On Anomaly Ranking and Excess-Mass Curves

Nicolas Goix
UMR LTCI No. 5141
Telecom ParisTech/CNRS
Institut Mines-Telecom
Paris, 75013, France

Anne Sabourin
UMR LTCI No. 5141
Telecom ParisTech/CNRS
Institut Mines-Telecom
Paris, 75013, France

Stéphan Clémençon
UMR LTCI No. 5141
Telecom ParisTech/CNRS
Institut Mines-Telecom
Paris, 75013, France

Abstract

Learning how to rank multivariate unlabeled observations depending on their degree of abnormality/novelty is a crucial problem in a wide range of applications. In practice, it generally consists in building a real valued "scoring" function on the feature space so as to quantify to which extent observations should be considered as abnormal. In the 1-d situation, measurements are generally considered as "abnormal" when they are remote from central measures such as the mean or the median. Anomaly detection then relies on tail analysis of the variable of interest. Extensions to the multivariate setting are far from straightforward and it is precisely the main purpose of this paper to introduce a novel and convenient (functional) criterion for measuring the performance of a scoring function regarding the anomaly ranking task, referred to as the *Excess-Mass* curve (EM curve). In addition, an adaptive algorithm for building a scoring function based on unlabeled data $X_1, \ldots, X_n$ with a nearly optimal EM is proposed and is analyzed from a statistical perspective.

1 Introduction

In a great variety of applications (*e.g.* fraud detection, distributed fleet monitoring, system management in data centers), it is of crucial importance to address anomaly/novelty issues from a ranking point of view. In contrast to novelty/anomaly detection (*e.g.* [4, 13, 10, 12]), novelty/anomaly ranking is very poorly documented in the statistical learning literature (see [14] for instance). However, when confronted with massive data, being enable to rank observations according to their supposed degree of abnormality may significantly improve operational processes and allow for a prioritization of actions to be taken, especially in situations where human expertise required to check each observation is time-consuming. When univariate, observations are usually considered as "abnormal" when they are either too high or else too small compared to central measures such as the mean or the median. In this context, anomaly/novelty analysis generally relies on the analysis of the tail distribution of the variable of interest. No natural (pre-) order exists on a d-dimensional feature space, $\mathcal{X} \subset \mathbb{R}^d$ say, as soon as $d > 1$. Extension to the multivariate setup is thus far from obvious and, in practice, the optimal ordering/ranking must be *learned* from training data $X_1, \ldots, X_n$, in absence of any parametric assumptions on the underlying probability distribution describing the "normal" regime. The most straightforward manner to define a preorder on the feature space $\mathcal{X}$ is to transport the natural order on the real half-line through a measurable *scoring function* $s : \mathcal{X} \to \mathbb{R}_+$: the "smaller" the score $s(X)$, the more "abnormal" the observation X is viewed. Any scoring function defines a preorder on $\mathcal{X}$ and thus a ranking on a set of new observations. An important issue thus concerns the definition of an adequate performance criterion, $\mathcal{C}(s)$ say, in order to compare possible candidate scoring function and to pick one eventually: optimal scoring functions s^* being then defined as those optimizing $\mathcal{C}$. Throughout the present article, it is assumed that the distribution F of the observable r.v. X is absolutely continuous w.r.t. Lebesgue measure *Leb* on $\mathcal{X}$, with density $f(x)$. The criterion should be thus defined in a way that the collection of level sets of an optimal scoring function $s^*(x)$ coincides with that related to f. In other words, any nondecreasing transform of the density should be optimal regarding the ranking

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.
Copyright 2015 by the authors.

performance criterion $\mathcal{C}$. According to the Empirical Risk Minimization (ERM) paradigm, a scoring function will be built in practice by optimizing an empirical version $\mathcal{C}_n(s)$ of the criterion over an adequate set of scoring functions $\mathcal{S}_0$ of controlled complexity (*e.g.* a major class of finite VC dimension). Hence, another desirable property to guarantee the universal consistency of ERM learning strategies is the uniform convergence of $\mathcal{C}_n(s)$ to $\mathcal{C}(s)$ over such collections $\mathcal{S}_0$ under minimal assumptions on the distribution $F(dx)$. In [1, 2], a functional criterion referred to as the Mass-Volume (MV) curve, admissible with respect to the requirements listed above has been introduced, extending somehow the concept of ROC curve in the unsupervised setup. Relying on the theory of *minimum volume* sets (see *e.g.* [8, 11] and the references therein), it has been proved that the scoring functions minimizing empirical and discretized versions of the MV curve criterion are accurate when the underlying distribution has compact support and a first algorithm for building nearly optimal scoring functions, based on the estimate of a finite collection of properly chosen minimum volume sets, has been introduced and analyzed. However, by construction, learning rate bounds are rather slow (of the order $n^{-1/4}$ namely) and cannot be established in the unbounded support situation, unless very restrictive assumptions are made on the tail behavior of $F(dx)$. See Figure 3 and related comments for an insight into the gain resulting from the concept introduced in the present paper in contrast to the MV curve minimization approach.

Given these limitations, it is the major goal of this paper to propose an alternative criterion for anomaly ranking/scoring, called the *Excess-Mass* curve (EM curve in short) here, based on the notion of *density contour clusters* [7, 3, 6]. Whereas minimum volume sets are solutions of volume minimization problems under mass constraints, the latter are solutions of mass maximization under volume constraints. Exchanging this way objective and constraint, the relevance of this performance measure is thoroughly discussed and accuracy of solutions which optimize statistical counterparts of this criterion is investigated. More specifically, rate bounds of the order $n^{-1/2}$ are proved, even in the case of unbounded support. Additionally, in contrast to the analysis carried out in [1], the model bias issue is tackled, insofar as the assumption that the level sets of the underlying density $f(x)$ belongs to the class of sets used to build the scoring function is relaxed here.

The rest of this paper is organized as follows. Section 3 introduces the notion of EM curve and that of optimal EM curve. Estimation in the compact support case is covered by section 4, extension to distributions with non compact support and control of the model bias are tackled in section 5. A simulation study is performed in section 6. All proofs are deferred to the Appendix section.

