# PhD Day, CREST–ENSAI

## Mercredi 4 décembre 2024

## Résumés

---

**13h30 - 13h45.** Louis Allain (SAFRAN-CREST)

Encadrants : Sébastien Da Veiga (ENSAI, Rennes), Brian Staber (Safran, Magny-les-Hameaux).
Titre : *Quantification d'incertitudes par méthodes à noyaux.*
Résumé : Cet exposé présente la quantification d'incertitudes des prédictions d'un modèle de machine learning. Nous commencerons par étudier le cas d'usage Safran. Nous verrons ensuite que les méthodes traditionnelles ne suffisent pas et que la théorie des prédictions conforme, malgré ses avantages, n'est pas adaptée. Nous allons alors explorer une méthode récente pour dépasser les limitations des méthodes actuelles et nous introduisons des améliorations pour le temps de calcul et le choix d'hyper-paramètres.

**13h45 - 14h00.** Koffi Amezouwui (CREST)

Encadrants : Brigitte Gelein, (ENSAI, Rennes), Matthieu Marbac-Lourdelle (ENSAI, Rennes), Anthony Sorel (Univ. Rennes 2, Rennes).
Titre : *Analyse et classification des situations de jeu au football pour peupler des environnements virtuels.*
Résumé : L'objectif de ce travail est d'analyser et de classer les situations de jeu au football afin de peupler les environnements virtuels. Nous nous intéressons à la classification des possessions, c'est-à-dire des phases de jeu ininterrompues. Chaque possession est caractérisée par les coordonnées spatiales du ballon. La possession s'arrête lorsque l'équipe perd le ballon (tir, interception, touche...). Nous proposons de modéliser la trajectoire du ballon par une chaîne de Markov cachée, caractérisée par la présence d'un état absorbant ainsi que par une distribution initiale distincte de la distribution pseudo-stationnaire. L'état absorbant correspond à une perte de balle, marquant la fin de la trajectoire de la balle et de la possession. La distribution initiale modélise la distribution lors de la prise de balle et n'a donc pas de raison d'être identique à la distribution pseudo-stationnaire. Les modèles de Markov cachés sont peu étudiés en présence conjointe d'un état absorbant et d'une distribution initiale libre. L'étude de ce modèle, dans le contexte de la modélisation des trajectoires de balle dans les séquences de football, constituera notre contribution en statistique mathématique. Pour vérifier l'intérêt de notre modélisation nous disposons d'une base de données de foot riche à analyser, constituée des enregistrements de 38 matchs pour lesquels nous avons les coordonnées du ballon enregistrées toutes les 0,04 secondes, ainsi que les positions des joueurs, etc. Après le clustering des situations de jeu, nous allons générer en utilisant des modèles d'IA (Intelligence Artificielle) de nouvelles séquences réaliste de situations de jeu pour peupler des environnements virtuels. Ces environnements permettront aux joueurs de s'entraîner sur le plan cognitif sans nécessiter d'efforts musculaires.

**14h00 - 14h15.** Daphné Aurouet (CREST)

Encadrant : Valentin Patilea (ENSAI, Rennes).
Titre : *Estimation en ligne de la fonction moyenne de données fonctionnelles ajustées aux covariables.*
Résumé : Certaines applications de l'analyse des données fonctionnelles nécessitent de prendre en compte des données fonctionnelles ajustées aux covariables pour permettre une meilleure compréhension du phénomène d'intérêt. Les estimateurs des fonctions de moyenne et de covariance ont été étudiés à l'aide de méthodes basées sur le noyau. Toutefois, ces méthodes ne permettent pas de traiter de grands volumes de données, des flux en temps réel, et des erreurs de mesure dans un contexte d'échantillonnage épars et aléatoire. Je présenterai un estimateur récursif de la fonction moyenne qui intègre les covariables continues et discrètes, en mettant en œuvre une stratégie de micro-pooling pour permettre le traitement en temps quasi réel quand le nombre de points par courbe est faible.

**14h15-14h30.** Jérémy Bettinger (CREST)

Encadrants : François Portier (ENSAI, Rennes), Adrien Saumard (ENSAI, Rennes).
Titre : *Cartes à régularité de forme en régression locale.*
Résumé : Dans cet exposé, nous introduirons le concept de "cartes de régression à régularité de forme" comme cadre pour établir des vitesses de convergence optimales pour divers estimateurs de régression locale non paramétriques. À l'aide de la théorie de Vapnik-Chervonenkis, nous établirons des bornes supérieures et inférieures sur l'erreur d'estimation sous des hypothèses légères sur le modèle de régression. Nous démontrerons que la régularité de forme des cartes de régression est à la fois suffisante et nécessaire pour atteindre ces vitesses optimales. De plus, nous établirons de nouvelles bornes de concentration pour des méthodes usuelles de régression locale, comme les arbres purement aléatoires uniformes, centrés et de type Mondrian.

