Proposition de sujet de thèse : apprentissage statistique pour des données fonctionnelles

Encadrants : Christophe Denis[(1)], Charlotte Dion[(2)], Viet-Chi Tran[(1)]
[(1)] Université Gustave Eiffel, LAMA, 77454 Marne-la-Vallée Cedex 2
[(2)] Sorbonne Université, LPSM, 75005 Paris

Contacts : christophe.denis@u-pem.fr, charlotte.dion_blanc@sorbonne-universite.fr, chi.tran@u-pem.fr

Etudiant Eddy ELLA MINTSA

Mots clés : Processus de diffusion, classification, estimation non-paramétrique.

Encadrement

- Christophe Denis, Maître de conférence à l'Université Gustave Eiffel, laboratoire LAMA, depuis 2013 https://perso.math.u-pem.fr/denis.christophe/
- Charlotte Dion, Maître de conférence à Sorbonne Université , laboratoire LPSM, depuis 2017 https://sites.google.com/site/charlottedionblanc/
- Viet-Chi Tran, Professeur à l'Université Gustave Eiffel, laboratoire LAMA, depuis 2019 https://perso.math.u-pem.fr/tran.viet-chi/

Présentation du candidat

Stage M2 à décrire
Passé : Master 2 Probabilités et Statistiques de nouvelles données, Université Gustave Eiffel
TER encadré par Claire Lacour

1 Introduction

La classification de trajectoires est un domaine important de recherche en Statistique, les récentes avancées technologiques (notamment l'utilisation de capteurs) rendant très facile la collecte de ce type de données [15]. En particulier, certains type de données fonctionnelles peuvent être modélisées par des processus de diffusion. Par exemple, la dynamique de la vitesse cellulaire en biologie [16] ou le prix d'une action financière [13] sont décrits généralement par des équations différentielles stochastiques.

Cette problématique s'inscrit dans le cadre général de la classification de données fonctionnelles. Dans ce contexte, des méthodes générales ont été étudiées [1, 6]. Peu de travaux étudient le cas où les données sont modélisées par des processus de diffusion [2, 3]. C'est pourquoi la mise en place d'une procédure de classification adaptée à ce type de modèles est un enjeu majeur. Une procédure de classification basée sur cette modélisation a été proposée

dans [3] dans le cas où les classes sont discriminées par le coefficient de dérive (*drift*). Plus précisément, deux classifieurs et leur convergence ont été obtenus dans le cas où la fonction de dérive est supposée dépendre d'un paramètre.

L'objectif de cette thèse est d'étendre les résultats existants notamment dans [3]. Le développement de procédures non-paramétriques adaptées à ce problème sera un axe important de la thèse. En outre, l'implémentation de ces procédures ainsi que leur évaluation numérique seront au cœur du travail de thèse. Un autre axe de la thèse sera d'élargir le cadre d'étude à des modèles plus généraux tel que les processus de Levy. Enfin une application à l'étude de la dynamique cellulaire basée sur des données réelles permettra de tester les différentes procédures sur un problème concret. Les méthodes proposées seront étudiées dans le contexte où les données sont des observations discrètes (de processus continus) sur un intervalle de temps fixe. En revanche, on supposera, disposer d'assez d'observations par classe afin d'en avoir une représentation précise. Tout le long de l'étude, il faudra prendre en compte la discrétisation du temps et l'étude des vitesses de convergence optimales dans ce cadre seront des enjeux importants du travail de thèse.

Ce travail est très ambitieux d'un point de vue statistique, car il faudra proposer des méthodes statistiques adaptées à chaque modèle. Dans un premier temps, la thèse aura pour but d'étudier d'un point de vue théorique les estimateurs développés, notamment au travers de l'étude des vitesses de convergence. Des méthodes paramétriques et non-paramétriques sont envisagées, exigeant la maîtrise d'une large palette d'outils statistiques tels que les inégalités de concentration. Ensuite, viendra l'implémentation des méthodes et leur évaluation sur données simulées nécessitant la mise en œuvre de procédures numériques d'optimisation. Enfin, l'analyse de données réelles et la validation (ou non) des modèles probabilistes sera un point fondamental de la thèse.

