Estimation d'un modèle générique pour un parc de machines

Farah Ankoud [1], Gilles Mourot [1], Roger Chevalier [2], Nicolas Paul [2], José Ragot [1]

***Résumé*—Dans certains domaines industriels, on dispose d'un parc de machines identiques, soumises au même type d'exploitation mais qui fonctionnent dans des conditions différentes. Dans cette communication, on propose une méthode pour déterminer un modèle générique de ce type de machines, s'il existe. Tout d'abord, la méthode procède par construction du modèle linéaire de chaque machine, à partir des données mesurées sur celle-ci et indépendamment les unes des autres. La seconde étape de la méthode consiste à identifier la partie commune aux différents modèles, si elle existe. Dans ce but, l'approche proposée détermine les coefficients des modèles jugés identiques à partir d'un critère de ressemblance. De nouveaux modèles sont ensuite obtenus en imposant des contraintes de type égalité sur ces coefficients. La troisième étape consiste alors à valider le choix de la partie commune et les modèles ainsi obtenus. Cette validation est réalisée par un test statistique basé sur la comparaison des estimations obtenues avec les nouveaux et les anciens modèles. La méthode est illustrée sur un exemple de simulation.**

***Mots clés*—parc de machines, modélisation, estimation de paramètres communs, indice de ressemblance**

I. Introduction

Dans certains secteurs industriels tels que ceux de la production d'énergie, on dispose d'un parc de machines identiques, soumises au même type d'exploitation mais qui peuvent fonctionner dans des conditions différentes (parcs éoliens, parcs nucléaires, etc.). Le problème à résoudre est de déterminer s'il est possible d'établir un modèle générique de fonctionnement de ce type de machines à partir des mesures réalisées sur les différentes machines. C'est un problème d'apprentissage simultané de modèles plus connus sous le vocable multi-task learning en anglais.

Deux hypothèses peuvent être faites sur la manière dont l'environnement peut affecter le fonctionnement des machines et donc la structure de ce modèle. La première est une hypothèse multiplicative dans le sens où les variables d'environnement n'interviennent pas de façon explicite comme variables explicatives dans le modèle recherché mais affectent dans le modèle uniquement les coefficients des variables propres aux machines. Une formulation mathématique possible de cette hypothèse est proposée dans [4] où les modèles ont la même structure mais des coefficients qui varient autour d'une valeur moyenne selon les conditions d'environnement dans lesquelles se trouve chaque machine. Cependant, ce formalisme ne permet pas d'établir le lien entre les variations des paramètres et les variables d'environnement.

Par contre, la seconde hypothèse est plutôt de type additif dans le sens où les variables d'environnement apparaissent comme étant des variables explicatives qui viennent se rajouter aux variables propres aux machines pour former le modèle. De plus, on peut assez logiquement supposer que les variables d'environnement ont un pouvoir explicatif moindre que les variables propres à chaque machine. On constate également que, contrairement à l'hypothèse précédente, dans le cas additif, les modèles peuvent avoir une structure différente. En résumé et compte tenu de cette hypothèse, on peut définir qu'un modèle générique d'une machine est composé d'une partie commune constituée des variables propres à la machine (structure et paramètres identiques) et d'une partie distincte liée aux variables d'environnement. Concernant cette dernière partie, on peut distinguer les variables d'environnement agissant sur le comportement de toutes les machines (coefficients du modèle différents suivant les machines) de celles qui affectent le comportement d'au moins une machine. Lors de l'ajout d'une machine nouvelle, l'établissement de son modèle est alors grandement simplifié puisque, bénéficiant du modèle générique, il suffit de réadapter la partie liée aux variables traduisant les effets de l'environnement local.

Les travaux [1], [10] présentent une méthode pour identifier à partir de toutes les données disponibles sur les différentes machines, à la fois la structure et les coefficients des modèles de régression linéaire en utilisant une méthode de type LASSO. Cependant, avec cette approche, la partie commune des modèles obtenus n'est pas mise en évidence. Les travaux de [6] se situent dans le cadre de l'estimation des coefficients de deux modèles linéaires ayant la même structure et pour lesquels tous les coefficients sont supposés communs. Dans [8], connaissant a priori les termes communs à deux modèles, les auteurs présentent une méthode d'estimation des paramètres de ces modèles. Dans [11], les auteurs traitent le problème d'estimation des paramètres des modèles ayant une partie commune ainsi que le problème de partition en classes de ces modèles, la partie commune aux modèles étant connue a priori.

