix established after reconstructions in different directions $\xi_i$, or using the following isolation index:

$$A_\Upsilon(k) = \frac{\Upsilon^{(i)}(k)}{\Gamma^{(i)2}} \tag{2.33}$$

The variable for which the isolation index $A_\Upsilon(k)$ is lower than one is declared faulty.

## 2.4 Simulation Example

To demonstrate the application of the FDI strategy based on PCA model according to the scheme explained in 2.1, let us recall the previous simulation example in 1. After constructing a PCA model in the first phase, based on $\ell = 2$ principal components (as detailed in section 1.3.4), we compute the off-line thresholds for the used statistics ($SPE$, $T^2$ and $SWE$). A new set of 500 data samples are generated based on relations of example 1, and a fault is added to variable 3 from sample time 350 to 500, as a bias corresponding to 10% of the variable variation.

The first computed statistic is the $SPE$ indicator which is represented by Figure 2.2. We notice that the injected fault is correctly detected based on this index with around 5 missed detections, and with the presence of a small amount of false alarms. Figure 2.3 represents the detection of faults based on the $SWE$ indicator, which exhibits a a slightly better results than the $SPE$, with barely one missed detection. This is due to the sensitivity of the $SWE$, because of its weighting by the eigenvalues. Finally, the $T^2$ statistic is presented in Figure 2.4, in which we note that the statistic could not detect the injected fault. As stated before, the $T^2$ by definition measures variations in

the principal space, thus can only detect faults that are much higher in amplitude than the ones detected by the $SPE$ or the $SWE$ statistic.

Once the fault is detected, the next step is to isolate the fault, in this case only using the $SPE$ indicator. We emphasis in our case on the isolation based on the reconstruction approach. Figure 2.5 represents reconstructions in different directions of the $SPE$ index, denoted $SPE^{(i)}$, $i = 1, \ldots, 6$. By definition, the reconstruction of variables gives an estimate of the variables while eliminating the effect of a certain variable of rank $i$, thus the reconstructions will be fault free if the fault is in this $i$th variable. According to Figure 2.5, we notice the absence of the fault in $SPE^{(3)}$, i.e. when eliminating the effect of third variable, which subsequently means that this variable is the faulty one. This statement is obviously true given that the fault was initially injected to the 3rd variable during our simulation.

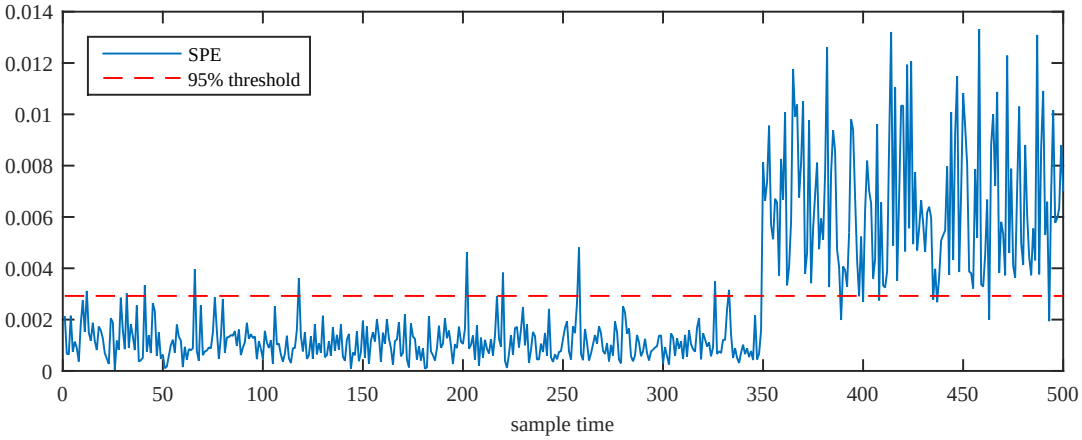



FIGURE 2.2 – $SPE$ indicator for example 1

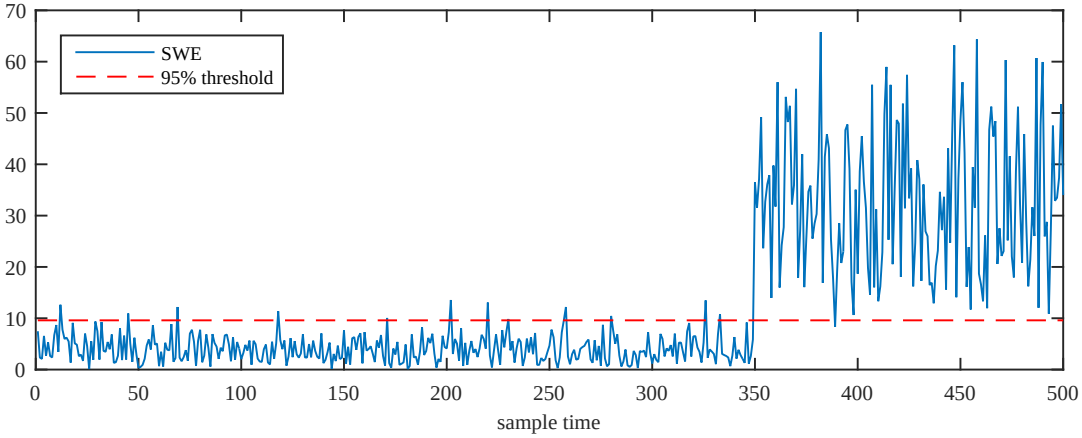



FIGURE 2.3 – $SWE$ indicator for example 1

## 2.5 Conclusion

Principal components analysis reduces the data representation subspace and enables the determination of the redundancy relationships (linear relations among the variables).

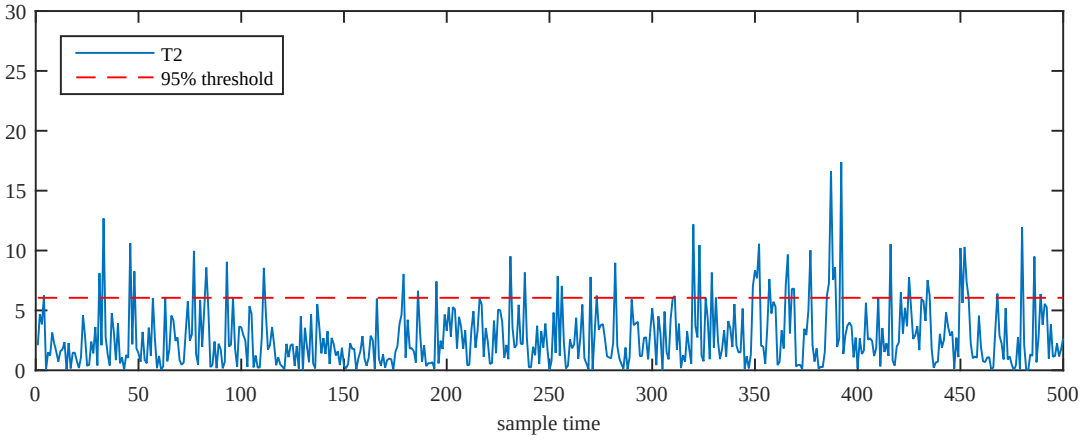



FIGURE 2.4 – $T^2$ indicator for example 1

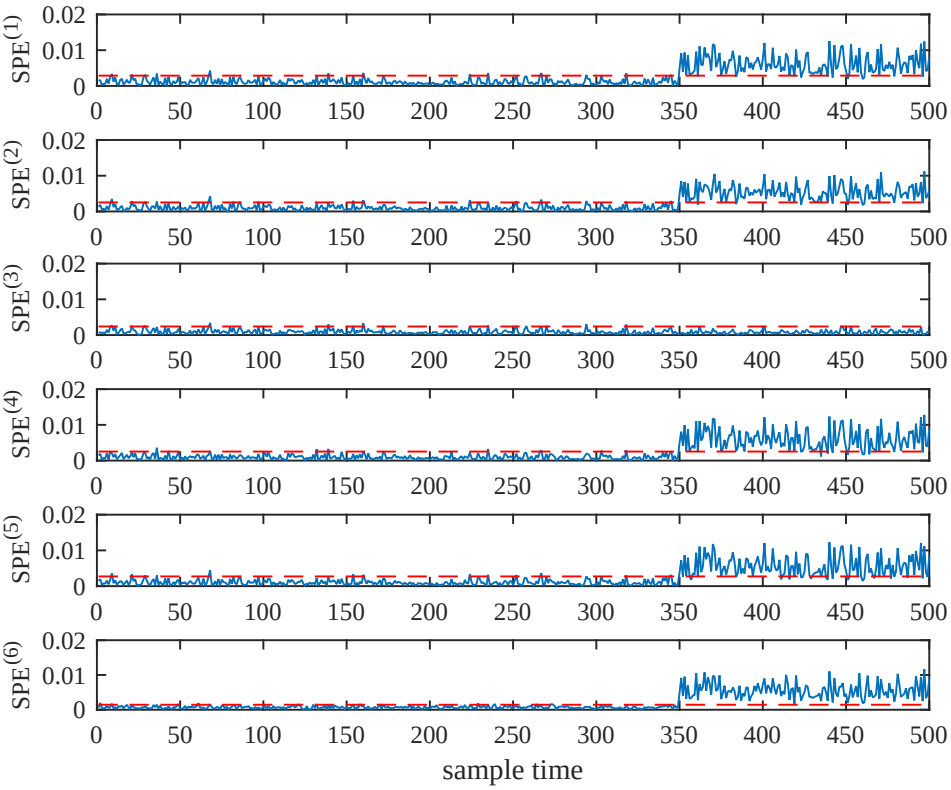



FIGURE 2.5 – Reconstructions based on SPE indicator of example 1 for fault isolation

The redundancy relations are then used to detect and isolate the faulty data. PCA is constructed with fault-free data from a decomposition in eigenvalues and eigenvectors of a covariance matrix. Then, the new data is referenced to this model where any faulty condition can be detected. The detection of faults based on a PCA model is mainly performed using a certain statistic or indicator, then the fault is isolated.
In this chapter, we presented the fault detection and isolation scheme based on a PCA model. Various statistics are presented, namely the $SPE$ , $T^2$ and $SWE$ indicators

and their corresponding thresholds. The isolation methods presented in this chapter are the contribution plots method, the partial PCA method, and the reconstruction based approach. The FDI scheme is illustrated based on a simulation example where the different statistics ($SPE$ , $T^2$ and $SWE$) are used for fault detection. The fault isolation phase is performed based on the reconstruction based approach, which demonstrates its effectiveness for the task.

PCA is a rather popular approach for sensor fault detection and isolation. It is however limited, in its classical off-line case to stationary processes, and can not handle dynamic processes, or non-linear processes. Therefore, other PCA variants should be used depending on the process at hand, such as Recursive PCA, Moving window PCA, Non-linear PCA, or Kernel PCA. In addition to that a more robust approach is required in the case of uncertain processes, which is mostly the case, as the real data sets always include uncertainties. This particular case motivated the research presented in this manuscript for using PCA for interval-valued data for diagnosis.

# PART 2

# PCA FOR INTERVAL-VALUED DATA AND APPLICATION TO FDI

# 3 | Interval-Valued PCA for FDI



## 3.1 Introduction

The main idea of Multivariate Statistical Process Control (MSPC) approach is to extract the useful data information from the original data samples obtained in large quantities from a process under normal operating conditions, and construct some statistics for monitoring the state or condition of the process. So far, the most widely used MSPC method may be the principal component analysis approach. By implicitly providing a model of the system while reducing the dimension of the used data, PCA can handle high-dimensional and correlated process variables, and provide detailed monitoring results for each process data samples. Along the past several decades, different improvements and modifications have been made to the traditional PCA in order to handle different natures of processes/data. Most of these PCA methods were originally proposed in other areas, such as neural PCA (Kramer, 1991) and kernel PCA (Scholkopf et al., 1998) for non-linear processes, as well as recursive PCA (Li et al., 2000) and moving-window PCA (Wang et al., 2005) for time-varying processes, among many others. These PCA variations have been widely investigated for use in MSPC as a tool for sensor fault detection and isolation. However, none of these

approaches treats the particular case where process data are highly noisy due to sensor uncertainties. i.e. there is no sufficiently robust approach with regard to uncertainties of sensor measurements, though it is mostly the case in practical cases.

Indeed, real life data are approximate values given by the process sensors, and are generally stained with uncertainties due to different factors, including measurement noise, degradation from age, environmental exposure, and even maintenance interventions, which affects the measured signals and can lead ultimately to a mass of misstatement and false alarms. A contemporary way of representing the influence of these uncertainties on sensors is the representation in the form of interval-valued data. In this case, each measurement is no longer represented as a value, but as a set of values limited by the minimum and maximum recorded values. Interval-valued data can be assumed as special case of Symbolic Data (SD) that were first introduced by (Diday, 1987). Unlike a classical observation which takes a single value, a symbolic observation takes multiple values. A symbolic observation can represent a category, an interval, a sequence of categorical values, a sequence of weighted values, a bar chart, a histogram, or a distribution. As a consequence, symbolic data have an internal structure which does not exist in classical data and traditional methods of analysis of classical data do not account for this structure. Therefore, new analytical methods need to be developed to account for this special characteristic of symbolic data.

Statistical methods for handling symbolic data data have been considered in the context of Symbolic Data Analysis. SDA's emphasis is the analysis of data sets where individuals are described by variables that can represent variation. In the framework of SDA, PCA was adapted to the statistical treatment of symbolic data, and mainly in the context of interval-valued data. In the last two decades, various approaches were proposed. The first PCA adaptations for interval-valued data are known as the vertices PCA (VPCA) and the centers PCA (CPCA) introduced by (Cazes et al., 1997) and (Chouakria, 1998). Both methods involve a transformation of the interval input matrix. The first approach uses the vertices matrix to compute principal components, whereas the second uses centers matrix. (Lauro and Palumbo, 2000) proposed a multi-stage method called the symbolic object PCA (SO-PCA) attempting to account dependency among the vertices of each observation. Other PCA for interval-valued data methods include the midpoints-radii PCA (MRPCA) introduced by (Palumbo and Lauro, 2003), which treats midpoints and interval ranges as two separate variables to enhance CPCA by incorporating radius. (D'Urso and Giordani, 2004), proposed an alternative approach using least squares for MRPCA. (Gioia and Lauro, 2006) proposed an analytical approach based on an interval-valued covariance matrix. A hybrid approach to PCA of multidimensional interval-valued data has been proposed by (Irpino, 2006) so-called Spaghetti PCA. (Le-Rademacher and Billard, 2012) employed symbolic covariance for interval-valued data to extend the classical PCA. (Wang et al., 2012) proposed a new PCA for interval-valued data method with an enhanced covariance matrix calculation , and is called the complete-information PCA (CIPCA). (Chen et al., 2015) proposed the probabilistic symbolic PCA for symbolic data with general probability distributions of data.

As an alternative to classical PCA, (Benaicha et al., 2013), (Ait-Izem et al., 2014a),(Ait-Izem et al., 2014b), (Ait-Izem et al., 2015a), have shown that it is most likely possible to apply PCA for interval-valued data methods in monitoring uncertain systems by modeling sensors uncertainties in the form of intervals. Thus, ensuring a maximum robustness of the diagnosis routine.

## 3.2 Interval-valued Data

The true value of the quantity is a concept. In almost all cases, the true value cannot be measured and the collected data on a process are only approximations given by the sensors, and are thus imprecise. This is due mainly to the uncertainties induced by measurement errors or determined by specific experimental conditions.
Let $X \in \mathbb{R}^{n\times m}$ be the conventional data matrix containing $n$ samples of $m$ process variables collected under NOC, where $x_j(k)$ is the $k$-th observation of the $j$-th variable. In order to preserve the variable information, it is more appropriate to represent such measure by an uncertain or interval value rather than a single value. The uncertainty being unknown; we suppose that its variation is limited and can be represented by an interval of the form $[\underline{x}_j(k), \overline{x}_j(k)]$, where $\underline{x}_j(k)$ and $\overline{x}_j(k)$ denote, respectively, the lower bound (LB) and the upper bound (UB) of the interval, precisely, they represent the minimum and maximum values registered for the $j$-th interval valued variable with respect to the $k$-th observation unit. To take into consideration the parameter uncertainties and thus to provide strict bounds of the estimated measurements, only real intervals are considered. Thus, the $k$-th real interval observation for the $j$-th variable is denoted $[x_j(k)]$, and the new global interval matrix $[X]$ is formalized as a set of interval valued observation for the different description variables by the following:

$$[X] = \begin{pmatrix} [\underline{x}_1(1), \overline{x}_1(1)] & . & . & [\underline{x}_m(1), \overline{x}_m(1)] \\ . & . & & . \\ . & & . & . \\ [\underline{x}_1(n), \overline{x}_1(n)] & . & . & [\underline{x}_m(n), \overline{x}_m(n)] \end{pmatrix} \tag{3.1}$$

Let $\delta x_j(k)$ be the error in a measurement, where $x_j(k)$ is the quantity of interest. $\delta x_j(k)$ is the difference between the true value $x_j^*(k)$ and the value reported as a result of a measurement $x_j^c(k)$.

$$\delta x_j(k) = x_j^c(k) - x_j^*(k) \tag{3.2}$$

Equivalently, when defining an interval model of the measured data, the radius of the interval observation is given by the measurement error $\delta x_j(k)$, also denoted $x_j^r(k)$, while the center of the interval is given by the measurement $x_j^c(k)$. Thus, the standard intervation construction formula is given by:

$$[x_j(k)] = \left[x_j^c(k) - x_j^r(k),\ x_j^c(k) + x_j^r(k)\right] \tag{3.3}$$

The interval $[x_j(k)]$ can also be expressed in the form known as the midpoint-radius form by the couple (midpoint,radius), as:

$$[x_j(k)] = \left\{x_j^c(k), x_j^r(k)\right\} \tag{3.4}$$

Choosing the bounds on a measurement, i.e. the width of the interval, depends on the corresponding uncertainty, and is an important step for constructing accurate data of the process. Neither measurements nor estimates are 100 % accurate, so in reality, the actual value $x_j^*(k)$ of a quantity can differ from the result $x_j^c(k)$ obtained by measurement. The measurement errors being $\delta x_j(k) = x_j^c(k) - x_j^*(k)$. The

manufacturer of the measuring instrument must supply us with an upper bound $\delta x_j(k)$ on the measurement error. If no such upper bound is supplied, this means that no accuracy is guaranteed, and the corresponding measurement instrument is practically useless. In this case, once we perform a measurement result $x_j^c(k)$, we know that the actual (unknown) value $x_j^*(k)$ of the measured quantity belong to the interval $x_j^*(k) = [\underline{x}_j(k) \;\; \overline{x}_j(k)]$, where $\underline{x}_j(k) = x_j^c(k) - \delta x_j(k)$ and $\overline{x}_j(k) = x_j^c(k) + \delta x_j(k)$.

Dealing with quantitative variables, there are many other cases in which a more complete information can be surely achieved by describing a set of statistical units in terms of interval data. For example, daily temperatures registered as minimum and maximum values offer a more realistic view on the weather conditions variations with respect to the simple average values. Another example can be given by air quality data where each concentration measurement is taken as a mean of several measurements over 15 minutes (sample time), the minimum and the maximum concentration, recorded over 15 minutes, represent a more relevant information for experts in order to evaluate tendency and variability of pollutants concentrations.

### 3.2.1 Interval Arithmetic and Statistics

This section reviews basic interval arithmetic operations and statistics for interval computations used in this dissertation. Modern developments on interval analysis and algebra go back to Moore's work (Moore, 1966). Considering the generic interval-valued datum $[x]$, the definitions of real arithmetic operators and functions are extended to interval-valued case.

**Definition 3.1 : Interval-valued object**
*A real interval-valued object* $[x]$*, is a closed and connected subset in* $\mathbb{R}$*, defined by:*
$[x] = [\underline{x}, \overline{x}] = \{x \in \mathbb{R} / \underline{x} < x < \overline{x}\}$

**Definition 3.2 : Trivial interval**
*An interval* $[x] = [\underline{x}, \overline{x}]$ *is defined to be trivial if* $\underline{x} = \overline{x}$*.* $[x]$ *is then reduced to a single value and becomes a classical observation.*

**Definition 3.3 : Interval midpoint**
*The midpoint of a generic interval* $[x]$*, or center, is given by:*
$x^c = \dfrac{\underline{x} + \overline{x}}{2}$

**Definition 3.4 : Interval radius and range**
*The radius of a generic interval* $[x]$ *is given by:*
$x^r = \dfrac{\overline{x} - \underline{x}}{2}$
*the range of an interval* $[x]$ *is its width given by:*
$w([x]) = \overline{x} - \underline{x}$

**Definition 3.5 : Interval Sum**
*the sum of two generic intervals* $[x]$ *and* $[y]$ *is the set* $\{x + y : x \in [x], y \in [y]\}$ *where* $\underline{x} + \underline{y} < x + y < \overline{x} + \overline{y}$*, and is implemented as* $[x] + [y] = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$

**Definition 3.6 : Interval Difference**
*the difference between two generic intervals* $[x]$ *and* $[y]$ *is the set* $\{x - y : x \in [x], y \in [y]\}$ *where* $\underline{x} - \overline{y} < x - y < \overline{x} - \underline{y}$*, and is implemented as* $[x] - [y] = [\underline{x} - \overline{y}, \overline{x} - \underline{y}]$

**Definition 3.7 : Interval Product**
*the product of two generic intervals* $[x]$ *and* $[y]$ *is the set* $\{x \times y : x \in [x], y \in [y]\}$ *where* $Min\, S < x \times y < Max\, S$*, given that* $S = \left\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\right\}$*, and is computed as* $[x] \times [y] = [Min\, S, Max\, S]$

**Definition 3.8 : interval-valued vector and matrix**
*An n-dimensional interval vector* $[\mathbf{x}]$*, is the ordered n-tuple of intervals given by:* $[\mathbf{x}] = [[x_1], [x_2], \ldots, [x_n]]$.
*The* $n \times m$ *dimensional matrix* $X$ *is the ordered m-tuple of interval-valued vectors* $[\mathbf{x_i}], i = 1, \ldots, m$ *given by:* $X = \left[[\mathbf{x_1}]^T, [\mathbf{x_2}]^T, \ldots, [\mathbf{x_m}]^T\right]$

**Definition 3.9 : Interval mean**
*the interval mean* $[m]$ *of an interval-valued vector* $[\mathbf{x}]$ *is defined as:* $[m] = \frac{1}{n}\sum_i [x_i]$

**Definition 3.10 : Interval distance**
*The distance between two generic intervals* $[x]$ *and* $[y]$ *is given by:*
$d([x], [y]) = |x^c - y^c| + |x^r - y^r|$
*where* $d([x], [y])$ *satisfies the Euclidean distance properties.*

This distance is based on a quadratic extension of the distance proposed by (Neumaier, 1990), and is better known as the Hausdorff distance between two intervals.

## 3.2.2 Interval-valued Data Normalisation

Typically, some normalisation must be performed prior to processing data in order to objective or scale-invariant result. Three alternative normalisation methods for the case of interval-valued data have been proposed in (Carvalho et al., 2006), and are described below. Interval valued variables are normalised according to the procedure described below, proposed by (Lauro et al., 2008). These results were obtained by referring to some basic interval arithmetic concepts in section 3.2.1.

### 3.2.2.a normalisation using the dispersion of interval center and range

Taking into account definition 3.10, the following properties can be verified. Let $\{[x_j(1)], [x_j(2)], ..., [x_j(k)]\}$ be a set of finite interval, so that $[x_j(k)] \subset \mathbb{R} \forall k \in \{1, ..., n\}$ and $[m_j]$ is their corresponding mean interval, then:
$\sum_{k=1}^{n}\left[\left(x_j^c(k) - m_j^c\right) + \left(x_j^r(k) - m_j^r\right)\right] = 0$. and $\sum_{k=1}^{n} d^2([x_j(k)], [m_j])$ is minimized. Where $x_j^c(k)$ and $x_j^r(k)$ are the midpoints and the radius of $[x_j(k)]$ according to definitions 3.3 and 3.4, and $m_j^c$,$m_j^r$ are their respective means. Definition 3.10 allows to introduce the notation of scalar variance for interval valued data. We define the variance as

the sum of squared distances with respect to the mean interval, as a consequence the variance $\sigma_j^2$ for interval valued data is defined as follows: $\sigma^2 = \frac{1}{n}\sum_{k=1}^{n} d^2\left([x_j(k)], [m_j]\right)$.

Variance definition can also be written according to the following formula:

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\left|x_j^c(k) - m_j^c\right| + \left|x_j^r(k) - m_j^r\right|\right)^2 \tag{3.5}$$

where $[m_j] = \left[\begin{array}{cc} \frac{1}{n}\sum_{k=1}^{n}\underline{x}_j(k) & \frac{1}{n}\sum_{k=1}^{n}\overline{x}_j(k) \end{array}\right]$, $m_j^c = \frac{1}{n}\sum_{k=1}^{n} x_j^c(k)$ and $m_j^r = \frac{1}{n}\sum_{k=1}^{n} x_j^r(k)$.

With a little algebra we obtain:

$$\sigma^2 = \frac{1}{n}\left[\sum_{k=1}^{n}\left(x_j^c(k) - m_j^c\right)^2 + \sum_{k=1}^{n}\left(x_j^r(k) - m_j^r\right)^2 + 2\sum_{k=1}^{n}\left|x_j^c(k) - m_j^c\right|\left|x_j^r(k) - m_j^r\right|\right] \tag{3.6}$$

The expression in 3.6 affirms that the variance for interval valued data can be decomposed into three components: variance among midpoints, variance among ranges and twice the connection between midpoints and ranges, given by $\sum_{k=1}^{n}\left|x_j^c(k) - m_j^c\right|\left|x_j^r(k) - m_j^r\right| \geq 0$.

The remarked properties in the definition 3.10 indicate that the distance between intervals can be generalized to the Euclidean distance in the space $\mathbb{R}^m$. A normalised interval is

$$\left[\frac{1}{\sigma}\left(x_j^c(k) - m_j^c - \left|x_j^r(k) - m_j^r\right|\right), \frac{1}{\sigma}\left(x_j^c(k) - m_j^c + \left|x_j^r(k) - m_j^r\right|\right)\right] \tag{3.7}$$

#### 3.2.2.b normalisation using the dispersion of the interval centers

The first method considers the mean and the dispersion of the interval centers $(\underline{x}_j(k) + \overline{x}_j(k))/2$ and normalises such that the resulting transformed midpoints have zero mean and unit variance for each variable.

