الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière Électronique
Spécialité Instrumentation

Présenté par

Kerarsi Mohamed Zakaria

&

Rahmani AbdeRaouf

Validation des signaux par méthode statistique (PLS)

Proposé par : Khentout Nourddine & Ykhlef Farid

Année Universitaire 2020-2021

Remerciements

A l'issue de cette fin de travail, nous remercions le Dieu tout puissant pour la volonté, la santé et la patience qu'il nous a donné durant toutes ces longues années d'études. Le mémoire de master présenté ici est le résultat d'un travail continu de plus d'une année.

Il n'aurait pas pu être élaboré sans l'aide et la collaboration, de près ou de loin, de nombreuses personnes que nous tenons ici à remercier.

Nous remerciements chaleureux à nos parents, qui ont été toujours à nos côtés pour nous encourager à atteindre nos objectifs académique et personnel. Nous souhaitons qu'ils soient fiers de nous.

Nous tenons à exprimer nos remerciements à nos grandes familles, qu'ont été aussi parmi les partenaires qui nous encourager.

Nous exprimons nos sincères remerciements à Monsieur **Khentout Nour eddine** qui a dirigé et initié cette activité de recherche. Sa connaissance du sujet, ses conseils et son aide continus ont été essentiels. Pour les résultats de notre travail.

Nous souhaiterons exprimer nos sincères remerciements à Monsieur **Ykhlef Farid**, notre professeur et notre guide de ce travail. Nous voudrions aussi lui exprimer nos profondes reconnaissances pour la confiance qu'il nous a toujours témoignée.

Nous adressons nos remerciements à tous les professeurs qui nous ont enseigné pendant notre périple académique.

Nous avons été sensibles à l'honneur que nous fait l'ensemble des professeurs de participer aux jurys et nous les remerciant sincèrement.

Enfin, nos reconnaissances s'adressent à tous nos camarades d'Université pour les bons moments que nous avons passés en leur compagnie et nos ami(e)s.

ملخص: الهدف الأساسي لهذه الذاكرة هو اقتراح طريقة إحصائية (مربع أقل جزئية) تسمح بالتأكد من صحة الإشارة وكشف العيوب في النظام. هو تقنية الحد من الأبعاد. وتتألف هذه الطريقة من تقدير البيانات بمقارنة هذا التقدير بالبيانات الموجودة باستخدام الخوارزمية مربع أقل جزئية.

كلمات مفتاحية: كشف العيوب؛ تحقق(تصديق) من الإشارات؛ تقدير؛ انحدار؛ المنهجية الإحصائية؛ المربعات الجزئية الأقل.

Résumé : L'objectif fondamental de ce mémoire est de proposer une méthode statistique (moindre carré partiel, PLS) permettant la validation du signal et de détecter des défauts dans un système. PLS est une technique de réduction des dimensions. Cette méthode consiste à l'estimation des données en comparent cette estimation avec les donnes existantes en utilisant l'algorithme PLS.

Mots clés : Détection de défauts ; Validation des signaux ; Estimation ; Régression ; Méthodes statistiques ; Moindres carrés partiels.

Abstract: The fundamental objective of this memory is to propose a statistical method (less partial square, PLS) allowing the validation of the signal and to detect defects in a system. PLS is a dimension reduction technique. This method consists of estimating the data by comparing this estimate with the existing data using the PLS algorithm.

Keywords: Fault detection; Signal validation; Estimation; Regression; Statistical method; Partial least squares.

Listes des acronymes et abréviations

FD/DD : Fault Detection / Détection des défauts

PLS/MCP : Partial Least Squares / Moindres Carrés Partiels

PLSR : Partial Least Squares Regression

PCA/ACP : Principal Components Analysis / Analyse en composantes principales

SLR/RLS : Simple Linear Regression / Régression linéaire simple

FDA/ADF : Fisher Discriminant Analysis / L'analyse Discriminante de Fisher

MLR/RLM : Multiple Linear Regression / Régression linéaire Multiple

SVD/DVS : Singular Value Decomposition / Décomposition en Valeurs Singulières

EVD/DVP : Eigenvalue Value Decomposition / Décomposition des valeurs propres

MRA/ARM : Maximum Redundancy Analysis / Analyse de redondance maximale

LCI : Limite de Contrôle minimale Inférieure

LCS : Limite de Contrôle maximale Supérieure

VL : Variables Latentes

Var : Variance

E : Espérance

Cov : Covariance

Table des matières

Liste des figures

Introduction générale

La détection de défauts est étudiée depuis longtemps. Il existe de nombreux articles scientifiques sur la détection et le diagnostic des défauts. L'importance est que la détection des pannes au fil du temps alors que le système est toujours en cours d'exécution peut empêcher des événements anormaux et des pertes, évitant ainsi les pannes et les catastrophes majeures du système. Il existe de nombreuses façons de détecter les pannes. Ils peuvent être divisés en deux catégories, basés sur des modèles et basés sur des données. Dans la détection de défauts basée sur un modèle, le modèle de processus exact ou le modèle mathématique du système doit être connu. Dans ce cas, la technologie basée sur les données peut être utilisée pour la détection et le diagnostic des défauts. De nombreuses usines de contrôle de processus ont un grand nombre d'enregistrements de paramètres d'usine, tels que la pression, la température, le débit, et beaucoup plus. Dans des conditions de fonctionnement normales et défectueuses. Ces mégadonnées peuvent être utilisées pour construire des modèles statistiques pour la détection future de défauts. Plus les données sont grandes, plus les résultats du modèle statistique construit sont précis.

L'architecture générale du système de diagnostic à base de redondance analytique pour la détection et la localisation de défaillances de capteurs est représentée sur la figure 1 (sachant qu'on ne s'intéresse qu'à des défaillances d'instrumentation). Dans une première étape, il s'agit de comparer les observations avec les connaissances sur le comportement normal du système contenu dans un modèle afin de vérifier sa cohérence. Cette comparaison impliquera la génération d'indicateurs de défauts appelés résidus. Ces indicateurs sont souvent des écarts entre les caractéristiques observées et les caractéristiques de références qui définissent le comportement normal du système.

La Figure 2 présente le principe de base de génération de résidus. Généralement, les méthodes de génération des résidus sont basées soit sur une estimation d'état ou une estimation paramétrique.

Figure 1-Schéma de principe de la détection et de la localisation de défauts à base de modèles.

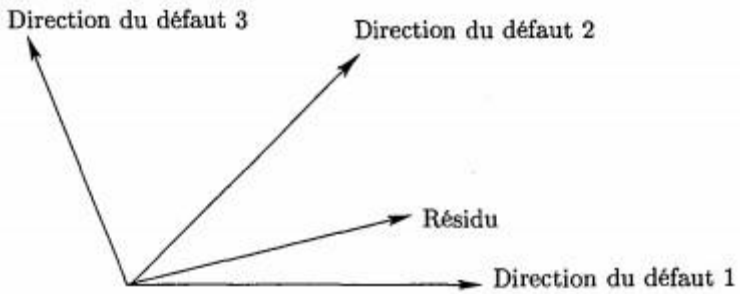

Figure 2- Exemple de résidu directionnel ou le résidu est dans la direction du défaut 1.

Dans ce travail, on s'intéresse à utilisation des méthodes statistiques et les moindres carrés partiels (PLS) pour construire un modèle de détection des défauts. Cela peut être utilisé là où nous avons des usines avec des processus presque constants et où nous avons beaucoup de données pendant le fonctionnement normal et les opérations d'usine défectueuses. Ceci est généralement observé dans les installations de contrôle de processus qui sont courantes dans les systèmes de contrôle réels. La tâche principale du contrôleur est d'assurer la stabilité de l'installation. L'influence des interférences, du bruit et des interférences doit être minimisée et des performances de contrôle optimisées doivent être obtenues.

Il existe de nombreux processus dans le monde réel, tels que : les réacteurs, les pompes d'échange de chaleur et les compresseurs. Les variables à contrôler dans ce cas sont principalement la température, la pression, la vitesse du réacteur et le niveau d'eau. En raison de tous ces processus et variables de processus, les données dans l'installation de contrôle de processus seront énormes et il est difficile de détecter les défauts au bon moment.

Objectif

La détection et le diagnostic précis et efficaces des défauts dans les systèmes d'ingénierie modernes sont essentiels pour garantir la fiabilité, la sécurité et le maintien de la qualité requise du produit. Dans ce mémoire, nous proposons une méthode innovante pour détecter des défauts dans des données multivariées fortement corrélées. La méthode développée utilise la méthode des moindres carrés partiels (PLS) comme approche de modélisation.

Les performances de l'algorithme de détection de défauts, PLS, serons illustrées et comparées avec d’autres méthodes de détection de défauts.

Chapitre 1 Notions générales

1.1 Introduction

Les défis économiques ont conduit, de plus en plus, à des restrictions sur la production. Dans l'environnement où les performances sont critiques, une moindre défaillance du processus est préjudiciable. Par conséquent, il est nécessaire de s'assurer, en permanence, que ce processus est exécuté dans des conditions et d'une façon meilleure. Grâce aux mesures des variables et paramètres de ce processus, des informations du comportement du système peuvent être obtenues. Alors, la qualité de ces mesures est un élément essentiel pour des performances meilleures des processus. Cette qualité des mesures est, particulièrement, fonction de la précision de l'instrument et au nombre de capteurs.

La redondance matérielle, plusieurs capteurs mesurent la même grandeur, est conçue pour une utilisation dans l'industrie de haute technologie. Cependant, ce type de redondance matérielle ne peut empêcher la défaillance de certains composants communs de la chaîne de mesure. L'avantage de la redondance analytique n'augmente pas les coûts d'installation et s'affranchit des contraintes matérielles.

Dans le domaine du détection et diagnostic des défauts, des méthodes basées sur le concept de redondance de l'information ont été développées. Leur principe repose généralement sur le test de cohérence entre le comportement observé du processus fourni par le capteur et le comportement attendu fourni par la représentation mathématique du processus. Par conséquent, l'analyse de ces méthodes de redondance nécessite la disponibilité d'un modèle de système défini en fonctionnement normal. La comparaison entre le comportement réel du système et le comportement attendu donné par le modèle fournit une quantité appelée le résidu, qui sera utilisée pour déterminer si le système est défaillant [1].

1.2 Opération sur les matrices

La matrice est un tableau rectangulaire de nombres ou d'expressions disposés en lignes et en colonnes. Des applications importantes des matrices peuvent être trouvées en mathématiques. Les opérations matricielles impliquent principalement trois opérations algébriques qui sont l'addition, la soustraction et la multiplication. En plus, d'autres opérations sur les matrices sont souvent utilisés telles que la moyenne centrée et la décomposition.

1.2.1 Operations arithmétiques

L'addition, la soustraction et la multiplication sont les opérations de base sur la matrice. Pour ajouter ou soustraire des matrices, celles-ci doivent être d'ordre identique et pour la multiplication, le nombre de colonnes de la première matrice est égal au nombre de lignes de la deuxième matrice.

1.2.2 Moyenne centrée d'une matrice

Soit une matrice de taille n t définie comme la matrice $n \times n$:

$$Cn = In - \ 1nJn \tag{1}$$

Où $I_n$est la matrice d'identité de taille n et $J_n$est une matrice $n \times n$ de tous les 1.

Étant donné un vecteur-colonne, Vde taille n, la propriété de centrage de C_n peut-être exprimé comme [2] :

$$C_n V = V - \left(\frac{1}{n} J_{n,1}^T V\right) J_{n,1} \tag{2}$$

- $\mathrm{J_{n,1}}$: est un vecteur colonne d'uns et$\frac{1}{\mathrm{n}}\mathrm{J_{n,1}^T V}$ est la moyenne des composantes de V.
- $\mathrm{C_n}$ est une matrice :
    - Symétrique positive semi-définie .
    - Idempotence , de sorte que $\mathrm{C_n^k = C_n}$, pour $k = 1,2,$ ….. Une fois que la moyenne a été supprimée, elle est nulle et la supprimer à nouveau n'a aucun effet.
    - Singulier . Les effets de l'application de la transformation $\mathrm{C_n V}$ne peut pas être inversé.
    - La valeur propre 1 de multiplicité $n \ - \ 1$ et la valeur propre 0 de multiplicité 1.
    - Un espace nul de dimension1, le long du vecteur $\mathrm{J_{n,1}}$.
    - Une matrice de projection orthogonale . C'est-à-dire,$\mathrm{C_n V}$est une projection de V sur le sous - espace ($n \ - 1$) dimensionnel orthogonal à l'espace nul$\mathrm{J_{n,1}}$. (C'est le sous-espace de tous les n -vecteurs dont la somme des composants est nulle.)
    - De trace : $\mathrm{n} * (\mathrm{n} - 1)/\mathrm{n} = \mathrm{n} - 1.$

1.2.3 Décomposition des matrices

En mathématique de l'algèbre linéaire, une décomposition matricielle est une factorisation en un produit de matrices. Il existe de nombreuses décompositions, chacune trouve son utilité dans une classe particulière de problèmes. En analyse numérique, différentes décompositions sont utilisées pour implémenter des algorithmes matriciels efficaces.

En raison du contexte de ce projet, nous limitons notre étude aux matrices réelles. De plus, nous limitons ici aux décompositions en valeurs propres (EVD) et en valeurs singulières (SVD).

a Décomposition en valeurs singulières (SVD)

Pour certaines matrices carrées, une décomposition en valeurs propres peut se faire :

$$A = XDX^{-1} \tag{3}$$

Où D est une matrice diagonale de valeurs propres et X est une matrice inversible de vecteurs propres. Quand on fait le produit matrice-vecteur $Ax = (XDX^{-1})x$, on prend x, on l'exprime dans la base donnée par les vecteurs propres $(X^{-1}x)$, on multiplie les éléments de ce vecteur par les valeurs propres D une à une, et on refait le changement de base inverse en multipliant par X.

Si on a un système linéaire $Ax = b$, on peut effectuer les changements de base $\hat{x} = X^{-1}x, \hat{b} = X^{-1}b$ et on obtient le système $D\hat{x} = \hat{b}$. Pour effectuer cette transformation, la matrice A doit être carrée et diagonalisable, ce qui présente une limitation de cette décomposition.

L'idée du SVD est similaire à EVD, seulement que SVD fonctionne avec n'importe quelle matrice A de taille $m \times n$: on factorise A en produit de trois matrices [3] :

$$A = U\Sigma V^* \tag{4}$$

Avec U une matrice $m \times m$ unitaire, V une matrice $n \times n$ unitaire et Σ une matrice $m \times n$ diagonale avec coefficients réels et positifs.

- **Décomposition en valeurs singulières réduite**

Soit S la boule (sphère) unité dans $\mathbb{R}^n$ (ou $\mathbb{C}^n$), et soit une matrice $A\, m \times n$ avec $m \geq n$ à coefficients dans $\mathbb{R}$ (ou $\mathbb{C}$). La matrice A définit une application de $\mathbb{R}^n$dans $\mathbb{R}^m$. On suppose que $rang\ A = n$. L'image de la boule S par A, notée AS est une hyperellipse dans $\mathbb{R}^m$.

On définit les n valeurs singulières de A comme les longueurs σ_i des n semi-axes principaux de AS.

Par convention, on numérote les valeurs singulières par ordre décroissant :

$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$. On définit $n$les vecteurs singuliers à gauche de A ou les n vecteurs de sortie de A, $\{u_1, u_2, \ldots, u_n\}$ orientés dans les directions des semi-axes principaux, où u_i est la direction de du semi-axe de longueur σ_i.

On définit les n vecteurs singuliers à droite de A ou les n vecteurs d'entrée de A, $\{v_1, v_2, \ldots, v_n\}$, $vi \in S$ qui sont les pré-images des semi-axes principaux : $Av_i = \sigma_i u_i$ pour $i = 1, \ldots, n$. Sous forme matricielle :

$$A \begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & | & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix}$$

Ou bien :

$$AV = \hat{U}\hat{\Sigma} \tag{5}$$

La matrice $\hat{\Sigma}$ est une matrice $n \times n$, diagonale avec coefficients positifs et réels. $\hat{U}$ Est une matrice $m \times n$ avec colonnes orthonormales et V est une matrice $n \times n$ avec colonnes orthonormales.

