Scoring anomalies: a M-estimation formulation

Stéphan Clémençon, Jérémie Jakubowicz

Scoring anomalies: a M-estimation formulation

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

Jérémie Jakubowicz
UMR SAMOVAR No. 5157
Telecom Sud Paris/CNRS
Institut Mines-Telecom
Evry, 91011, France

Abstract

It is the purpose of this paper to formulate the issue of scoring multivariate observations depending on their degree of abnormality/novelty as an unsupervised learning task. Whereas in the 1-d situation, this problem can be dealt with by means of tail estimation techniques, observations being viewed as all the more "abnormal" as they are located far in the tail(s) of the underlying probability distribution. In a wide variety of applications, it is desirable to dispose of a scalar valued "scoring" function allowing for comparing the degree of abnormality of multivariate observations. Here we formulate the issue of scoring anomalies as a M-estimation problem. A (functional) performance criterion is proposed, whose optimal elements are, as expected, nondecreasing transforms of the density. The question of empirical estimation of this criterion is tackled and preliminary statistical results related to the accuracy of partition-based techniques for optimizing empirical estimates of the empirical performance measure are established.

1 INTRODUCTION

In a wide variety of applications, ranging from the monitoring of aircraft engines in aeronautics to non destructive control quality in the industry through fraud detection, *anomaly/novelty detection* is of crucial importance. In practice, its very purpose is to rank observations by degree of abnormality/novelty, rather than simply assigning them a binary label, "abnormal" *vs* "normal". Whereas in the case of univariate observations generally, abnormal values are those which are extremes, *i.e.* "too large" or "too small" in regard to central quantities such as the mean or the median, and anomaly detection may then derive from standard tail distribution analysis, it is far from easy to formulate the issue in a multivariate situation. In high-dimensional situations, a variety of statistical techniques, relying on the concept of *minimum volume set* proposed in the seminal contribution of Polonik [1997], have been developed in order to split the feature space, $\mathcal{X} \subset \mathbb{R}^d$ with $d \geq 1$ say, into two halves and decide whether observations should be considered as normal or not (see also Scott and Nowak [2006] and Koltchinskii [1997] for closely related notions). The problem considered here is of different nature, the goal pursued is not to assign to all possible observations a label "normal" *vs* "abnormal", but to rank them according to their level of "abnormality". The most natural way to define a preorder on the feature space $\mathcal{X}$ is to transport the natural order on the real line through some (measurable) *scoring function* $s : \mathcal{X} \to \mathbb{R}_+$: the "smaller" the score $s(X)$, the more likely the observation X is viewed as "abnormal". This problem shall be here referred to as *anomaly scoring* and can be related to the literature dedicated to *statistical depth functions* in nonparametric statistics and operations research, see Zuo and Serfling [2000] and the references therein. Such functions are generally proposed *ad hoc* to defined a "center" for the probability distribution of interest and a notion of distance to the latter. The angle embraced in this paper is quite different, it is that of statistical learning theory. Our objective is indeed twofold: 1) propose a performance criterion for the anomaly scoring problem so as to formulate it in terms of M-estimation 2) investigate the accuracy of scoring rules which optimize empirical estimates of the criterion thus tailored.

Due to the global nature of the problem, the criterion we promote is *functional* and is referred to as the

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

Mass-Volume curve (MV-curve in abbreviated form). The latter induces a partial preorder on the set of scoring functions: the collection of optimal elements is defined as the set of scoring functions whose MV-curve is minimum everywhere and is shown to coincide, as expected, as increasing transforms of the underlying probability density. The issue of estimating the criterion based on data is then tackled and statistical results for learning strategies based on the minimization of such empirical estimates are then established.

The remainder of the article is organized as follows. Section 2 sets out the main notations and briefly recalls the crucial notions related to anomaly detection on which the results of the paper rely. Section 3 first provides an informal description of the anomaly scoring problem and then introduces a criterion dedicated to evaluate the performance of any scoring function in this context. The set of optimal elements is described and statistical estimation of the criterion proposed is also tackled. Statistical learning of an anomaly scoring function is then formulated as a functional M-estimation problem in section 4. Section 5 is devoted to the study of (partition-based) learning techniques for the design of optimal anomaly scoring functions based on empirical estimates of the performance measure, according to the *empirical risk minimization* paradigm. Technical proofs are postponed to the Appendix section.

2 THEORETICAL BACKGROUND

As a first go, we start off with describing the mathematical setup and recalling key concepts in anomaly detection involved in the subsequent analysis.