Le projet de thèse de Eddy Ella Minsta est divisé en deux grandes parties : l'étude de procédures de classification non-paramétriques pour des processus de diffusion homogène et l'extension de celles-ci à d'autres modèles de processus stochastiques.

2 Classification pour des solutions d'équations différentielles stochastiques (E.D.S.) homogènes en temps

2.1 Présentation générale du problème

La classification pour des processus de diffusion a été étudiée dans [3] (et dans [2] et [6] dans un modèle de bruit blanc). Le modèle est le suivant :

$$\begin{cases} X_0 &= x_0 \\ dX_t &= b^*_Y(X_t)dt + \sigma(X_t)dW_t, \end{cases} \tag{1}$$

où $(W_t)_{t\geq 0}$ est un mouvement Brownien standard et tel que l'étiquette (*label*) Y est indépendant de $(W_t)_{t\geq 0}$ de loi inconnue $\mathbb{P}$ données par $(p_i)_{i\in\mathcal{Y}} = \{1, \dots, K\}$. Enfin, on note $(\mathcal{F}_t^X)_{t\geq 0}$ la filtration naturelle associée au processus X notée g_t. Supposons que le processus X est observé sur l'intervalle de temps $[0, T]$. Dans ce cadre, un classifieur (ou règle de classification) au temps $0 \leq t \leq T$ est une application $\mathcal{F}_t^X$-mesurable à valeur dans $\mathcal{Y}$. La performance d'une

règle de classification g_t est évaluée au travers de le mesure de risque suivante

$$R(g_t) = \mathbb{P}\left(g_t(X) \neq Y\right).$$

En particulier, si on note $\mathcal{G}_t$ l'ensemble des classifieurs au temps t, le classifieur de Bayes est défini comme le minimiseur du risque

$$g_t^* \in \underset{g_t \in \mathcal{G}_t}{\operatorname{argmin}}\, R(g_t),$$

et caractérisé par

$$g_t^*(X) = \underset{i \in \mathcal{Y}}{\operatorname{argmax}}\, \pi_t^*(i), \quad \text{avec} \quad \pi_t^*(i) = \mathbb{P}(Y = i | \mathcal{F}_t^X).$$

On peut montrer la formule suivante (voir [2, 3])

$$\pi_t^*(i) = \varphi_i(F_t) \quad \mathbb{P} - p.s. \tag{2}$$

où $F_t^i := \int_0^t \frac{b_i^*}{\sigma^2}(X_s)dX_s - \frac{1}{2}\int_0^t \frac{(b_i^*)^2}{\sigma^2}(X_s)ds.$ et $F_t = (F_t^1, \ldots, F_t^K)$, $\varphi_i : (x_1, \ldots, x_K) \mapsto \frac{p_i \mathrm{e}^{x_i}}{\sum_{j=1}^K p_j \mathrm{e}^{x_j}}$. Ainsi, pour construire une procédure de classification, on peut considérer des estimateur $\hat{\pi}_t(i)$ des probabilités $\pi_t^*(i)$ et définir le prédicteur associé

$$\hat{g}_t \in \underset{i \in \mathcal{Y}}{\operatorname{argmax}}\, \hat{\pi}_t(i).$$

Ces estimateurs vont être construit à partir d'un échantillon d'apprentissage de taille N noté $D_N = (X^{(j)}, Y^{(j)})_{j=1,\ldots,N}$ constitué de copies indépendantes de (X, Y). En outre, Nous supposons que les trajectoires sont observées à temps discret sur $[0, 1]$ et on utilise toute l'observation ($t = T$). Dans [3], deux procédures sont proposées, s'appuyant sur l'hypothèse selon laquelle le coefficient de dérive b^* inconnue est en fait une fonction connue à un paramètre près θ qui est alors le paramètre discriminant les classes ($b_Y^*(x) = b(\theta_Y^*, x)$). Par ailleurs, l'étude est effectuée dans le cas où le coefficient de diffusion σ est connu.