Vest unitaire et on peut réécrire l'équation.

$$A = \hat{U}\hat{\Sigma}V^* \tag{6}$$

Cette factorisation est appelée décomposition en valeurs singulières réduite

$$\begin{bmatrix} & A & \end{bmatrix} = \begin{bmatrix} & \hat{U} & \end{bmatrix} \begin{bmatrix} & \hat{\Sigma} & \end{bmatrix} \begin{bmatrix} & V^* & \end{bmatrix}$$

- **Décomposition en valeurs singulières complète**

$\hat{U}$ Est une matrice $m \times n$ et, sauf si $m = n$, les colonnes de $\hat{U}$ ne forment pas une base de $\mathbb{C}^m$. En ajoutant $m - n$ colonnes orthonormales manquantes à U, on peut en faire une matrice unitaire, que l'on appellera U. Si $\hat{U}$ est remplacée par U dans la factorisation, la matrice $\hat{\Sigma}$ doit être augmentée, en ajoutant $m - n$ lignes de zéros. On obtient alors une décomposition en valeurs singulières complète [4].

Ou encore :

$$A = U\Sigma V^* \tag{7}$$

b Décomposition des valeurs propres d'une matrice (EVD)

Avec une matrice carrée $n \times n$, un nombre λ (complexe) est appelé une valeur propre de A s'il existe un vecteur de colonne non nul n-dimensionnel X tel que :

$$AX = \lambda X, \qquad X \neq 0 \tag{8}$$

Le vecteur AX satisfaisant (59) est appelé un vecteur propre de A correspondant à la valeur propre λ.

Nous décrivons maintenant comment trouver les valeurs propres d'une matrice donnée. Les valeurs propres de A s'avèrent être précisément les racines du polynôme caractéristique de la matrice $p_A(t) \coloneqq \det(A - tI_n)$, où $I_n$est l'identité $n \times n$ matrice :

λEst valeur propre pour $A \Longleftrightarrow p_A(\lambda) = det(A - \lambda I_n) = 0$

Le théorème fondamental de l'algèbre garantit que tout polynôme avec des coefficients réels, tels que p_A, peut être pris en compte dans les facteurs linéaires.

$$p_A(t) = (-1)^n (t - \lambda_1)^{m_1} \dots (t - \lambda_k)^{m_k} \tag{9}$$

Où $\lambda_1, \dots, \lambda_k$sont précisément les valeurs propres distinctes (complexes) de A. L'entier positif $m_j$est appelé la multiplicité algébrique de la valeur propre $\lambda_j , j = 1, \dots, k$. Les valeurs propres non résiduelles, le cas échéant, se présentent en paires conjuguées complexes. L'autre information importante sur chaque valeur propre $\lambda = \lambda_j$ est sa multiplicité géométrique, qui est définie comme $dim\, E_\lambda^A$, la dimension de l'espace propre E_λ^A, ou le nombre maximum de vecteurs propres linéairement indépendants de A correspondant à λ. Un puits-fait connu dans l'algèbre linéaire lit : $multiplicité\ géométrique\ de\ \lambda \leq multiplicité\ algébrique\ de\ \lambda$ [5].

1.3 Probabilité et statistiques

1.3.1 Espérance

En théorie des probabilités, l'espérance mathématique d'une variable aléatoire réelle est, intuitivement, la valeur que l'on s'attend à trouver, en moyenne, si l'on répète un grand nombre de fois la même expérience aléatoire.

Elle se note $E(x)$ et se lit « espérance de X ».

Elle correspond à une moyenne pondérée des valeurs que peut prendre cette variable. Dans le cas où celle-ci prend un nombre fini de valeurs, il s'agit d'une moyenne pondérée par les probabilités d'apparition de chaque valeur. Dans le cas où la variable aléatoire possède une densité de probabilité, l'espérance est la moyenne des valeurs pondérées par cette densité.

De manière plus théorique, l'espérance d'une variable aléatoire est l'intégrale de cette variable selon la mesure de probabilité de l'espace probabilisé de départ [6].

L'espérance E(X) d'une variable aléatoire discrète X est donnée par la formule 10 [7] :

$$E(X) = \sum_i x_i P(x_i) \tag{10}$$

L'espérance $E(X)$ d'une variable aléatoire continu X est donnée par la formule 11 [8] :

$$E(X) = \int x f(x) dx \tag{11}$$

Soit Xune variable aléatoire prenant un nombre fini de valeurs$x_1, x_2, \ldots, x_n$,avec les probabilités respectives $P_1, P_2, \ldots, P_n$.

La somme$P_1 x_1 + P_2 x_2 + \cdots + P_n x_n$peut être interprétée de deux manières :

- **Mathématiquement :** c'est une moyenne pondérée. Plus précisément, c'est le barycentre du système (x_i, P_i) $(i = 1, \ldots, n)$, le « point »$x_i$étant affecté de la « masse » P_i.
- **Heuristiquement :** supposons qu'on répète N fois l'expérience aléatoire attachée à X et soit $N_i$le nombre de fois où X prend la valeur $x_i (i = 1, \ldots, n)$. La moyenne arithmétique des valeurs de X observées au cours des N essais est :

$$\frac{N_1 x_1 + N_2 x_2 + \cdots + N_n x_n}{N} = \frac{N_1}{N} x_1 + \frac{N_2}{N} x_2 + \cdots + \frac{N_n}{N} x_n \tag{12}$$

Qui puisque $\frac{N_i}{N}$ est proche de P_i si N est grand (loi empirique des grands nombres5), est pratiquement égale à $P_1 x_1 + P_2 x_2 + \cdots + P_n x_n$ si N est grand [9].

Espérance d'une fonction de variable aléatoire :

- Cas discret : Si X a une distribution discrète et $Y = r(X)$,alors :

$$E\big(r(x)\big) = E(Y) = \sum_{y} yP(Y = y) = \sum_{x} r(x)P(X = x) \quad (13)$$

- Cas continu : Si X a pour densité fx et Y = r(X), alors [8] :

$$E\big(r(x)\big) = E(Y) = \int yf_Y(y)d_B = \int r(x)f_X(x)d_x \quad (14)$$

1.3.2 Écart-type

En mathématiques, l'écart type est une mesure du degré de dispersion de la valeur médiane d'un échantillon statistique ou d'une distribution de probabilité. Il est défini comme la racine carrée de la variance, ou de manière équivalente, la déviation quadratique moyenne de la moyenne. Il est généralement écrit dans la lettre grecque σ « sigma », après le nom anglais « standard deviation ». Elle est homogène avec la grandeur mesurée.

Les écarts-types se rencontrent dans tous les domaines d'application des probabilités et des statistiques, notamment dans les domaines de l'investigation, de la physique, de la biologie ou de la finance. Ils peuvent généralement synthétiser les résultats numériques d'expériences répétées. En probabilité et en statistique, il est utilisé pour exprimer d'autres concepts importants, tels que le coefficient de corrélation, le coefficient de variation ou la distribution optimale de Naiman.

En statistique, l'écart type est une mesure du degré de variation ou de dispersion d'un ensemble de valeurs. Un écart type faible indique que la valeur a tendance à être proche de la moyenne de l'ensemble (également appelée valeur attendue), tandis qu'un écart type élevé indique que la valeur est répartie sur une plage plus large [10].

L'écart type est une statistique qui mesure le degré de dispersion entre un ensemble de données et sa moyenne. En déterminant l'écart de chaque point de données par rapport à la moyenne, l'écart type est calculé comme la racine carrée de la variance. Si le point de données est plus éloigné de la moyenne, l'écart de l'ensemble de données sera plus grand, donc plus les données sont dispersées, plus l'écart type est grand.

- La formule de l'écart-type :
  - x_{Je} =Valeur du ith point dans l'ensemble de données
  - X =La valeur moyenne de l'ensemble de données
  - n =Le nombre de points de données dans l'ensemble de données

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_{Je} - X)^2}{n-1}} \quad (15)$$

- L'écart type est calculé comme suit :
  - La moyenne est calculée en additionnant tous les points de données et en divisant par le nombre de points de données.
  - La variance de chaque point de données est calculée en soustrayant la valeur moyenne de la valeur du point de données. Ensuite, placez chacune de ces valeurs de résultat au carré et additionnez les résultats. Divisez ensuite le résultat par le nombre de points de données moins un.
  - Le résultat de la racine carrée de la variance est ensuite utilisé pour trouver l'écart-type [11].

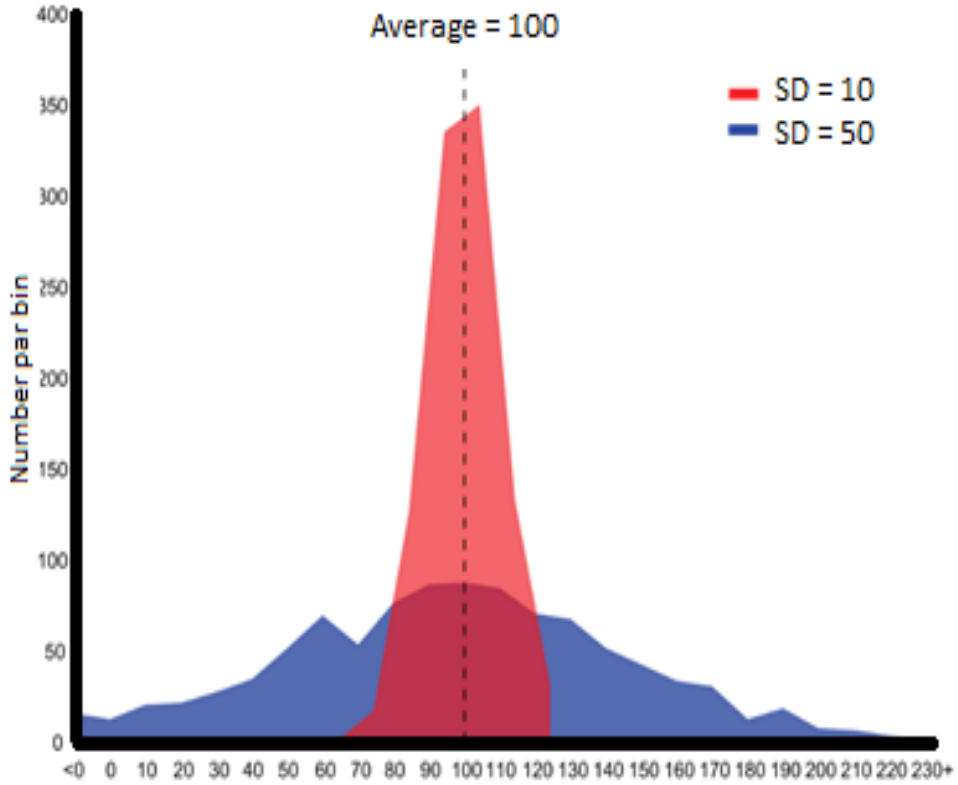

Figure 3-Exemple de deux échantillons ayant la même moyenne mais des écarts-types différents (11).

1.3.3 Variance

En statistique et en théorie des probabilités, la variance est une mesure de la dispersion des valeurs d'un échantillon ou d'une distribution de probabilité. Elle exprime la moyenne des carrés des écarts à la moyenne, aussi égale à la différence entre la moyenne des carrés des valeurs de la variable et le carré de la moyenne, selon le théorème de König-Huygens. Ainsi, plus l'écart à la moyenne est grand plus il est prépondérant dans le calcul total de la variance qui donnerait donc une bonne idée sur la dispersion des valeurs.

La variance est toujours positive, et ne s'annule que si les valeurs sont toutes égales. Sa racine carrée définit l'écart type σ, d'où la notation [12] :

$$\sigma^2 = V = V(X) = Var(X) \tag{16}$$

Pour une variable unique X ayant une distribution $P(x)$avec une moyenne de population *connue*, la variance de population, communément également écrite, est définie comme :

$$\sigma^2 \equiv \prec (X - \mu)^2 \succ \tag{17}$$

Où μ est la moyenne de la population et $\prec X \succ$ désigne la valeur attendue de X. Pour une distribution discrète avec N des valeurs possibles de x_i, la variance de la population est donc :

$$\sigma^2 = \sum_{i=1}^{N} P(x_i)(x_i - \mu)^2 \tag{18}$$

Alors que pour une distribution continue, elle est donnée par :

$$\sigma^2 = \int P(x)(x - \mu)^2 dx \tag{19}$$

La variance est donc égale au deuxième moment central μ^2.

Notez que certains soins est nécessaire pour interpréter σ^2 comme une variance, car le symbole σ est aussi couramment utilisé comme un paramètre lié à, mais pas équivalent à la racine carrée de la variance, par exemple dans la distribution log-normale, distribution de Maxwell, et la distribution de Rayleigh.

Si la distribution sous-jacente n'est pas connue, la variance de l'échantillon peut être calculée comme suit :

$$s_N^2 \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{20}$$

La variance est dérivée en prenant la moyenne des points de données, en soustrayant la moyenne de chaque point de données individuellement, en mettant au carré chacun de ces résultats, puis en prenant une autre moyenne de ces carrés. L'écart type est la racine carrée de la variance.

La variance aide à déterminer la taille de propagation des données par rapport à la valeur moyenne. Au fur et à mesure que la variance augmente, la variation des valeurs de données augmente et il peut y avoir un écart plus important entre une valeur de données et une autre. Si les valeurs de données sont toutes proches les unes des autres, la variance sera plus petite. Cependant, cela est plus difficile à saisir que l'écart type, car les variances représentent un résultat au carré qui peut ne pas être exprimé de manière significative sur le même graphique que l'ensemble de données d'origine [13].

La variance est une mesure de la distribution des points de données à partir de la moyenne, une faible variance indique que les points de données sont généralement similaires et pas très différents de la moyenne, une variance plus élevée indique que les valeurs de données ont une plus grande variance et sont réparties sur une plage plus large que la moyenne, un calculateur de variance est disponible et qui trouve la variance, l'écart type, la taille de l'échantillon n, la moyenne et la somme des carrés Vous pouvez également visualiser le travail organisé du calcul Il suffit d'entrer un ensemble de données avec des valeurs séparées par des espaces, des virgules ou des sauts de ligne Vous pouvez copier et coller vos données à partir d'un document ou d'une feuille de calcul, ce qui peut être observé pour connaître les avantages et les inconvénients du contraste.

La variance est une statistique utilisée pour mesurer l'asymétrie d'une distribution de probabilité L'écart est la tendance des résultats à différer de la valeur attendue L'étude de la variance permet à un individu de mesurer la variance dans une distribution de probabilité Les distributions de probabilité avec des résultats différents auront une différence significative Possible les essais avec des résultats proches les uns des autres auront peu de différence [14].

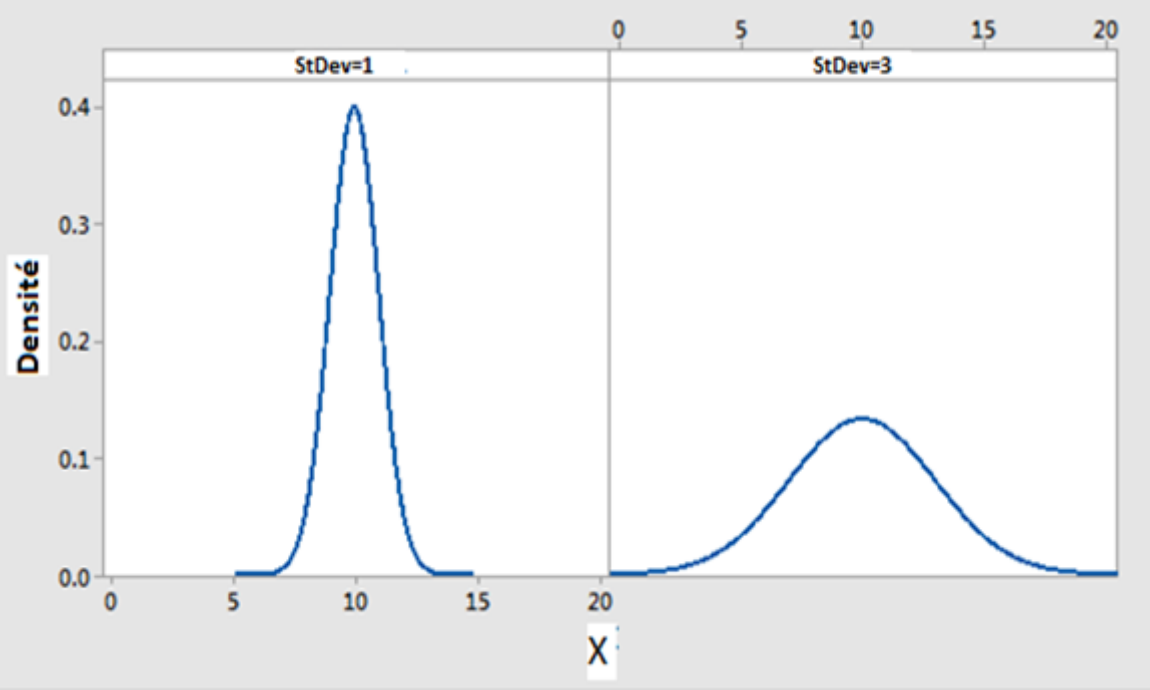

Figure 4-variation dans les échantillons de variabilité faible et élevée (14).