Les travaux existants dans le cadre de l'estimation des paramètres des modèles linéaires tenant compte de l'existence d'une partie commune partent tous de l'hypothèse que cette dernière est connue a priori. Dans cette communication, on propose à la section II une méthode qui détermine la partie commune aux différents modèles. A la section III, on donne les résultats de l'application de la méthode proposée sur un exemple de simulation. On termine ce papier par des conclusions et des perspectives.

F. Ankoud, G. Mourot, J. Ragot : Centre de Recherche en Automatique de Nancy, UMR 7039 - Nancy-Université, CNRS, 2, avenue de la Forêt de Haye, 54516 Vandoeuvre-les-Nancy Cedex France farah.ankoud, gilles.mourot, jose.ragot@ensem.inpl-nancy.fr

R. Chevalier, N. Paul : Electricité de France (EDF-R&D), 6 quai Watier, 78401 Chatou Cedex France roger.chevalier, nicolas.paul@edf.fr

Symbole	Signification	Dimension
B.D.D	Base de données	
K	Nombre des B.D.D	
k	Numéro de la B.D.D	
m	Nombre de variables mesurées	
n_k	Nombre d'observations dans la $k^{ème}$ B.D.D	
y^k	Variable à estimer dans la $k^{ème}$ B.D.D	n_k
p_k	Nombre de variables explicatives constituant le modèle pour estimer y^k	
W^k	Matrice des variables pouvant contribuer au modèle pour estimer y^k	$n_k \times m$
S^k	Matrice de sélection des variables contribuant au modèle pour estimer y^k	$m \times p_k$
X^k	Matrice contenant les p_k variables contribuant au modèle pour estimer y^k	$n_k \times p_k$
x_i^k	Colonne numéro i de X^k	n_k
$\hat{c}_i^k$	Coefficient de x_i^k dans le modèle estimant y^k	
$\hat{c}^k$	Vecteur regroupant les coefficients $\hat{c}_i^k$ $(i = 1, \ldots, p_k)$	p_k
$\hat{\sigma}_i^k$	Ecart-type estimé du coefficient $\hat{c}_i^k$	
p	Nombre de coefficients identiques	
S_p^k	Matrice de sélection des variables de X^k à coefficients identiques	$p_k \times p$
U^k	Matrice formée par les variables de X^k à coefficients identiques	$n_k \times p$
$S_{\overline{p}}^k$	Matrice de sélection des variables de X^k à coefficients non identiques	$p_k \times (p_k - p)$
V^k	Matrice formée par les variables de X^k à coefficients non identiques	$n_k \times (p_k - p)$

TABLE I
NOTATIONS ET DIMENSIONS DES VARIABLES

II. DÉTERMINATION DE LA PARTIE COMMUNE DES MODÈLES

Avant de décrire notre approche, on donne à la table I la liste des notations utilisées et les dimensions des différents vecteurs et matrices. On se situe dans le cas général où l'on ne dispose pas des équations physiques décrivant le comportement du système mais pour lequel on peut trouver des modèles linéaires à partir des mesures. La première étape de la méthode proposée consiste à trouver les différents modèles linéaires à partir des données collectées sur chaque machine indépendamment les unes des autres. C'est un problème classique d'identification de modèles linéaires pour lequel il faut déterminer la structure et les paramètres du modèle à partir des données disponibles, puis valider le modèle ainsi obtenu (le lecteur peut se référer à [9], [3], [12] pour plus de détails sur cette partie). Dans une deuxième étape, les coefficients des modèles potentiellement identiques sont identifiés en faisant une analyse de ressemblance. De nouveaux modèles sont ensuite obtenus en imposant des contraintes de type égalité sur ces coefficients. La troisième étape consiste alors à valider le choix de la partie commune et les modèles ainsi obtenus. Cette validation est réalisée par un test statistique basé sur la comparaison des estimations obtenues avec les nouveaux et les anciens modèles.