The mean value and the dispersion of all interval midpoint are given by:

$$m_j = \frac{1}{n}\sum_{k=1}^{n}\frac{\left(\underline{x}_j(k) + \overline{x}_j(k)\right)}{2} \quad \text{and} \quad \sigma_j^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\underline{x}_j(k) + \overline{x}_j(k)}{2} - m_j\right)^2 \tag{3.8}$$

with this notation the normalised interval is defined with boudaries

$$\left[\frac{\underline{x}_j(k) - m_j}{\sigma_j}, \frac{\overline{x}_j(k) - m_j}{\sigma_j}\right] \tag{3.9}$$

#### 3.2.2.c normalisation using the dispersion of the interval boundaries

The second normalisation method transforms, for each variable $[X_j]$, the $n$ intervals $[x_j(k)]$ such that the mean and the joint dispersion of the rescaled interval boundaries are 0 and 1, respectively. The joint dispersion of a variable $[X_j]$ is defined by:

$$\sigma_j^2 = \frac{1}{n}\sum_{k=1}^{n} \frac{\left(\underline{x}_j(k) - m_j\right)^2 + \left(\overline{x}_j(k) - m_j\right)^2}{2} \tag{3.10}$$

Then, for $k = 1, ..., n$, the intervals $[x_j(k)] = [\underline{x}_j(k), \overline{x}_j(k)]$ are transformed into:

$$\left[\frac{\underline{x}_j(k) - m_j}{\sigma_j}, \frac{\overline{x}_j(k) - m_j}{\sigma_j}\right] \tag{3.11}$$

#### 3.2.2.d normalisation using the global range

The third normalisation method transforms, for a given variable, the intervals $[x_j(k)] = [\underline{x}_j(k), \overline{x}_j(k)]$, $(k = 1, ..., n)$ such that the range of the $n$ rescaled intervals is the unit interval $[0, 1]$. Let $Min_j = \min\left\{\underline{x}_j(k), ..., \underline{x}_j(k)\right\}$ and $Max_j = \max\left\{\overline{x}_j(k), ..., \overline{x}_j(k)\right\}$ be the extremal lower and upper boundary values. With this notation, the interval is transformed into normalised interval with boundaries:

$$\frac{\underline{x}_j(k) - Min_j}{Max_j - Min_j} \quad and \quad \frac{\overline{x}_j(k) - Min_j}{Max_j - Min_j} \tag{3.12}$$

## 3.3 PCA for Interval-valued Data

One key step of performing PCA on a set of data is the calculation of correlation (or covariance) matrix and it's eigen-decomposition. Computing eigenvalues of a matrix is a basic linear algebra task used in several engineering fields. However, real life applications makes this problem more complicated by imposing uncertainties and errors on collected measurements. Thus, representing these measurements as interval imposes using interval computation to solve the eigenvalue problem, which tends to be a very difficult task. The first results on interval-valued eigenvalues are due to (Deif, 1991) where bounds for real and imaginary parts for complex eigenvalues were studied, and (Rohn and Deif, 1991) where only real eigenvalues considered. Many works on the eigenvalue problem were developed afterwards, and several approximation algorithms were proposed. Unfortunately, the theoretical background for the eigenvalue problem of symmetric interval-valued matrices is not wide enough and there are few practical methods. Few existing approaches to interval-valued PCA use these methods, and are known as interval algebra based PCA's, but only work for very narrow intervals, especially with large sample size data, hence limiting their applicability. Alternatively, and given the absence of well established mathematics in the matter, several interval-valued PCA methods perform a codification of the initial interval-valued data set based on a certain geometrical representation.

In a multidimensional space, each observation unit can be visualized as a hyper-cube (or a hyper-rectangle) having $2^m$ vertices corresponding to all the possible combinations of lower bounds and upper bounds, while the length of its segments are given by the intervals of every description variable. Various interval-valued PCA approaches are based on this hyper-cube representation of interval-valued data, which furnishes adequate description to the symbolic objects case given by interval-valued variables. Figure 3.1 is a graphical representation of a set of three principal components (PC's) obtained from a three dimensional object $H_k$ described by (j = 3) interval-valued variables (Bertrand and Goupil, 2000). In a perfect PCA modeling case, the shapes formed by the projections on the two interval-valued PC's, $(PC_1, PC_2)$ and $(PC_2, PC_3)$, constitutes a maximal envelope of the projection points from $H_k$. Thus, every point in the hyper-cube $H_k$ lies inside this envelope when projected. However, there can be some points within the envelope that may not be projections of points from $H_k$, i.e. an overestimation of the hyper-cube by the PC's, or some projections of the hyper-cube that are not covered by the envelope, i.e. an underestimation of the hyper-cube by the PC's.

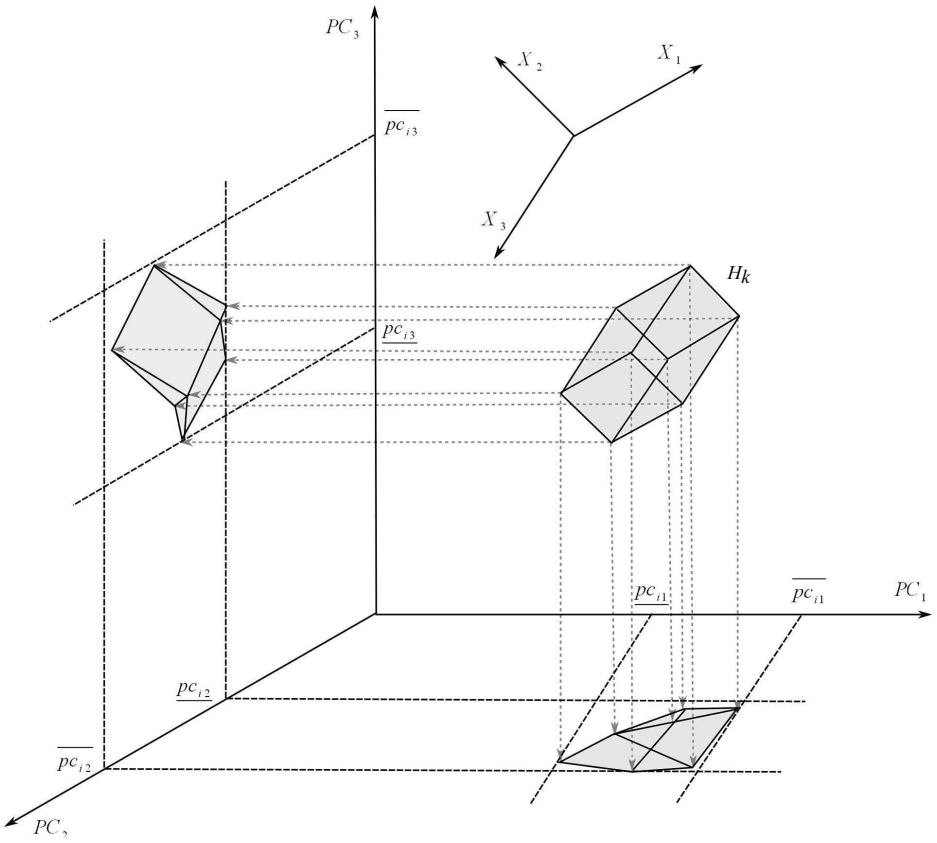



FIGURE 3.1 – Interval PCA (projection of a hyper-rectangle on principal components)

Undoubtedly, for an FDI adaptation of interval-valued PCA, an accurate PCA model have to be used. In this section, four most known interval PCA approaches: Vertices PCA, Centers PCA,Midpoints-Radii PCA and the Complete information PCA are presented, discussed , and investigated for use in process modelling with a comparison in terms of effectiveness for FDI purpose.

As for ordinary PCA, it is advisable to pre-process the data, for all hereby presented PCA for interval-valued data methods, in order to avoid unwanted differences among the variables. Without loss of generality, we assume in the following that interval-valued observations have been normalized according to methods of section 3.2.2.

### 3.3.1 Vertices PCA

One of the first proposed approaches to implement PCA for interval-valued data is Vertices Principal Component Analysis (VPCA) (Cazes et al., 1997),(Chouakria, 1998). VPCA proceeds in two steps. In the first step, classical PCA is applied to vertices coded matrix corresponding to interval-valued data, and in the second step, a range of each symbolic object on the scores is reconstructed to visualize the configuration of the interval data. According to VPCA, each object $H(k)$, described by an interval valued vector $[\mathbf{x}(k)] = [[\underline{x}_1(k), \overline{x}_1(k)] \dots [\underline{x}_m(k), \overline{x}_m(k)]]$ in an $m$ dimensional space, can be represented by a numerical data matrix $V(k)$ of $2^m$ lines and $m$ columns containing the vertices of the associated hyper-cubes $H_k$. Vertices are given by all the possible combinations between the bounds of the vector $[\mathbf{x}(k)]$. As an example, given a two dimensional interval vector $[\mathbf{x}(k)] = [[\underline{x}_1(k), \overline{x}_1(k)], [\underline{x}_2(k), \overline{x}_2(k)]]$, vertices matrix $V(k)$ is constructed as follows:

$$V(k) = \begin{pmatrix} \underline{x}_1(k) & \underline{x}_2(k) \\ \underline{x}_1(k) & \overline{x}_2(k) \\ \overline{x}_1(k) & \underline{x}_2(k) \\ \overline{x}_1(k) & \overline{x}_2(k) \end{pmatrix} \tag{3.13}$$

In other words, each bound of the interval-valued observations is a vertex of the corresponding hyper-cube $H_k$. Thus, a classical observation, i.e. a point has 1 vertex, a segment has 2 vertices, a rectangle 4 vertices, and so on. The global VPCA vertices matrix $V$ is then obtained by concatenating all vertices matrices $V(k)$ for all $H(k)$ objects as:

$$V = \begin{pmatrix} V(1) \\ \cdot \\ \cdot \\ V(k) \\ \cdot \\ \cdot \\ V(n) \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \underline{x}_1(1) & . & . & \underline{x}_m(1) \\ . & . & & . \\ . & & . & . \\ \overline{x}_1(1) & . & . & \overline{x}_m(1) \end{bmatrix} \\ \cdot \\ \cdot \\ \begin{bmatrix} \underline{x}_1(k) & . & . & \underline{x}_m(k) \\ . & . & & . \\ . & & . & . \\ \overline{x}_1(k) & . & . & \overline{x}_m(k) \end{bmatrix} \\ \cdot \\ \cdot \\ \begin{bmatrix} \underline{x}_1(n) & . & . & \underline{x}_m(n) \\ . & . & & . \\ . & & . & . \\ \overline{x}_1(n) & . & . & \overline{x}_m(n) \end{bmatrix} \end{pmatrix} \tag{3.14}$$

Therefore, PCA can be performed on the classical vertices matrix $V$ as:

$$V = P\Lambda P^T \tag{3.15}$$

Given that $\Lambda$ and $P$ are the eigenvalues and eigenvectors of covariance matrix $\Sigma_V$ computed for the vertices matrix $V$. Prior to computing the covariance matrix $\Sigma_V$, (Cazes et al., 1997) and (Chouakria et al., 2011) suggest to apply a weighting scheme on the vertices matrix $V$ according to the nature of the application at hand. Three main symbolic weighting schemes are proposed, which are presented in the following.

First, let us denote the weight of observation $[\mathbf{x}_j(k)]$ by $w_j(k)$, and let $w_j^l(k)$,$l = 1, \ldots, 2^m$ be the weight of vertex $l$. The weighting scheme should satisfy:

$$w_j(k) = \sum_{l=1}^{2^m} w_j^l(k), \quad \sum_{j=1}^{m} w_j(k) = 1 \tag{3.16}$$

A First and frequent choice of weight $w_j(k)$ gives equal weight to all observations, as in , $w_j(k) = 1/m, j = 1, ..., m$.

A second choice of weights gives importance to differing internal variations of hyper-cubes is given by:

$$w_j(k) = \frac{\nu_j(k)}{\sum\limits_{j=1}^{m} \nu_j(k)} \tag{3.17}$$

where $\nu_j(k)$ is the volume of hyper-cube $H_k$ associated with the observation vector $[\mathbf{x}(k)]$ computed as:

$$\nu_j(k) = \prod_{\underline{x}_j(k) \neq \overline{x}_j(k)} \left( \overline{x}_j(k) - \underline{x}_j(k) \right) \tag{3.18}$$

The third weighting scheme is where the weights are inversely proportional to volume, given by:

$$w_j(k) = \frac{1 - \frac{\nu_j(k)}{\sum\limits_{j=1}^{m} \nu_j(k)}}{\sum\limits_{j=1}^{m} \frac{1 - \nu_j(k)}{\sum\limits_{j=1}^{m} \nu_j(k)}} \tag{3.19}$$

Hence, the weight matrix $D$ is the $n \times n$ diagonal matrix of weights $w_j(k)$ associated with the vertices matrix $V$. The weighted covariance matrix of vertices is then given by:

$$\Sigma_V = VDV^T \tag{3.20}$$

VPCA is then performed by conducting a classical PCA on vertices Matrix $V$ using covariance matrix $\Sigma_V$. Let $V(k)$ be the vertices matrix of $[\mathbf{x}(k)]$, then principal components vertices matrix $T(k) = [t_1(k), t]$, is given by:

$$T(k) = V(k)P \tag{3.21}$$

where $P = [\mathbf{p}_1, ..., \mathbf{p}_m]$ is the eigenvectors of the covariance matrix $\Sigma_V$. (Cazes et al., 1997) proposed a computation method for the interval-valued components

$[t_1(k)], ..., [t_m(k)]$ using Moore algorithm (see appendix A), from the principal components vertices matrix calculated in (3.21), using the following formulas:

$$\begin{cases} \underline{t}_j(k) = \min\ (T_j(k)) \\ \overline{t}_j(k) = \max\ (T_j(k)) \end{cases} \tag{3.22}$$

Consequently, estimates of interval measurements are thus computed. Let $\hat{V}_j(k)\,(j = 1, \ldots, m)$ be the $j$th column of the the estimated vertices matrix $\hat{V}(k)$ given by:

$$\hat{V}(k) = V(k)P_\ell P_\ell^T = V(k)C_\ell \tag{3.23}$$

the estimated interval-valued measurements for the $\ell$ first principal components are given by:

$$\begin{cases} \underline{\hat{x}}_j(k) = \min\ \left(\hat{V}_j(k)\right) \\ \overline{\hat{x}}_j(k) = \max\ \left(\hat{V}_j(k)\right) \end{cases} \tag{3.24}$$

where $\hat{V}_j(k)$ is the $j$th column of $\hat{V}(k)$ matrix.
Given that an interval-valued observation is represented by its $2^m$ vertices, calculating the variance–covariance matrix $\Sigma_V$ is of complexity order $O(n2^m)$. However, (Cazes et al., 1997) and (Chouakria, 1998) provided a short-cut for calculation of covariance matrix $\Sigma_V$ of complexity order $O(n)$, as:

$\Sigma_V =$

$$\begin{pmatrix} \frac{1}{2n}\sum_{k=1}^{n}(\underline{x}_1^2(k)+\overline{x}_1^2(k)) & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_1(k)+\overline{x}_1(k))(\underline{x}_2(k)+\overline{x}_2(k)) & \cdots & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_1(k)+\overline{x}_1(k))(\underline{x}_m(k)+\overline{x}_m(k)) \\ \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_2(k)+\overline{x}_2(k))(\underline{x}_1(k)+\overline{x}_1(k)) & \frac{1}{2n}\sum_{k=1}^{n}(\underline{x}_2^2(k)+\overline{x}_2^2(k)) & \cdots & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_2(k)+\overline{x}_2(k))(\underline{x}_m(k)+\overline{x}_m(k)) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_m(k)+\overline{x}_m(k))(\underline{x}_1(k)+\overline{x}_1(k)) & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_m(k)+\overline{x}_m(k))(\underline{x}_2(k)+\overline{x}_2(k)) & \cdots & \frac{1}{2n}\sum_{k=1}^{n}(\underline{x}_m^2(k)+\overline{x}_m^2(k)) \end{pmatrix} \tag{3.25}$$

### 3.3.2 Centers PCA

PCA for interval-valued data methods presented in literature are mainly based on the analysis of coded/transformed matrices from initial interval-valued data as vertices of hyper-cubes or as centers and radii. As an alternative to using the vertices of the hyper-cubes $H_k$ in VPCA case, the centers of the hyper-cubes could be used. The Centers PCA (CPCA) (Cazes et al., 1997),(Chouakria, 1998), decomposes the correlation matrix of the centers coded matrix $X^c$ and it projects the vertices as supplementary points in the factorial subspace. The interval-valued data matrix in 3.1 is coded in terms of centers by the following :

$$X^c = \begin{pmatrix} x_1^c(1) & . & . & x_m^c(1) \\ . & . & & . \\ . & & . & . \\ x_1^c(n) & . & . & x_m^c(n) \end{pmatrix} \tag{3.26}$$

Where $x_j^c(k)$ is the midpoint or center of the interval at hand given in definition 3.3. CPCA is then performed by conducting a classical PCA on the centers matrix $X^c$, the covariance matrix $\Sigma_C$ is calculated as:

$$\Sigma_C = \frac{1}{n-1} X^{cT} X^c \tag{3.27}$$

(Cazes et al., 1997) also provided a straightforward way of calculating covariance $\Sigma_C$ for the interval-valued matrix $[X]$, as:

$$\Sigma_C = \begin{pmatrix} \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_1(k)+\overline{x}_1(k))^2 & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_1(k)+\overline{x}_1(k))(\underline{x}_2(k)+\overline{x}_2(k)) & \cdots & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_1(k)+\overline{x}_1(k))(\underline{x}_m(k)+\overline{x}_m(k)) \\ \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_2(k)+\overline{x}_2(k))(\underline{x}_1(k)+\overline{x}_1(k)) & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_2(k)+\overline{x}_2(k))^2 & \cdots & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_2(k)+\overline{x}_2(k))(\underline{x}_m(k)+\overline{x}_m(k)) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_m(k)+\overline{x}_m(k))(\underline{x}_1(k)+\overline{x}_1(k)) & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_m(k)+\overline{x}_m(k))(\underline{x}_2(k)+\overline{x}_2(k)) & \cdots & \frac{1}{4n}\sum_{k=1}^{n}(\underline{x}_m(k)+\overline{x}_m(k))^2 \end{pmatrix} \tag{3.28}$$

The principal components of classical PCA on centers matrix, ${T_1}^c, ..., {T_m}^c$, are obtained via mapping using $T^c = X^c P$, given that $P = p_1, ..., p_m$ are the eigenvectors from spectral decomposition of matrix $\Sigma_C$. The interval-valued principal components are then constructed as:

$$\begin{cases} \underline{t}_j(k) = \sum\limits_{i=1,p_{qj}<0}^{m} \overline{x}_i(k)p_{ij} + \sum\limits_{i=1,p_{ij}>0}^{m} \underline{x}_i(k)p_{ij} \\ \overline{t}_j(k) = \sum\limits_{i=1,p_{ij}<0}^{m} \underline{x}_i(k)p_{ij} + \sum\limits_{i=1,p_{ij}>0}^{m} \overline{x}_i(k)p_{ij} \end{cases} \tag{3.29}$$

Where $p_{ij}$ is the $i$th element of the $j$th column of eigenvector matrix $P$. Similarly, the interval-valued estimates based on CPCA model for the first $(\ell)$ components are obtained as:

$$\begin{cases} \underline{\hat{x}}_j(k) = \sum\limits_{i=1,c_{ij}<0}^{m} \overline{x}_i(k)c_{ij} + \sum\limits_{i=1,c_{ij}>0}^{m} \underline{x}_i(k)c_{ij} \\ \overline{\hat{x}}_j(k) = \sum\limits_{i=1,c_{ij}<0}^{m} \underline{x}_i(k)c_{ij} + \sum\limits_{i=1,c_{ij}>0}^{m} \overline{x}_i(k)c_{ij} \end{cases} \tag{3.30}$$

Where $c_{ij}$ is the $i$th element of the $j$th column of matrix $C_\ell$.

Summing up, CPCA performed on interval input data is based on a numerical centers codification of the data, a treatment with classical PCA analysis technique, and finally a transformation of classical results into interval description.

### 3.3.3 Symbolic Covariance

An enhanced method based on symbolic covariance was introduced in (Le-Rademacher and Billard, 2012), that furthermore improves the performance of the above interval-valued PCA methods. The Idea is replacing the formal

covariance calculation of covariance (as in VPCA and CPCA approaches) with a new so-called symbolic covariance which is the maximum likelihood estimator of the overall variance of an interval-valued variable. Thus, the principal component space computed from the symbolic covariance matrix correctly accounts for the total variation of interval-valued data, and it is consistent with the symbolic data analysis framework.
Let $[X]$ be the interval data matrix defined in (3.1), and $W_j$ be a point of $[X_j]$, where $[X_j]\,(j = 1, ..., m)$ denotes the $j$-th variable of $[X]$. (Bertrand and Goupil, 2000) present the formulas for sample mean and sample variance of $W_j$, respectively, as:

$$\bar{W}_j = \sum_{k=1}^{n} \left(\underline{x}_j(k) + \overline{x}_j(k)\right) / (2n) \tag{3.31}$$

$$\mathrm{Var}(j,j) = \sum_{k=1}^{n} \frac{\left(\underline{x}_j^2(k) + \underline{x}_j(k)\overline{x}_j(k) + \overline{x}_j^2(k)\right)}{3n} - \left(\sum_{k=1}^{n} \frac{\left(\underline{x}_j(k) + \overline{x}_j(k)\right)}{2n}\right)^2 \tag{3.32}$$

The sample covariance of of $W_j$ and $W_j'$ defined by (Billard, 2008) is given by:

$$\begin{aligned} Cov(j,j') = \frac{1}{6n} \sum_{k=1}^{n} [2 \left(\underline{x}_j(k) - \bar{W}_j\right) \left(\underline{x}_{j'}(k) - \bar{W}_{j'}\right) + \left(\underline{x}_j(k) - \bar{W}_j\right) \left(\overline{x}_{j'}(k) - \bar{W}_{j'}\right) \\ + \left(\overline{x}_j(k) - \bar{W}_j\right) \left(\underline{x}_{j'}(k) - \bar{W}_{j'}\right) + 2 \left(\overline{x}_j(k) - \bar{W}_j\right) \left(\overline{x}_{j'}(k) - \bar{W}_{j'}\right)] \end{aligned} \tag{3.33}$$

For numerical data, or in other words for a trivial interval (definition 3.2), the symbolic covariance in (3.33) is equal to the classical covariance. The symbolic covariance computational complexity is of O(n), the same as for classical PCA

### 3.3.4 Midpoints-Radii PCA

The above presented methods of PCA for interval-valued data share the same approach, which is a suitable coding (vertices and centers coding) in order to define data structures to be handled with classical algorithms. However, The midpoints-radii PCA (MRPCA), introduced in (Palumbo and Lauro, 2003), approaches this matter in a different way. Problems with statistical analysis of interval data using standard interval arithmetic can be avoided by representing them using interval midpoints and ranges, i.e. $[X] = \{X^c, X^r\}$. The midpoints-radii PCA (MRPCA) on interval-valued data is a hybrid method that is resolved in terms of midranges ($X^r$) and midpoints ($X^c$), given by definitions 3.3 and 3.4, and their interconnection. MRPCA can be considered as an improvement over CPCA by including radius.

Considering the definition 3.10, the variance for interval-valued data, (Palumbo and Lauro, 2003), can be expressed by:

$$\sigma^2 = \frac{1}{n}\left[\sum_{k=1}^{n}\left(x_j^c(k) - m_j^c\right)^2 + \sum_{k=1}^{n}\left(x_j^r(k) - m_j^r\right)^2 + 2\sum_{k=1}^{n}\left|x_j^c(k) - m_j^c\right|\left|x_j^r(k) - m_j^r\right|\right] \tag{3.34}$$

The variance in 3.34 is thus defined as the sum of three components: the variance between midpoints, the variance between radii, and twice the measure of congruence between midpoints and radii. It follows that the global covariance matrix $\Sigma$ is given by:

$$\Sigma = \frac{1}{N}\left(X^{cT}X^c\right) + \frac{1}{N}\left(X^{rT}X^r\right) + \frac{1}{N}\left(\left|X^{cT}X^r\right| + \left|X^{rT}X^c\right|\right) \tag{3.35}$$

Thus, according to MRPCA, two independent PCA's are singly exploited on the two matrices of midpoints and radius. The solutions are given by the following eigen-systems:

$$X^c{\Sigma_c}^{-1}P^c = \Lambda^c P^c \tag{3.36}$$

$$X^r{\Sigma_r}^{-1}P^r = \Lambda^r P^r \tag{3.37}$$

Where $\Lambda^c, P^c$ and $\Lambda^r, P^r$ are, respectively, the eigenvalues and eigenvectors of the two partial eigen-decomposition of midpoints and radii matrices, and given that:

$$\Sigma_c = \frac{1}{N}\left(X^{cT}X^c\right) \;\; and \;\; \Sigma_r = \frac{1}{N}\left(X^{rT}X^r\right) \tag{3.38}$$

The two independent PCA's on midpoints and radius do not however cover the whole variance in 3.35. (Palumbo and Lauro, 2003) proposed a solution to include variance of the interconnection between midpoints and raddi given by the term $\left(\left|X^{cT}X^r\right| + \left|X^{rT}X^c\right|\right)$, and, at the same time, giving a graphical interpretation of the interval-valued units. The radius coordinates are thus rotated and then superimposed on the midpoints PC's as supplementary points, in order to get a logical graphical representation of the statistical units based on MRPCA model. Two choices for rotation of radius are defined below:

A first choice for computing the radius rotation matrix can be achieved by maximizing the Tucker congruence coefficient between midpoints and radii (Lauro et al., 2008), which is given by:

$$f(A) = \sum_{j=1}^{m}\left(\frac{\mathbf{a}_j^T X^{cT} X_j^r}{\left(\mathbf{a_j^T} X^{cT} X^c \mathbf{a_j}\right)^{1/2}\left(X_j^{rT} X_j^r\right)^{1/2}}\right) \tag{3.39}$$

Under the constraint $A^TA = 1$, and given that $A = [\mathbf{a}_1, \ldots \mathbf{a}_m]$ is the rotation matrix of radii, and $X_j^r$ denotes the radii of $j$-th interval-valued variable $[X_j]$.

