Reconstruction d'écoulements par modélisation réduite et algorithme EM (Expectation-Maximization)

Romain LEROUX[1], Ludovic CHATELLIER[2] et Laurent DAVID[3]

[1,2,3]Institut P', CNRS- Université de Poitiers - ENSMA, UPR 3346 - 11 Boulevard Marie et Pierre Curie - 86962 Futuroscope

1 Introduction

L'acquisition, l'analyse et le stockage de données issues de techniques expérimentales de type Vélocimétrie par Images de Particules conduisent à des coûts mémoire et de calcul toujours plus élevés à mesure que les résolutions spatiales et temporelles des dispositifs d'acquisition progressent. Dans ce contexte, la problématique d'une mesure à la fois statistiquement convergée et capable de résoudre la dynamique spatio-temporelle d'écoulements instationnaires, est posée. Sur la base d'une approche stochastique, utilisant un modèle à espaces d'états, l'estimation de l'état du système dynamique considéré au cours du temps est réalisée en se basant sur les observations, mesurées de manière directe d'une partie des variables d'états. A cet effet, le traitement de séries de données sur une base de temps irrégulière faite de blocs disjoints est effectué. Dans le cas où des observations sont manquantes, l'information utilisée séquentiellement pour l'estimation est incomplète et par conséquent moins fiable. L'estimation liée au modèle à espace d'état devient alors d'avantage prépondérante. La reconstruction des données manquantes s'avère donc nécessaire afin de compenser la réduction du nombre d'observations utilisées pour l'estimation. Le problème de reconstruction de séries temporelles avec des données manquantes est équivalent à l'estimation de paramètres dans les modèles à espaces d'état linéaires à erreurs gaussiennes. La méthode la plus utilisée pour estimer les paramètres d'un modèle à espace d'état à variables latentes est de déterminer l'estimateur du maximum de vraisemblance à l'aide de l'algorithme EM (*Expectation-Maximisation*).

Après avoir défini le modèle à espace d'état tenant compte des observations manquantes, l'expression de la vraisemblance est rappelée. L'algorithme EM, basé sur la maximisation de la vraisemblance en utilisant le filtre et le lisseur de Kalman, est détaillé en tenant compte des modifications matricielles à apporter à l'algorithme pour tenir compte des observations manquantes. L'algorithme EM est ensuite utilisé afin de reconstruire des coefficients de prédiction temporels à partir de coefficients de projection temporels issus d'une POD sous-échantillonnée sur l'ensemble des snapshots avec différents types de sous-échantillonnages. Cette méthode est appliquée à la reconstruction de l'écoulement autour d'un profil NACA0012 à angle d'incidence de 20° et à nombre de Reynolds Re=1000.

2 Formalisme

La réduction de modèle obtenue par projection de Galerkin d'une base Snapshot-POD (Proper Orthogonal Decomposition [1][2]) tronquée sur les équations de Navier Stokes incompressibles conduit à l'écriture d'un modèle réduit POD-Galerkin de type :

$$\dot{a}_i(t) = D_i + \sum_{i=1}^{N_{pod}} L_{ij} a_j(t) + \sum \sum_{i,j=1}^{N_{pod}} Q_{ijk} a_j(t) a_k(t) \qquad (1)$$

où les $a_i(t)$ représentent les coefficients temporels de projection des champs de vitesse d'écoulement dans leur base snapshot-POD $\vec{\Phi}_i(\vec{x})$ [3]. Dans le cas de mesures PIV résolues en temps, une évaluation directe des coefficients $\dot{a}_i(t)$ permet l'identification empirique de paramètres

[1] romain.leroux@u-bordeaux1.fr

[2] ludovic.chatellier@cnrs.pprime.fr

[3] laurent.david@cnrs.pprime.fr

D, *L* et *Q* du modèle, dont le comportement peut alors être simulé par simple intégration temporelle.