2.1 Framework and Notations

Here and throughout, we suppose that we observe independent and identically distributed realisations $X_1, \ldots, X_n$ of an unknown continuous probability distribution $F(dx)$, copies of a generic random variable X, taking their values in a (possibly very high dimensional) feature space $\mathcal{X} \subset \mathbb{R}^d$, with $d \geq 1$. The density of the random variable X with respect to $\lambda(dx)$, Lebesgue measure on $\mathbb{R}^d$, is denoted by $f(x)$, its support by $suppF$ and the indicator function of any event $\mathcal{E}$ by $\mathbb{I}\{\mathcal{E}\}$. The pseudo-inverse of any càd-làg function $H(x)$ on $\mathbb{R}$ is defined by $H^{-1}(u) = \inf\{t \in \mathbb{R} : H(t) \geq u\}$. The essential supremum of any nonnegative r.v. $|Y|$ is denoted by $||Y||_\infty$. A natural way of defining preorders on $\mathcal{X}$ is to map its elements onto $\mathbb{R}_+$ and use the natural order on the real half-line. For any measurable function $s : \mathcal{X} \mapsto \mathbb{R}_+$, we denote by $\preceq_s$ the preorder on $\mathcal{X}$ defined by

$$\forall (x, x') \in \mathcal{X}^2 \text{: } x \preceq_s x' \text{ iff } s(x) \leq s(x').$$

We denote the level sets of any scoring function s by:

$$\Omega_{s,t} = \{x \in \mathcal{X} : \ \ s(x) \geq t\}, \ \ t \in [-\infty, \ +\infty].$$

Observe that the sequence is decreasing (for the inclusion, as t increases from $-\infty$ to $+\infty$):

$$\forall (t, t') \in \mathbb{R}^2, \ \ t \geq t' \Rightarrow \Omega_{s,t} \subset \Omega_{s,t'}$$

and that $\lim_{t \to +\infty} \Omega_{s,t} = \emptyset$ and $\lim_{t \to -\infty} \Omega_{s,t} = \mathcal{X}$[1]. When the function $s(x)$ is additionally integrable w.r.t. Lebesgue measure, it is called a *scoring function*. The set of all scoring functions is denoted by $\mathcal{S}$.

The following quantities shall also be used in the sequel. For any scoring function s and threshold level $t \geq 0$, define:

$$\begin{aligned} \alpha_s(t) &= \mathbb{P}\{s(X) \geq t\}, \\ \lambda_s(t) &= \lambda\left(\{x \in \mathcal{X} : \ \ s(x) \geq t\}\right). \end{aligned}$$

The quantity $\alpha_s(t)$ is referred to as the *mass* of the level set $\Omega_{s,t}$, while $\lambda_s(t)$ is generally termed the *volume* (w.r.t. Lebesgue measure). Notice that the volumes are finite on $\mathbb{R}_+^*$: $\forall t > 0$, $\lambda_s(t) \leq \int_{\mathcal{X}} s(x)dx/t < +\infty$. Reciprocally, for any (total) preorder $\preceq$ on $\mathcal{X}$, one may define a scoring function s such that $\preceq$ coincides with $\preceq_s$. Indeed, for all $x \in \mathcal{X}$, consider the (supposedly measurable) set $\Omega_{\preceq}(x) = \{x' \in \mathcal{X} : \ \ x' \preceq x\}$. Then, for any finite positive measure $\mu(dx)$ with $\mathcal{X}$ as support, define the scoring function $s_{\mu, \preceq}(x) = \int_{x' \in \mathcal{X}} \mathbb{I}\{x' \in \Omega(x)\}\mu(dx')$. It is immediate to check that $s_{\mu, \preceq}$ induces the preorder $\preceq$ on $\mathcal{X}$.

For any $\alpha \in (0, 1)$ and any scoring function s, $Q(s, \alpha) = \inf\{u \in \mathbb{R} : \ \ \mathbb{P}\{s(X) \leq u\} \geq 1 - \alpha\}$ denotes the quantile at level $1 - \alpha$ of $s(X)$'s distribution throughout the paper. We also set $Q^*(\alpha) = Q(f, \alpha)$ for all $\alpha \in (0, 1)$.