A second choice for computing rotation matrix is expressed in terms of singular value decomposition, (Palumbo and Lauro, 2003).

$$X^{cT}X^{r} = P\Lambda^{cr}Q^{T} \tag{3.40}$$

where $\Lambda^{cr}$ is the eigenvalues matrix of covariance $X^{cT}X^{r}$. The matrix $A$ is defined as:

$$A = QP^{T} \tag{3.41}$$

So, let us consider that we have performed two classical PCA's on $X^{c}$ and $X^{r}$, then principal components of midpoints radii are given by:

$$\begin{cases} T^{c} = X^{c}P^{c} \\ T^{r} = X^{r}P^{r} \end{cases} \tag{3.42}$$

It follows that the construction of the interval-valued component vector $[\mathbf{t}(k)]$ using rotated ranges via matrix $A$ is:

$$\begin{cases} \underline{\mathbf{t}}(k) = \mathbf{t}^{c}(k) - A\mathbf{t}^{\mathbf{r}}(k) \\ \overline{\boldsymbol{t}}(k) = \mathbf{t}^{c}(k) + A\mathbf{t}^{\mathbf{r}}(k) \end{cases} \tag{3.43}$$

Hence, for a number of components $\ell$, the interval-valued estimates of vector $[\mathbf{x}(k)]$ are obtained, similarly, by projecting rotated ranges estimates $\hat{\mathbf{x}}^{r}(k)$ on the estimated midpoints $\hat{\mathbf{x}}^{c}$, as:

$$\begin{cases} \underline{\hat{\mathbf{x}}}(k) = \hat{\mathbf{x}}^{c}(k) - A\hat{\mathbf{x}}^{r}(k) \\ \overline{\hat{\mathbf{x}}}(k) = \hat{\mathbf{x}}^{c}(k) + A\hat{\mathbf{x}}^{r}(k) \end{cases} \tag{3.44}$$

### 3.3.5 Complete Information PCA

In order to seize more information within interval observations, Huiwen et al. (Wang et al., 2012) proposed a new interval PCA method called complete information based PCA (CIPCA). By defining the inner product of hyper-cubes divided into informative grid data, and based on a rather analytic approach, CIPCA accomplishes the derivation of interval-valued principal components and transforms PCA modeling into the computation of some inner products. Thus, leading to more accurate results and providing an efficient and effective way for conducting PCA on large-scaled interval data.

According to (Wang et al., 2012), we define the following:

**Definition 3.11 : Inner product of interval-valued variables**
*given any two interval-valued variables* $[X_j]$ *and* $[X_{j'}]$*, the inner product is defined*

*as:* $\langle [X_j], [X_{j'}] \rangle = \sum_{k=1}^{n} \langle [x_j(k)], [x_{j'}(k)] \rangle$ *where:*
$\langle [x_j(k)], [x_{j'}(k)] \rangle = \frac{1}{4} \left( \underline{x}_j(k) + \overline{x}_j(k) \right) \left( \underline{x}_{j'}(k) + \overline{x}_{j'}(k) \right)$

**Definition 3.12 : Interval squared norm**
*In the case of auto-correlation given by* $\langle [X_j], [X_j] \rangle$ *the inner product is equal to the squared norm extended to the interval case* $\|[X_j]\|^2$*, defined by:* $\|[X_j]\|^2 = \sum_{k=1}^{n} \|[x_j(k)]\|^2$*, where:*
$\|[x_j(k)]\|^2 = \frac{1}{3} \left( \underline{x}^2{}_j(k) + \underline{x}_j(k)\overline{x}_j(k) + \overline{x}^2{}_j(k) \right)$

Based on the above definitions of interval norm and inner product, and with Matrix $[X]$ being normalized, the covariance matrix $\Sigma$ of $X_{n \times m}$ is given by:

$$\Sigma = \frac{1}{n} \begin{pmatrix} \langle [X_1], [X_1] \rangle & \langle [X_1], [X_2] \rangle & \cdots & \langle [X_1], [X_m] \rangle \\ \langle [X_2], [X_1] \rangle & \langle [X_2], [X_2] \rangle & \cdots & \langle [X_2], [X_m] \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle [X_m], [X_1] \rangle & \langle [X_m], [X_2] \rangle & \cdots & \langle [X_m], [X_m] \rangle \end{pmatrix} \tag{3.45}$$

The determination of the interval-valued principal components $[T]$ in CIPCA method is based on a linear combination algorithm for interval-valued variables first developed by Moore (Moore, 1966), (see appendix A), also adopted in VPCA and CPCA for the projection procedure.

Let us consider that we performed an eigen-decomposition of the covariance matrix $\Sigma$, where $\lambda_1, ..., \lambda_m$ and $p_1, ..., p_m$ are the resulting eigenvalues and eigenvectors respectively. The interval-valued principal components are thus computed as

$$\begin{cases} \underline{t}_j(k) = \sum_{i=1}^{m} p_{ij} \left( \tau \underline{x}_i(k) + (1 - \tau) \overline{x}_i(k) \right) \\ \overline{t}_j(k) = \sum_{i=1}^{m} p_{ij} \left( (1 - \tau) \underline{x}_i(k) + \tau \overline{x}_i(k) \right) \end{cases} \tag{3.46}$$

with

$$\tau = \begin{cases} 0, & p_{ij} \leq 0, \\ 1, & p_{ij} \geq 0 \end{cases}$$

It follows that interval-valued estimates from CIPCA model are obtained as:

$$\begin{cases} \underline{\hat{x}}_j(k) = \sum_{q=1}^{m} C_{\ell qj} \left( \tau \underline{x}_q(k) + (1 - \tau) \overline{x}_q(k) \right) \\ \overline{\hat{x}}_j(k) = \sum_{q=1}^{m} C_{\ell qj} \left( (1 - \tau) \underline{x}_q(k) + \tau \overline{x}_q(k) \right) \end{cases} \tag{3.47}$$

with the same condition on $\tau$, and given that $C_\ell = P_\ell P_\ell{}^T$, and $\ell$ is the number of chosen principal components.

### 3.3.6 Determining the Number of PC's for interval-Valued PCA

The number of principal components has a significant impact on each step of the PCA based sensor FDI scheme as well as it's performances. Various methods for determining the number of PC's exist in the literature that are detailed in section 1.3, which include: the predicted residual error sum of squares ($PRESS$) or cross validation criterion, the average eigenvalues, The cumulative percentage of variance ($CPV$), and the variance of reconstruction error ($VRE$). Only, these methods were developed for single-valued PCA approach and do not cover the interval-valued PCA case. However, The interval-valued PCA methods detailed above are based on various approximations, and the resulting covariance matrix, eigenvalues and eigenvector are single-valued. Thus, classical methods for determining the number of principal components could be employed for interval-valued PCA, that is, due to the fact that all the methods for selecting PC's depend partially or globally on the triplet {Covariance, Eigenvalues, Eigenvectors}. The only exclusion is the variance of reconstruction error criterion (VRE), which more-or-less does not necessarily apply for the interval-valued PCA case, for a reason that we will discuss further.

Perhaps the most suitable and accurate approach for the determination of the PC's, for diagnosis purpose, is the variance of reconstruction error introduced in (Dunia and Joe Qin, 1998). It's principle consists in estimating a variable from other process variables using the PCA model. The reconstruction accuracy is thus related to the capacity of the PCA model to reveal the redundancy relations among the variables. The VRE criterion defines the variance $\rho_i(\ell)$ which is the variance of reconstruction error with respect to the number of components $\ell$, given by:

$$\rho_i(\ell) = Var\left\{\xi_j^T(\mathbf{x}(k) - \mathbf{x}^{(i)}(k))\right\} = \frac{\tilde{\xi}_j^T \Sigma \tilde{\xi}_j}{(\tilde{\xi}_j^T \tilde{\xi}_j)^2} \tag{3.48}$$

However, in the case of interval-valued data, the equation 3.48 does not account for the actual reconstruction error of the interval-valued PCA model. The reason is that the estimation process for classical PCA, and that of interval-valued PCA (in its different variations) are not the same. In other words, the term $\frac{\tilde{\xi}_j^T \Sigma \tilde{\xi}_j}{(\tilde{\xi}_j^T \tilde{\xi}_j)^2}$ does not correspond to the variance of the reconstruction error. Thus, an actual reconstruction based on interval-valued PCA have to be performed via the relation $\left\{\xi_j^T(\mathbf{x}(k) - \mathbf{x}^{(i)}(k))\right\}$.

In this section, we propose an extension of the reconstruction principle for the interval-valued PCA case, and we derive the variance of interval reconstruction error criterion ($VIRE$) for the determination of the number of PC's to be retained in interval-valued PCA model.

For classical PCA, the reconstruction of the ith variable of a vector $\mathbf{x}(k) = [x_1(k)\, x_2(k)\, \ldots\, x_m(k)]^T$ is given by:

$$z_i\,(k) = \frac{[c_{-i}^T \quad 0 \quad c_{+i}^T]}{1 - c_{ii}}\mathbf{x}(k) \tag{3.49}$$

Where $z_i$ is the reconstructed value of measurement $x_i$ and $c_{-i}$, $c_{+i}$ denote, respectively, the first $(i-1)$ and last $(m-i)$ elements of the ith column of matrix $C_\ell$. Let us consider $\mathbf{x}^{(i)}(k) = [x_1(k) \ldots x_{i-1}(k)\, z_i(k)\, x_{i+1}(k) \ldots x_m(k)]^T$, the reconstructed vector can also be formulated as:

$$\mathbf{x}^{(i)}(k) = G^{(i)}\mathbf{x}(k) \tag{3.50}$$

The projection matrix $G^{(i)}$ is expressed in terms of $C_\ell$ as:

$$G^{(i)} = I + \frac{\xi_i {\xi_i}^T}{1 - {\xi_i}^T C_\ell \xi_i}\,(C_\ell - I) \tag{3.51}$$

Where $\xi_i$ is the vector of reconstruction direction with all elements equal to 0 except the ith which value is 1.

For the interval case, calculation of the elements of $[\mathbf{x}^{(i)}(k)]$ depends on the used interval PCA method. In the following, we introduce the extension of the variable reconstruction, for the four presented interval PCA methods, using the projection matrix $G^{(i)}$ for the first $(\ell)$ PCs.

#### 3.3.6.a Reconstruction of variables for VPCA model

For VPCA, a classical PCA is conducted on the the vertices matrix V given in (3.14). Let us consider an interval vector $[\mathbf{x}(k)] = [[x_1(k)]\,[x_2(k)], \ldots, [x_m(k)]]$, and $V(k)$ be its corresponding vertices matrix of $2^m$ lines and m columns. Let $V^{(i)}(k)$ be the reconstructed matrix, in the classical way given in (3.50), performed on the vertices matrix $V(k)$ as:

$$V^{(i)}(k) = G^{(i)}V(k) \tag{3.52}$$

$$\begin{aligned} V(k) &= [V_1(k) \ldots V_i(k) \ldots V_m(k)] \\ &= \begin{bmatrix} \underline{x}_1(k) & \ldots & \underline{x}_i(k) & \ldots & \underline{x}_m(k) \\ . & . & . & . & . \\ . & . & . & . & \\ \overline{x}_1(k) & . & \overline{x}_i(k) & . & \overline{x}_m(k) \end{bmatrix} \end{aligned} \tag{3.53}$$

The interval reconstruction $[z_i(k)] = [\underline{z}_i(k), \overline{z}_i(k)]$ of an interval measurement $[x_i(k)]$ is then given by:

$$\begin{cases} \underline{z}_i(k) = \min V_i^{(i)}(k) \\ \overline{z}_i(k) = \max V_i^{(i)}(k) \end{cases} \tag{3.54}$$

where $V_i^{(i)}(k)$ is the $i$th column of the reconstructed vertices matrix $V^{(i)}(k)$.

#### 3.3.6.b Reconstruction of variables for CPCA model

Similarly, and based on projection matrix $G^{(i)}$ obtained from classical PCA on centers matrix $X^c$ using (3.51). The reconstructed measurement $[z_i(k)]$ for C-PCA is given by the following relations:

$$\begin{cases} \underline{z}_i(k) = \sum\limits_{q=1, G_q^{(i)}<0}^{m} \overline{x}_q(k) G_q^{(i)} + \sum\limits_{q=1, G_q^{(i)}>0}^{m} \underline{x}_q(k) G_q^{(i)} \\ \overline{z}_i(k) = \sum\limits_{q=1, G_q^{(i)}<0}^{m} \underline{x}_q(k) G_q^{(i)} + \sum\limits_{q=1, G_q^{(i)}>0}^{m} \overline{x}_q(k) G_q^{(i)} \end{cases} \tag{3.55}$$

#### 3.3.6.c Reconstruction of variables for MRPCA model

For the MR-PCA case, and by adding the rotated radii points using rotation matrix $A$, we compute the reconstructed interval measurement $[z_i(k)]$. Let us consider that $c_{c-i}$ and $c_{r-i}$ denote, respectively, the first $(i-1)$ elements of the $i$-th column of matrices ${C_\ell}^c$ and ${C_\ell}^r$. And let $c_{c+i}$ and $c_{r+i}$ denote, respectively, the last $(m-i)$ elements of the $i$-th column of matrices ${C_\ell}^c$ and ${C_\ell}^r$.

$$\begin{cases} \mathbf{g}_c^{(i)} = \frac{[c_{c-i}^T \quad 0 \quad c_{c+i}^T]}{1 - c_{ii}^c} \\ \mathbf{g}_r^{(i)} = \frac{[c_{r-i}^T \quad 0 \quad c_{r+i}^T]}{1 - c_{ii}^r} \end{cases} \tag{3.56}$$

Thus, reconstructed interval-valued measurement $[z_i(k)]$ for MRPCA model, is defined as

$$\begin{cases} \underline{z}_i(k) = \mathbf{g}_c^{(i)} \mathbf{x}^c(k) - \mathbf{g}_r^{(i)} (A \mathbf{x}^{\mathbf{r}}(k)) \\ \overline{z}_i(k) = \mathbf{g}_r^{(i)} \mathbf{x}^c(k) + \mathbf{g}_r^{(i)} (A \mathbf{x}^{\mathbf{r}}(k)) \end{cases} \tag{3.57}$$

#### 3.3.6.d Reconstruction of variables for CIPCA model

For the CIPCA model the reconstruction $[z_i(k)]$ is given by:

$$\begin{cases} \underline{z}_i(k) = \sum\limits_{q=1}^{m} G_q^{(i)} \left( \tau \underline{x}_q(k) + (1-\tau)\, \overline{x}_q(k) \right) \\ \overline{z}_i(k) = \sum\limits_{q=1}^{m} G_q^{(i)} \left( (1-\tau)\, \underline{x}_q(k) + \tau \overline{x}_q(k) \right) \end{cases} \tag{3.58}$$

with

$$\tau = \begin{cases} 0, \ G_q^{(i)} \leq 0, \\ 1, \ G_q^{(i)} \geq 0 \end{cases}$$

#### 3.3.6.e Variance of interval reconstruction error

Once the interval reconstruction $[z_i(k)] = [\underline{z}_i(k), \overline{z}_i(k)]$ is obtained for the $i$th variable and for a fixed number of components $\ell$, the interval reconstruction error $[e_i(k)]$ can be calculated as:

$$[e_i(k)] = \{ \xi_i^T \left( [\mathbf{x}(k)] - [\mathbf{x}^{(i)}(k)] \right) \} \tag{3.59}$$

Where $[\mathbf{x}(k)]$ is an interval data sample of $m$ sensors,
$[\mathbf{x}^{(i)}(k)] = [[x_1(k)] \ \ldots \ [z_i(k)] \ \ldots \ [x_m(k)]]$ is the reconstructed vector with $[z_i(k)]$ its $i$th interval reconstruction, and $[e_i(k)] = [\underline{e}_i(k), \overline{e}_i(k)]$. Afterwards, a calculation of the variance for the interval reconstruction error is performed as:

$$\rho_i(\ell) = \frac{1}{n} \sum_{k=1}^{n} \| [e_i(k)] \|^2 \tag{3.60}$$

where

$$\| [e_i(k)] \|^2 = \frac{1}{3} (\underline{e}_i^2(k) + \underline{e}_i(k) \overline{e}_i(k) + \overline{e}_i^2(k)) \tag{3.61}$$

Note that the variance of reconstruction error can also be computed based on the symbolic variance in 3.32. Accordingly, and based on the new $VIRE$ criterion, the determination of the number of components can be formulated by the following optimization problem with respect to the number of PC's $\ell$ :

$$VIRE = \sum_{i=1}^{m} \rho_i(\ell) \tag{3.62}$$

As in the classical VRE case, the number of principal components to be kept in the interval-valued PCA model corresponds to the minimum value of VIRE with respect to $\ell$, which ensures the minimum reconstruction error of the interval-valued PCA model.

## 3.4 Interval-valued PCA for Fault Detection

In the case of single-valued data, PCA based FDI aims to detect deviations from typical process behaviour by evaluating whether a process is statistically in control or not. This is done by analysing different projections of the new process data, into the principal or residual sub-spaces, which are obtained according to the PCA model, established based on data acquired from the different process sensors in NOC. Application of PCA-based fault detection involves using several statistics for monitoring the condition of the process. The most used statistics are the Hotteling's $T^2$ test, the squared prediction error $SPE$ and the squared weighted error $SWE$ (Qin, 2003).

In terms of data mining, a sampling period of time is usually determined according to the nature of the process, and a measurement is considered as the average of the collected measured values during that sampling period. When this type of data is used for a PCA based monitoring, an excessive rate of false alarms and missing detection may occur depending on the averaging period of time, due to information loss. In addition to that, to obtain a reliable PCA model representing the process, we need to use a good amount of samples, which are not always available.
However, describing the measurement as an interval-valued unit helps covering the whole averaging period, and thus yields more information about the process than the averaged measure, even for a few samples. This quality of information gathered from the process positively impacts on the PCA model for interval-valued data, which becomes consequently more robust to uncertainties and even to slight process variations.

The first attempts of adapting PCA for interval-valued data to FDI go back to the work of (Benaicha et al., 2013). The authors developed a dedicated interval-valued PCA method which is based on an analytic approach. The model is constructed based on modelling of eigenvalues and eigenvectors variations, and has somehow limited applicability, as is the case for analytic interval-valued PCA methods. In this section, we propose several fault detection strategies and fault detection indices based on PCA for interval-valued data methods, to achieve a maximum robustness toward uncertainties.

### 3.4.1 Univariate Chart

One way of approaching PCA based fault detection, is by separately analysing the residuals. Let us first define the interval-valued PCA estimation error, or residuals. Based upon the different methods presented above, interval-valued residuals $\underline{\mathbf{e}}(k)$ and $\overline{\mathbf{e}}(k)$ can be obtained via the difference between interval-valued estimates given in (3.24)(3.30)(3.44)(3.47) and original variables, or by

projection on last $(m - \ell)$ components, (Ait-Izem et al., 2014a),(Ait-Izem et al., 2014b), (Ait-Izem et al., 2015a), as:

$$\begin{cases} \underline{\mathbf{e}}(k) = \underline{\mathbf{x}}(k) - \overline{\hat{\mathbf{x}}}(k) \\ \overline{\mathbf{e}}(k) = \overline{\mathbf{x}}(k) - \underline{\hat{\mathbf{x}}}(k) \end{cases} \tag{3.63}$$

In classical PCA, unusual events are projected onto residual space and can therefore be detected using several statistics. The same principle can be applied in the interval-valued PCA case. However, an interesting property of interval-valued residuals makes it possible to achieve proper fault detection in univariate case, i.e. without use of detection statistics, that will be discussed further.

In order to demonstrate the use of the interval-valued residuals in fault detection, let us consider the previous simulation example 1 based on 6 variables $j = 1, \ldots, 6$, let us also recall that this data are based on two variables generated from two normal distributions, and four linear analytic redundancy relations between the remaining variables. To simulate the presence of measurement uncertainties, a variation $\delta X$ was added to the generated data. This variation is a realization of centered variables corresponding to 10% of the variation range of each variable. Thus, we obtain the new interval data matrix $[X] = [X - \delta X, X + \delta X]$ as defined in equation 3.1. The interval-valued variables of data in example 1 are presented in Figure 3.2 for 100 samples, where we can clearly see the interval tendency of data.

After measurements normalisation to zero mean and unit variance, according to section 3.2.2, the eigen-decomposition of the covariance matrix of $[X]$ allows the determination of its eigenvalues and eigenvectors, i.e. the interval-valued PCA model according to the various methods described previously. The number $\ell$ of components to be retained for the interval-valued PCA models are determined using the VIRE criterion, where $\ell$ is chosen as the value that minimizes $VIRE(\ell)$. Figure 3.3 represents the evolution of the variance of interval reconstruction error in terms of the number of PC's $\ell$. Accordingly, $\ell = 2$ principal axes are kept in the interval PCA model, which proves the accuracy of the method, since the simulation example is based on 2 variables.

Since we are using the a PCA for interval-valued model, let us first inspect the structure or nature of residuals. Figure 3.4 represent the interval-valued residuals obtained through an interval-valued PCA model. We note that the upper and lower bounds of residuals , which are, respectively, residuals of upper and lower estimations of interval data matrix, are forming an envelope around the zero line.

**Remark.** *the nature of interval-valued residuals is the same for all the interval-valued PCA models presented in this manuscript. Thus, we choose to represent results for a CIPCA model due to its high accuracy, a comparison between the various models in terms of performances for fault detection will be further discussed in chapter 4.*

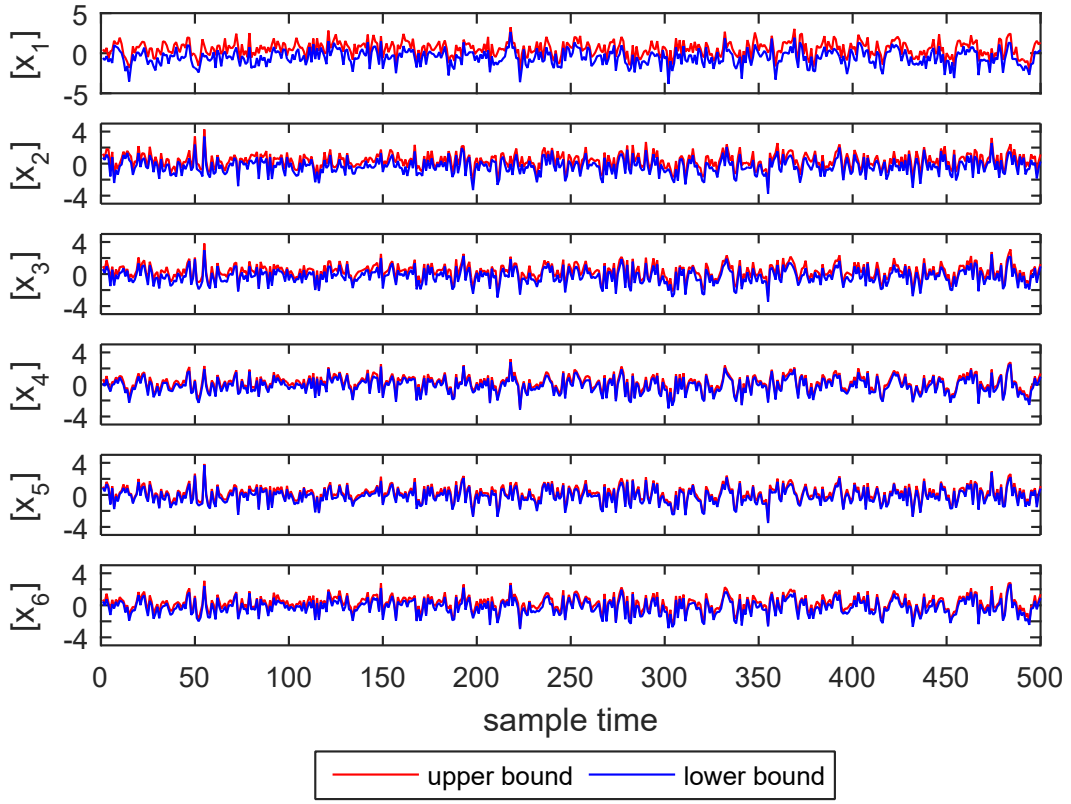



FIGURE 3.2 – Generated Interval-valued variables of Example 1

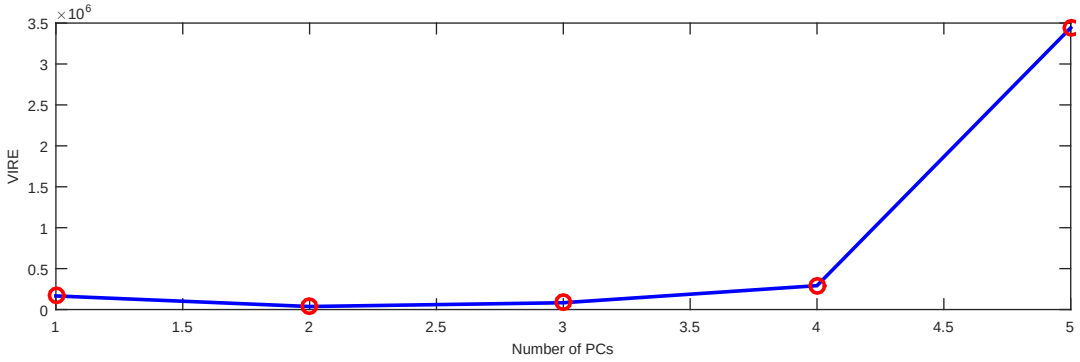



FIGURE 3.3 – Evolution of VIRE in terms of the number of PC's

The system in example 1 is simulated a second time on 500 samples, then, two offsets are added to the data, as follows:

— an offset $\zeta_1(k)$ added to variable $x_1(k)$ from time sample 120 to 150, simulated as constant bias of amplitude equal to 5% of variation range of variable $x_1(k)$, i.e. $\zeta_1(k) = x_1(k) + (x_1(k) \times 15\%)$. This offset is 5% lower than the variable range and represents then an uncertainty of measurement. Or, $-\delta x_1(k) < \zeta_1(k) < \delta x_1(k)$.