L'instabilité numérique caractéristique de tels modèles peut être efficacement contrôlée par une approche stochastique exploitant le filtrage de Kalman [4][5][6]. Dans le cadre de champs de vitesse répartis sur une base de temps irrégulière, on superpose à ce traitement un processus dit de « données manquantes » permettant l'application de l'algorithme EM (Expectation-Maximization [7]) dans le but de reconstruire l'écoulement sur une base de temps régulière résolue temporellement.

Modèle à espace d'état utilisé

On suppose que le processus latent est observé à plusieurs instants $(t_1, t_2, .., t_N)$ et que le modèle à espace d'état utilisé pour la prédiction du processus latent non observé est linéaire gaussien et constitué des deux équations suivantes :

$$(S)\begin{cases} X_k = F_k X_{k-1} + w_k \\ Y_k = H_k X_k + v_k \end{cases} \qquad (2)$$

où F est la matrice de propagation des états qui décrit la dynamique du système. On considère que le vecteur d'état $Y_k=0$ si l'une de ses composantes est nulle. Les observations manquantes sont remplacées par des zéros dans le vecteur Y_k et les lignes (ou colonnes) correspondantes de H_k sont également remplacées par des zéros. Les bruits w_k et v_k sont des bruits blancs gaussiens décorrélés, de matrices de covariance respectives Q et R. Le vecteur d'état initial x_0 est considéré comme distribué selon une loi normale de moyenne μ_0 et de covariance Σ_0 Les paramètres du modèle (S) sont notés $\Theta_k=\{F_k, H_k, R_k, Q_k\}$. Etant donné le modèle statistique (S), l'objectif est d'estimer les variables latentes Θ_k et le vecteur d'état X_k à chaque instant k, conditionnellement à l'ensemble des observations disponibles Y_k . Les paramètres Θ_k sont estimés par la méthode du maximum de vraisemblance.
La fonction de log-vraisemblance complète est basée sur les séquences des variables latentes observées et un *a priori* :

$$\log(p(x,y;\theta)) = \log(p(x_0)) + \sum_{k=1}^{N} \log(p(x_k|x_{k-1};\theta) + \sum_{k=1}^{N} \log(p(y_k|x_{k-1};\theta) \qquad (3)$$

où les distributions conditionnelles $p(x, y; \theta)$ et $p(y_k|x_{k-1}\theta)$ suivent respectivement des lois normales données par :

$$p(x,y;\theta) = \exp\left(-\frac{1}{2}(x_{k+1} - F_k x_{k-1})^t Q^{-1}(x_{k+1} - F_k x_{k-1})\right)(2\pi)^{-\frac{n}{2}}(\det(Q))^{-\frac{1}{2}} \qquad (4)$$

$$p(y_k|x_{k-1};\theta) = \exp\left(-\frac{1}{2}(y_k - H_k x_k)^t R^{-1}(y_k - H_k x_k)\right)(2\pi)^{-\frac{p}{2}}(\det(R))^{-\frac{1}{2}} \qquad (5)$$

où n est la taille du vecteur d'état et p la taille du vecteur des observations. On suppose ici que $p(x_0)$ suit une distribution de moyenne μ_0 et de variance Σ_0 correspondant à une connaissance *a priori* du système. La log-vraisemblance du système s'écrit (à une constante près)

$$\log L_Y(\Theta) = \log L(X,Y,\Theta) = -\frac{1}{2}\{\log|\Sigma_0| + (x_0-\mu_0)^t\Sigma_0^{-1}(x_0-\mu_0) + N\log|Q| + \sum_{k=1}^{N}(x_k - F_k x_{k-1})^t Q^{-1}(x_k - F_k x_{k-1}) + N\log|R| + \sum_{k=1}^{N}(y_k - H_k x_k)^t R^{-1}(y_k - H_k x_k)\} \quad (6)$$