A. Détermination des coefficients identiques

On suppose qu'on dispose de K bases de données contenant chacune n_k $(k = 1, \ldots, K)$ mesures d'une variable y^k à estimer à partir d'un ensemble de m variables regroupées dans une matrice W^k. Dans la $k^{ème}$ base de données, l'estimation de y^k à partir des éléments de W^k peut s'écrire sous la forme :

$$\hat{y}^k = X^k \hat{c}^k \tag{1}$$

avec :

$$X^k = W^k S^k \tag{2}$$

$\hat{y}^k$ est l'estimé de y^k dans la $k^{ème}$ base de données. S^k est une matrice de sélection qui permet de choisir dans W^k les colonnes associées aux variables explicatives contribuant à estimer y^k. La matrice S^k est constituée de 0 et de 1, où 1 indique les variables à sélectionner. Par exemple, pour sélectionner les variables numéro 2 et 4 parmi $m = 5$ variables, la matrice S^k est définie par :

$$S^k = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}^T$$

$\hat{c}^k$ est l'estimé du vecteur des coefficients c^k du modèle donné par :

$$\hat{c}^k = ({X^k}^T X^k)^{-1} {X^k}^T y^k \tag{3}$$

sous la condition que la matrice ${X^k}^T X^k$ soit inversible c'est-à-dire que la matrice X^k soit de plein rang colonne (il faut donc que toutes les colonnes de X^k soient non nulles et qu'il n'y ait pas deux vecteurs colinéaires).
On cherche à trouver la partie commune aux modèles estimant la variable y^k dans toutes les bases de données. En d'autres termes on cherche à trouver les variables x_i^k ayant sensiblement le même pouvoir explicatif. On propose alors de considérer les coefficients c_i^k $(\forall \quad k)$ identiques si une intersection non vide existe entre leurs intervalles de confiance notés I_i^k. On sait que l'intervalle de confiance à 99% du coefficient $\hat{c}_i^k$ peut être obtenu par [7] :

$$I_i^k = [\hat{c}_i^k - 2.32\hat{\sigma}_i^k ; \hat{c}_i^k + 2.32\hat{\sigma}_i^k] \tag{4}$$

l'écart-type $\hat{\sigma}_i^k$ étant estimé à partir des données et du modèle construit. En fait, après avoir estimé les différents paramètres du modèle, on calcule le vecteur résidu du modèle selon :

$$e^k = y^k - \hat{y}^k \tag{5}$$

La variance de ce résidu peut être estimée par :

$$\hat{\sigma}_{e^k}^2 = \frac{1}{n_k - p_k} {e^k}^T e^k \tag{6}$$

La matrice de variance-covariance du vecteur $\hat{c}^k$ peut être estimée par :

$$\hat{\Sigma}_c^k = \hat{\sigma}^2_{\epsilon^k}(X^{k^T}X^k)^{-1} \tag{7}$$

Finalement, l'estimé de la variance $\hat{\sigma}_i^{k^2}$ d'un coefficient $\hat{c}_i^k$ est le $i^{ème}$ terme de la diagonale de $\hat{\Sigma}_c^k$.
Pour simplifier les notations, on note $I_i^k = [BI_i^k; BS_i^k]$.
L'intersection des intervalles de confiance I_i^k ($\forall\ k$) se fait de la façon suivante :
Procédure 1

- on calcule $BI_i = \max\limits_k (BI_i^k)$ et $BS_i = \min\limits_k (BS_i^k)$
- si $BI_i < BS_i$, l'intersection des intervalles de confiance est égale à l'intervalle $[BI_i; BS_i]$; sinon l'intersection est vide

Les coefficients pour lesquels il existe une intersection non vide de leurs intervalles de confiance sont supposés potentiellement identiques dans toutes les bases de données.

B. Identification des nouveaux paramètres tenant compte de la partie commune

Une fois identifiée la partie commune aux modèles, on fait une estimation des paramètres du modèle estimant y^k dans chaque base de données en imposant l'égalité des coefficients jugés identiques (soit p le nombre de ces coefficients). Pour cela, on partitionne les variables de chaque matrice X^k en deux matrices notées U^k et V^k telles que :

$$U^k = X^k S_p^k \qquad et \qquad V^k = X^k S_{\overline{p}}^k$$

où $S_p^k \in \mathbb{R}^{p_k \times p}$ est la matrice de sélection des p variables de X^k formant la partie commune au modèle estimant y^k dans toutes les bases de données. $S_{\overline{p}}^k \in \mathbb{R}^{p_k \times (p_k - p)}$ permet la sélection des variables de X^k pour lesquelles le coefficient n'est pas le même dans toutes les bases de données.
Le problème d'identification des nouveaux paramètres sous contraintes de type égalité sur les coefficients jugés identiques peut être résolu par une méthode de moindres carrés classique en écrivant :

$$\underbrace{\begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^K \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} U^1 & V^1 & 0 & \cdots & 0 \\ U^2 & 0 & V^2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ U^K & 0 & \cdots & 0 & V^K \end{bmatrix}}_{Z} \underbrace{\begin{bmatrix} \alpha \\ \beta^1 \\ \beta^2 \\ \vdots \\ \beta^K \end{bmatrix}}_{\theta} \tag{8}$$

