Apprentissage Automatique et Extrêmes pour la Détection d'Anomalies

– RESUME en Français –

Nicolas Goix

1 Introduction

Une *anomalie*, du grec $\alpha\nu\omega\mu\alpha\lambda\iota\alpha$, aspérité, irrégularité, "non-semblable" (an-homalos), désigne un écart par rapport à une certaine normalité, par rapport à un comportement attendu. On appelle *anomalie* l'objet qui induit cet espace, l'observation qui s'écarte de la normalité. Dans beaucoup de domaines, la situation suivante se présente : un expert cherche à prédire un phénomène sur la base d'observations antérieures. Le cas le plus fondamental est lorsque l'on veut prédire certaines caractéristiques binaires d'observations nouvelles, compte tenu des précédentes. Par exemple, on peut penser à un médecin voulant prédire si un nouveau patient présente ou non une certaine pathologie, en utilisant les données des patients précédents (comme l'âge, l'histoire, le sexe, la pression artérielle) associées à leur véritable **étiquette/label** : avoir ou non la pathologie en question. Ce cas est un exemple de *classification binaire*, où le médecin cherche à trouver une règle pour **prédire** l'étiquette d'un nouveau patient (ce dernier étant caractérisé par son dossier médical, contenant toutes les mesures qui lui ont été faites). Cette règle est appelée un *classifieur* et doit être construite, *apprise*, sur des dossiers médicaux précédents. Intuitivement, le classificateur prédit le même diagnostic pour des dossiers médicaux similaires, dans un sens qui doit être appris avec précision.

On peut distinguer deux cas. Si les étiquettes des patients antérieurs sont connues (porteur ou non de la pathologie), on dit que la tâche de classification est **supervisée**. Si les étiquettes de ces données d'entrainement sont inconnues, la classification est dite **non-supervisée**. Suivant notre exemple, le médecin doit trouver deux formes (ou cluster) distinctes dans les données, correspondant aux deux étiquettes, "en bonne santé" - "malade", formes qui contiennent chacune des dossiers de patients similaires.

La détection d'anomalies survient lorsqu'une étiquette est fortement sous-représentée dans les données d'entrainement, par exemple si très peu de patients ont la pathologie dans les données d'entrainement. Ainsi, la **détection d'anomalies supervisée** se résume à la classification supervisée de classes fortement déséquilibrées. En ce qui concerne la **détection d'anomalies non supervisée** (également appelée simplement **détection d'outliers**), elle suppose généralement que la

base de données a un modèle "normal" caché, et les anomalies sont des observations qui s'écartent de ce modèle. Le médecin veut trouver des dossiers médicaux qui s'écartent de la grande majorité de ceux de ses patients précédents. Sa tâche est en quelque sorte simplifiée s'il sait que tous ses patients antérieurs sont en bonne santé : il est plus facile pour lui d'apprendre le modèle "normal", c'est-à-dire le dossier médical typique d'un patient en bonne santé, à confronté avec les dossiers médicaux de ses nouveaux patients. Ce cadre est celui de la **détection de nouveauté** – également appelé classification à une classe ou détection d'anomalies semi-supervisées : les données d'entraînement ne contiennent que des instances normales.

Ce résumé introductif est organisé de la façon suivante. Section 2, la détection d'anomalies est formellement introduite, ainsi que la notion de fonction de score. Deux critères sur la qualité d'une fonction de score sont ensuite présentés section 3. La section 4 se concentre sur la théorie des valeurs extrêmes (TVE) pour gagner en précision sur les régions extrêmes. Après avoir introduit la STDF (stable tail deviation function) représentant la structure de dépendance des événements rares (section 4.1), on montre que la théorie des extrêmes multivariées peut être utile pour produire des fonctions de score précise sur les régions de faible probabilité (section 4.2). La section 5 regroupe deux contributions de nature heuristique portant d'une part sur l'évaluation / la sélection d'algorithmes de détection d'anomalies non supervisés (section 5.1) et d'autre part sur l'extension des forêts aléatoires à la classification à une classe (section 5.2). La section 6 présente les contributions relatives à la librairie open-source scikit-learn. La section 7 énumère les productions scientifiques et conclut.

Notations. A travers ce document, $\mathbb{N}$ désigne l'ensemble des entiers naturels, $\mathbb{R}$ and $\mathbb{R}_+$ désignent respectivement l'ensemble des nombres réels et celui des nombres réels positifs. Les ensembles sont généralement écrit en lettre calligraphiques comme $\mathcal{G}$, et $|\mathcal{G}|$ désigne le nombre d'éléments dans $\mathcal{G}$.

Les vecteurs sont écrits en minuscules et en gras. Pour un vecteur $\mathbf{x} \in \mathbb{R}^d$ et $i \in \{1, \ldots, d\}$, x_i désigne la i^{me} composante de $\mathbf{x}$. Le produit scalaire entre deux vecteurs est noté $\langle \cdot, \cdot \rangle$. $\| \cdot \|$ désigne une norme arbitraire (sur des vecteurs ou sur des matrices) et $\| \cdot \|_p$ la norme L_p.

Au long de cette thèse, $\mathbb{P}[A]$ représente la probabilité de l'évènement $A \in \Omega$, l'espace de probabilité sous-jacent étant $(\Omega, \mathcal{F}, \mathbb{P})$. Nous utilisons la notation $\mathbb{E}[X]$ pour indiquer l'espérance de la variable aléatoire X. La notation $X \overset{d}{=} Y$ signifie que X et Y sont égales en distribution et $X_n \overset{d}{\to} Y$ signifie que (X_n) converge vers Y en distribution. Nous utilisons souvent l'abréviation $\mathbf{X}_{1:n}$ pour désigner un échantillon $i.i.d.$ $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$.

2 Détection d'anomalies, ranking d'anomalies et fonctions de scores

D'un point de vue probabiliste, il existe différentes façons de modéliser les comportements normaux et anormaux, ce qui conduit à différentes méthodologies. Un modèle probabiliste naturel consiste à supposer deux processus de génération différents pour les données normales et anormales. Les données normales (resp. données anormales) sont générées selon une distribution F (respectivement G). La distribution sous-jacente générale est alors un mélange de F et G. L'objectif est de déterminer si une nouvelle observation $\mathbf{x}$ a été générée à partir de F ou de G. Le meilleur moyen de résoudre théoriquement ce problème est le test du rapport de vraisemblances, également appelé test de Neyman-Pearson. Si $(\mathrm{d}F/\mathrm{d}G)(\mathbf{x}) > t$ avec un certain seuil $t > 0$, alors $\mathbf{x}$ a été généré selon F. Sinon, $\mathbf{x}$ a été généré selon G. Cela revient à estimer l'*ensemble de niveau de densité* $\{\mathbf{x}, (\mathrm{d}F/\mathrm{d}G)(\mathbf{x}) > t\}$ (Schölkopf et al., 2001; Steinwart et al., 2005; Scott & Nowak, 2006; Vert & Vert, 2006). Comme les anomalies sont très rares, leur structure ne peut être observée dans les données, en particulier leur distribution G. Il est courant et commode de remplacer G dans le problème ci-dessus par la mesure de Lebesgue, de sorte qu'il se résume à l'estimation du niveau de densité de F. (Schölkopf et al., 2001; Scott & Nowak, 2006; Vert & Vert, 2006).

Cela revient à supposer que les anomalies sont uniformément réparties sur le support de la distribution normale. Cette hypothèse est donc implicitement faite par une majorité d'ouvrages sur la détection de nouveauté / classification à une classe. Nous observons les données dans $\mathbb{R}^d$ à partir de la classe normale seulement, avec une distribution sous-jacente F et avec une densité $f : \mathbb{R}^d \to \mathbb{R}$. Le but est d'identifier les caractéristiques de cette classe normale, telles que son support $\{\mathbf{x}, f(\mathbf{x}) > 0\}$ ou un certain niveau de densité fixé $\{\mathbf{x}, f(\mathbf{x}) > T\}$ avec $t > 0$ près de 0.

La détection d'anomalies non supervisée est souvent considérée comme un problème de classification à une classe, où les données d'entraînement sont polluées par quelques éléments de la classe anormale : elle fait appel à des algorithmes à une classe *robustes* aux anomalies.

Une idée naturelle pour estimer les ensembles de niveau de densité est de calculer une estimation de la densité et de considérer les ensembles de niveau associés (Tsybakov, 1997; Cuevas & Fraiman, 1997; Baillo et al., 2001; Baillo, 2003; Cadre, 2006; Rigollet & Vert, 2009; Mason & Polonik, 2009). La densité est généralement estimée à l'aide d'un estimateur à noyau non paramétrique ou d'un estimateur de maximum de vraisemblance à partir d'une famille paramétrique de fonctions. Mais ces méthodes ne s'adaptent pas bien à la grande dimension. D'une certaine manière, ces méthodes cherchent à capturer plus d'information que né-

cessaire pour la tâche d'estimation d'ensemble de niveau, comme les propriétés locales de la densité qui sont inutiles pour cette tache. En effet, il s'avère que pour toute transformation croissante T, les ensembles de niveau de $T \circ f$ sont exactement ceux de f. Ainsi, il suffit d'estimer n'importe quel représentant de la classe des transformées croissantes de f, pour obtenir des estimés d'ensemble de niveau. Intuitivement, il suffit d'estimer le pré-ordre (le *scoring*) induit par f sur $\mathbb{R}^d$. Définissons une *fonction de score* comme toute fonction mesurable $s : \mathbb{R}^d \to \mathbb{R}_+$ intégrable par rapport à la mesure de Lebesgue Leb$(.)$ et $\mathcal{S}$ l'espace de toutes les fonctions de score. Toute fonction de score définit un pré-ordre sur $\mathbb{R}^d$ et donc un classement sur un ensemble de nouvelles observations. Ce classement peut être interprété comme un degré d'anormalité, plus $s(x)$ est petit, plus x est normal. Notons que la plupart des algorithmes de détection d'anomalies renvoient plus qu'une étiquette binaire, normale / anormale. Ils renvoient une fonction de score, qui peut être convertie en prédiction binaire, généralement en imposant un seuil basé sur sa distribution statistique.

