Proposition de Sujet de Thèse

Algorithmes stochastiques pour les mélanges et les processus de déformation

Bernard Bercu et Jérémie Bigot
Université de Bordeaux, Institut de Mathématiques de Bordeaux
UMR 5251, 351 cours de la libération, 33405 Talence
Bernard.Bercu@u-bordeaux.fr
Jeremie.Bigot@u-bordeaux.fr

Les méthodes d'approximation stochastique sont des algorithmes récursifs qui permettent d'approcher les zéros ou les extréma de fonctions inconnues pouvant seulement être évaluées à partir d'observations bruitées. Les algorithmes stochastiques les plus célèbres et performants sont les algorithmes de Robbins-Monro et de Kiefer-Wolfowitz. Ils sont très largement utilisés en gestion de stocks, en ingénierie mathématique et financière, ainsi qu'en apprentissage compétitif. A titre d'exemple, l'agorithme de Kiefer-Wolfowitz permet d'estimer le maximum d'une fonction f strictement concave. Il satisfait la relation récursive

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n - \frac{\gamma_{n+1}}{2a_{n+1}}\big(Y_{n+1} - Z_{n+1}\big)$$

avec $\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = f(\widehat{\theta}_n - a_n)$ et $\mathbb{E}[Z_{n+1}|\mathcal{F}_n] = f(\widehat{\theta}_n + a_n)$, où les suites (γ_n), (a_n) sont déterministes, positives et décroissantes vers zéro avec

$$\sum_{n=1}^{\infty} \gamma_n = +\infty, \qquad \sum_{n=1}^{\infty} \gamma_n a_n < +\infty, \qquad \sum_{n=1}^{\infty} \Big(\frac{\gamma_n}{a_n}\Big)^2 < +\infty.$$

L'objectif de cette thèse est de proposer de nouveaux algorithmes stochastique de type Robbins-Monro et Kiefer-Wolfowitz afin d'estimer les paramètres inconnus des modèles de mélange ou des processus de déformation.

Partie I: Estimation semi-paramétrique pour les modèles de mélange.
On considère le modèle de mélange à deux composantes donné par

$$g(x) = pf(x - \theta_1) + (1 - p)f(x - \theta_2)$$

où la densité de mélange f est paire, la probabilité de mélange $0 < p < 1/2$ et les paramètres $(\theta_1, \theta_2) \in \mathbb{R}^2$ avec $\theta_1 \neq \theta_2$. Soit $(X_1, \ldots, X_n)$ une suite de variables aléatoires indépendantes et de même loi de densité g. Via la formule d'inversion

$$f(x) = \frac{1}{1-p} \sum_{k=0}^{\infty} \Big(\frac{-p}{1-p}\Big)^k g(x + \theta_2 + k(\theta_2 - \theta_1)),$$

Bordes, Mottelet, et Vandekerkhove [**?**], [**?**] ont proposé une procédure d'estimation statistique semi-paramétrique, basée sur $(X_1, \ldots, X_n)$, afin d'estimer p, θ_1, θ_2 ainsi que la densité de mélange f. Cette procédure d'estimation statistique est judicieuse mais elle fait appel à une fonction de contraste dont l'évaluation est loin d'être aisée. Laurent, Marteau, et Maugis-Rabusseau [**?**] ont récemment étudié le cas particulier du mélange gaussien sparse à deux composantes, c'est-à-dire pour lesquel f correspond à la densité de la loi $\mathcal{N}(0, 1)$ et la probabilité p est très petite devant $1/\sqrt{n}$. La première partie de la thèse consistera à proposer une nouvelle procédure d'estimation statistique des paramètres θ_1 et θ_2, basée sur l'algorithme stochastique de Kiefer-Wolfowitz, associée à l'estimation non-paramétrique de f et g par des estimateurs à noyau. On commencera par étudier le cas particulier du mélange gaussien à deux composantes.

Partie II: Estimation semi-paramétrique pour les processus de déformation.
Une classe importante de phénomènes réels survient de façon périodique. On peut penser à la médecine avec les signaux ECG, la biologie avec les rythmes circadiens, la météorologie avec les cycles de températures, le trafic routier avec les heures de pointe le matin et le soir, l'économétrie avec la consommation d'électricité, etc. Les processus de déformation périodiques sont couramment utilisés pour modéliser ce type de phénomènes. Ils sont donnés, pour tout $1 \leq i \leq n$ et $1 \leq j \leq p$, par

