Détection de défauts à l'aide d'une analyse en composantes principales robuste

Mohamed Faouzi Harkat[(1)], Gilles Mourot[(2)], José Ragot[(2)]
[(1)] Université Badji Mokhtar - Annaba. Faculté des Sciences de l'Ingénieur
Département d'Electronique. BP. 12, 23000 Annaba, Algérie
[(2)] Institut National Polytechnique de Lorraine.
Centre de Recherche en Automatique de Nancy. UMR 7039 CNRS-UHP-INPL.
2, Avenue de la forêt de Haye. F - 54 516 Vandoeuvre les Nancy Cedex.
mharkat@univ-annaba.org,{gilles.mourot,jose.ragot}@ensem.inpl-nancy.fr

Résumé— **L'Analyse en Composantes Principales (*ACP*) est un outil statistique largement utilisé pour l'analyse de données collectées sur des systèmes en cours de fonctionnement afin de surveiller leur comportement. Cependant, d'un point de vue statistique, l'un des inconvénients majeurs de l'approche *ACP* résulte de son utilisation de techniques d'estimation par moindres carrés, techniques qui échouent à prendre en compte les biais de mesures accidentels ce qui est malheureusement assez fréquent sur le plan pratique. Cette communication présente : 1) la formulation d'une estimation robuste (vis-à-vis des valeurs aberrantes) de l'état d'un système basée sur l'analyse en composantes principales, 2) une procédure de détection et de localisation de défauts de mesures. La méthode proposée ne nécessite pas d'étude préliminaire relative à la détection et au rejet de valeurs aberrantes ou de grosses erreurs dans les données utilisées pour la conception du modèle *ACP*. Elle présente l'intérêt d'utiliser directement les données brutes, éventuellement entachées de grosses erreurs, et le modèle *ACP* est construit à partir de ces données sans filtre préalable, cette construction étant robuste vis-à-vis de la présence de grosses erreurs. Le modèle *ACP* obtenu étant sain, c'est-à-dire non contaminé par les valeurs aberrantes, son utilisation pour le diagnostic (détection et localisation de défauts de mesure) est alors efficace.**

Mots-clés— **ACP, redondance, valeurs aberrantes, robustesse, diagnostic, détection / localisation.**

I. Introduction

L'analyse en composantes principales est une technique numérique bien éprouvée dans le domaine du traitement de données pour réduire la dimension de l'espace de représentation d'un système [18]. En plus de cet aspect qui concerne sa première application, elle a été enrichie d'autres développements et a trouvé des domaines d'application très variés : compression de données [21], extraction de caractères [6], reconnaissance de formes [11], détection de fautes [24], [13], diagnostic de fonctionnement de systèmes [7], [8], [12], validation de données [19], modélisation de système [20], traitement de signal [10], supervision [25]. En dépit de ces extensions et applications variées, l'*ACP* est essentiellement basée sur la mise en évidence de relations linéaires entre les variables et présente un caractère d'optimalité uniquement au sens d'un critère portant sur l'erreur quadratique d'estimation en valeur moyenne (*MSE*). Il est bien connu que l'estimation basée sur l'utilisation de critère de type *MSE* est moins robuste que celle issue d'autres critères comme celui de l'erreur en valeur absolue.

De plus, rappelons que l'approche classique de l'*ACP* utilise un calcul préliminaire de la moyenne des données et de leur matrice de covariance en vue de la normalisation de ces données ; comme la moyenne et la variance sont sensibles à la présence de valeurs aberrantes, les résultats obtenus s'avèrent souvent inexploitables car trop biaisés par l'influence de ces valeurs aberrantes.

Pour tolérer la présence de valeurs aberrantes, une analyse en composantes principales robuste peut être conduite en calculant les valeurs propres et les vecteurs propres de la matrice de covariance des données (ou de la matrice de corrélation) évaluée au moyen d'une technique robuste. Pour cela, dans [5], les auteurs construisent des fonctions d'influence particulières et les variances asymptotiques qui en découlent ; le comportement du modèle *ACP* obtenu à partir de cette matrice de variance a ensuite été largement testé par des simulations.

Dans [9] et [17], les auteurs proposent l'approche robuste *ROBPCA* qui combine la poursuite de projection à une estimation robuste de la matrice de variance. Cette technique produit à la fois des estimations précises pour des données non contaminées par des valeurs aberrantes, ainsi que des estimations qui se révèlent robustes en présence de valeurs aberrantes. Dans [2], les auteurs se sont focalisés sur l'estimation robuste de la matrice de covariance pour des systèmes multi-dimensionnels, la valeur de cette matrice de covariance étant un point clef pour la recherche du modèle *ACP*. D'autres approches permettant d'appréhender le problème de robustesse ont été proposées dans [1], [15] en utilisant une loi de distribution dite contaminée des erreurs de mesure et dans [26] où les auteurs développent une approche basée sur un calcul de moments.