1.3.4 Covariance

En théorie des probabilités et en statistique, la covariance entre deux variables aléatoires est un nombre qui peut quantifier leur écart conjoint par rapport à leurs attentes respectives. Il est également utilisé pour deux séries de données numériques (écarts par rapport à la moyenne). La covariance de deux variables aléatoires indépendantes est nulle, mais l'inverse n'est pas toujours vrai.

La covariance est une extension du concept de variance. La corrélation est la forme normalisée de covariance (la dimension de covariance entre deux variables est le produit de leurs dimensions, et la corrélation est une quantité sans dimension).

Ce concept se généralise naturellement à plusieurs variables (vecteur aléatoire) par la matrice de covariance (ou matrice de variance-covariance) qui, pour un ensemble de p variables aléatoires réelles $X_1, \ldots\ldots, X_p$ est la matrice carrée dont l'élément de la ligne i et de la colonne j est la covariance des variables X_i et X_j. Cette matrice permet de quantifier la variation de chaque variable par rapport à chacune des autres. La forme normalisée de la matrice de covariance est la matrice de corrélation.

Par exemple, la dispersion d'un ensemble de points aléatoires dans un espace à deux dimensions ne peut pas être complètement caractérisée par un seul nombre, ni par la variance dans les directions X et Y seules ; la matrice 2×2 fournit une compréhension de la nature bidimensionnelle des changements [15].

La covariance mesure la relation directionnelle entre les rendements de deux actifs. Une covariance positive signifie que les rendements des actifs se déplacent ensemble alors qu'une covariance négative signifie qu'ils se déplacent inversement. La covariance est calculée en analysant les surprises au retour (écarts-types par rapport au rendement prévu) ou en multipliant la corrélation entre les deux variables par l'écart-types de chaque variable.

La covariance évalue la façon dont les valeurs moyennes de deux variables se déplacent ensemble. Si le rendement des actions Aaugmenté chaque fois que le rendement des actions B augmente et que la même relation est établie lorsque le rendement de chaque action diminue, on dit que ces actions ont une covariance positive. En finance, la covariance est calculée pour aider à diversifier les avoirs en titres [16].

La *covariance* est définie :

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{21}$$

- La covariance peut prendre des valeurs positives, négatives ou nulles.
- Quand $x_i = y_i$, pour tout $i = 1, . , n$, la covariance est égale à la variance.

La covariance peut également s'écrire [17] :

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} \tag{22}$$

a Covariance d'une Matrice

La matrice de covariance est une matrice semi-définie positive, qui peut être diagonalisée et l'étude des valeurs propres et des vecteurs propres permet d'utiliser des bases orthogonales pour caractériser la distribution : cette méthode fait l'objet d'une analyse en composantes principales et peut être considéré comme une sorte de compression de l'information.

- La matrice de covariance d'un vecteur de $p$variables aléatoires $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$dont chacune possède une variance, est la matrice carrée dont le terme générique est donné par :

$$a_{ij} = Cov\ (X_i, X_j) \tag{23}$$

- La matrice de covariance, notée parfois Σ, est définie par :

$$Var(\vec{X})\ = E[(\vec{X} -\ E(\vec{X}))(\vec{X} - E(\vec{X}))^T] \tag{24}$$

- En développant les termes :

$$Var(\vec{\mathrm{X}}) = \begin{pmatrix} \mathrm{Var(X_1)} & \mathrm{Cov(X_1, X_2)} & \cdots & \mathrm{Cov(X_1, X_P)} \\ \mathrm{Cov(X_2, X_1)} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov(X_P, X_1)} & \cdots & \cdots & \mathrm{Var(X_P)} \end{pmatrix} = \begin{pmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} & \cdots & \sigma_{x_1 x_p} \\ \sigma_{x_2 x_1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_p x_1} & \cdots & \cdots & \sigma_{x_p}^2 \end{pmatrix} \tag{25}$$

Propriétés de la matrice de covariance :

- La matrice de covariance est symétrique, ses éléments diagonaux sont les variances et les éléments extra-diagonaux sont les covariances des couples de variables.
- La matrice de covariance est semi-définie positive (ses valeurs propres sont positives ou nulles). Elle est définie positive (valeurs propres strictement positives) s'il n'existe aucune relation affine presque sûre entre les composantes du vecteur aléatoire.
- Soit une application linéaire F de $M_{n,m}(R)$ de matrice M.
- Soit $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ un vecteur aléatoire de matrice de covariance C de $M_n(R)$. Alors le vecteur aléatoire $F(X)$ a pour matrice de covariance $M\ C\ M^T$.
- L'inverse de la matrice de covariance est parfois désignée « matrice de précision » [15].

b Covariance de deux variables

La covariance de deux variables aléatoires réelles $X \; et \; Y$ ayant chacune une variance, notée $Cov(X, Y)$ ou parfois σ_{XY}, est la valeur :

$$Cov(X, Y) \equiv E[(X - E[X])(Y - E[Y])] \tag{26}$$

Où E désigne l'espérance mathématique. La variance de X est donc $Var(X) = Cov(X, X)$.

Intuitivement, la covariance caractérise les variations simultanées de deux variables aléatoires : elle sera positive lorsque les écarts entre les variables et leurs moyennes ont tendance à être de même signe, négative dans le cas contraire.

Conformément à l'expression de sa définition, la dimension de la covariance est le produit des dimensions des variables. En revanche, la corrélation, qui s'exprime à l'aide de la variance et de la covariance, prend ses valeurs dans $[-1, 1]$ et reste adimensionnelle.

Deux variables aléatoires dont la covariance est nulle sont dites non corrélées : leur corrélation est également nulle.

Pour deux variables aléatoires discrètes $X \; et \; Y$ prenant respectivement leurs valeurs dans deux ensembles finis $\{x_i | 1 \leq i \leq n\}$et $\{y_j | 1 \leq j \leq m\}$on a :

$$Cov(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_i P(X = x_i et Y = y_i) - E[X]E[Y] \tag{27}$$

Tandis que :

$$\sigma_X^2 = \sum_{i=1}^{n} x_i^2 P(X = x_i) - E[X]^2$$

Et (28)

$$\sigma_Y^2 = \sum_{j=1}^{m} y_j^2 P(Y = y_i) - E[Y]^2$$

1.3.5 Corrélations variables

En probabilités et en statistique, la corrélation entre plusieurs variables aléatoires ou statistiques est une notion de liaison qui contredit leur indépendance.

Cette corrélation est souvent simplifiée comme une corrélation linéaire entre des variables quantitatives, c'est-à-dire que la relation affine obtenue par régression linéaire ajuste une variable par rapport à une autre. Pour cela, on calcule le coefficient de corrélation linéaire 1, qui est le quotient du produit de la covariance et de l'écart type. Son signe indique si une valeur supérieure "en moyenne" correspond à une autre valeur supérieure ou inférieure. La valeur absolue du coefficient est toujours comprise entre 0 et 1. Il ne mesure pas la force du lien, mais mesure plutôt l'avantage de la relation affine sur les changements internes de la variable. Un coefficient nul ne signifie pas indépendance, car d'autres types de corrélation sont également possibles.

D'autres indicateurs nous permettent de calculer le coefficient de corrélation des variables ordinales.

Le fait que deux variables soient « fortement corrélées » ne prouve pas qu'il existe une relation causale entre une variable et l'autre. Le contre-exemple le plus typique est qu'ils sont en fait liés par une relation de cause à effet commune. Cette confusion est appelée "Cum hoc ergo propter hoc" [18].

a Droite de régression

Calculer le coefficient de corrélation entre deux variables numériques revient à essayer d'utiliser une ligne droite pour résumer la relation entre les variables. C'est ce qu'on appelle l'ajustement linéaire.

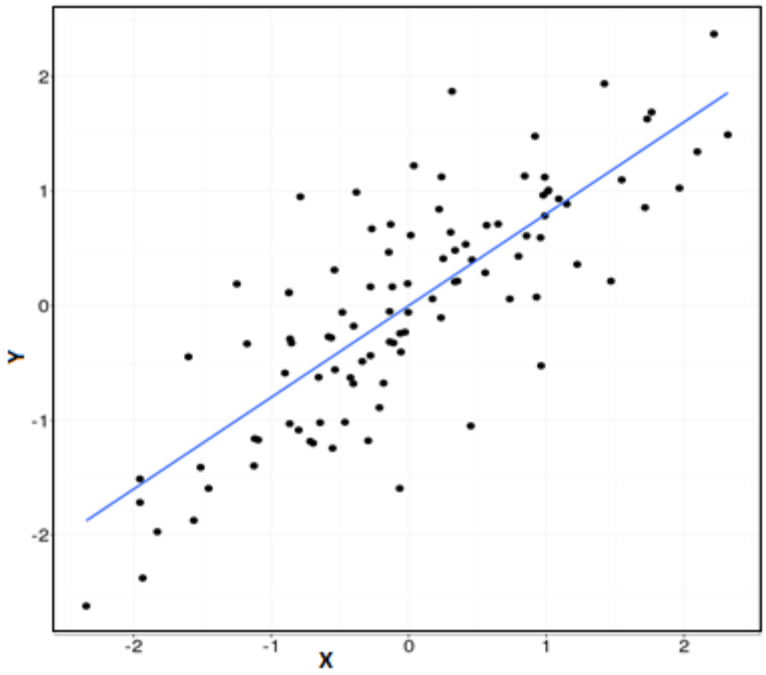

Figure 5-Droite de régression [18].

Comment calculer les caractéristiques de cette droite ? En veillant à utiliser des lignes droites pour représenter les liens entre les variables, les erreurs que nous commettons sont aussi petites que possible. Le critère formel le plus couramment utilisé, mais pas le seul possible, est de minimiser la somme de toutes les erreurs carrées réelles. C'est ce qu'on appelle l'ajustement par les moindres carrés ordinaires. La ligne produite par cet ajustement est appelée ligne de régression. Cette droite indique que plus la qualité globale du lien entre les variables est bonne, meilleur est le coefficient de corrélation linéaire de la corrélation. Il existe une équivalence formelle entre ces deux concepts.

b Coefficient de Corrélation

Le coefficient de corrélation entre deux variables aléatoires réelles X et Y ayant chacune une variance notée $Cor(X,Y)$ou parfois ρxy ou $r_p$ou simplement r , est défini par :

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{29}$$

Ou $Cov(X,Y)$désigne la covariance des variables X et Y,σ_X et $\sigma_Y$désignent leurs écarts types.

De manière équivalente :

$$r = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} \quad (30)$$

c *Matrice de Corrélation*

La matrice de corrélation d'un vecteur de p variables aléatoires $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ dont chacune possède une variance, est la matrice carrée dont le terme générique est donné par :

$$r_{i,j} = Cor(X_i, X_j) \quad (31)$$

Les termes diagonaux de cette matrice sont égaux à 1, elle est symétrique, semi-définie positive et ses valeurs propres sont ou nulles [18].

Exemple représentez graphiquement pour trouver des corrélations :

Les nuages de points sont un excellent moyen de vérifier rapidement les relations entre les paires de données continues . Le nuage de points ci-dessous affiche la taille et le poids des pré-adolescentes. Chaque point sur le graphique représente une fille individuelle et sa combinaison de taille et de poids. Ces données sont des données réelles que j'ai collectées lors d'une expérience [19].

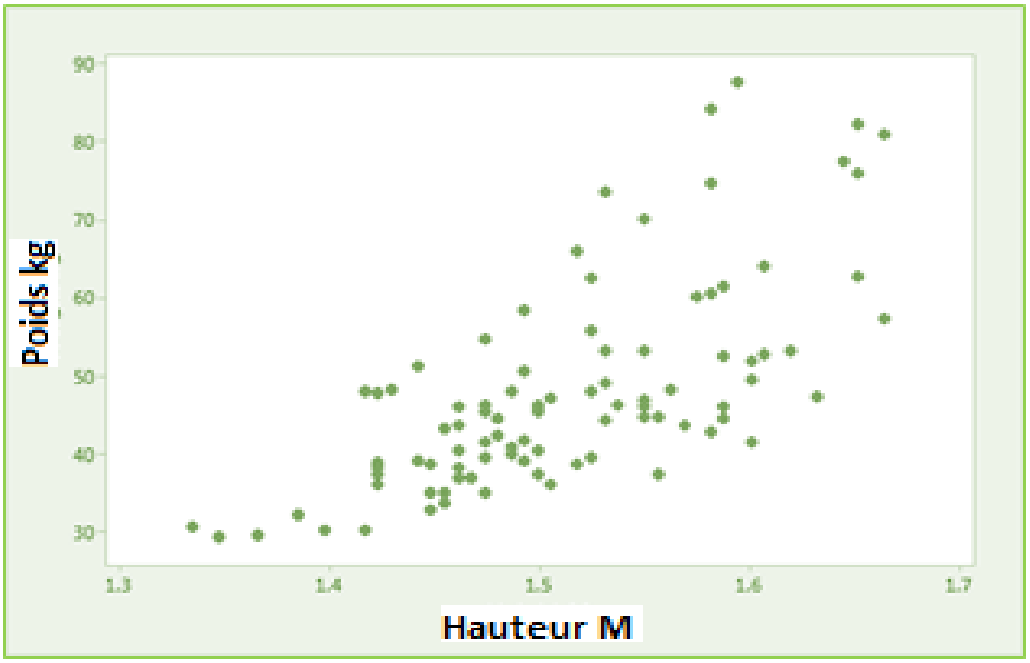

Figure 6-Représentez graphiquement vos données pour trouver des corrélations (19).

1.4 Réduction des dimensions

La réduction de dimensionnalité est la conversion de données d'un espace de grande dimension vers un espace de faible dimension, de sorte que la représentation de faible dimension conserve certains attributs importants des données d'origine et se rapproche idéalement de ses dimensions intrinsèques. Pour de nombreuses raisons, travailler dans un espace de grande dimension peut être indésirable en raison du désastre de la dimensionnalité, les données d'origine sont généralement rares et l'analyse des données est souvent difficile à gérer sur le plan informatique. La réduction de la dimensionnalité est courante dans les domaines qui traitent un grand nombre d'observations et/ou un grand nombre de variables, tels que le traitement du signal, la reconnaissance vocale, la neuro-informatique et la bio-informatique [20].

Ces méthodes sont généralement divisées en méthodes linéaires et méthodes non linéaires. La méthode peut également être divisée en sélection et extraction de caractéristiques. La réduction de la dimensionnalité peut être utilisée pour la réduction du bruit, la visualisation des données, l'analyse de cluster ou comme étape intermédiaire pour faciliter une analyse plus approfondie [21].

Il existe deux méthodes principales de réduction de dimensionnalité.

- La première est appelée projection linéaire, qui consiste à projeter linéairement des données d'un espace de grande dimension vers un espace de faible dimension. Cela inclut des techniques telles que l'analyse en composantes principales, la décomposition en valeurs singulières et la projection aléatoire.
- La deuxième méthode est appelée apprentissage multiple, également appelée réduction de dimensionnalité non linéaire. Cela implique des techniques telles que la cartographie d'équivalence, la mise à l'échelle multidimensionnelle (MDS) et l'analyse de composants indépendants.

Dans ce projet, je me concentrerai sur la régression des moindres carrés partiels, qui est une technique de réduction de dimensionnalité linéaire [22].