— an offset $f_4(k)$ added to variable $x_4(k)$ from time sample 300 to 340, simulated as a constant bias of amplitude equal to 15% of variation range of variable $x_4(k)$. i.e. $f_4(k) = x_4(k) + (x_4(k) \times 5\%)$. $f_4(k)$ is 5% higher

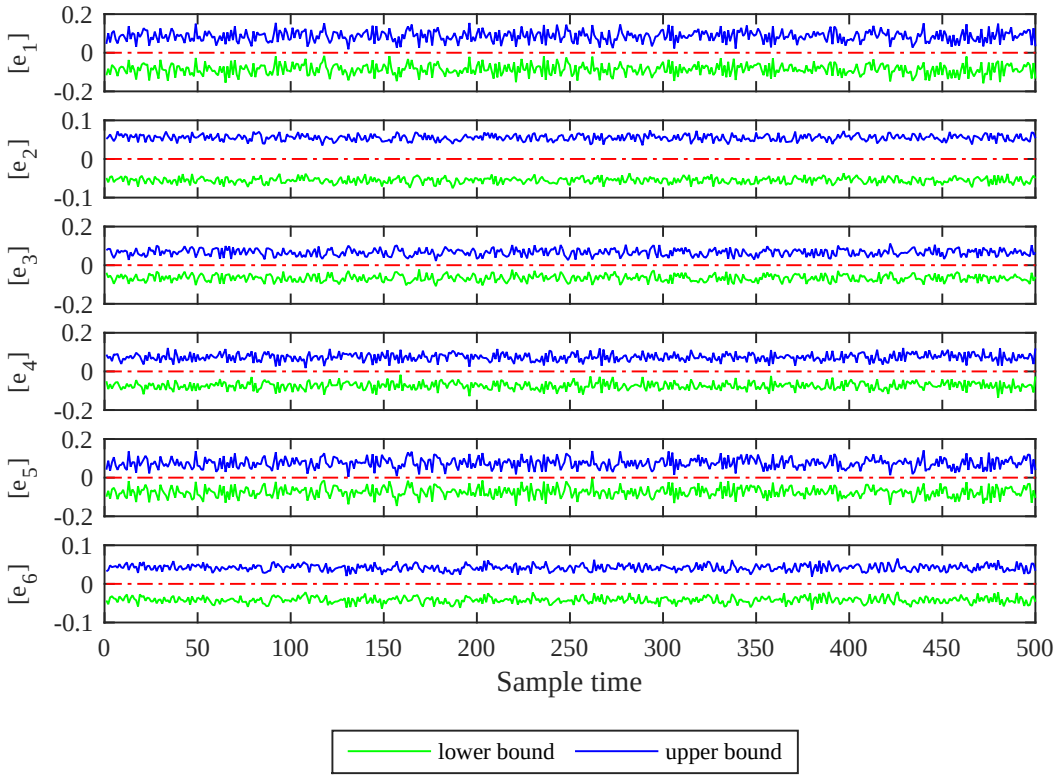
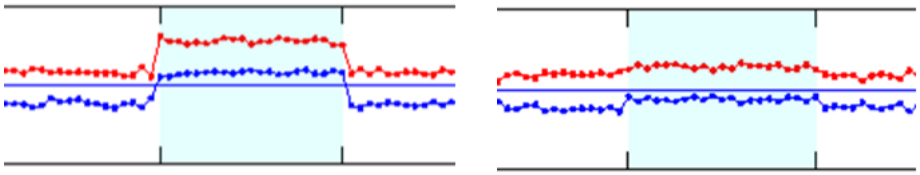



FIGURE 3.4 – Interval-valued Residuals for Example 1

than the variable range and represents then a fault. In other words, $\delta x_4(k) < f_4(k) < -\delta x_4(k)$

For the new univariate fault detection case, an abnormal behavior cannot be considered as a fault unless one of the residuals bounds changes sign. In other words, the envelope created by residuals is a safe zone spanned by the interval of uncertainty in which every unusual event is not considered as a fault but as an uncertainty, as presented in Figure 3.5.

We can clearly notice the behavior of the residuals in Figure 3.6, which represents



(A) Real fault (B) Uncertainty

FIGURE 3.5 – Univariate Fault detection method using PCA for interval-valued data

the interval-valued residuals in faulty case. The highlighted areas correspond to the zones where offsets have been added, where as stated previously, an uncertainty is present from moment 120 to 150, and a fault from 300 to 340. In

the uncertainty zone, the residual envelope slightly changes in different directions (positive or negative). In other words, both bounds of the residual shift in a direction, but do not however exceed the middle zero line (no sign change in residuals). In the fault occurrence zone, we note the sign change of the lower bound in residual $[e_4(k)]$, which confirms a faulty condition. in mathematical terms we can express the univariate interval-valued PCA detection method as follows :

— Process is faulty, if:

$$0 \notin \left[ \begin{array}{cc} \underline{e}_j(k) & \overline{e}_j(k) \end{array} \right] , \quad j = 1, ..., m \tag{3.64}$$

or

$$\underline{e}_j(k) \times \overline{e}_j(k) > 0 , \quad j = 1, ..., m \tag{3.65}$$

— Process is healthy, if:

$$0 \in \left[ \begin{array}{cc} \underline{e}_j(k) & \overline{e}_j(k) \end{array} \right] , \quad j = 1, ..., m \tag{3.66}$$

or

$$\underline{e}_j(k) \times \overline{e}_j(k) < 0 , \quad j = 1, ..., m \tag{3.67}$$

However, this method is restricted to processes with small number of variables, as it can be computationally cumbersome for large processes.

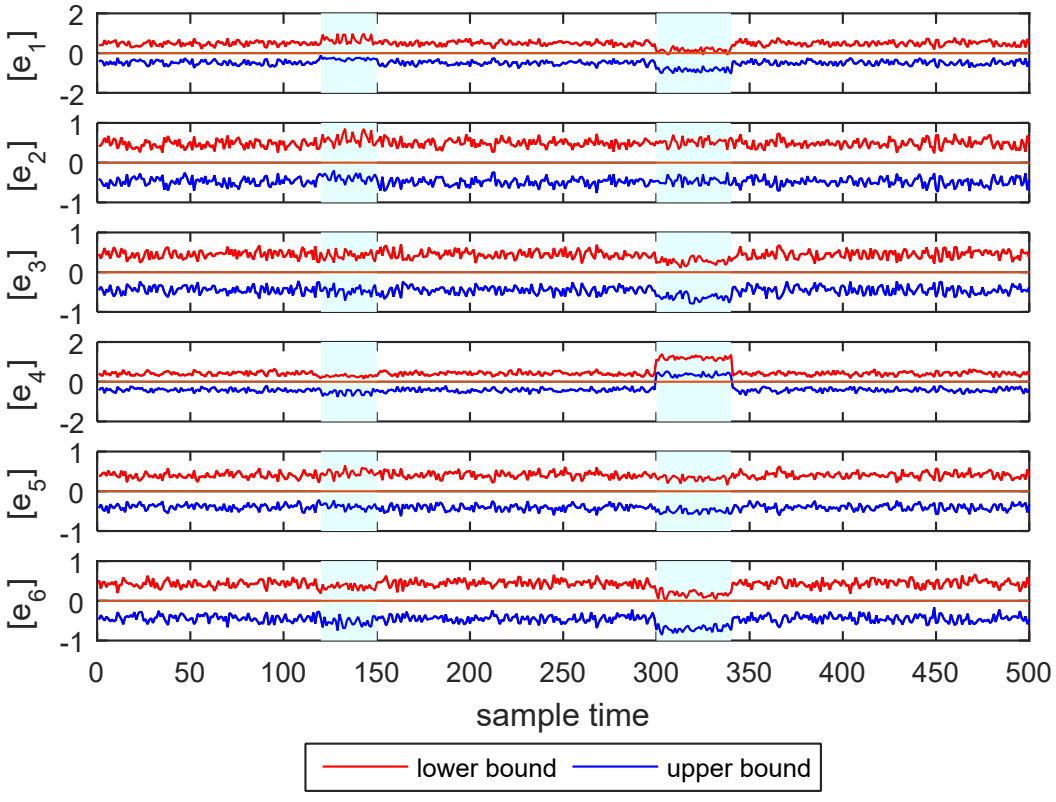



FIGURE 3.6 – Interval-valued Residuals for Example 1 in faulty case

### 3.4.2 Multivariate Charts

In classical PCA based fault detection, several statistics that measure variations in different projection sub-spaces are used in multivariate fault detection based on PCA model. The most common statistics used as fault indicator being the $SPE$, $T^2$ and $SWE$, as detailed in section 2.2. For interval-valued data, and based on an interval-valued PCA model, an extension of these indicators is required in order to handle the new nature of data.

#### 3.4.2.a Interval-valued SPE

For the interval-valued case, the proposed form of interval-valued $SPE$ calculation have been introduced in (Ait-Izem et al., 2015c). This statistic is calculated based on residuals as in the classical case, thus, yielding an interval $[SPE]$ with an upper and a lower bound, corresponding respectively to the upper and lower bounds of the calculated residuals, as:

$$\begin{cases} \underline{SPE}(k) = \|\underline{\mathbf{e}}(k)\|^2 = \underline{\mathbf{e}}(k)^T\underline{\mathbf{e}}(k) \\ \overline{SPE}(k) = \|\overline{\mathbf{e}}(k)\|^2 = \overline{\mathbf{e}}(k)^T\overline{\mathbf{e}}(k) \end{cases} \tag{3.68}$$

Given that $\underline{\mathbf{e}}(k) = [\underline{e}_1(k), \ldots, \underline{e}_j(k), \ldots, \underline{e}_m(k)]$ and $\overline{\mathbf{e}}(k) = [\overline{e}_j(k), \ldots, \overline{e}_j(k), \ldots, \overline{e}_m(k)]$, and $[SPE](k) = [\underline{SPE}(k), \overline{SPE}(k)]$.

#### 3.4.2.b Interval-valued $T^2$ and $SWE$

Based on the interval-valued PCA model, the decomposition of principal components, explained for each interval-valued PCA model in (3.22) (3.29) (3.42) (3.46), is given by:

$$[\hat{\mathbf{t}}(k)] = [[\underline{t}_1(k), \overline{t}_1(k)] \ \ldots \ [\underline{t}_\ell(k), \overline{t}_\ell(k)]] \tag{3.69}$$

$$[\tilde{\mathbf{t}}(k)] = [[\underline{t}_{\ell+1}(k), \overline{t}_{\ell+1}(k)] \ \ldots \ [\underline{t}_m(k), \overline{t}_m(k)]] \tag{3.70}$$

It is well known that the calculation of the Hotelling's $T^2$ and the $SWE$ statistics require both a weighting by the first $\ell$ and last $(m-\ell)$ eigenvalues, respectively. One may argue that the available eigenvalues from the interval PCA models presented hereby are single-valued, which brings inconsistency to the theory of the interval statistics. However, the real meaning of the eigenvalues in this case is to represent residuals variances. Thus, for the interval case, and for each statistic ($[T^2]$ and $[SWE]$), eigenvalues are calculated as the corresponding empirical variances for the given bound, i.e. the variance of the interval principal components $\underline{\hat{\mathbf{t}}}(k)$ and $\overline{\hat{\mathbf{t}}}(k)$ for the Hotelling $[T^2]$ statistic, and the variance of the interval residuals $\underline{\tilde{\mathbf{t}}}(k)$ and $\overline{\tilde{\mathbf{t}}}(k)$for the $[SWE]$ statistic.

For the first interval principal components, the interval $[T^2]$ statistic is calculated for the two bounds trough a combination of interval eigenvalues and interval principal components as in the classical way described in (2.7) by:

$$\begin{cases} \underline{T}^2(k) = \hat{\underline{\mathbf{t}}}(k)^T \underline{\Lambda}_\ell^{-1} \hat{\underline{\mathbf{t}}}(k) \\ \overline{T}^2(k) = \hat{\overline{\mathbf{t}}}(k)^T \overline{\Lambda}_\ell^{-1} \hat{\overline{\mathbf{t}}}(k) \end{cases} \tag{3.71}$$

Similarly, and according to (2.10), the interval $[SWE]$ statistic is given by:

$$\begin{cases} \underline{SWE}(k) = \tilde{\underline{\mathbf{t}}}(k)^T \underline{\Lambda}_{m-\ell}^{-1} \tilde{\underline{\mathbf{t}}}(k) \\ \overline{SWE}(k) = \tilde{\overline{\mathbf{t}}}(k)^T \overline{\Lambda}_{m-\ell}^{-1} \tilde{\overline{\mathbf{t}}}(k) \end{cases} \tag{3.72}$$

#### 3.4.2.c Thresholds for interval-valued statistics

We propose to use the limit presented in (Jackson and Mudholkar, 1979) and extend it to interval-valued data case in order to compute the $[SPE]$, $[T^2]$ and $[SWE]$ control limits based on Box's quadratic form approximation (Box, 1954). as in the classical case given in 2.14. Hence, the limits for the corresponding index can be computed based on its estimate mean $(a)$ and estimate variance $(b)$ so that,

$$\delta_\alpha^2 = g^{index} \chi^2_{h^{index},\alpha} \tag{3.73}$$

where $g$ and $h$ can be approximated as $b/2a$ and $2a^2/b$, respectively. Note that control limits for the presented interval statistic are calculated for each bound of the statistic at hand. However, the obtained thresholds (for upper and lower bound) tend to be equal due to the symmetry of bounds, which leaves us with one control limit for the interval index.
To illustrate the use of these statistics, let us consider the previous interval-valued data generated from example 1. We then inject two faults in two different moments: to variable $x_1(k)$ from time sample 120 to 150, and to variable $x_4(k)$ from time sample 300 to 340. These two faults are represented, respectively, by a constant bias of amplitude equal to 20% of variation range of the corresponding variable.

In interval PCA based FDI, or generally speaking, in interval approaches based FDI, a fault is detected if both bounds of the interval statistics exceed the detection threshold. However, the case where only one bound of the statistic exceeds the control threshold is not considered as a certain case of fault occurrence (Benaicha et al., 2013). This scenario can be illustrated in Figure 3.7 regarding both faults, where only the lower $[SPE]$ bound allows the detection of the fault. Also the second fault scenario in Figure 3.9, for the $[SWE]$, which presents the same ambiguous decision. Thus, the detection rate of these interval statistics drastically decreases, once we consider this detection condition based on both bounds of the control statistic. A quick analysis of fault detection using the three interval statistics, represented by Figures 3.7, 3.8 and 3.9 reveals some

inaccuracy of the used control charts, as we note a great number of missed detection and false alarms.

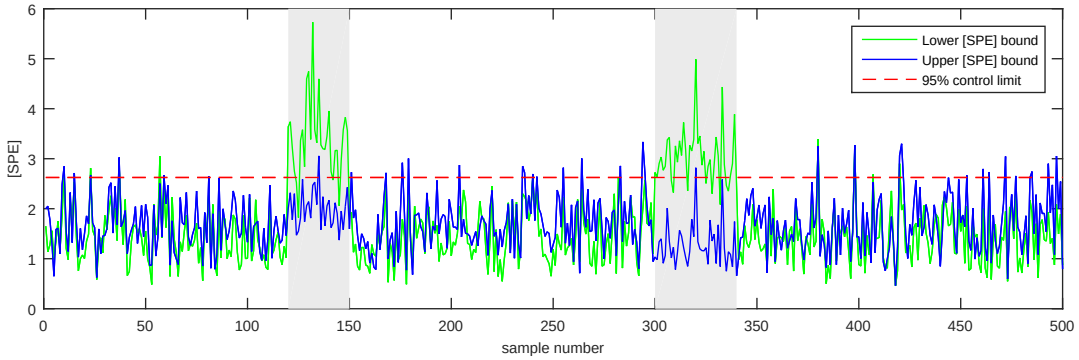



FIGURE 3.7 – Time evolution of [SPE] index

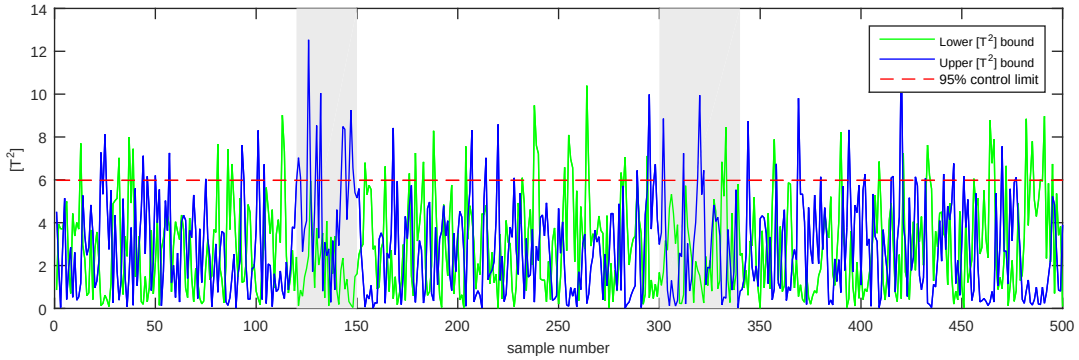



FIGURE 3.8 – Time evolution of $[T^2]$ index

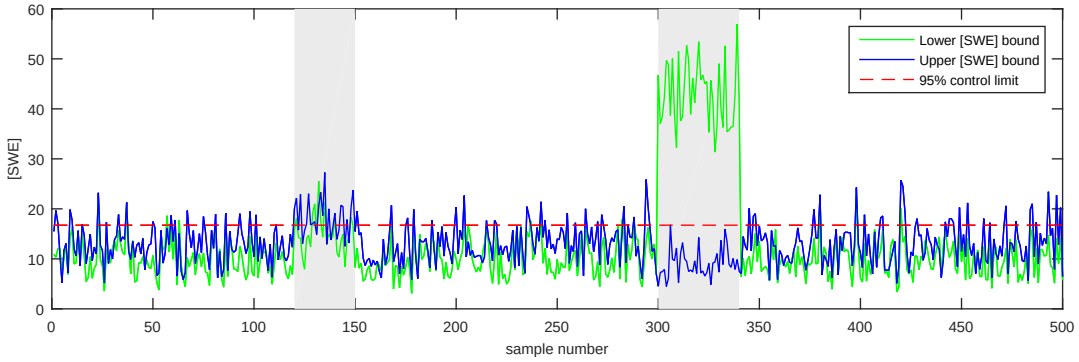



FIGURE 3.9 – Time evolution of [SWE] index

**Remark:** It has been verified that based on an interval-valued PCA model, constructed from NOC data, the calculated thresholds (upper and lower threshold) for interval statistics $[SPE]$, $[T^2]$ and $[SWE]$ tend to be equal. Detection decisions are then made regarding the two bounds of the statistic but using only one threshold.

#### 3.4.2.d New Interval fault detection indices

The motivation behind developing a new interval fault detection index, is mainly the ambiguity in the fault detection decision due to the interval nature of the statistics presented above. That is, when each bound yields a different decision from the other regarding the occurrence of faults.

Since, the classical squared prediction error $SPE$, used as an index for fault detection based PCA approach, is a norm of residuals vector, the alternative form of this statistic norm for the interval-valued data may be achieved using the interval squared norm formula (3.12).

Let us consider a new interval measurement vector given by:

$$[\mathbf{x}(k)] = [[\underline{x}_1(k), \overline{x}_1(k)]\ [\underline{x}_2(k), \overline{x}_2(k)]\ \ldots\ [\underline{x}_m(k), \overline{x}_m(k)]] \tag{3.74}$$

The proposed interval fault detection index is defined as the interval squared prediction error ($ISPE$) and is given by:

$$ISPE(k) = \|[\mathbf{e}(k)]\|^2 = \sum_{j=1}^{m} \|[e_j(k)]\|^2 \tag{3.75}$$

where

$$\|[e_j(k)]\|^2 = \frac{1}{3}\left(\underline{e}_j^2(k) + \underline{e}_j(k)\overline{e}_j(k) + \overline{e}_j^2(k)\right) \tag{3.76}$$

In the same way, the interval Hotelling $T^2$ statistic ($IT^2$) is calculated as:

$$IT^2(k) = \left\| \frac{\left[\hat{\mathbf{t}}(k)\right]}{[\Lambda_\ell]^{1/2}} \right\|^2 \tag{3.77}$$

and the new interval squared weighted error ($ISWE$) is computed as:

$$ISWE(k) = \left\| \frac{\left[\tilde{\mathbf{t}}(k)\right]}{[\Lambda_{m-\ell}]^{1/2}} \right\|^2 \tag{3.78}$$

Where $[\hat{t}(k)]$ and $[\tilde{t}(k)]$ are respectively the first $\ell$ and last $m-\ell$ interval-valued components from data projection using interval PCA model, with $[\Lambda_\ell]$ and $[\Lambda_{m-\ell}]$ being the corresponding eigenvalues which are calculated as in the $[T^2]$ and $[SWE]$ case.

Based on the previous simulation example (1), we calculated the $ISPE$ and $ISWE$ indices in faulty case represented in Figures 3.10 and 3.11, respectively. We note an improvement compared to the $[SPE]$ and $[SWE]$ indices in terms of fault detection and false alarms, i.e. enhancement of fault detection and reduction of false alarms. Both indices $[T^2]$ and $IT^2$ fail to detect the fault in

this case (1). However, $IT^2$ outperforms the interval $[T^2]$ index in the case of high amplitude faults usually detected by Hotelling's statistic.
A standard statistical test have to be realized in order to validate these control charts. Thus, a distribution fitting procedure using normality tests have been lead, which can be done by a calculation of the probability density function for a number of populations or based on other tests. Under the assumption that the $n$ samples are independent and the joint distribution of the $m$ variables is the multivariate normal, the new $ISPE$ statistic follows a chi-squared ($\chi^2$) distribution as in the classical case. The same statement has also been verified for the the interval Hotelling $IT^2$ and the $ISWE$. This will furthermore allow to compute statistical control thresholds. Hence, the example of the confidence limit $\delta^2_\alpha$ for the $ISPE$ can be computed from its approximate distribution, based on Box's formula (Box, 1954), as:

$$\delta^2_\alpha = g\chi^2_{h,\alpha} \tag{3.79}$$

Where $g$ is a weighting parameter included to account for the magnitude of the $ISPE$ and $h$ accounts for the degrees of freedom with a significance level of $1-\alpha$, typically selected to be 95% to 99%.

Accordingly, the control limits of the $IT^2$ and the $ISWE$ are calculated similarly, based on Box's formula, for the corresponding degrees of freedom and significance level.

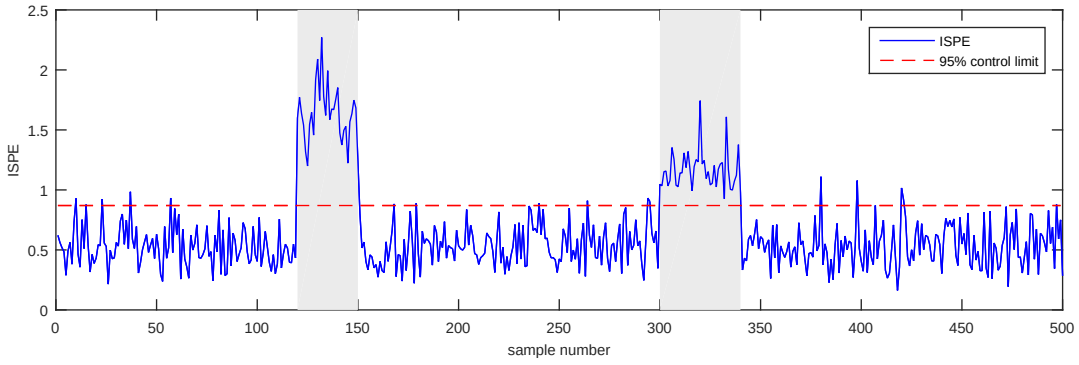



FIGURE 3.10 – Time evolution of ISPE index

To further enhance the performances of these statistics, we investigate the use of filtering to reduce noise's impact on the data. In this framework, the exponentially weighted moving average (EWMA) control scheme can be combined with the concerned index. Thus, If we note by '$S$' the interval statistic ($ISPE$, $IT^2$ or $ISWE$), and '$S_{(f)}$' its EWMA filtered version, the general expression of EWMA filtering applied to the interval index at hand is given by:

$$S_{(f)}(k) = (1-\beta)S_{(f)}(k-1) + \beta S \tag{3.80}$$

where $0 < \beta < 1$ is a forgetting factor.

For the same conditions of the simulation example given in (1), we demonstrate the performances of the filtered $ISPE$ based on $EWMA$ filter denoted $ISPE_{(f)}$.

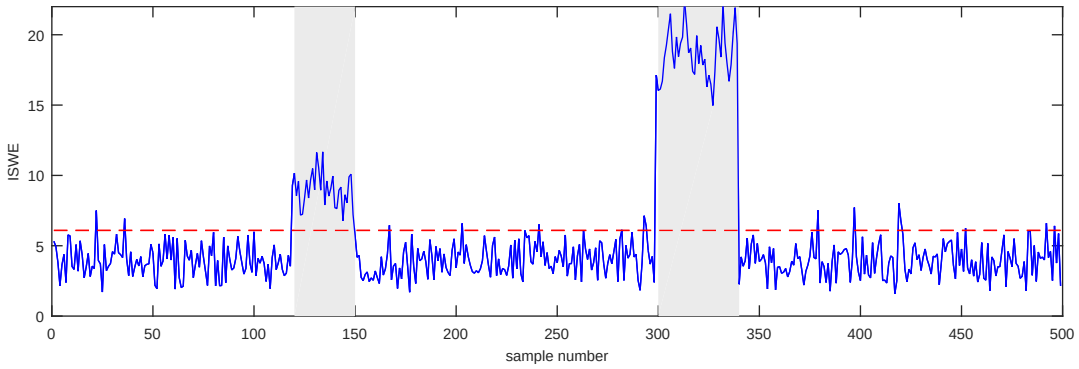



FIGURE 3.11 – Time evolution of ISWE index in fault free and faulty cases

Time evolution of the $ISPE_{(f)}$ and $ISWE_{(f)}$ are depicted in Figures 3.12 and 3.13. Simulation results show that the enhanced interval charts tend to improve the quality of detection and reduce the rate of false alarms but also introducing a small detection delay.

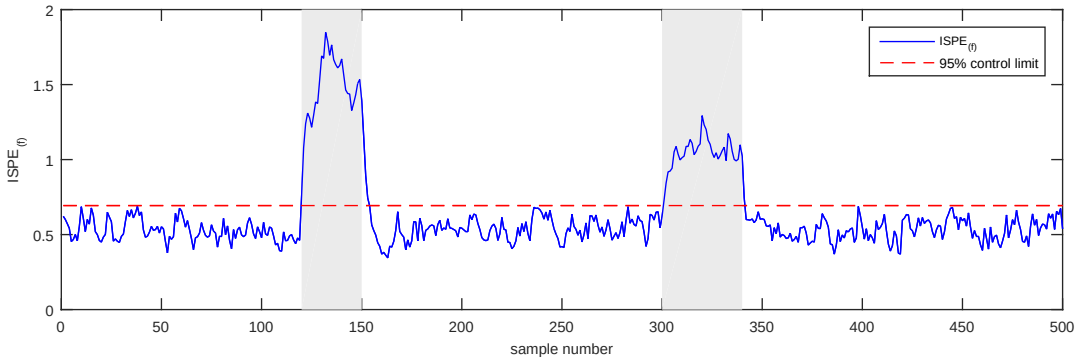



FIGURE 3.12 – Time evolution of the filtered $ISPE_{(f)}$ with EWMA filtering

## 3.5 Fault Isolation Using PCA for Interval-Valued Data

After the presence of fault has been detected, it is important to identify the faulty variable and apply the necessary corrective actions to eliminate the abnormal data. Among the various PCA based isolation techniques, we can find the variable reconstruction approach proposed in (Dunia and Joe Qin, 1998) which is an effective approach that is largely employed for that matter in many related works (Harkat et al., 2006) (Alcala and Qin, 2009).