Dans la mise en œuvre de l'*ACP*, en plus des données aberrantes, un autre handicap à surmonter apparaît lorsque les données disponibles sont incomplètes ; c'est une situation assez fréquente lors de l'acquisition de mesures sur un processus en cours d'exploitation où, pendant certaines périodes de temps, telle ou telle grandeur n'est pas accessible à la mesure ; ce point est abordé dans [3], [4], [23], le principe généralement adopté étant l'utilisation d'indicateurs de présence ou d'absence des données lors du calcul du modèle *ACP*.

Notre présentation est essentiellement consacrée au problème de détection et de localisation de défauts dans des données. Dans ce domaine, on distingue habituellement deux familles de méthodes. La première repose sur la redondance d'informations qui résulte du couplage fonc-

tionnel entre les variables du système dont les données sont issues. Dans cette situation, la validation de données et plus généralement le diagnostic sont basés sur l'utilisation d'un modèle du système ; la procédure la plus classique consiste à générer des résidus indicateurs de défauts en comparant la sortie du système à celle de son modèle, tous deux étant soumis aux mêmes entrées d'excitation. La seconde approche utilise directement les données disponibles sans connaissance a priori du modèle du système. Les outils utilisés relèvent alors de l'analyse typologique des données, de la classification de données, de la reconnaissance de forme, de l'analyse de rupture de séquences temporelles. Dans la suite de cette présentation, cette approche à base de traitement de données sera systématiquement utilisée et notre contribution porte essentiellement sur la détection de valeurs aberrantes et leur localisation en utilisant deux outils complémentaires : la reconstruction de données et l'analyse de résidus.

Pour aborder ce point de vue lié au diagnostic des systèmes, la première étape de l'*ACP* consiste à déterminer les composantes principales utiles, c'est-à-dire celles traduisant le contenu informationnel de la matrice des données. Les dernières composantes principales expliquent l'information résiduelle non prise en compte par les premières ; l'analyse des données dans cet espace résiduel révèle la présence des défauts c'est-à-dire ici les valeurs aberrantes. Pour cela, nous nous sommes inspirés de ce qui est proposé en identification paramétrique et notamment dans les méthodes de moindres carrés robustes [16]. La section 2 est un bref rappel sur la procédure d'extraction des composantes principales. La présence de valeurs aberrantes et l'extraction robuste de composantes principales fait l'objet de la section 3. Une procédure de détection et de localisation des valeurs aberrantes est ensuite proposée en section 4, puis appliquée à un exemple de synthèse en insistant sur la génération de signatures de défaut.

II. Principe de l'*ACP*

Soit x un vecteur aléatoire constitué de n variables aléatoires caractérisant le fonctionnement d'un système à analyser. En l'absence de relations fonctionnelles entre les variables, on a généralement besoin des n variables pour décrire le comportement du système ; par contre, si l'une des variables x_i est liée aux autres, on peut cependant examiner si, au détriment d'une perte d'explicabilité quantifiable et admissible, il est possible de réduire le nombre de ces variables. Un des objectifs de l'*ACP* est précisément de décrire le système à l'aide d'un nombre restreint de variables et comme objectif secondaire de détecter les variables redondantes. Il est bien connu, [18], [22] que l'analyse des valeurs propres de Σ renseigne sur le nombre de variables explicatives à retenir.

Dans sa conception, l'*ACP* s'applique à des variables aléatoires. Cependant, dans la pratique, on est souvent confronté à analyser des données issues de chaînes de mesures collectant des informations sur le fonctionnement d'un système physique. Ces données sont rarement des variables aléatoires, mais la procédure précédente est néanmoins appliquée. On dispose alors d'une matrice de données $X \in \mathcal{R}^{N\times n}$, de vecteurs lignes x_i^T, qui rassemble les N mesures effectuées sur les n variables du système. Le critère de réduction de dimension est défini comme la somme des carrés des écarts entre les variables et leurs projections respectives sur une droite ad-hoc :

$$\phi(p) = \sum_{i=1}^{N} \parallel x_i^T - \pi x_i^T \parallel^2 \tag{1}$$

π étant la matrice caractérisant cette projection (définie à partir d'une direction p) sous la forme $\pi = pp^T$. On peut exprimer le minimum de ϕ vis-à-vis de p ce qui conduit à rechercher le vecteur propre de la matrice $\mathcal{S} = \sum_{i=1}^{N} x_i x_i^T$ associée à la valeur propre la plus petite de cette matrice. On peut compléter ce calcul en recherchant une deuxième direction, orthogonale à la première, et minimisant les écarts entre les variables et leurs projection sur ce deuxième axe. De façon générale, pour rechercher l'ensemble des directions ou axes principaux, on procède de la façon suivante :

- évaluer la matrice de covariance des données :

$$\mathcal{S} = X^T \left(I - \frac{UU^T}{U^T U} \right) X$$

  U, le vecteur de dimension N dont toutes les composantes sont égales à 1
- Résoudre l'équation :

$$\Sigma P = P\Lambda \tag{2}$$

  $P \in \mathcal{R}^{n.n}$ étant la matrice des vecteurs propres p_i de Σ et $\Lambda \in \mathcal{R}^{n.n}$ celle, diagonale, de ses valeurs propres.

