Proposition de sujet de thèse CIFRE en apprentissage statistique

Apprentissage statistique pour la prédiction dans un modèle de régression.

Statistical learning for regression model prediction.

Contacts :

Olivier Wintenberger
Professeur
Université Pierre et Marie Curie
olivier.wintenberger@upmc.fr

Yannig Goude

A. Contexte et problématique

Présentation DataLab EDF

B. Verrous scientifiques

Le projet consiste à mettre en œuvre une méthode de prédiction peu coûteuse en le nombre de variables explicatives (linéaire), souple car pouvant s'adapter à une évolution du nombre de variables explicatives au cours du temps et suffisamment adaptative pour pouvoir prendre en compte des comportements très différents. Il se composera de deux étapes successives

Modèles à espace d'états

Le modèle de base sera un modèle espace-état généralisé avec

- une équation d'espace Y_t = \theta_t X_t + \epsilon_t,
- une équation d'état \theta_t = \theta_{t-1} + \eta_t,

où (\epsilon_t) et (\eta_t) sont deux niveaux de bruits, Y_t sont les variables à expliquer et X_t les variables explicatives. Dans le cadre généralisé, l'équation d'espace doit être vu comme un modèle linéaire généralisé, voir Fahrneir (1994).

Des extensions de ce modèle, appelé Modèle Linéaire Dynamique, a rencontré de nombreux succès pratiques, voir Petris et al. (2009). Toutefois, supporté par des articles théoriques comme Klüppelberg et Pergamenchtchikov (2004), il est apparu que la forme très simple ci-dessus est suffisante pour capturer un très grand nombre de comportements différents. La dynamique théorique sur le coefficient aléatoire étant celle d'une marche aléatoire. En pratique, n'importe quelle dynamique suffisamment régulière dans le temps du coefficient peut être capturée, voir Prado et West (2010).

Ce phénomène observé en pratique provient de la méthode d'optimisation utilisée. Pour le modèle espace-état linéaire, il est optimal d'utiliser une récurrence de Kalman. Même dans le cadre généralisé, la récurrence de Kalman sera utilisée afin de maintenir une complexité linéaire en le nombre de variables explicatives. On parle alors de filtre de Kalman généralisé. Cette procédure en

ligne va réussir à se mettre en poursuite de la valeur (non observée) de \theta_t et ainsi l'estimer même si celle-ci a une dynamique arbitraire.

Le premier objectif de cette thèse va être l'analyse de la stabilité de ces algorithmes en pratique et en théorie.

Agrégation par poids exponentiels

Le modèle espace-état ci-dessus peut être décliner infiniment en rajoutant des variables explicatives. On peut penser à

- des variables explicatives Y passées, le modèle espace état est alors similaire à des modèles de type ARMA,
- des fonctions des variables explicatives X afin d'obtenir des modèles additifs généralisés.

Le temps de calcul peut rapidement devenir prohibitif. La parallélisation de l'algorithme de Kalman sera mis en place afin de remédier à cette difficulté.

La prédiction de la variable en sortie sera alors obtenu grâce à une méthode d'agrégation en ligne par poids exponentiels, voir Cesa-Bianchi et Lugosi (2006). Chaque filtre de Kalman fournit gratuitement un indicateur de la confiance de la prédiction sous forme d'une variance conditionnelle. Cet indicateur sera pris en compte pour adapter les algorithmes de poids exponentiels et obtenir une agrégation optimale pour le risque quadratique conditionnel du type de celle obtenue par Leung et Barron (2006). L'intégration de cet indicateur permettra de ne pas tenir en compte dans la prédiction finale d'états trop instables.

Références :

Cesa-Bianchi, N., Lugosi, G. (2006). Prediction, learning, and games. Cambridge university press.
Fahrmeir, L. (1994). Dynamic modelling and penalized likelihood estimation for discrete time survival data. Biometrika, 81(2), 317–330.
Klüppelberg, C., Pergamenchtchikov, S. (2004). The tail of the stationary distribution of a random coefficient AR (q) model. The Annals of Applied Probability, 14(2), 971-1005.
Leung, G., & Barron, A. R. (2006). Information theory and mixing least-squares regressions. IEEE Transactions on Information Theory, 52(8), 3396-3410.
Petris, G., Petrone S., Campagnoli P. (2009). Dynamic Linear Models with R, Springer Textbook.
Prado, R., West, M. (2010). Time series: modeling, computation, and inference. CRC Press.

C. Calendrier prévisionnel

Phases	Université	Entreprise
Prise en main du sujet (Automne 2017)	Renforcement des connaissances en apprentissage statistique (suivi de cours), Exploration de la bibliographie.	
Premiers résultats (Jusque fin 2018)	Rédaction d'un article de conférence sur la traçabilité des filtres de Kalman à état marche aléatoire.	
Phase finale (Jusque fin 2019)	Rédaction du document final et de l'article sur l'agrégation avec critère modifié.	

D. Retombées

Les retombées du projet de thèse sont attendues à deux niveaux :

1. Au niveau fondamental, il est envisagé de mettre au point un package sous R ou Python incluant l'agrégation de filtres de Kalman pénalisant ceux qui sont trop instables. Ce package

fera l'objet d'un article descriptif dans une revue spécialisée. Les autres retombées seront des publications et des communications dans des conférences et séminaires.

2. Au niveau industriel,

E. Bénéfices pour les deux parties

Le partenariat envisagé entre EDF et le Laboratoire de Statistique Théorique et Appliquée de l'Université Pierre et Marie Curie nous semble apporter des bénéfices significatifs pour les deux parties.

Pour le Laboratoire de Statistique Théorique et Appliquée, ce travail doctoral représente non seulement un moyen de valider expérimentalement des modèles statistiques théoriques développés au sein de l'unité, mais également un potentiel fort de contrats de recherches puisque le Data Lab d'EDF est un acteur incontournable de la recherche statistique appliquée en France.

F. Moyens envisagés

Côté EDF, le doctorant sera encadré par Y Goude.

Au niveau de l'Université Pierre et Marie Curie, le doctorant sera encadré par Olivier Wintenberger et intégré au Laboratoire de Statistique Théorique et Appliquée.

Enfin, tous les moyens nécessaires au bon déroulement de la thèse (bureau, ordinateurs, logiciels, visite sur le terrain, etc.) seront bien entendu déployés. Le candidat-doctorant dispose déjà d'un bureau équipé à EDF d'une part et au sein du LSTA d'autre part.