In conventional PCA, the isolation of faults is achieved via a comparison between the variables values before and after reconstruction. The variable reconstruction approach assumes that each variable may be faulty and suggests to reconstruct the assumed faulty variable using the PCA model from the remaining variables

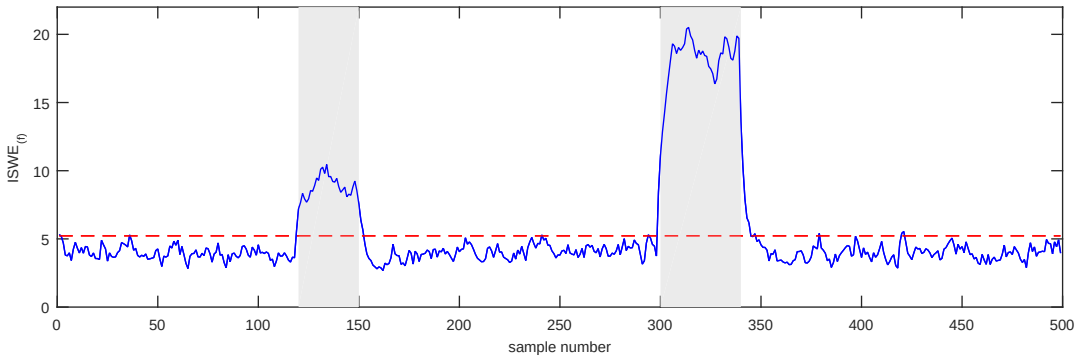



FIGURE 3.13 – Time evolution of the filtered $ISWE_{(f)}$ with EWMA filtering

(Dunia and Joe Qin, 1998). This reconstructed variable is then used to isolate the detected faults. In other words, the reconstruction of the $ith$ variable uses all the other variables data except the $ith$ one. Thus, if only this variable is faulty; its reconstruction eliminates the fault affecting it.

In this section, an extension of this approach for interval-valued PCA model is used for fault isolation and reconstruction of the faulty measurements. The reconstructions for the various interval PCA models are obtained via the projection matrix $G^{(i)}$, which provides reconstructed projections on principal or residual sub-spaces based on the chosen number of components, as defined in equations (3.54) (3.55) (3.57) (3.58).

So, let $ISPE^{(i)}_{(f)}(k)$ be the filtered variation of index $ISPE(k)$ calculated after reconstruction of the $i^{th}$ variable. Therefore, if the reconstructed variable is faulty, the $ISPE^{(i)}_{(f)}(k)$ index is in the control limit because the fault is eliminated by reconstruction. In the other hand, if the reconstructed variable is not faulty, the $ISPE^{(i)}_{(f)}(k)$ index is outside its control limit because it is affected by the fault. In summary, when a fault is detected, and in order to isolate it, all indices $ISPE^{(i)}_{(f)}(k)$, $(i = 1, ..., m)$ are computed, and if $ISPE^{(i)}_{(f)}(k) \leq \delta^2_\alpha$, the $i^{th}$ sensor is considered as the faulty one.

The general expression of the $ISPE^{(i)}$ calculated after the reconstruction of the $i$th variable, is given by:

$$ISPE^{(i)}(k) = \left\| \left[ \mathbf{e}^{(i)}(k) \right] \right\|^2 = \sum_{j=1}^{m} \left\| \left[ e^{(i)}_j(k) \right] \right\|^2 \tag{3.81}$$

Figure 3.14 illustrates the time evolution of the different $ISPE^{(i)}_{(f)}$ $(i = 1, ..., 6)$ indices, calculated after the reconstruction of the different variables. From this figure, we can show that between sample time 120 and 150 all indices $ISPE^{(i)}_{(f)}$ are out of their thresholds except $ISPE^{(1)}_{(f)}$, which indicates that variable $[x_1]$ is the faulty one. The same reasoning can be made on $ISPE^{(4)}_{(f)}$ between sample time 300 and 340 where variable $[x_4]$ is identified as the faulty variable.

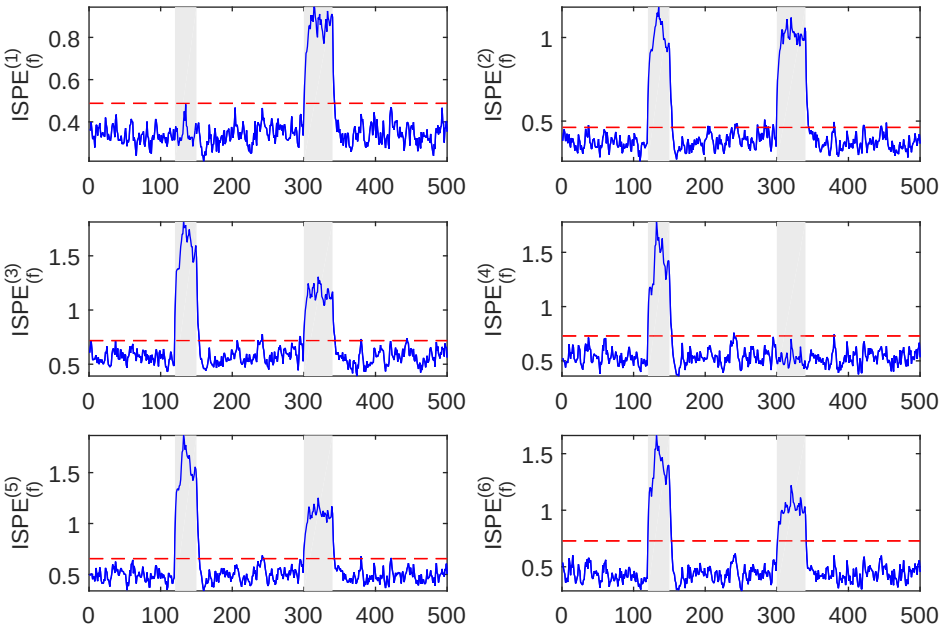



FIGURE 3.14 – Time evolution of $SPE_{(f)}^{(i)}$ calculated after the reconstruction of different variable $i = 1, 2, \ldots, 6$

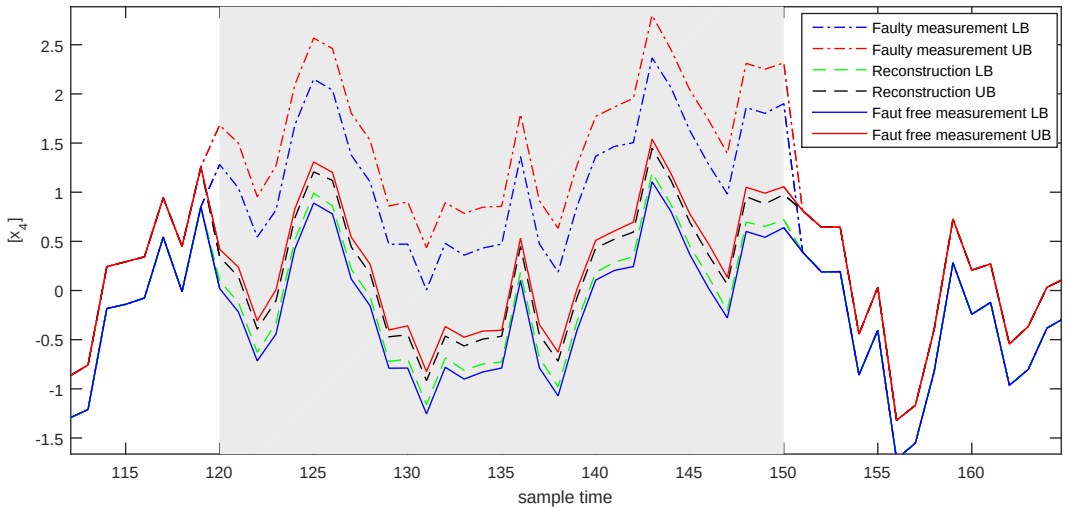



FIGURE 3.15 – Reconstruction of the faulty variable $[x_1]$

Once the faulty sensors are identified, the following task consists in rectifying the aberrant values of sensors by estimating the present state of the process. The aim of the variable reconstruction approach is to best estimate the faulty variable $[x_i]$, using the PCA model and fault direction $\xi_i$. Figures 3.15 and 3.16 represents the fault free variable, the faulty variable and its estimate values given by reconstructions, for the contaminated variables $[x_1]$ and $[x_4]$, respectively. We can clearly see that the reconstructed measurements are good replacement values for the faulty measurements.

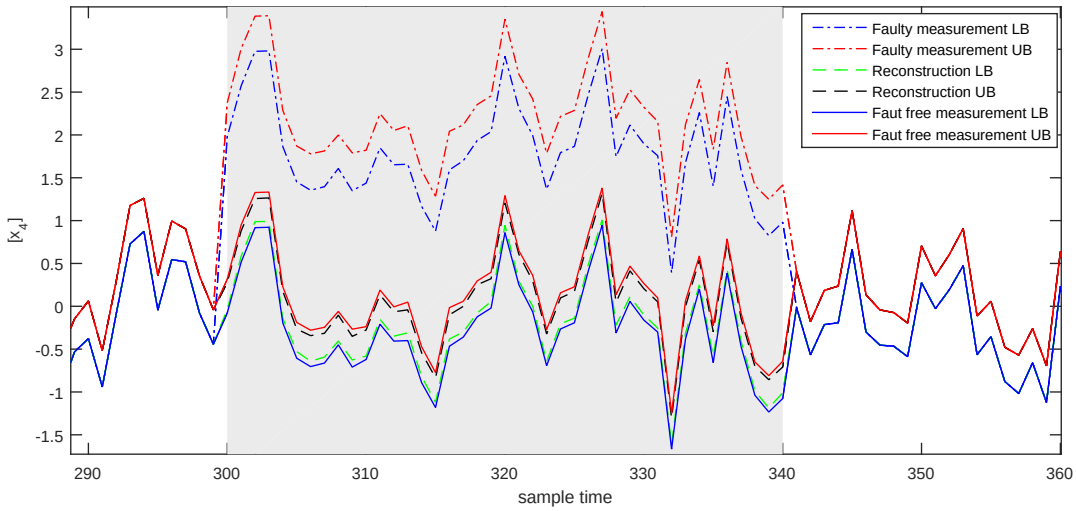



FIGURE 3.16 – Reconstruction of the faulty variable $[x_4]$

## 3.6 Conclusion

Introducing interval notion to diagnosis of systems, and more precisely using principal component analysis, is a novel technique that emphasizes on uncertainties of measurement. The interval nature of the projections ensures the elimination of uncertainties during fault detection and isolation procedure. Undoubtedly, taking benefits of any knowledge about uncertainties is one of the fundamental points of current research and development in multivariate statistical process monitoring. Hence, several interval-valued PCA models have been developed for more efficiency in analyzing the huge-volume of data continuously emerging from various computerized industries

In this section, we presented different Interval-valued PCA methods, that are geometric approximations based on the hyper-rectangles corresponding to the interval-valued data. The most known methods being VPCA, CPCA, MRPCA and CIPCA approaches. Within the framework of diagnosis, that lead us to define the theory of Interval-valued PCA based FDI, which includes defining fault indicators of interval type, and extension of isolation methods to the interval case. Throughout this chapter, we have introduced new interval-valued statistics for use with interval-valued PCA along with their threshold calculation. For the isolation of faults, we introduced an extension of the reconstruction principle for interval-valued data, which was also used to define a new criterion for the determination of the number of components to be retained in and interval-valued PCA model. the new criterion is the variance or interval reconstruction error (VIRE) which ensure the minimum reconstruction error of the interval-valued PCA model. The different indices are illustrated using a simulation example and proven efficient in detecting and isolating faults in uncertain processes. However, it is necessary to view and compare the performances of such interval-valued models and statistics in real processes, which will be discussed in the next chapter.

# 4 | Comparative Studies and Applications



## 4.1 Comparative Study and Validation on Synthetic Data-sets

In this section, we demonstrate the effectiveness of the proposed control charts and interval diagnosis techniques using simulated data. The performance evaluation includes comparison in the case of univariate and multivariate detection method using Monte-Carlo simulation. This simulation, also known as the repeated random sub-sampling validation, randomly splits the dataset into training and validation data. For each such split, the model is fit to the training data, and predictive accuracy is assessed using the validation data for a number of iterations (folds). This method also exhibits Monte-Carlo variation, meaning that the results will vary if the analysis is repeated with different random splits.

### 4.1.1 Monte-Carlo Simulation

Let us first describe how the data for the Monte-Carlo simulation are constructed. Each data set in this experience is an $n \times m$ interval valued matrix $[X]$ with $m = 6$ variables and $n = 500$ samples, which are generated based on example (1). In this simulation example, radius $\delta X$ is chosen to be strictly positive and having

different configurations varying from 10% to 20% of the variables variations. i.e. $x_i \times 10\% < \delta x_i < x_i \times 20\%$, $i = 1, \ldots, 6$

To carry on with diagnosis routine, a second set of data, i.e. validation set, is generated afterwards. In order to test the robustness of the IVD-PCA methods in terms of modeling and detection abilities, we deliberately choose to simulate two types of offsets, real faults $\delta x_i < \zeta_i < -\delta x_i$, and uncertainties $-\delta x_i < \zeta_i < \delta x_i$, $(i = 1, ..., 6)$. In the framework of IVD-PCA based monitoring, the radius $\delta X$ of data is considered as a safe zone, were small magnitude offsets, i.e. uncertainties $-\delta x_i < \zeta_i < \delta x_i$ are not detected because considered as normal process variation, while high magnitude offsets $\delta x_i < \zeta_i < -\delta x_i$ are considered as faults. Thus, for the $m$ variables of the process we then have $\zeta_1, \ldots, \zeta_m$ added offsets, which are randomly generated as uncertainties or real faults. Finally, a Monte-Carlo simulation will be performed for $k = 5000$ runs or iterations.

The evaluated performances are calculated in order to compare between several IVD-PCA approaches and are given by the following:

— The good detection rate (GDR) is given by the number of faulty data points (violated samples) that exceed the control threshold (good detection) over the total number of faulty data points.

$$GDR = \left( \frac{\text{violated samples}}{\text{faulty data}} \right) \% \tag{4.1}$$

— The uncertainties detection rate (UDR) is, roughly speaking, a GDR calculated for the moments where uncertainties $-\delta x_i < \zeta_i < \delta x_i$ are injected. From a robustness point of view, the IVD-PCA method with the best performances tends to neglect $-\delta x_i < \zeta_i < \delta x_i$, hence having a very low UDR.

### 4.1.2 Univariate Case

In the univariate detection case, the interval-valued residuals are used to assess the existence or not of faults. This method exploits the interval-valued nature of residuals, such that, a fault is detected if one of the bounds of the residual changes sign. Or, mathematically, if the product of bounds is negative then there is no fault $(\underline{e}(k) \times \overline{e}(k) < 0)$, if the product of bounds is positive then there is a fault $(\underline{e}(k) \times \overline{e}(k) > 0)$. In other words, both bounds of residuals are by nature of opposite signs, i.e. $\underline{e}(k) < 0$, $\overline{e}(k) > 0$, and their product is then always negative. So, given that in the case of fault presence only one bound changes sign (depending on positive or negative nature of fault). Thus, when this occurs, the product of bounds will be that of two positives or two negatives, resulting in a positive quantity.

Therefore, based on the data generated using a Monte-Carlo simulation, we computed the two performances criteria (GDR and UDR) for the four (4) presented IVD-PCA methods, namely, VPCA, CPCA, MRPCA and CIPCA. The results are represented by the histogram graph in Figure 4.1

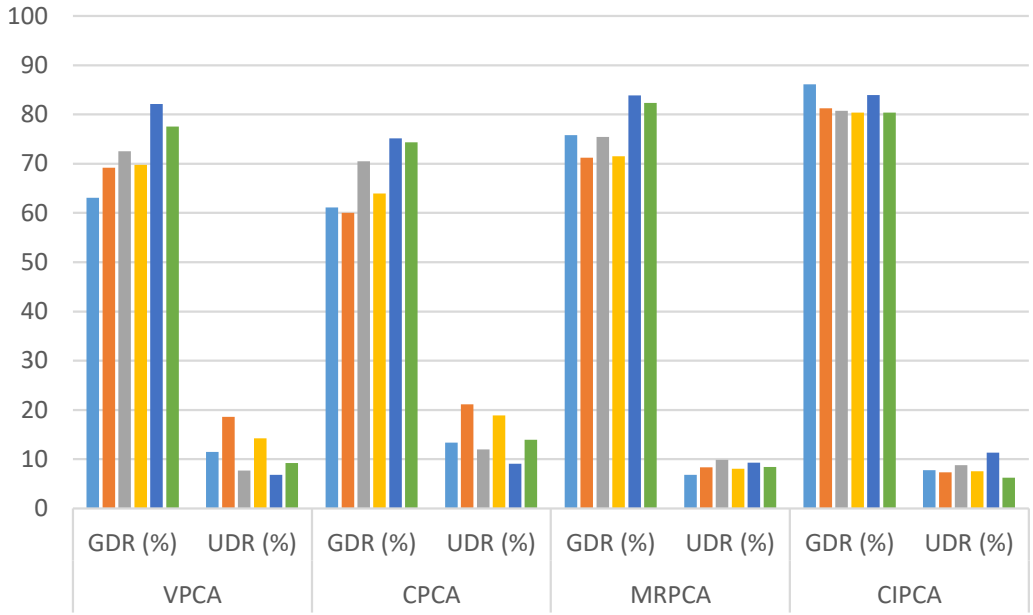



FIGURE 4.1 – Univariate chart comparison for interval-valued PCA model

We note from Figure 4.1, that all the models give fairly good results for the detection of faults, with the CIPCA model being the one with the highest GDR rates, followed by the MRPCA model, VPCA model and CPCA model. When it comes to handling uncertainties, the MRPCA model yields the best results (lowest UDR ratios) followed by the CIPCA model, the VPCA model and the CPCA model. Most balanced detection results are then obtained via CIPCA or MRPCA models, with good performances for the VPCA model, and rather poor performances of the CPCA model. However, the univariate detection method is computationally cumbersome, as it depends on singly analysing variables (residuals) and is therefore limited to processes of small number of variables. In the case of large scale processes (with huge amount of variables) one might best consider the multivariate detection approach.

### 4.1.3 Multivariate Case

#### 4.1.3.a Comparison between IVD PCA's for diagnosis

A first comparison between the four presented IVD-PCA methods: VPCA, CPCA, MRPCA and CIPCA is performed in terms of good detection and handling of uncertainties, based on the different variants of the squared prediction error (SPE) for interval-valued data: $[SPE]$, $ISPE$ and $ISPE_{(f)}$ for a significance level of 95%. Results are represented by the histogram graphs in Figures 4.2, 4.3, 4.4 and 4.5, where the different colored bars account for the percentage of the calculated performance for injected offsets in variables 1 to 6 (from left to right).

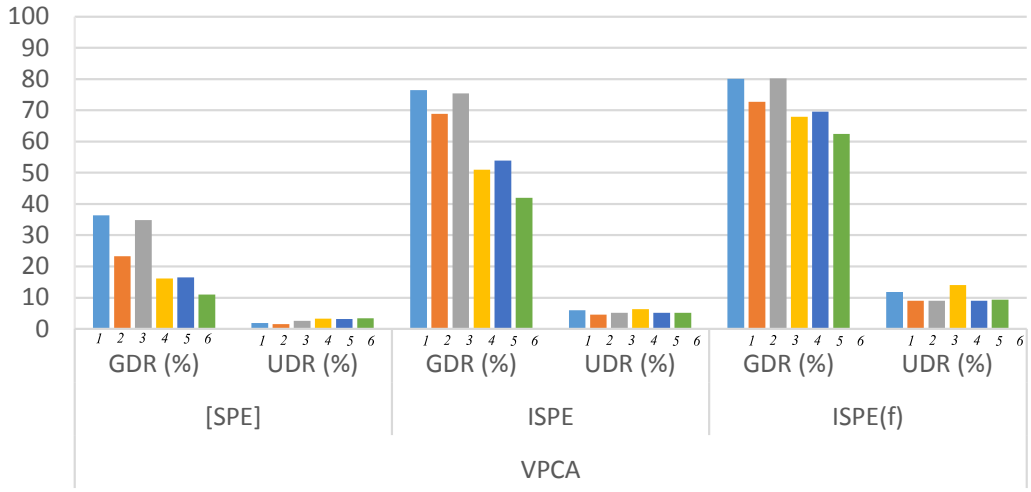



FIGURE 4.2 – Interval SPE variations comparison using VPCA model

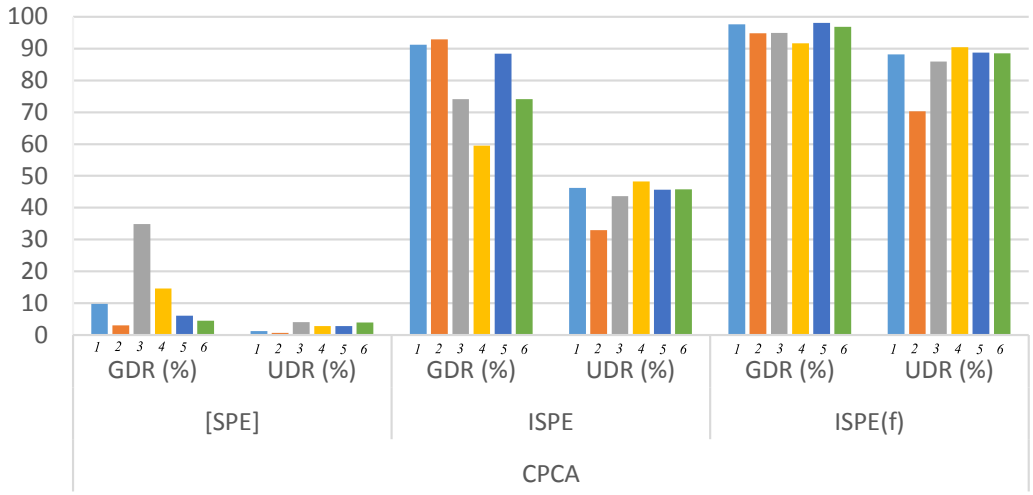



FIGURE 4.3 – Interval SPE variations comparison using CPCA model

In general, all the four methods yield fairly well results for detecting real faults, especially based on the $ISPE$ and its filtered version $ISPE_{(f)}$. The actual difference showed between the IVD-PCA methods lies mainly in the handling of measurement uncertainties, which is described on the graphs by the UDR ratios. The best performances are achieved using CIPCA model which tends to neglect uncertainties in approximately 95% of the cases (corresponding to a UDR of 5%). This is a rather satisfying result considering that thresholds are calculated for a significance level of 95%. Other well behaving methods are VPCA and MRPCA with respectively 92% and 80% of neglected uncertainties. In the other hand, CPCA method almost detects all the uncertainties as if they are actually real faults, making it a poor model for diagnosis purpose. This comparison is conducted based on the $ISPE$ statistic which performs best when combined with the EWMA filter. However, a global comparison has been held for all the

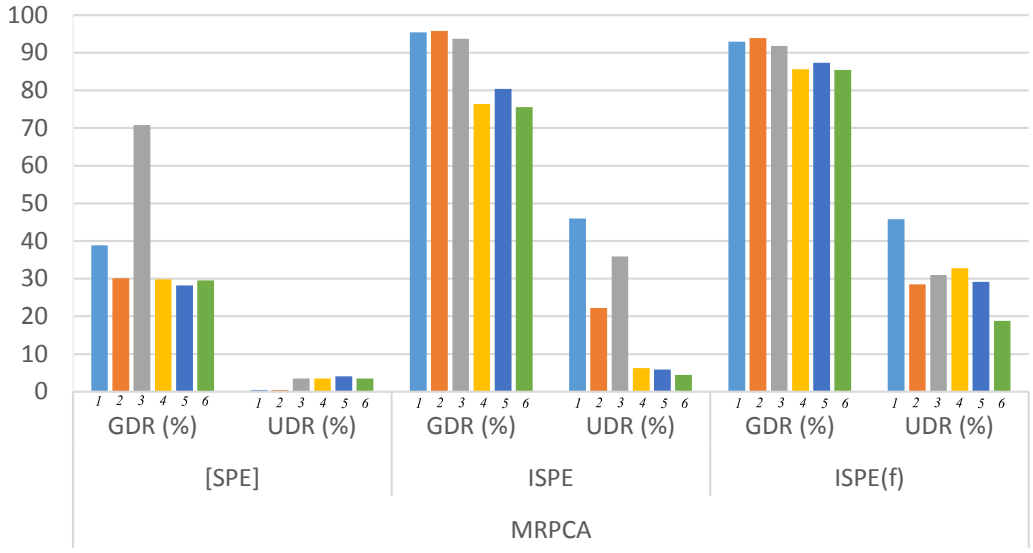



FIGURE 4.4 – Interval SPE variations comparison using MRPCA model

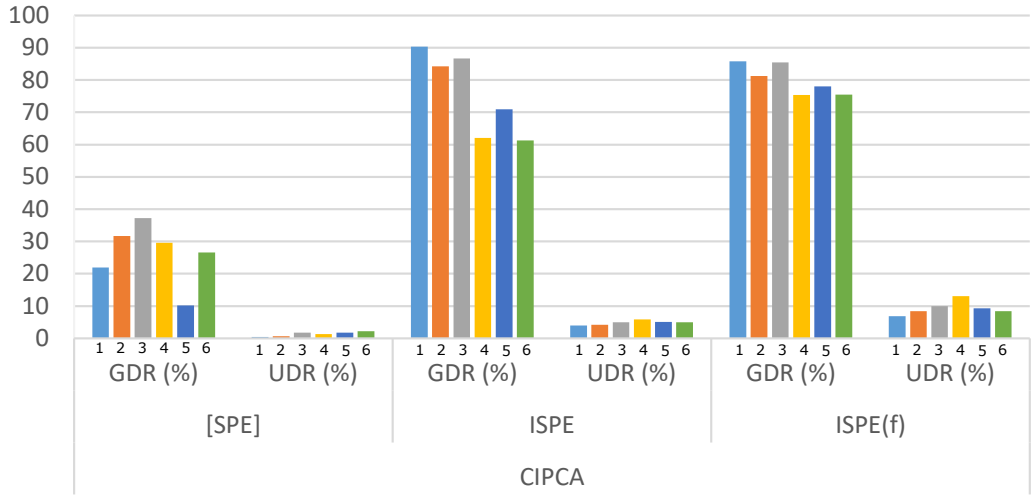



FIGURE 4.5 – Interval SPE variations comparison using CIPCA model

statistics and IVD-PCA's with the same overall results.

#### 4.1.3.b Comparison between interval statistics

Based on the engaged simulations on the various IVD-PCA's presented hereby, the performance comparison results declare the CIPCA model as the most accurate method for diagnosis, with the highest fault detection and uncertainties rejection ratios. Therefore, we choose to demonstrate the performances of the interval statistics only for CIPCA method.