2 Background and related work

As a first go, we first provide a brief overview of the scoring approach based on the MV curve criterion, as a basis for comparison with that promoted in the present paper.

Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{1}_{\mathcal{E}}$, the Dirac mass at any point x by δ_x, $A \Delta B$ the symmetric difference between two sets A and B and by $\mathcal{S}$ the set of all scoring functions $s : \mathcal{X} \to \mathbb{R}_+$ integrable w.r.t Lebesgue measure. Let $s \in \mathcal{S}$. As defined in [1, 2], the MV-curve of s is the plot of the mapping $\alpha \in (0, 1) \mapsto MV_s(\alpha) = \lambda_s \circ \alpha_s^{-1}(\alpha)$, where $\alpha_s(t) = \mathbb{P}(s(X) \geq t)$, $\lambda_s(t) = Leb(\{x \in \mathcal{X}, s(x) \geq t\})$ and H^{-1} denotes the pseudo-inverse of any cdf $H : \mathbb{R} \to (0, 1)$. This induces a partial ordering on the set of all scoring functions: s is preferred to s' if $MV_s(\alpha) \leq MV_{s'}(\alpha)$ for all $\alpha \in (0, 1)$. One may show that $\mathrm{MV}^*(\alpha) \leq \mathrm{MV}_s(\alpha)$ for all $\alpha \in (0, 1)$ and any scoring function s, where $MV^*(\alpha)$ is the optimal value of the constrained minimization problem

$$\min_{\Gamma \ borelian} Leb(\Gamma) \text{ subject to } \mathbb{P}(X \in \Gamma) \geq \alpha. \qquad (1)$$

Suppose now that $F(dx)$ has a density $f(x)$ satisfying the following assumptions:

$\mathbf{A_1}$ *The density f is bounded, i.e. $||f(X)||_\infty < +\infty$.*
$\mathbf{A_2}$ *The density f has no flat parts: $\forall c \geq 0$, $\mathbb{P}\{f(X) = c\} = 0$* . One may then show that the curve MV^* is actually a MV curve, that is related to (any increasing transform of) the density f namely: $\mathrm{MV}^* = \mathrm{MV}_f$. In addition, the minimization problem (1) has a unique solution Γ^*_α of mass α exactly, referred to as *minimum volume set* (see [8]): $\mathrm{MV}^*(\alpha) = Leb(\Gamma^*_\alpha)$ and $F(\Gamma^*_\alpha) = \alpha$. Anomaly scoring can be then viewed as the problem of building a scoring function $s(x)$ based on training data such that MV_s is (nearly) minimum everywhere, *i.e.* minimizing $\|\mathrm{MV}_s - \mathrm{MV}^*\|_\infty \overset{def}{=} \sup_{\alpha \in [0,1]} |\mathrm{MV}_s(\alpha) - \mathrm{MV}^*(\alpha)|$. Since F is unknown, a minimum volume set estimate $\widehat{\Gamma}^*_\alpha$ can be defined as the solution of (1) when F is replaced by its empirical version $F_n = (1/n) \sum_{i=1}^n \delta_{X_i}$, minimization is restricted to a collection $\mathcal{G}$ of borelian subsets of $\mathcal{X}$ supposed not too complex but rich enough to include all density level sets (or reasonable approximants of the latter) and α is replaced by $\alpha - \phi_n$, where the *tolerance parameter* ϕ_n is a probabilistic upper bound for the supremum $\sup_{\Gamma \in \mathcal{G}} |F_n(\Gamma) - F(\Gamma)|$. Refer to [11] for further details. The set $\mathcal{G}$ should ideally offer statistical and computational advantages both at the same time. Allowing

for fast search on the one hand and being sufficiently complex to capture the geometry of target density level sets on the other. In [1], a method consisting in preliminarily estimating a collection of minimum volume sets related to target masses $0 < \alpha_1 < \ldots < \alpha_K < 1$ forming a subdivision of $(0, 1)$ based on training data so as to build a scoring function $s = \sum_k \mathbb{1}_{x \in \hat{\Gamma}^*_{\alpha_k}}$ has been proposed and analyzed. Under adequate assumptions (related to $\mathcal{G}$, the perimeter of the $\Gamma^*_{\alpha_k}$'s and the subdivision step in particular) and for an appropriate choice of $K = K_n$ either under the very restrictive assumption that $F(dx)$ is compactly supported or else by restricting the convergence analysis to $[0, 1-\epsilon]$ for $\epsilon > 0$, excluding thus the tail behavior of the distribution F from the scope of the analysis, rate bounds of the order $\mathcal{O}_{\mathbb{P}}(n^{-1/4})$ have been established to guarantee the generalization ability of the method.

Figure 3 illustrates the problems inherent to the use of the MV curve as a performance criterion for anomaly scoring in a "non asymptotic" context, due to the prior discretization along the mass-axis. In the 2-d situation described by Fig. 3 for instance, given the training sample and the partition of the feature space depicted, the MV criterion leads to consider the sequence of empirical minimum volume sets A_1, $A_1 \cup A_2$, $A_1 \cup A_3$, $A_1 \cup A_2 \cup A_3$ and thus the scoring function $s_1(x) = \mathbb{I}\{x \in A_1\} + \mathbb{I}\{x \in A_1 \cup A_2\} + \mathbb{I}\{x \in A_1 \cup A_3\}$, whereas the scoring function $s_2(x) = \mathbb{I}\{x \in A_1\} + \mathbb{I}\{x \in A_1 \cup A_3\}$ is clearly more accurate.

In this paper, a different functional criterion is proposed, obtained by exchanging objective and constraint functions in (1), and it is shown that optimization of an empirical discretized version of this performance measure yields scoring rules with convergence rates of the order $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$. In addition, the results can be extended to the situation where the support of the distribution F is not compact.