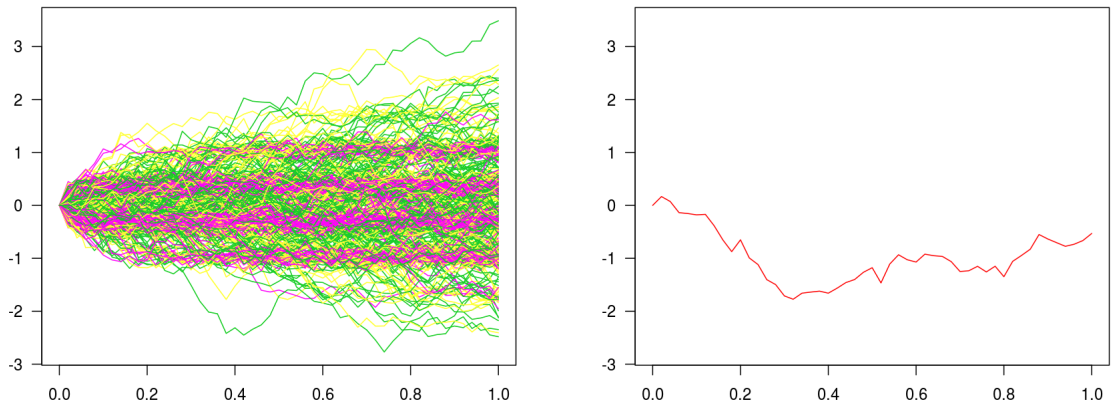

FIGURE 1 – Exemple d'échantillon d'apprentissage (étiqueté) à gauche, nouvelle observation à classifier à droite.

2.2 Développements théoriques et numériques

Une première partie de la thèse va consister à étendre les résultats obtenus par [3] au cas non-paramétrique c'est à dire que l'on ne fait pas d'hypothèse sur la forme de la fonction b_Y qui discrimine les classes. Pour construire un classifieur $\widehat{g}$ dans ce cadre deux méthodes seront explorées : le *plug-in* et la minimisation de risque empirique.

2.2.1 Estimation par *plug-in*

La première idée est de construire un classifieur *plug-in* c'est à dire s'appuyant sur un estimateur $\hat{b} = \{\hat{b}_1, \dots, \hat{b}_K\}$ où $\hat{b}_i$ est un estimateur de la dérive du processus issu de la classe i pour $i = 1, \dots, K$. Pour cela on pourra dans une première approche considérer l'estimateur proposé dans [4] qui s'appuie sur l'observation de N trajectoires discrètes du processus de diffusion. La procédure classification sera alors de la forme :

$$g_{\widehat{b}}(X) := \underset{i \in \mathcal{Y}}{\operatorname{argmax}}\ \pi_{\widehat{b}}(i),$$

où $\pi_{\widehat{b}}(i)$ est un estimateur de $\pi^*(i)$ au temps T fondé sur l'estimateur $\hat{b}$ et la discrétisation des trajectoires. L'enjeu de l'étude théorique de cette procédure sera de déterminer si les propriétés théoriques satisfaites par l'estimateur de b choisi sont suffisantes pour garantir la consistance de la procédure vis-à-vis du risque de mauvaise classification. En particulier, l'un des points clé sera l'étude de l'optimalité de l'estimateur obtenu du point de vue minimax.

