

---

# Estimation de systèmes peu observables en grande dimension par apprentissage statistique

Lionel Mathelin\*<sup>1</sup>

<sup>1</sup>Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS) – CNRS :  
UPR3251 – rue John von Neumann, Campus Universitaire d'Orsay, Bât. 508, 91405 ORSAY cedex,  
France

## Résumé

Dans un contexte d'assimilation de données, l'efficacité de l'inférence Bayésienne est typiquement limitée par le coût de l'évaluation numérique du modèle et le peu d'information disponible. Une réduction de la dimension de l'espace d'inférence permet de significativement réduire la quantité de variables à identifier. Dans le cas de champs spatialement distribués, les approches classiques consistent à utiliser des bases de modes spatiaux, typiquement issus d'une hypothèse sur un noyau de corrélation statistique (décomposition de Karhunen-Loève) ou une séquence de champs d'apprentissage (modes dominants d'une décomposition en Analyse en Composantes Principales). D'autres approches reposent sur des modes à supports compacts (e.g., ondelettes), permettant une représentation typiquement creuse du champ à inférer.

Des algorithmes efficaces existent pour déterminer une base de modes (dictionnaire) conduisant à une représentation creuse à partir d'une séquence d'apprentissage, par exemple K-SVD.

La représentation du champ inconnu étant, avec une bonne probabilité, creuse, le problème d'identification peut alors être efficacement résolu par une formulation de type Compressed Sensing, impliquant une régularisation basée sur la parcimonie de la représentation du champ dans la base du dictionnaire.

Néanmoins, cette classe d'approches ne prend pas en compte la chaîne d'acquisition lors de la détermination du dictionnaire, qui est appris à partir de champs complets, et non de mesures. Se pose donc un problème d'observabilité des modes résultants dont le support peut ne pas inclure les capteurs utilisés in fine sur site. Le dictionnaire permet alors une très bonne approximation (creuse) du champ, avec une bonne probabilité, mais les mesures ne permettent pas forcément d'informer correctement ces modes, résultant en une identification, potentiellement, de très mauvaise qualité.

Dans cet exposé, la détermination d'une base d'approximation par apprentissage statistique sous contrainte d'observabilité et de parcimonie, sera discutée. On montre que le problème se formule comme un problème de double optimisation sous contraintes. Une approche Bayésienne est adoptée et une formulation hiérarchique permet, dans le cas particulier d'un système physique régi par un opérateur d'évolution linéaire et d'un modèle de vraisemblance Gaussien, l'apprentissage de dictionnaires par une approche progressive. Cette approche est très efficace numériquement, et conduit à des expressions explicites pour l'estimation du champ aléatoire considéré (maximum a posteriori, matrice de covariance). L'estimation du

---

\*Intervenant

champ est alors immédiate à partir de ces résultats. Dans le cas plus général, une approche MCMC peut-être utilisée avec le dictionnaire appris.

Des exemples illustrant ces différents points seront présentés et discutés.