3 The Excess-Mass curve

The performance criterion we propose in order to evaluate anomaly scoring accuracy relies on the notion of *excess mass* and *density contour clusters*, as introduced in the seminal contribution [7]. The main idea is to consider a Lagrangian formulation of a constrained minimization problem, obtained by exchanging constraint and objective in (1): for $t > 0$,

$$\max_{\Omega \text{ borelian}} \left\{ \mathbb{P}(X \in \Omega) - t Leb(\Omega) \right\}. \qquad (2)$$

We denote by Ω^*_t any solution of this problem. As shall be seen in the subsequent analysis (see Proposition 3 below), compared to the MV curve approach, this formulation offers certain computational and theoretical advantages both at the same time: when letting (a discretized version of) the Lagrangian multiplier t increase from 0 to infinity, one may easily obtain solutions of empirical counterparts of (2) forming a *nested* sequence of subsets of the feature space, avoiding thus deteriorating rate bounds by transforming the empirical solutions so as to force monotonicity.

Definition 1. (Optimal EM curve) *The optimal Excess-Mass curve related to a given probability distribution $F(dx)$ is defined as the plot of the mapping*

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

Equipped with the notation above, we have: $EM^*(t) = \mathbb{P}(X \in \Omega^*_t) - t Leb(\Omega^*_t)$ for all $t > 0$. Notice also that $\mathrm{EM}^*(t) = 0$ for any $t > \|f\|_\infty \overset{def}{=} \sup_{x \in \mathcal{X}} |f(x)|$.

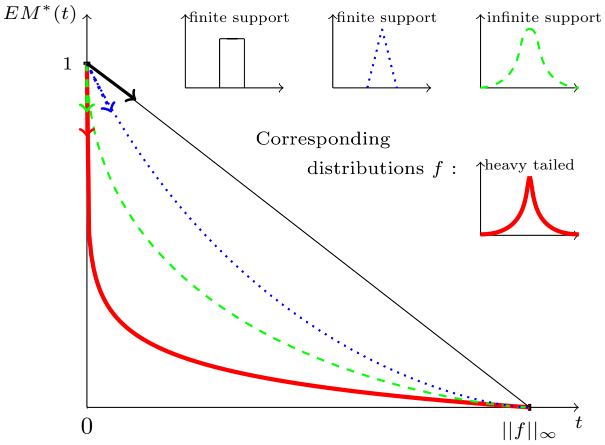

Figure 1: EM curves depending on densities

Lemma 1. (On existence and uniqueness) *For any subset Ω^*_t solution of (2), we have*

$$\{x, f(x) > t\} \subset \Omega^*_t \subset \{x, f(x) \geq t\} \textit{almost-everywhere},$$

and the sets $\{x, f(x) > t\}$ and $\{x, f(x) \geq t\}$ are both solutions of (2). In addition, under assumption $\mathbf{A_2}$,

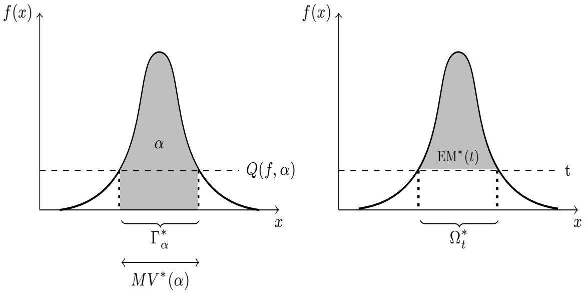

Figure 2: Comparison between $MV^*(\alpha)$ and $EM^*(t)$

the solution is unique:

$$\Omega_t^* = \{x, f(x) > t\} = \{x, f(x) \geq t\}.$$

Observe that the curve EM^* is always well-defined, since $\int_{f \geq t}(f(x) - t)dx = \int_{f>t}(f(x) - t)dx$. We also point out that $\mathrm{EM}^*(t) = \alpha(t) - t\lambda(t)$ for all $t > 0$, where we set $\alpha = \alpha_f$ and $\lambda = \lambda_f$.

Proposition 1. (Derivative and convexity of EM^*) *Suppose that assumptions* $\mathbf{A_1}$ *and* $\mathbf{A_2}$ *are fullfilled. Then, the mapping* EM^* *is differentiable and we have for all* $t > 0$*:*

$$\mathrm{EM}^{*'}(t) = -\lambda(t).$$

In addition, the mapping $t > 0 \mapsto \lambda(t)$ *being decreasing, the curve* EM^* *is convex.*

We now introduce the concept of Excess-Mass curve of a scoring function $s \in \mathcal{S}$.

Definition 2. (EM curves) *The* EM *curve of* $s \in \mathcal{S}$ *w.r.t. the probability distribution* $F(dx)$ *of a random variable* X *is the plot of the mapping*

$$\mathrm{EM}_s : t \in [0, \infty[\mapsto \sup_{A \in \{(\Omega_{s,l})_{l>0}\}} \mathbb{P}(X \in A) - tLeb(A), \tag{3}$$

where $\Omega_{s,t} = \{x \in \mathcal{X}, s(x) \geq t\}$ *for all* $t > 0$*. One may also write:* $\forall t > 0$, $\mathrm{EM}_s(t) = \sup_{u>0} \ \alpha_s(u) - t\lambda_s(u)$*. Finally, under assumption* $\mathbf{A_1}$*, we have* $\mathrm{EM}_s(t) = 0$ *for every* $t > \|f\|_\infty$*.*

Regarding anomaly scoring, the concept of EM curve naturally induces a partial order on the set of all scoring functions: $\forall(s_1, s_2) \in \mathcal{S}^2$, s_1 is said to be more accurate than s_2 when $\forall t > 0, \mathrm{EM}_{s_1}(t) \geq \mathrm{EM}_{s_2}(t)$. Observe also that the optimal EM curve introduced in Definition 1 is itself the EM curve of a scoring function, the EM curve of any strictly increasing transform of the density f namely: $\mathrm{EM}^* = \mathrm{EM}_f$. Hence, in the unsupervised framework, optimal scoring functions are those maximizing the EM curve everywhere. In addition, maximizing EM_s can be viewed as recovering a collection of subsets $(\Omega_t^*)_{t>0}$ with maximum mass when penalized by their volume in a linear fashion. An optimal scoring function is then any $s \in \mathcal{S}$ with the Ω_t^*'s as level sets, for instance any scoring function of the form

$$s(x) = \int_{t=0}^{+\infty} \mathbb{1}_{x \in \Omega_t^*} a(t) dt, \tag{4}$$

with $a(t) > 0$ (observe that $s(x) = f(x)$ for $a \equiv 1$).