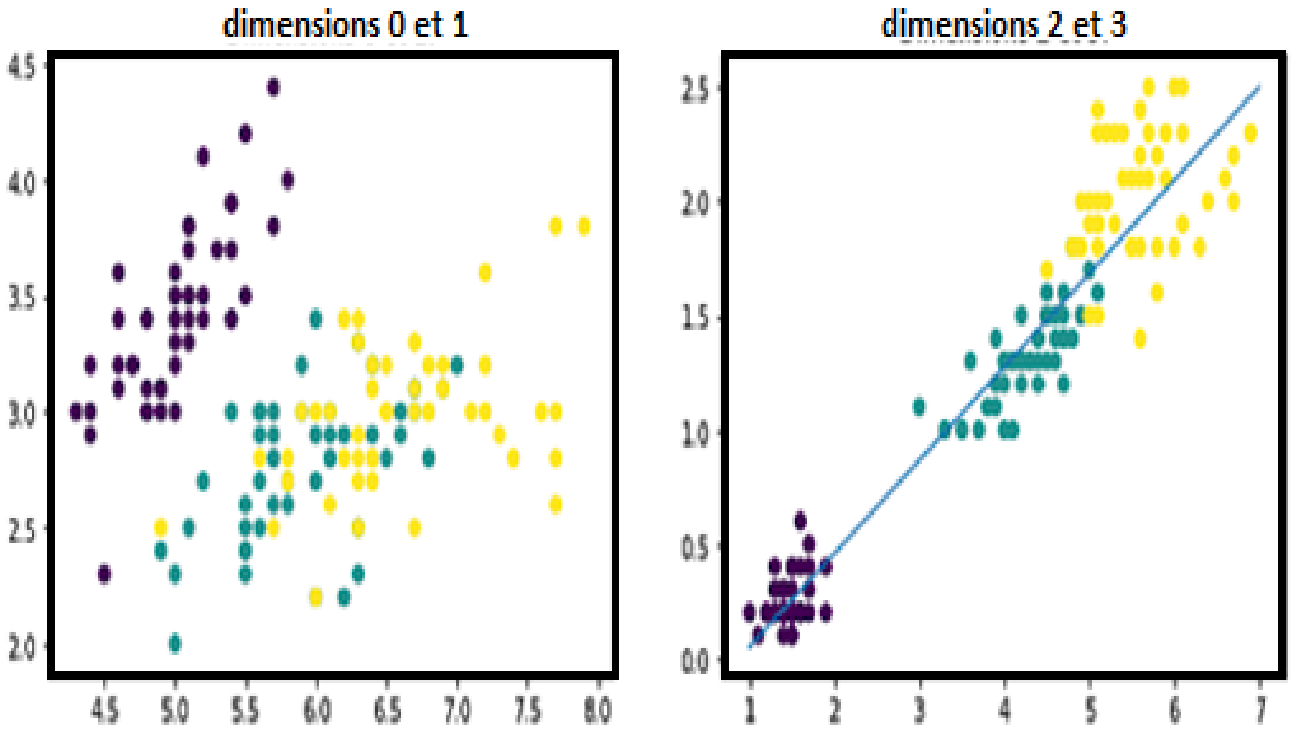

Figure 7-Exemple dans la réduction de dimensionnalité [22].

1.5 Prédiction (Estimation)

En général, la prédiction est le processus de détermination de l'ampleur des variables statistiques à un moment futur. Dans les contextes statistiques, le mot peut également avoir des significations légèrement différentes.

Par exemple, dans une équation de régression exprimant une variable dépendante y en termes de x dépendants, la valeur donnée pour y par des valeurs spécifiées de x est appelée valeur « prédite » même si aucun élément temporel n'est impliqué [23].

Pour estimer θ on ne dispose que des données $x_1, \ldots, x_n$, donc une estimation de θ sera une fonction de ces observations.

Une statistique t est une fonction des observations $x_1, \ldots, x_n$:

$$t : \mathbb{R}^n \to \mathbb{R}^m$$
$$(x_1, \ldots, x_n) \rightarrow t\,(x_1, \ldots, x_n) \tag{32}$$

Par exemple :

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\,, x_1^*, (x_3, x_4 + x_4, 2lnx_6) \tag{33}$$

Sont des statistiques.

Puisque les observations $x_1, \ldots, x_n$ sont des réalisations des variables aléatoires$X_1, \ldots, X_n$, la quantité calculable à partir des observations $t(x_1, \ldots, x_n)$ est une réalisation de la variable aléatoire $t(X_1, \ldots, X_n)$. Et on retrouve par exemple le fait que :

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{34}$$

Une réalisation :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{35}$$

Pour simplifier les écritures, on note souvent :

$$t_n = t(x_1, \ldots, x_n) \; et T_n = t(x_1, \ldots, x_n) \tag{36}$$

Par abus, on donne le même nom de statistique aux deux quantités, mais dans une perspective d'estimation, on va nommer différemment t_n et T_n.

Un estimateur d'une grandeur θ est une statistique T_n a valeurs dans l'ensemble des valeurs possibles de θ. Une estimation de θ est une réalisation t_n de l'estimateur T_n.

Un estimateur est donc une variable aléatoire, alors qu'une estimation est une valeur déterministe.

Dans l'exemple des ampoules, l'estimateur de λ est $1/\bar{X}_n$et l'estimation de λ est 0.012 [24].

Chapitre 2 Régression des Moindres Carrés (PLS)

2.1 Introduction

La recherche scientifique et technique implique généralement l'utilisation de variables (facteurs) contrôlables et/ou facilement mesurables pour expliquer, ajuster ou prédire le comportement d'autres variables (réponses). Lorsque le nombre de facteurs est petit, n'est pas significativement redondant (colinéaire) et a une relation bien comprise avec la réponse, la régression linéaire multiple (MLR) peut être un bon moyen de convertir les données en informations. Cependant, si l'une de ces trois conditions se détériore, la MLR peut être invalide ou inappropriée. Dans ces applications dites soft science, les chercheurs sont confrontés à de nombreuses variables et relations mal comprises, le but étant de construire un bon modèle prédictif.

Estimer la teneur en différents composés de l'échantillon chimique. Dans ce cas, les facteurs sont les mesures qui composent le spectre ; ils peuvent être des centaines, mais ils sont susceptibles d'être très colinéaires. La réponse est le nombre de composants que le chercheur espère prédire dans les futurs échantillons.

Les moindres carrés partiels (PLS) sont une méthode de construction d'un modèle prédictif lorsque les facteurs sont nombreux et fortement colinéaires. Notez que l'accent est mis sur la prédiction de la réponse, pas nécessairement sur la compréhension des relations sous-jacentes entre les variables. Par exemple, le PLS n'est généralement pas adapté pour éliminer les facteurs qui ont un impact négligeable sur la réponse. Cependant, lorsque la prédiction est l'objectif et qu'il n'est pas nécessaire de limiter le nombre de facteurs de mesure, le PLS peut être un outil utile.

Le PLS a été développé par « Herman Wold » en tant que technique économétrique dans les années 1960, mais certains de ses plus ardents partisans (dont le fils de Wold Svante) sont des ingénieurs chimistes. En plus de l'étalonnage spectral discuté ci-dessus, PLS est également appliqué à la surveillance et au contrôle des processus industriels, un grand processus peut facilement avoir des centaines de variables contrôlables et des dizaines de sorties.

La section suivante décrit brièvement le fonctionnement du PLS et le relie à d'autres techniques multivariées telles que la régression en composantes principales et l'analyse de redondance maximale. Un exemple chimio-métrique étendu est fourni qui montre comment évaluer les modèles PLS et comment interpréter leurs composants. La dernière partie traite des alternatives et des extensions à PLS. Les annexes présentent la procédure expérimentale du PLS pour l'exécution des moindres carrés partiels et des techniques de modélisation connexes.

2.2 Objectif

Le but de la régression du PLS est de prédire Y à partir de X et de décrire leur structure commune. Lorsque Y est un vecteur et que X est en rang plein, cet objectif pourrait être atteint en utilisant une régression multiple ordinaire. Lorsque le nombre de prédicteurs est élevé par rapport au nombre d'observations, X est susceptible d'être singulier et l'approche de régression n'est plus possible (c.-à-d. en raison de la multi-colinéarité). Plusieurs approches ont été développées pour faire face à ce problème. Une approche consiste à éliminer certains prédicteurs (e.g.., en utilisant des méthodes par étapes) une autre, appelée régression des composantes principales, consiste à effectuer une analyse des composantes principales (PCA) de la matrice X, puis à utiliser les composantes principales (c.-à-d. les vecteurs propres). De X comme régression sur Y. Techniquement dans PCA, X est décomposé en utilisant sa valeur singulière décomposition comme :

$$X = S\Delta V^T \tag{37}$$

Avec :

$$S^T S = V^T V = I \tag{38}$$

(Ce sont les matrices des vecteurs singuliers gauche et droite) et Δ étant une matrice diagonale avec les valeurs singulières comme éléments diagonales. Les vecteurs singuliers sont ordonnés selon leurs valeurs singulières correspondantes qui correspondent à la racine carrée de la variance de X expliquée par chaque vecteur singulier. Les vecteurs singuliers de gauche (c'est-à-dire les colonnes de S) sont ensuite utilisés pour prédire Y en utilisant la régression standard parce que l'orthogonalité des vecteurs singuliers élimine le problème de multi-colinéarité. Mais, le problème de choisir un sous-ensemble optimal de prédicteurs reste. Une stratégie possible est de garder seulement quelques-uns des premiers composants. Mais ces composantes sont choisies pour

expliquer X plutôt que Y, et rien ne garantit donc que les principales composantes, qui « expliquent » X, sont pertinentes.

En revanche, la régression PLS trouve des composants de X qui sont également pertinents pour Y. Plus précisément, la régression PLS recherche un ensemble de composants (appelés vecteurs latents) qui effectue une décomposition simultanée de X et Y avec la contrainte que ces composants expliquent autant que possible de la covariance entre X et Y. Cette étape généralise PCA. Elle est suivie d'une étape de régression où la décomposition de X est utilisée pour prédire Y [25].

2.3 Principe

La régression PLS décompose X et Y comme un produit d'un ensemble commun de facteurs orthogonaux et un ensemble de charges spécifiques. Ainsi, les variables indépendantes sont décomposées comme $X = TP^T$ avec $T^TT = I$ avec I étant la matrice d'identité (certaines variations de la technique n'exigent pas que T ait des normes unitaires). Par analogie avec PCA, T est appelé la matrice de score, et P la matrice de chargement (dans la régression PLS les charges ne sont pas orthogonales). De même, Y est estimé comme $\hat{Y} = TBC^T$ où B est une matrice diagonale avec les « poids de régression » comme éléments diagonales et C est la « matrice de poids » des variables dépendantes. Les colonnes de T sont les vecteurs latents. Lorsque leur nombre est égal au rang de X, ils effectuent une décomposition exacte de X. Notez cependant qu'ils n'estiment que Y. (c'est-à-dire qu'en général $\hat{Y}$ n'est pas égal à Y) [26].

2.4 Méthodes Statistiques

Une méthode d'analyse de données est nécessaire pour la mise au point. Elle permet en effet, de rechercher une relation entre une activité et une ou plusieurs variables quantitatives. Plusieurs approches sont envisageables, il s'agit alors de choisir la plus adaptée et celle permettant au mieux caractériser le système pour obtenir un modèle fiable. Dans l'ensemble de notre étude, nous avons principalement utilisé comme techniques pour l'analyse des données, la régression linéaire simple et multiple (SLR, MLR), la régression des moindres carrés partiels (PLS) et l'analyse en composantes principales (PCA).

2.4.1 Analyse en composantes principales (ACP)

(PCA : Principal Composent Analysais) est une méthode qui consiste à transformer des variables corrélées (liées entre elles) en nouvelles variables non corrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet de réduire le nombre de variables et de rendre l'information moins redondante.

Etape de la PCA :

- Centrer et réduire les variables.
  - Center : retrancher la moyenne$(X - \bar{X})$. Cette opération a pour but de donner le même poids à toutes les variables.
  - Réduire : diviser la variable par l'écart type $(X - \bar{X})/\bar{V}$ dans le but d'analyser des variables (données) sans unité.
- Matrice de corrélation

$$\begin{pmatrix} 1 & R_{X1\,X2} & R_{X1\,X3} \\ R_{X2\,X1} & 1 & R_{X2\,X3} \\ R_{X3\,X1} & R_{X3\,X2} & 1 \end{pmatrix} \quad (39)$$

- Diagonaliser la matrice de corrélation

  Cette étape va nous permettre d'avoir les valeurs et les vecteurs propres. A partir des vecteurs propres, on construit les composantes.

- Construction des composantes principales.

Composantes	Valeurs propres	Vecteurs propres	Proportion
1	λ_1	$V_1\begin{pmatrix} a_1 \\ b_1 \end{pmatrix}$	%
2	λ_2	$V_2\begin{pmatrix} a_2 \\ b_2 \end{pmatrix}$	%

$$PC_1 = a_1X_1 + b_1X_2 et PC_1 = a_2X_1 + b_2X_2 \quad (40)$$

La première composante principale traduit la plus grande part de la variance globale du système, les composantes successives n'ayant pour but que d'expliquer la variance résiduelle, non expliquée par les composantes qui les précèdent.

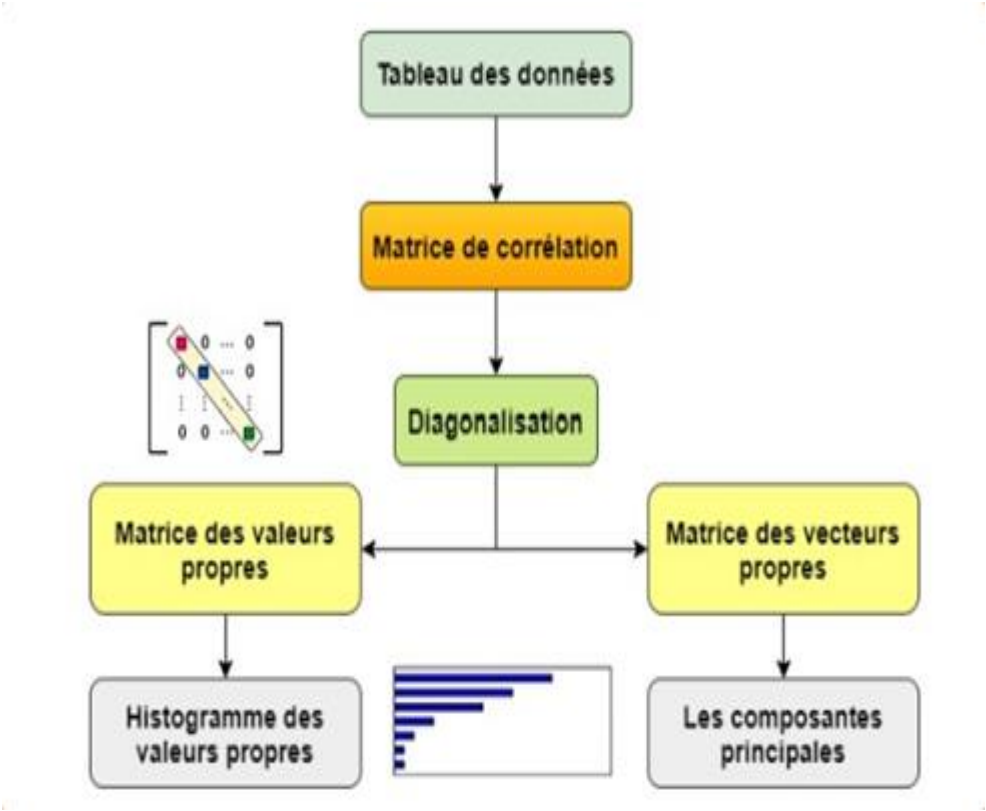

Figure 8-Etapes de la méthode PCA (27).

Des représentations à 2 dimensions suivant les composantes PC_1 et PC_2peuvent être utilisées, ces deux composantes étant celles qui caractérisent la plus grande part de la variance dans le système. La matrice des coordonnées nous permet d'analyser la dispersion des individus dans le nouvel espace défini. Ainsi, deux échantillons proches graphiquement portent une information très similaire.

