

Diagnostic de défauts précoces pour des signaux à distribution Gamma : Analyse théorique

Abdulrahman YOUSSEF^{1,2}, Claude DELPHA¹, Demba DIALLO²

¹Laboratoire des Signaux et Systèmes (L2S) (Univ. Paris Sud, CNRS, CentraleSupélec)
3, rue Joliot Curie, 91192, Gif-sur-Yvette, France

²Laboratoire de Génie Électrique et Électronique de Paris (GEEPS) (Univ. Paris Sud, CNRS, CentraleSupélec, UPMC)
11, rue Joliot Curie, 91192, Gif-sur-Yvette, France

Abdulrahman.youssef@centralesupelec.fr, claude.delpha@l2s.centralesupelec.fr
demba.diallo@geeps.centralesupelec.fr

Résumé – Le diagnostic des défauts précoces (naissants) est devenu ces dernières années une préoccupation majeure pour assurer la sûreté de fonctionnement des processus industriels. Parmi les approches récentes, les méthodes pilotées par les données ont prouvé leur efficacité pour les systèmes complexes. Elles imposent tout de même l'utilisation de techniques d'extraction et d'analyse de caractéristiques efficace pour le diagnostic telles que les approches statistiques multivariées. En plus de la détection de défaut, l'estimation de sa sévérité est essentielle pour prendre la décision appropriée (arrêt du processus, maintenance,...). Dans cet article, nous développons un modèle analytique permettant la détection de défaut et s'appuyant sur la divergence de Kullback-Leibler (DKL) lorsque l'on considère que les données suivent une distribution Gamma. L'objectif de ce modèle étant son utilisation pour l'estimation de la sévérité d'un défaut de capteur. Dans l'espace de l'Analyse en Composantes Principales (ACP), le modèle proposé est analysé et comparé aux valeurs estimées de la DKL lorsqu'on utilise l'estimateur de Monte-Carlo. Les résultats montrent que pour des défauts naissants ($< 10\%$ et $SNR > 40dB$), le modèle analytique a une précision relative de 90% qui permet son utilisation pour l'estimation de la sévérité du défaut.

Abstract – Incipient fault diagnosis has become a key issue for the safety of industrial processes. In the last years, data-driven methods have become more and more attractive for diagnosis purpose. However they require effective feature extraction and analysis. Recently the use of multivariate statistical techniques has proved its efficiency. Within the fault diagnosis scope, beside fault detection, fault estimation is also essential to make the best decision (safe stop, maintenance, fault accommodation). An analytical model of the Kullback-Leibler Divergence (KLD) for noisy and fault gain is therefore developed with the assumption of Gamma distributed. This model is evaluated and its accuracy is studied with variable noise conditions by a comparison to Monte-Carlo estimator. The results show that for incipient case (fault amplitude $< 10\%$, and $SNR > 40dB$), the analytical model is accurate enough with a relative error lower than 10%.

1 Introduction

Les systèmes industriels modernes ont des exigences croissantes de sûreté de fonctionnement. Pour assurer la fiabilité et la sécurité, le système doit être surveillé de façon à détecter, au plus tôt tout défaut, à le localiser et à estimer sa sévérité. Les méthodes de détection et d'estimation de défauts doivent être robustes aux bruits, aux incertitudes inattendues et aux perturbations. Si un modèle analytique du système en défaut est disponible, des approches d'optimisation peuvent être utilisées pour estimer précisément l'amplitude de ce défaut. Mais dans la plupart des cas, c'est un modèle piloté par les données qui est utilisé. Il est bien souvent combiné avec des méthodes statistiques pour la détection et l'estimation.

Dans cet article, nous adoptons une approche basée sur les données utilisant des caractéristiques descriptives dans

l'espace de l'Analyse en Composantes Principales (ACP) [1] [2] combinée avec des techniques d'analyse statistique pour développer une méthode de détection et d'estimation de défaut efficace. L'avantage des méthodes de surveillance basées sur l'ACP est qu'elles peuvent traiter des données de grandes dimensions, bruitées et fortement corrélées. Elles permettent généralement d'obtenir de bonnes performances dans le cadre de la surveillance de systèmes industriels complexes.