On peut également montrer la décomposition suivante :

$$X = TP^T \tag{3}$$

$$T = XP \tag{4}$$

Ces relations trouvent uniquement leur intérêt lorsqu'on diminue la dimension de l'espace de représentation. Une fois déterminé le nombre ℓ de composantes à retenir, la matrice X des données peut être approximée de la façon suivante. Tout d'abord, la matrice des vecteurs propres est partitionnée sous la forme :

$$P = \begin{pmatrix} \hat{P} & \tilde{P} \end{pmatrix} \qquad \hat{P} \in \mathcal{R}^{n\times \ell} \tag{5}$$

On peut alors expliciter la partie des données expliquées par les ℓ premiers vecteurs propres et la partie résiduelle expliquée par les composantes restantes :

$$\hat{X} = X\hat{P}\hat{P}^T = X\sum_{i=1}^{\ell} p_i p_i^T \tag{6}$$

$$\tilde{X} = X - \hat{X} = X(I - \hat{P}\hat{P}^T) \tag{7}$$

Une redondance d'information (liée à des liaisons physiques entre des variables du système) est détectable par la présence de valeurs propres de Σ nulles. Dans la pratique, la présence d'erreurs de mesure affectant les données ou l'existence de liaisons "complexes" entre variables masquent partiellement ces valeurs nulles ; les redondances sont alors mises en évidence à partir d'un seuil approprié sur les valeurs propres.

III. Approche robuste de l'ACP

Ainsi que nous l'avons indiqué en introduction, une difficulté majeure de l'*ACP* provient de sa formalisation au moyen d'un critère des moindres carrés qui reste très sensible à la présence de valeurs aberrantes. Une solution consiste à utiliser une technique itérative. Dans une première étape, la présence éventuelle des valeurs aberrantes n'est pas prise en compte et l'*ACP* est effectuée sur l'ensemble des données disponibles. Puis, grâce au modèle *ACP* obtenu, les données sont analysées (analyse de leurs projections dans l'espace résiduel) afin de déceler la présence de valeurs aberrantes ; ces dernières sont alors retirées et l'*ACP* est reconduite sur les données restantes. Le processus peut être itéré plusieurs fois ; il est assez efficace lorsque peu de données sont contaminées mais échoue dans la plupart des cas. Pour tolérer la présence de valeurs aberrantes, une autre méthode couramment utilisée dans le domaine de la statistique consiste à remplacer l'estimation standard de la moyenne et de la variance par une estimation robuste [2], [3]. Une autre solution consiste à utiliser une approche par projection dans laquelle les vecteurs propres de la matrice de covariance sont progressivement estimés de façon à réduire l'influence des valeurs aberrantes. Dans ce qui suit, nous proposons une Analyse en Composantes Principales Robuste basée sur un M-estimateur. Deux points retiennent notre attention : 1) le modèle *ACP* doit être insensible aux valeurs aberrantes, 2) afin de satisfaire aux exigences de surveillance du système, les valeurs aberrantes doivent être détectées et isolées.

A. Hypothèses sur la nature des erreurs de mesure

L'approche proposée repose sur la prise en compte de deux type d'erreurs dans les données. Le premier type concerne les erreurs de faibles amplitudes provenant de variations à caractère aléatoire du comportement des capteurs. On fait souvent l'hypothèse que ces erreurs sont de valeur moyenne nulle et pour les caractériser on les considère comme des réalisations de variables aléatoires de distribution gaussienne (afin de faciliter les calculs d'estimation !). Le deuxième type d'erreur est relatif aux valeurs accidentelles, se manifestant sous forme de données dites aberrantes dans les bases de données ; elles peuvent être dues à des défauts passagers des capteurs. Ces valeurs aberrantes sont généralement de forte amplitude (vis-à-vis des erreurs du premier type) et peuvent aussi, pour faciliter les calculs d'estimation être considérés comme des réalisations de variables aléatoires. Dans ce qui suit nous utiliserons également une loi de distribution gaussienne pour les représenter. Pour distinguer les deux types d'erreur, les deux distributions seront caractérisées par des variances différentes. De façon plus globale, pour représenter ces erreurs, on propose d'utiliser une distribution dite contaminée correspondant au mélange, avec des poids μ et $1-\mu$, des distributions des deux types d'erreur :

$$\begin{cases} f_1(\varepsilon) = \dfrac{1}{\sqrt{2\pi}\sigma_1} exp\left(-\dfrac{1}{2}\left(\dfrac{\varepsilon}{\sigma_1}\right)^2\right) \\ f_2(\varepsilon) = \dfrac{1}{\sqrt{2\pi}\sigma_2} exp\left(-\dfrac{1}{2}\left(\dfrac{\varepsilon}{\sigma_2}\right)^2\right) \\ f(\varepsilon) = \mu f_1(\varepsilon) + (1-\mu) f_2(\varepsilon) \end{cases} \tag{8}$$

Le paramètre μ est à mettre en regard de la proportion de mesures saines vis-à-vis du nombre total de mesures, tandis que σ_1 et σ_2 caractérisent les distributions des erreurs affectant les mesures saines et de celles affectant les mesures aberrantes. Les figures 1 et 2 représentent deux situations, caractérisées par les paramètres $\{\sigma_1 = 1,\ \sigma_2 = 5\}$ et pour deux valeurs du paramètre de mélange $\mu = 0.1$ et $\mu = 0.5$. Chaque figure montre (de haut en bas) les distributions f_1 et f_2 puis la distribution mélange. Le rôle du paramètre de mélange μ apparaît alors clairement en comparant ces deux figures.

