Rapport sur le Mémoire de Nicolas Goix

le 29 Septembre 2016, à Paris

La dissertation de Nicolas Goix est intitulée *Apprentissage automatique et Extrêmes pour la détection d'anomalies*. Elle est organisée en cinq parties :

i) un résumé de l'ensemble du travail en une vingtaine de pages (*Summary*)
ii) une partie destinée à présenter des outils utilisés de manière récurrente dans le reste du mémoire (*Preliminaries*)
iii) le développement d'un critère *excès de masse* pour le rangement d'anomalies (*Anomaly ranking*)
iv) la confrontation des techniques d'apprentissage et de la théorie des valeurs extrêmes
v) le développement d'approches heuristiques efficaces

Le résumé

Le résumé donne un aperçu bref et fidèle du contenu de la thèse. Il décrit un travaille qui couvre des domaines très différents, va du théorique à la programmation et à l'évaluation expérimentale. Ce résumé donne une bonne idée de l'ampleur de la thèse.

Les préliminaires

Ce chapitre d'introduction est remarquablement bien écrit. Il fait de la thèse de Nicolas Goix un outil précieux pour les lecteurs qui s'intéressent à la fois aux valeurs extrêmes et à l'apprentissage. Ils trouveront là une présentation homogène bien organisée qui leur fera gagner un temps précieux.

N. Goix décrit les trois axes de son travail :

i) l'étude des méthodes de M-estimation dans le développement des *scoring functions* (la détection d'anomalies, sujet central, peut se concevoir comme l'usage d'une fonction de *scoring*).
ii) la confrontation entre la Statistique des extrêmes et les méthodes de la théorie de l'apprentissage.
iii) le développement logiciel et l'expérimentation (les approches heuristiques).

Cet aperçu rapide montre que N. Goix a abordé dans sa thèse des spécialités qui sont habituellement séparées par des parois étanches. La présentation élégante et efficace suggère que N. Goix s'est formé une culture véritable à propos des méthodes de processus empiriques (de celles qui sont pratiquées en apprentissage, qui produisent des énoncés non-asymptotiques), de la statistique des extrêmes multivariés (domaine difficile et classiquement très asymptotique), et, ce qui est remarquable, en développement logiciel.

Le chapitre 2 intitulé *Concentration inequalities from the method of bounded differences* introduit les inégalités classiques utilisés par N. Goix dans ses travaux originaux. Nicolas Goix prend un parti minimaliste. Il énonce et parfois démontre les versions « martingales » des inégalités exponentielles classiques (Hoeffding, Bernstein et Bennett) et l'inégalité de McDiarmid qui s'ensuit. Il décrit en fin de chapitre les inégalités de Vanik-Chervonenkis et en particulier les versions normalisées qui lui seront utiles par la suite.

Le chapitre 3 intitulé *Extreme Value Theory*. N. Goix survole la théorie univariée en énonçant le théorème de Fisher-Tippett-Gnedenko (mais pas celui de Balkema-de Haan-Pickands, pourtant plus en harmonie avec son approche de la statistique des extrêmes). Il décrit ensuite les lois d'extrêmes multivariées (celles qu'on obtient comme loi limite des maximas par coordonnées) et leur caractérisations (mesures exposants, L-fonctions, etc).

Les préliminaires s'achèvent par une présentation des algorithmes de détection d'anomalies. La construction d'un algorithme de détection d'anomalies est présentée classiquement comme la construction d'un test entre une hypothèse nulle simple formée d'une loi (inconnue, il faut l'« apprendre ») et une alternative constituée par les autres lois ou plutôt par un ensemble de lois (peu ou mal spécifié) absolument continues par rapport à une loi qui domine l'hypothèse nulle. Pour fournir un test de niveau α, une littérature fournie suggère d'estimer un niveau de densité. Cette approche a été développée en apprentissage [Voir Steinwart et al., 2005, Vert and Vert, 2006, Vert and Rigollet, 2006, entre autres]. N. Goix observe que l'on peut détecter les anomalies en utilisant une fonction de notation (*scoring function*), idéalement une transformation monotone de la densité. N. Goix conduit une discussion pointue des mérites des différentes approches classiques de la détection d'anomalie.

N. Goix décrit trois techniques de détection d'anomalies : les SVM à une classe, la méthode *local outlier factor* (une méthode de plus proches voisins) et les *isolation forest* (inspirée des forêts aléatoires). N. Goix a codé ces deux méthodes dans module `python` `scikit-learn`.