Il a déjà été montré, que dans l'espace de l'ACP, la divergence de Kullback-Leibler (DKL) [3] est conceptuellement plus robuste et aussi plus sensible pour la détection de défauts naissants que les indices de détection habituels, comme le test de Hotelling (T^2) et l'erreur quadratique de prédiction (SPE) [4] [5].

Le but de ce travail est d'élaborer à partir des données un modèle analytique de la DKL pour l'estimation de dé-

faut naissant. Nous supposons que les données suivent une distribution Gamma. Cette hypothèse n'est pas trop restrictive puisque cette distribution englobe les distributions Gaussienne et χ^2 que l'on retrouve dans beaucoup d'applications.

Dans ce travail, nous allons comparer l'estimation de la DKL obtenue par le modèle analytique avec la DKL approximée par la simulation de Monte-Carlo ($D\hat{K}L$) en tenant compte de l'influence du bruit et de la taille de l'échantillon de données considéré.

2 Définitions

2.1 Notations

Considérons la matrice $X_{[N \times m]}$ des données collectées sur un système à contrôler, tel que $X = [x_1, \dots, x_i, \dots, x_m]$ où $x_i = [x_{1i} \dots x_{Ni}]'$ est le vecteur colonne de N mesures sur la $i^{\text{ème}}$ variable. Soit S la matrice de covariance des données, $P_{[m \times l]}$ la matrice des vecteurs propres générant le sous espace principal de dimension l . Soient $\lambda_1, \dots, \lambda_l$ telles que $\lambda_1 > \lambda_2 > \dots > \lambda_l$ les valeurs propres associées à P . Considérons la matrice des composantes principales $T_{[N \times l]}$, telle que $T = [t_1, \dots, t_l] = X \times P$.

Le bruit considéré dans le système est un bruit blanc Gaussien additif ($v \sim \mathcal{N}(0, \sigma_v^2)$). On suppose que le bruit ici représente un changement d'environnement dans le système considéré. Du coup, le bruit est ajouté seulement sur les signaux à tester et non sur les signaux de référence qui sont considérés comme sains et sans défauts. On peut donc écrire :

$$x_i = x_i^* + v_i \quad (1)$$

Où x_i^* représente les signaux de référence qui ont une distribution Gamma ($x_i^* \sim \Gamma(\alpha_i^*, \beta_i^*)$). Alors, les variables bruitées x_i ont également une distribution Gamma telle que ($x_i \sim \Gamma(\alpha_i, \beta_i)$) avec

$$\alpha_i = \frac{(\alpha_i^* \beta_i^*)^2}{\alpha_i^* (\beta_i^*)^2 + \sigma_v^2} \quad \text{et} \quad \beta_i = \frac{\alpha_i^* (\beta_i^*)^2 + \sigma_v^2}{\alpha_i^* \beta_i^*}.$$

On suppose que le défaut dans notre cas affecte une seule variable x_j . On modélise ce défaut comme un biais additif caractérisé par une amplitude a (défaut de gain) tel que :

$$x_j = (x_j^* + v_j) + a \times (x_j^* + v_j) \quad (2)$$

2.2 La divergence de Kullback-Leibler

Dans notre étude, nous nous intéressons à la divergence de Kullback-Leibler (DKL) entre deux distributions de probabilité continues $f(x)$ et $g(x)$ d'une variable aléatoire x . Elle est définie comme étant la version symétrique de l'information de Kullback-Leibler I entre ces deux distributions [6] :

$$I(f||g) = \int f(w) \log \frac{f(w)}{g(w)} dw. \quad (3)$$

$$DKL(f, g) = I(f||g) + I(g||f). \quad (4)$$

Dans le cas de distributions Gamma, $f \sim \Gamma(\alpha_1, \beta_1)$ et $g \sim \Gamma(\alpha_2, \beta_2)$, la DKL [3] s'écrit :

$$DKL(\alpha_1, \beta_1; \alpha_2, \beta_2) = (\alpha_1 - \alpha_2)\psi(\alpha_1) - \log \Gamma(\alpha_1) + \log \Gamma(\alpha_2) + \alpha_2(\log \beta_2 - \log \beta_1) + \alpha_1 \frac{\beta_1 - \beta_2}{\beta_2} \quad (5)$$

où $\psi(\alpha)$ est la fonction digamma.