Supposons que nous voulons apprendre une fonction de score s dont le pré-ordre induit est "proche" de celui de f, ou de manière équivalente dont les ensembles de niveau induits sont proches de ceux de f. Le problème est de transformer cette notion de proximité en critère $\mathcal{C}$, les fonctions de score optimales s^* étant alors définies comme celles qui optimisent $\mathcal{C}$. Dans le cadre de l'estimation de la densité, la différence uniforme $\|f - \hat{f}\|_\infty$ est un critère commun pour évaluer la qualité de l'estimation. Nous aimerions un critère similaire, mais qui est invariant par transformé croissante de $\hat{f}$. En d'autres termes, le critère doit être défini de telle sorte que la collection d'ensemble de niveau d'une fonction de score optimale $s^*(x)$ coïncide avec celle relative à f, et toute transformation croissante de la densité devrait être optimale au sens de $\mathcal{C}$. Plus formellement, nous allons considérer $\mathcal{C}^\Phi(s) = \|\Phi(s) - \Phi(f)\|$ (au lieu de $\|s - f\|$) avec $\Phi : \mathbb{R} \to \mathbb{R}_+$ vérifiant $\Phi(T \circ s) = \Phi(s)$ pour toute fonction de score s et transformation croissante T. Ici $\Phi(s)$ désigne soit la courbe masse-volume MV_s de s, soit sa courbe en excès-masse EM_s, définies dans la section suivante.

Ce critère qui mesure la qualité d'une fonction de score est alors un outil pour construire / apprendre une bonne fonction de score. Selon le paradigme de la minimisation du risque empirique, une fonction de score est construite en optimisant une version empirique $\mathcal{C}_n(s)$ du critère sur un ensemble adéquat de fonctions de score $\mathcal{S}_0$ de complexité contrôlée (par exemple une classe de dimension VC finie).

La section suivante décrit deux critères fonctionnels au vue de la nature globale du problème, tout comme les courbes ROC (*Receiver Operating Characteristic*) et PR (*Precision-Recall*), et qui sont admissibles par rapport aux exigences énumérées ci-dessus. Ces critères fonctionnels étendent en quelque sorte le concept de la courbe ROC au cadre non-supervisé.

Remarque 1. Terminologie : détection d'anomalies, ranking d'anomalies. *À proprement parler, les critères que nous recherchons sont des critères de* ranking *d'anomalies, de la même manière que la courbe ROC est essentiellement un critère de* ranking *bipartite. En pratique comme mentionné ci-dessus, tous les algorithmes de détection d'anomalies sont candidats à la tâche de ranking d'anomalie. Ils produisent tous une fonction de score, même ceux qui traitent à l'origine du cadre de "classification des anomalies", c'est à dire cherchent à être optimal sur un seul ensemble de niveau ou pour un taux de faux positifs fixe. Dans la littérature, la terminologie "détection d'anomalies" est largement utilisée, au lieu de la terminologie plus précise de "ranking d'anomalies". Par exemple, Liu et al. (2008) écrit "*Le but de la détection d'anomalies est de fournir un ranking qui reflète le degré d'anomalie*". Dans le cadre de ce travail, nous optons de même pour la convention que la détection d'anomalies se réfère au ranking d'anomalies : si les labels sont disponibles pour l'étape d'évaluation, l'objectif est de maximiser l'aire sous la courbe ROC. Si aucune donnée labelisée n'est disponible, l'objectif est de maximiser les critères non-supervisées définis dans la section suivante.*

3 M-estimation et critères de performance pour les fonctions de scores

Cette section est basée sur le travail Goix et al. (2015c). Nous fournissons un bref aperçu du critère de la courbe masse-volume introduit dans Clémençon & Jakubowicz (2013), qui est basé sur la notion d'ensembles de volume minimum. Nous exposons ensuite les principaux inconvénients de cette approche et proposons un autre critère, la courbe d'excès de masse.

3.1 Ensembles à volume minimal

La notion d'ensemble de volume minimal (Polonik (1997); Einmahl & Mason (1992)) a été introduite pour décrire des régions où une variable aléatoire multivariée $\mathbf{X} \in \mathbb{R}^d$ se trouve avec très grande ou très petite probabilité. Soit $\alpha \in (0, 1)$, un ensemble de volume minimal Γ_α^* de masse au moins α est une solution du problème de minimisation sous contrainte

$$\min_{\Gamma \text{ borelien}} \text{Leb}(\Gamma) \text{ tel que } \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha, \tag{1}$$

le minimum étant pris sur tous les sous-ensembles mesurables Γ de $\mathbb{R}^d$. On peut montrer que chaque niveau de densité est un ensemble de volume minimal pour une certaine masse et que la réciproque est vraie si la densité n'a pas de

partie plate. Dans le reste de cette section, on suppose que F a une densité $f(x)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ satisfaisant les hypothèses suivantes :
$\mathbf{A_1}$ *La densité f est bornée.*
$\mathbf{A_2}$ *La densité f n'a pas de partie plate :* $\forall c \geq 0$, $\mathbb{P}\{f(\mathbf{X}) = c\} = 0$.

Sous les hypothèses précédentes, pour n'importe quel $\alpha \in (0,1)$, il existe un unique ensemble de volume minimal Γ_α^*, dont la masse est égale à α. La fonction quantile (généralisée) est alors définie par :

$$\forall \alpha \in (0,1), \quad \lambda^*(\alpha) := \text{Leb}(\Gamma_\alpha^*).$$

En outre, l'application λ^* est continue sur $(0,1)$ et uniformément continue sur $[0, 1-\epsilon]$ pour tout $\epsilon \in (0,1)$ - quand le support de F est compact, la continuité uniforme est valable sur l'intervalle fermé $[0,1]$.

Les estimés $\widehat{\Gamma}_\alpha^*$ des ensembles de volume minimal sont construits en remplaçant la distribution de probabilité inconnue F par sa version empirique $F_n = (1/n)\sum_{i=1}^N \delta_{\mathbf{X}_i}$ et en restraignant l'optimisation à une collection $\mathcal{A}$ de sous-ensembles boréliens de $\mathbb{R}^d$. $\mathcal{A}$ est supposée suffisamment riche pour inclure tous les ensembles de niveau de la densité f, ou au moins des approximations raisonnables de ceux-ci.

Dans Polonik (1997), des résultats limites sont prouvés pour le processus quantile empirique généralisé $\{\text{Leb}(\widehat{\Gamma}_\alpha^*) - \lambda^*(\alpha)\}$ – sous l'hypothèse en particulier que $\mathcal{A}$ est une classe de Glivenko-Cantelli pour F. Dans Scott & Nowak (2006), il est proposé de remplacer le niveau α par $\alpha - \phi_n$ où ϕ_n joue le rôle d'un paramètre de tolérance (du même ordre que le supremum $\sup_{\Gamma\in\mathcal{A}} |F_n(\Gamma) - F(\Gamma)|$), la complexité de la classe $\mathcal{A}$ étant contrôlée par sa dimension VC, afin d'établir des bornes. La version statistique du problème du volume minimal est alors

$$\min_{\Gamma\in\mathcal{A}} \text{Leb}(\Gamma) \text{ subject to } F_n(\Gamma) \geq \alpha - \phi_n.$$

La classe $\mathcal{A}$ de sous-ensembles boréliens de $\mathbb{R}^d$ offre dans l'idéal des avantages statistiques et computationnels, permettant une recherche rapide tout en étant suffisamment complexe pour capturer la géométrie des ensembles de niveau de la densité – en d'autre termes, le "biais de modèle" $\inf_{\Gamma\in\mathcal{A}} \text{Leb}(\Gamma \Delta \Gamma_\alpha^*)$ doit être petit.

3.2 La courbe Masse-Volume

Soit $s \in \mathcal{S}$ une fonction de score. Comme défini dans Clémençon & Jakubowicz (2013); Clémençon & Robbiano (2014), la courbe masse-volume de s est le tracé de la fonction

$$MV_s : \alpha \in (0,1) \mapsto MV_s(\alpha) = \lambda_s \circ \alpha_s^{-1}(\alpha),$$

où H^{-1} désigne la pseudo-inverse de n'importe quelle fonction de répartition $H : \mathbb{R} \to (0,1)$ et où α_s et λ_s sont définis par

$$\begin{aligned} \alpha_s(t) &:= \mathbb{P}(s(\mathbf{X}) \geq t), \\ \lambda_s(t) &:= \text{Leb}(\{\mathbf{x} \in \mathbb{R}^d, s(\mathbf{x}) \geq t\}) \, . \end{aligned} \tag{2}$$

Ceci induit un ordre partiel sur l'ensemble de toutes les fonctions de score : s est préférée à s' si $MV_s(\alpha) \leq MV_{s'}(\alpha)$ pour tout $\alpha \in (0,1)$. De plus, la courbe masse-volume reste inchangée lors de l'application d'une transformation croissante sur s. On peut prouver que $MV^*(\alpha) \leq MV_s(\alpha)$ pour tout $\alpha \in (0,1)$ et toute fonction de score s, où $MV^*(\alpha)$ est la valeur optimale du problème de minimisation sous contrainte (1), à savoir

$$MV^*(\alpha) = \text{Leb}(\Gamma^*_\alpha) = \min_{\Gamma \; mes.} \text{Leb}(\Gamma) \text{ subject to } \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha \, . \tag{3}$$

Sous les hypothèses $\mathbf{A_1}$ et $\mathbf{A_2}$, on peut montrer que la courbe MV^* est bien une courbe masse volume, associée à (toute transformation croissante de) la densité f à savoir : $MV^* = MV_f$.

L'objectif est alors de construire une fonction de score $\hat{s}$ en fonction des données d'entraînement $\mathbf{X}_1, ...\mathbf{X}_n$ telle que $MV_{\hat{s}}$ soit minimale partout, c'est-à-dire minimisant $\|MV_{\hat{s}} - MV^*\|_\infty := \sup_{\alpha \in [0,1]} |MV_{\hat{s}}(\alpha) - MV^*(\alpha)|$.