Un critère d'excès de masse pour les anomalies

Cette partie est constituée par le chapitre 5 (*On anomaly ranking and excess mass curves*). Plutôt qu'à la détection d'anomalies, N. Goix s'intéresse au tri des anomalies (*anomaly ranking*).

N. Goix rappelle les inconvénients du critère *mass-volume curve* pour la construction de fonctions de notation et défend le critère *excess mass curve* en s'appuyant sur les travaux historiques de [Polonik, 1995, Mueller and Sawitzki, 1991]. N. Goix défend cette approche duale d'un point de vue statistique et d'un point de vue calculatoire.

Les ensembles qui optimisent les critères d'excès de masse (ou de volume minimal) sont, sans surprise, comparables aux régions critiques des tests de Neyman-Pearson. Le critère proposé par N. Goix l'emporte sur le critère masse-volume dans presque toutes les catégories (sans toutefois contourner la difficulté qui consiste à approcher le volume d'ensembles complexes).

Le chapitre culmine avec les théorèmes 13 et 14 qui fournissent des bornes de type Vapnik-Chervonenkis pour l'écart maximal entre la courbe d'excès de masse optimale et la courbe d'excès de masse construite à partir des données selon la méthode proposée par N. Goix.

Ce chapitre a fait l'objet d'une communication au colloque AISTATS en 2015.

Apprentissage et théorie des valeurs extrêmes

Le chapitre 6 intitulé *Learning the dependence structure of rare events: a non-asymptotic study* confronte les outils de la théorie de l'apprentissage (les inégalités de Vapnik-Chervonenkis) et un problème important en Statistique des extrêmes : l'estimation de la fonction L (c'est à dire une estimation d'une caractérisation de mesure exposant d'une loi max-stable multivariée).

Dans le cas bivarié

$$L(x,y) := -\log G\left(\frac{x^{-\gamma_1}-1}{\gamma_1}, \frac{y^{-\gamma_2}-1}{\gamma_2}\right)$$

où G est la fonction de répartition d'une loi d'extrême bivariée d'indices marginaux γ_1, γ_2 qui définit le max-domaine d'attraction de la loi échantillonnée. La standardisation par les rangs permet de définir une version empirique de L (cette version empirique a cependant l'inconvénient de ne pas respecter les contraintes de forme satisfaites par les L-fonctions comme la convexité). Cette version empirique $\widehat{L}$ sert d'estimateur de la fonction L.

En observant justement que les processus empiriques impliqués dans la définition de $\widehat{L} - L$ sont des classes de Vapnik (dont la dimension coïncide avec la dimension de l'espace où vivent les observations), N. Goix *et al.* fournissent (pour la première foix à ma connaissance) une inégalité de déviation non-asymptotique pour

$$\sup_{\mathbf{0}\leq\mathbf{x}\leq T} \left|\widehat{L}(\mathbf{x}) - L(\mathbf{x})\right| .$$

Le théorème 6.6 constitue un complément non-trivial et bienvenu de résultats asymptotiques comme par exemple la consistance décrite dans le théorème 7.2.1 de de Haan and Ferreira [2006]. La preuve utilise une version raffinée des inégalités de Vapnik-Chervonenkis (Théorème 6.1) qui convient bien aux classes de parties dont la réunion est de petite probabilité (ce qui est le cas dans le scenario envisagé par N. Goix *et al.*). Il serait intéressant de produire des résultats de concentration est combinant l'inégalité de Talagrand (qui est sensible à la normalisation) et les majorant fins de l'espérance des suprema de processus empiriques indexés par les classes de Vapnik (comme par exemple celui décrit dans le théorème 13.7 de [Boucheron et al., 2013]). On aimerait aussi confronter le théorème de N. Goix *et al.* avec les théorèmes centraux limites fonctionnels pour $\widehat{L}$ comme le théorème 7.2.2 dans [de Haan and Ferreira, 2006].

La démarche mise en place par N. Goix *et al.* ouvre la voie à l'élaboration de techniques d'estimation adaptative pour la fonction L (ou de manière équivalente pour l'estimation de la structure de dépendance de la loi d'extrêmes multivariée).

Dans le chapitre 7, intitulé *Sparse representation of multivariate extremes*, N. Goix revient sur les lois d'extrêmes multivariées, un ensemble infini dimensionnel de lois et se demande quelles lois admettent une représentation parcimonieuse (*sparse*).