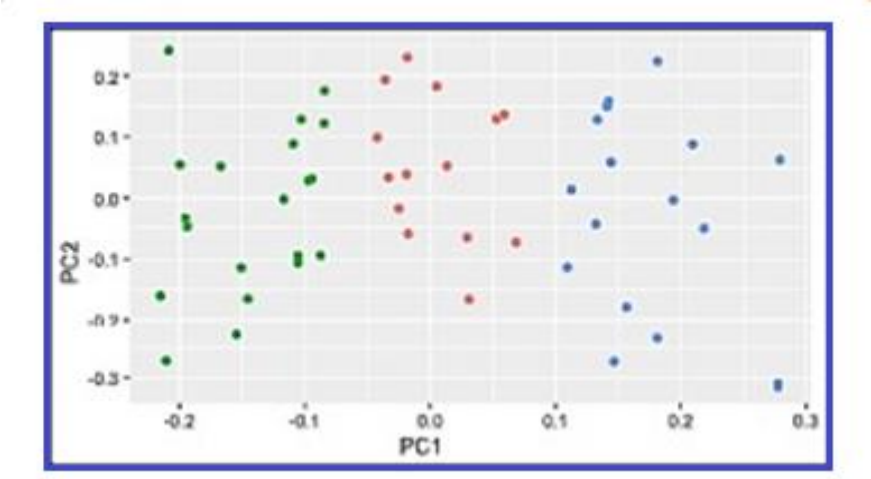

Figure 9-Matrice des coordonnées dans le plan des 2 premières composantes principales d'une analyse PCA [27].

2.4.2 Méthodes de régression linéaire (SLR, MLR)

Sont les plus utilisées, elles permettent de mettre en évidence une relation linéaire entre une réponse ou variable à expliquer (dépendante), notée Y, et une ou plusieurs variables explicatives (indépendantes) notées X.

- *Régression linéaire simple (SLR) :* est une méthode de régression permettant de relier linéairement une variable dépendante Y avec une variable indépendante X. La relation entre ces deux variables s'écrit de la manière suivante :

$$Y = a_0 + aX \tag{41}$$

Cependant, ce modèle est une forme simplifiée, car en réalité, il est perturbé par un terme d'erreur (résidu) noté ε que l'on doit introduire.

La relation devient alors :

$$Y = a_0 + aX + \varepsilon \tag{42}$$

L'intercepte a_0 et a sont les coefficients de régression, constantes inconnues qu'on cherche à estimer et ε est le terme d'erreur.

Pour déterminer les paramètres a et a_0, n observations de la variable X, notées $xi(i = 1, \ldots, n)$ et n valeurs de la variable Y notées yi sont alors mesurées, c'est-à-dire une collecte de n couple de données $(xi\ ;\ yi)$. Cela se traduit par l'équation suivante :

$$Yi = a_0 + aXi + \varepsilon i\ (i = 1, .., n) \tag{43}$$

- *Critère des Moindres Carrés* :

Pour trouver la meilleure droite, il faut chercher les meilleures valeurs des coefficients de régression a eta_0.

$$\widehat{Y}_i = \hat{a}_0 + \hat{a}X_i \tag{44}$$

$\hat{a}_0 et \hat{a}$ Sont des estimateurs de $a0\ et\ a$

Les écarts ou les différences entre les valeurs $\widehat{Y}i$ estimées (obtenues à partir de l'équation de régression) et Yi observées (expérimentales) sont appelés les moindres (résidus).

$$\hat{\varepsilon}_i = Yi - \widehat{Y}i \tag{45}$$

Si l'on prend la somme des écarts (moindres), ces derniers se compensent et la somme totale peut être nulle, ce qui ne reflète pas la réalité.

Pour éviter ce problème de compensation, il faut prendre le carrée des écarts (moindres) d'où l'appellation moindres carrées.

La somme des carrés des écarts (SCE) est donnée par :

$$S(\hat{a}, \hat{a}_0) = \sum_{i=1}^{n} \hat{\varepsilon}i^2 = \sum_{i=1}^{n} (\mathrm{Yi} - \hat{a}_0 - \hat{a}X_i)^2 \quad (46)$$

Le principe de la méthode des moindres carrées consiste à minimiser la quantité S c'est-à-dire la somme des moindres carrés. En effet cette fonction est minimale lorsque ses dérivées par rapport à $a_0$et a s'annulent.

$$\frac{\partial S}{\partial \hat{a}_0} = -2\sum_{i=1}^{n} (\mathrm{Yi} - \hat{a}_0 - \hat{a}\mathrm{Xi})^2 = 0 \quad (47)$$

$$\frac{\partial S}{\partial \hat{a}} = -2\sum_{i=1}^{n} \mathrm{X}_i(\mathrm{Yi} - \hat{a}_0 - \hat{a}\mathrm{Xi})^2 = 0 \quad (48)$$

L'équation 47 donne :

$$\sum_{i=1}^{n} \mathrm{Yi} - n\hat{a}_0 - \hat{a}\sum_{i=1}^{n} \mathrm{Xi} = 0 \quad (49)$$

En utilisant la formule de la moyenne $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Xi}$ et en divisant par$n$, on obtient :

$$\hat{a}_0 = \bar{Y} - \hat{a}\bar{X} \quad (50)$$

L'équation 48 donne :

$$\sum_{i=1}^{n} \mathrm{X}_i\mathrm{Yi} - \hat{a}_0\sum_{i=1}^{n} \mathrm{Xi} - \hat{a}\sum_{i=1}^{n} \mathrm{Xi}^2 = 0 \quad (51)$$

En replaçant $\hat{a}_0$ par sa formule obtenue en équation 50, on obtient :

$$\sum_{i=1}^{n} \mathrm{X}_i\mathrm{Yi} - (\bar{Y} - \hat{a}\bar{X})\sum_{i=1}^{n} \mathrm{Xi} - \hat{a}\sum_{i=1}^{n} \mathrm{Xi}^2 = 0 \quad (52)$$

A partir d'équation 52, on peut tirer l'expression du deuxième coefficient de régression $\hat{a}$.

$$\hat{a} = \frac{\sum XiYi - \sum Xi\bar{Y}}{\sum Xi^2 - \sum Xi\,\bar{X}} = \frac{\sum 氟i(Yi - \bar{Y})}{\sum Xi(Xi - \bar{X})} = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sum (Xi - \bar{X})(Xi - \bar{X})} \quad (53)$$

$$\hat{a} = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sum (Xi - \bar{X})^2} \quad (54)$$

$$\hat{a}_0 = \bar{Y} - \left(\frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sum (Xi - \bar{X})^2} \right) \bar{X} \quad (55)$$

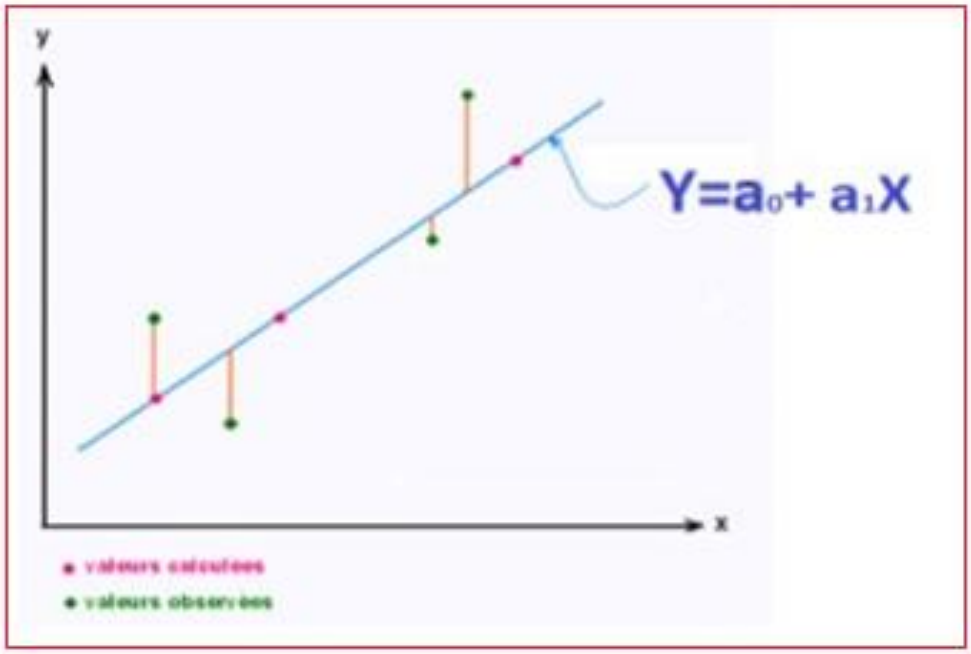

Figure 10-Représentation graphique de la régression linéaire simple (27)

- *Régression linéaire Multiple (MLR) :* est une méthode de régression permettant de relier linéairement une réponse ou variable à expliquer (dépendante), notée Y_i, et plusieurs variables explicatives (indépendantes) notées X_i. Cela se traduit par l'équation suivante :

$$Yi = a_0 + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_p X_{ip} + \varepsilon_i i = 1, \ldots, n \quad (56)$$

Dans notre étude, Y_i et $X_i$représentent les activités observées et les descripteurs calculés respectivement.

Les n échantillons (observations) des variables dépendantes et indépendantes sont connues. Il s'agit donc de considérer un système d'équations qui peut être donné sous la forme matricielle suivante :

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \cdots \\ Y_p \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ 1 & X_{31} & X_{32} & \ldots & X_{3p} \\ \cdots & \cdots & \ldots & \ldots & \ldots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \ldots \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \ldots \\ \varepsilon_p \end{pmatrix} \quad (57)$$

Dans le cas d'un modèle à p variables, le critère des moindres carres s'écrit :

$$S(a_0, \ldots, a_p) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 1 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a_0 - a_1 X_{i1} - \cdots - a_p X_{ip})^2 \quad (58)$$

La valeur prédite de la variable dépendante Y (estimée par le modèle de régression) s'écrit :

$$\hat{Y} = X\hat{a} = Xa \quad (59)$$

Les valeurs des a qui minimisent ce critère seront les solutions a_0, a_1, a_p du système linéaire de $(p+1)$ équations à $(p+1)$ inconnues obtenues comme suit :

$$a = (X'X)^{-1} X'Y \quad (60)$$

Avec (27):

- X' : la matrice transposée de la matrice des variables explicatives X.
- $(X'X)^{-1}$: la matrice inverse de la matrice $(X'X)$.
- Y : vecteur des valeurs de la variable à expliquer.

2.4.3 Méthode des moindres carrés partiels (PLS)

Est une généralisation et combinaison de la régression linéaire multiple et de l'analyse en composantes principales. Elle peut être utilisée lorsque le nombre de descripteurs est élevé et que ceux-ci sont fortement corrélés.

Le principe de La régression PLS consiste à réduire le nombre de prédicteurs à un nombre plus petit de composantes non corrélées et qui effectue la régression par les moindres carrées sur ces composantes plutôt que sur les données initiales.

a Etapes de la régression PLS

- *Etape 1 :* Construction de la première composante $t_1(h = 1)$
  - Maximiser la corrélation entre la première composante t_1 et la variable à expliquer Y.
  - Maximiser la variance de la première composante t_1 afin qu'elle représente au mieux toutes les variables explicatives.

$$t_1 = WX = W_{11}X_1 + W_{12}X_2 + W_{13}X_3 + \dots + W_{1p}X_p \quad (61)$$

W : poids de chaque variable explicative X dans la première composante t_1, il peut être obtenu comme suit :

$$W_{1j} = \frac{Cov(X_j, Y)}{\sqrt{\sum_{k=1}^{p} Cov^2(X_k, Y)}} \quad (62)$$

L'équation de la régression simple de Y sur t_1 est la suivante :

$$Y = C_1 t_1 + Y^{[1]} \quad (63)$$

Avec :

- C_1: coefficient de régression de Y sur la première composante t_1.
- $Y^{[1]}$: résidu nonexpliqué par la première composante t_1.

- *Etape 2 :* Construction des composantes suivantes $h = 2$ (deuxième composante t_2)
  - Les résidus de la 1ére composante vont servir comme données pour construire la deuxième composante t_2, en faisant p régressions simples des variables $X_1 \dots X_p$ sur t_1.
  - De la même manière, la deuxième composante sera construite en utilisant les résidus des X et de Y provenant de la première composante.

$$W_{2j} = \frac{Cov(X_j^{[1]}, Y^{[1]})}{\sqrt{\sum_{k=1}^{p} Cov^2(X_k^{[1]}, Y^{[1]})}} \quad (64)$$

$$Y = C_1 t_1 + C_2 t_2 + Y^{[2]} \quad (65)$$

Avec (27):

- $Y^{[2]}$: le résidu de Y non expliqué par la deuxième composante.
- $C_1 et C_2$: sont les coefficients de régression de Y sur t_1 et sur t_2 respectivement.

La procédure est répétée jusqu'à obtenir $h = A$ composantes= ?

. *Etape 3 :* Choix du nombre de composantes

Le problème qui se pose après avoir construit un modèle est de connaître son aptitude à prédire les réponses de nouvelles variables d'entrée $[X]$. Pour le cas de PLS, ceci consiste surtout à évaluer le nombre optimal de composantes à inclure dans le modèle.

En effet, si beaucoup de composantes sont incluses, il peut y avoir un phénomène de surévaluation (over-fitting) : les composantes sans importance peuvent fausser les prédictions.

Par contre, si trop peu de composantes sont incluses, on risque d'avoir peu d'information pour expliquer le Y.

. *Etape 4 :* Critère de choix de nombre de composantes

Pour déterminer à quel moment le nombre de composantes est suffisant pour traduire la variance du système, une démarche de validation interne est réalisée et la composante est considérée utile si elle contribue de manière significative à améliorer la robustesse de l'analyse.

Critère de « validation croisée » :

1. Calcul de R^2_{cv} après l'ajout de chaque composante.
   - Si $R^2_{cv}(h) > R^2_{cv}(h-1)$, cela signifie que la nouvelle composante ajoutée a un effet sur l'explication de Y.
   - Si $R^2_{cv}(h) \simeq R^2_{cv}(h-1)$, cela signifie que la composante ajoutée n'a pas d'effet sur l'explication de Y et l'équation doit contenir dans ce cas $(h-1)$ composantes.

2.Calcul du $PRESS$:

$$PRESS_h = \sum_{i=1}^{n} (Y_i - \hat{Y}^h_{(-1)})^2 \quad (66)$$

   - Cette quantité décroît en fonction du nombre de composantes pour atteindre une valeur minimale et se stabiliser par la suite. C'est ce minimum qui détermine le nombre de composantes à retenir pour le modèle [27].

2.4.4 Définition

La régression des moindres carrés partiels (régression PLS) est une méthode statistique qui a une certaine relation avec la régression en composantes principales, au lieu de trouver des hyperplans de variance maximale entre la réponse et les variables indépendantes, il trouve un modèle de régression linéaire en projetant les variables prédites et les variables observables dans un nouvel espace. Étant donné que les données X et Y sont projetées dans de nouveaux espaces, la famille de méthodes PLS est connue sous le nom de modèles à facteurs bilinéaires. L'analyse discriminante des moindres carrés partiels (PLS-DA) est une variante utilisée lorsque le Y est catégorique.

PLS est utilisé pour trouver les relations fondamentales entre deux matrices (X et Y), c'est-à-dire une approche à variables latentes pour modéliser les structures de covariance dans ces deux espaces. Un modèle PLS essaiera de trouver la direction multidimensionnelle dans l'espace X qui explique la direction de la variance multidimensionnelle maximale dans l'espace Y. La régression PLS est particulièrement adaptée lorsque la matrice de prédicteurs comporte plus de variables que d'observations et lorsqu'il existe une multi-colinéarité entre les valeurs X. En revanche, la régression standard échouera dans ces cas (sauf si elle est régularisée) [28].

2.5 Fonctionnement

En principe, le MLR peut être utilisé avec de nombreux facteurs. Toutefois, si le nombre de facteurs devient trop important (par exemple, plus grand que le nombre d'observations), vous obtiendrez probablement un modèle qui correspond parfaitement aux données échantillonnées, mais qui ne permettra pas de bien prévoir les nouvelles données. Ce phénomène est appelé surajustement.

