Diagnostic de fonctionnement d'une station d'épuration par analyse en composantes principales

Yvon THARRAULT*, Gilles MOUROT*, José RAGOT*, David FIORELLI** et Serge Gillé**

*Centre de Recherche en Automatique de Nancy
UMR7039 Nancy-université CNRS
2, avenue de la Forêt de Haye, 54516 Vandoeuvre-lès-Nancy, France
{yvon.tharrault, gilles.mourot, jose.ragot}@ensem.inpl-nancy.fr

*Centre de recherche public Henri Tudor
2A, rue Kalchesbrück, L-1852 Luxembourg-Kalchesbrück
{david.fiorelli, serge.gille}@tudor.lu

I. INTRODUCTION

L'*ACP* est essentiellement basée sur la mise en évidence de relations linéaires entre les variables mais elle est sensible aux valeurs aberrantes. Pour tolérer la présence de valeurs aberrantes, une analyse en composantes principales robuste est conduite en calculant une matrice de covariance des données robuste, cette matrice étant un point clef pour la recherche du modèle *ACP*. Notre présentation est consacrée à l'application de l'analyse en composantes principales robuste à une station d'épuration pour la détection et la localisation de défauts capteurs.

II. PRINCIPE DE L'ANALYSE EN COMPOSANTES PRINCIPALES

Soit une matrice de données $X \in \Re^{N\times n}$, de vecteurs lignes x_i^T, qui rassemble les N mesures effectuées sur les n variables du système.

II.1. Approche classique

Dans le cas de l'*ACP* classique, les données sont supposées être recueillies sur un système ayant un fonctionnement normal (absence de défauts), ce qui est parfois difficile à vérifier dans le cas de données issues d'un système complexe.
L'*ACP* détermine une transformation optimale (vis-à-vis d'un critère de variance) de la matrice de données X :

$$T = XP \quad \text{et} \quad X = TP^T \tag{1}$$

$T \in \Re^{N\times n}$ est la matrice des composantes principales et $P = [p_1\ p_2\ \ldots\ p_n] \in \Re^{n\times n}$, où les vecteurs orthogonaux p_i sont les vecteurs propres correspondant aux valeurs propres λ_i de la décomposition en valeurs et vecteurs propres de la matrice de covariance (ou de corrélation) Σ de X :

$$\Sigma = P\Lambda P^T \qquad \text{avec} \qquad PP^T = P^TP = I_n \tag{2}$$

avec $\Lambda = diag(\lambda_1 \ldots \lambda_n)$ une matrice diagonale où les termes diagonaux sont ordonnés dans l'ordre décroissant.
Différentes approches pour déterminer le nombre de composantes principales sont comparées dans (Valle *et coll.* 1999). Une fois déterminé le nombre ℓ de composantes à retenir, la matrice X des données peut être approximée. Pour cela la matrice des vecteurs propres est partitionnée sous la forme :

$$P = \begin{pmatrix}\hat{P} & \tilde{P}\end{pmatrix} \qquad \hat{P} \in \mathcal{R}^{n\times\ell} \tag{3}$$

A partir de la décomposition (1), on peut alors expliciter la partie principale $\hat{X}$ des données expliquées par les ℓ premiers vecteurs propres et la partie résiduelle $\tilde{X}$ expliquée par les composantes restantes :

$$\hat{X} = X\hat{P}\hat{P}^T = XC_\ell \qquad E = X - \hat{X} = X(I - C_\ell) \tag{4}$$

II.2. Approche robuste

Une difficulté majeure de l'*ACP* provient de sa sensibilité aux valeurs aberrantes. Dans (Fekri *et al.*, 2003) les auteurs définissent une matrice de covariance "locale" qui tend à privilégier la contribution d'observations proches au détriment d'observations éloignées dues à la présence de valeurs aberrantes. Cette matrice est définie par :

$$V = \frac{\displaystyle\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} w_{i,j}(x_i - x_j)(x_i - x_j)^T}{\displaystyle\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} w_{i,j}} \tag{5}$$

où les poids $w_{i,j}$ sont eux-mêmes définis par :

$$w_{i,j} = \exp\left(-\frac{\beta}{2}(x_i - x_j)^T\Sigma^{-1}(x_i - x_j)\right) \tag{6}$$

β étant un paramètre à régler pour obtenir effectivement une réduction de l'influence des observations éloignées, les auteurs préconisant une valeur voisine de 2. L'*ACP* peut alors être conduite sur cette "nouvelle" matrice de covariance réputée robuste vis-à-vis des valeurs aberrantes grâce à la présence de poids adaptés $w_{i,j}$.

III. APPLICATION A UNE STATION D'EPURATION DES EAUX USEES

Nous allons maintenant appliquer l'analyse en composantes principales robuste sur les données hydrauliques de la station d'épuration de bleesbrück (Luxembourg). Le but est de valider les informations issues des capteurs utilisés pour la conduite du système. Lors du fonctionnement normal de la station, il existe quelques points avec un fonctionnement très différents. Ces points sont considérés comme des valeurs aberrantes.