The comparative results for the interval SPE variations are shown in Table 4.1. As can be seen, the $ISPE$ variations have satisfying results based on GDR and UDR ratios, compared to the relatively poor standard $[SPE]$ indicator. We note however that the EWMA filter enhances the detection results of the statistic but reduces its robustness by slightly increasing the UDR ratios.

| | $[SPE]$ | | $ISPE$ | | $ISPE_{(f)}$ | |
|---|---|---|---|---|---|---|
| Variables | GDR(%) | UDR(%) | GDR(%) | UDR(%) | GDR(%) | UDR(%) |
| $[x_1]$ | 21.92 | 0.47 | 90.36 | 4 | 85.82 | 6.86 |
| $[x_2]$ | 31.68 | 0.61 | 84.23 | 4.14 | 81.24 | 8.36 |
| $[x_3]$ | 37.25 | 1.72 | 86.71 | 4.97 | 85.44 | 9.91 |
| $[x_4]$ | 29.55 | 1.26 | 62.02 | 5.86 | 75.33 | 13.07 |
| $[x_5]$ | 10.16 | 1.8 | 70.92 | 5.1 | 78 | 9.3 |
| $[x_6]$ | 26.62 | 2.2 | 61.32 | 4.97 | 75.44 | 8.45 |

TABLE 4.1 – Diagnosis performances using $[SPE]$, $ISPE$ and $ISPE_{(f)}$ indices

For the interval $T^2$ case, presented by Table 4.2, we notice the superiority of the EWMA filter based $IT^2_{(f)}$ over the other statistics. However, the overall performance of all the interval $T^2$ statistics is not satisfying given that faults are only detectable on the first two variables. This is due to the nature of the $T^2$ indicator which basically calculates distance in the principal space from the origin and can only detect high magnitude faults.

| | $[T^2]$ | | $IT^2$ | | $IT^2_{(f)}$ | |
|---|---|---|---|---|---|---|
| Variables | GDR(%) | UDR(%) | GDR(%) | UDR(%) | GDR(%) | UDR(%) |
| $[x_1]$ | 1.1 | 0 | 23.43 | 7.9 | 54.8 | 15.53 |
| $[x_2]$ | 2.81 | 0 | 61.45 | 2.53 | 77.19 | 0.33 |
| $[x_3]$ | 5 | 0 | 6.18 | 3.94 | 11.52 | 4.76 |
| $[x_4]$ | 1 | 0 | 5.68 | 4.33 | 9.17 | 5.73 |
| $[x_5]$ | 0 | 0 | 7.51 | 4.09 | 18.65 | 5.35 |
| $[x_6]$ | 0 | 0 | 5.82 | 4.35 | 10.18 | 6.04 |

TABLE 4.2 – Diagnosis performances using $[T^2]$, $IT^2$ and $IT^2_{(f)}$ indices

Next, we compare the interval squared weighted error ($SWE$) statistics presented in this manuscript based on interval CIPCA model, the results are gathered in Table 4.3. A slight improvement is perceived in detection ratios compared to the interval $SPE$ charts, especially for the $ISWE_{(f)}$. Nevertheless, we note a considerable increase of detected uncertainties explained by the UDR ratios. By definition, the $SWE$ chart takes into account the residual variances for more sensibility, which explains the increase in the detection of uncertainties.

| | $[SWE]$ | | $ISWE$ | | $ISWE_{(f)}$ | |
|---|---|---|---|---|---|---|
| Variables | GDR(%) | UDR(%) | GDR(%) | UDR(%) | GDR(%) | UDR(%) |
| $[x_1]$ | 15.9 | 1.59 | 79.5 | 4.51 | 81.22 | 8.19 |
| $[x_2]$ | 1.82 | 0.22 | 67.22 | 5.21 | 75.47 | 12.74 |
| $[x_3]$ | 32.7 | 1.11 | 88.77 | 6.67 | 86.6 | 23.6 |
| $[x_4]$ | 7.25 | 1.49 | 88.45 | 5.91 | 87.48 | 16.32 |
| $[x_5]$ | 7.58 | 1.23 | 79.02 | 6.4 | 82.22 | 20.34 |
| $[x_6]$ | 21.35 | 2.97 | 88.21 | 5.27 | 87.54 | 12.06 |

TABLE 4.3 – Diagnosis performances using $[SWE]$, $ISWE$ and $ISWE_{(f)}$ indices

Summing up, the interval norm based control charts are a real improvement over the classical interval-valued charts, especially when combined with a robust filtering method, which is done in our case using the EWMA filter. Most balanced results are obtained for the $ISPE_{(f)}$ indicator which demonstrates precision in distinguishing uncertainties from actual faults. The $ISWE_{(f)}$ indicator has the highest detection rate but is oversensitive to uncertainties, and the $IT^2_{(f)}$ yield rather unsatisfying results. However, it performs at its best in some particular situations as in the conventional PCA case.

Another important matter to discuss in diagnosis is the isolation of faults. In this work, we emphasis on the isolation based on the reconstruction principle, extended for interval valued data. Table 4.4 presents the isolation results for $ISPE_{(f)}$, $ISWE_{(f)}$ and $IT^2_{(f)}$, that are considered as the most effective of the statistics studied hereby. The calculated performance is the good isolation ratio (GIR) for each sensor/variable, which indicates the ability of the approach to isolate the detected faults. Good results are obtained in general for all statistics, thus, demonstrating the effectiveness of the used interval-valued data reconstruction to isolate the detected faults based on IVD-PCA model.

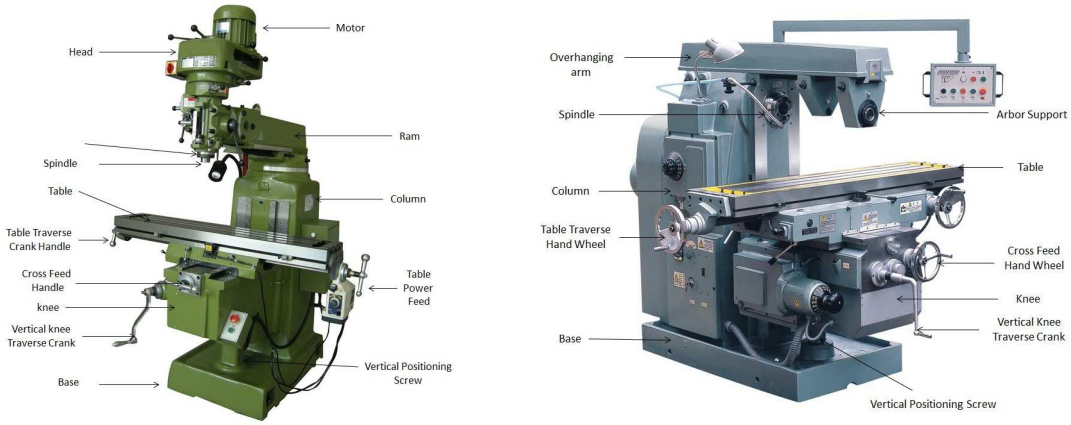
| | GIR (%) | | |
|---|---|---|---|
| Variables | $ISPE_{(f)}$ | $IT^2_{(f)}$ | $ISWE_{(f)}$ |
| $[x_1]$ | 89.09 | 92.3 | 89.11 |
| $[x_2]$ | 82.88 | 83.23 | 82.99 |
| $[x_3]$ | 88.86 | 71.1 | 88.62 |
| $[x_4]$ | 74.98 | 83.3 | 71.34 |
| $[x_5]$ | 77.52 | 89.42 | 76.19 |
| $[x_6]$ | 74.08 | 76.23 | 70.65 |

TABLE 4.4 – Isolation performances using $ISPE_{(f)}$, $IST^2_{(f)}$ and $ISWE_{(f)}$ indices

## 4.2 Application on Milling Machine Data

### 4.2.1 Description of the process

Milling is the most common form of machining, a material removal process, which can create a variety of features on a part by cutting away the unwanted material. The milling process requires a milling machine, work-piece, fixture, and cutter. Milling is typically used to produce parts that are not axially symmetric and have many features, such as holes, slots, pockets, and even three dimensional surface contours. Parts that are fabricated completely through milling often include components that are used in limited quantities, perhaps for prototypes, such as custom designed fasteners or brackets. Another application of milling is the fabrication of tooling for other processes. For example, three-dimensional molds are typically milled. Milling is also commonly used as a secondary process to add or refine features on parts that were manufactured using a different process. Due to the high tolerances and surface finishes that milling can offer, it is ideal for adding precision features to a part whose basic shape has already been formed.




(A) Vertical Milling Machine

(B) Horizontal Milling Machine

FIGURE 4.6 – Different Parts of Milling Machines

In milling, the speed and motion of the cutting tool is specified through several parameters. These parameters are selected for each operation based upon the work-piece material, tool material, tool size, and more.

— Cutting feed: The distance that the cutting tool or work-piece advances during one revolution of the spindle and tool, measured in inches per revolution (IPR).

— Cutting speed: The speed of the work-piece surface relative to the edge of the cutting tool during a cut, measured in surface feet per minute (SFM).

— Spindle speed: The rotational speed of the spindle and tool in revolutions per minute (RPM).

— Feed rate: The speed of the cutting tool's movement relative to the work-piece as the tool makes a cut. The feed rate is measured in inches per minute (IPM) and is the product of the cutting feed (IPR) and the spindle speed (RPM).

— Axial depth of cut: The depth of the tool along its axis in the work-piece as it makes a cut.

— Radial depth of cut: The depth of the tool along its radius in the work-piece as it makes a cut.

In this section, we present and application of PCA and IVD-PCA based sensor fault detection and isolation on the milling machine data-set from the NASA Prognostics Center of Excellence (Agogino and Goebel, 2007).

### 4.2.2 The Data

The data-set used in this application represents experiments from runs on a milling machine under various operating conditions. Data were sampled by three different types of sensors: acoustic emission sensor, vibration sensor and current sensor, and were recorded at several positions. The experimental conditions of data used in this application are as in the first case, explained in (Agogino and Goebel, 2007), with a depth of cut of $1.5mm$ and a feed of $413mm/min$ for cast iron material, the cutting speed was set to $200m/min$. The acquired sensors/variables are explained in Table.4.5. An excellent reference on machining operations and different possible strategies for process monitoring is given in (Teti et al., 2010).

| | |
|---|---|
| $AC_{smc}$ | AC spindle motor current |
| $DC_{smc}$ | DC spindle motor current |
| $Vib_T$ | Table vibration |
| $Vib_S$ | Spindle vibration |
| $AE_T$ | Acoustic emission at table |
| $AE_S$ | Acoustic emission at spindle |

TABLE 4.5 – Milling machine variables

A correlation analysis, between the different process sensors reveals that the first variable, which corresponds to the AC spindle motor current ($AC_{smc}$), has a very weak correlation with the other variables. This can clearly be identified in Eq. 4.2 which represents the correlation coefficients between different sensors, calculated as in 1.20, 1.21. The off-diagonal coefficients of the first line/column (highlighted elements) are close to zero, explaining that there is almost no relationship between the $AC_{smc}$ and the other sensors. Hence, the $AC_{smc}$ is excluded from the PCA model, and the monitored variables are reduced to the

5 remaining variables, explained in Table 4.6.

$$\begin{bmatrix} 1.0000 & -0.0022 & 0.0006 & 0.0069 & -0.0849 & -0.0090 \\ -0.0022 & 1.0000 & 0.8220 & 0.8162 & 0.6133 & 0.5775 \\ 0.0006 & 0.8220 & 1.0000 & 0.8664 & 0.7350 & 0.7703 \\ 0.0069 & 0.8162 & 0.8664 & 1.0000 & 0.7907 & 0.7418 \\ -0.0849 & 0.6133 & 0.7350 & 0.7907 & 1.0000 & 0.8620 \\ -0.0090 & 0.5775 & 0.7703 & 0.7418 & 0.8620 & 1.0000 \end{bmatrix} \tag{4.2}$$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|
| $DC_{smc}$ | $Vib_T$ | $Vib_S$ | $AE_T$ | $AE_S$ |

TABLE 4.6 – Monitored milling machine variables

The data are first normalized to zero mean and unit variance, as depicted in Figure 4.7, the normalization rescales the variables in order to have the same range of values for each of the variables. The normalized data variables are presented in Figure 4.8. The number of PC's to be retained for the PCA model is chosen based on the VRE criterion, which corresponds to the value of $\ell = 2$ for a minimum of VRE as in Figure 4.9.

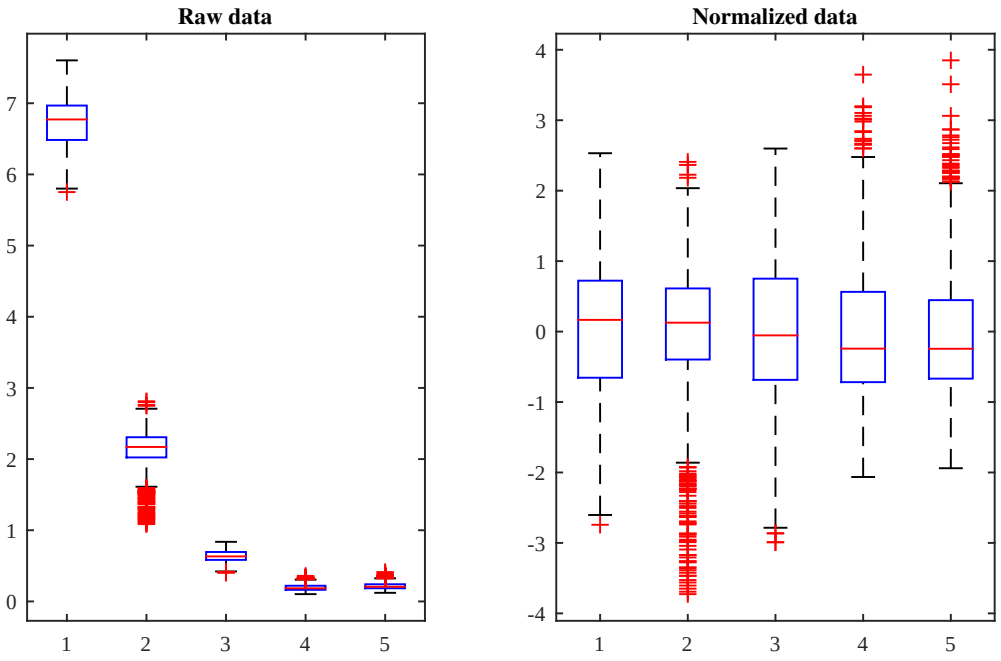



FIGURE 4.7 – Tendency of variables before/after normalization for the milling machine data-set

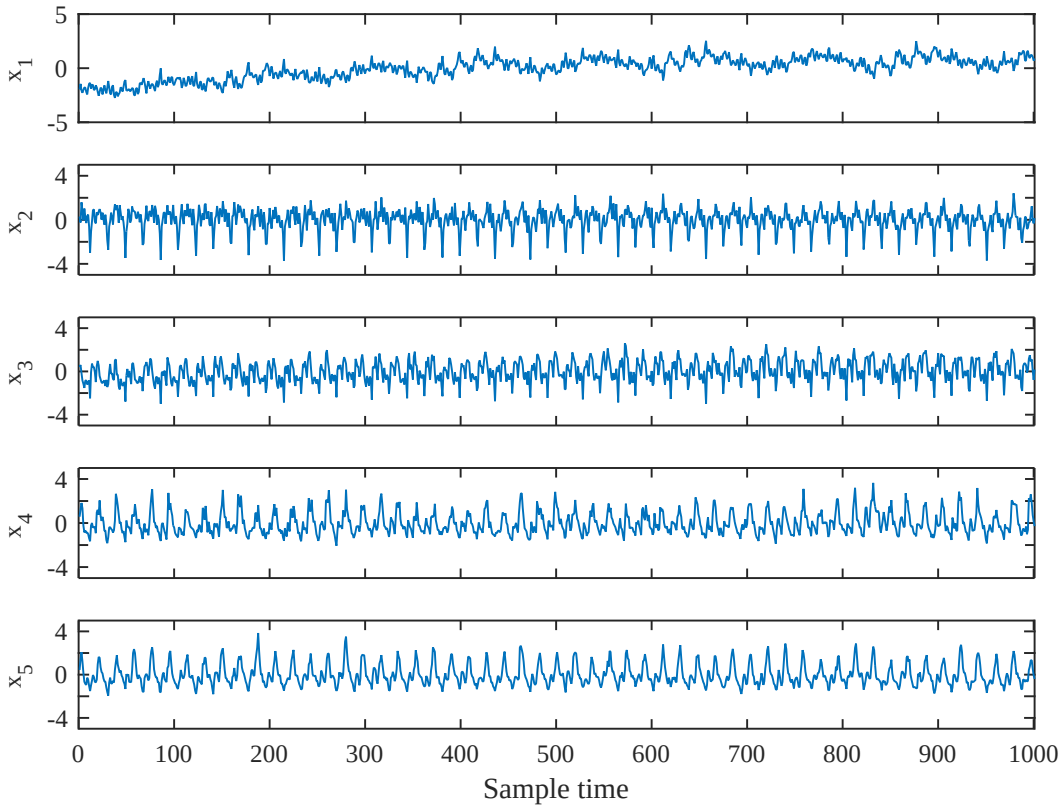



FIGURE 4.8 – Normalized variables of milling machine data-set

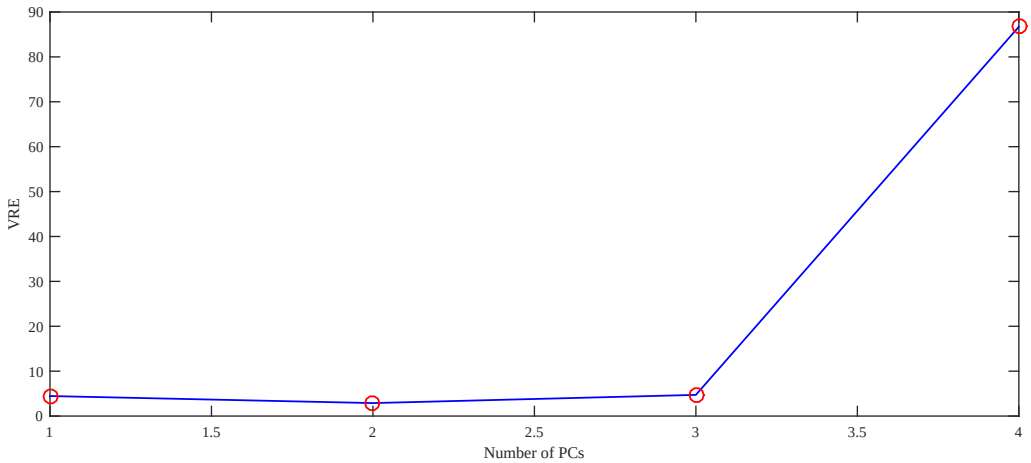



FIGURE 4.9 – Number of PCs According to the VRE criterion for the milling machine data-set

To further test the robustness of an interval-valued PCA model for FDI, and to show the improvements perceived using PCA for interval-valued data, compared to classical PCA. We perform an empirical study using the different fault detection statistics presented in this manuscript based on the well-established CIPCA model, as it is the most precise. This empirical study will also be carried

out in the framework of Monte-Carlo simulation experiment. Using the milling machine dataset, we repeat the experiment for $K = 5000$ times to avoid biased results. In the $k$th replication, the following steps will be conducted:

1. Randomly choose the uncertainty range $\delta x_j$, also called radius $x_j^r$, for each variable/sensor. $\delta x_j$ is randomly selected as 5% to 20% of the corresponding variable $x_j$ from the mill dataset at each iteration $k$, allowing for different interval configurations to test. As an example, for a 5% range, $\delta x_j = x_j \times 0.05$
2. Construct the interval valued variables $[x_j] = [(x_j - \delta x_j), (x_j + \delta x_j)]$.
3. Construct a linear PCA model based on the experimental single-valued milling machine data set, and a PCA for interval-valued data (CIPCA model) based on the interval-valued milling machine data from step 2. 500 data samples are used as training data, where the principal components to be retained for each model are determined according to the $VRE$ and $VIRE$ criteria.
4. The validation samples from the mill data set are selected as $n = 1000$. For each variable $x_j$, uncertainties $\zeta x_j$ are added between sample time 300 and 400, and faults $df_{x_j}$ are added from sample time 800 to the end. Both offsets have random magnitudes at each iteration.
5. Calculate the corresponding fault detection statistics: $SPE$ for classical PCA, as well as interval $[SPE]$, $ISPE$ and its $EWMA$ filtered version for the interval CIPCA model.
6. Determine the performances of both models, and their corresponding statistics.

The evaluated performances of PCA for interval-valued data and classical PCA, based on the different detection statistics presented in this manuscript, are the Good Detection Ratio (GDR%), which represents the ability of the approach to detect faults, and the Uncertainties Detection Ratio (UDR%), which represents the sensitivity of the approach to uncertainties, as presented in the previous section. For a random case in the conducted Monte-Carlo simulation, we give a graphical representation of the behavior of the different fault detection statistics: classical $SPE$, interval $[SPE]$, $ISPE$ and their $EWMA$ filtered versions, illustrated in Figures 4.10, 4.11, 4.12 and 4.13, respectively. Performances results using Monte-Carlo simulation are presented in Table. 4.7

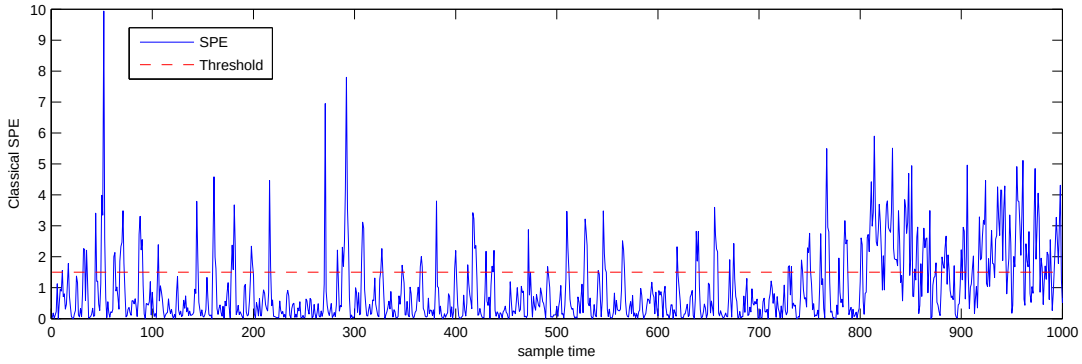



FIGURE 4.10 – Milling machine fault detection using classical $PCA$ and $SPE$ indicator

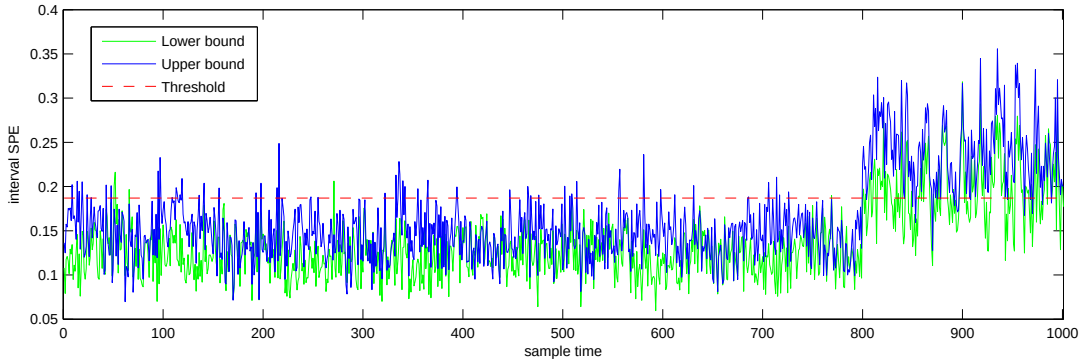



FIGURE 4.11 – Milling machine fault detection using interval $[SPE]$ indicator

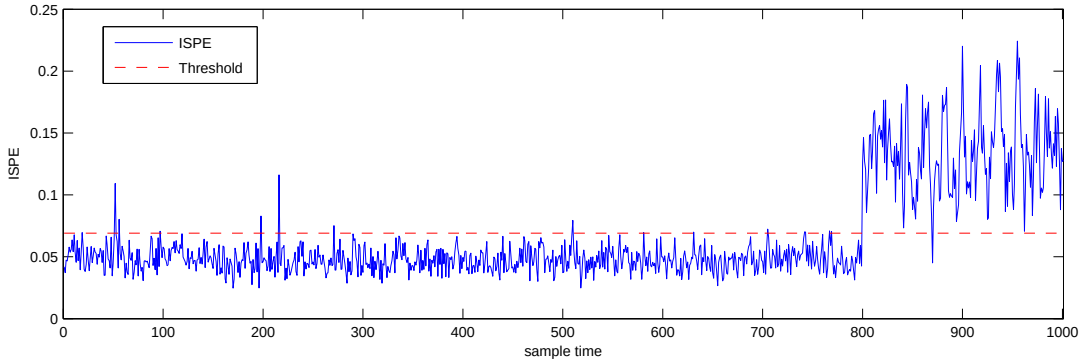



FIGURE 4.12 – Milling machine fault detection using $ISPE$ indicator

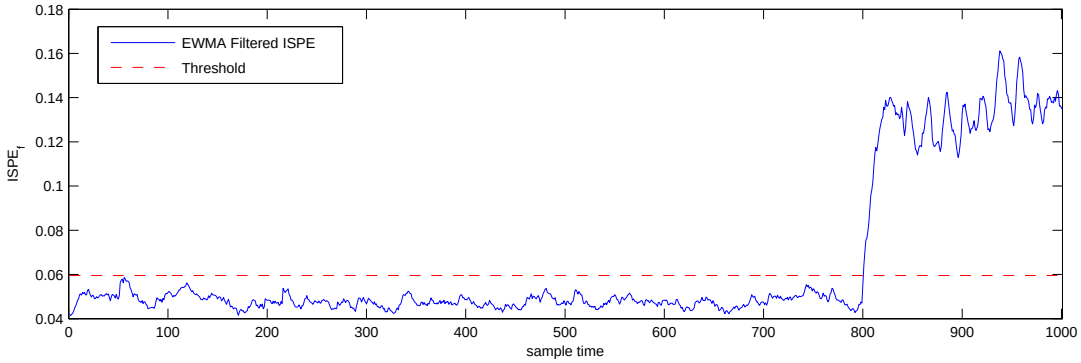



FIGURE 4.13 – Milling machine fault detection using EWMA filtered $ISPE$ indicator

| | $SPE$ | | $[SPE]$ | | $ISPE$ | | $ISPE_f$ | |
|---|---|---|---|---|---|---|---|---|
| | GDR(%) | UDR(%) | GDR(%) | UDR(%) | GDR(%) | UDR(%) | GDR(%) | UDR(%) |
| $DC_{smc}$ | 9.89 | 8.27 | 99.83 | 3.52 | 99.87 | 2.87 | 99.82 | 0.32 |
| $Vib_T$ | 50.70 | 8.97 | 74.68 | 5.31 | 75.55 | 5.61 | 91.63 | 6.78 |
| $Vib_S$ | 44.22 | 13.91 | 87.46 | 5.98 | 89.46 | 6.16 | 99.14 | 6.87 |
| $AE_T$ | 14.58 | 9.03 | 12.88 | 6.57 | 32.21 | 14.87 | 81.10 | 18.27 |
| $AE_S$ | 41.94 | 13.55 | 66.05 | 6.45 | 65.27 | 6.75 | 93.13 | 6.61 |

TABLE 4.7 – Diagnosis performances based on Monte-Carlo simulation using $SPE$, $[SPE]$, $ISPE$ and $ISPE_f$ for the mill data set

The highest fault detection results, given by GDR's presented in Table. 4.7, are mainly obtained using PCA for interval-valued data based on the $ISPE_f$ index. Whilst the lowest UDR's are mostly for the interval-valued case using $[SPE]$ index. However, we need to chose the most balanced results between GDR's and UDR's for these fault detection statistics for a better monitoring application. From the simulation results in Table. 4.7, we note the following:

— Classical PCA based $SPE$ has a rather poor performance in detecting faults and neglecting uncertainties. This is due to the high influence of uncertainties on the constructed classical PCA model. Thus, the $SPE$ index exhibits a huge amount of missed detection and false alarms as presented in Figure 4.10.