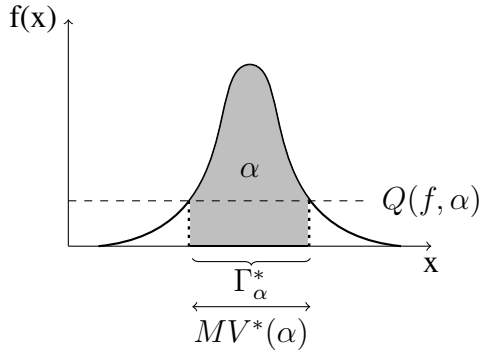

FIGURE 1: Masse-Volume au niveau α

Pour ce faire, il faut d'abord estimer une collection d'ensembles de volume minimal relatifs aux masses cibles $0 < \alpha_1 < \ldots < \alpha_K < 1$ formant une subdivision de $(0,1)$ sur la base des données d'entrainement afin de définir $s = \sum_k \mathbb{1}_{\{x \in \hat{\Gamma}^*_{\alpha_k}\}}$. L'analyse se fait sous des hypothèses adéquates (relatives à $\mathcal{G}$, au périmètre de $\Gamma^*_{\alpha_k}$ et au pas de la subdivision en particulier) et pour un choix approprié de $K = K_n$. Cependant, par construction, les vitesse d'apprentissage sont plutôt

lentes (de l'ordre $n^{-1/4}$) et ne peuvent pas être établies lorsque le support n'est pas borné.

Les quatre principaux inconvénients de ce critère de courbe masse-volume sont les suivants.

1) Lorsqu'il est utilisé comme critère de performance, la mesure de Lebesgue d'ensembles pouvant être très complexes doit être calculée.

2) Lorsqu'il est utilisé comme critère de performance, il n'existe pas de méthode directe pour comparer les courbes MV puisque l'aire sous la courbe est potentiellement infinie.

3) Lorsqu'il est utilisé comme critère d'apprentissage, il produit des ensembles de niveau qui ne sont pas nécessairement imbriqués, puis des fonctions de score imprécises.

4) Lorsqu'il est utilisé comme un critère d'apprentissage, les taux d'apprentissage sont plutôt lents (de l'ordre $n^{-1/4}$), et ne peuvent pas être établis dans le cas d'un support non borné.

Dans la section suivante, et comme contribution de cette thèse, un autre critère fonctionnel est proposé, obtenu en échangeant objectif et contrainte dans (1). Les inconvénients du critère de la courbe masse-volume sont résolus à l'exception du premier, et l'on montre que l'optimisation d'une version discrète empirique de cette mesure de performance donne des fonctions de score avec des taux de convergence de l'ordre $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$. En outre, les résultats peuvent être étendus à la situation où le support de la distribution F n'est pas compact. De plus, lorsqu'on relaxe l'hypothèse faite dans l'analyse de la courbe masse-volume que tous les ensembles de niveau de f sont inclus dans notre classe de minimisation $\mathcal{A}$, un contrôle du biais du modèle est établi. Enfin, nous déduisons des propriétés théoriques (non statistiques) vérifiées par ce critère, ce qui corrobore sa qualité de métrique sur les ensembles de niveau contenus dans les fonctions de score.

3.3 Le critère d'excès de masse

Nous proposons un autre critère de performance qui s'appuie sur la notion de *d'excès de masse* et d'*ensemble de contours de densité*, comme introduits dans la contribution Polonik (1995). L'idée principale est de considérer une formulation lagrangienne d'un problème de minimisation sous contrainte, obtenu en échangeant la contrainte et l'objectif dans (1) : pour $t > 0$,

$$\max_{\Omega \text{ borelien}} \{\mathbb{P}(\mathbf{X} \in \Omega) - t\text{Leb}(\Omega)\}. \tag{4}$$

On désigne par Ω_t^* une solution de ce problème. Cette formulation offre certains avantages à la fois computationnels et théoriques : en laissant (une version discrétisée) du multiplicateur lagrangien t augmenter de 0 à l'infini, on peut facilement

obtenir des solutions à la contrepartie empirique de (4) formant une suite *imbriquée* d'ensembles, évitant ainsi une dégradation du taux de convergence – due à la transformation des solutions empiriques pour forcer la monotonie.

La **courbe d'excès de masse optimale** d'une distribution de probabilité F est définie comme le graphe de la fonction

$$t > 0 \;\; \mapsto \;\; EM^*(t) := \max_{\Omega \text{ borelian}} \{\mathbb{P}(\mathbf{X} \in \Omega) - t\text{Leb}(\Omega)\}.$$

Avec les notations précédentes, nous avons : $EM^*(t) = \mathbb{P}(\mathbf{X} \in \Omega_t^*) - t\text{Leb}(\Omega_t^*)$ pour tout $t > 0$. Remarquons que $EM^*(t) = 0$ pour tout $t > \|f\|_\infty := \sup_{x \in \mathbb{R}^d} |f(x)|$. La **courbe d'excès de masse** de $s \in \mathcal{S}$ par rapport à la distribution de probabilité F d'une variable aléatoire $\mathbf{X}$ est le graphe de la fonction

$$EM_s : t \in [0, \infty[\mapsto \sup_{A \in \{(\Omega_{s,l})_{l>0}\}} \{\mathbb{P}(\mathbf{X} \in A) - t\text{Leb}(A)\}, \tag{5}$$

où $\Omega_{s,t} = \{x \in \mathbb{R}^d, s(x) \geq t\}$ pour tout $t > 0$.

On peut également écrire EM_s en termes de λ_s et α_s définis en (2), $EM_s(t) = \sup_{u>0} \alpha_s(u) - t\lambda_s(u)$. Enfin, sous l'hypothèse $\mathbf{A_1}$, nous avons $EM_s(t) = 0$ pour tout $t > \|f\|_\infty$.

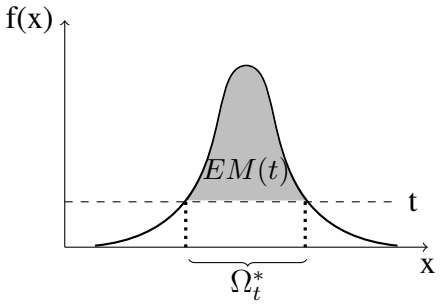

Figure 2 : Excess-Mass curve

Maximiser EM_s peut être vue comme trouver une collection de sous-ensembles $(\Omega_t^*)_{t>0}$ avec une masse maximale lorsqu'ils sont pénalisés par leur volume de façon linéaire. Une fonction de score optimale est alors n'importe quel $s \in \mathcal{S}$ admettant Ω_t^* 's comme ensembles de niveau, par exemple une fonction de score de la forme

$$s(x) = \int_{t=0}^{+\infty} \mathbb{1}_{x \in \Omega_t^*} a(t) dt,$$

avec $a(t) > 0$ (notons que $s(x) = f(x)$ pour $a \equiv 1$). La fonction EM_s est décroissante sur $(0, +\infty)$, à valeurs dans $[0, 1]$ et satisfait, $EM_s(t) \leq EM^*(t)$

pour tout $t \geq 0$. De plus, pour $t \geq 0$ et pour n'importe quel $\epsilon > 0$, nous avons

$$\inf_{u>0} \epsilon \text{Leb}(\{s > u\} \Delta_\epsilon \{f > t\}) \;\leq\; EM^*(t) - EM_s(t) \\ \leq\; \|f\|_\infty \inf_{u>0} \text{Leb}(\{s > u\} \Delta \{f > t\})$$

avec $\{s > u\}\Delta_\epsilon\{f > t\} := \{f > t+\epsilon\} \setminus \{s > u\} \bigsqcup \{s > u\} \setminus \{f > t-\epsilon\}$. Ainsi la quantité $EM^*(t) - EM_s(t)$ mesure avec quelle qualité les ensembles de niveau de s peuvent-il approchés ceux de la densité sous-jacente. Sous des hypothèse raisonnables, (voir Goix et al. (2015c), Prop.1), nous avons aussi pour $\epsilon > 0$,

$$\sup_{t \in [\epsilon, \|f\|_\infty]} |EM^*(t) - EM_s(t)| \;\leq\; C \inf_{T \in \mathcal{T}} \|f - T \circ s\|_\infty$$

où l'infimum est pris sur l'ensemble $\mathcal{T}$ de toutes les transformations croissantes mesurables $T : \mathbb{R}_+ \to \mathbb{R}_+$. Les inégalités précédentes révèlent que $\|EM^* - EM_s\|_\infty$ peut être interprété comme une pseudo distance, soit entre les ensembles de niveau de s et ceux de la densité sous-jacente f, soit entre les préordres induits par s et f.

Le concept de la courbe EM fournit un moyen simple de comparer les fonctions de score, mais l'optimisation d'un tel critère fonctionnel est loin d'être simple. Comme proposé dans Clémençon & Jakubowicz (2013) pour le critère MV, l'optimisation est faite sur une certaine classe de fonctions de score, que nous espérons assez riche pour fournir une bonne approximation (biais de modèle petit) tout en étant assez simple pour contrôler le taux de convergence. Nous considérons ici les fonctions de score de la forme

$$s_N(x) := \sum_{k=1}^{N} a_k \mathbb{1}_{x \in \hat{\Omega}_{t_k}}, \quad \text{avec} \quad \hat{\Omega}_{t_k} \in \mathcal{G}$$

où $\mathcal{G}$ est une class VC de sous-ensembles de $\mathbb{R}^d$. Nous choisissons de manière arbitraire $a_k := (t_k - t_{k+1})$ de telle sorte que les $\hat{\Omega}_{t_k}$ correspondent exactement aux ensembles de niveau t_k, $\{s \geq t_k\}$. Ensuite, la maximisation du critère fonctionnel d'excès de masse s'effectue en résolvant de manière séquentielle, pour $k = 1, \ldots, N$,

$$\hat{\Omega}_{t_k} \in \arg\max_{\Omega \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \in \Omega} \;-\; t_k \text{Leb}(\Omega).$$

Les solution $\hat{\Omega}_{t_k}$ de ces problèmes d'optimisation peuvent toujours être choisies de manière à être imbriquées (contrairement au problème d'optimisation analogue pour le critère masse-volume). En d'autres termes, une contrainte d'inclusion peut être incorporée dans le problème d'optimisation précédent, sans affecter

la qualité de la solution obtenue. Dans le cadre du critère masse-volume, des hypothèses sont faites stipulant que le support de la distribution doit être compact et que la classe VC $\mathcal{G}$ doit contenir les ensembles de niveau de la densité f. Ici, nous relaxons ces hypothèses, la première en choisissant des niveaux adaptatifs t_k, et la seconde en dérivant une étude de biais.