Dans de tels cas, bien qu'il existe de nombreux facteurs manifestes, il peut n'y avoir que quelques facteurs sous-jacents ou latents qui expliquent la majeure partie de la variation de la réponse. L'idée générale du PLS est d'essayer d'extraire ces facteurs latents, en tenant compte de la plus grande partie possible de la variation manifeste des facteurs tout en modélisant bien les réponses.

Pour cette raison, l'acronyme PLS a également été interprété comme signifiant « projection vers une structure latente ». Il convient toutefois de noter que le terme « latent » n'a pas la même signification technique dans le contexte du PLS que pour d'autres techniques multivariées. En particulier, le PLS ne donne pas d'estimations cohérentes de ce qu'on appelle les « variables latentes » dans la modélisation formelle des équations structurelles.

La figure 11 donne un schéma de la méthode. L'objectif global est d'utiliser les facteurs pour prédire les réponses dans la population. On y parvient indirectement en extrayant les variables latentes T et U des facteurs échantillonnés et des réponses, respectivement. Les facteurs extraits T (aussi appelés scores X) sont utilisés pour prédire les scores Y U, puis les scores Y prédits sont utilisés pour construire des prédictions pour les réponses. Cette procédure couvre en fait diverses techniques, en fonction de la source de variation considérée comme la plus cruciale.

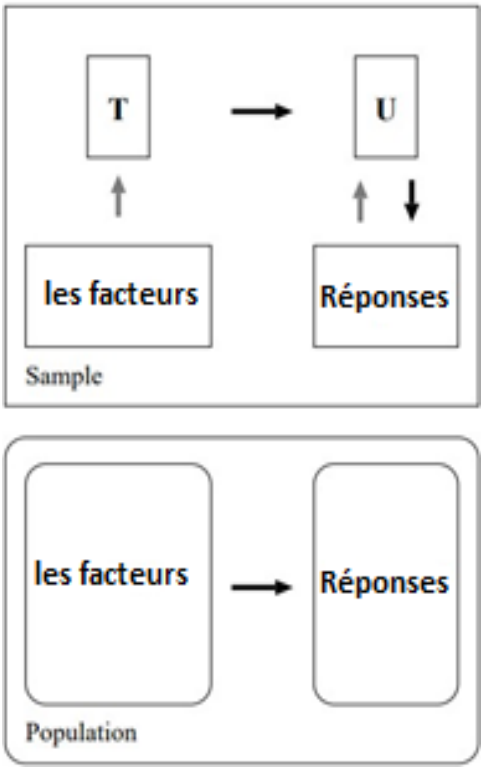

Figure 11-Modélisation indirecte (23).

- *Régression des composantes principales (PCR) :* Les scores X sont choisis pour expliquer le plus possible la variation des facteurs. Cette approche donne des directions informatives dans l'espace des facteurs, mais elles peuvent ne pas être associées à la forme de la surface prévue.
- *Analyse de redondance maximale (MRA) :* Les scores Y sont choisis pour expliquer le plus possible la variation Y prévue. Cette approche cherche des directions dans l'espace factoriel qui sont associées à la plus grande variation dans les réponses, mais les prédictions peuvent ne pas être très précises.
- *Moindres carrés partiels :* Les scores X et Y sont choisis de sorte que la relation entre les paires successives de scores soit aussi forte que possible. En principe, c'est comme une forme robuste d'analyse de redondance, cherchant des directions dans l'espace de facteurs qui sont associées

à une grande variation dans les réponses, mais les biaisant vers des directions qui sont prédites avec précision.

Une autre façon de relier les trois techniques est de noter que la PCR est basée sur la décomposition spectrale de $X'X$, où X est la matrice des valeurs de facteur. MRA est basée sur la décomposition spectrale de $\hat{Y}'\hat{Y}$, où $\hat{Y}$ est la matrice de (prédite) les valeurs de réponse, et PLS est basé sur la décomposition des valeurs singulières de $'Y$. Dans le logiciel SAS, la procédure REG et le logiciel SAS/INSIGHT mettent en œuvre des formes de régression des composantes principales, l'analyse de redondance peut être effectuée à l'aide de la procédure TRANSREG.

Si le nombre de facteurs extraits est supérieur ou égal au rang de l'espace du facteur d'échantillonnage, le PLS est équivalent au MLR. Une caractéristique importante de la méthode est qu'habituellement beaucoup moins de facteurs sont nécessaires. Le nombre précis de facteurs extraits est généralement choisi par une technique heuristique basée sur la quantité de variation résiduelle. Une autre approche consiste à construire le modèle du PLS pour un nombre donné de facteurs sur un ensemble de données, puis à le tester sur un autre, en choisissant le nombre de facteurs extraits pour lesquels l'erreur de prédiction totale est minimisée. Par ailleurs, van der Voet (1994) suggère de choisir le plus petit nombre de facteurs extraits dont les résidus ne sont pas significativement supérieurs à ceux du modèle avec une erreur minimale. Si aucun ensemble de test pratique n'est disponible, chaque observation peut être utilisée à son tour comme ensemble de test, ce qui est appelé validation croisée [25].

2.6 Régression et covariance

Les vecteurs latents pourraient être choisis de différentes façons. En fait, dans la formulation précédente, tout ensemble de vecteurs orthogonaux couvrant l'espace de la colonne de X pourrait être utilisé pour jouer le rôle de T. Afin de spécifier T, des conditions supplémentaires sont nécessaires. Pour la régression PLS, cela revient à trouver deux ensembles de poids w et c afin de créer (respectivement) une combinaison linéaire des colonnes de X et Y de sorte que leur covariance soit maximale. Plus précisément, l'objectif est d'obtenir une première paire de vecteurs $t = Xw$ et $u = Yc$ avec les contraintes que $w^T w = 1$, $t^T t = 1$ et$t^T u$ soit maximal. Lorsque le premier vecteur latent est trouvé, il est soustrait de X et Y et la procédure est réitérée jusqu'à ce que X devienne une matrice nulle (voir la section sur l'algorithme pour plus d'informations).

2.7 Algorithme

Les propriétés de la régression PLS peuvent être analysées à partir d'un croquis de l'algorithme original. La première étape consiste à créer deux matrices : $E = X$ et $F = Y$. Ces matrices sont ensuite centrées sur les colonnes et normalisées (c'est-à-dire transformées en Z). La somme de squaresof de ces matrices sont dénotées SS_X et SS_Y. Avant de commencer le processus d'itération, le vecteur u est initialisé avec des valeurs aléatoires. (Dans ce qui suit, le symbole α signifie « pour normaliser le résultat de l'opération »).

- Étape 1. $w \alpha E^T u$(Estimation X poids).
- Étape 2. $t \alpha Ew$ (Estimation des scores de facteur X).
- Étape 3. $c \alpha F^T t$ (Estimation des poids Y).
- Étape 4. $u = Fc$ (Estimation des scores Y).

Si t n'a pas convergé, passer à l'étape 1, si t a convergé, puis calculer la valeur de b qui est utilisée pour prédire Y à partir de t comme $b = t^T u$, et calculer les charges factorielles pour X comme $p = E^T t$. Maintenant, soustraire (c.-à-d. partiel) l'effet de t de E et de F comme suit $E = E - tp^T$ et $F = F - btc^T$. Les vecteurs t, u, w, c et p sont ensuite stockés dans les matrices correspondantes, et le scalaire b est stocké en tant qu'élément diagonal de B. La somme des carrés de X (respectivement Y) expliquée par le vecteur latent est calculée en $p^T p$ (respectivement b^2), et la proportion de variance expliquée est obtenue en divisant la somme expliquée des carrés par la somme totale correspondante des carrés (c.-à-d. SS_X et SS_Y).

Si E est une matrice nulle, alors tout l'ensemble des vecteurs latents a été trouvé, sinon la procédure peut être répétée à partir de l'étape 1.

2.7.1 Décomposition en valeurs singulières

L'algorithme itératif présenté ci-dessus est similaire à la méthode power-method qui trouve les vecteurs propres. Ainsi, la régression du PLS est susceptible d'être étroitement liée à la décomposition des valeurs propres et singulières, et c'est effectivement le cas. Par exemple, si nous partons de l'étape 1 qui calcule : $w \propto E^T u$, et que nous remplaçons le terme le plus à droite itérativement, nous trouvons la série suivante d'équations : $w \propto E^T u \propto E^T F c \propto E^T F F^T t \propto E^T F F^T E w$ Cela montre que le premier vecteur de poids w est le premier vecteur singulier droit de la matrice $X^T Y$. De même, le premier vecteur de poids c est le vecteur singulier gauche de $X^T Y$.. Le même argument montre que les premiers vecteurs t et u sont les premiers vecteurs propres de $XX^T YY^T$ et $YY^T XX^T$ [26].

2.7.2 Prévision des variables dépendantes

Les variables dépendantes sont prédites à l'aide de la formule de régression multivariée $\hat{Y} = TBC^T = XB_{PLS}$ avec $B_{PLS} = (P^{T+})BC^T$ (où P^{T+}est le pseudo-inverse de Moore-Penrose de P^T). Si toutes les variables latentes de X sont utilisées, cette régression est équivalente à la régression de la composante principale. Lorsque seul un sous-ensemble de variables latentes est utilisé, la prédiction de Y est optimale pour ce nombre de prédicteurs.

Une question évidente est de trouver le nombre de variables latentes nécessaires pour obtenir la meilleure généralisation pour la prédiction de nouvelles observations. Ceci est généralement obtenu par des techniques de validation croisée telles que l'amorçage.

L'interprétation des variables latentes est souvent facilitée par l'examen de graphiques apparentés aux graphiques PCA [26].

Chapitre 3 Détection des défauts

3.1 Introduction

La sécurité des processus et la qualité des produits sont deux questions cruciales pour les processus industriels modernes. La détection des défauts et le diagnostic jouent un rôle important du point de vue de l'amélioration de la qualité des produits et de la sécurité des processus. Grâce à une surveillance adéquate des processus, les temps d'arrêt sont réduits au minimum, la sécurité des opérations des processus est améliorée et les coûts de fabrication sont réduits. Bien entendu, le processus de surveillance peut être défini comme l'ensemble des actions menées pour détecter, isoler les sources de mesure défectueuses et les supprimer avant qu'elles n'affectent les performances du processus [29].

Le but de la détection de défaut est d'identifier tout événement de défaut indiquant une distance du comportement du processus par rapport à son comportement nominal. Alors que l'isolement des pannes est utilisé pour déterminer l'emplacement de la panne détectée [30]. Ce travail se concentre sur la détection des défauts. Les méthodes de surveillance des processus fondées sur les données, aussi connues sous le nom de méthodes fondées sur l'historique des processus ou de méthodes sans modèle [31]. Peuvent extraire des informations utiles des données historiques, en calculant la relation entre les variables sans avoir besoin d'un modèle analytique. À cette fin, les méthodes de surveillance fondées sur les données reposent sur la disponibilité des données historiques obtenues à partir du processus surveillé dans des conditions d'exploitation nominales [32]. Les données sans défaut sont d'abord utilisées pour construire un modèle empirique qui décrit le comportement du processus nominal, qui est ensuite utilisé pour détecter les défauts dans les données futures. Ensuite, le modèle empirique est utilisé pour estimer les valeurs réelles des nouvelles mesures, et les défauts sont détectés et diagnostiqués. Comme aucun modèle explicite n'est requis, dont le développement est habituellement coûteux ou chronophage, les méthodes fondées sur les données sont devenues très populaires dans les procédés industriels. Toutefois, la

performance des méthodes basées sur les données dépend principalement de la disponibilité de la quantité et de la qualité des données d'entrée.

Diverses techniques de détection des défauts fondées sur des données et peuvent être catégorisées en deux grandes catégories : les techniques variées et multi-variées. Les méthodes de surveillance statistique unidimensionnelle comme le graphique EWMA (moyenne mobile pondérée exponentiellement) et le graphique CUSUM (somme cumulative) sont essentiellement utilisées pour surveiller une seule variable de processus. Cependant, les procédés industriels modernes présentent souvent un grand nombre de variables hautement corrélées [33].

C'est le domaine où les méthodes un variées de détection des défauts sont incapables d'expliquer différents aspects du processus et, par conséquent, elles ne conviennent pas aux processus modernes. De plus, plusieurs variables de processus différentes ont été mises au point dans le cadre d'un suivi statistique multi-varié en même temps. Les méthodes détection des défauts multi-variées tiennent compte de la corrélation entre les variables d'un processus, mais pas les méthodes détection des défauts uni-variées. En particulier, pour la surveillance multi-variée des processus, la méthode de régression des variables latentes (LVR) a reçu beaucoup d'attention au cours des dernières décennies. L'idée principale de l'approche de surveillance fondée sur les LVR (e.g.., régression partielle du moindre carré (PLS), analyse des composantes principales (PCA)) consiste à extraire les données utiles de l'ensemble de données original et à construire certaines statistiques pour la surveillance [34][35] [36].

Le PLS, également connu sous le nom de projection à structure latente, est l'une des méthodes de surveillance multi-variée des processus statistiques (MSPM) les plus couramment utilisées pour surveiller les processus multi-variés [37]. Le PLS tente de décomposer les données de manière à maximiser la corrélation entre le prédicteur et les variables prévues [38]. En extrayant les données utiles de l'ensemble de données original, puis en utilisant des indices de surveillance tels que les statistiques T^2 et Q, le PLS a été utilisé avec succès pour la détection de défauts dans un processus multi-varié avec des variables fortement corrélées. La méthode de surveillance des processus basée sur le PLS ainsi que ses variantes ont été largement exploitées et utilisées pour différentes applications d'ingénierie [39][40].

La détection des défauts est d'une grande importance principalement lorsque les opérations sont effectuées à distance ou dans un environnement dangereux. Détecter les défauts dans le temps permet d'économiser beaucoup de temps et d'argent dans la réparation de l'équipement ou du produit fabriqué [41].

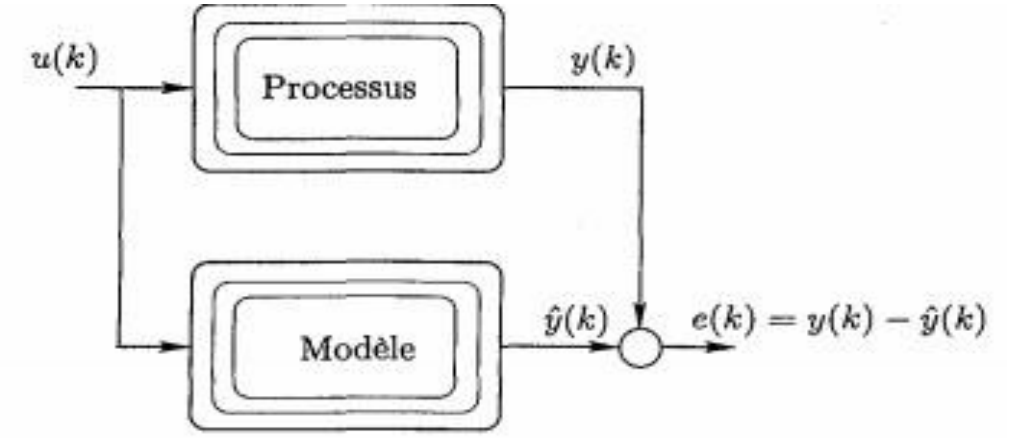

Figure 12-Principe de génération des résidus

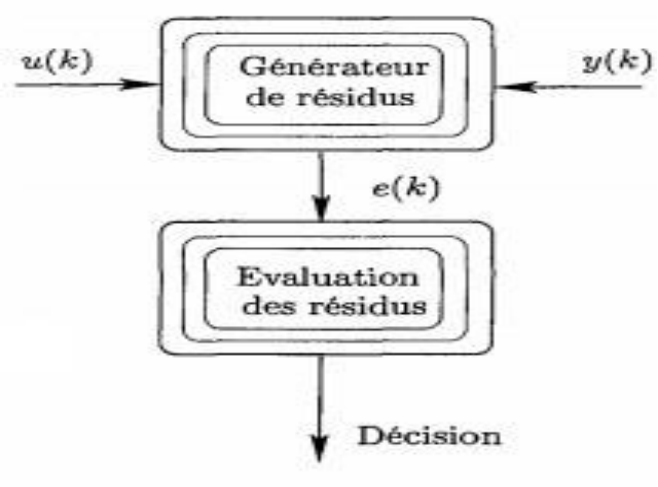

Figure 13-Structure générale d'un système de détection

3.2 Validation des signaux

La validation du signal est une partie importante du contrôle et de la surveillance des processus. Jusqu'à présent, de nombreuses technologies ont été développées et sont en cours de développement pour vérifier l'intégrité du signal. Le but de cette enquête est de déterminer la faisabilité de l'utilisation de réseaux de neurones à rétropropagation d'anticipation dans les capacités de détection de défauts de signal [42].