— Interval $[SPE]$ handles the uncertainties of measurement very well, but demonstrates a certain lack in detecting faults, mainly due to the need of both bounds to determine the presence of faults, as can be seen in Figure 4.11.

— $ISPE$ is an improvement over the Interval $[SPE]$ in fault detection results, with slight increase in UDR's. As illustrated in Figure 4.12, it is of single valued nature, and has very few false alarms .

— $ISPE_f$ has most satisfying detection results compared to the other statistics, and well balanced UDR's. According to Figure 4.13, false alarms

are mostly eliminated using this index, which is due to the positive effect of the application of $EWMA$ filtering. A negative effect of this filter is introducing a slight time delay in fault detection which doesn't degrade the overall quality of the statistic.

In summary, the overall results of the Monte-Carlo simulation confirm the synthetic data simulation results. This simulation also demonstrates the very good performances of the new fault detection strategy based on PCA for interval-valued data, for handling large scale complex and uncertain systems, compared to a classical PCA based FDI scheme. Among the different indices used in the new interval fault detection scheme, and according to the calculated GDR's and UDR's, the $ISPE_f$ index shows the more balanced performances. However, $[SPE]$ and $ISPE$ indices can also be good choices, and can both have good performances in fault detection using PCA for interval-valued data.

The fault isolation is performed based on the interval-valued reconstruction approach. For the case of a fault on the third variable $Vib_s$, the reconstructions based on the $ISPE_f$ indicator are given in Figure 4.14. We can clearly identify the fault as being in variable 3 due to its absence from the corresponding reconstructed indicator $ISPE^3_{(f)}$.

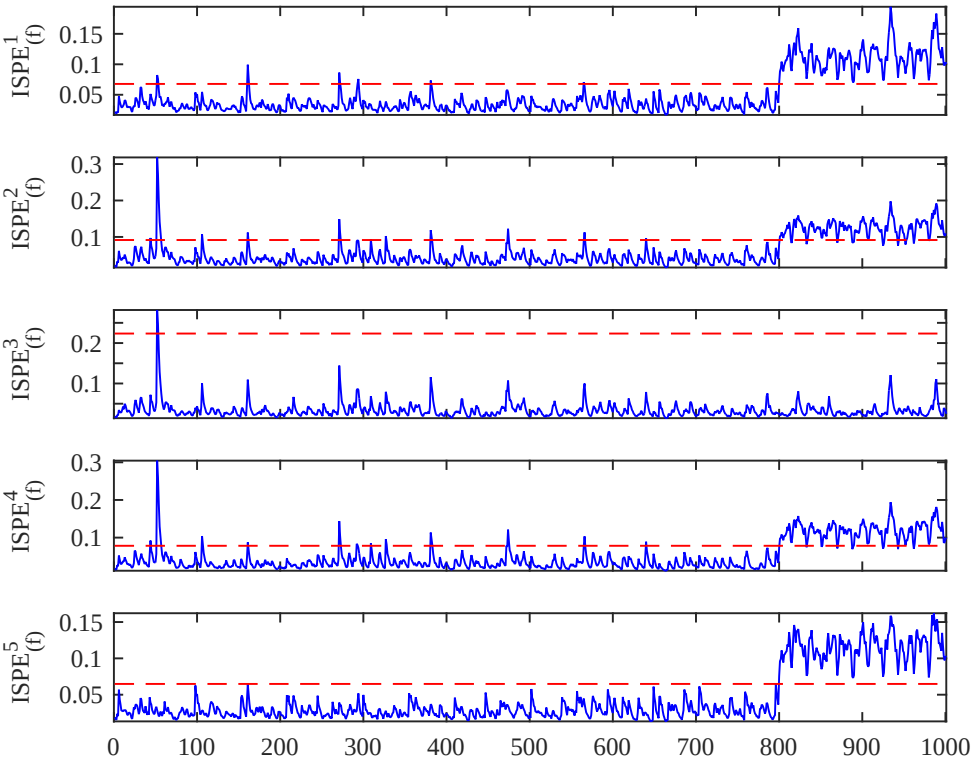



FIGURE 4.14 – Milling machine fault isolation using reconstrcutions principle and the EWMA filtered $ISPE$ indicator

## 4.3 Application to the Distilation Collumn Benchmark

The FDI strategy presented in this thesis has been tested in a simulated distillation column process. The plant is a linearized dynamic model of a continuous distillation column, the process model and simulation conditions are similar to the ones provided by (Skogestad, 1997). The so-called "column A" with LV-configuration have 41 theoretical stages and separates a binary mixture with relative volatility of 1.5 into products of 99% purity. Figure 4.15 represents the diagram of a simple distillation column, where the corresponding manipulated variables are summarized in Table 4.8.

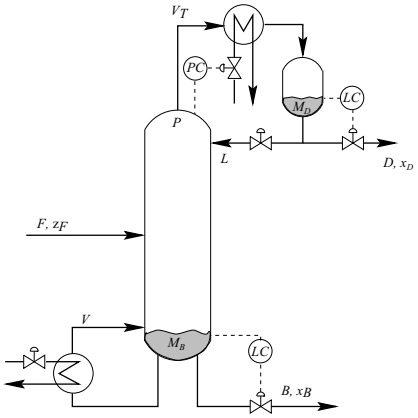



FIGURE 4.15 – Basic distillation column controlled with LV-configuration

| | |
|---|---|
| $F$ | feed rate [kmol/min] |
| $qF$ | fraction of liquid in feed |
| $z_F$ | feed composition [mole fraction] |
| $T_F$ | feed temperature [°C] |
| $F_M$ | feed molar flow [kmol/min] |
| $F_V$ | feed volumetric flow [kmol/min] |
| $D$, $B$ | distillate and bottom flow [kmol/min] |
| $L$, $V$ | reflux and boilup flow [kmol/min] |
| $M_D$, $M_B$ | condenser and reboiler holdup [kmol] |
| $x_D$, $x_B$ | distillate and bottom composition [mole fraction] |
| $T_i$ | tray temperature ($i = 1, \ldots, 41$) |

TABLE 4.8 – Distillation column process variables

The linear dynamic model of continuous distillation column, for use with MATLAB and/or SIMULINK, is provided by S. Skogestad (Skogestad, 1997) in "open-loop" form , i.e. uncontrolled. The column is "column A" studied in several

papers in the literature. A controlled and optimized version of this model is given in (P-M. Villalba, 2012) which will be used in this section. The distillation column benchmark is given in Figure 4.16.

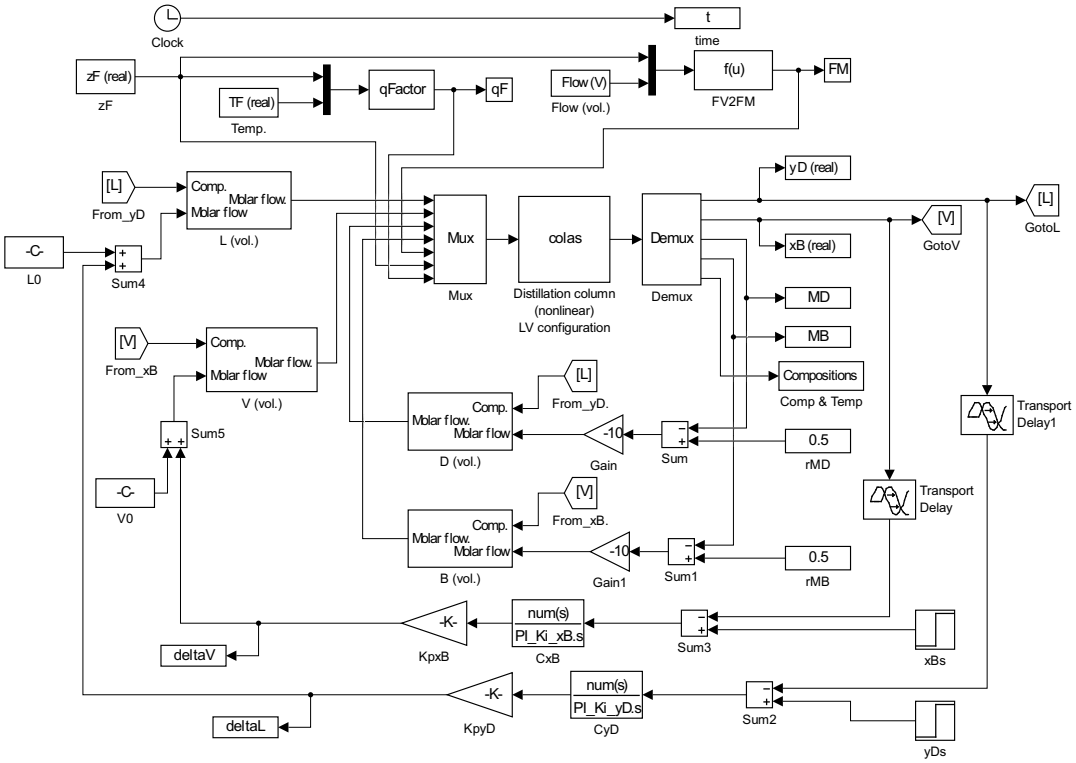



FIGURE 4.16 – Distillation column Benchmark

The column has NT=41 stages including reboiler and total condenser, the feed is at stage NF=21 counted from the bottom. The nominal conditions of the distillation column are as follows:

- Feed rate $F = 1$ [kmol/min],
- Feed composition $z_F = 0.5$ [mole fraction units]
- Feed liquid fraction $q_F = 1$ (i.e., saturated liquid)
- Reflux flow $L_T = 2.706$ [kmol/min]
- Boilup $V = 3.206$ [kmol/min].
- The nominal liquid holdup on all 41 stages is $M_{0i} = 0.5$ [kmol] (including the reboiler and condenser).
- The time constant for the liquid flow dynamics on each stage (except the reboiler and condenser) is $\tau = 0.063$ min. We assume that the vapor flow does not effect the liquid holdup, i.e. lambda=0.

This results in a distillate product with $D = 0.5$ [kmol/min] and composition $x_{NT} = 0.99$ [mole fraction units], and a bottoms product with $B = 0.5$ [kmol/min] and composition $x_B = x_1 = 0.01$ [mole fraction units].

We simulated the distillation column plant for 2 hours, under NOC, and collected 2000 samples, the monitored variables being the 12 variables of the process explained in Table 4.9. Assuming that each sensor measurements are stained with noise and are imprecise, with the uncertainty ratio for each variable supposedly represented by 20% of its variability, we construct the new interval data of the process. Afterwards, and from the NOC interval data, the interval-valued PCA model with four principal axes ($\ell = 4$) was chosen based on the $VIRE$ criterion. In Figure 4.17 the evolution of the different variables and their estimation based on interval-valued PCA model is presented, this shows that the estimation error is relatively small which demonstrates the accuracy of the used model.

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $B$ | $M_B$ | $D$ | $M_D$ | $F_M$ | $F_V$ | $L$ | $T_F$ | $V$ | $q_F$ | $x_B$ | $z_F$ |

TABLE 4.9 – Monitored Distillation Column Process Variables

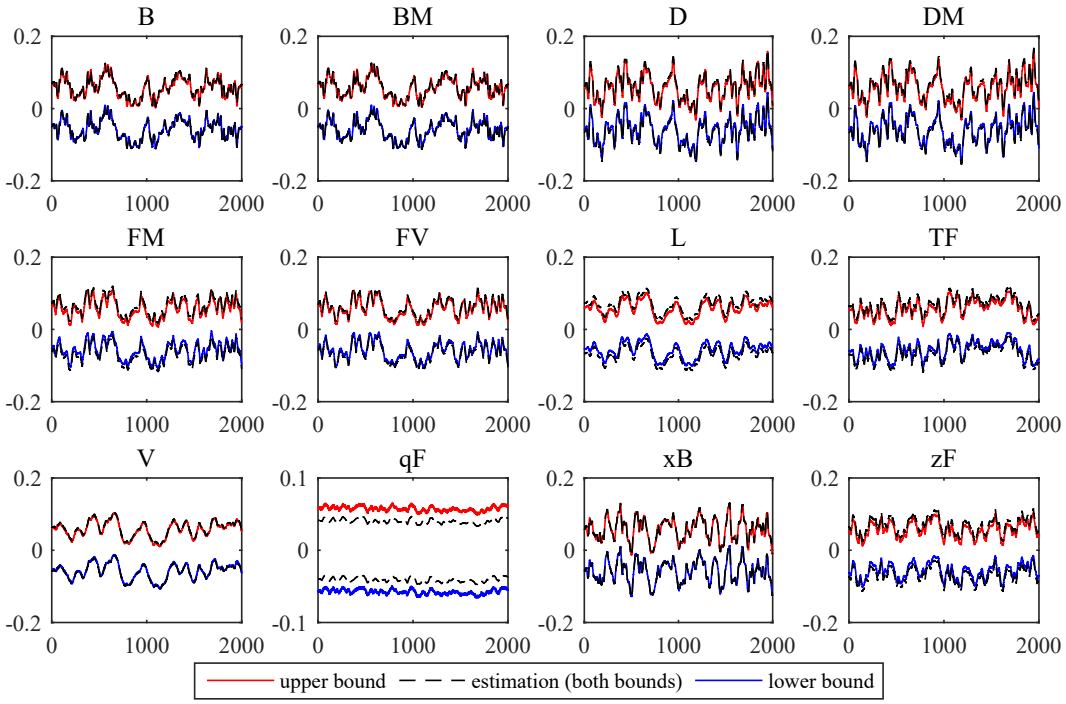



FIGURE 4.17 – Interval estimation based on PCA for interval-valued data model

To carry on with the monitoring procedure, a second simulation is performed where two different offsets are introduced to the system: the first is a sensor fault simulated as a bias in the feed molar flow ($F_M$), i.e. variable number 5, that starts at moment 1600 and ends at 2000, and represents 30% of the variable variation. The second offset is an uncertainty in feed composition $z_F$, , i.e. variable number 12, explained by 10% of the variable variation lasting from moment 500 to 1000.

In order to visualize the improvements perceived using interval-valued PCA in comparison with the classical PCA case, a conventional PCA model is built based on the NOC data of the plant. Figure 4.18 represents the time evolution of $SPE$ and $SWE$ charts. Because of the influence of uncertainties, the monitoring charts of conventional PCA in Figure 4.18 barely capture the fault in this uncertain process. Moreover, they exhibit a massive number of false alarms in the normal condition area, thus deteriorating the reliability of detection decisions made using this PCA model.

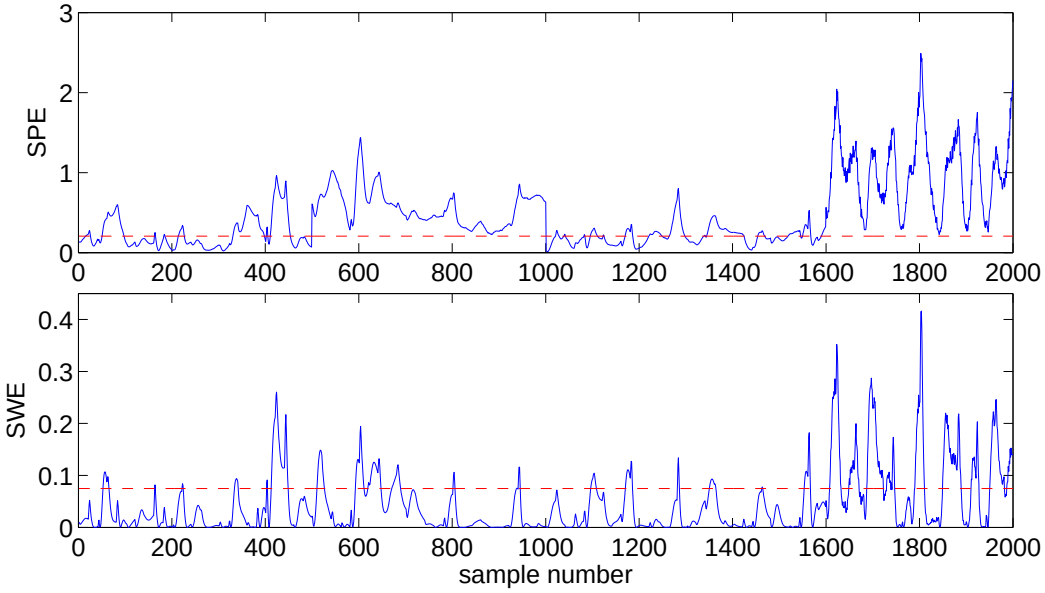



FIGURE 4.18 – Classical PCA fault detection charts (SPE and SWE)

For the interval case, results for fault detection are shown in Figures 4.19 and 4.20 for different variations of $ISPE$ and $ISWE$ respectively. On one hand, we notice the complete absence of the injected uncertainties, which are *a priori* to be detected from moment 500 to 1000, the explanation being that the interval-valued PCA model used hereby considers it as normal process variation and not a real fault. In the other hand, the real fault, introduced accounting from moment 1600, is correctly detected for the two indicators with all variations. However, the $ISPE$ and $ISWE$ charts contain a light number of false alarms, that are more frequent in the $ISWE$ case, but are mostly eliminated in the enhanced versions based on the EWMA filter.

In terms of process monitoring, a sampling period of time is usually determined, and the measurement is considered as the average of the collected measured values during that period. When this type of data is used, an excessive rate of false alarms and missing detection may occur depending on the averaging period of time. In addition to that, to obtain a reliable PCA model representing the NOC, one needs to use a good amount of samples. However, describing the measurement as an interval roughly covers the whole averaging period, and thus yields more information about the process than the averaged measure, even for

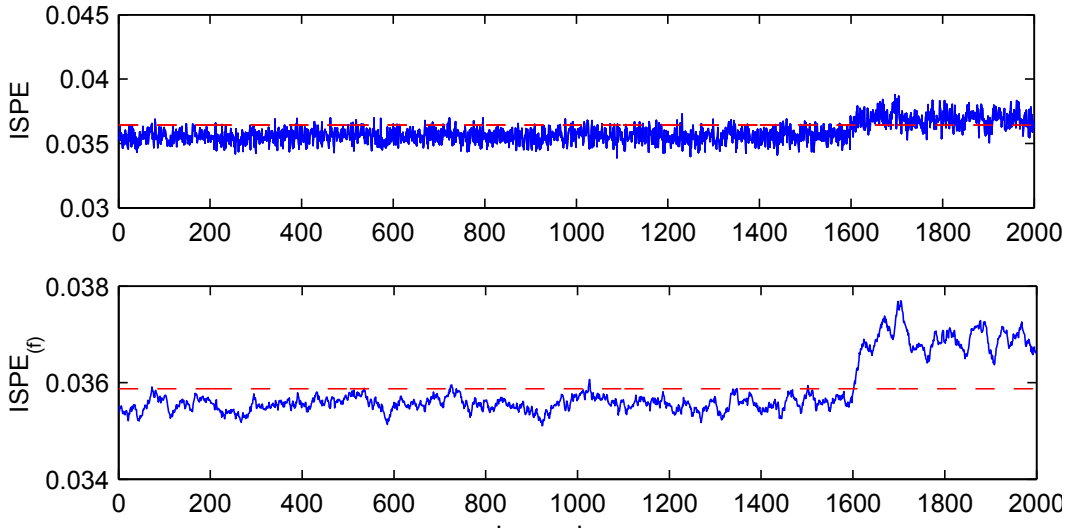



FIGURE 4.19 – Fault detection based on ISPE variations for interval-valued PCA

a few samples. This quality of information gathered from the process positively impacts on the interval-valued PCA model, which becomes consequently more robust to uncertainties and even to slight process variations.

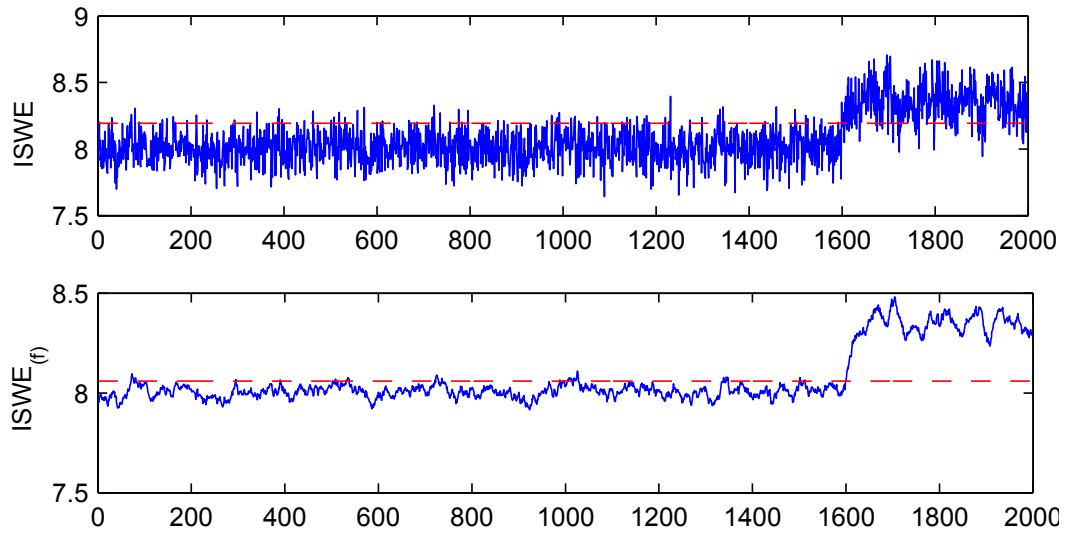



FIGURE 4.20 – Fault detection based on ISWE variations for interval-valued PCA

The next step in diagnosis routine is the isolation of the detected faults, which will be performed based on the interval reconstruction principle. Figure 4.21 represents the reconstructions for the 12 monitored variables based on the $ISPE$ chart, where the exponent stands for the reconstruction without the corresponding variable. We note that for different directions, the fault is detected

except in the 5th variable/direction which corresponds to the faulty variable $F_m$ according to the reconstruction principle. The isolation procedure is achieved based on the comparison between theoretical signatures of faults known *a priori* and experimental signatures determined from different reconstruction directions, given in Figure 4.21.

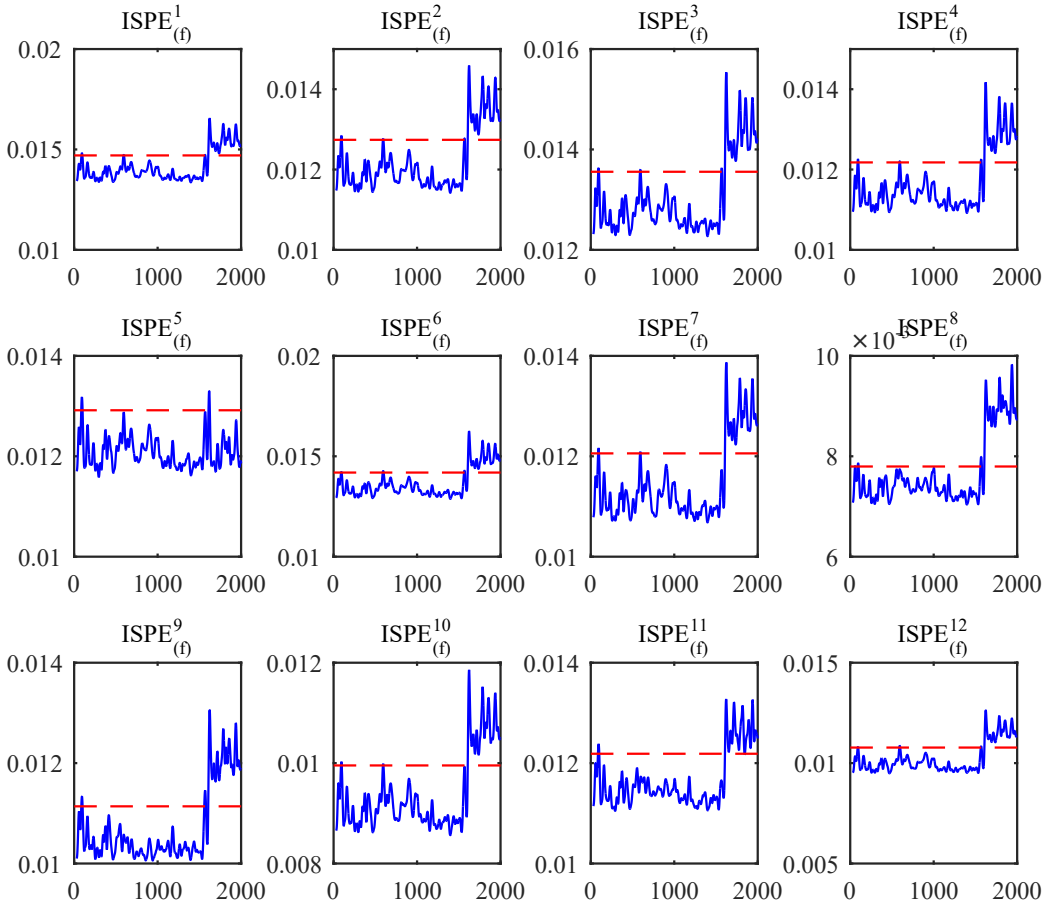



FIGURE 4.21 – Reconstructions based on $ISPE_{(f)}^{(i)}$

## 4.4 Conclusion

In this section, the proposed strategy for FDI using PCA for interval-valued data is applied for process monitoring, which ensures a robust FDI strategy as it combines sensor measurements and uncertainties using an interval representation. Throughout the manuscript, we have introduced different interval statistics based on interval-valued PCA model, namely, $ISPE$, $IT^2$ and $ISWE$, with improvements using EWMA filter for less noisy charts. In diagnosis framework, these control charts are validated and compared in terms of good detection and isolation ratios, along with their ability to neglect uncertainties. This validation was done based on a simulation example and using a Monte-Carlo simulation. In addition to that, a comparison between the four most known interval-valued

PCA methods is performed, namely VPCA, CPCA, MRPCA and CIPCA, where the latter is chosen as the most accurate model. Two real applications are presented: on a milling machine data and on a simulated distillation column process. The application demonstrate the good performances obtained using PCA for interval-valued data for FDI.