α correspond au vecteur des p coefficients jugés identiques dans toutes les bases de données, β^k est le vecteur des $p_k - p$ coefficients restants.
Le vecteur θ peut être estimé par :

$$\hat{\theta} = \left(Z^T Z\right)^{-1} Z^T Y \tag{9}$$

N.B : La matrice $Z^T Z$ est inversible car la matrice Z est de plein rang colonnes, cette dernière étant construite à partir des colonnes des matrices X^k qui ont été supposées indépendantes.

C. Critère pour la validation du choix des coefficients identiques

De nouvelles estimations des vecteurs y^k ($k = 1, \ldots, K$) sont alors disponibles. Ces estimations sont données par :

$$\tilde{y}^k = \begin{bmatrix} U^k & V^k \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}^k \end{bmatrix} \tag{10}$$

Les nouveaux vecteurs résidus sont donnés par :

$$\tilde{e}^k = y^k - \tilde{y}^k \tag{11}$$

Afin de valider notre choix de la partie considérée commune à tous les modèles, on peut utiliser un critère s'appuyant sur les statistiques définies selon les formules (12) et (13) (voir [7]) :

$$\mathscr{L} = \sum_{k=1}^{K} e^{kT} e^k \tag{12}$$

$$\tilde{\mathscr{L}} = \sum_{k=1}^{K} \tilde{e}^{kT} \tilde{e}^k \tag{13}$$

$\frac{1}{\sigma^2}\mathscr{L}$ suit une loi χ^2 à $N-P$ degrés de liberté où $N = \sum\limits_{k=1}^{K} n_k$ est le nombre total d'observations disponibles, $P = \sum\limits_{k=1}^{K} p_k$ est le nombre de variables explicatives constituant les modèles dans toutes les bases de données et σ^2 est la variance résiduelle tenant compte de toutes les données. $\frac{1}{\sigma^2}(\tilde{\mathscr{L}} - \mathscr{L})$ suit une loi χ^2 à $(K-1)p$ degrés de liberté.
On peut en déduire que :

$$\frac{N-P}{(K-1)p} \cdot \frac{\tilde{\mathscr{L}} - \mathscr{L}}{\mathscr{L}} \sim \mathscr{F}_a((K-1)p, N-P) \tag{14}$$

où $\mathscr{F}_a((K-1)p, N-P)$ désigne la loi de Fisher pour un seuil de confiance $(1-a)$ avec $((K-1)p, N-P)$ degrés de liberté. Le critère de validation consiste à accepter l'hypothèse qu'il s'agit bien de la partie commune à tous les modèles si :

$$\tilde{\mathscr{L}} \leq (1 + \frac{(K-1)p}{N-P} F_a)\mathscr{L} \tag{15}$$

F_a est une valeur qu'on peut déterminer selon le seuil de confiance souhaité.

D. Algorithme

Il comporte les étapes suivantes :

1) Identifier le modèle permettant d'estimer la variable y^k dans chacune des bases de données indépendamment les unes des autres
2) Trouver les coefficients potentiellement identiques en faisant l'intersection de leurs intervalles de confiance sur toutes les bases de données (équation (4) et Procédure 1)
3) Estimer les paramètres du modèle dans chaque base de données en imposant des contraintes d'égalité sur les coefficients jugés identiques (équations (8) à (10))
4) Comparer la somme des critères résiduels issus des deux estimations pour confirmer ou non la présence de partie commune du modèle sur toutes les bases de données (équation (15))

III. Application

On a généré 3 modèles et 3 bases de données à 250 observations chacune selon :

$$\begin{aligned} y^1 &= x_1^1 + 5x_3^1 + 5.5x_5^1 - 10 + \varepsilon^1 \\ y^2 &= x_2^2 + 5x_3^2 + 0.6x_4^2 + 5.68x_5^2 - 12 + \varepsilon^2 \\ y^3 &= 0.5x_1^3 + 1.2x_2^3 + 5.1x_3^3 + 0.7x_4^3 + 5.3x_5^3 - 14 + \varepsilon^3 \end{aligned} \quad (16)$$

