Projet de thèse : Régression quantile extrême en grande dimension

Directeurs de thèse : Gilles Stupfler (Université d'Angers, 50%)
et Antoine Usseglio-Carleve (Avignon Université, 50%)

Contexte Bien que dans la majorité des applications de la statistique, on souhaite estimer et interpréter des quantités dites « centrales » (une moyenne, une médiane, des quartiles...), certaines applications requièrent au contraire de se concentrer sur des paramètres ou indicateurs « extrêmes » de la distribution sous-jacente au phénomène étudié. Par exemple, en gestion de risque en finance et en assurance, la directive Solvabilité II de l'Union Européenne impose qu'une compagnie d'assurance ayant ses activités au sein de l'Union calcule et provisionne une quantité de fonds suffisante pour couvrir ses obligations sur les 12 mois à venir avec probabilité 0.995, ce qui revient à calculer un quantile extrême (au niveau 0.995) de la distribution de ses pertes. Dans ce cadre réglementaire, il est crucial pour l'entreprise d'étudier avec précision les indemnisations qu'elle verse, et tout particulièrement le comportement probabiliste des « grandes » indemnisations, afin d'éviter non seulement une sous-estimation du risque qui la mettrait en difficulté vis-à-vis du régulateur, mais aussi une sur-estimation de ce risque qui pourrait l'inciter à se montrer trop prudente et donc non compétitive. Ce type d'applications constitue la motivation de la théorie moderne des valeurs extrêmes.

Dans l'exemple actuariel précédent, le calcul final de la prime d'assurance se fait avec des données client, souvent de grande dimension. En finance, modéliser le cours d'un actif boursier peut se faire au moyen de processus aléatoires, ou de modèles de régression comme les modèles de type GARCH. De façon générale, une meilleure compréhension des extrêmes d'une variable aléatoire passe souvent par l'intégration d'une covariable à l'analyse statistique : pourtant, les travaux existants sur l'analyse statistique de valeurs extrêmes en présence d'une covariable dans des modèles de régression explicites restent la plupart du temps sans fondement théorique solide. Deux exceptions sont les travaux récents de [1] et [4], qui fournissent un cadre théorique et pratique à l'analyse conditionnelle de valeurs extrêmes, grâce à des résultats généraux obtenus sur les résidus de modèles de régression. Ces travaux comportent néanmoins des conditions de type uniforme assez fortes, une hypothèse de queue lourde sur les erreurs des modèles, et leur utilisation pratique se limite à la petite dimension.

Objectifs Ce projet de thèse vise à lever ces restrictions afin de pouvoir développer des procédures d'estimation et d'inférence de valeurs extrêmes conditionnelles dans un cadre très général. On se concentrera sur des modèles de régression dits location-scale, qui permettent non seulement de modéliser des jeux de données de grande dimension mais aussi des séries temporelles complexes. On étudiera en particulier comment intégrer des méthodes de réduction de dimension type LASSO pour obtenir des estimateurs valides dans un cadre de grande dimension. Ceci pourra demander l'utilisation et/ou le développement d'inégalités de concentration qui s'inscriront dans la lignée de résultats récents, comme ceux de [2, 3]; il sera alors important

d'obtenir des versions non asymptotiques de résultats existants sur des estimateurs de valeurs extrêmes construits sur des résidus d'un modèle de régression. On s'attachera à appliquer les résultats obtenus à l'obtention de procédures d'inférence (intervalles de confiance, tests) dans un ou plusieurs jeux de données réelles, par exemple en détection d'anomalies industrielles, en sciences climatiques, ou en assurance/finance.

Profil recherché Master de Mathématiques, de préférence comprenant une composante statistique et/ou machine learning substantielle. Une appétence pour la statistique mathématique est indispensable. Une compétence en `R` et/ou `Python` sera appréciée mais n'est pas obligatoire.

Bibliographie

[1] A. Ahmad, E. Deme, A. Diop, S. Girard, A. Usseglio-Carleve (2020). Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model, *Electronic Journal of Statistics* **14**: 4421–4456.

[2] S. Boucheron, M. Thomas (2012). Concentration inequalities for order statistics, *Electronic Communications in Probability* **17**: 1–12.

[3] S. Boucheron, M. Thomas (2015). Tail index estimation, concentration and adaptivity, *Electronic Journal of Statistics* **9**: 2751–2792.

[4] S. Girard, G. Stupfler, A. Usseglio-Carleve (2021). Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models, *Annals of Statistics* **49**: 3358–3382.