2.2 Minimum Volume Sets

The notion of minimum volume sets has been introduced in the seminal contribution Polonik [1997] in order to describe regions where a multivariate r.v. X takes its values with highest/smallest probability. Let $\alpha \in (0, 1)$, a minimum volume set Ω_α^* of mass at least α is any solution of the constrained minimization problem

$$\min_\Omega \lambda(\Omega) \text{ subject to } \mathbb{P}\{X \in \Omega\} \geq \alpha,$$

[1] Recall that a sequence $(A_n)_{n \geq 1}$ of subsets of an ensemble E is said to converge iff $\limsup A_n = \liminf A_n$. In such a case, one defines its limit, denoted by $\lim A_n$ as $\limsup A_n = \liminf A_n$.

where the minimum is taken over all measurable subsets Ω of $\mathcal{X}$. Application of this concept includes in particular novelty/anomaly detection: for large values of α, abnormal observations are those which belong to the complementary set $\mathcal{X} \setminus \Omega_\alpha^*$. In the continuous setting, it can be shown that there exists a threshold value $t_\alpha^* \stackrel{def}{=} Q(f, \alpha) \geq 0$ such that the level set Ω_{f,t_α^*} is a solution of the constrained optimization problem above. The (generalized) quantile function is then defined by:

$$\forall \alpha \in (0,1), \;\; \lambda^*(\alpha) \stackrel{def}{=} \lambda(\Omega_\alpha^*).$$

The following assumptions shall be used in the subsequent analysis.

$\mathbf{A_1}$: the density f is bounded, *i.e.* $||f(X)||_\infty < +\infty$.

$\mathbf{A_2}$: the density f has no flat parts, *i.e.* for any constant $c \geq 0$,

$$\mathbb{P}\{f(X) = c\} = 0.$$

Under the hypotheses above, for any $\alpha \in (0,1)$, there exists a unique minimum volume set Ω_{f,t_α^*} (up to subsets of null F-measure), whose mass is equal to α exactly. Additionally, the mapping λ^* is continuous on $(0,1)$ and uniformly continuous on $[0, 1-\epsilon]$ for all $\epsilon \in (0,1)$ (when the support of $F(dx)$ is compact, uniform continuity holds on the whole interval $[0,1]$).

From a statistical perspective, estimates $\widehat{\Omega}_\alpha^*$ of minimum volume sets are built by replacing the unknown probability distribution F by its empirical version $F_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ and restricting optimization to a collection $\mathcal{A}$ of borelian subsets of $\mathcal{X}$, supposed rich enough to include all density level sets (or reasonable approximants of the latter). In Polonik [1997], functional limit results are derived for the generalized empirical quantile process $\{\lambda(\widehat{\Omega}_\alpha^*) - \lambda^*(\alpha)\}$ under certain assumptions for the class $\mathcal{A}$ (stipulating in particular that $\mathcal{A}$ is a Glivenko-Cantelli class for $F(dx)$). In Scott and Nowak [2006], it is proposed to replace the level α by $\alpha - \phi_n$ where ϕ_n plays the role of tolerance parameter (of the same order as the supremum $\sup_{\Omega \in \mathcal{A}} |F_n(\Omega) - F(\Omega)|$ roughly, complexity of the class $\mathcal{A}$ being controlled by the VC dimension or by means of the concept of Rademacher averages, so as to establish rate bounds at $n < +\infty$ fixed.

Alternatively, so-termed *plug-in* techniques, consisting in computing first an estimate $\widehat{f}$ of the density f and considering next level sets $\Omega_{\widehat{f},t}$ of the resulting estimator have been investigated in several papers, among which Tsybakov [1997] or Rigollet and Vert [2009] for instance. Such an approach however yields significant computational issues even for moderate values of the dimension, inherent to the curse of dimensionality phenomenon.

3 SCORING ANOMALIES

In this section, the issue of scoring observations depending on their level of novelty/abnormality is first described in an informal manner and next formulated quantitatively, as a functional optimization problem.

3.1 Overall Objective

The idea promoted through this article is to learn a scoring function s, based on training data $X_1, \ldots, X_n$, so as to describe extremal behavior of the (high dimensional) random vector X by that of the univariate variable $s(X)$, which can be summarized by its tail behavior near 0: hopefully, the smaller the score $s(X)$, the more abnormal/rare the observation X should be considered. Hence, an optimal scoring function should naturally permit to rank observations X by increasing order of magnitude of $f(X)$. The set of optimal scoring functions is then given by:

$$\mathcal{S}^* = \{T \circ f : \; T : \; Imf(X) \to \mathbb{R}_+ \text{ strictly increasing}\},$$

denoting by $Imf(X)$ the image of the mapping $f(X)$.

The result below connects the notion of optimal scoring function to minimum volume sets. It shows that solving the anomaly scoring problem boils down to recovering all density level sets and can be seen as an overlaid collection of minimum volume set estimation problems. This observation shall turn out to be very useful when designing practical learning strategies, see section 5.