Les variables x_i^k sont générées comme étant des signaux de télégraphiste filtrés dont les valeurs ont toutes le même ordre de grandeur. Ainsi, dans l'exemple donné aux équations (16), comme les coefficients des variables x_3^k et x_5^k $(k = 1, \ldots, 3)$ sont les plus grands, ce sont ces variables qui ont le plus grand pouvoir explicatif de y^k. Les variables ε^k sont les vecteurs associés au bruit de mesure. On les a supposé de moyenne nulle et de variance respective proportionnelle à l'étendue de la variable y^{*k}, cette dernière correspondant à la valeur exacte de y^k en absence de bruit de mesure ($y^{*k} = y^k - \varepsilon^k$).
Les variables y^k sont obtenues après avoir additionné aux vecteurs y^{*k} des vecteurs ε^k ayant chacun une variance égale à 10% de l'étendue de y^{*k}.
La figure 1 donne les tracés des signaux de la première base de données utilisés dans les simulations.

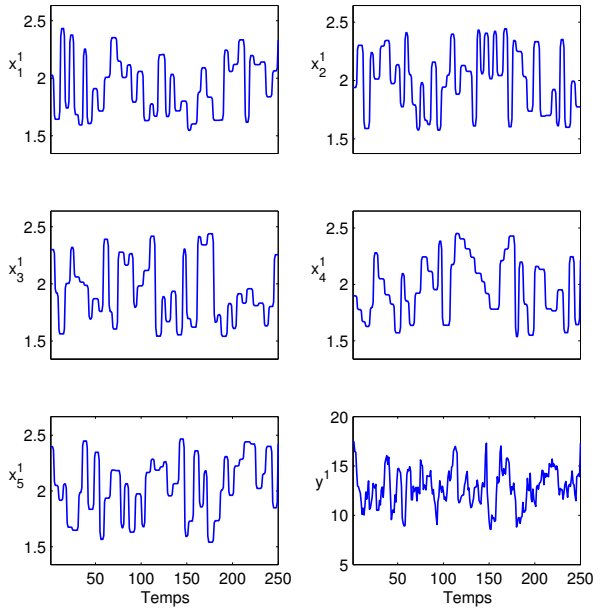

Figure 1. Les signaux de la première base de données

On cherche à identifier les paramètres communs aux trois modèles dans les trois bases de données. Dans l'exemple présenté, les coefficients des variables x_3^k et x_5^k sont quasiment identiques dans les trois modèles : ce sont ces coefficients qui seront jugés identiques d'après la méthode proposée à la section II comme on le démontrera dans les paragraphes suivants.

A. Etape 1 : Identification des paramètres du modèle sur chaque base de données

La première étape de la méthode consiste à estimer les coefficients pour le modèle de y^k à partir de chaque base de données indépendamment des autres. Les modèles choisis étant linéaires, on a estimé les coefficients en question à l'aide de la méthode des moindres carrés classique. Les expressions des estimés des y^k obtenues sont :

$$\begin{aligned} \hat{y}^1 &= 0.89x_1^1 + 5.09x_3^1 + 5.66x_5^1 - 10.29 \\ \hat{y}^2 &= 1.15x_2^2 + 4.89x_3^2 + 0.54x_4^2 + 5.40x_5^2 - 11.52 \\ \hat{y}^3 &= 0.58x_1^3 + 1.18x_2^3 + 4.98x_3^3 + 0.66x_4^3 + 5.31x_5^3 - 13.84 \end{aligned} \quad (17)$$

Les tracés des variables y^k et leurs estimés sont donnés à la figure 2.

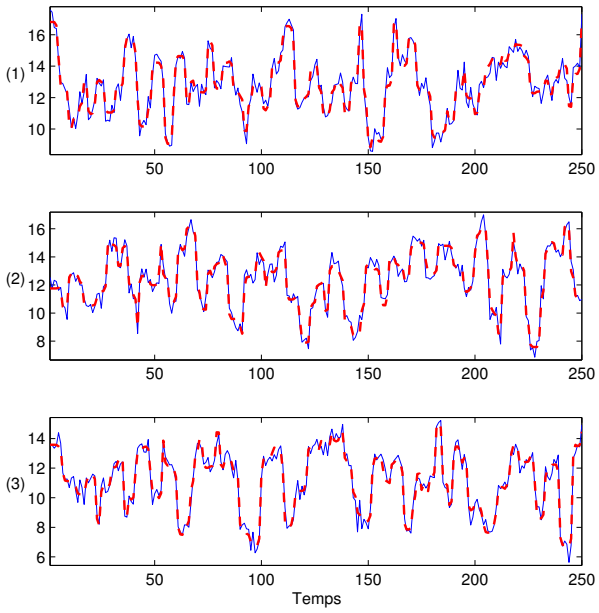

Figure 2. La variable y^k (trait bleu plein) et son estimé $\hat{y}^k$ (trait pointillé rouge) pour chaque base de données