Proposition 2. *(*Nature of anomaly scoring*) Let* $s \in \mathcal{S}$*. The following properties hold true.*

(i) The mapping EM_s *is non increasing on* $(0, +\infty)$*, takes its values in* $[0, 1]$ *and satisfies,* $\mathrm{EM}_s(t) \leq \mathrm{EM}^*(t)$ *for all* $t \geq 0$*.*

(ii) For $t \geq 0$*, we have:* $0 \leq \mathrm{EM}^*(t) - \mathrm{EM}_s(t) \leq \|f\|_\infty \inf_{u>0} Leb(\{s > u\} \Delta \{f > t\})$*.*

(iii) Let $\epsilon > 0$*. Suppose that the quantity* $\sup_{u>\epsilon} \int_{f^{-1}(\{u\})} 1/\|\nabla f(x)\| \quad d\mu(x)$ *is bounded, where* μ *denotes the* $(d-1)$*-dimensional Hausdorff measure. Set* $\epsilon_1 := \inf_T \|f - T \circ s\|_\infty$*, where the infimum is taken over the set* $\mathcal{T}$ *of all borelian increasing transforms* $T : \mathbb{R}_+ \to \mathbb{R}_+$*. Then*

$$\sup_{t \in [\epsilon + \epsilon_1, \|f\|_\infty]} |\mathrm{EM}^*(t) - \mathrm{EM}_s(t)| \leq C_1 \inf_{T \in \mathcal{T}} \|f - T \circ s\|_\infty$$

where $C_1 = C(\epsilon_1, f)$ *is a constant independent from* $s(x)$*.*

Assertion (ii) provides a control of the pointwise difference between the optimal EM curve and EM_s in terms of the error made when recovering a specific minimum volume set Ω_t^* by a level set of $s(x)$. Assertion (iii) reveals that, if a certain increasing transform of a given scoring function $s(x)$ approximates well the density $f(x)$, then $s(x)$ is an accurate scoring function w.r.t. the EM criterion. As the distribution $F(dx)$ is generally unknown, EM curves must be estimated. Let $s \in \mathcal{S}$ and $X_1, \ldots, X_n$ be an i.i.d. sample with common distribution $F(dx)$ and set $\widehat{\alpha}_s(t) = (1/n)\sum_{i=1}^n \mathbb{1}_{s(X_i) \geq t}$. The empirical EM curve of s is then defined as

$$\widehat{\mathrm{EM}}_s(t) = \sup_{u>0}\{\widehat{\alpha}_s(u) - t\lambda_s(u)\} \ .$$

In practice, it may be difficult to estimate the volume $\lambda_s(u)$ and Monte-Carlo approximation can naturally be used for this purpose.

4 A general approach to learn a scoring function

The concept of EM-curve provides a simple way to compare scoring functions but optimizing such a functional criterion is far from straightforward. As in [1], we propose to discretize the continuum of optimization problems and to construct a nearly optimal scoring function with level sets built by solving a finite collection of empirical versions of problem (2) over a subclass $\mathcal{G}$ of borelian subsets. In order to analyze the accuracy of this approach, we introduce the following additional assumptions.

$\mathbf{A_3}$ *All minimum volume sets belong to* $\mathcal{G}$*:*

$$\forall t > 0, \ \Omega_t^* \in \mathcal{G} \ .$$

$\mathbf{A_4}$ *The Rademacher average*

$$\mathcal{R}_n = \mathbb{E}\left[\sup_{\Omega\in\mathcal{G}} \frac{1}{n}\left|\sum_{i=1}^{n} \epsilon_i \mathbb{1}_{X_i\in\Omega}\right|\right]$$

is of order $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$, where $(\epsilon_i)_{i\geq 1}$ is a Rademacher chaos independent of the X_i's.

Assumption $\mathbf{A_4}$ is very general and is fulfilled in particular when $\mathcal{G}$ is of finite VC dimension, see [5], whereas the zero bias assumption $\mathbf{A_3}$ is in contrast very restrictive. It will be relaxed in section 5.

Let $\delta \in (0,1)$ and consider the complexity penalty $\Phi_n(\delta) = 2\mathcal{R}_n + \sqrt{\frac{log(1/\delta)}{2n}}$. We have for all $n \geq 1$:

$$\mathbb{P}\left(\left\{\sup_{G\in\mathcal{G}} (|P(G) - P_n(G)| - \Phi_n(\delta)) > 0\right\}\right) \leq \delta, \quad (5)$$

see [5] for instance. Denote by $F_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ the empirical measure based on the training sample $X_1, \ \ldots, \ X_n$. For $t \geq 0$, define also the signed measures:

$$\begin{aligned} &H_t(\,\cdot\,) = F(\,\cdot\,) - tLeb(\,\cdot\,) \\ \text{and} \qquad &H_{n,t}(\,\cdot\,) = F_n(\,\cdot\,) - tLeb(\,\cdot\,). \end{aligned}$$

Equipped with these notations, for any $s \in \mathcal{S}$, we point out that one may write $\mathrm{EM}^*(t) = \sup_{u\geq 0} H_t(\{x \in \mathcal{X}, f(x) \geq u\})$ and $\mathrm{EM}_s(t) = \sup_{u\geq 0} H_t(\{x \in \mathcal{X}, s(x) \geq u\})$. Let $K > 0$ and $0 < t_K < t_{K-1} < \ldots < t_1$. For k in $\{1, \ \ldots, \ K\}$, let $\hat{\Omega}_{t_k}$ be an *empirical t_k-cluster*, that is to say a borelian subset of $\mathcal{X}$ such that

$$\hat{\Omega}_{t_k} \in arg \max_{\Omega\in\mathcal{G}} H_{n,t_k}(\Omega).$$

The empirical excess mass at level t_k is then $H_{n,t_k}(\hat{\Omega}_{t_k})$. The following result reveals the benefit of viewing density level sets as solutions of (2) rather than solutions of (1) (corresponding to a different parametrization of the thresholds).