4 Précision sur les régions extrêmes

4.1 Analyse du point de vue de la théorie des valeurs extrêmes par l'estimation de la STDF

Cette section est basée sur l'article Goix et al. (2015b).

Rappelons que les fonctions de score sont construites en approchant les ensembles de niveau de densité / ensembles de volume minimal de la densité "normale" sous-jacente. Comme nous l'avons mentionné précédemment, dans le cadre de la détection d'anomalie, nous souhaitons être précis sur des ensembles de niveau correspondant à des quantiles élevés, à savoir avec un niveau t près de 0 – ou de manière équivalente, être précis sur des ensembles de volume minimal avec une contrainte de masse α proche de 1.

Dans le cas univarié, supposons que nous voulons considérer le quantile d'ordre $(1-p)$ de la distribution F d'une variable aléatoire X, pour une probabilité donnée de dépassement p, c'est-à-dire $x_p = \inf\{x \in \mathbb{R},\ \mathbb{P}(X > x) \leq p\}$. Pour les valeurs modérées de p, une contrepartie empirique naturelle est $x_{p,n} = \inf\{x \in \mathbb{R},\ 1/n \sum_{i=1}^{n} \mathbb{1}_{X_i > x} \leq p\}$. Cependant, si p est très petit, l'échantillon fini $X_1, \ldots, X_n$ ne contient pas suffisamment d'informations et $x_{p,n}$ devient inutile. Ce problème devient dans le cas multivarié celui d'estimer des ensembles de niveau de densité avec un niveau très faible ou de manière equivalente celui d'estimer les fonctions de score associées à ces ensembles de niveau. La théorie des valeurs extrêmes traite spécialement de ces problèmes, aussi bien dans le cadre unidimensionnel que multidimensionnel.

Préliminaires. La théorie des valeurs extrêmes (nommée dans la suite TVE) développe des modèles pour apprendre l'insolite plutôt que l'habituel. Ces modèles sont largement utilisés dans les domaines de la gestion des risques comme celui de la finance, de l'assurance, des télécommunications ou des sciences de l'environnement. Une application majeure de la TVE est de fournir une évaluation raisonnable de la probabilité d'occurrence d'événements rares.

Pour illustrer ce point, supposons que nous voulons gérer le risque d'un portefeuille contenant d actifs différents, $\mathbf{X} = (X_1, \ldots, X_d)$. Un but assez général est alors d'évaluer la probabilité d'événements du type $\{X_1 \geq x_1 \text{ or } \ldots \text{ or } X_d \geq$

$x_d\}$, pour des seuils multivariés grands $\mathbf{x} = (x_1, \ldots, x_d)$. Dans des conditions pas trop strictes sur la régularité de la distribution $\mathbf{X}$, la TVE montre que pour des seuils suffisamment importants,

$$\mathbb{P}\{X_1 \geq x_1 \text{ or } \ldots \text{ or } X_d \geq x_d\} \simeq l(p_1, \ldots, p_d),$$

où l est la STDF (*stable tail dependence function*) et où les p_j sont les probabilités de dépassement marginal, $p_j = \mathbb{P}(X_j \geq x_j)$. La fonction l caractérise la *dépendance* entre les extrêmes. La distribution *jointe* (sur des seuils importants) peut donc être récupérée à partir de la connaissance des distributions marginales avec la STDF l. Dans la pratique, l peut être estimée à partir de données "modérément extrêmes", typiquement les k 'plus grandes 'parmi un échantillon de taille n, avec $k \ll n$.

L'estimation des p_j peut s'effectuer suivant un chemin bien pavé : dans le cas univarié, la TVE consiste essentiellement à modéliser la distribution des maxima (*resp.* la queue de distribution) par un élément des familles paramétriques de Gumbel, Fréchet ou Weibull (*resp.* par une distribution de Pareto généralisée).

Par contre, dans le cas multivarié, il n'y a pas de paramétrisation finidimensionnelle de la structure de dépendance. Cette dernière étant caractérisée par la STDF, l'estimation de cette fonctionnelle est l'un des principaux problèmes de la TVE multivariée. Les propriétés asymptotiques de la contrepartie empirique de la STDF ont été largement étudiées, voir Huang (1992); Drees & Huang (1998); Embrechts et al. (2000); De Haan & Ferreira (2007) pour le cas bivarié et Qi (1997); Einmahl et al. (2012) pour le cas général multivarié sous des hypothèses de régularité.

Cependant, aucune borne n'existe sur l'estimation non-asymptotique. La contribution résumée dans la section suivante et publiée dans Goix et al. (2015b) dérive de telles bornes non asymptotiques. Nos résultats ne nécessitent aucune hypothèse autre que l'existence de la STDF.

Apprentissage de la structure de dépendance des événements rares. Les inégalités de VC classiques visent à borner la déviation des quantités empiriques par rapport aux quantités théoriques sur des classes d'ensemble relativement simples, appelées classes VC. Assez souvent, ces classes recouvrent tout le support de la distribution sous-jacente. Cependant, lorsqu'il s'agit d'événements rares, il est intéressant d'avoir de telles bornes sur une classe d'ensembles qui ne couvre qu'une région de faible probabilité et contient donc (très) peu d'observations. Cela donne des bornes plus fines, puisque seules les différences entre de très petites quantités sont impliquées dans l'analyse. Le point de départ est l'inégalité VC énoncée ci-dessous.

Théoreme 1. *Soit* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *des réalisations* $i.i.d.$ *d'une variable aléatoire* $\mathbf{X}$*,* $\mathcal{A}$ *une classe VC de VC-dimension* $V_{\mathcal{A}}$*.*

Considerons la classe $\mathbb{A} = \cup_{A \in \mathcal{A}} A$*, et posons* $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$*. Alors il existe une constante absolue* C *de sorte que pour tout* $0 < \delta < 1$*, avec probabilité au moins* $1 - \delta$*,*

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}\big[\mathbf{X} \in A\big] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \;\leq\; C \left[\sqrt{p} \sqrt{\frac{V_{\mathcal{A}}}{n} \log \frac{1}{\delta}} + \frac{1}{n} \log \frac{1}{\delta} \right].$$

L'idée principale est la suivante. L'estimateur empirique de la STDF est basé sur la mesure empirique des régions "extrêmes", qui sont touchées seulement avec une faible probabilité. Il suffit donc de borner les déviations maximales sur ces régions à faible probabilité. La clé consiste à choisir une classe VC adaptative, qui ne couvre que ces régions là, et à dériver des inégalités de type VC qui intègrent p, la probabilité de toucher la classe. La borne obtenue sur l'erreur non asymptotique est alors :

Théoreme 2. *Soit* T *un nombre positif tel que* $T \geq \frac{7}{2}(\frac{\log d}{k} + 1)$*,* δ *tel que* $\delta \geq e^{-k}$ *et soit* $k = k(n)$ *une suite d'entiers strictement positifs telle que* $k \to \infty$ *et* $k = o(n)$ *quand* $n \to \infty$*. Alors il existe une constante absolue* C *telle que pour chaque* $n > 0$*, avec probabilité au moins* $1 - \delta$ *:*

$$\sup_{0 \leq \mathbf{x} \leq T} |l_n(\mathbf{x}) - l(\mathbf{x})| \;\leq\; Cd \sqrt{\frac{T}{k} \log \frac{d+3}{\delta}} \;+\; \textit{bias}(k, n, T),$$

où l *est la* STDF *et* l_n *sa version empirique standard. Le second terme dans la borne est un biais issu de la nature asymptotique de* l*.*

Dans cette section, nous avons introduit et étudié, dans un cadre non-paramétrique, une fonctionnelle particulière caractérisant la structure de dépendance des extrêmes. Une autre caractérisation pratique (non paramétrique) de cette dépendance dans le cadre de la TVE multivariée est la *mesure angulaire* qui fournit des informations directes sur les "directions" probables des extrêmes, c'est-à-dire la contribution relative de chaque coordonnée dans les "grandes" observations.

Dans de nombreuses applications, il est plus commode de travailler avec la mesure angulaire elle-même. Cette dernière donne des informations plus directes sur la structure de dépendance et est capable de refléter des propriétés structurelles (par exemple la sparsité/parcimonie comme détaillé ci-dessous) qui n'apparaîtraient pas dans les copules ou dans la STDF, ces derniers étant des versions intégrées de la mesure angulaire. Cependant, la modélisation non paramétrique de la mesure angulaire est confrontée à des difficultés majeures, dues à sa

structure potentiellement complexe, en particulier dans un cadre de grande dimension. D'autre part, d'un point de vue théorique, l'estimation non paramétrique de la mesure angulaire n'a été étudiée que dans le cas bidimensionnel et dans un cadre asymptotique Einmahl et al. (2001); Einmahl & Segers (2009). La section ci-dessous résume une nouvelle méthodologie visant à représenter parcimonieusement la structure de dépendance des extrêmes.

4.2 Représentation parcimonieuse des extrêmes multivariés

Cette section résume les travaux publiés Goix et al. (2016c), ainsi que sa version longue Goix et al. (2016b) en cours de révision.

La TVE a été intensivement utilisée en détection d'anomalies dans le cas unidimensionnelle, voir par exemple Roberts (1999, 2000); Clifton et al. (2011, 2008); Lee & Roberts (2008). Dans le cas multivariée, cependant, il n'existe – à notre connaissance – aucune méthode de détection d'anomalies reposant sur la TVE *multivariée*. Jusqu'à présent, le cas multidimensionnel n'a été abordé que par l'usage de statistiques basées sur la TVE univariée. La raison majeure est la difficulté du passage à l'échelle des modèles multivariés avec la dimension. Dans le présent travail, nous comblons l'écart entre la détection d'anomalies et la TVE multivariée en proposant une méthode qui est capable d'apprendre un "profil normal" parcimonieux des extrêmes multivariés et, en tant que tel, peut être mis en œuvre pour améliorer la précision de tout algorithme de détection d'anomalies.