On cherche les paramètres θ_k qui maximisent la fonction de log-vraisemblance complète. On note θ_k les approximations successives des estimateurs de maximum de vraisemblance au cours des différents cycles k d'assimilation de l'algorithme EM. L'opérateur H_k est la matrice identité pour les cas où on dispose d'observations à l'instant k et la matrice nulle sinon. La variable y_k correspond aux coefficients de projection temporels obtenus par la POD sous échantillonnée, l'opérateur F_k correspond au modèle d'évolution des coefficients de prédiction temporels, c'est-à-dire au modèle réduit POD-Galerkin. La variable x_{k-1} contient les coefficients de prédiction temporels à l'instant k et la variable $F_k x_{k-1}$ contient les coefficients de prédiction temporels à l'instant k. Les matrices Q_k et R_k sont les matrices de covariance des bruits gaussiens w_k et v_k ,supposés indépendants entres eux, aussi on a $Q_k = \sigma_w I$ et $R_k = \sigma_v I$ où σ_w et σ_v sont les écarts-types des bruits des équations d'états et de mesure.

Estimation de paramètres avec l'algorithme EM

Les paramètres estimés $\hat{F}, \hat{H}, \hat{Q}, \hat{R}$ sont les suivants [8][9] :

$$\hat{F} = A_4 A_3^{-1} \qquad \hat{Q} = A_2 - A_4 A_3^{-1} A_4^t \qquad \mu_0 = x_0$$
$$\hat{H} = A_6 A_1^{-1} \qquad \hat{R} = A_5 - A_6 A_1^{-1} A_6^t \qquad \sigma_0 = P_0 \quad (7)$$

où les matrices $A_1, A_2, \ldots, A_6$ sont définies de la manière suivante :

$$A_1 = \frac{1}{N+1}\sum_{k=0}^{N} x_k x_k^t \quad A_2 = \frac{1}{N}\sum_{k=1}^{N} x_k x_k^t \quad A_3 = \frac{1}{N}\sum_{k=1}^{N} x_{k-1} x_{k-1}^t$$
$$A_4 = \frac{1}{N}\sum_{k=1}^{N} x_k x_{k-1}^t \quad A_5 = \frac{1}{N+1}\sum_{k=0}^{N} y_k y_k^t \quad A_6 = \frac{1}{N+1}\sum_{k=0}^{N} y_k x_k^t \quad (8)$$

Dans le cas d'observations manquantes, les quantités $\hat{F}, \hat{H}, \hat{Q}, \hat{R}$ sont calculées à l'étape E de l'algorithme EM en utilisant les espérances conditionnelles suivantes :

$$
\begin{aligned}
\mathbb{E}[x_k|Y,\Theta^{(r)}] &= x_k^s \\
\mathbb{E}[x_k x_k^t|Y,\Theta^{(r)}] &= P_k^{(s)} + x_k^{(s)}(x_k^{(s)^{(t)}} \\
\mathbb{E}[x_k x_{k-1}^t|Y,\Theta^{(r)}] &= P_{k,k-1}^{(s)} + x_k^{(s)}(x_{k-1}^{(s)})^t \\
\mathbb{E}[y_k x_k^t|Y,\Theta^{(r)}] &= H^{(m)} E[x_k x_k^t|Y,\Theta^{(r)}] \\
\mathbb{E}[y_k y_k^t|Y,\Theta^{r)}] &= R^{(m)} + H^{(m)} E[x_k x_k^t|Y,\Theta^{(r)}](H^{(m)})^t \\
\mathbb{E}[y_k|Y,\Theta^{(r)}] &= H^{(m)} E[x_k|Y,\Theta^{(r)}]
\end{aligned}
\qquad (9)
$$

Ces espérances conditionnelles peuvent être calculées $\forall\, k = 1,2,\ldots,N$ à partir du lisseur de Kalman, connu aussi sous le nom de lisseur RTS (*Rauch-Tung-Striebel*) en utilisant les paramètres estimés à l'itération k. Le lisseur RTS est constitué d'une étape *forward* où le filtre de Kalman linéaire est appliqué et une étape *backward* où le lisseur de Kalman est appliqué. Les estimations lissées $x_k^{(s)}$ et $P_k^{(s)}$ sont utilisées pour estimer les espérances conditonnelles utilisées ensuite dans l'étape M de l'algorithme EM pour la maximisation de la vraisemblance.