3 Détection et estimation de défaut

La procédure générale de surveillance statistique est de rassembler un grand nombre d'échantillons de données sains utilisés comme l'ensemble de données de référence. Toutes les nouvelles données mesurées sont alors comparées aux signaux sains pour vérifier si un comportement anormal s'est produit.

Dans notre cas, une fois que le modèle de l'ACP est établi, une distribution de probabilité de référence est évaluée pour chaque composante principale. Alors pour chaque nouvel ensemble d'observations, les composantes principales associées sont calculées par le modèle de l'ACP et leur distribution de probabilité est évaluée. Ensuite, la DKL est utilisée pour mesurer les dissimilarités entre les fonctions de densité de probabilité des données saines et des données mesurées. Les étapes principales de la méthode sont résumées dans la figure 1.

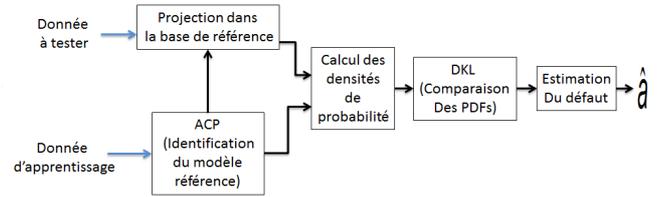


FIGURE 1 – Schéma de diagnostic utilisant la DKL

3.1 Modèle théorique

Si on admet que les vecteurs de mesures suivent une distribution Gamma, on peut obtenir une expression analytique de la divergence. En effet, chacune des l composantes principales t_k , ($k = 1, 2, \dots, l$), qui est une combinaison linéaire des variables originales, suit aussi une distribution Gamma que nous notons f de telle sorte que $f \sim \Gamma(\alpha_k, \beta_k)$. On propose alors de comparer f avec sa référence saine notée $f^* \sim \Gamma(\alpha_k^*, \beta_k^*)$

- Calculons les paramètres α_k^*, β_k^* de f^* :

$$t_k^* = X_{[N \times m]}^* \times p_{k[m \times 1]} \quad (6)$$

où p_k est le $k^{\text{ème}}$ vecteur propre de la matrice X^* .

$$t_k^* = p_{1k}x_1^* + p_{2k}x_2^* + \dots + p_{ik}x_i^* + \dots + p_{mk}x_m^* \quad (7)$$

On a :

$$x_i^* \sim \Gamma(\alpha_i^*, \beta_i^*) \implies p_{ik} x_i^* \sim \Gamma(\alpha_i^*, p_{ik} \times \beta_i^*)$$

$$t_k^* = \sum_{i=1}^{i=m} p_{ik} x_i^* \sim \Gamma(\alpha_k^*, \beta_k^*)$$

Nous obtenons alors une somme de variables aléatoires Gamma. Ainsi, nous pouvons approcher la somme par une distribution Gamma $\Gamma(\alpha_k^*, \beta_k^*)$ où α_k^* et β_k^* sont obtenus par l'identification des 2 premiers moments statistiques (moyenne et variance) [7] :

$$\alpha_k^* = \frac{\mu_{t_k^*}^2}{\sigma_{t_k^*}^2} \quad \beta_k^* = \frac{\sigma_{t_k^*}^2}{\mu_{t_k^*}} \quad (8)$$

Pour la moyenne et la variance de la composante principale t_k^* , nous calculons :

$$\mu_{t_k^*} = \mathbb{E}\left(\sum_{i=1}^{i=m} p_{ik} x_i^*\right) = \sum_{i=1}^{i=m} \mathbb{E}(p_{ik} x_i^*) = \sum_{i=1}^{i=m} p_{ik} \alpha_i^* \beta_i^* \quad (9)$$

$$\sigma_{t_k^*}^2 = \mathbb{E}\left(\left(\sum_{i=1}^{i=m} p_{ik} x_i^*\right) - \mu_{t_k^*}\right)^2 \quad (10)$$