On définit le résidu relatif ε^k et le critère résiduel J^k comme étant :

$$\varepsilon^k = \frac{e^k}{y^k} \quad (18)$$

$$J^k = \frac{1}{n^k} e^{kT} e^k \quad (19)$$

n^k étant le nombre d'observations disponibles pour la $k^{ème}$ base de données.
On résume à la table II les résultats obtenus pour les estimations $\hat{y}^k$.
k est le numéro de la base de données ; e^k_{min}, e^k_{max} et σ^k_e (respectivement ε^k_{min}, ε^k_{max} et σ^k_ε) sont les valeurs minimale, maximale et l'écart-type du vecteur résidu e^k (5) (respectivement du résidu relatif calculé selon la formule (18)) pour la $k^{ème}$ base de données ; $|e|^k_{moy}$ (respectivement $|\varepsilon|^k_{moy}$) est la

k	e^k_{min}	e^k_{max}	$\|e\|^k_{moy}$	σ^k_e	ε^k_{min}	ε^k_{max}	$\|\varepsilon\|^k_{moy}$	σ^k_ε	corr$(y^k,\hat{y}^k)$	J^k
1	-1.34	1.14	0.034	0.43	-0.10	0.09	0.03	0.03	0.97	0.18
2	-1.22	1.29	0.32	0.40	-0.13	0.13	0.03	0.04	0.98	0.16
3	-1.31	1.14	0.32	0.41	-0.18	0.12	0.03	0.04	0.98	0.17

TABLE II
RÉSULTATS OBTENUS POUR CHAQUE MODÈLE $\hat{y}^k$, k= 1,. . .,3

moyenne de la valeur absolue du résidu (respectivement du résidu relatif) ; corr$(y^k,\hat{y}^k)$ correspond à la corrélation entre y^k et son estimé et J^k est le critère résiduel (19).
Les résultats obtenus après estimation des paramètres du modèle pour y^k dans chaque base de données indépendamment des autres sont satisfaisants et les modèles ainsi trouvés sont considérés comme des modèles valides.

B. Etape 2 : Recherche des coefficients identiques

Les intervalles de confiance à 99% des coefficients estimés sont donnés à la table III.

coefficient	intervalle
$\hat{c}^1_0$	[-11.13 ; -9.46]
$\hat{c}^1_1$	[0.62 ; 1.15]
$\hat{c}^1_3$	[4.86 ; 5.34]
$\hat{c}^1_5$	[5.41 ; 5.91]
$\hat{c}^2_0$	[-12.38 ; -10.66]
$\hat{c}^2_2$	[0.89 ; 1.39]
$\hat{c}^2_3$	[4.66 ; 5.12]
$\hat{c}^2_4$	[0.31 ; 0.78]
$\hat{c}^2_5$	[5.19 ; 5.61]
$\hat{c}^3_0$	[-14.87 ; -12.81]
$\hat{c}^3_1$	[0.34 ; 0.81]
$\hat{c}^3_2$	[0.87 ; 1.49]
$\hat{c}^3_3$	[4.73 ; 5.22]
$\hat{c}^3_4$	[0.39 ; 0.93]
$\hat{c}^3_5$	[5.09 ; 5.53]

TABLE III
INTERVALLES DE CONFIANCE DES COEFFICIENTS ESTIMÉS SUR LES DONNÉES DE CHAQUE BASE DE DONNÉES INDÉPENDAMMENT LES UNES DES AUTRES

$\hat{c}^k_i$ $(i = 1,\ldots,5)$ correspond au coefficient de la variable x^k_i et $\hat{c}^k_0$ correspond à la constante du modèle dans la $k^{ème}$ base de données. Une intersection non vide existe entre les intervalles de confiance du coefficient de la variable x^k_3. Elle est égale à [4.86 ; 5.12]. Il en est de même pour les intervalles de confiance de $\hat{c}^k_5$, l'intersection étant égale à [5.41 ; 5.53]. Pour les autres coefficients on ne trouve pas d'intersection entre les intervalles de confiance sur les trois bases de données. En conclusion, les coefficients des variables x^k_3 et x^k_5 sont considérés potentiellement identiques.