Lemma 1. (Optimal scoring functions) *A bounded scoring function s^* belongs to $\mathcal{S}^*$ iff there exists a nonnegative borelian function ω and a continuous positive r.v. V such that:*

$$\forall x \in \mathcal{X}, \;\; s^*(x) = \sup_{u \in \mathcal{X}} s^*(u) + \mathbb{E}\left[\omega(V) \cdot \mathbb{I}\{f(x) \geq V\}\right].$$

The proof is straightforward and is left to the reader, due to space limitations. Notice that the equation above is in particular fulfilled for $s^*(x) = f(x)$, when $\omega \equiv ||f(X)||_\infty$ and $V = ||f(X)||_\infty \cdot U$, where U is uniformly distributed on $(0,1)$.

3.2 A Functional Criterion

We now introduce the concept of Mass-Volume curve and shows that it is a natural criterion to evaluate the accuracy of decision rules in regard to anomaly scoring.

Definition 2. (True Mass-Volume curve) *Let $s \in \mathcal{S}$. Its Mass-Volume curve (*MV *curve in abbreviated form) with respect to X's probability distribution is the parametrized curve:*

$$t \in \mathbb{R}_+ \mapsto (\alpha_s(t), \lambda_s(t)) \in [0,1] \times [0, +\infty].$$

In addition, if α_s has no flat parts, the MV *curve can also be defined as the plot of the mapping*

$$\mathrm{MV}_s: \quad \alpha \in (0,1) \mapsto \mathrm{MV}_s(\alpha) \stackrel{def}{=} \lambda_s \circ \alpha_s^{-1}(\alpha).$$

By convention, points of the curve corresponding to possible jumps are connected by line segments, so that the MV *curve is continuous and can always be seen as the graph of a mapping MV_s on $(0, \bar{\alpha}_s)$ with $\bar{\alpha}_s = \sup_{t>0} \alpha_s(t)$. When $\bar{\alpha}_s < 1$, one sets $\mathrm{MV}_s(\alpha) = +\infty$ for $\alpha \in [\bar{\alpha}_s, 1)$.*

This functional criterion induces a partial order over the set of all scoring functions. Let s_1 and s_2 be two scoring functions on $\mathcal{X}$, the ordering provided by s_1 is better than that induced by s_2 when

$$\forall \alpha \in (0,1), \;\; \mathrm{MV}_{s_1}(\alpha) \leq \mathrm{MV}_{s_2}(\alpha).$$

We point out that, in certain situations, some parts of the MV curve may be of interest solely, corresponding to large values of α when focus is on extremal observations and to small values of α when modes of the underlying distributions are investigated. For instance, the more concentrated around its modes X's distribution, the closer to the right lower corner of the MV space the MV curve.

The result below shows that optimal scoring functions are those whose MV curves are minimum everywhere.

Proposition 3. (Optimal MV curve) *Let assumptions $\mathbf{A_1}-\mathbf{A_2}$ be fulfilled. The elements of the class $\mathcal{S}^*$ have the same* MV *curve and provide the best possible ordering of $\mathcal{X}$'s elements in regard to the* MV *curve criterion:*

$$\forall (s, \alpha) \in \mathcal{S} \times (0,1), \;\; \mathrm{MV}^*(\alpha) \leq \mathrm{MV}_s(\alpha), \qquad (1)$$

where $\mathrm{MV}^*(\alpha) = \mathrm{MV}_f(\alpha)$ for all $\alpha \in (0,1)$.

In addition, we have: $\forall (s, \alpha) \in \mathcal{S} \times (0,1)$,

$$0 \leq \mathrm{MV}_s(\alpha) - \mathrm{MV}^*(\alpha) \leq \lambda\left(\Omega_\alpha^* \Delta \Omega_{s, Q(s,\alpha)}\right),$$

where Δ denotes the symmetric difference.

The proof is immediate based on the results recalled in subsection 2.2 and is left to the reader. Incidentally, notice that, equipped with the notations introduced in 2.2, $\lambda^*(\alpha) = \mathrm{MV}^*(\alpha)$ for all $\alpha \in (0,1)$.

Example 1. (Gaussian distribution) *In the case when X is a Gaussian variable $\mathcal{N}(0,1)$, we have $\mathrm{MV}^*(\alpha) = 2\Phi^{-1}((1+\alpha)/2)$, where $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} \exp(-u^2/2) du$.*

The following result reveals that the optimal MV curve is convex and provides a closed analytical form for its derivative.