En s'appuyant sur les propriétés de la covariance, l'expression de $\sigma_{t_k^*}^2$ est :

$$\sigma_{t_k^*}^2 = \sum_{i=1}^{i=m} \alpha_i^* p_{ik}^2 (\beta_i^*)^2 + 2 \sum_{i=1}^{i=m-1} \sum_{q=i+1}^{q=m} p_{ik} p_{iq} S_{iq} \quad (11)$$

où S_{iq} est un élément de la matrice de covariance S . Enfin, on obtient :

$$\alpha_k^* = \frac{(\sum_{i=1}^{i=m} \alpha_i^* p_{ik} \beta_i^*)^2}{\sum_{i=1}^{i=m} \alpha_i^* p_{ik}^2 (\beta_i^*)^2 + 2 \sum_{i=1}^{i=m-1} \sum_{q=i+1}^{q=m} p_{ik} p_{iq} S_{iq}} \quad (12)$$

$$\beta_k^* = \frac{\sum_{i=1}^{i=m} \alpha_i^* p_{ik}^2 (\beta_i^*)^2 + 2 \sum_{i=1}^{i=m-1} \sum_{q=i+1}^{q=m} p_{ik} p_{iq} S_{iq}}{\sum_{i=1}^{i=m} \alpha_i^* p_{ik} \beta_i^*} \quad (13)$$

qui sont les paramètres de la distribution des signaux de référence sains avec :

$$\alpha_i^* = \frac{\mu_{x_i^*}^2}{\sigma_{x_i^*}^2} = \frac{(\frac{1}{N} \sum_{r=1}^N x_{ri}^*)^2}{\frac{1}{N} \sum_{r=1}^N (x_{ri}^* - \frac{1}{N} \sum_{r=1}^N x_{ri}^*)^2} \quad (14)$$

$$\beta_i^* = \frac{\sigma_{x_i^*}^2}{\mu_{x_i^*}} = \frac{\frac{1}{N} \sum_{r=1}^N (x_{ri}^* - \frac{1}{N} \sum_{r=1}^N x_{ri}^*)^2}{\frac{1}{N} \sum_{r=1}^N x_{ri}^*} \quad (15)$$

• Calculons les paramètres α_k, β_k de f :

Nous rappelons ici que le défaut n'affecte qu'une variable x_j comme décrit par l'équation (2). Si la variable bruitée $(x_j^* + v_j) \sim \Gamma(\alpha_j, \beta_j)$, alors la variable en défaut résultante est $x_j \sim \Gamma(\alpha_j, (1+a) \times \beta_j)$.

En se basant sur l'approximation de la somme des variables aléatoires Gamma, les expressions de α_k et β_k sont obtenues par les équations (16) et (17) pour lesquelles on note :

$$\alpha_i = \frac{\mu_{x_i}^2}{\sigma_{x_i}^2} = \frac{(\frac{1}{N} \sum_{r=1}^N x_{ri}^*)^2}{\frac{1}{N} \sum_{r=1}^N (x_{ri}^* - \frac{1}{N} \sum_{r=1}^N x_{ri}^*)^2 + \sigma_v^2} \quad (18)$$

$$\beta_i = \frac{\sigma_{x_i}^2}{\mu_{x_i}} = \frac{\frac{1}{N} \sum_{r=1}^N (x_{ri}^* - \frac{1}{N} \sum_{r=1}^N x_{ri}^*)^2 + \sigma_v^2}{\frac{1}{N} \sum_{r=1}^N x_{ri}^*} \quad (19)$$

Définissons les termes suivants :

$$\delta_y = \sum_{i=1}^{i=m} \alpha_i p_{ik} \beta_i \quad \delta_y^* = \sum_{i=1}^{i=m} \alpha_i^* p_{ik} \beta_i^* \quad (20)$$

$$\delta_w = \sum_{i=1}^{i=m} \alpha_i p_{ik}^2 (\beta_i)^2 \quad \delta_w^* = \sum_{i=1}^{i=m} \alpha_i^* p_{ik}^2 (\beta_i^*)^2$$

$$\theta_y = \alpha_j p_{jk} \beta_j \quad \theta_w = \alpha_j p_{jk}^2 (\beta_j)^2$$

On obtient donc le modèle de la DKL tel que décrit dans l'équation (21).