C. Etape 3 : Validation du choix des coefficients identiques

On a imposé des contraintes de type égalité sur les coefficients de x^k_3 et x^k_5 et on a ré-estimé les coefficients du modèle estimant y^k. Les expressions des nouveaux estimés de y^k sont donnés par :

$$\begin{aligned}\tilde{y}^1 &= 0.94x^1_1 + 4.98x^1_3 + 5.44x^1_5 - 9.72\\ \tilde{y}^2 &= 1.13x^2_2 + 4.98x^2_3 + 0.54x^2_4 + 5.44x^2_5 - 11.75\\ \tilde{y}^3 &= 0.58x^3_1 + 1.16x^3_2 + 4.98x^3_3 + 0.66x^3_4 + 5.44x^3_5 - 14.09\end{aligned} \quad (20)$$

Les tracés de y^k et $\tilde{y}^k$ donnés à la figure 3 indiquent une bonne qualité d'estimation de y^k ce qui est confirmé par la table IV (la notation tilde est utilisée pour rappeler qu'il s'agit des résultats issus de (20)).

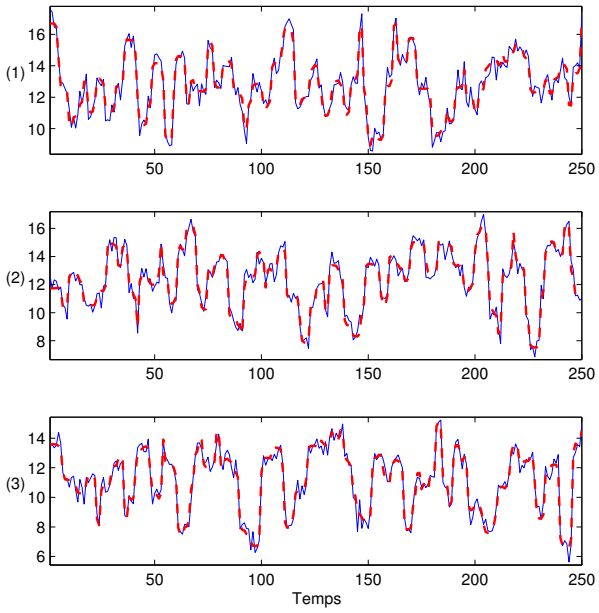

FIGURE 3. La variable y^k (trait bleu plein) et son estimé $\tilde{y}^k$ (trait pointillé rouge) pour chaque base de données

k	$\tilde{e}^k_{min}$	$\tilde{e}^k_{max}$	$\|\tilde{e}\|^k_{moy}$	$\sigma^k_{\tilde{e}}$	$\tilde{\varepsilon}^k_{min}$	$\tilde{\varepsilon}^k_{max}$	$\|\tilde{\varepsilon}\|^k_{moy}$	$\sigma^k_{\tilde{\varepsilon}}$	corr$(y^k,\tilde{y}^k)$	$\tilde{J}^k$
1	-1.31	1.19	0.35	0.43	-0.11	0.09	0.03	0.04	0.97	0.18
2	-1.22	1.34	0.32	0.41	-0.13	0.14	0.03	0.04	0.98	0.16
3	-1.30	1.17	0.32	0.41	-0.17	0.12	0.03	0.04	0.98	0.17

TABLE IV
RÉSULTATS OBTENUS POUR CHAQUE MODÈLE $\tilde{y}^k$, k= 1,. . .,3

Les résultats de la table IV sont quasiment identiques à ceux donnés à la table II.
Les valeurs des variables $\mathscr{L}$ et $\tilde{\mathscr{L}}$ obtenues d'après les équations (12) et (13) sont respectivement : 127.95 et 129.31. Afin de pouvoir confirmer l'hypothèse que les coefficients de x^k_3 et de x^k_5 peuvent être considérés identiques dans toutes les bases de données il faut que $\mathscr{L}$ et $\tilde{\mathscr{L}}$ vérifie (15). Dans notre application : $N = 750$, $P = 15$ (12 variables explicatives et une constante dans le modèle de chaque base de données), $K = 3$ et $p = 2$. Avec un niveau de confiance de 99%, la valeur de $\mathscr{F}_a((K-1)p, N-P)$ est égale à 3.34. Les valeurs $\mathscr{L}$ et $\tilde{\mathscr{L}}$

doivent alors vérifier :

$$\tilde{\mathscr{L}} \leq (1+\frac{2\times 2}{750-15}3.34)\mathscr{L}$$

Il faut donc avoir :

$$\tilde{\mathscr{L}} \leq 130.51$$

Comme la valeur obtenue de $\tilde{\mathscr{L}}$ vérifie la dernière inéquation, le choix de la partie commune aux modèles de y^k dans les différentes bases de données est alors justifié.