Proposition 3. (Monotonicity) *For any k in $\{1, \ \ldots, \ K\}$, the subsets $\cup_{i\leq k}\hat{\Omega}_{t_i}$ and $\cap_{i\geq k}\hat{\Omega}_{t_i}$ are still empirical t_k-clusters, just like $\hat{\Omega}_{t_k}$:*

$$H_{n,t_k}(\cup_{i\leq k}\hat{\Omega}_{t_i}) = H_{n,t_k}(\cap_{i\geq k}\hat{\Omega}_{t_i}) = H_{n,t_k}(\hat{\Omega}_{t_k}).$$

The result above shows that monotonous (regarding the inclusion) collections of empirical clusters can always be built. Coming back to the example depicted by Fig. 3, as t decreases, the $\hat{\Omega}_t$'s are successively equal to A_1, $A_1 \cup A_3$, and $A_1 \cup A_3 \cup A_2$, and are thus monotone as expected. This way, one fully avoids the problem inherent to the prior specification of a subdivision of the mass-axis in the MV-curve minimization approach (see the discussion in section 2).

Consider an increasing sequence of empirical t_k clusters $(\hat{\Omega}_{t_k})_{1\leq k\leq K}$ and a scoring function $s \in S$ of the form

$$s_K(x) := \sum_{k=1}^{K} a_k \mathbb{1}_{x\in\hat{\Omega}_{t_k}} \ , \qquad (6)$$

where $a_k > 0$ for every $k \in \{1, \ \ldots, \ K\}$. Notice that the scoring function (6) can be seen as a Riemann sum approximation of (4) when $a_k = a(t_k) - a(t_{k+1})$. For simplicity solely, we take $a_k = t_k - t_{k+1}$ so that the $\hat{\Omega}_{t_k}$'s are t_k-level sets of s_K, *i.e* $\hat{\Omega}_{t_k} = \{s \geq t_k\}$ and $\{s \geq t\} = \hat{\Omega}_{t_k}$ if $t \in]t_{k+1}, t_k]$. Observe that the results established in this paper remain true for other choices. In the asymptotic framework considered in the subsequent analysis, it is stipulated that $K = K_n \to \infty$ as $n \to +\infty$. We assume in addition that $\sum_{k=1}^{\infty} a_k < \infty$.

Remark 1. (Nested sequences) *For $L \leq K$, we have $\{\Omega_{s_L,l}, l \geq 0\} = (\hat{\Omega}_{t_k})_{0\leq k\leq L} \subset (\hat{\Omega}_{t_k})_{0\leq k\leq K} = \{\Omega_{s_K,l}, l \geq 0\}$, so that by definition, $EM_{s_L} \leq EM_{s_K}$.*

Remark 2. (Related work) *We point out that a very similar result is proved in [9] (see Lemma 2.2 therein) concerning the Lebesgue measure of the symmetric differences of density clusters.*

Remark 3. (Alternative construction) *It is noteworthy that, in practice, one may solve the optimization problems $\tilde{\Omega}_{t_k} \in \arg\max_{\Omega\in\mathcal{G}} H_{n,t_k}(\Omega)$ and next form $\hat{\Omega}_{t_k} = \cup_{i\leq k}\tilde{\Omega}_{t_i}$.*

The following theorem provides rate bounds describing the performance of the scoring function s_K thus built with respect to the EM curve criterion in the case where the density f has compact support.

Theorem 1. (Compact support case) *Assume that conditions $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ and $\mathbf{A_4}$ hold true, and that f has a compact support. Let $\delta \in]0,1[$, let $(t_k)_{k\in\{1, \ \ldots, \ K\}}$ be such that $\sup_{1\leq k<K}(t_k - t_{k+1}) = \mathcal{O}(1/\sqrt{n})$. Then, there exists a constant A independent from the t_k's, n and δ such that, with probability at least $1 - \delta$, we have:*

$$\begin{aligned} &\sup_{t\in]0,t_1]} |\mathrm{EM}^*(t) - \mathrm{EM}_{s_K}(t)| \\ &\qquad\qquad \leq \ \left(A + \sqrt{2\log(1/\delta)} + Leb(suppf)\right)\frac{1}{\sqrt{n}}. \end{aligned}$$

Remark 4. (Localization) *The problem tackled in this paper is that of scoring anomalies, which correspond to observations lying outside of "large" excess mass sets, namely density clusters with parameter t close to zero. It is thus essential to establish rate bounds for the quantity $\sup_{t\in]0,C[} |\mathrm{EM}^*(t) - \mathrm{EM}_{s_K}(t)|$, where $C > 0$ depends on the proportion of the "least normal" data we want to score/rank.*

5 Extensions - Further results

This section is devoted to extend the results of the previous one. We first relax the compact support assumption and next the one stipulating that all density level sets belong to the class $\mathcal{G}$, namely $\mathbf{A_3}$.

5.1 Distributions with non compact support

It is the purpose of this section to show that the algorithm detailed below produces a scoring function s such that EM_s is uniformly close to EM^* (Theorem 2). See Figure 3 as an illustration and a comparaison with the MV formulation as used as a way to recover empirical minimum volume set $\hat{\Gamma}_\alpha$.