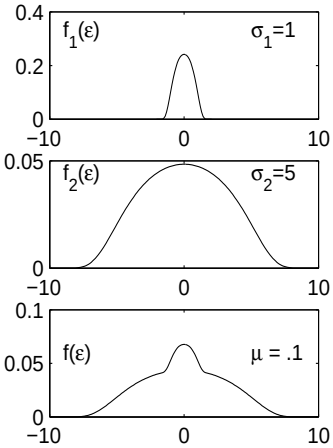

Fig. 1. Distribution contaminée ($\mu = 0.1$)

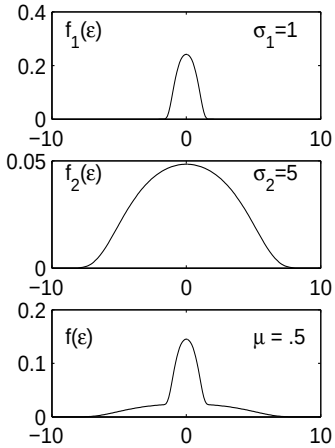

Fig. 2. Distribution contaminée ($\mu = 0.5$)

B. Formulation de l'ACP robuste

Dans cette section, on reprend la détermination des composantes principales d'une matrice de données X compte tenu de l'hypothèse qui vient d'être proposée sur la distribution des erreurs affectant les données. Soit $p \in \mathcal{R}^{\sqsubseteq}$ un vecteur unitaire sur lequel les données projetées sont définies par :

$$t = Xp, \quad t \in \mathcal{R}^{\mathcal{N}} \tag{9}$$

Compte tenu de (7), l'erreur d'estimation des données à partir de cet axe est :

$$\tilde{X} = X(I - pp^T) \tag{10}$$

Comme indiqué précédemment, l'existence d'au moins une relation de redondance entre les variables est traduite par le fait que cette erreur d'estimation est nulle en l'absence d'erreur de mesure. En présence d'erreurs de faibles amplitudes sur les variables, cette erreur sera elle aussi de faible amplitude. La présence de mesures aberrantes se traduit par une erreur de reconstruction significativement non nulle sauf si la direction de projection a été évaluée de façon robuste c'est-à-dire en s'affranchissant de la présence de ces valeurs aberrantes. C'est précisément cette évaluation robuste que nous envisageons. Pour cela, en utilisant le concept de distribution contaminée, nous formulons l'hypothèse suivante de distribution des erreurs de reconstruction :

$$\tilde{X} \sim \mu\mathcal{N}(0, \sigma_1^2 I) + (1-\mu)\mathcal{N}(0, \sigma_2^2 I) \tag{11}$$

chaque loi partielle prenant en compte l'un des deux types d'erreur. Ansi, l'erreur de reconstruction relative à l'obser-

vation de rang k, définie à partir de (10) :

$$\tilde{x}(k) = (I - pp^T)x(k) \tag{12}$$

est considérée comme la réalisation d'une variable aléatoire de densité de probabilité :

$$\begin{cases} f(\tilde{x}(k)) = \mu f_1(\tilde{x}(k)) + (1-\mu) f_2(\tilde{x}(k)) \\ f_1(\tilde{x}(k)) = \dfrac{1}{(2\pi)^{\frac{v}{2}}\sigma_1^v} \exp\left(-\dfrac{\| \tilde{x}(k) \|^2}{2\sigma_1^2}\right) \\ f_2(\tilde{x}(k)) = \dfrac{1}{(2\pi)^{\frac{v}{2}}\sigma_2^v} \exp\left(-\dfrac{\| \tilde{x}(k) \|^2}{2\sigma_2^2}\right) \end{cases} \tag{13}$$

Afin d'estimer à partir des mesures la direction de projection p, on forme la fonction de log-vraisemblance de l'ensemble des erreurs de'estimation, sous l'hypothèse d'indépendance statistique de ces erreurs :

$$\mathcal{V} = \sum_{k=1}^{N} \log\left(f(\tilde{x}(k))\right) \tag{14}$$

Le recherche du maximum de $\mathcal{V}$ vis-à-vis de p conduit à :

$$(X^T W(p) X - \lambda I)p = 0 \tag{15}$$

$$\begin{cases} W(p) = diag(w(1) \dots w(N)) \\ w(k) = \dfrac{\mu \dfrac{f_1(\tilde{x}(k))}{\sigma_1^2} + (1-\mu)\dfrac{f_2(\tilde{x}(k))}{\sigma_2^2}}{\mu f_1(\tilde{x}(k)) + (1-\mu) f_2(\tilde{x}(k))} \\ \tilde{x}(k) = (I - pp^T)x(k) \end{cases} \tag{16}$$