◦ Filtre de Kalman - *Forward Pass*

Conditions initiales du filtre

$$\hat{x}_0 = \bar{x}_0 \quad \text{et} \quad p_0 = Q_0 \qquad (10)$$

A tout instant $k \geq 1$ Étape de prédiction

$$
\begin{aligned}
x_k^f &= F_k(x_{k-1}) \\
P_k^f &= F_k P_{k-1}^a F_k^t + H_k Q_k H_k^t
\end{aligned}
\qquad (11)
$$

Étape de correction

$$
\begin{aligned}
x_k^a &= x_k^f + K_k[y_k - H_k x_k^f] \\
P_k^a &= [I - K_k H_k] P_k^f \\
K_k &= P_k^f H_k^t [H_k P_k^f H_k^t + R_k]^{-1}
\end{aligned}
\qquad (12)
$$

◦ Lisseur de Kalman - *Backward Pass*

Conditions initiales du filtre

$$x_k^{(s)} = x_k^{(a)} \quad \text{et} \quad P_k^{(s)} = P_k^{(a)}$$

Etape de lissage : Pour k allant de n-1 à 1 :

$$
\begin{aligned}
K_k^{(s)} &= P_k^{(a)} F_k^t (P_{k+1}^{(f)})^{-1} \\
x_k^{(s)} &= x_k^{(a)} + K_k^{(s)}(x_{k+1}^{(s)} - x_{k+1}^{(f)}) \\
P_k^{(s)} &= P_k^{(a)} + K_k^{(s)}(P_{k+1}^{(s)} - P_{k+1}^{(f)})(K_k^s)^t \\
P_{k-1,k}^{(s)} &= (I - K_k H_k) F_k P_{k-1}^{(a)} + (P_k^{(s)} - P_k^{(a)})(P_k^{(a)})^{-1}(I - K_k H_k) F_k P_{k-1}^{(a)}
\end{aligned}
\qquad (13)
$$

L'algorithme EM est un processus itératif qui utilise la distribution des données complètes pour calculer les estimateurs du maximum de vraisemblance lorsque les données observées sont incomplètes. Cet algorithme se déroule en deux étapes. La première est l'étape E qui consiste à calculer l'espérance conditionnelle de la fonction de log-vraisemblance des données complètes sachant les données observées. La seconde étape est l'étape M, qui est la maximisation de la log-vraisemblance obtenue à l'étape E. On détermine ainsi l'estimateur qui maximise l'équation trouvée à l'étape E. Ces étapes E et M sont répétées itérativement jusqu'à convergence de la log-vraisemblance.

- **Etape E à l'itération r+1**

1. Utilisation des paramètres $\theta^{(r)}$ estimés à l'itération r et application des filtres et lisseurs de Kalman (équations 5.20 à 5.29) pour obtenir les valeurs lissées $x_k^{(N)}$ et $P_k^{(N)}$. Obtention des espérances conditionnelles 5.13 à 5.18
2. Calcul de la nouvelle log-vraisemblance $\log(\theta^{(r+1)})$
3. Calcul des matrices A_1 à A_6 en utilisant les quantités définies au (1)

- **Etape M à l'itération r+1**

1. Re-estimation de l'ensemble des paramètres $\theta^{(r+1)}$ en utilisant les équation 5.7 à 5.12 et les statistiques déterminées en (2) de l'étape E.