Context : Extrèmes multivariés en grande dimension. L'estimation paramétrique ou semi-paramétrique de la structure des extrêmes multivariés est relativement bien documentée dans la littérature, voir par exemple Coles & Tawn (1991); Fougères et al. (2009); Cooley et al. (2010); Sabourin & Naveau (2014) et leurs références. Cependant, des hypothèses structurelles restrictives doivent être faites, stipulant par exemple que seuls quelques sous-groupes pré-définis de composantes peuvent être extrêmes ensemble. En outre, leur utilisation pratique est limitée à des problèmes en dimension modérée (par exemple, $d \leq 10$), sinon des choix de modélisation simplifiés sont nécessaires – comme dans Stephenson (2009). Enfin, l'évaluation de l'incertitude concernant les quantités produites par ces modèles est faite sous l'hypothèse que les données d'entraînement sont "asymptotiques", au sens où l'on suppose que, quand elles excèdent un grand seuil fixé, les données sont exactement réparties selon la distribution limite des extrêmes . En d'autres termes, l'erreur de modélisation est ignorée.

L'estimation non-paramétrique de la mesure angulaire n'a été traitée que dans le cas bidimensionnel, dans Einmahl et al. (2001); Einmahl & Segers (2009), et dans un cadre asymptotique. Nous allons étendre l'étude non-asymptotique sur

l'estimation de la STDF (section précédente) à la mesure angulaire des extrêmes, restreinte à une classe bien choisie d'ensembles. L'objectif est d'apprendre une représentation de la mesure angulaire, assez simple pour contrôler la variance en grande dimension et suffisamment précise pour obtenir des informations sur les "directions probables" des extrêmes. Ceci donne une première estimation non paramétrique de la mesure angulaire en dimension quelconque, limitée à une classe de sous-cones, avec une borne non asymptotique sur l'erreur. Notons que ce procédé peut également être utilisé comme étape de prétraitement, dans un cadre de réduction de dimension, avant de procéder à une estimation paramétrique ou semi-paramétrique qui pourrait bénéficier des informations de structure émises lors de la première étape. De telles applications dépassent le cadre de cette thèse.

Le cadre que nous développons est non paramétrique et se trouve à l'intersection de l'estimation de support, de l'estimation de densité et de la réduction de dimension : il consiste à apprendre le support d'une distribution (à partir des données d'apprentissage), qui peut être décomposé en sous-cones, potentiellement de dimension faible et auxquels une certaine masse est assignée.

Ceci produit une fonction de score définie sur les régions extrêmes, qui peut ainsi être exploitée pour détecter les anomalies parmi les extrêmes. En raison de sa complexité modérée - d'ordre $dn \log n$ - cet algorithme convient au traitement de problèmes d'apprentissage à grande échelle, et les résultats expérimentaux révèlent une performance significativement accrue sur les régions extrêmes par rapport aux approches de détection d'anomalies standards.

Dans un large éventail de situations, on peut s'attendre à l'apparition de deux phénomènes :

1- Il y a seulement un "petit" nombre de groupes de coordonnées pouvant être extrêmes ensemble, de sorte que seul un "petit" nombre d'hyper-cubes (ceux correspondant à ces sous-ensembles de coordonnées) ont une masse non nulle – "petite" est relatif au nombre total de groupes, 2^d.

2- Chacun de ces groupes contient un nombre limité de coordonnées (par rapport à la dimension initiale), de sorte que les hyper-cubes correspondants (de masse non nulle) ont une petite dimension par rapport à d.

Le but principal de ce travail est d'introduire une méthodologie pilotée par les données pour identifier ces groupes, afin de réduire la dimension du problème et ainsi apprendre une représentation parcimonieurse des comportements extrêmes.

Dans le cas où l'hypothèse **2-** n'est pas vérifiée, une telle représentation peut tout de même être apprise, mais perd la propriété que les hyper-cubes la supportant sont de faible dimension.

Un problème majeur est que les données réelles ne se concentrent généralement pas sur les sous-espaces de mesure Lebesgue nulle. Ceci peut être résolu en mettant à zéro n'importe quelle coordonnée inférieure à un seuil $\epsilon > 0$, de sorte que "l'angle" correspondant soit affecté à une face de dimension inférieure.

Plus formellement, les figures 2 et 3 représentent l'espace initial des données transformé, résultant de la standardisation classique des marginales. Après cette transformation non linéaire, la représentation des données extrêmes est pour sa part linéaire et apprise en estimant la masse portée par les sous-cônes

$$\mathcal{C}_\alpha = \{\mathbf{v} \geq 0,\ \|\mathbf{v}\|_\infty \geq 1,\ v_j > 0 \ \text{ for } j \in \alpha,\ v_j = 0 \ \text{ for } j \notin \alpha\},$$

ou plus précisément, la masse de la mesure angulaire Φ sur les sous-sphères correspondantes

$$\Omega_\alpha = \{\mathbf{x} \in S_\infty^{d-1} : x_i > 0 \text{ for } i \in \alpha\ ,\ x_i = 0 \text{ for } i \notin \alpha\} = S_\infty^{d-1} \cap \mathcal{C}_\alpha,$$

représentées Figure 2.

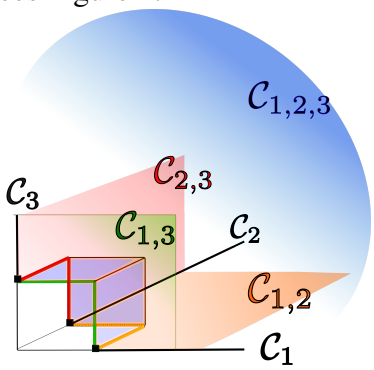

FIGURE 2: Truncated cones in 3D

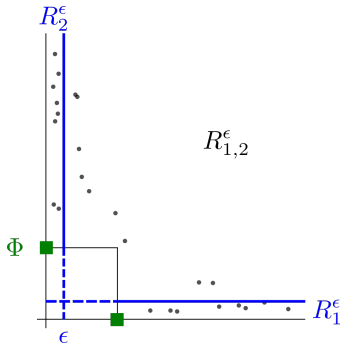

FIGURE 3: Truncated ϵ-cones in 2D

Cette estimation est faite en utilisant les sous-cones ϵ-épaissis $\mathcal{C}_\alpha^\epsilon$, correspondant aux sous-sphères ϵ-épaissies Ω_α^ϵ, comme le montre la Figure 3 dans le cas de la dimension deux. Nous obtenons ainsi un estimateur $\widehat{\mathcal{M}}$ de la représentation

$$\mathcal{M} = \{\Phi(\Omega_\alpha) :\ \emptyset \neq \alpha \subset \{1,\ \ldots,\ d\}\}.$$

Théoriquement, retrouver le vecteur inconnu $\mathcal{M}$ de dimension $(2^d - 1)$ revient à peu près à approximer le support de Φ en utilisant la partition $\{\Omega_\alpha, \alpha \subset \{1, \ldots, d\}, \alpha \neq \emptyset\}$, c'est-à-dire, déterminer quels sont les Ω_α qui ont une masse non nulle – et évaluer cette masse $\Phi(\Omega_\alpha)$. Cette estimation de support est potentiellement parcimonieuse – si seul un petit nombre d'Ω_α ont une masse non nulle, *i.e.* Phénomène **1-** – et potentiellement de faible dimension – si les dimensions des sous-sphère Ω_α ayant une masse non nulle sont faibles comparées à d, *i.e.* Phénomène **2-**.

Détection d'anomalies. L'algorithme que nous proposons, DAMEX (Detecting Anomalies with Extremes), apprend une représentation $\widehat{\mathcal{M}}$ (éventuellement parcimonieuse et de faible dimension) de la mesure angulaire, à partir de laquelle

une fonction de score peut être définie dans le contexte de la détection des anomalies. L'hypothèse sous-jacente est qu'une observation est potentiellement anormale si sa "direction" (après une normalisation de chaque marginal) est particulière par rapport aux autres observations extrêmes. En d'autres termes, si elle n'appartient pas à la représentation (parcimonieuse) $\widehat{\mathcal{M}}$. Selon les expériences obtenus dans ce chapitre, DAMEX améliore significativement les performances (en terme de précision et de courbes ROC) dans les régions extrêmes, ce qui induit des courbes ROC plus verticales près de l'origine.

Garanties théoriques. A partir des travaux sur l'estimation de la STDF résumés dans la sous-section précédente 4.1, en particulier à partir du Théorème 1 et des idées utilisées pour prouver le Théorème 2, nous sommes en mesure de prouver quelques garanties théoriques relative à cette approche. Sous des hypothèses non-restrictives standards en TVE (existence de la mesure angulaire et fonctions de répartition des marges continues), on obtient une borne non asymptotique de la forme

$$\sup_{\emptyset \neq \alpha \subset \{1,\,...,\,d\}} |\widehat{\mathcal{M}}(\alpha) - \mathcal{M}(\alpha)| \;\leq\; Cd\left(\sqrt{\frac{1}{\epsilon k}\log\frac{d}{\delta}} + Md\epsilon\right) + \text{bias}(\epsilon, k, n),$$

avec probabilité plus grande que $1-\delta$, où $k = k(n) \to \infty$ avec $k(n) = o(n)$ pouvant être interprété comme le nombre de données considérées extrêmes. Le terme de biais tend vers zéro quand $n \to \infty$, pour tout ϵ fixé.

5 Approches heuristiques

Les deux contributions de cette section sont de nature heuristique et ne sont pas encore étayées par des résultats théoriques statistiquement solides. Bien que ces travaux en cours n'aient pas encore été publiés et seront certainement achevés dans un proche avenir, nous pensons qu'ils ont leur place dans ce manuscrit, étant donné les nombreuses expériences numériques convaincantes qui ont été menées et la justification des approches promues. Ces deux contributions abordent deux défis majeurs en détection d'anomalies :

— Comment évaluer la détection d'anomalies non supervisée en pratique ?
— Comment créer des forêts aléatoires efficaces sur une seule classe de données ?

Le premier point a été partiellement traité dans la section 3 avec les courbes MV et EM. Cependant, ces deux critères ont été initialement introduits pour construire des fonctions de scores *via* minimisation du risque empirique (ERM), et aucune étude n'a été faite sur leur utilisation pour évaluer les fonctions de scores comme le font les critères ROC ou PR dans le cas où des données labélisées sont

disponibles. En outre, leur utilisation pour mesurer la qualité d'une fonction de score s_n implique le calcul de la mesure de Lebesgue $\text{Leb}(s_n \geq u)$, ce qui est très difficile quand la dimension est grande.