**14h30-14h45.** Raphaël Carpintero-Perez (SAFRAN-X)

Encadrants : Sébastien Da Veiga (ENSAI, Rennes), Josselin Garnier (X, Palaiseau), Brian Staber (Safran, Magny-les-Hameaux).
Titre : *Régression par processus Gaussiens pour des entrées graphes en grande dimension.*
Résumé : Dans cet exposé, on s'intéresse à l'apprentissage de simulations basées sur des maillages en présence de variabilités géométriques et paramétriques. On cherche ainsi à résoudre un problème de régression dont les entrées sont des graphes en grande dimension correspondant à des maillages, et les sorties sont des champs. Les processus Gaussiens représentent un outil de choix car ils sont puissants lorsque la taille de l'échantillon est faible et qu'il est nécessaire de quantifier les incertitudes. Une première partie du travail concerne la définition de noyaux entre graphes utilisés pour de la régression par processus Gaussiens à sorties scalaires. J'introduis les noyaux de Sliced Wasserstein Weisfeiler-Lehman (SWWL) combinant une représentation des graphes à du transport optimal. Une seconde partie traite de l'extension de la régression à des sorties sous formes de champs. On repose à nouveau sur du transport optimal appliqué dans l'espace des sorties combiné avec de la réduction de dimension. Les méthodes sont illustrées sur des jeux de données en mécanique des fluides et du solide.

**14h45 - 15h00.** Mohamed El Hasnaoui (CREST)

Encadrant : Matthieu Marbac-Lourdelle (ENSAI, Rennes).
Titre : *Estimation semi-paramétrique des densités de mélanges.*
Résumé : Nous proposons un estimateur semi-paramétrique pour des densités de mélange comportant deux ou plusieurs composantes. La densité estimée appartient à une famille de mélanges, où les densités des composantes, au sein d'un mélange, sont égales à une translation et un facteur d'échelle près. Nous estimons les proportions du mélange, les paramètres de position et

d'échelle (paramètres de dimension finie) ainsi que la densité des composantes (paramètre de dimension infinie) par la méthode du maximum de vraisemblance. La consistance découle de l'identifiabilité du vrai paramètre et de la compacité de l'espace des paramètres, tandis que l'identifiabilité et la vitesse de convergence proviennent de la projection de la densité des composantes sur un sous-espace de dimension finie de $L^2(\mathbb{R})$. En se basant sur l'algorithme EM, nous menons des expériences numériques qui confirment nos résultats théoriques.

---

## – PAUSE –

---

**15h15 - 15h30.** Omar Kassi (CREST)

Encadrants : Valentin Patilea (ENSAI, Rennes), Matthieu Marbac-Lourdelle (ENSAI, Rennes).
Titre : *Inférence optimale pour la fonction moyenne.*
Résumé : Le problème de l'estimation de la moyenne des fonctions aléatoires à partir de données échantillonnées discrètement se manifeste naturellement dans l'analyse des données fonctionnelles. Dans cet exposé, nous étudions l'estimation et l'inférence de la fonction moyenne dans le cadre d'un schéma d'observations aléatoire. des taux de convergence optimaux basés sur les séries de Fourier sont proposés, et des limites non asymptotiques sont données pour la norme L2 et la norme uniforme. L'inférence est également construite.

**15h30 - 15h45.** Hassan Maissoro (CREST)

Encadrants : Valentin Patilea (ENSAI, Rennes), Myriam Vimond (ENSAI, Rennes), Julien Bretteville (Datastorm, Palaiseau).
Titre : *Prédiction adaptative pour les séries temporelles fonctionnelles.*
Résumé : Une procédure adaptative de prédiction de courbe pour une série temporelle fonctionnelle stationnaire est proposée. Les trajectoires des séries temporelles fonctionnelles sont supposées être irrégulières et sont observées avec erreur à des instants discrets. Notre prédicteur linéaire est basé sur le meilleur prédicteur linéaire sans biais (BLUP) et sur les estimateurs non paramétriques adaptatifs des fonctions de moyenne et d'autocovariance du processus. En d'autres termes, les fenêtres de lissage de ces estimateurs sont choisies de manière adaptative en fonction de la régularité locale des trajectoires. L'avantage d'une telle procédure est une réduction du risque de prédiction par rapport aux procédures existantes. Des simulations ainsi qu'une application sur des données réelles illustrent les performances de la méthode proposée.