IV. Conclusions et perspectives

Dans cette communication, on a proposé une méthode de détermination de la partie commune d'un modèle générique de machine à partir des mesures recueillies sur un parc de machines identiques soumises au même type d'exploitation mais fonctionnant dans des conditions différentes. La première étape de la méthode proposée consiste à trouver les différents modèles linéaires à partir des données collectées sur chaque machine indépendamment les unes des autres. C'est un problème classique d'identification de modèles linéaires pour lequel il faut déterminer la structure et les paramètres du modèle à partir des données disponibles, puis valider le modèle ainsi obtenu. Dans une deuxième étape, les coefficients des modèles potentiellement identiques sont identifiés en faisant une analyse de ressemblance. De nouveaux modèles sont ensuite obtenus en imposant des contraintes d'égalité sur ces coefficients. La troisième étape consiste à valider le choix de la partie commune et les modèles ainsi obtenus. Cette validation est réalisée par un test statistique basé sur la comparaison des estimations obtenues avec les nouveaux et les anciens modèles. La méthode a été appliquée avec succès sur un exemple académique de petites dimensions. Par ailleurs, la méthode proposée peut être utilisée avec les méthodes proposées par [11] et [1] pour déterminer la partie commune aux différents modèles. En perspective, on développera une méthode permettant de concevoir le modèle d'une nouvelle machine ajoutée au parc à partir du modèle générique obtenu et de données issues de cette machine, c'est-à-dire identifier la partie associée à l'environnement (variables d'environnement communes aux différentes machines et ensemble des variables d'environnement propres à chaque machine). De plus, l'apport de ce type de modélisation en termes de diagnostic d'un parc de machines identiques sera aussi étudié. Plus précisément, comme l'approche de diagnostic par redondance analytique [2], [5] repose sur la génération de résidus structurés, on peut profiter de la partie commune pour générer des résidus qui soient insensibles aux variables de cette partie et sensibles aux autres variables du modèle générique. Ainsi, un premier pas vers l'obtention d'un système de diagnostic générique sera franchi. Finalement, on testera la méthode sur des données réelles issues des différentes tranches de centrales nucléaires d'EDF.

Références

[1] A. Argyriou, T. Evgeniou et M. Pontil : Multi-task feature learning. *In Advances in Neural Information Processing Systems 19*, Vancouver, Canada, 2007. MIT Press.

[2] E. Y. Chow et A. S. Willsky : Analytical redundancy and the design of robust failure detection system. *IEEE Transactions on Automatic Control*, 29(7):603 – 614, 1984.

[3] N. R. Draper et H. Smith : *Applied regression analysis*. Wiley Series in Probability and Mathematical Statistics. New York, second édition, 1981.

[4] T. Evgeniou, M. Pontil et O. Toubia : A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. Rapport technique 2006/62/TOM/DS, INSEAD, 2006.

[5] P. M. Frank : Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy : A survey and some new results. *Automatica*, 26(3):459 – 474, 1990.

[6] T. Kubokawa : Double shrinkage estimation of common coefficients in two regression equations with heteroscedasticity. *Journal of Multivariate Analysis*, 67:169–189, 1998.

[7] L. Lebart et J. P. Fénelon : *Statistique et informatique appliquées*. Bordas, Paris, troisième édition, 1975.

[8] A. Liu : Estimation of the parameters in two linear models with only some of the parameter vectors identical. *Statistics & Probability Letters*, 29(4):369 – 375, 1996.

[9] L. Ljung : *System identification : theory for the user*. Prentice-Hall, Inc., New Jersey, 1987.

[10] K. Lounici, M. Pontil, A. B. Tsybakov et S.A. Van de Geer : Taking advantage of sparsity in multi-task learning. *In Conference on Computational Learning Theory*, 2009.

[11] R. Porreca et G. Ferrari-Trecate : Partitioning datasets based on equalities among parameters. *Automatica*, 46(2):460–465, 2010.

[12] R. Tibshirani : Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.