$$Y_{i,j} = a_j f(X_i - \theta_j) + v_j + \varepsilon_{i,j}$$

où f est une fonction inconnue de forme périodique tandis que $a = (a_1, \ldots, a_p)$, $\theta = (\theta_1, \ldots, \theta_p)$ et $v = (v_1, \ldots, v_p)$ sont des paramètres inconnus d'échelle, de translation et de centrage. La suite (X_n) correspond aux instants d'observation et les suites $(Y_{i,j})$, $(\varepsilon_{i,j})$ sont les sorties et les bruits associés au processus de déformation. Dans le cas particulier où $p = 1$, Bercu et Fraysse [**?**], ont proposé une procédure d'estimation efficace du paramètre de translation θ par l'algorithme stochastique projeté de Robbins-Monro. La fonction de régression f a été estimée simultanément par un estimateur à noyau de type Nadaraya-Watson récursif. Ce travail a récemment été généralisé par Fraysse [**?**], [**?**] dans le cadre multidimensionnel avec une application à la détection d'arythmie cardiaque par l'étude de signaux ECG. Les processus autorégressifs de déformation s'avèrent mieux adaptés pour modéliser certains phénomènes réels. La seconde partie de la thèse portera sur l'étude des processus autorégressifs de déformation défini, pour tout $1 \leq i \leq n$ et $1 \leq j \leq p$, par

$$X_{i,j} = a_j f(X_{i-1,j} - \theta_j) + v_j + \varepsilon_{i,j}$$

où les paramètres d'échelle et de translation seront estimés par des algorithmes stochastiques de type Robbins-Monro. Il est important d'observer qu'aucune analyse statistique n'a encore été réalisée sur ce type de processus.

Partie III: Estimation jointe de la variabilité en phase et en amplitude de mesures ponctuelles.
Dans cette troisième partie, on considère le cadre de données enregistrées chez un ou plusieurs individus ou bien au cours d'essais répétés. Ces observations peuvent être modélisées sous la forme d'un ensemble de variables aléatoires réelles $(X_{i,j})$ où l'indice i, variant de 1 à n, modélise un individu ou bien un numéro d'essai, tandis que l'indice j représente la j-ième observation pour le i-ème individu, pour lequel p_i mesures ont été effectuées. Pour de nombreux domaines d'applications liés aux neurosciences ou à la bio-informatique [**?**], [**?**], un cadre de modélisation possible est de considérer que, pour chaque $1 \leq i \leq n$, les observations $(X_{i,j})_{1 \leq j \leq p_i}$ sont des variables aléatoires indépendantes et de même loi de densité de probabilité f_i satisfaisant, pour tout $x \in \mathbb{R}$,

$$f_i(x) = \gamma_i'(x) f_0\left(\gamma_i(x)\right).$$

Dans cette modélisation,

(a) γ_i est une fonction croissante inconnue qui représente une source de variabilité en phase,

(b) f_0 est une densité de référence inconnue qui modélise une variabilité en amplitude.

Le cadre paramétrique des translations aléatoires inconnues pour lesquelles $\gamma_i(x) = x - \theta_i$, a récemment été étudié par Bigot et al. [**?**]. L'objectif de cette partie portera sur l'estimation jointe de la variabilité en phase et en amplitude. Dans ce contexte, une approche appropriée consiste à utiliser des outils basés sur la théorie du transport optimal et sur la distance de Wasserstein entre mesures de probabilités, voir Bigot et al. [**?**]. Les méthodes d'estimation statistique basées sur des algorithmes stochastiques du type Robbins-Monro et Kiefer-Wolfowitz n'ont pas encore été développées. Il s'agit donc d'une approche novatrice avec de nombreuses perspectives d'applications en biologie et neurosciences.

Références

[1] B. Bercu and P. Fraysse, A Robbins-Monro procedure for estimation in semiparametric regression models, Ann. Statist. 40, pp. 666-693, 2012.

[2] J. Bigot, S. Gadat, T. Klein, and C. Marteau, Intensity estimation of non-homogeneous poisson processes from shifted trajectories, Electronic Journal of Statistics 7, pp. 881-93, 2013.

[3] J. Bigot, R. Gouet, T. Klein, and A. Lopez, Geodesic PCA in the Wasserstein space by Convex PCA, To appear in Annales de l'Institut Henri Poincaré B: Probability and Statistics, 2016.

[4] L. Bordes, S. Mottelet, and P. Vandekerkhove, Semiparametric estimation of a two-component mixture model, Ann. Statist. 34, pp. 1204-1232, 2006.

[5] L. Bordes and P. Vandekerkhove, Semiparametric two-component mixture model with a known component: an asymptotically normal estimator, Math. Methods Statist. 19 pp. 22-41, 2010

[6] P. Fraysse, Recursive estimation in a class of models of deformation, J. Statist. Plann. Inference 147, pp. 132-158, 2014.

[7] P. Fraysse, Estimation of the shift parameter in regression models with unknown distribution of the observations, Electron. J. Stat. 8, pp. 998-1028, 2014.

[8] D. Johnson, A. Mortazavi, R. Myers, and B. Wold, Genome-wide mapping of in vivo protein-DNA interactions, Science 316, pp. 1497-1502, 2007.

[9] B. Laurent, C. Marteau, and C. Maugis-Rabusseau, Non-asymptotic detection of two-component mixtures with unknown means, Bernoulli 22, pp. 242-274, 2016.

[10] W. Wu and A. Srivastava, An information-geometric framework for statistical inferences in the neural spike train space, Journal of Computational Neuroscience 31, pp. 725-748, 2011.