Les deux approches proposées sont heuristiques, et aucune garantie théorique de consistence ou de taux de convergence n'est dérivée. Cependant, de nombreuses expériences montrent la pertinence de ces approches.

5.1 Évaluer un algorithme de détection d'anomalies

Cette partie est basée sur un article de workshop (Goix, 2016) et sur un article à soumettre (Goix & Thomas, 2016).

Lorsque suffisamment de données labélisées sont disponibles, les critères classiques basés sur les courbes ROC (Provost et al., 1997, 1998; Fawcett, 2006) ou PR (Davis & Goadrich, 2006; Clémençon & Vayatis, 2009) peuvent être utilisés pour comparer les performances d'algorithmes de détection d'anomalies non supervisés. Cependant, dans de nombreuses situations, pas ou peu de données sont étiquetées. C'est dans ce cas qu'un critère alternatif pouvant être calculé sur des données non-étiquetées trouve toute son utilité.

Alors que les courbes d'excès de masse et masse-volume ont été initialement introduites pour construire des fonctions de score *via* minimisation du risque empirique (ERM), la courbe MV a été utilisée récemment pour la calibration du SVM à une classe (Thomas et al., 2015). Lorsque ce critère est utilisé pour attester la qualité d'une fonction de score, les volumes induits deviennent inconnus et doivent être estimés, ce qui est difficile en grande dimension si aucune connaissance préalable sur la forme de ces ensembles de niveau n'est disponible. De plus, l'efficacité des courbes EM ou MV comme critères d'évaluation n'a pas encore été étudiée. Dans cette section et en tant que contribution de cette thèse, on montre que des scores numériques associées aux critères EM et MV (qui ne nécessitent pas d'étiquettes) sont aptes à discriminer avec précision les algorithmes suivants leurs performances. Une méthodologie basée sur le sous-échantillonnage et l'agrégation de features est également décrite et testée. Elle étend l'utilisation de ces critères à des ensembles de données de grande dimension et résout les principaux inconvénients inhérents aux courbes EM et MV classiques.

Rappelons que les courbes MV et EM d'une fonction de score s peuvent être écrites comme

$$MV_s(\alpha) = \inf_{u \geq 0} \ \text{Leb}(s \geq u) \ \ \textit{s.t.} \ \ \mathbb{P}(s(\mathbf{X}) \geq u) \geq \alpha \tag{6}$$

$$EM_s(t) = \sup_{u \geq 0} \ \mathbb{P}(s(\mathbf{X}) \geq u) \ - \ t\text{Leb}(s \geq u) \tag{7}$$

pour tout $\alpha \in (0,1)$ et $t > 0$. Les courbes optimales sont $MV^* = MV_f = MV_{T\circ f}$ et $EM^* = EM_f = EM_{T\circ f}$ pour toute transformation croissante T : Im(f) $\to \mathbb{R}$. Comme il n'existe pas de manière triviale pour comparer deux courbes, considérons la norme $\|.\|_{L^1(I)}$ avec $I \subset \mathbb{R}$ un interval. Comme $MV^* = MV_f$ est en dessous de MV_s en tout point, $\arg\min_s \|MV_s - MV^*\|_{L^1(I)} = \arg\min \|MV_s\|_{L^1(I)}$. Nous définissons donc $\mathcal{C}^{MV}(s) = \|MV_s\|_{L^1(I^{MV})}$, ce qui revient à considérer $\|MV_s - MV^*\|_{L^1(I^{MV})}$. Etant donné que nous sommes intéressés par les grands ensembles de niveau, un interval naturel serait par exemple $I^{MV} = [0.9, 1]$. Cependant, la courbe MV diverge en 1 quand le support est infini, ce qui nous amène à consider arbitrairement $I^{MV} = [0.9, 0.999]$. Plus la valeur de $\mathcal{C}^{MV}(s)$ est petite, meilleure est la fonction de score s. De même, nous considérons $\mathcal{C}^{EM}(s) = \|EM_s\|_{L^1(I^{EM})}$, cette fois avec $I^{EM} = [0, EM^{-1}(0.9)]$, où $EM_s^{-1}(0.9) := \inf\{t \geq 0,\ EM_s(t) \leq 0.9\}$, puisque $EM_s(0)$ est fini (égal à 1).

Comme la distribution F des données normales est généralement inconnue, les courbes MV et EM doivent être estimées. Soit $s \in \mathcal{S}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n$ un échantillon i.i.d. de distribution F. Utilisons la notation $\mathbb{P}_n(s \geq t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{s(\mathbf{X}_i)\geq t}$. Les courbes empiriques MV et EM de s sont alors simplement définies comme la version empirique de (6) ou de (7),

$$\widehat{MV}_s(\alpha) = \inf_{u\geq 0} \{\text{Leb}(s \geq u) \ \textit{s.t.}\ \mathbb{P}_n(s \geq u) \geq \alpha\} \tag{8}$$

$$\widehat{EM}_s(t) = \sup_{u\geq 0} \mathbb{P}_n(s \geq u) \ - \ t\text{Leb}(s \geq u) \tag{9}$$

Enfin, nous obtenons les critères de performance empiriques relatifs à EM et MV :

$$\widehat{\mathcal{C}}^{EM}(s) = \|\widehat{EM}_s\|_{L^1(I^{EM})} \qquad I^{EM} = [0, \widehat{EM}^{-1}(0.9)], \tag{10}$$

$$\widehat{\mathcal{C}}^{MV}(s) = \|\widehat{MV}_s\|_{L^1(I^{MV})} \qquad I^{MV} = [0.9, 0.999]. \tag{11}$$

La méthodologie pour faire passer à l'échelle l'utilisation des critères EM et MV (aux données en grande dimension) consiste à sous-échantillonner les données d'entrainement *et* de test, en utilisant un paramètre d' contrôlant le nombre de dimensions choisies au hasard pour le calcul du score (EM ou MV). Ce tirage se fait sans remplacement – le remplacement se fait seulement après chaque tirage $F_1, \ldots, F_m$.

Un score partiel $\widehat{\mathcal{C}}_k^{MV}$ (resp. $\widehat{\mathcal{C}}_k^{EM}$) est calculé à chaque tirage F_k en utilisant (10) (resp. (11)). Le critère de performance final est obtenu en moyennant ces critères partiels. Cette méthodologie est décrite par l'algorithme 1.

Les critères EM/MV pour dimension faible et pour dimension élevée sont testés à l'aide de trois algorithmes de détection d'anomalies classiques. Une large gamme de jeux de données étiquetés réels est utilisée à titre de comparaison. Les

Algorithm 1 EM/MV en grande dimension : évaluation d'algorithmes de détection d'anomalies sur des donnéees de dimension élevée

Entrées : algorithme de détection d'anomalies $\mathcal{A}$, jeu de données $X = (x_i^j)_{1\leq i\leq n, 1\leq j\leq d}$, taille de sous-échantillonage d', nombre de tirages m.
for $k = 1, \ldots, m$ **do**
  sélectionner aléatoirement un sous-groupe F_k de d' coordonnées
  calculer la fonction de score associée $\widehat{s}_k = \mathcal{A}\big((x_i^j)_{1\leq i\leq n,\ j\in F_k}\big)$
  calculer $\widehat{\mathcal{C}}_k^{EM} = \|\widehat{EM_{\widehat{s}_k}}\|_{L^1(I^{EM})}$ en utilisant (10) ou $\widehat{\mathcal{C}}_k^{MV} = \|\widehat{MV_{\widehat{s}_k}}\|_{L^1(I^{MV})}$ en utilisant (11)
end for
Sortie : critère de performance sur $\mathcal{A}$:

$$\widehat{\mathcal{C}}^{EM}_{high\_dim}(\mathcal{A}) = \frac{1}{m}\sum_{k=1}^{m} \widehat{\mathcal{C}}_k^{EM} \quad \text{(idem for MV)}$$

expériences montrent que lorsqu'un algorithme a de meilleures performances que d'autres sur un certain jeu de données ("meilleures" selon les AUC des courbes ROC et PR), on peut s'attendre à le retrouver sans utiliser d'étiquettes avec une précision de 82% dans le cadre de détection de nouveauté, et avec une précision de 77% dans le cadre non-supervisé.

5.2 Forêts aléatoires à une classe

Cette partie est basée sur un travail en voie de soumission (Goix et al., 2016a).

Construire des fonctions de score précises en optimisant les critères EM ou MV est très difficile en pratique, de manière analogue à la construction de classifieurs en optimisant la courbe ROC (Clémençon & Vayatis (2010)) dans le cadre supervisé. Il faut davantage de travail pour que ces méthodes soient efficaces dans la pratique, en particulier au niveau du choix de la classe d'ensembles sur lesquels l'optimisation est effectuée. En effet, cette classe doit être *assez riche pour fournir une bonne approximation tout en étant assez simple pour contrôler le taux de convergence*. Ce compromis est difficile à réaliser, en particulier en dimension élevée, lorsqu'aucune connaissance préalable sur la forme des ensembles de niveau n'est disponible.

Dans cette section, nous proposons une approche heuristique pour construire des fonctions de score en utilisant des forêts aléatoires (dans la suite abrégées RF pour "random forests") (Breiman, 2001; Genuer et al., 2008; Biau et al., 2008; Biau & Scornet, 2016). Plus formellement, nous adaptons les RFs au cadre de la

classification à une classe en introduisant des critères de séparation à une classe.

Les RFs standards sont des estimateurs qui entrainent un certain nombre de classificateurs d'arbre de décision sur différents sous-échantillons aléatoires de l'ensemble de données. Chaque arbre est construit récursivement, selon un critère de séparation/scission basé sur une certaine mesure d'impureté définie sur chaque noeud de l'arbre – un noeud est en fait une cellule rectangulaire de l'espace des observations. La prédiction est faite en moyennant les prédictions de chaque arbre. Dans le cadre de la classification (à deux classes), les prédictions des arbres sont moyennés à travers un vote majoritaire. Peu de tentatives pour transférer l'idée des RFs à la classification à une classe ont déjà été faites (Désir et al., 2012; Liu et al., 2008; Shi & Horvath, 2012). Aucun algorithme ne prolonge structurellement (sans échantillonnage de seconde classe et sans estimateurs de base alternatifs) les RFs au cadre de la classification à une classe.