Critère d'arrêt de l'algorithme EM

Le critère d'arrêt de l'algorithme EM utilisé ici est basé sur la convergence de la log-vraisemblance :

$$\left| \frac{Q(\theta^{(r+1)},\theta^{(r+1)}) - Q(\theta^{(r)},\theta^{(r)})}{Q(\theta^{(r)},\theta^{(r)})} \right| < \varepsilon \qquad (14)$$

avec $10^{-6} < \varepsilon < 10^{-4}$. Si le critère d'arrêt est vérifié alors les itérations de l'algorithme EM sont stoppées et l'ensemble des paramètres est retenu comme l'ensemble des estimés des paramètres du modèle sinon on pose $\theta^{(r)} = \theta^{(r+1)}$ et $\log(\theta^{(r)}) = \log(\theta^{(r+1)})$ et l'algorithme continue.

3 Résultats

L'écoulement autour d'un profil NACA 0012 de corde *c*=60mm est étudié dans un tunnel hydrodynamique de section 160mm*160mm pour des nombres de Reynolds *Re*=1000 et *Re*=2000 aux angles d'incidence α =0°, 10°, 15°, 20° et 30°. Des séries de mesures PIV résolues en temps de 2050 et 10240 échantillons sont réalisées afin de simuler des bases de temps irrégulières et de valider la reconstruction des données manquantes.

Le processus des données manquantes est simulé par un masquage de la base de temps selon un sous-échantillonnage par paquets paramétré par les nombres d'échantillons disponibles, l_1, et manquants, l_2, résultant en un taux de vide $\tau_{vide} = l_2/(l_1 + l_2)$ (Figure 1).

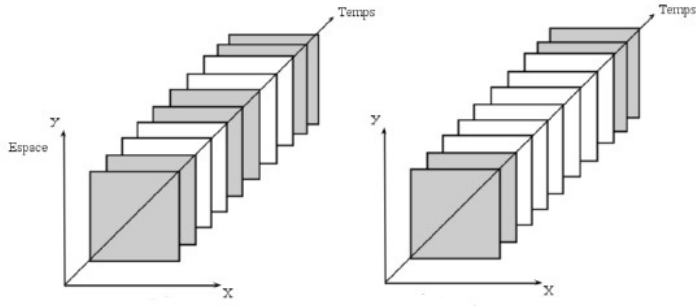

Figure 1 - Exemples de sous-échantillonnage par paquets : (l_1,l_2)=(2,2) et (l_1,l_2)=(2,6), correspondant à des taux de vide respectifs de 50% et 75%