La figure 1 décrit la partie hydraulique de la station ainsi que la position des différents capteurs. Les mesures avec un numéro grisé (2, 5, 7, 8, 9) représentent les différentes commandes de la station avec la localisation des actionneurs, les autres numéros (1, 3, 4, 6, 10) représentent la position des capteurs à valider. L'eau arrive ensuite dans un puisard qui permet de la stocker, on mesure la hauteur de l'eau dans le puisard. Puis une station de relevage relève les eaux usées pour permettre un écoulement gravitaire dans le reste de la station. A la sortie de la station de relevage, il y a une mesure du débit. Les eaux usées sont alors traitées par boues activées. Il y a un circuit de recirculation et d'extraction (Boues en excès et boues de surface) des boues. Puis un capteur mesure la hauteur de la surverse situé après le décanteur, cette surverse permet de limiter le débit arrivant dans la 2ème biologie et cette mesure permet d'estimer le débit rejeté directement dans la rivière Sûre. Pour les différents débits de recirculation et d'extraction des boues tout comme pour la station de relevage, on connaît les commandes des pompes et les débits nominaux. Une mesure est enregistrée toutes les 2 minutes.

La matrice de données est construite à partir des différents signaux disponibles décalés ou non dans le temps. La figure 2 montre la mesure après le dégrilleur (numéro 3), avec son estimation

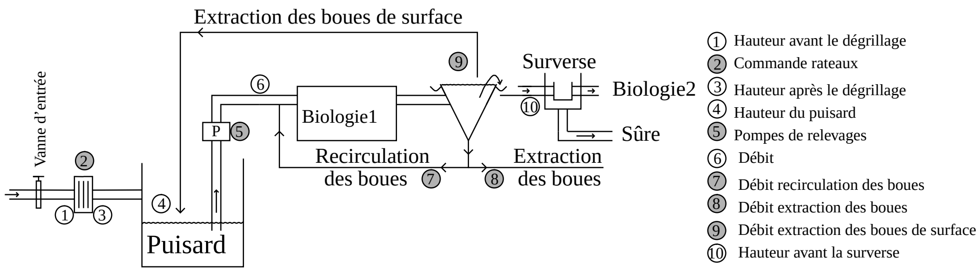

Figure 1. Description de la station

par ACP classique et par ACP robuste ainsi que les résidus (mesure - estimation) correspondants. L'ACP classique veut modéliser l'ensemble des données, y compris les valeurs aberrantes. Alors que l'ACP robuste veut modéliser seulement les données saines. On trouve alors dans le cas de l'ACP robuste un résidu proche de zéro sauf pour les valeurs aberrantes (points 125 et 130) contrairement au cas de l'ACP classique.

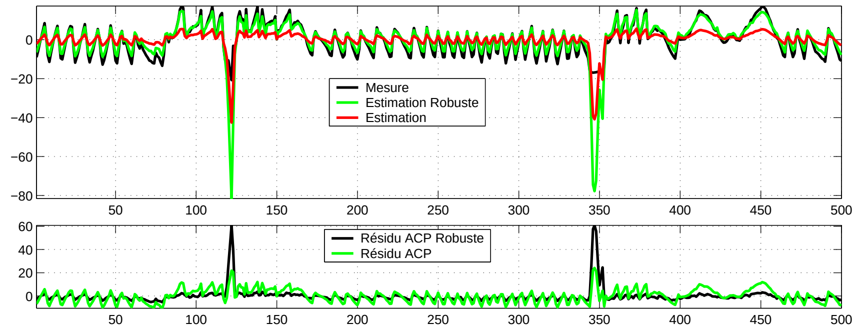

Figure 2. Hauteur après le dégilleur

Les principes de reconstruction et d'analyse des signatures des défauts (Tharrault *et coll.* 2007) sont ensuite appliqués pour détecter et localiser les valeurs aberrantes multiples.

IV. CONCLUSION

Lorsque des valeurs aberrantes corrompent les données, l'*ACP* classique s'avère inefficace, alors que sa version robuste donne des résultats tout à fait satisfaisants. On peut donc construire un modèle *ACP* directement à partir des données disponibles contenant d'éventuels défauts. L'utilisation conjointe des principes de reconstruction et d'analyse des signatures des défauts a permis de détecter et d'isoler de façon efficace les données aberrantes.

Références

Fekri, M. et A. Ruiz-Gazen (2003). Robust weighted orthogonal regression in the errors-in-variables model. *Journal of Multivariate Analysis*, **88**, 89–108.

Tharrault, Y., G. Mourot, J. Ragot, D. Fiorelli et S. Gillé (2007). Détection et isolation de défauts par analyse en composantes princpales. In : *Journées doctorales MACS*. Reims.

Valle, S., W. Li et S.Joe Qin (1999). Selection of the Number of Principal Components : The Variance of the Reconstruction Error Criterion with a Comparaison to Other Methods. *Industrial & Engineering Chemistry Research*, **38**, 4389–4401.