**15h45 - 16h00.** Gabriel Mastrilli (INRIA-CREST)

Encadrants : Frédéric Lavancier (ENSAI, Rennes), Bartłomiej Błaszczyszyn (INRIA, Paris).
Titre : *Estimation non asymptotique des propriétés spectrales des processus ponctuels.*
Résumé : La question de l'estimation des propriétés spectrales des champs aléatoires a été largement étudiée et continue de l'être, mais son équivalent pour les processus ponctuels demeure peu exploré. Des travaux récents ont permis de formaliser certains estimateurs spectraux utilisés en physique, notamment la famille polyvalente des estimateurs multi-tapers. Cependant, ces études se concentrent principalement sur leur propriétés asymptotiques. Dans cet exposé, nous adoptons un cadre non asymptotique qui met en lumière les choix pratiques des paramètres pour ces estimateurs multi-tapers. Après avoir établi une borne inférieure minimax, nous montrons que ces estimateurs atteignent cette borne, garantissant leur optimalité. Enfin, nous présentons leurs propriétés de concentration.

**16h00-16h15.** Mahamat Nassouradine (CEA-CREST)

Encadrants : Clément Gauchy (CEA, Saclay), Pierre-Emmanuel Angeli (CEA, Saclay), Sébastien Da Veiga (ENSAI, Rennes).

Titre : *Prédiction de champs physiques sous contraintes linéaires, application en mécanique des fluides.*

Résumé : Cet exposé présente un modèle d'apprentissage statistique de champs physiques sous contraintes linéaires, combinant la régression par processus gaussien et l'ACP. L'objectif de ce couplage est d'une part de gérer la grande dimensionnalité des champs, puis de construire des noyaux de covariance adaptés pour prendre en compte les contraintes linéaires. Le modèle proposé a été testé et validé sur des simulations de mécanique des fluides et des champs de vitesse. Enfin, nous discuterons des axes de recherche prévus pour la thèse, ainsi que des défis posés par la haute dimensionnalité liée à la taille du maillage, notamment pour la modélisation du tenseur de Reynolds, quantité primordiale pour la simulation d'écoulements turbulents.

**16h15-16h30.** Jean Rubin (INSEE)

Encadrant : Guillaume Chauvet (ENSAI, Rennes).

Titre : *Construction de plans de sondage équilibrés avec remise.*

Résumé : L'équilibrage correspond à une méthode d'échantillonnage, tirant avantage d'une information connue afin d'avoir de meilleures estimations. Elle est notamment employée en France pour la conception de l'Echantillon Maître, qui est un échantillon géographique utilisé entre autres pour le recensement de la population française. La manière classique pour produire un échantillon équilibré, appelée méthode "du cube", se concentre toutefois sur la production d'échantillons sans remise. Dans l'optique de quantifier l'incertitude des estimations par des méthodes bootstrap, nous proposons de généraliser la méthode du cube pour produire des échantillons équilibrés avec remise. Nous étudierons ensuite les propriétés de ces méthodes par simulation.

**16h30-16h45.** Simon Viel (CREST)

Encadrants : Lionel Truquet (ENSAI, Rennes), Ikko Yamane (ENSAI, Rennes).

Titre : *Analyse du biais pour des estimateurs d'appariement.*

Résumé : Estimer l'espérance d'une fonction de régression est un problème central dans différents domaines dont l'adaptation de domaine et l'analyse des effets de traitement. Les estimateurs basés sur l'algorithme des $k$ plus proches voisins, également appelés estimateurs d'appariement, sont largement utilisés dans ces contextes. Il est connu que la variance de ces estimateurs converge à une vitesse paramétrique, mais leur biais peut avoir une décroissance moins rapide lorsque la dimension des covariables est supérieure ou égale à 3. Dans ma présentation, je fournirai des propriétés du biais en mettant l'accent sur deux conditions géométriques sur le support qui permettent d'éliminer le biais lié au bord.

**16h45-17h00.** Sunny Wang (CREST)

Encadrant : Valentin Patilea (ENSAI, Rennes).

Titre : *Inférence accélérée pour les intégrales de fonctions aléatoires multivariées.*

Résumé : Le problème de l'estimation des intégrales de fonctions aléatoires définies sur un intervalle compact ou un rectangle multidimensionnel se manifeste naturellement dans l'analyse des données fonctionnelles. Ces intégrales sont nécessaires pour faire des prédictions à l'aide de modèles de régression fonctionnelle, pour calculer les scores d'une fonction aléatoire dans une base, pour calculer la profondeur des données, etc. En utilisant une approche récente d'intégration linéaire issue de la littérature de Monte Carlo, nous proposons des procédures d'inférence dans le cas où les fonctions aléatoires sont observées, éventuellement avec un bruit de mesure, sur un ensemble discret et aléatoire de points d'observations. En l'absence de bruit, les estimateurs convergent plus rapidement que les moyennes empiriques. Des intervalles de prédiction et de confiance étroits sont proposés avec et sans bruit de mesure.