2.2.2 Minimisation de risque empirique

Une seconde idée est de construire une règle de classification par minimisation du risque empirique. On va donc considérer $\hat{g}$ défini comme suit

$$\widehat{g} \in \underset{b \in \mathcal{B}}{\operatorname{argmin}}\ \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{\widehat{g}_b(X^{(j)}) \neq Y^{(j)}\}}.$$

Cette minimisation est effectuée sur une classe de fonctions $\mathcal{B}$ avec $b \in \mathcal{B}$ dont le choix est une question importante. L'étude théorique de cet estimateur a par exemple été effectué dans [2], néanmoins il n'est implémentable en pratique du fait que le problème de minimisation n'est ni convexe ni *smooth*. Pour palier ce problème, on s'appuie sur la méthode de convexification de risque [18]. Dans [3], cette méthode est considérée dans le cas paramétrique.

L'enjeu du travail de thèse serait d'étendre cette procédure au cas non-paramétrique. Il faudra pour cela faire appel à la théorie des recouvrements [9]. Une application possible de ce travail sera l'étude de procédures d'agrégation convexe de prédicteurs. Si là encore, l'étude des performances théoriques de l'estimateur obtenu sera primordiale, l'implémentation de cette méthode nécessitera une maitrise des outils d'optimisation. Il sera, en particulier, très intéressant d'étudier l'efficacité des algorithmes d'optimisation convexe de type gradient stochastique. Ce travail pourra être valoriser par le développement d'une librairie `Python` ou d'un package `R`.

2.3 Difficulté supplémentaire

Par ailleurs, un des points importants sera de considérer le cas où le coefficient de diffusion σ est inconnu. Il est connu que l'on peut estimer σ à partir d'observations discrètes du

processus lorsque l'on observe une trajectoire en haute fréquence (et les vitesses sont meilleures que celles obtenues pour le coefficient de dérive voir [7, 8, 12]). Cette difficulté additionnelle devra être étudiée dans chacun des cas (et pourra éventuellement lever une autre question : que se passe-t-il si les classes sont discriminées par b et σ ?)

3 Extension à des modèles plus généraux

Dans un deuxième temps, l'objectif de la thèse sera d'étendre à des modèles plus complexes, les résultats obtenus dans le cadre de la problématique présentée en Section 2.1.

3.1 Dimension supérieure

En vue d'applications en biologie ou en finance, il semble qu'il serait important d'étudier le cas de processus de $\mathbb{R}^d$ avec $d > 1$ (jusqu'ici nous avions supposé $d = 1$). Les procédures présentées devraient s'adapter en dimension supérieure. Cependant naturellement certaines inégalités sont plus difficiles à étudier, et certains outils doivent être utilisés spécialement pour la dimension supérieure.

3.2 Classification : cas de diffusion in-homogènes (en temps)

On pourra également considérer des solutions d'équations différentielles stochastiques du type :

$$dX_t = b_Y(X_t, t)dt + \sigma(X_t, t)dW_t$$

sous de bonnes hypothèses de régularité des coefficients. La formule (2) repose sur le rapport de vraisemblance, cela ne devrait pas poser de problème de se placer dans ce cadre plus général (voir par exemple [11]). Cette classe de processus de diffusion, beaucoup plus grande, permet de décrire des phénomènes plus complexes. Par exemple, en neuroscience, il semble que le potentiel de membrane d'un neurone entre deux *spikes* serait bien décrit par une équation différentielle stochastique dont le terme de dérive contient une composante dépendant du temps uniquement (voir par exemple [10]).

3.3 Classification : cas de processus de diffusion à sauts

Les processus de diffusion à sauts ont leur dynamique décrite par une équation différentielle stochastique comme (1) mais à laquelle on a ajouté un terme de "saut", généralement un processus de Lévy (voir par exemple [17]).

Ce type de processus est utilisé dans de nombreux domaines, notamment en finance. Ici encore, des estimateurs des coefficients fondés, par exemple, sur la vraisemblance existent. Par contre, les techniques diffèrent quelques peu des techniques pour les EDS classiques car il faut tenir compte des sauts (voir par exemple [14]).