Pour résoudre ce système, notons que si $W(p)$ était connue, alors l'équation (15) exprime que λ et p sont les valeurs et vecteurs propres de la matrice $X^T W(p) X$. Comme la matrice de pondération W dépend implicitement de la direction p, la résolution de (15) est généralement réalisée par une procédure itérative. Ainsi, une procédure dite à itération directe s'explicite, en fonction de l'indice i d'itération, de la façon suivante :

$$\begin{cases} E_1 \quad \textit{Initialisation : } i = 0, w^{(i)}(k) = 1 \\ E_2 \quad i \leftarrow i+1 \textit{ Solution à l'étape } \text{i} \\ \quad W(p^{(i-1)}) = diag\left(w^{(i-1)}(1) \dots w^{(i-1)}(N)\right) \\ \quad \left(X^T W(p^{(i-1)}) X - \lambda^{(i)} I\right) p^{(i)} = 0 \\ \quad \tilde{x}^{(i)}(k) = (I - p^{(i)} p^{(i)T}) x(k) \\ \quad w^{(i)}(k) = \dfrac{\mu \dfrac{f_1^{(i)}(\tilde{x}^{(i)}(k))}{\sigma_1^2} + (1-\mu)\dfrac{f_2^{(i)}(\tilde{x}^{(i)}(k))}{\sigma_2^2}}{\mu f_1^{(i)}(\tilde{x}^{(i)}(k)) + (1-\mu) f_2^{(i)}(\tilde{x}^{(i)}(k))} \\ \quad f_1^{(i)}(\tilde{x}^{(i)}(k)) = \dfrac{1}{(2\pi)^{\frac{v}{2}}\sigma_1^v} \exp\left(-\dfrac{\| \tilde{x}^{(i)}(k) \|^2}{2\sigma_1^2}\right) \\ \quad f_2^{(i)}(\tilde{x}^{(i)}(k)) = \dfrac{1}{(2\pi)^{\frac{v}{2}}\sigma_2^v} \exp\left(-\dfrac{\| \tilde{x}^{(i)}(k) \|^2}{2\sigma_2^2}\right) \\ E_3 \quad \textit{test de convergence} \\ \quad | \left(X^T W(p^{(i)}) X - \lambda^{(i)} I\right) p^{(i)} | \leq eps? \end{cases} \tag{17}$$

En l'absence de connaissance particulière sur la présence de valeurs aberrantes, l'initialisation de cet algorithme itératif peut être faite en mettant tous les poids $w^{(0)}(k)$ à 1. Ainsi, à chaque itération, le vecteur p retenu correspond au vecteur propre de la matrice $XW(p^{(i)})X$ associé à la plus grande valeur propre de cette matrice. A la fin de cette procédure itérative, on dispose ainsi de la première composante principale déterminée de façon robuste vis-à-vis des valeurs aberrantes.

Pour rechercher les directions principales suivantes, on propose d'utiliser la méthode classique de déflation qui consiste à soustraire progressivement de la matrice X initiale les informations expliquées par la direction principale qui vient d'être trouvée. Ainsi à l'étape ℓ, notons $X^{(\ell)}$ la matrice X dont on a soustrait l'influence des directions propres précédentes. On a alors la mise à jour suivante :

$$X^{(\ell+1)} = X^{(\ell)} - Xp^{(\ell)}p^{(\ell T)}$$

Cette procédure de déflation est appliquée jusqu'à avoir extrait l'ensemble de l'information utile de la matrice des données X. La mise en œuvre proposée ici suppose que les paramètres de forme de la distribution contaminée soient a priori fixés. La connaissance même grossière de la proportion de valeurs aberrantes peut permettre un choix raisonnable du facteur de mélange μ ; dans le cas contraire quelques essais permettent de trouver la valeur adéquate, sachant que la technique tolère des valeurs de ce coefficient de mélange pouvant être assez différentes de la proportion de valeurs aberrantes. L'optimisation de ces paramètres a été également effectuée, mais le gain obtenu, au regard du volume de calcul supplémentaire, reste négligeable.

C. Exemple

La technique précédente est illustrée à partir d'un exemple de faibles dimensions de façon à pouvoir présenter l'ensemble des résultats numériques. Les données utilisées ont été générées à partir d'un modèle caractérisé par 6 relations linéaires entre 7 variables :

$$\begin{cases} x_1^*(k) \sim \mathcal{N}(0,1) \\ x_2^*(k) = -x_1^*(k) \\ x_3^*(k) = x_1^*(k) - x_2^*(k) \\ x_4^*(k) = x_1^*(k) - 3x_2^*(k) \\ x_5^*(k) = 0.5x_2^*(k) + x_3^*(k) \\ x_6^*(k) = x_1^*(k) - 4x_2^*(k) \\ x_7^*(k) = 0.2x_2^*(k) + x_3^*(k) \end{cases} \tag{18}$$

Les coefficients des variables apparaissant dans les différentes équations du modèle sont rassemblés dans la partie supérieure de la table (I). La partie inférieure de cette table présente le même modèle mais en faisant apparaître ces coefficients après combinaison linéaire des équations du modèle ; cela fait clairement apparaître la dépendance des six premières variables vis-à-vis de la septième. A chaque variable ainsi générée a été superposé un bruit issu de la réalisation d'une variable aléatoire à densité gaussienne :

$$x_i(k) = x_i^*(k) + b_i(k), \quad b_i(k) \sim \mathcal{N}(0, \sigma)$$

De plus, des valeurs aberrantes (chacune d'amplitude arbitrairement égale à 2) ont été ajoutées à différentes variables