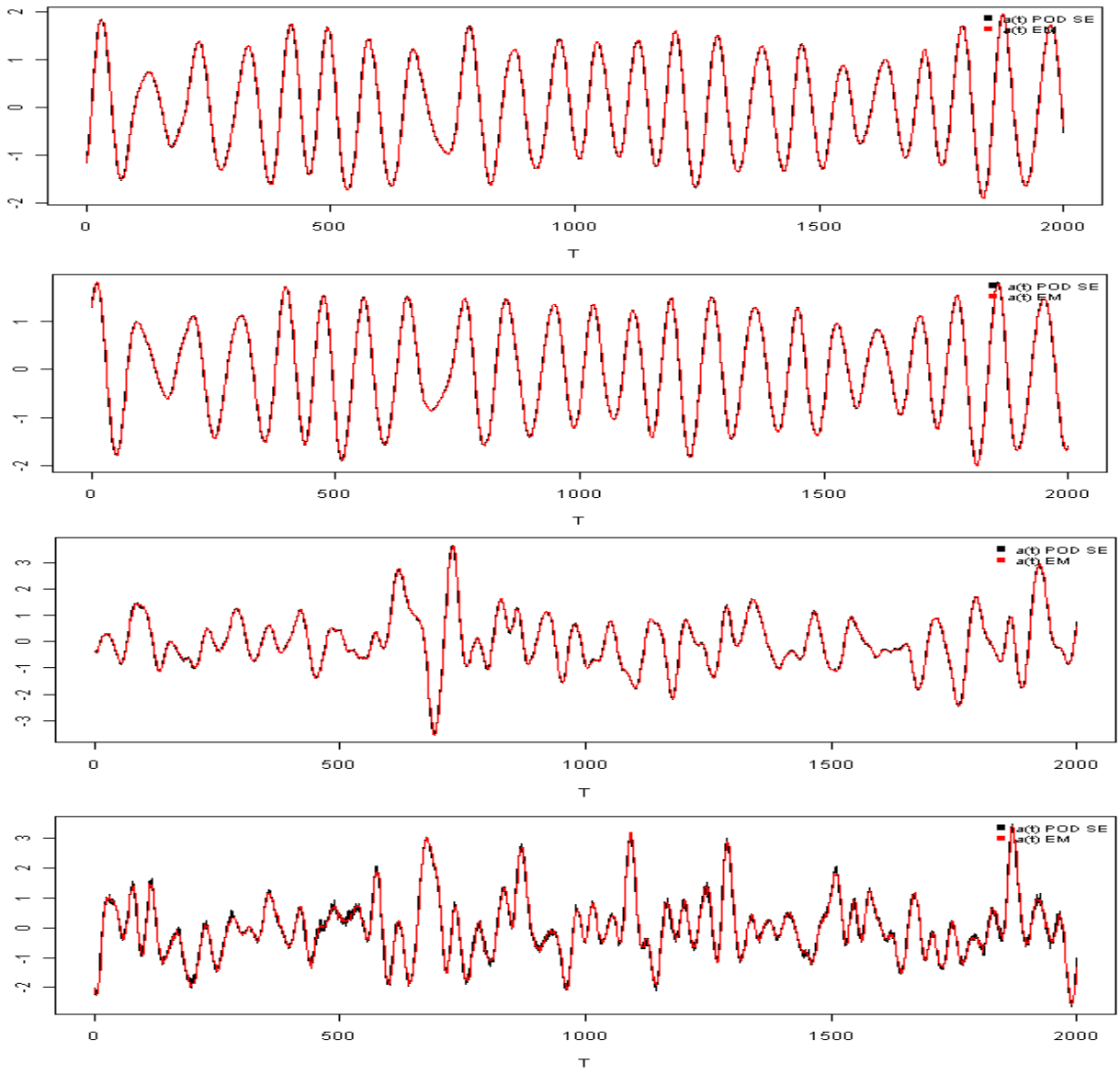

Figure 2 - Modes temporels $a_1(t), a_2(t), a_5(t), a_{10}(t)$ POD (en noir) et reconstruits par modèle POD-Galerkin et algorithme EM (en rouge) - Taux de données manquante de 75%. Cas α =20°, Re=1000 ; (l_1,l_2)=(2,6).