Ceci nous permet alors d'écrire l'estimation théorique de l'amplitude du défaut (sévérité) qui dépend de la valeur de la divergence :

$$\hat{a} = \frac{\theta_y - 2\theta_w + \sqrt{(2\theta_w - \theta_y)^2 - 4(\theta_w \delta_w - \xi \delta_y)}}{2\theta_w} \quad (22)$$

où

$$\xi = \frac{\beta_k^*}{e^{-\sqrt{\frac{2D_{\hat{K}L}}{\alpha_k^*}}}} \quad (23)$$

3.2 Validation du modèle

Considérons un exemple numérique décrivant le fonctionnement d'un système avec $m = 7$ variables de distribution Gamma, définies à l'instant i par les équations suivantes :

$$\begin{aligned} x_1(i) &\propto \Gamma(1, 2), & x_2(i) &\propto \Gamma(4, 5), & x_3(i) &\propto \Gamma(1.5, 3) \\ x_4(i) &= 2 \times x_1(i), & x_5(i) &\propto \Gamma(3, 2), & x_6(i) &= 3 \times x_2(i) \\ x_7(i) &\propto \Gamma(7, 5) \end{aligned}$$

La projection de la matrice de données X formée de N observations conduit à une seule composante principale t_1^* , $t_1^* = X \times p_1^*$, dont la direction est déterminée par le vecteur propre $p_1^* = [0.138 \quad -0.138 \quad 0.2761 \quad 0.5521 \quad 0.2070 \quad 0.6901 \quad 0.2484]$ et de variance $\lambda_1^* = 998$.

Nous représentons sur la figure 2 l'évolution de la DKL obtenue avec le modèle théorique et celle estimée par simulation de Monte-Carlo ($D\hat{K}L$) en fonction de l'amplitude du défaut a . Les performances obtenues avec ce modèle sont satisfaisantes pour les défauts naissants avec un faible niveau de bruit. Néanmoins, plus la puissance du bruit est importante, plus la différence entre la DKL estimée et celle obtenue avec le modèle analytique sera grande.

La figure 3 montre l'erreur relative $\epsilon_{DKL} = \frac{D\hat{K}L - DKL}{D\hat{K}L}$ entre le modèle analytique et la DKL estimée ($D\hat{K}L$) pour des niveaux de bruit et un nombre d'échantillons différents. On peut noter que dans des conditions de faible niveau de bruit ($SNR \geq 40dB$), le modèle est précis avec une erreur relative inférieure à 10% indépendamment du nombre d'échantillon.

$$\alpha_k = \frac{((\sum_{i=1}^{j-1} \alpha_i p_{ik} \beta_i) + \alpha_j p_{jk} (1+a) \beta_j + (\sum_{i=j+1}^m \alpha_i p_{ik} \beta_i))^2}{(\sum_{i=1}^{j-1} \alpha_i p_{ik}^2 \beta_i^2) + \alpha_j p_{jk}^2 (1+a)^2 (\beta_j)^2 + (\sum_{i=j+1}^m \alpha_i p_{ik}^2 \beta_i^2) + 2 \sum_{i=1}^{m-1} \sum_{q=i+1}^m p_{ik} p_{qk} S_{iq}} \quad (16)$$

$$\beta_k = \frac{(\sum_{i=1}^{j-1} \alpha_i p_{ik}^2 \beta_i^2) + \alpha_j p_{jk}^2 (1+a)^2 (\beta_j)^2 + (\sum_{i=j+1}^m \alpha_i p_{ik}^2 \beta_i^2) + 2 \sum_{i=1}^{m-1} \sum_{q=i+1}^m p_{ik} p_{qk} S_{iq}}{(\sum_{i=1}^{j-1} \alpha_i p_{ik} \beta_i) + \alpha_j p_{jk} (1+a) \beta_j + (\sum_{i=j+1}^m \alpha_i p_{ik} \beta_i)} \quad (17)$$

$$DKL = \left(\frac{\delta_y^{*2}}{\delta_w^*} - \frac{(\delta_y + \theta_y a)^2}{\delta_w + \theta_w a^2 + \theta_w 2a} \right) \psi \left(\frac{\delta_y^{*2}}{\delta_w^*} \right) - \log \Gamma \left(\frac{\delta_y^{*2}}{\delta_w^*} \right) + \log \Gamma \left(\frac{(\delta_y + \theta_y a)^2}{\delta_w + \theta_w a^2 + \theta_w 2a} \right) + \frac{(\delta_y + \theta_y a)^2}{\delta_w + \theta_w a^2 + \theta_w 2a} \left(\log \left(\frac{\delta_w + \theta_w a^2 + \theta_w 2a}{\delta_y + \theta_y a} \right) - \log \left(\frac{\delta_w^*}{\delta_y^*} \right) \right) + \frac{\delta_y^{*2}}{\delta_w^*} \left(\frac{\delta_w^* (\delta_y + \theta_y a)}{\delta_y^* (\delta_w + \theta_w a^2 + 2\theta_w a)} - 1 \right) \quad (21)$$