Algorithm 1. *Suppose that assumptions* $\mathbf{A_1}$*,* $\mathbf{A_2}$*,* $\mathbf{A_3}$*,* $\mathbf{A_4}$ *hold true. Let* t_1 *such that* $\max_{\Omega\in\mathcal{G}} H_{n,t_1}(\Omega) \geq 0$*. Fix* $N > 0$*. For* $k = 1, \ldots, N$*,*

1. *Find* $\tilde{\Omega}_{t_k} \in \arg\max_{\Omega\in\mathcal{G}} H_{n,t_k}(\Omega)$ *,*
2. *Define* $\hat{\Omega}_{t_k} = \cup_{i\leq k}\tilde{\Omega}_{t_i}$
3. *Set* $t_{k+1} = \frac{t_1}{(1+\frac{1}{\sqrt{n}})^k}$ *for* $k \leq N-1$*.*

In order to reduce the complexity, we may replace steps 1 and 2 with $\hat{\Omega}_{t_k} \in \arg\max_{\Omega\supset\hat{\Omega}_{t_{k-1}}} H_{n,t_k}(\Omega)$*. The resulting piecewise constant scoring function is*

$$s_N(x) = \sum_{k=1}^{N}(t_k - t_{k+1})\mathbb{1}_{x\in\hat{\Omega}_{t_k}} \ . \qquad (7)$$

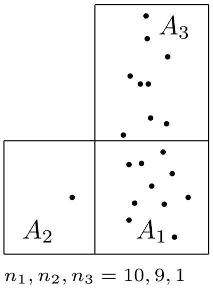

Figure 3: Sample of $n = 20$ points in a 2-d space, partitioned into three rectangles. As α increases, the minimum volume sets $\hat{\Gamma}_\alpha$ are successively equal to A_1, $A_1\cup A_2$, $A_1\cup A_3$, and $A_1\cup A_3\cup A_2$, whereas, in the EM-approach, as t decreases, the $\hat{\Omega}_t$'s are successively equal to A_1, $A_1\cup A_3$, and $A_1\cup A_3\cup A_2$.

The main argument to extend the above results to the case where $suppf$ is not bounded is given in Lemma 2 in the "Technical Details" section. The meshgrid (t_k) must be chosen adaptively, in a data-driven fashion. Let $h : \mathbb{R}_+^* \to \mathbb{R}_+$ be a decreasing function such that $\lim_{t\to 0} h(t) = +\infty$. Just like the previous approach, the grid is described by a decreasing sequence (t_k). Let $t_1 \geq 0$, $N > 0$ and define recursively $t_1 > t_2 > \ldots > t_N > t_{N+1} = 0$, as well as $\hat{\Omega}_{t_1}, \ldots, \hat{\Omega}_{t_N}$, through

$$t_{k+1} = t_k - (\sqrt{n})^{-1}\frac{1}{h(t_{k+1})} \qquad (8)$$

$$\hat{\Omega}_{t_k} = \arg\max_{\Omega\in\mathcal{G}} H_{n,t_k}(\Omega), \qquad (9)$$

with the property that $\hat{\Omega}_{t_{k+1}} \supset \hat{\Omega}_{t_k}$. As pointed out in Remark 3, it suffices to take $\hat{\Omega}_{t_{k+1}} = \tilde{\Omega}_{t_{k+1}} \cup \hat{\Omega}_{t_k}$, where $\tilde{\Omega}_{t_{k+1}} = \arg\max_{\Omega\in\mathcal{G}} H_{n,t_k}(\Omega)$. This yields the scoring function s_N defined by (7) such that by virtue of Lemma 2 (see the Technical Deails), with probability at least $1-\delta$,

$$\sup_{t\in]t_N,t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left(A + \sqrt{2\log(1/\delta)} + \sup_{1\leq k\leq N}\frac{\lambda(t_k)}{h(t_k)}\right)\frac{1}{\sqrt{n}} \ .$$

Therefore, if we take h such that $\lambda(t) = \mathcal{O}(h(t))$ as $t \to 0$, we can assume that $\lambda(t)/h(t) \leq B$ for t in $]0, t_1]$ since λ is decreasing, and we obtain:

$$\sup_{t\in]t_N,t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left(A + \sqrt{2\log(1/\delta)}\right)\frac{1}{\sqrt{n}} \ . \qquad (10)$$

On the other hand from $tLeb(\{f > t\}) \leq \int_{f>t} f \leq 1$, we have $\lambda(t) \leq 1/t$. Thus h can be chosen as $h(t) := 1/t$ for $t \in]0, t_1]$. In this case, (9) yields, for $k \geq 2$,

$$t_k = \frac{t_1}{(1+\frac{1}{\sqrt{n}})^{k-1}} \ . \qquad (11)$$

Theorem 2. *(*Unbounded support case*) Suppose that assumptions* $\mathbf{A_1}$*,* $\mathbf{A_2}$*,* $\mathbf{A_3}$*,* $\mathbf{A_4}$ *hold true, let* $t_1 > 0$ *and for* $k \geq 2$*, consider* t_k *as defined by* (11)*,* Ω_{t_k} *by (8), and* s_N *(7). Then there is a constant* A *independent from* N*,* n *and* δ *such that, with probability larger than* $1-\delta$*, we have:*

$$\sup_{t\in]0,t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left[A + \sqrt{2\log(1/\delta)}\right]\frac{1}{\sqrt{n}} + o_N(1),$$

where $o_N(1) = 1 - EM^*(t_N)$*. In addition,* $s_N(x)$ *converges to* $s_\infty(x) := \sum_{k=1}^{\infty}(t_{k+1} - t_k)\mathbb{1}_{\hat{\Omega}_{t_{k+1}}}$ *as* $N \to \infty$ *and* s_∞ *is such that, for all* $\delta \in (0,1)$*, we have with probability at least* $1-\delta$*:*

$$\sup_{t\in]0,t_1]} |EM^*(t) - EM_{s_\infty}(t)| \leq \left[A + \sqrt{2\log(1/\delta)}\right]\frac{1}{\sqrt{n}}$$

5.2 Bias analysis

In this subsection, we relax assumption $\mathbf{A_3}$. For any collection $\mathcal{C}$ of subsets of $\mathbb{R}^d$, $\sigma(\mathcal{C})$ denotes here the σ-algebra generated by $\mathcal{C}$. Consider the hypothesis below.