# Conclusions and Perspectives

This thesis presented a new sensor fault detection and isolation strategy for systems subject to uncertainties, which is based on principal component analysis for interval-valued data. PCA is a faithful data exploratory tool that has long been used in several engineering fields, and had known several improvements over time. Among the newest versions of PCA, is PCA for interval-valued data. Our purpose in this work, was to investigate the applicability of this powerful method for the purpose of fault detection and isolation. Indeed, for an FDI scheme, one can believe that any additional information on the process can only be beneficial, and especially when this information is about uncertainties.

We showed throughout this manuscript, that PCA for interval-valued data does apply for FDI, and performs better then conventional PCA approach, in addition to providing an ultimate robustness towards uncertainties of measurements. However, most of the theory behind PCA based FDI has been developed for single-valued data, so we had to forge most of the theory for an interval-valued PCA case. So, starting from the four most know PCA models for interval-valued data, known as VPCA, CPCA, MRPCA and CIPCA, we had to define fault indicators of interval type, and extensions of isolation methods to the interval case, in order to come up with a new strategy for FDI using interval-valued PCA. Overall, the main aspects of this dissertation can be summarized as follows.

— The key performance of PCA for interval-valued data for sensor FDI is its ability to neglect uncertainties of measurement by considering them as normal process variation. In other words, due to the interval nature of data, any information inside the interval is considered as normal variation of the process, while any data outside this interval is considered as a fault.

— Two strategies are presented. The first strategy is a univariate fault detection and isolation approach, i.e. FDI performed on generated residuals from the model. The second strategy is a multivariate strategy which is based on fault detection statistics.

— Several extensions for detection statistics are proposed for the interval-valued PCA case, from interval extensions of well known $SPE$, $T^2$ and $SWE$, to the new interval-norm based $ISPE$, $IT^2$ and $ISWE$. These new statistics give considerably better results in terms of decision preciseness, and are furthermore enhanced based on EWMA filtering method.

— Isolation of faults is presented based on an extension of the reconstruction principle for the interval-valued data case. This extension is different for

every interval-valued PCA model, and is detailed, then demonstrated for fault isolation.

— For the determination of the number of PC's, a new criterion is introduced, which is based on the different extensions of reconstruction principle. the so-called variance of interval reconstruction error (VIRE) ensures the minimum reconstruction error in the interval-valued PCA model for an accurate model, and consequently better performances in FDI.

— The proposed strategy is tested on a simulation example, then a Monte-Carlo simulation is lead in order to compare different interval-valued models and statistics. Then, a real application is presented for the milling machine data, and the distillation column process, showing the substantial improvements in precision for detecting faults, while perfectly handling uncertainties by considering them as process variations.

This research axis is proving itself very promising, and several open issues deserve further study for interval valued variables. One is the extension of the diagnosis strategy to the dynamic case, which is limited by the updating abilities of the interval model. Another possibility is the use of interval distances to construct kernel functions in order to handle the high dimensional non-linear data.

# A | Moore's Algorithm

An interval-valued variable $x$ is represented by an interval of the form $[x] = [\underline{x}\overline{x}]$, where $\underline{x} < \overline{x}$ (The upper bound of the interval is always greater than the lower bound).

Let $\gamma \in \mathcal{R}$ be a real scalar, then the interval-valued variable $[x]$ times $\gamma$ is given by:

$$\gamma \left[ \begin{array}{cc} \underline{x} & \overline{x} \end{array} \right] = \begin{cases} \left[ \begin{array}{cc} \gamma\underline{x} & \gamma\overline{x} \end{array} \right] & if \quad \gamma > 0 \\ \\ \left[ \begin{array}{cc} \gamma\overline{x} & \gamma\underline{x} \end{array} \right] & if \quad \gamma < 0 \end{cases} \tag{A.1}$$

This linear combination rule is called Moores's rule and used in order to avoid the problem in which the resulted lower bound value is greater than the upper bound value[1]. Equations for computation of interval components and estimates in VPCA, CPCA and CIPCA are obtained using the moore's rule to guarantee that the upper bound of the computed interval valued component or estimate is always greater than the lower bound.

1. Moore R. E., Interval Analysis, Prentice Hall, Englewood Cliffs, NJ., 1966.

# Bibliography

Agogino, A. and K. Goebel (2007). *Milling Data Set, NASA Ames Prognostics Data Repository*. URL: http://ti.arc.nasa.gov/project/prognostic-data-repositoryl.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2014a). "Fault Detection and Isolation of Uncertain Process Using Interval Principal Component Analysis". In: *International Conference on Technological Advances in Electrical Engineering (ICTAEE-2014)*. Skikda- Algeria.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2014b). "Fault Detection and Isolation Using Interval Vertices Principal Component Analysis". In: *3rd International Conference on Information Processing and Electrical Engineering (ICIPEE-2014)*. Tebessa- Algeria.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2015a). "Fault Detection and Isolation Using Interval Principal Component Analysis Methods". In: *IFAC-PapersOnLine* 48.21. 9th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS 2015, pp. 1402 –1407. DOI: https://doi.org/10.1016/j.ifacol.2015.09.721.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2015b). "Interval PCA Based Fault Detection and Isolation With New Interval SPE Statistic". In: *International Conference on Automatic control, Telecommunication and Signals (ICATS-2015)*. Annaba- Algeria.

Ait-Izem, T., W. Bougheloum, M-F. Harkat, and M. Djeghaba (2015c). "Vertices and Centers Principal Component Analysis for Fault Detection and Isolation". In: *2nd International Conference on Automation,Control, Engineering and Computer Science (ACECS-2015)*. Sousse- Tunisia: Proceedings of Engineering and Technology (PET).

Ait-Izem, T., M-F. Harkat, F. Kratz, and M. Djeghaba (2017a). "Approche Neuronale d'ACP par Intervalle Appliquèe au Diagnostic". In: *12 ème Congrès International Pluridisciplinaire en Qualité, Sûreté de fonctionnement et Développement durable, (Qualita-2017)*. Bourges- France.

Ait-Izem, T., M-F. Harkat, M. Djeghaba, and F. Kratz (2017b). "Sensor Fault Detection Based on Principal Component Analysis for Interval-Valued Data". In: *Quality Engineering*. DOI: 10.1080/08982112.2017.1391288.

Ait-Izem, T., M-F. Harkat, M. Djeghaba, and F. Kratz (2018). "On the Application of Interval PCA to Process Monitoring: A Robust Strategy for Sensor FDI with New Efficient Control Statistics". In: *Journal of Process Control* 63, pp. 29 –46. ISSN: 0959-1524. DOI: https://doi.org/10.1016/j.jprocont.2018.01.006.

Alcala, Carlos F. and S. Joe Qin (2009). "Reconstruction-based contribution for process monitoring". In: *Automatica* 45.7, pp. 1593 –1600. ISSN: 0005-1098. DOI: 10.1016/j.automatica.2009.02.027.

Bartlett, M. S. (1950). "Tests of significance in factor analysis". In: *British Journal of Statistical Psychology* 3.2, pp. 77–85. ISSN: 2044-8317. DOI: 10.1111/j.2044-8317.1950.tb00285.x.

Benaicha, A., G. Mourot, K. Benothman, and J. Ragot (2013). "Fault detection and isolation with Interval Principal Component Analysis". In: International Conference on Control, Engineering and Information Technology. Proceedings Engineering and Technology PET, pp. 162–167.

Bertrand, P. and F. Goupil (2000). "Descriptive Statistics for Symbolic Data". In: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Berlin Heidelberg, pp. 106–124.

Billard, L. (2008). "Sample Covariance Functions for Complex Quantitative Data". In: *in Proceedings World Conferences International Association of Statistica lComputing*. Japanese Society of Computational Statistics, pp. 157–163.

Box, G.E.P. (1954). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification". In: 25.

Carvalho, Francisco de A. T. de, Paula Brito, and Hans-Hermann Bock (2006). "Dynamic clustering for interval data based on L2 distance". In: *Computational Statistics* 21.2, pp. 231–250. ISSN: 1613-9658. DOI: 10.1007/s00180-006-0261-z.

Cattell, Raymond B. (1966). "The Scree Test For The Number Of Factors". In: *Multivariate Behavioral Research* 1.2, pp. 245–276. DOI: 10.1207/s15327906mbr0102_10.

Cazes, P., A. Chouakria, E. Diday, and Y Schektman (1997). "Extension de l'Analyse en Composantes Principales à des Données de Type Intervalle". In: *Revue de Statistique Appliquée* 45.3, pp. 5–24.

Chen, Gang and Thomas J. McAvoy (1998). "Predictive on-line monitoring of continuous processes". In: *Journal of Process Control* 8.5, pp. 409 –420. ISSN: 0959-1524. DOI: https://doi.org/10.1016/S0959-1524(98)00023-7.

Chen, Meiling, Huiwen Wang, and Zhongfeng Qin (2015). "Principal component analysis for probabilistic symbolic data: a more generic and accurate algorithm". In: *Advances in Data Analysis and Classification* 9.1, pp. 59–79. URL: https://EconPapers.repec.org/RePEc:spr:advdac:v:9:y:2015:i:1:p:59-79.

Chouakria, A. (1998). "Extension des Methodes d'analyse Factorielle à des Données de Type Intervalle". PhD thesis. Université Paris-Dauphine.

Chouakria, D. A., L. Billard, and E. Diday (2011). "Principal component analysis for interval-valued observations". In: *Statistical Analysis and Data Mining* 4.2, pp. 229–246. ISSN: 1932-1872. DOI: 10.1002/sam.10118.

Cooley, W.W. and P.R. Lohnes (1971). *Multivariate data analysis*. Wiley. ISBN: 9780471170600.

De Ketelaere, Bart, Mia Hubert, and Eric Schmitt (2015). *Overview of pca-based statistical process monitoring methods for time-dependent high-dimensional data*.

Deif, Assem (1991). "The Interval Eigenvalue Problem". In: *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 71.1, pp. 61–64. ISSN: 1521-4001. DOI: 10.1002/zamm.19910710117.

Diday, E. (1987). "Introduction à l'approche symbolique en analyse des Données". In: *Journées symboliques-numériques*.

Dong, D. and T.J. McAvoy (1996). "Nonlinear principal component analysis—Based on principal curves and neural networks". In: *Computers and Chemical Engineering* 20.1, pp. 65–78. ISSN: 0098-1354. DOI: 10.1016/0098-1354(95)00003-K.

Doymaz, Fuat, Jose A. Romagnoli, and Ahmet Palazoglu (2001). "A strategy for detection and isolation of sensor failures and process upsets". In: *Chemometrics and Intelligent Laboratory Systems* 55.1-2, pp. 109–123. ISSN: 01697439. DOI: 10.1016/S0169-7439(00)00126-X.

Dunia, Ricardo and S. Joe Qin (1998). "Subspace approach to multidimensional fault identification and reconstruction". In: *AIChE Journal* 44.8, pp. 1813–1831. ISSN: 1547-5905. DOI: 10.1002/aic.690440812.

Dunia, Ricardo and S. Joe Qin (1998). "A unified geometric approach to process and sensor fault identification and reconstruction: the unidimensional fault case". In: *Computers and Chemical Engineering* 22.7, pp. 927 –943. ISSN: 0098-1354. DOI: https://doi.org/10.1016/S0098-1354(97)00277-9.

Dunia, Ricardo, S. Joe Qin, Thomas F. Edgar, and Thomas J. McAvoy (1996a). "Identification of faulty sensors using principal component analysis". In: *AIChE Journal* 42.10, pp. 2797–2812. ISSN: 1547-5905. DOI: 10.1002/aic.690421011.

Dunia, Ricardo, S Joe Qin, Tf Edgar, and Tj McAvoy (1996b). "Use of principal component analysis for sensor fault identification". In: *Computers and chemical engineering* 20.96, pp. 713–718. ISSN: 00981354. DOI: 10.1016/0098-1354(96)00128-7. URL: http://www.sciencedirect.com/science/article/pii/0098135496001287.

D'Urso, P. and P. Giordani (2004). "A least squares approach to principal component analysis for interval valued data". In: *Chemometr Intell Lab Syst* 70.2, pp. 179,192.

Gertler, Janos, Weihua Li, Yunbing Huang, and Thomas McAvoy (1999). "Isolation enhanced principal component analysis". In: *AIChE Journal* 45.2, pp. 323–334. ISSN: 1547-5905. DOI: 10.1002/aic.690450213.

Gioia, F. and C Lauro (2006). "Principal component analysis on interval data". In: *Computational Statistics* 21, pp. 343–363.

Harkat, Mohamed-Faouzi (2002). "Détection et localisation de défauts par analyse en composantes principales". PhD thesis. Institut National Polytechnique de Lorraine-INPL.

Harkat, Mohamed-Faouzi, Gilles Mourot, and Jose Ragot (2006). "An improved PCA scheme for sensor FDI: Application to an air quality monitoring network". In: *Journal of Process Control* 16.6, pp. 625 –634. ISSN: 0959-1524. DOI: 10.1016/j.jprocont.2005.09.007.

Harmon, J. L. and Duboc (1995). "Factor analytical modeling of biochemical data". In: *Comput. Chem.* 19, p. 1287.

Harrou, F., M. Nounou, and H. Nounou (2013). "A statistical fault detection strategy using PCA based EWMA control schemes". In: *2013 9th Asian Control Conference (ASCC)*, pp. 1–4. DOI: 10.1109/ASCC.2013.6606311.

Horn, J. L (1965). "A rationale and test for the number of factors in factor analysis". In: *Psychometrica* 30, p. 73.

Hotelling, H (1933). "Analysis of a complex of statistical variables into principal components". In: *J. Educ. Psychol.* 24, p. 417.

Hotelling, H (1947). *Techniques of Statis-tical Analysis- Multivariate quality control-illustrated by air testing of sample bombsights*. Mcgraw-Hill, New York, pp. 11–148.

Huang, Yunbing, Janos Gertler, and Thomas J. McAvoy (1999). "Fault Isolation by Partial PCA and Partial NLPCA". In: *IFAC Proceedings Volumes* 32.2. 14th IFAC World Congress 1999, Beijing, Chia, 5-9 July, pp. 7647 –7652. ISSN: 1474-6670. DOI: https://doi.org/10.1016/S1474-6670(17)57305-X.

Irpino, Antonio (2006). ""Spaghetti" PCA analysis: An extension of principal components analysis to time dependent interval data". In: *Pattern Recognition Letters* 27.5, pp. 504 –513. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2005.09.013.

Isermann, R. (2005). *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer Berlin Heidelberg. ISBN: 9783540241126.

Jackson, J. Edward (1991). *A Users Guide to Principal Components*. John Wiley and Sons, Inc. ISBN: 9780471725336. DOI: 10.1002/0471725331.

Jackson, J. Edward and Govind S. Mudholkar (1979). "Control Procedures for Residuals Associated With Principal Component Analysis". In: *Technometrics* 21.3, pp. 341–349. DOI: 10.1080/00401706.1979.10489779.

Jia, F., E. B. Martin, and A. J. Morris (2000). "Non-linear principal components analysis with application to process fault detection". In: *International Journal of Systems Science* 31.11, pp. 1473–1487. DOI: 10.1080/00207720050197848.

Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer. ISBN: 9780387954424. URL: https://books.google.dz/books?id=_olByCrhjwIC.

Kaiser, Henry F. (1960). "The Application of Electronic Computers to Factor Analysis". In: *Educational and Psychological Measurement* 20.1, pp. 141–151. DOI: https://doi.org/10.1177/001316446002000116.

Kramer, Mark A. (1991). "Nonlinear principal component analysis using autoassociative neural networks". In: *AIChE Journal* 37.2, pp. 233–243. ISSN: 1547-5905. DOI: 10.1002/aic.690370209.

Kresta, James V., John F. Macgregor, and Thomas E. Marlin (1991). "Multivariate statistical monitoring of process operating performance". In: *The Canadian Journal of Chemical Engineering* 69.1, pp. 35–47. ISSN: 1939-019X. DOI: 10.1002/cjce.5450690105.

Ku, W., R. H. Storer, and C Georgakis (1995). "Disturbance detection and isolation by dynamic principal component analysis". In: *Chemom. Intell. Lab. Syst.* 30, p. 179.

Lauro, Carlo N. and Francesco Palumbo (2000). "Principal component analysis of interval data: a symbolic data analysis approach". In: *Computational Statistics* 15.1, pp. 73–87. ISSN: 09434062. DOI: 10.1007/s001800050038.

Lauro, N. Carlo, Rosanna Verde, and Antonio Irpino (2008). "Principal Component Analysis of Symbolic Data Described by Intervals". In: *Symbolic Data Analysis and the SODAS Software*. John Wiley and Sons, pp. 279–311. ISBN: 9780470723562. DOI: 10.1002/9780470723562.ch15.

Le-Rademacher, J. and L Billard (2012). "Symbolic Covariance Principal Component Analysis and Visualization for Interval-Valued Data". In: *Journal of Computational and Graphical Statistics* 21.2, pp. 413–432. DOI: 10.1080/10618600.2012.679895.

Li, Weihua, H.Henry Yue, Sergio Valle-Cervantes, and S.Joe Qin (2000). "Recursive PCA for adaptive process monitoring". In: *Journal of Process Control* 10.5, pp. 471 –486. ISSN: 0959-1524. DOI: https://doi.org/10.1016/S0959-1524(00)00022-6.

MacGregor, J. F. and T Kourti (1995). "Statistical process control of multivariate processes". In: *Control Eng. Practice* 3, p. 414.

MacGregor, John F., Christiane Jaeckle, Costas Kiparissides, and M. Koutoudi (1994). "Process monitoring and diagnosis by multiblock PLS methods". In: *AIChE Journal* 40.5, pp. 826–838. ISSN: 1547-5905. DOI: 10.1002/aic.690400509.

Miller, P., R. E. Swanson, and C. E. Heckler (1998). "Contribution plots: a missing link in multivariate quality control". In: *Applied Mathematics and Computer Science* Vol. 8, no 4, pp. 775–792.

Mnassri, Baligh (2012). "Analyse de données multivariées et surveillance des processus industriels par analyse en composantes principales". PhD thesis. 'Université d'Aix-Marseille.

Moore, R. (1966). *Interval Analysis*. Prentice Hal.

Nasri, Othman, Imen Gueddi, Philippe Dague, and Kamal Benothman (2015). "Spacecraft Actuator Diagnosis with Principal Component Analysis: Application to the Rendez-Vous Phase of the Mars Sample Return Mission". In: *Journal of Control Science and Engineering*, p. 11. DOI: 10.1155/2015/204918.

Neumaier, A. (1990). *Interval Methods for Systems of Equations*. Cambridge Middle East Library. Cambridge University Press. ISBN: 9780521331968.

Nomikos, P. and J. F MacGregor (1995). "Multivariate SPC charts for monitoring batch process". In: *Technometrics* 37, p. 414.

P-M. Villalba, Torán (2012). "Multivariate Statistical Process Monitoring of a Distillation Column". PhD thesis. Universidad Politecnica de Valencia.

Palumbo, F. and N.C Lauro (2003). "A PCA for Interval-Valued Data Based on Midpoints and Radii". In: *in New Developments in Psychometrics, eds. H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. Meulman, Tokyo*, pp. 641–648.

Pearson, K. (1901). "On lines and planes of closest fit to systems of points in space". In: *The london, Edinburg and Dublin Phylosophical Magazine and Journal of Science* 6, pp. 559–572.

Piovoso, M. J., K. A. Kosanovich, and R. K Pearson (1992). "Monitoring process performance in real-time. In Proceedings of ACC". In: *Chicago* 2359, p. 2363.

Qin, S. Joe (2003). "Statistical process monitoring: basics and beyond". In: *Journal of Chemometrics* 17.8-9, pp. 480–502. ISSN: 1099-128X. DOI: 10.1002/cem.800.

Qin, S. Joe (2012). "Survey on data-driven industrial process monitoring and diagnosis". In: *Annual Reviews in Control* 36.2, pp. 220 –234. ISSN: 1367-5788. DOI: 10.1016/j.arcontrol.2012.09.004.

Qin, S. Joe, Hongyu Yue, and Ricardo Dunia (1997). "Self-Validating Inferential Sensors with Application to Air Emission Monitoring". In: *Industrial and Engineering Chemistry Research* 36.5, pp. 1675–1685. DOI: 10.1021/ie960615y.

Rao, C. Radhakrishna (1964). "The Use and Interpretation of Principal Component Analysis in Applied Research". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26.4, pp. 329–358. ISSN: 0581572X. URL: http://www.jstor.org/stable/25049339.

Rato, Tiago, Marco Reis, Eric Schmitt, Mia Hubert, and Bart De Ketelaere (2016). "A systematic comparison of PCA-based Statistical Process Monitoring methods for high-dimensional, time-dependent Processes". In: *AIChE Journal* 62.5, pp. 1478–1493. ISSN: 1547-5905. DOI: 10.1002/aic.15062.

Rohn, J. and A. Deif (1991). "On the Range of Eigenvalues of an Interval Matrix". In: *Computing* 47.3-4, pp. 373–377. ISSN: 0010-485X. DOI: 10.1007/BF02320205.

Russell, E.L., L.H. Chiang, and R.D. Braatz (2012). *Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes*. Advances in Industrial Control. Springer London. ISBN: 9781447104094.

Scholkopf, B., A. Smola, and K. R. Muller (1998). "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". In: *Neural Computation* 10.5, pp. 1299–1319. ISSN: 0899-7667. DOI: 10.1162/089976698300017467.

Sharmin, Rumana, Sirish L. Shah, and Uttandaraman Sundararaj (2008). "A PCA Based Fault Detection Scheme for an Industrial High Pressure Polyethylene Reactor". In: *Macromolecular Reaction Engineering* 2.1, pp. 12–30. ISSN: 1862-8338. DOI: 10.1002/mren.200700023.

Sjöstrjöm, Michael, Svante Wold, and Bengt Sjöderstrjöm (1986). "PLS discriminant plots". In: *Pattern Recognition in Practice*. Elsevier, pp. 461 –470. ISBN: 978-0-444-87877-9. DOI: https://doi.org/10.1016/B978-0-444-87877-9.50042-X.

Skogestad, S (1997). "Dynamics and Control of Distillation Columns". In: *Chemical Engineering Research and Design* 75.6, pp. 539–562. ISSN: 0263-8762. DOI: 10.1205/026387697524092.

Teti, R., K. Jemielniak, G. O'Donnell, and D. Dornfeld (2010). "Advanced monitoring of machining operations". In: *CIRP Annals* 59.2, pp. 717 –739. ISSN: 0007-8506. DOI: https://doi.org/10.1016/j.cirp.2010.05.010.

Tien, D. X., K. W. Lim, and L. Jun (2004). "Comparative study of PCA approaches in process monitoring and fault detection". In: *30th Annual Conference of IEEE Industrial Electronics Society, 2004. IECON 2004*. Vol. 3, 2594–2599 Vol. 3. DOI: 10.1109/IECON.2004.1432212.

Tracy, Nola, John Young, and Robert Mason (1992). "Multivariate Control Charts for Individual Observations". In: *Journal of Quality Technology* 24.2, pp. 88–92.

Tulsyan, Aditya and Paul I Barton (2017a). "Interval enclosures for reachable sets of chemical kinetic flow systems. Part 1: Sparse transformation". In: *Chemical Engineering Science* 166, pp. 334–344.

Tulsyan, Aditya and Paul I Barton (2017b). "Interval enclosures for reachable sets of chemical kinetic flow systems. Part 2: Direct-bounding method". In: *Chemical Engineering Science* 166, pp. 345–357.

Tulsyan, Aditya and Paul I Barton (2017c). "Interval enclosures for reachable sets of chemical kinetic flow systems. Part 3: Indirect-bounding method". In: *Chemical Engineering Science* 166, pp. 358–372.

Valle, Sergio, Weihua Li, and S. Joe Qin (1999). "Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods". In: *Industrial and Engineering Chemistry Research* 38.11, pp. 4389–4401. DOI: 10.1021/ie990110i.

Velicer, Wayne F. (1976). "Determining the number of components from the matrix of partial correlations". In: *Psychometrika* 41.3, pp. 321–327. ISSN: 1860-0980. DOI: 10.1007/BF02293557.

Venkatasubramanian, Venkat, Raghunathan Rengaswamy, Kewen Yin, and Surya.N Kavuri (2003a). "A review of process fault detection and diagnosis part I: Quantitative model-based methods". In: *Computers and Chemical Engineering* 27, pp. 293–311. DOI: 10.1016/S0098-1354(02)00160-6.

Venkatasubramanian, Venkat, Raghunathan Rengaswamy, and Surya.N Kavur (2003b). "A review of process fault detection and diagnosis Part II : Qualitative models and search strategies". In: *Computers and Chemical Engineering* 27, pp. 313–326.

Venkatasubramanian, Venkat, Raghunathan Rengaswamy, Surya.N Kavuri, and Kewen Yin (2003c). "A review of process fault detection and diagnosis Part III : Process history based methods". In: *Computers and Chemical Engineering* 27, pp. 327–346.

Wang, H., R. Guan, and J. Wu (2012). "CIPCA: Complete-Information-based Principal Component Analysis for Interval-valued Data". In: *Neurocomput.* 86, pp. 158–169. DOI: 10.1016/j.neucom.2012.01.018.

Wang, Shengwei and Fu Xiao (2004). "Detection and diagnosis of AHU sensor faults using principal component analysis method". In: *Energy Conversion and Management* 45.17, pp. 2667–2686. DOI: 10.1016/j.enconman.2003.12.008.

Wang, Xun, Uwe Kruger, and George W. Irwin (2005). "Process Monitoring Approach Using Fast Moving Window PCA". In: *Industrial and Engineering Chemistry Research* 44.15, pp. 5691–5702. DOI: 10.1021/ie048873f.

Wax, M. and T Kailath (1985). "Detection of signals by information criteria". In: *IEEE Trans. Acoust. Speech Signal Process. ASSP-33* 387, p. 392.

Wold, S (1978). "Cross validatory estimation of the number of components in factor and principal components analysis". In: *Technometrics* 20, p. 406.

Wold, S. (1994). "Exponentially weighted moving principal components analysis and projections to latent structures". In: *Chemometrics and Intelligent Laboratory Systems* 23.1. Proceedings of the 3rd Scandinavian Symposium on Chemometrics (SSC3), pp. 149 –161. ISSN: 0169-7439. DOI: 10.1016/0169-7439(93)E0075-F.

Wold, S., K. Esbensen, and P Geladi (1987). "Principal Component Analysis". In: *Chemom. Intell. Lab. Syst.* 2, p. 52.

Yoon, Seongkyu and John F. MacGregor (2001). "Fault diagnosis with multivariate statistical models part I: using steady state fault signatures". In: *Journal of Process Control* 11.4, pp. 387 –400. ISSN: 0959-1524. DOI: 10.1016/S0959-1524(00)00008-1.

Zwick, W. R. and W. F Velicer (1986). "Comparison of five rules for determining the number of components to retain". In: *Psychol. Bull.* 99, p. 442.