Modèle vrai						
1	1	0	0	0	0	0
1	-1	1	0	0	0	0
1	-3	0	-1	0	0	0
0	0.5	1	0	-1	0	0
1	-4	0	0	0	-1	0
0	0.2	1	0	0	0	-1
Modèle vrai normalisé						
1	0	0	0	0	0	-0.556
0	1	0	0	0	0	0.556
0	0	1	0	0	0	-1.111
0	0	0	1	0	0	-2.222
0	0	0	0	1	0	-0.833
0	0	0	0	0	1	-2.778

TABLE I
COEFFICIENTS DU MODÈLE AYANT GÉNÉRÉ LES DONNÉES

(table II où la première ligne indique les variables corrompues, la deuxième ligne indiquant les instants où les valeurs aberrantes ont été appliquées).

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
k	2, 10	3, 4	6	10, 19, 28	19	28	37

TABLE II
LISTE DES VARIABLES CORROMPUES PAR DES VALEURS ABERRANTES

La table (III) regroupe les données ainsi générées, chaque variable ayant été cenrée par rapport à sa valeur moyenne, les valeurs aberrantes étant indiquées en caractère gras. Pour juger de l'influence des valeurs aberrantes sur le modèle ACP, la table (IV) donne les valeurs des composantes des vecteurs propres de la matrice X^TX les calculs ayant été conduits avec comme valeurs numériques pour les fonctions de distribution des erreurs : $\sigma_1 = .1, \quad \sigma_2 = 3, \quad \mu = .15$ La partie supérieure du tableau donne les composantes des 7 vecteurs propres de la matrice de variance-covariance des données calculées de façon non robuste. En ne retenant que les 6 vecteurs propres correspondant aux 6 valeurs propres les plus faibles, une combinaison linéaire de ces 6 vecteurs conduit à la partie centrale du tableau et de façon explicite aux équations du modèle. Ainsi, le premier modèle trouvé s'explicite :

$$x_1 - 0.608x_7 = 0$$

qui est à comparer au modèle ayant servi à générer les données (I), c'est-à-dire

$$x_1 - 0.556x_7 = 0$$

Plus généralement, la partie médiane du tableau (IV) est à comparer à sa partie inférieure du tableau (I) ce qui met en évidence la sensibilité importante des résultats obtenus vis-à-vis des valeurs aberrantes. Par contre, si l'on compare la troisième partie du tableau (IV) au tableau (I) l'utilisation de la méthode robuste proposée conduit à des résultats beaucoup plus proches du modèle théorique, le premier modèle étant en effet :

$$x_1 - 0.576x_7 = 0$$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	-0.59	0.31	-0.98	-2.02	-0.75	-2.33	-0.88
2	**2.70**	-0.01	-0.38	-0.83	-0.31	-0.83	-0.34
3	0.16	**2.52**	0.59	1.09	0.42	1.56	0.53
4	-0.16	**2.86**	-0.10	-0.26	-0.10	-0.13	-0.09
5	-0.34	0.04	-0.46	-0.99	-0.36	-1.02	-0.41
6	-0.50	0.20	**2.22**	-1.61	-0.62	-1.84	-0.70
7	-0.26	-0.02	-0.29	-0.69	-0.26	-0.67	-0.28
8	0.26	-0.54	0.72	1.37	0.53	1.91	0.65
9	-0.59	0.27	-0.95	-1.96	-0.72	-2.24	-0.84
10	**2.7**6	-0.06	-0.26	**2.42**	-0.20	-0.51	-0.23
11	0.15	-0.44	0.54	0.98	0.38	1.44	0.46
12	-0.67	0.37	-1.09	-2.28	-0.84	-2.64	-1.00
13	0.00	-0.27	0.21	0.31	0.12	0.58	0.18
14	0.27	-0.57	0.75	1.45	0.55	2.03	0.69
15	-0.44	0.12	-0.63	-1.31	-0.48	-1.45	-0.58
16	-0.56	0.27	-0.93	-1.92	-0.71	-2.19	-0.83
17	-0.02	-0.30	0.18	0.26	0.12	0.56	0.15
18	0.19	-0.50	0.62	1.18	0.45	1.68	0.55
19	-0.09	-0.20	0.02	**3.00**	**3.01**	0.19	0.03
20	0.20	-0.50	0.60	1.15	0.45	1.64	0.55
21	-0.52	0.21	-0.79	-1.67	-0.63	-1.91	-0.74
22	0.05	-0.36	0.34	0.57	0.22	0.92	0.26
23	-0.36	0.06	-0.50	-1.08	-0.39	-1.14	-0.46
24	-0.30	-0.01	-0.36	-0.81	-0.28	-0.78	-0.32
25	-0.54	0.25	-0.87	-1.82	-0.68	-2.08	-0.80
26	0.11	-0.42	0.44	0.82	0.30	1.24	0.41
27	-0.08	-0.22	0.08	0.04	0.03	0.26	0.04
28	0.02	-0.33	0.30	**3.49**	0.21	**3.85**	0.26
29	0.09	-0.41	0.41	0.75	0.31	1.18	0.36
30	-0.64	0.35	-1.06	-2.19	-0.81	-2.52	-0.97
31	0.24	-0.54	0.72	1.33	0.51	1.89	0.64
32	-0.19	-0.08	-0.19	-0.48	-0.16	-0.38	-0.20
33	0.04	-0.35	0.32	0.56	0.22	0.91	0.27
34	0.06	-0.35	0.28	0.52	0.20	0.84	0.25
35	0.14	-0.41	0.47	0.87	0.34	1.29	0.40
36	-0.57	0.25	-0.87	-1.85	-0.66	-2.08	-0.79
37	-0.27	-0.06	-0.28	-0.61	-0.25	-0.56	**2.74**
38	-0.10	-0.21	0.03	0.01	0.02	0.24	0.02
39	0.22	-0.54	0.67	1.29	0.48	1.80	0.59
40	0.13	-0.42	0.47	0.87	0.35	1.31	0.41