La vérification des signaux est la détection, l'isolement et la caractérisation des signaux de défaut. Du point de vue de l'amélioration de la disponibilité des installations et de la fiabilité du comportement des opérateurs, des signaux de processus correctement validés sont bénéfiques [43].

Un système logiciel complet de vérification des signaux a été développé, qui peut être appliqué aux centrales nucléaires. Le système combine certaines méthodes de détection de défauts établies et certains modules nouvellement développés. Ces technologies ont été mises en œuvre dans une architecture modulaire, permettant d'ajouter ou de supprimer des "modules" de vérification de signal selon les besoins. Le facteur intégré du module décrivant la validité d'un signal donné est dérivé à l'aide de fonctions d'appartenance floue. Le gestionnaire de système effectue l'évaluation finale de l'état du signal sur la base des résultats de chaque module de vérification de signal. Afin de prendre des décisions fiables dans ce système parallèle, un décideur actif a été développé [44].

3.3 Défaut

La défaillance est un écart inacceptable d'au moins un paramètre ou attribut du système par rapport aux conditions normales. Lorsque la variable mesurée dépasse le seuil, cela indique qu'il y a un défaut dans le système. Le seuil peut être connu à partir des données recueillies dans des conditions normales de travail avant.

3.4 Détection des défauts

La détection des défauts est la première étape de la surveillance multi-variée des processus. Généralement, les indices SPE (ou Q-statistique) et T^2 de Hotelling sont utilisés pour surveiller la variabilité normale dans RS et PCS, respectivement. Il convient de noter que la modélisation de l'PCA ou du PLS n'exige pas que les données soient gaussiennes. L'hypothèse gaussienne n'est nécessaire que pour déterminer les limites de contrôle appropriées pour les indices de détection de défauts [45].

De plus, l'indépendance temporelle des échantillons n'est pas requise pour déterminer les limites de contrôle, car seules les erreurs de type I sont spécifiées pour contrôler le taux de fausses alarmes. L'indépendance temporelle des mesures surveillées est nécessaire lorsqu'il s'agit d'erreurs de type II, c'est-à-dire le taux de défauts non détectés.

Lorsqu'une distribution gaussienne est supposée pour les mesures, il est généralement approprié d'utiliser la distance Mahalanobis pour définir la région normale de détection des défauts, par exemple, dans le sous-espace de la composante principale. Cependant, comme les données de processus sont généralement fortement corrélées, ce qui rend les variances des composants résiduels proches de zéro, il sera mal conditionné d'utiliser la distance de Mahalanobis dans RS. Par conséquent, la statistique Q ou SPE utilise la distance euclidienne pour définir la région normale de

détection de défaut. En raison de la nature complémentaire de ces deux indices, des indices combinés sont également proposés pour la détection et le diagnostic des pannes [46].

3.5 Méthodes de détection des défauts

Dans le domaine de recherche de la détection des défauts, un certain nombre de divisions sont décrites pour les différentes méthodes.

Les méthodes de détection des défauts sont divisées en deux grandes catégories : les méthodes fondées sur les données et les méthodes fondées sur des modèles.

3.5.1 Les méthodes fondées sur les données

Les méthodes de détection des défauts fondées sur les données n'utilisent que des données de processus (historiques) pour détecter les défauts dans le processus et ne nécessitent aucune connaissance préalable du processus. Ces méthodes sont formées avec des données non basiques, et sont utilisées pour reconnaître les écarts par rapport à ce comportement normal, qui sont causés par des défauts. L'exemple le plus simple est la vérification des limites. Cette méthode analyse les données historiques d'une seule variable et détermine ses limites normales. Ces limites sont généralement une limite de contrôle maximale supérieure (LCS) et une limite de contrôle minimale inférieure (LCI). Si la valeur de la variable du processus traverse une de ces limites, une erreur est déclarée. Ceci est illustré à la figure 14-a. Cependant, vérifier une seule variable n'est souvent pas suffisant pour détecter correctement les défauts. Les variables de processus sont souvent corrélées, ce qui signifie qu'une variable de processus peut avoir une influence significative sur les limites acceptables d'une autre variable. L'application de la vérification multi-variée des limites permet de déterminer les limites de fonctionnement acceptable pour plusieurs variables de processus simultanément. Cette situation est illustrée par la figure 14-b.

Il est également possible d'appliquer une sorte de transformation statistique, telle que l'analyse des composantes principales (PCA) ou les moindres carrés partiels (PLS), aux variables de processus pour réduire la dimensionnalité. Dans ce cas, la vérification des limites peut être appliquée à une ou plusieurs variables transformées. La vérification des limites est une méthode qui peut être utilisée pour détecter les écarts par rapport au comportement normal, mais il existe de nombreuses autres méthodes, comme la détection des changements, qui peuvent également être appliquées. Les écarts dans la moyenne, la variance ou la stationnarité en sont des exemples. Le principal

inconvénient des méthodes basées sur les données est que leur performance dépend fortement de la quantité et de la qualité des données de processus [47].

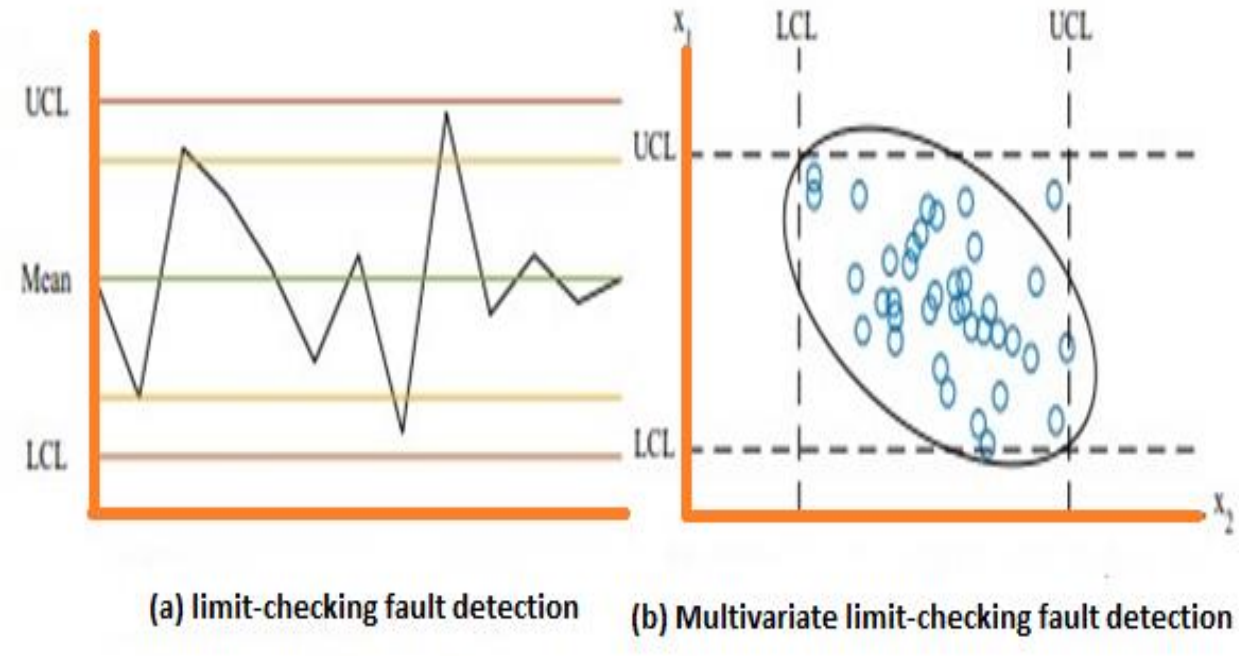

Figure 14-Exemples de méthodes de détection de pannes fondées sur des données [41]

3.5.2 Les méthodes fondées sur des modèles

Les méthodes de détection de défauts basées sur des modèles font des modèles d'un processus, et comparent leurs sorties aux sorties du processus réel. Ceci est illustré à la figure 15. La figure montre les signaux d'entrée mesurés U et les signaux de sortie Y, et leur relation est représentée par un modèle de processus mathématique. La méthode de détection de défaut correspondante extrait des caractéristiques, telles que les paramètres θ, les variables d'état x ou les résidus r. En appliquant une méthode de détection de changement pour comparer ces caractéristiques au comportement normal du système, les symptômes analytiques s sont générés. Le bruit dans le système est représenté par N.

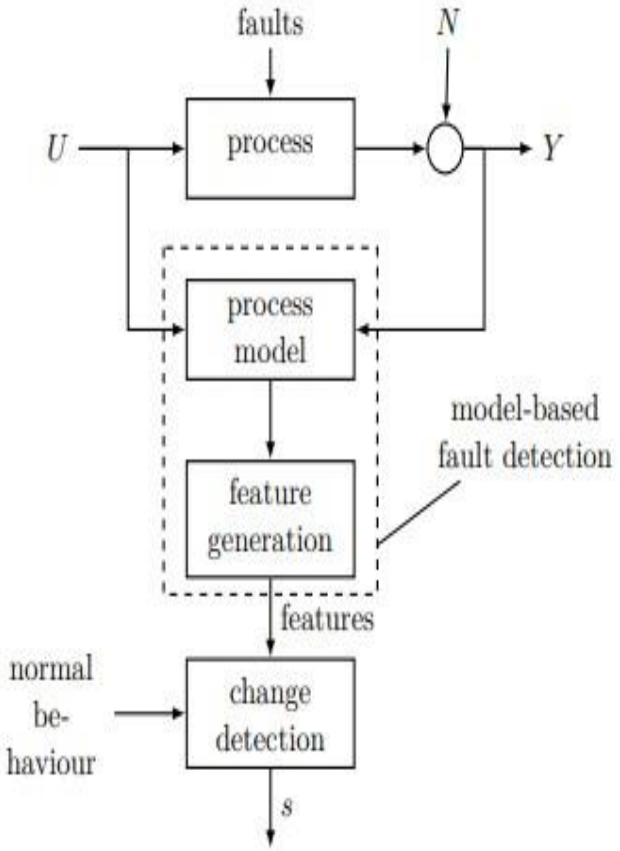

Figure 15-Schéma général de détection des défauts par modèle [41]

De nombreuses implémentations de méthodes de détection de défauts basées sur des modèles utilisent le résidu entre la sortie du processus réel et la sortie du modèle de processus comme caractéristique qu'ils utilisent pour la détection de défauts. La procédure générale est la suivante : un modèle du processus est construit, et ce modèle est utilisé, à un moment donné τ, pour générer une estimation de la variable de sortie $\hat{y}_\tau$ en utilisant les variables d'entrée de l'étape de temps précédente, $x_\tau - 1$. Le résidu peut alors être utilisé comme fonction de détection des pannes. Il est donné par :

$$r_\tau = y_\tau - \hat{y}_\tau \tag{67}$$

L'idée ici est que le modèle représente le processus dans des conditions normales de non-défectueuses, et que sa sortie représente la sortie du système si aucune défaillance n'était présente. On suppose donc que si aucune défaillance n'est présente dans le système, le résidu doit être égal à zéro. Cependant, si une défaillance se produit dans le système, la sortie du processus est affectée par la défaillance, tandis que la sortie du modèle n'est pas affectée. Dans ce cas, le résidu sera non nul, et il peut être conclu qu'une défaillance s'est produite. Dans la pratique, le bruit entraîne

également une valeur résiduelle supérieure à zéro, même s'il n'y a pas de défaut dans le système, c'est pourquoi un seuil est introduit $\bar{r}_\tau$. Si le résidu franchit ce seuil, une défaillance est détectée. Cette procédure est illustrée à la figure 16.

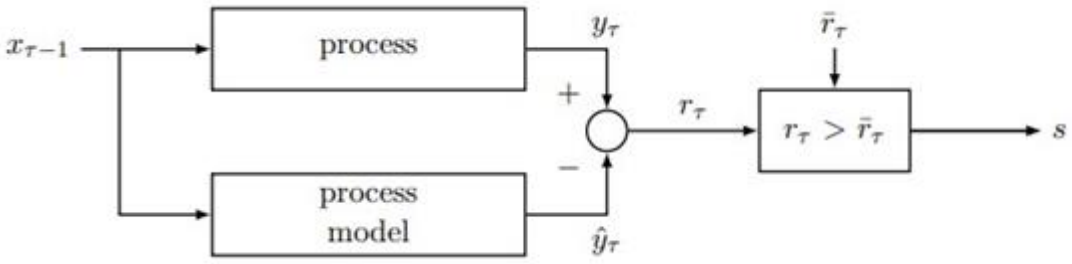

Figure 16-Exemple de méthode de détection des défauts [41]

Il existe de nombreuses méthodes différentes qui peuvent être utilisées pour construire un modèle de processus. Si la structure et les paramètres du processus sont connus, les premiers principes peuvent être utilisés. Dans ce cas, le modèle est construit en utilisant les propriétés physiques du système, qui sont régies par les lois de la nature. L'application de cette méthode aux grands systèmes se fait en commençant à modéliser chacun des sous-systèmes et en les combinant par la suite en un seul modèle global du système. Ce type de modélisation est appelé modélisation théorique et il commence toujours par des hypothèses sur le processus, qui simplifient la construction du modèle. Dans les grandes applications du monde réel, la modélisation théorique n'est pas toujours possible dans la pratique. Le système pourrait être si complexe que la construction de son modèle coûterait trop d'efforts ou pourrait même être trop complexe à modéliser. Dans ces cas, la modélisation expérimentale peut être appliquée. C'est ce qu'on appelle souvent l'identification, et ce type de modélisation permet d'obtenir le modèle mathématique du processus à l'aide de mesures. Les méthodes utilisent des mesures des signaux d'entrée et de sortie, qu'elle évalue de telle manière que leur relation soit exprimée dans un modèle mathématique. Des techniques telles que les régressions linéaires ou les réseaux neuronaux artificiels en sont des exemples. Si de telles méthodes sont utilisées, cela signifie que des techniques basées sur les données sont utilisées pour construire un modèle de processus ; cela diffère des techniques de détection de défauts basées sur les données qu'ils utilisent simplement des techniques basées sur les données pour analyser les données de processus.

Les modèles théoriques reposent sur une description fonctionnelle des données physiques du processus et de ses paramètres. Le modèle expérimental, en revanche, détermine ses paramètres

à partir de mesures, dont la relation avec les données physiques et le processus est inconnue. Par conséquent, ces derniers sont appelés modèles de boîte noire. En revanche, les modèles théoriques sont appelés modèles de boîte blanche. Cependant, il n'est pas toujours vrai qu'un modèle peut être classé comme boîte noire ou blanche. Par exemple, lorsqu'un modèle connaît les lois physiques, mais ne connaît pas les paramètres, et que les mesures du processus peuvent être utilisées pour les identifier. Ces modèles sont appelés modèles à boîte grise [47].