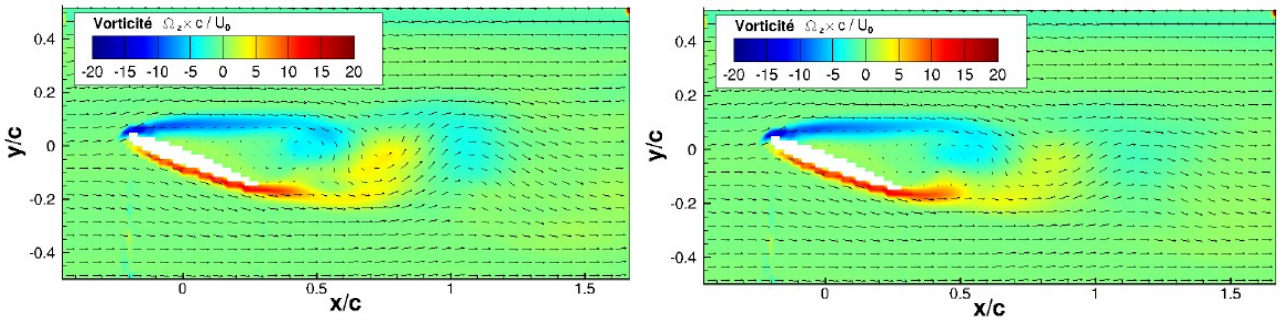

Figure 3 - (a) Champ POD (80%) et (b) reconstruction par modèle POD-Galerkin et algorithme EM dans le cas d'un taux de données manquantes de 75%.Cas α =20°, Re=1000 ; (I_1,I_2)=(2,6).

Les résultats obtenus montrent que la combinaison du modèle POD-Galerkin et de l'algorithme EM permet la reconstruction des coefficients temporels de la POD avec une précision qui dépend à la fois de la dynamique de l'écoulement reconstruit, de l'indice du mode POD, du taux de vide de la base de temps et de la taille caractéristique I_1+I_2 des blocs de données traités. Dans le cas présenté ici, l'erreur de reconstruction sur l'ensemble des coefficients est de l'ordre de 10^{-3}. L'écoulement est alors reconstruit avec une précision pondérée par le niveau d'énergie associé à chaque mode POD retenu (Figure 3).

4 Conclusion

En conclusion, des techniques issues de l'inférence bayésienne ont été appliquées sur le modèle réduit POD-Galerkin dans le cas d'observations manquantes. L'objectif est de réduire les coûts de calcul en sélectionnant un nombre limité d'observations pour reconstruire l'etat du système. A cet effet, une POD sous échantillonnéee de l'écoulement est tout d'abord realisée. L'algorithme EM est ensuite utilisé pour l'inférence sur le modèle réduit POD-Galerkin avec des observations manquantes. Cet algorithme est connu pour converger lentement, c'est pourquoi il a ete appliqué au modèle POD-Galerkin sur un cas où le RIC est sufiisamment élevé (80%) et correspondant à un nombre peu élevé de modes. L'algorithme EM permet de reconstruire les coefficients de prediction temporels et l'écoulement associé de maniére satisfaisante.

Remerciements Les auteurs remercient le programme européen FP7 AFDAR (Advanced Flow Diagnostics for Aeronautical Research).

5 Références

[1] J.L. Lumley « The structure of inhomogeneous turbulent flows », Atm. turb. and radio wave prop., 166–178 (1967)

[2] L. Sirovitch « Turbulence and the dynamics of coherent structures », Quarterly of Applied Mathematic (1987)

[3] B. R. Noack, P. Papas et P.A. Monkewitz « The need for a pressure-term representation in empirical Galerkin models of incompressible shear flow », J. Fluid Mech. 523:339-365 (2005)

[4] R.E. Kalman « A new approach to linear filtering and prediction problems » Journal of Basic Engineering 82 (1): 35–45 (1960)

[5] G. Evensen « Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics » Journal og geophysical research 99(C5) : 10,143-10,162 (1994)

[6] R. Leroux, L. Chatellier et L. David « Spatio-temporal reconstruction of flows around a NACA0012 airfoil using Kalman-filtered POD Galerkin LODS » PIV'11 - Ninth International Symposium on Particle Image Velocimetry (2011)

[7] A. P. Dempster et N. M. Laird et D. B. Rubin « Maximum Likelihood from Incomplete Data via the EM Algorithm » Journal of the Royal Statistical Society. Series B (Methodological) 39(1) :1-38 (1977)

[8] Digalakis, V., Rohlicek, J., and Ostendorf, M. (1993). « Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition . » IEEE Transactions on Speech and Audio Processing, 1(4) :431-442.

[9] Gyau-Boakye, P. and Schultz, G. (1994). « Filling gaps in runo time series in west africa ».Hydro.Sci.J, 39(6) :621-636.