TABLE III
DONNÉES UTILISÉES

Vecteurs propres estimés de façon non robuste						
0.13	0.75	-0.12	0.55	0.03	-0.12	-0.29
-0.10	0.09	0.98	0.09	-0.02	-0.04	-0.00
0.21	-0.27	0.01	0.20	0.49	-0.76	0.15
0.61	0.39	0.06	-0.36	-0.01	0.04	0.58
0.22	0.09	0.04	-0.59	-0.19	-0.38	-0.64
0.67	-0.36	0.10	0.26	0.17	0.42	-0.35
0.21	-0.23	-0.01	0.32	-0.83	-0.28	0.15
Modèle estimé de façon non robuste						
1	0	0	0	0	0	-0.61
0	1	0	0	0	0	0.49
0	0	1	0	0	0	-0.99
0	0	0	1	0	0	-2.89
0	0	0	0	1	0	-1.04
0	0	0	0	0	1	-3.19
Modèle estimé de façon robuste						
1	0	0	0	0	0	-0.576
0	1	0	0	0	0	0.527
0	0	1	0	0	0	-1.111
0	0	0	1	0	0	-2.232
0	0	0	0	1	0	-0.835
0	0	0	0	0	1	-2.763

TABLE IV
COEFFICIENTS ESTIMÉS DU MODÈLE

La figure (3) visualise les poids $w(k)$ affectant les mesures

et qui ont été obtenus par l'algorithme de la section 3.2 ; les valeurs numériques indiquent bien que les observations volontairement biaisées $2, 3, 4, 6, 10, 19, 28$ et 37 ne sont pas prises en compte dans l'estimation du modèle ACP. En fait, afin d'accentuer cet effet, un renforcement des poids a été utilisé en n'utilisant que les valeurs 0 (pour les poids inférieurs à 0.5) et 1 (pour les poids supérieures à 0.5) ce que traduit la partie inférieure de la figure. La figure (4) visualise les erreurs d'estimation des six premières variables au cours du temps. Sans ambiguïté les valeurs aberrantes ont été détectées et localisées à partir des poids, leurs influences sur l'estimation des relations de redondance du système ont été fortement réduites ce qui explique la bonne qualité d'estimation des paamètres des modèles.

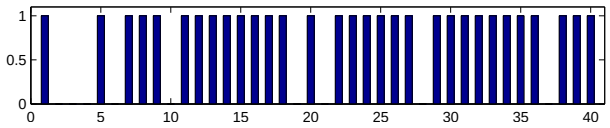

Fig. 3. Poids estimés

IV. Conclusion

En l'absence de modèle d'un système, la procédure proposée est bien adaptée à la recherche de relations de redondance à partir de mesures collectées sur ce système. Ces mesures pouvant être entachées de valeurs aberrantes, il convient de s'affranchir de leur influence lors de la recherche de ces redondances. L'approche robuste proposée ici réalise simultanément la recherche des redondances et le rejet des valeurs aberrantes grâce à l'estimation de poids appropriés. Par la suite, il est envisagé, toujours à partir de mesures, d'étendre cette procédure à la recherche de relations de redondance non linéaires en utilisant une approche de type courbes principales.

Références

[1] D. Böhning, R. Ruangroj. A note on the maximum deviation of the scale-contaminated normal to the best normal distribution. Metrika, 55, 177-182, 2002.

[2] C. L. Brown, R. F. Brcich and A. M. Zoubi. Adaptive M-Estimators For Robust Covariance Estimation. http ://asmda2005.enst-bretagne.fr/IMG/pdf/proceedings/872.pdf.

[3] H. Chen. Principal Component Analysis With Missing Data and Outliers. http ://www.caip.rutgers.edu/riul/research/tutorials / tutorialrpca.pdf

[4] J.M. Chen, B.S. Chen. System parameter estimation with input/output noisy data and missing measurements. IEEE Transactions on Signal Processing, 48 (6), 1548-1558, 2000.