La volonté d'estimer la mesure spectrale plutôt que la STDF peut se justifier de plusieurs façons, en particulier en notant que les estimateurs de la STDF ne sont pas toujours des STDF alors que toute mesure vérifiant quelques conditions marginales définit une STDF [de Haan and Ferreira, 2006, Théorème 6.1.14 et commentaires p. 247]. Une mesure spectrale est une mesure sur l'orthant positif d'une hypersphère (définie par une norme $\ell_\infty, \ell_1, \ldots$).

En s'interessant à la mesure spectrale plutôt qu'à la mesure exposant, N. Goix est en mesure de proposer une conception personnelle de la parcimonie. Une mesure spectrale est parcimonieuse si elle ne charge que quelques sous-cônes de dimension « petite »par rapport à la dimension des observations (supposées grandes). A ma connaissance, il s'agit d'une des premières tentatives d'importer le point de vue de la parcimonie dans le monde de la statistique des extrêmes. L'objectif proclamé est de décrire la structure de dépendance dans les régions extrêmes (une entreprise déjà bien engagée dans le cas bivarié). Estimer les sous-cônes de petite dimension sur lesquels la mesure spectrale est concentrée ne va pas de soi : il s'agit d'une concentration asymptotique qui n'a pas a être vérifiée par les lois d'excès empiriques. N. Goix propose une manière de contourner cet écueil en étudiant la mesure empirique d'épaississements des sous-cônes.

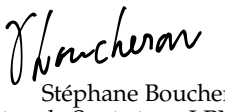
En s'appuyant sur les mêmes versions des inégalités de Vapnik que dans le chapitre 6, N. Goix aboutit (Théorème 7.12) à une borne de risque pour son estimateur de la mesure spectrale. La construction de cette borne démontre de la maitrise technique et de la ténacité. Ce résultat est d'autant plus remarquable que dans le cas des extrêmes multivariés, on a du mal à concevoir la possibilité d'utiliser des inégalités de concentration prêtes à l'emploi comme

dans le cas univarié où transformation quantile, représentation de Rényi et inégalité de concentration de Talagrand pour les échantillons exponentiels se combinent harmonieusement.

Il est encore tôt pour affirmer que cette approche de la parcimonie convient à l'étude des extrêmes multivariés et que c'est la seule pertinente. Les fonctions sur l'hypersphère peuvent être décrites sur des *frames* intéressants comme les *needlets*, et on peut concevoir qu'une focntion (par exemple la densité spectrale) est parcimonieuse lorsque la suite de ses coefficients est elle-même parcimonieuse. On peut aussi être tenté d'explorer les liens entre extrêmes et processus de Poisson et tenter d'importer le savoir faire disponible en matière d'estimation adaptative d'intensité de processus de Poisson [Reynaud-Bouret, 2003].

Conclusion

Le rapport est très bien écrit, instructif et agréable à lire. Les travaux de Nicolas Goix ont fait l'objet de trois communications dans des conférences sélectives (COLT, AISTATS), et sont soumis à des journaux de très bon niveau (comme *Journal of Multivariate Analysis*). A côté de son travail théorique, N. Goix a démontré un savoir-faire logiciel en participant au développement de `scikit-learn`, un module Python consacré à l'apprentissage, module réputé pour sa cohérence et sa solidité logicielle. Nicolas Goix a déployé une activité de recherche importante, diversifiée, originale et audacieuse. Pour toutes ces raisons, je soutiens avec enthousiasme la défense de la thèse de Nicolas Goix.



Stéphane Boucheron
Equipe de Statistique LPMA
Université Paris-Diderot
http://www.lpma-paris.fr/~boucheron

BIBLIOGRAPHY

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, 2013.

L. de Haan and A. Ferreira. *Extreme value theory*. Springer-Verlag, 2006.

D. Mueller and G. Sawitzki. Excess mass estimates and tests for multimodality. *J. Am. Stat. Assoc.*, 86(415):738–746, 1991.

W. Polonik. Measuring mass concentrations and estimating density contour clusters-an excess-mass approoach. *Annals of Statistics*, 23:855–881, 1995.

P. Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous poisson processes via concentration inequalities. *Probability Theory and Related Fields*, 126:103–153, 2003.

I. Steinwart, D. Hush, and C. Scovel. Density level detection is classification. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1337–1344. MIT Press, Cambridge, MA, 2005.

R. Vert and P. Rigollet. Plug-in methods for density level-set estimation: consistency and fast convergence rates. Technical report, LRI, 2006.

R. Vert and J. Vert. Consistency of one-class svm and related algorithms. *Journal of Machine Learning Research*, 2006. to appear.