$\tilde{\mathbf{A}}_\mathbf{3}$ *There exists a countable subcollection of $\mathcal{G}$, $F = \{F_i\}_{i\geq 1}$ say, forming a partition of $\mathcal{X}$ and such that $\sigma(F) \subset \mathcal{G}$.*

Denote by f_F the best approximation (for the L_1-norm) of f by piecewise functions on F,

$$f_F(x) := \sum_{i\geq 1} \mathbb{1}_{x\in F_i} \frac{1}{Leb(F_i)} \int_{F_i} f(y)dy \ .$$

Then, variants of Theorems 1 and 2 can be established without assumption $\mathbf{A_3}$, as soon as $\tilde{\mathbf{A}}_\mathbf{3}$ holds true, at the price of the additional term $\|f - f_F\|_{L^1}$ in the bound, related to the inherent bias. For illustration purpose, the following result generalizes one of the inequalities stated in Theorem 2:

Theorem 3. *(*Biased empirical clusters*) Suppose that assumptions $\mathbf{A_1}$, $\mathbf{A_2}$, $\tilde{\mathbf{A}}_\mathbf{3}$, $\mathbf{A_4}$ hold true, let $t_1 > 0$ and for $k \geq 2$ consider t_k defined by (11), Ω_{t_k} by (8), and s_N by (7). Then there is a constant A independent from N, n, δ such that, with probability larger than $1-\delta$, we have:*

$$\sup_{t\in]0,t_1]} |EM^*(t) - EM_{s_N}(t)|$$
$$\leq \ \left[A + \sqrt{2\log(1/\delta)}\right] \frac{1}{\sqrt{n}} + \|f - f_F\|_{L^1} + o_N(1),$$

where $o_N(1) = 1 - EM^*(t_N)$.

Remark 5. (Hypercubes) *In practice, one defines a sequence of models $F_l \subset \mathcal{G}_l$ indexed by a tuning parameter l controlling (the inverse of) model complexity, such that $\|f - f_{F_l}\|_{L^1} \to 0$ as $l \to 0$. For instance, the class F_l could be formed by disjoint hypercubes of side length l.*

6 Simulation examples

Algorithm 1 is here implemented from simulated 2-*d heavy-tailed* data with common density $f(x,y) = 1/2 \times 1/(1+|x|)^3 \times 1/(1+|y|)^2$. The training set is of size $n = 10^5$, whereas the test set counts 10^6 points. For $l > 0$, we set $\mathcal{G}_l = \sigma(F)$ where $F_l = \{F_i^l\}_{i\in\mathbb{Z}^2}$ and $F_i^l = [li_1, li_1+1] \times [li_2, li_2+1]$ for all $i = (i_1, i_2) \in \mathbb{Z}^2$. The bias of the model is thus bounded by $\|f - f_F\|_\infty$, vanishing as $l \to 0$ (observe that the bias is at most of order l as soon as f is Lipschitz for instance). The scoring function s is built using the points located in $[-L, L]^2$ and setting $s = 0$ outside of $[-L, L]^2$. Practically, one takes L as the maximum norm value of the points in the training set, or such that an empirical estimate of $\mathbb{P}(X \in [-L, L]^2)$ is very close to 1 (here one obtains 0.998 for $L = 500$). The implementation of our algorithm involves the use of a sparse matrix to store the data in the partition of hypercubes, such that the complexity of the procedure for building the scoring function s and that of the computation of its empirical EM-curve is very small compared to that needed to compute f_{F_l} and $EM_{f_{F_l}}$, which are given here for the sole purpose of quantifying the model bias.

Fig. 4 illustrates as expected the deterioration of EM_s for large l, except for t close to zero: this corresponds to the model bias. However, Fig. 5 reveals an "overfitting" phenomenon for values of t close to zero, when l is fairly small. This is mainly due to the fact that subsets involved in the scoring function are then tiny in regions where there are very few observations (in the tail of the distribution). On the other hand, for the largest values of t, the smallest values of l give the best results: the smaller the parameter l, the weaker the model bias and no overfitting is experienced because of the high local density of the observations. Recalling the notation $EM^*_{\mathcal{G}}(t) = \max_{\Omega\in\mathcal{G}} H_t(\Omega) \leq EM^*(t) = \max_{\Omega\ meas.} H_t(\Omega)$ so that the bias of our model is $EM^* - EM^*_{\mathcal{G}}$, Fig. 6 illustrates the variations of the bias with the wealth of our model characterized by l the width of the partition by hypercubes. Notice that partitions with small l are not so good approximation for large t, but are performing as well as the other in the extreme values, namely when t is close to 0. On the top of that, those partitions have the merit not to overfit the extreme datas, which typically are isolated.

This empirical analysis demonstrates that introducing a notion of adaptivity for the partition F, with progressively growing bin-width as t decays to zero and as the hypercubes are being selected in the construction of s (which crucially depends on local properties of the empirical distribution), drastically improves the accuracy of the resulting scoring function in the EM curve sense.

7 Conclusion

Prolongating the contribution of [1], this article provides an alternative view (respectively, an other parameterization) of the anomaly scoring problem, leading to another adaptive method to build scoring functions, which offers theoretical and computational advantages both at the same time. This novel formulation yields a procedure producing a nested sequence of empirical density level sets, and exhibits a good performance, even in the non compact support case. In addition, the model bias has been incorporated in the rate bound analysis.