[5] C. Croux, G. Haesbroeck. Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix : Influence Functions and Efficiencies, Biometrika, 87, 603-618, 2000.

[6] R. P.W. Duin, R. Haeb-Umbach. Multi-Class Linear Feature Extraction by Nonlinear PCA. 15th International Conference on Pattern Recognition, 2, 239, 2000.

[7] R. Dunia and Qin, S.J. Joint diagnosis of process and sensor faults using principal component analysis. Control Engineering Practice, 6 (4), 457-469, 1998.

[8] R. Dunia and S. J. Qin. A subspace approach to multidimensional fault identification and reconstruction. AIChE Journal, 44 (8), 1813-1831, 1998.

[9] S. Engelen, M. Hubert, K. Vanden Branden. A comparison of three procedures for robust PCA in high dimensions. Austrian Journal of Statistics, 34 (2), 117-126, 2005.

[10] R Hong Enríquez, M.S. Castellanos, J.F. Rodríguez and J.L. Hernández Cáceres. Analysis of the photoplethysmographic signal by means of the decomposition in principal components. Physiol. Meas. 23 N17-N29, 2002.

[11] J. Fortuna, D. Sc huurman, D. Capson. A Comparison of PCA and ICA for Object Recognition Under Varying Illumination. 16th International Conference on Pattern Recognition, 3, 300-311, 2002.

[12] J. Gertler and J. Cao. PCA-based process diagnosis in the presence of control. IFAC Safeprocess Symposium, Arlington, VA, 2003.

[13] M.F. Harkat, G. Mourot, J. Ragot. Nonlinear PCA combining principal curves and RBF-networks for process monitoring. 42th IEEE Conference on Decision and Control, Hawaii, USA, 2003.

[14] M.F. Harkat, Mourot G., Ragot J. Variable reconstruction using RBF-NLPCA for process monitoring. 5th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes, Safeprocess'2003, Washington, D.C., USA, 2003.

[15] I. Higuchi and S. Eguchi. Robust principal component analysis with adaptive selection for tuning parameters. J. Machine Learning Research 5, 453-471, 2004.

[16] Paul W. Holland and Roy E. Welsch. Robust regression using iteratively reweighted least- squares. Commun. Statist.-Theor. Meth., A6 :813-827, 1977.

[17] M. Hubert, P. J. Rousseeuw and K. Vanden Branden. ROBPCA : a New Approach to Robust Principal Component Analysis. http ://www.wis.kuleuven.ac.be/stat/Papers/robpca.pdf

[18] I. Jollife. Principal Component Analysis. Springer-Verlag, New York, 1986

[19] G. Kerschen, P. De Boe, J.C. Golinval, K. Worden Sensor validation using principal component analysis Smart Materials and Structures 14, 36-42, 2005.

[20] Yu Qian, H. Cheng, X. Li, and Y. Jiang. Dynamic Process Modelling using a PCA-based Output Integrated Recurrent Neural Network. The canadian journal of chemical engineering, 80 (4), 2002.

[21] N. Roy, G. Gordon. Exponential Family PCA for Belief Compression in POMDPs. In Advances in Neural Information Processing Systems 15, S. Becker, S. Thrun and K. Obermayer (eds.), MIT Press, 2003.

[22] G. Saporta. Probabilité, analyse des données et statistiques. Edition Technip - Paris. 493 p, 1990.

[23] H. Shum, K. Ikeuchi, R. Reddy. Principal component analysis with missing data and its application to polyhedral object modeling. Modelling from reality, 3-39, Kluwer Academic, 2001.

[24] H. Tong, C. M. Crowe. Detection of gross errors in data reconciliation by principal component analysis. AIChE Journal, 41 (7), 1712 - 172, 1995.

[25] A. Wachs and D. R. Lewin. Process monitoring using model based PCA. 5th IFAC Symposium of Dynamics and Control of Processing Systems, 1998.

[26] G. Willems, S. Van Aelst and M. Salibian-Barrera. Robust PCA with bootstrap based on MM-estimators.

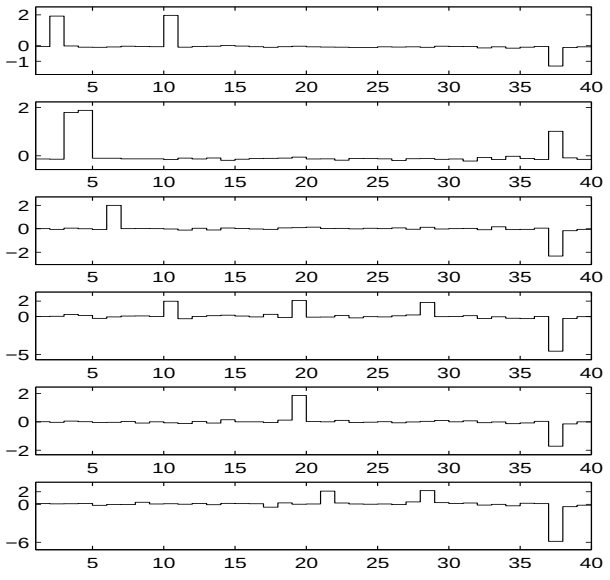

Fig. 4. Projection des données au cours du temps