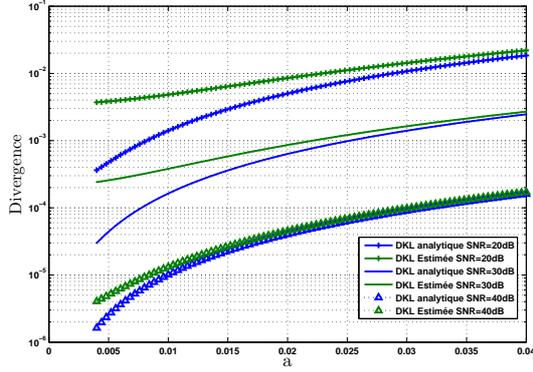


FIGURE 2 – Comparaison de la DKL avec la \widehat{DKL}

4 Conclusion

Dans le processus de diagnostic de défaut, la détection puis l'estimation de la sévérité de défaut est cruciale pour permettre une prise de décision appropriée. Nous avons développé ici un modèle analytique de la DKL dans l'espace de l'ACP en considérant des signaux à distribution Gamma. Le modèle obtenu a été validé par comparaison avec l'estimation obtenue en utilisant la simulation Monte-Carlo. Ce modèle a été évalué pour un défaut de type gain, en fonction du nombre d'échantillons et du bruit. Pour le cas de défaut naissant (faible amplitude et SNR important), le modèle permet une surestimation de la sévérité du défaut avec une erreur relative inférieure à 10%.

Références

- [1] I.T. Jolliffe, „ in *Principle Component Analysis*. Second Edition, Aberdeen U.K, 2002.
- [2] J. F. MacGregor and T. Kourti, “Statistical process control of multivariate processes,” *Control Engineering Practice*, vol. 3, no. 3, pp. 403–414, 1995.

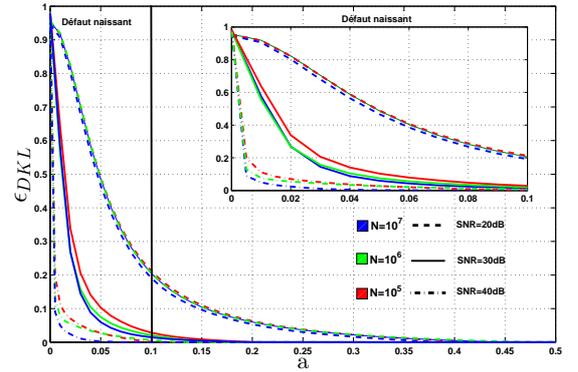


FIGURE 3 – Evolution de l'erreur relative ϵ_{DKL} en fonction de la sévérité du défaut

- [3] S. Kullback and R.A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [4] J. Harmouche, C. Delpha, and D. Diallo, “A theoretical approach for incipient fault severity assessment using the kullback-leibler divergence,” in *EUSIPCO 2013*, Marrakech, Morocco, September 9-13 2013.
- [5] J. Harmouche, C. Delpha, and D. Diallo, “Incipient fault detection and diagnosis based on kullback-leibler divergence using principal component analysis : Part I,” *Elsevier Journal on Signal Processing*, vol. 94, pp. 278–287, January 2014.
- [6] M. Basseville, “Distances measures for signal processing and pattern recognition,” *Elsevier Signal Processing*, vol. 18, no. 4, pp. 349–369, 1989.
- [7] C. Florent, P. Borgnat, J. Tourneret, and P. Abry, “Parameter estimation for sums of correlated gamma random variables. application to anomaly detection in internet traffic,” in *ICASSP*, Las Vegas US, 2008.