Nous introduisons précisément une telle méthodologie. Elle s'appuie sur une adaptation naturelle des critères de scission à deux classes au contexte de classification à une classe, ainsi qu'une adaptation du vote majoritaire. De plus, il s'avère que le modèle à une classe que nous promouvons ici correspond au comportement asymptotique d'une méthode pour générer des outliers qui s'adapte au données (plus précisement qui s'adapte au processus de croissance des arbres, qui lui dépend des données).

Modèle à une classe de paramètres (n, α), $\mathbf{M}(n, \alpha)$. Considérons un variable aléatoire $X : \Omega \to \mathbb{R}^d$ par rapport à un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. La loi de X est supposée dépendre d'une autre variable $y \in \{0, 1\}$, qui vérifie $\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = 0) = \alpha$. Nous supposons que conditionellement à $y = 0$, X suit une loi F, et conditionellement à $y = 1$, une loi G :

$$\begin{aligned} X \mid y = 0 &\sim F, \qquad \mathbb{P}(y = 0) = 1 - \alpha, \\ X \mid y = 1 &\sim G, \qquad \mathbb{P}(y = 1) = \alpha. \end{aligned}$$

Nous modélisons le cadre à une classe comme suit. Parmi les n observations $i.i.d.$ nous observons seulement celles avec $y = 0$ (le comportement normal), c'est-à-dire N réalisations de $(X \mid y = 0)$ où N est lui-même la réalisation d'une variable aléatoire $\mathbf{N}$ de loi $\mathbf{N} \sim \text{Bin}\big(n, (1 - \alpha)\big)$, la distribution binomiale de paramètres (n, P). Comme les outliers ne sont pas observés, on suppose classiquement que G suit une distribution uniforme sur l'hyper-rectangle $\mathcal{X}$ contenant toutes les observations, de sorte que G a une densité constante $g(x) \equiv 1/\text{Leb}(\mathcal{X})$ sur $\mathcal{X}$. Cette hypothèse *sera supprimée* dans l'approche adaptative, où aucune distribution préalable n'est supposée pour les outliers.

On obtient alors des analogues empiriques à une classe des mesures d'impuretés à deux classes en remplaçant les quantités relatives au comportement normal par leurs versions empiriques. Les quantités relatives à la deuxième classe non

observée (comportement anormal) sont exprimées naturellement en utilisant l'hypothèse de distribution uniforme.

De cette façon, notre fonction de différence d'impuretés à une classe correspond à celle à deux classe, où les quantités empiriques de la seconde classe ont été remplacées par leur espérance supposant une distribution uniforme.

Mais elle induit également un problème majeur : ces espérances, qui sont proportionnelles au volume du noeud en jeu, deviennent très petites lorsqu'on descend de plus en plus profondément dans l'arbre. Dans le cadre à deux classes, le problème analogue est lorsque la seconde classe est fortement sous-représentée au voisinage des observations.

Comme nous supposons que la deuxième classe est uniforme sur un hyperrectangle contenant toutes les observations, ce fait était attendu, surtout en grande dimension (malédiction de la dimension). Quand les quantités relatives à la seconde classe sont très proches de zéro, on observe que le critère d'impureté devient constant, indépendemment de la scission du noeud, donc inutile.

Approche adaptative Une solution consiste à choisir de façon adaptative (par rapport au volume de chaque noeud) le nombre αn, qui peut être interprété comme le nombre d'outliers (cachés). Rappelons que ni n ni α ne sont observés dans le modèle à une classe M(n, α) défini ci-dessus.

L'idée est de faire $\alpha(t) \to 1$, $n(t) \to \infty$ quand le volume du noeud t tend vers zéro. En d'autres termes, au lieu de considérer un modèle général fixe M(n, α), nous l'adaptons à chaque noeud t, en considérant M($n(t)$, $\alpha(t)$) *avant de chercher la meilleure partition*. Nous considérons encore les N observations normales comme une réalisation de ce modèle. Lors de la croissance de l'arbre, l'utilisation de M($n(t)$, $\alpha(t)$) permet de maintenir une proportion espérée non négligeable d'outliers dans le noeud à diviser, même lorsque son volume devient très petit. Bien sûr, des contraintes doivent être imposées pour assurer la cohérence entre ces modèles. Par exemple, rappelant que le nombre N d'observations normales est une réalisation de $\mathbf{N}$ suivant une distribution binomiale de paramètres $(n, 1-\alpha)$, une première contrainte naturelle sur $\big(n(t), \alpha(t)\big)$ est

$$(1-\alpha)n = \big(1-\alpha(t)\big) \cdot n(t) \quad \text{for all } t, \tag{12}$$

de sorte que l'espérance de $\mathbf{N}$ soit inchangée. Alors le modèle asymptotique (quand le volume de t tend vers 0) consiste en fait à supposer que le nombre N de données normales que nous avons observées est une réalisation d'une distribution de Poisson $\mathcal{P}\big((1-\alpha)n\big)$, et qu'un nombre infini d'outliers ont été cachés. Dans le cadre de la classification à deux classes, ceci correspond à l'observation d'un nombre infini d'outliers répartis étroitement autour et à l'intérieur du support de la distribution normale, rompant la malédiction de la dimension due à l'utilisation d'outliers uniformément répartis.

Remarque 2 (**Idée fondamentale de l'approche adaptative**). *Ce travail correspond en fait à l'idée simple suivante qui nous permet de diviser un noeud sans observations de la deuxième classe. Chaque fois que nous recherchons la meilleure scission pour un noeud t, nous remplaçons simplement (dans la diminution d'impuretés à deux classes que nous voulons maximiser) la proportion de la seconde classe dans le noeud gauche t_L par la proportion espérée $volume(t_L)/volume(t)$ (idem pour le noeud droit). Cela implique qu'un noeud enfant (t_L ou t_R) essaie de capturer le nombre maximum d'observations avec un volume minimal, alors que l'autre enfant cherche le contraire.*

Remarque 3 (**Absence d'échantillonage**). *La méthode d'échantillonnage correspondante est la suivante : pour chaque note t à diviser contenant n_t observations (inliers), générer n_t outliers uniformément sur le noeud correspondant pour ensuite pouvoir optimiser un critère de division à deux classes. Nous* évitons précisément de générer *ces n_t outliers en utilisant la proportion espérée $volume(t_L)/volume(t)$.*

RFs à une classe. Résumons l'algorithme dans sa version la plus générique. Il y a 7 paramètres : $max\_samples$, $max\_features\_tree$, $max\_features\_node$, γ, max_depth, n_trees, s_k.

Chaque arbre est classiquement construit sur un sous-ensemble aléatoire d'observations et de coordonnées/variables (Ho, 1998; Panov & Džeroski, 2007). Ce sous-ensemble aléatoire est un sous-échantillon de taille $max\_samples$, avec $max\_features\_tree$ variables choisies au hasard sans remplacement. L'arbre est construit en minimisant une version à une classe du critère de Gini (Gini, 1912), obtenue en remplaçant les quantités empiriques liées à la seconde classe (non observée) par les versions théoriques. Ceux-ci correspondent à une distribution uniforme pondérée, le poids augmentant lorsque le volume du noeud diminue, afin d'éviter des classes fortement déséquilibrées (volume vs. observations). En effet, lorsque leur profondeur augmente, les noeuds ont tendance à avoir des volumes plus petits tout en gardant un nombre d'observations (normales) relativement élevé.

De nouveaux noeuds sont construits (en minimisant ce critère) jusqu'à ce que la profondeur maximale max_depth soit atteinte. La minimisation est effectuée comme introduit dans (Amit & Geman, 1997), en définissant un grand nombre $max\_features\_node$ de variables et en recherchant sur une sélection aléatoire de celles-ci la meilleure division à chaque noeud. La forêt est composée d'un nombre n_trees d'arbres. Le score (prédiction) d'un point x est alors donné par $s_k(x)$, qui est la profondeur moyenne de x parmis la forêt.

6 Contributions sur scikit-learn

Comme autre contribution de cette thèse, deux algorithmes classiques de détection d'anomalies, Isolation Forest et Local Outlier Factor ont été implémentés et fusionnés sur scikit-learn.

Scikit-learn (Pedregosa et al., 2011), est une bibliothèque open-source fournissant des méthodes de machine learning bien établies. Il s'agit d'un module Python, ce dernier étant très populaire pour la programmation scientifique, du fait de son caractère interactif de haut niveau. Scikit-learn fournit un mécanisme de composition (à travers un objet *Pipeline*) pour combiner des estimateurs, des outils de prétraitement et des méthodes de sélection de modèle de façon à ce que l'utilisateur puisse facilement construire des algorithmes complexes. Le développement se fait sur *Github* [1], un service d'hébergement de référentiel Git qui facilite la collaboration, car le codage se fait en forte interaction avec d'autres développeurs. En raison du grand nombre de développeurs, l'accent est mis sur la préservation de la maintenabilité du projet, par exemple en évitant la duplication de code au prix d'une perte raisonnable de performance de calcul.

Ces contributions ont été supervisées par Alexandre Gramfort et financées par le Center for Data Science de Paris-Saclay. Il inclut également du travail sur la maintenance de scikit-learn comme la résolution de problèmes et la relecture de code en construction par d'autre contributeurs.

7 Conclusion and production scientifique

Les contributions de cette thèse peuvent être résumées comme suit. Tout d'abord, un critère de performance adéquat appelé courbe d'excès de masse est proposée (partie 3.3), afin de comparer les fonctions de score candidates. La publication correspondante est Goix et al. (2015c) :

— On Anomaly Ranking and Excess-Mass Curves. (AISTATS 2015).
Auteurs : Goix, Sabourin, and Clémençon.

Deuxièmement, des avancées dans la théorie des valeurs extrêmes multivariée sont apportées en fournissant des bornes non asymptotiques pour l'estimation de la STDF, fonctionnelle caractérisant la structure de dépendance des extrêmes (partie 4.1). La publication correspondante est Goix et al. (2015b) :

— Learning the dependence structure of rare events : a non-asymptotic study. (COLT 2015).
Auteurs : Goix, Sabourin, and Clémençon.