3.6 Modèles d'équations PLS

PLS est une technique statistique multi-variée réputée pour la réduction dimensionnelle des données de processus. Le rôle clé d'un PLS linéaire est basé sur sa capacité à traiter les données colinéaires, et plusieurs variables dans la matrice d'entrée (prédicteur) X et la matrice de sortie (réponse) Y. Dans sa forme générale, le PLS trouve les variables latentes des données de processus en saisissant la plus grande variabilité des données et obtient la corrélation croisée maximale entre le prédicteur et les variables de réponse. Étant donné une matrice de données d'entrée $X \in R^{n\times m}$ ayant n observations et m variables, et une matrice de données de sortie $Y \in R^{n\times p}$ composée de variables de réponse p, un modèle PLS est formellement déterminé par deux ensembles d'équations linéaires :

Le modèle intérieur et le modèle extérieur. Le modèle interne représente les relations entre les variables latentes (VL), et le modèle externe représente les relations reliant les VL et les variables observées qui leur sont associées

Le modèle extérieur, qui relie les VL, et les matrices de réponse et de prédicteur, peut être exprimé comme [48] :

$$\begin{cases} X = \hat{X} + E = \sum_{i=1}^{l} t_i P_i^T + E = TP^T + E \\ Y = \hat{Y} + F = \sum_{i=1}^{l} u_i q_i^T + F = UQ^T + F \end{cases} \quad (68)$$

Où $\hat{X}$et $\hat{Y}$ représentent les matrices de modélisation de X et Y, respectivement, les matrices $T \in R^{n\times l}$et $U \in R^{n\times q}$ consistent en l, VL conservés des données de prédicteur et de réponse, respectivement, les matrices $E \in R^{n\times m}$et $F \in R^{n\times p}$ représentent les résidus approximatifs des données de prédicteur et de réponse, respectivement, et les matrices $P \in R^{m\times l}$et $Q \in R^{p\times q}$ représentent respectivement les matrices de chargement des prédicteurs et des réponses. Le nombre de VL, l, peut être estimé en utilisant la validation croisée ou d'autres techniques.

$$U = TB + H \tag{69}$$

Où B représente une matrice de régression composée des paramètres du modèle reliant les VL's du prédicteur et de la réponse, et H représente une matrice résiduelle. La réponse Y peut maintenant être exprimée comme suit [49] :

$$Y = TBQ^T + F^* \tag{70}$$

Bien entendu, la méthode du PLS projette les données jusqu'à un certain nombre de VL qui expliquent la plupart des variations des prédicteurs et des réponses, puis modélise les VL au moyen de régressions linéaires.

Pour la surveillance basée sur le PLS deux statistiques, T^2 de Hotelling et Q ou erreur de prédiction au carré (SPE), sont généralement utilisées [50].

Différents indices de détection de défaut peuvent être utilisés pour les techniques linéaires PCA et PLS. Les deux indices les plus populaires sont les statistiques T^2 et Q. T^2 Mesure la variation du modèle, tandis que la statistique Q mesure la variation de l'espace résiduel, et ces statistiques seront décrites ensuite [51].

3.7 Méthodes classiques de PLS

3.7.1 Statistique $\mathbf{T^2}$

La statistique T^2 mesure la variation des composantes principales à différents échantillons de temps et se définit comme suit :

$$T^2 = X^T \hat{P} \hat{\Lambda} \widehat{P^T} X \tag{71}$$

Où $\hat{\Lambda} = diag\ (\lambda_1, \lambda_2, \dots, \lambda_l)$est la matrice diagonale qui contient les valeurs propres associées aux composantes principales retenues. Pour les données de test, une erreur est déclarée lorsque la valeur T^2 dépasse la valeur du seuil comme suit :

$$T^2 \geq T_\alpha^2 = \frac{(n^2-1)l}{n(n-l)} F(l, n - \ ^{\circ}_{\square}) \tag{72}$$

Où α est le niveau de signification, généralement attribué une valeur entre 90 et 99%, et $F(a, n - a)$ est la valeur critique de la distribution Fisher-Snedecor avec n et $n - a$ degrés de liberté.

3.7.2 Statistique Q

La statistique Q mesure la projection des données sur le sous-espace résiduel et permet à l'utilisateur de mesurer dans quelle mesure les données correspondent au modèle de l'PCA. La statistique Q est définie comme suit :

$$Q = \left\|\tilde{X}\right\| = \left\|\left(I - \hat{P}\hat{P}T\right)X\right\|^2 \tag{73}$$

Pour les données de test, une erreur est déclarée lorsque la valeur seuil est violée comme suit :

$$Q \geq Q_\alpha = \varphi_1 \frac{h_0 c_\alpha \sqrt{2\varphi_2}}{\varphi_1} + 1 + \frac{\varphi_2 h_0 (h_0 - 1)}{\varphi_1^2} \tag{74}$$

Où $\varphi_i = \sum_{j=l+1}^{m} \lambda_j^i i = 1,2,3, h_0 = 1 - \frac{2\varphi_1\varphi_3}{3\varphi_2^2}$, où $c_\alpha$est la valeur obtenue à partir de la distribution normale de signification α [52].

L'algorithme de détection des erreurs basé sur le PLS est ensuite résumé.

- Donné :
    - Un ensemble de données de formation sans défaut (X et Y) qui représente les opérations normales du processus et un ensemble de données de test (éventuellement des données défectueuses).
- Prétraitement des données :

- Mettre à l'échelle les données utilisées pour la construction du modèle de processus, à zéro moyenne et variance unitaire.

- Créer le modèle du PLS à l'aide des données de formation :
  - Sélectionner le nombre de variables latentes en utilisant la validation croisée ou toute autre méthode de sélection de modèle.
  - Exprimer la matrice de données en une somme de matrices approximatives et résiduelles, comme indiqué dans l'équation (68).
  - Calculer les limites de contrôle pour le modèle statistique (par exemple, les limites statistiques Qα).

- Tester les nouvelles données :
  - Redimensionner les nouvelles données avec l'écart moyen et l'écart-type obtenus à partir des données de formation.
  - Calculer les résidus des variables de réponse F.
  - Calculer la statistique de surveillance (statistiques Q ou T^2) pour les nouvelles données à l'aide des équations (71) ou (73).

- Vérification des défauts :
  - Déclarer une erreur lorsque de nouvelles données dépassent les limites de contrôle (par exemple, $Q > Q_\alpha$) [53].

Conclusion générale

La mission de la validation des signaux et la détection des pannes consiste à traiter les données d'archive afin de signaler les pannes et de déterminer le temps de leurs apparitions. La défaillance peut être associée à l'usure de l'équipement ou à des défauts de processus extrêmes. La précision de la détection des défauts à l'aide des données de processus peut être améliorée en utilisant des techniques de réduction de dimension des données telles que PLS, PCA, FDA, etc. Dans ce mémoire, nous avons étudié en détail la méthode PLS ainsi que ses avantages par rapport aux autres méthodes statistiques de détection de défauts. L'utilisation des techniques de réduction dimension, à la place d'observation direct des variables, résulte une réduction considérablement du nombre des défauts non détectés et par conséquent, la sensibilité et l'efficacité de détection de défauts est améliorée. Il a été observé également, que lorsque l'algorithme PLS est appliqué, le délai de détection des défauts est fortement réduit et par conséquent, la vitesse de détection des défauts est augmentée par rapport aux autres méthodes, particulièrement, lorsque les données sont fortement corrélées et que les petits défauts sont ciblés.

Le processus complet d'isolation des pannes comprend deux étapes, comme le montre la figure 17 : la détection et l'identification des défauts.

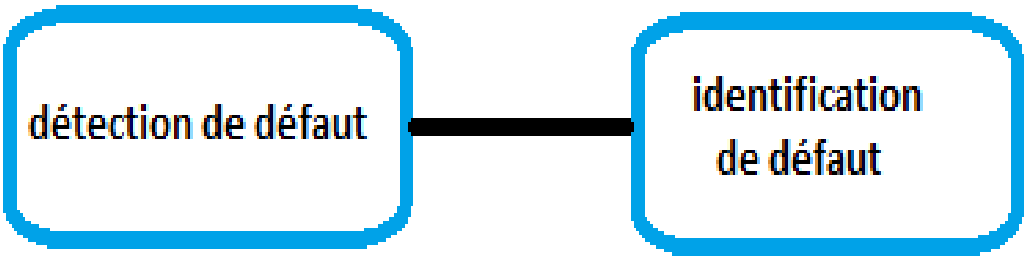

Figure 17-Procédure d'isolement des pannes.

Dans ce mémoire, nous n’avons traité que la première étape qui est la recherche des pannes. Une fois qu’un défaut est détecté dans le système, l'étape suivante (qui sera traité prochainement), consiste à trouver le type et l'emplacement exact du défaut dans le système (identification des défauts), en utilisant d'autres méthodes statistiques multi-variées telles que FDA, PCA, etc. Enfin, des actions correctives appropriées seront prises en fonction du type des pannes diagnostiquées.

Bibliographie

1. Mohamed-Faouzi, Harkat. Détection et Localisation de Défauts par Analyse en Composantes Principales. France : s.n., 2003.
2. Analyse et modélisation des données de classement. Marden, John I. 0-412-99521-2, 1995, Chapman & Hall, p. 59.
3. Fresnel, Jean. Algebre des matrices. s.l. : Hermann, 2013.
4. III, Lloyd N Trefethen and David Bau. Numerical linear algebra. s.l. : Siam, 1997. Vol. 50.
5. Cascaval, Radu C. Eigenvalues, Singular Value Decomposition. Colorado Springs : s.n.
6. wikipédia"espirance". [En ligne] https://fr.m.wikipedia.org/wiki/Esp%C3%A9rance_math%C3%A9matique?fbclid=IwAR2RbwBeK1qs-RI6nQCnBhCeCxHljt2Sy0paRlxLf773GlYqwLnRdMwU_V8.
7. DUSART, Pierre. Cours de Statistiques inférentielles. s.l. : Licence 2-S4 SI-MASS , 2018.
8. Mathématiques Générales B Université de Genève Sylvain Sardy. 22 mai 2008.
9. Gallardo, L. Notes du cours de Probabilités de M1 . s.l. : Université de Tours,, 2008-2009.
10. Notes statistiques : erreur de mesure. Bland, JM et Altman, DG. 7047, 1996, BMJ, Vol. 7047, p. 312.
11. MARÉCHAL HARGRAVE. Écart-type. Investopedia. [En ligne] 15 Avr 2021. https://www.investopedia.com/terms/s/standarddeviation.asp.
12. Variance (mathématiques). wikipedia. [En ligne] 11 Avr 2021. https://fr.wikipedia.org/wiki/Variance_(math%C3%A9matiques)#cite_ref-RAFisher_14-0.
13. Weisstein, Eric W. Variance. MathWorld. [En ligne] https://mathworld.wolfram.com/Variance.html.
14. najib, Yasmin. Qu'est-ce que la variance en statistique et ses caractéristiques ? 05 février 2021.
15. Covariance. Wikipedia. [En ligne] https://fr.wikipedia.org/wiki/Covariance#cite_ref-1.
16. JAMES CHEN, GORDON SCOTT. covariance. investopedia. [En ligne] 4 Mar 2020. https://www.investopedia.com/terms/c/covariance.asp.
17. Tille, Yves. Résume du Cours de Statistique Descriptive. [En ligne] 15 décembre 2010.
18. Wikipedia "Corrélation (statistiques)". [En ligne] [Citation : 12 mai 2021.] https://fr.wikipedia.org/wiki/Corr%C3%A9lation_(statistiques)#cite_ref-armatte_2-0.
19. Jim Frost. Introduction to Statistics: An Intuitive Guide for Analyzing Data and Unlocking Discoveries. 2020.

20. Van der Maaten, Laurens, Postma, Eric et van den Herik, Jaap. "Réduction de dimensionnalité : un examen comparatif" (PDF). 26 October 2009. Vol. 10, pp. 66–71.

21. Pudil, P. et Novovičová, J. Méthodes nouvelles pour la sélection de sous-ensemble de caractéristiques en ce qui concerne la connaissance des problèmes. [éd.] Huan Liu et Hiroshi Motoda. 1998. p. 101.

22. Stability and generalizatio. Bousquet, Olivier,André Elisseeff. 2.Mar (2002), Journal of machine learning research, pp. 499-526.

23. Methodological information (metadata). Thursday, May 23, 2002.

24. M, Lejeune. Statistique: la théorie et ses applications. s.l. : Springer, 2004.

25. Fatiha, LARBAOUI Djazia. Etude écotoxicologique des composés organiques sur les . Tlemcen : s.n., 2020.

26. PLS-régression : un outil de base de la chimiométrie. Wold, S, Sjöström, M. et Eriksson. (2), 2001, Chimiométrie et systèmes de laboratoire intelligents, Vol. 58 , pp. 109-130.

27. Abdi1, Hervé. Partial Least Square Regression PLS-Regression.

28. Tobias, Randall D. An Introduction to Partial Least Squares Regression. Cary,NC : SAS Institute Inc.

29. Model-based fault-detection and diagnosis : status and applications. Isermann, R. 2005, Annual Reviews in Control, Vol. 29, pp. 71–85.

30. Fault detection and diagnosis in industrial systems. L. Chaing, E. Russel,R. Braatz. 2001, Springer.

31. A review of process fault detection and diagnosis part III: Process history based methods. V. Venkatasubramanian, R. Rengaswamy, S. Kavuri,K. Yin. 2003, Computers and Chemical Engineering, Vol. 27, pp. 327–346.

32. Introduction to statistical quality control. Sons, John Wiley&. 2005, D. C. Montgomery.

33. Monitoring linear antenna arrays using an exponentially weighted moving average-based fault detection scheme. F. Harrou, M. Nounou. 1, 2014, Systems Science & Control Engineering: An Open Access Journal, Vol. 2, pp. 433–443.

34. Improved principal component analysis for anomaly detection: Application to an emergency department. F. Harrou, F. Kadri, S. Chaabane, C. Tahon,Y. Sun. 2015, Computers & Industrial Engineering, Vol. 88, pp. 63–77.

35. PLSbased EWMA fault detection strategy for process monitoring. F. Harrou, M. N. Nounou, H. N. Nounou,M. Madakyaru. 1, 2015, Journal of Loss Prevention in the Process Industries, Vol. 36, pp. 108-119.

36. Statistical fault detection using PCA-based GLR hypothesis testing. F. Harrou, M. Nounou, H. Nounou, M. Madakyaru. 1, 2013, Journal of Loss Prevention in the Process Industries, Vol. 26, pp. 129-139.

37. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. S. Wold, H. Ruhe, H. Wold,W. D. III. 3, 1984, SIAM Journal on Scientific and Statistical Computing, Vol. 5, pp. 735-743.

38. Partial least-squares regression: a tutorial. P. Geladi, B. Kowalski. 1986, Analytica chimica acta, Vol. 185, pp. 1-17.

39. A PLS-based statistical approach for fault detection and isolation of robotic manipulators. R. Muradore, P. Fiorini. 8, 2012, IEEE Transactions on Industrial Electronics, Vol. 59, pp. 3167-3175.

40. Partial least squares-based dimension reduction with gene selection for tumor classification. G. Li, X. Zeng, J. Yang,M. Yang. 2007, in Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2007. IEEE,, pp. 1439-1444.

41. Fault detection in process control plants using principal component analysis. Kandula, Vamshi Krishna. 2011.

42. Electric Power Research for the 90's, Actes de la deuxième conférence annuelle du programme de partenariat industriel . Holbert, Keith E. s.l. : Actes de la deuxième conférence annuelle du programme de partenariat industriel , 9 mars 1992.

43. Transactions IEEE sur la science nucléaire . Holbert, Keith E. 2, 1991, Vol. 38, pp. 803-811.

44. Technologie nucléaire. Keith E. Holbert, Belle R. Upadhyaya. 3, 1990, Vol. 92, pp. 411-427.

45. Mudholkar, Jackson et. 1979, Qin.

46. Cinar, Raich et. 1996, Yue et Qin.

47. Fault-diagnosis systems: An introduction from fault detection to fault tolerance. Isermann, R. 2006.

48. Process analysis, monitoring and diagnosis using multivariate projection methods: A tutorial. T. Kourti, J. MacGregor. 3, 1995, Chemometrics and Intelligent Laboratory Systems, pp. 3-21.

49. Model selection for partial least squares regression. B. Li, J. Morris,E. Martin. 1, 2002, Chemometrics and Intelligent Laboratory Systems, Vol. 64, pp. 79-89.

50. Analysis of a complex of statistical variables into principal components. Hotelling, H. 1933, Journal of Educational Psychology, Vol. 24, pp. 417–441.

51. Quality control methods for several related variables. Jackson, J. E. 4, 1959, Technometrics, Vol. 1, pp. 359–377.

52. Control procedures for residuals associated with principal component analysis. Mudholkar, J. Jackson and G. 1979, Technometrics, Vol. 21, pp. 341–349.

53. Enhanced Anomaly Detection Via PLS Regression Models and Information Entropy Theory. Fouzi, H., & Sun, Y. Cap, Afrique du Sud : s.n., 2015. Symposium Series on Computational Intelligence.