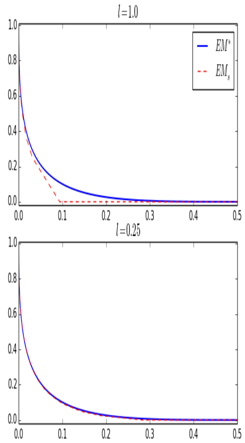

Figure 4: Optimal and realized EM curves

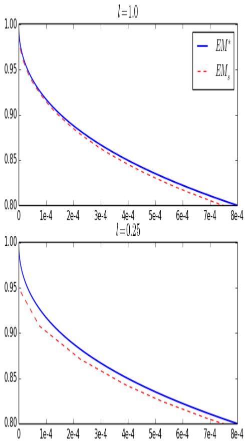

Figure 5: Zoom near 0

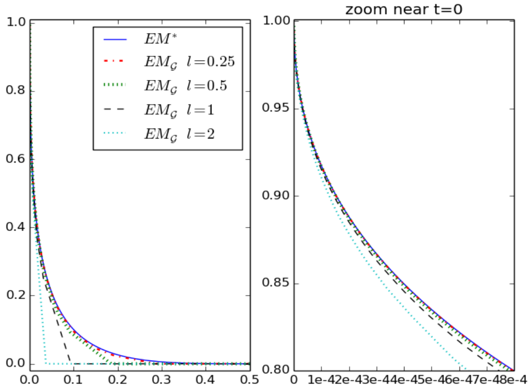

Figure 6: $EM_{\mathcal{G}}$ for different l

Technical Details

Proof of Theorem 1 (Sketch of) The proof results from the following lemma, which does not use the compact support assumption on f and is the starting point of the extension to the non compact support case (section 5.1).

Lemma 2. *Suppose that assumptions* $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ *and* $\mathbf{A_4}$ *are fulfilled. Then, for* $1 \le k \le K-1$, *there exists a constant* A *independent from* n *and* δ, *such that, with probability at least* $1-\delta$, *for* t *in* $]t_{k+1}, t_k]$,

$$|\mathrm{EM}^*(t) - \mathrm{EM}_{s_K}(t)| \;\le\; \left(A + \sqrt{2log(1/\delta)}\right)\frac{1}{\sqrt{n}} \\ +\; \lambda(t_{k+1})(t_k - t_{k+1}).$$

The detailed proof of this lemma is in the supplementary material, and is a combination on the two following results, the second one being a straightforward consequence of the derivative property of EM^* (Proposition 1):

- With probability at least $1-\delta$, for $k \in \{1, ..., K\}$,

  $$0 \le EM^*(t_k) - EM_{s_K}(t_k) \le 2\Phi_n(\delta)\ .$$

- Let k in $\{1, ..., K-1\}$. Then for every t in $]t_{k+1}, t_k]$,

  $$0 \le EM^*(t) - EM^*(t_k) \le \lambda(t_{k+1})(t_k - t_{k+1})\ .$$

Proof of Theorem 2 (Sketch of) The first assertion is a consequence of (10) combined with the fact that

$$\sup_{t\in]0,t_N]} |EM^*(t) - EM_{s_N}(t)| \;\le\; 1 - EM_{s_N}(t_N) \\ \le\; 1 - EM^*(t_N) + 2\Phi_n(\delta)$$

holds true with probability at least $1-\delta$. For the second part, it suffices to observe that $s_N(x)$ (absolutely) converges to s_∞ and that, as pointed out in Remark 1, $EM_{s_N} \le EM_{s_\infty}$.

Proof of Theorem 3 (Sketch of) The result directly follows from the following lemma, which establishes an upper bound for the bias, with the notations $\mathrm{EM}^*_{\mathcal{C}}(t) := \max_{\Omega\in\mathcal{C}} H_t(\Omega) \le \mathrm{EM}^*(t) = \max_{\Omega\ meas.} H_t(\Omega)$ for any class of measurable sets $\mathcal{C}$, and $\mathcal{F} := \sigma(F)$ so that by assumption $\mathbf{A_3}$, $\mathcal{F} \subset \mathcal{G}$. Details are omitted due to space limits.

Lemma 3. *Under assumption* $\tilde{\mathbf{A}}_\mathbf{3}$, *we have for every* t *in* $[0, \|f\|_\infty]$,

$$0 \le \mathrm{EM}^*(t) - \mathrm{EM}^*_{\mathcal{F}}(t) \le \|f - f_F\|_{L^1}\ .$$

The model bias $\mathrm{EM}^* - \mathrm{EM}^*_{\mathcal{G}}$ *is then uniformly bounded by* $\|f - f_F\|_{L^1}$.

To prove this lemma (see the supplementary material for details), one shows that:

$$\mathrm{EM}^*(t) - \mathrm{EM}^*_{\mathcal{F}}(t) \le \int_{f>t} (f - f_F) \\ + \int_{\{f>t\}\setminus\{f_F>t\}} (f_F - t) \\ - \int_{\{f_F>t\}\setminus\{f>t\}} (f_F - t)\ ,$$

where we use the fact that for all $t > 0$, $\{f_F > t\} \in \mathcal{F}$ and $\forall F \in \mathcal{F}$, $\int_G f = \int_G f_F$. It suffices then to observe that the second and the third term in the bound are non-positive.

References

[1] S. Clémençon and J. Jakubowicz. Scoring anomalies: a M-estimation approach. 2013.

[2] S. Clémençon and S. Robbiano. Anomaly ranking as supervised bipartite ranking. In *Proceedings of ICML 2014*, 2014.

[3] J.A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82(397):267–270, 1987.

[4] V. Koltchinskii. M-estimation, convexity and quantiles. *The Annals of Statistics*, 25(2):435–477, 1997.

[5] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *The Annals of Statistics*, 34:2593–2706, 2006.

[6] D.W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415):738–746, 1991.

[7] W. Polonik. Measuring mass concentrations and estimating density contour cluster-an excess mass approach. *The annals of Statistics*, 23(3):855–881, 1995.

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

[9] W. Polonik. The silhouette, concentration functions and ml-density estimation under order restrictions. *The Annals of Statistics*, 26(5):1857–1877, 10 1998.

[10] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[11] C. Scott and R. Nowak. Learning Minimum Volume Sets. *Journal of Machine Learning Research*, 7:665–704, 2006.

[12] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *J. Machine Learning Research*, 6:211–232, 2005.

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

[14] K. Viswanathan, L. Choudur, V. Talwar, C. Wang, G. Macdonald, and W. Satterfield. Ranking anomalies in data centers. In R.D.James, editor, *Network Operations and System Management*, pages 79–87. IEEE, 2012.