1. https ://github.com/scikit-learn

La troisième contribution consiste à développer une méthode statistique qui produit une représentation (possiblement parcimonieuse) de la structure de dépendance des extrêmes, tout en dérivant des bornes non asymptotiques pour évaluer la précision de la procédure d'estimation (partie 4.2). Cette contribution inclut également un algorithme basé sur la théorie des extrêmes multivariés qui retourne une fonction de score définie sur les régions extrêmes. Cette méthodologie s'applique donc directement à la détection d'anomalies. Les publications correspondantes sont Goix et al. (2016c), Goix et al. (2015a) et Goix et al. (2016b) :

— Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. (AISTATS 2016 and NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning).
  Auteurs : Goix, Sabourin, and Clémençon.
— Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection. (Under review for Journal of Multivariate Analysis).
  Auteurs : Goix, Sabourin, and Clémençon.

Comme quatrième contribution, nous montrons (empiriquement) que les critères EM ou MV sont capables de discriminer avec précision (relativement aux critères ROC ou PR) parmis les fonctions de score, en faible dimension. Par ailleurs, nous proposons une méthodologie basée sur du sous-échantillonnage des variables et de l'agrégation, pour faire passer à l'échelle l'utilisation de ces critères. Les publications correspondantes sont Goix (2016) et Goix & Thomas (2016) :

— How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms ? (ICML 2016, Workshop on Anomaly Detection).
  Auteur : Goix.
— How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms ? (to be submitted).
  Auteurs : Goix and Thomas.

La cinquième contribution de cette thèse est de développer une heuristique efficace pour construire des fonctions de score précises. Cela se fait en généralisant les forêts aléatoires au cadre de la classification à une classe. Le travail correspondant (à soumettre) est Goix et al. (2016a) :

— One-Class Splitting Criteria for Random Forests with Application to Anomaly Detection. (to be submitted).
  Auteurs : Goix, Brault, Drougard and Chiapino.

Enfin, deux algorithmes de détection d'anomalies classiques ont été implémentés et fusionnés sur scikit-learn. Ils sont utilisés dans cette dissertation à des fins de comparaison empirique pour attester de la pertinence des approches mentionnées

ci-dessus. Les pull requests correspondant à ces deux contributions sont disponibles à l'adresse suivante :

— https ://github.com/scikit-learn/scikit-learn/pull/4163 (Isolation Forest)
— https ://github.com/scikit-learn/scikit-learn/pull/5279 (LOF)

Contexte de ce travail. Cette thèse a été réalisée dans l'équipe STA (Statistiques et Applications) du département Traitement du Signal et de l'Image (TSI) de Télécom ParisTech. Les contributions présentées dans cette thèse ont été soutenues financièrement par l'Ecole Normale Supérieure de Cachan via un contrat doctoral pour normalien ainsi que par la chaire industrielle "Machine Learning for Big Data" de Telecom ParisTech. Les contributions scikit-learn ont été financées par le Center for Data Science de Paris Saclay pour ce qui est de la collaboration avec Alexandre Gramfort et par la chaire industrielle mentionnée ci-dessus en ce qui concerne la collaboration à l'Université de New York avec Andreas Müller.

Références

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9 :1545–1588, 1997.

A. Baillo. Total error in a plug-in estimator of level sets. *Statistics & probability letters*, 65 :411–417, 2003.

A. Baillo, J. A Cuesta-Albertos, and A. Cuevas. Convergence rates in nonparametric estimation of level sets. *Statistics & probability letters*, 53 :27–35, 2001.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *JMLR*, 9 :2015–2033, 2008.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25 :197–227, 2016.

L. Breiman. Random forests. *Machine learning*, 45 :5–32, 2001.

B. Cadre. Kernel estimation of density level sets. *JMVA*, 97 :999–1023, 2006.

S. Clémençon and J. Jakubowicz. Scoring anomalies : a m-estimation formulation. In *Proc. AISTATS*, volume 13, pages 659–667, 2013.

S. Clémençon and S. Robbiano. Anomaly Ranking as Supervised Bipartite Ranking. In *Proc. ICML*, 2014.

S. Clémençon and N. Vayatis. Nonparametric estimation of the precision-recall curve. In *Proc. ICML*, pages 185–192, 2009.

S. Clémençon and N. Vayatis. Overlaying classifiers : a practical approach to optimal scoring. *Constr Approx*, 32 :619–648, 2010.

D. A. Clifton, S. Hugueny, and L. Tarassenko. Novelty detection with multivariate extreme value statistics. *J Signal Process Syst.*, 65 :371–389, 2011.

D.A. Clifton, L. Tarassenko, N. McGrogan, D. King, S. King, and P. Anuzis. Bayesian extreme value statistics for novelty detection in gas-turbine engines. In *AEROSP CONF PROC*, pages 1–11, 2008.

S. Coles and J.A Tawn. Modeling extreme multivariate events. *JR Statist. Soc. B*, 53 :377–392, 1991.

D. Cooley, R.A. Davis, and P. Naveau. The pairwise beta distribution : A flexible parametric multivariate model for extremes. *JMVA*, 101 :2103–2117, 2010.

A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *Ann. Stat.*, pages 2300–2312, 1997.

J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proc. ICML*, 2006.

L. De Haan and A. Ferreira. *Extreme value theory : an introduction*. Springer Science & Business Media, 2007.

H. Drees and X. Huang. Best attainable rates of convergence for estimators of the stable tail dependence function. *JMVA*, 64 :25–47, 1998.

C. Désir, S. Bernard, C. Petitjean, and L. Heutte. A new random forest method for one-class classification. In *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 2012.

J. H. J. Einmahl, A. Krajina, and J. Segers. An m-estimator for tail dependence in arbitrary dimensions. *Ann. Stat.*, 40 :1764–1793, 2012.

J. H. J. Einmahl and J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Stat.*, 37 :2953–2989, 2009.

J. HJ Einmahl, L. de Haan, and V. I Piterbarg. Nonparametric estimation of the spectral measure of an extreme value distribution. *Ann. Stat.*, 29 :1401–1423, 2001.

J. HJ Einmahl and D. M Mason. Generalized quantile processes. *Ann. Stat.*, 20 : 1062–1078, 1992.

P. Embrechts, L. de Haan, and X. Huang. Modelling multivariate extremes. *Extremes and integrated risk management*, pages 59–67, 2000.

T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27 :861–874, 2006.

A-L. Fougères, J. P Nolan, and H Rootzén. Models for dependent extremes using stable mixtures. *Scandinavian Journal of Statistics*, 36 :42–59, 2009.

R. Genuer, J.-M. Poggi, and C. Tuleau. Random forests : some methodological insights. *arXiv :0811.3619*, 2008.

C. Gini. Variabilita e mutabilita. *Memorie di metodologia statistica*, 1912.

N. Goix. How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms ? In *ICML Workshop on Anomaly Detection*, 2016.

N. Goix, R. Brault, N. Drougard, and M. Chiapino. One Class Splitting Criteria for Random Forests with Application to Anomaly Detection. Submitted to AISTATS, 2016a.

N. Goix, A. Sabourin, and S. Clémençon. Sparse Representation of Multivariate Extremes. NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning, 2015a.

N. Goix, A. Sabourin, and S. Clémençon. Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection. In the reviewing process of JMVA, July 2016b.

N. Goix, A. Sabourin, and S. Clémençon. Learning the dependence structure of rare events : a non-asymptotic study. In *Proc. COLT*, 2015b.

N. Goix, A. Sabourin, and S. Clémençon. On Anomaly Ranking and Excess-Mass Curves. In *Proc. AISTATS*, 2015c.

N. Goix, A. Sabourin, and S. Clémençon. Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. In *Proc. AISTATS*, 2016c.

N. Goix and A. Thomas. How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms ? To be submitted, 2016.

T.K. Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20 :832–844, 1998.

X. Huang. Statistics of bivariate extreme values. *PhD thesis*, 1992.

H.J. Lee and S.J. Roberts. On-line novelty detection using the kalman filter and extreme value theory. In *ICPR*, pages 1–4, 2008.

F.T. Liu, K.M. Ting, and Z.H. Zhou. Isolation Forest. In *ICDM*, pages 413–422, 2008.

D. M. Mason and W. Polonik. Asymptotic normality of plug-in level set estimates. *The Annals of Applied Probability*, 19 :1108–1142, 2009.

P. Panov and S. Džeroski. *Combining bagging and random subspaces to create better ensembles*. Springer, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn : Machine learning in Python. *JMLR*, 12 :2825–2830, 2011.

W. Polonik. Measuring Mass Concentrations and Estimating Density Contour Cluster-An excess Mass Approach. *Ann. Stat.*, 23 :855–881, 1995.

W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69 :1–24, 1997.

FJ Provost, T. Fawcett, et al. Analysis and visualization of classifier performance : comparison under imprecise class and cost distributions. In *KDD*, volume 97, pages 43–48, 1997.

FJ Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proc. ICML*, volume 98, pages 445–453, 1998.

Y. Qi. Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics. *Acta Mathematicae Applicatae Sinica*, 13 : 167–175, 1997.

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15 :1154–1178, 2009.

S.J. Roberts. Novelty detection using extreme value statistics. *IEE P-VIS IMAGE SIGN*, 146 :124–129, Jun 1999.

S.J Roberts. Extreme value statistics for novelty detection in biomedical data processing. *IEE P-SCI MEAS TECH*, 147 :363–367, 2000.

A. Sabourin and P. Naveau. Bayesian dirichlet mixture model for multivariate extremes : A re-parametrization. *Comput. Stat. Data Anal.*, 71 :542–567, 2014.

B. Schölkopf, J.C Platt, J. Shawe-Taylor, A.J Smola, and R.C Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13 : 1443–1471, 2001.

C.D Scott and R.D Nowak. Learning minimum volume sets. *JMLR*, 7 :665–704, 2006.

T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *J. Comp. Graph. Stat.*, 15, 2012.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *JMLR*, 6 :211–232, 2005.

A.G. Stephenson. High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics*, 51 :77–88, 2009.

A. Thomas, V. Feuillard, and A. Gramfort. Calibration of One-Class SVM for MV set estimation. In *DSAA*, pages 1–9, 2015.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Stat.*, 25 :948–969, 1997.

J.-P. Vert and R. Vert. Consistency and convergence rates of one-class svms and related algorithms. *JMLR*, 6 :828–835, 2006.