4 Application à la dynamique cellulaire

L'objectif dans cette partie sera d'appliquer les méthodes de classification développées à des données de trajectoires cellulaires, en collaboration avec Christèle Etchegaray (CR Inria Bordeaux Sud-Ouest). Nous disposons de données de dynamiques cellulaires partagées

en différents groupes, soumis à diverses altérations pharmacologiques. Les trajectoires ainsi obtenues ont des morphologies bien distinctes. Il est classique de modéliser des trajectoires cellulaires au moyen de processus de diffusion ou de processus de Lévy [16]. Il est cependant pertinent de confronter ces modèles à des dynamiques cellulaires variées, afin de tester également leur capacité à discriminer entre différents comportements.

Nous travaillerons tout d'abord avec des modèles décrivant la dynamique du module de la vitesse en 1D. Le cas 2D pourra être abordé par la suite. Dans un premier temps, nous utiliserons un processus de diffusion dont le terme de dérive dépend d'un paramètre servant à quantifier le caractère persistant de la dynamique [5]. Cela reviendra à tester la procédure développée en [3]. Puis nous pourrons considérer les modèles plus généraux, abordés dans la thèse, qui sont particulièrement adaptés à la dynamique étudiée (dérive non paramétrique, coefficient de diffusion inconnu, processus de Lévy). Cela nous permettra éventuellement d'identifier le degré de généralisation nécessaire pour discriminer entre ces différents états cellulaires.

Références

[1] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 51(6) :2163–2172, 2005.

[2] B. Cadre. Supervised classification of diffusion paths. *Math. Methods Statist.*, 22(3) :213–225, 2013.

[3] Christophe Denis, Charlotte Dion, and Miguel Martinez. Consistent procedures for multiclass classification of discrete diffusion paths. *Scandinavian Journal of Statistics, to appear*, 2019.

[4] Christophe Denis, Charlotte Dion, and Miguel Martinez. A ridge estimator of the drift from discrete repeated observations of the solutions of a stochastic differential equation. *hal-02528092*, 2020.

[5] Christeèle Etchegaray. *Modélisation mathématique et numérique de la migration cellulaire*. PhD thesis, Paris Saclay, 2016.

[6] Sébastien Gadat, Sébastien Gerchinovitz, and Clément Marteau. Optimal functional supervised classification with separation condition. *arXiv preprint arXiv :1801.03345*, 2018.

[7] Valentine Genon-Catalot and Jean Jacod. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'IHP Probabilités et statistiques*, 29(1) :119–151, 1993.

[8] Arnaud Gloter. Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient. *ESAIM : Probability and Statistics*, 4 :205–227, 2000.

[9] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[10] R Höpfner. A time inhomogeneous cox-ingersoll-ross diffusion with jumps. *ArXiv e-prints*, 2009.

[11] Jean Jacod. Inference for stochastic processes. In *Handbook of Financial Econometrics : Applications*, pages 197–239. Elsevier, 2010.

[12] N/ Jakobsen, Munkholt and M Sørensen. Efficient estimation for diffusions sampled at high frequency over a fixed time interval. *Bernoulli*, 23(3) :1874–1910, 2017.

[13] Damien Lamberton and Bernard Lapeyre. *Introduction to stochastic calculus applied to finance*. Chapman and Hall/CRC, 2011.

[14] Cecilia Mancini. Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 36(2) :270–296, 2009.

[15] James O Ramsay and Bernard W Silverman. *Applied functional data analysis : methods and case studies*. Springer, 2007.

[16] P. Romanczuk, M. Bär, W. Ebeling, B. Lindner, and L. Schimansky-Geier. Active brownian particles. *The European Physical Journal Special Topics*, 202(1) :1–162, 2012.

[17] Ken-iti Sato. Basic results on lévy processes. In *Lévy processes*, pages 3–37. Springer, 2001.

[18] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct) :1225–1251, 2004.