Proposition 4. (Convexity and Derivative) *Suppose that hypotheses $\mathbf{A_1}-\mathbf{A_2}$ are satisfied. Then, the optimal curve $\alpha \in (0,1) \mapsto \mathrm{MV}^*(\alpha)$ is convex. In addition, if f is differentiable with a gradient taking nonzero values on the boundary $\partial\Omega_\alpha^* = \{x \in \mathcal{X} : f(x) = Q^*(\alpha)\}$, MV^* is differentiable at $\alpha \in [0,1[$ and:*

$$\mathrm{MV}^{*\prime}(\alpha) = \frac{1}{f(Q^*(\alpha))} \int_{x \in \partial\Omega_\alpha^*} \frac{1}{||\nabla f(x)||} d\mu(dx),$$

where μ denotes the Hausdorff measure on $\partial\Omega_\alpha^*$.

See the Appendix section for the technical proof. Elementary properties of MV curves are summarized in the following proposition.

Proposition 5. (Properties of MV curves) *for any $s \in \mathcal{S}$, the following assertions hold true.*

1. **Limit values.** *We have $\mathrm{MV}_s(0) = 0$ and $\lim_{\alpha \to +1} \mathrm{MV}_s(\alpha) = \lambda(suppF)$.*
2. **Invariance.** *For any strictly increasing function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, we have $\mathrm{MV}_s = \mathrm{MV}_{\psi \circ s}$.*
3. **Monotonicity.** *The mapping $\alpha \in (0,1) \mapsto \mathrm{MV}_s(\alpha)$ is strictly increasing on $(0, \bar{\alpha}_s)$.*
4. **Differentiability.** *Suppose that $s(X)$'s distribution is continuous and has no plateau. Then, if $t > 0 \mapsto \lambda_s(t)$ is differentiable and s is Lipschitz, then MV_s is differentiable on $(0,1)$ and:*

$$\mathrm{MV}_s'(\alpha) = -\frac{1}{\alpha_s'(\alpha_s^{-1}(\alpha))} \int_{s^{-1}(\{\alpha_s^{-1}(\alpha)\})} \frac{\mu(\mathrm{d}y)}{|\nabla f(y)|},$$

for all $\alpha \in (0,1)$, denoting by $\mu(dy)$ the Hausdorff measure on the boundary $s^{-1}(\{\alpha_s^{-1}(\alpha)\})$.

These assertions straightforwardly result from Definition 2. Except those related to the derivative formula, details are omitted.

3.3 Piecewise Constant Functions

We now focus on scoring functions of the simplest form. Let $K \geq 1$ and consider a partition $\mathcal{P}$ of the feature space $\mathcal{X}$ defined by K pairwise disjoint subsets $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of finite Lebesgue measure plus the subset $\mathcal{X} \setminus \cup_{k \leq K} \mathcal{C}_k$. When $suppF$ is compact, one may suppose $\mathcal{X}$ of finite Lebesgue measure and take $\mathcal{X} = \cup_{k \leq K} \mathcal{C}_k$. Then, for any permutation σ in the symmetric group $\mathfrak{S}_K$ of $\{1, \ldots, K\}$, define the piecewise constant scoring function given by: $\forall x \in \mathcal{X}$,

$$s_{\mathcal{P},\sigma}(x) = \sum_{k=1}^{K} (K - k + 1) \cdot \mathbb{I}\{x \in \mathcal{C}_{\sigma(k)}\}.$$

Its piecewise linear MV curve $(\alpha, \mathrm{MV}_{s_{\mathcal{P},\sigma}}(\alpha))$ connects the knots $(0,0) = (\alpha_{s_{\mathcal{P},\sigma}}(K+1), \lambda_{s_{\mathcal{P},\sigma}}(K+1)), \ldots, (\alpha_{s_{\mathcal{P},\sigma}}(1), \lambda_{s_{\mathcal{P},\sigma}}(1))$, where

$$\alpha_{s_{\mathcal{P},\sigma}}(k) = \sum_{j=1}^{k} F(\mathcal{C}_{\sigma(k)}) \text{ and } \beta_{s_{\mathcal{P},\sigma}}(k) = \sum_{j=1}^{k} \lambda(\mathcal{C}_{\sigma(k)}).$$

We also have $\mathrm{MV}_{s_{\mathcal{P},\sigma}}(\alpha) = +\infty$ for $\alpha > \sum_{k=1}^{K} F(\mathcal{C}_k)$. The following lemma describes the best scoring functions among those of the type defined above in the sense of the MV curve criterion. Its proof is omitted.

Lemma 6. (Optimality) *Let $\sigma^* \in \mathfrak{S}_K$ such that*

$$\frac{\lambda(\mathcal{C}_{\sigma^*(1)})}{F(\mathcal{C}_{\sigma^*(1)})} \leq \cdots \leq \frac{\lambda(\mathcal{C}_{\sigma^*(K)})}{F(\mathcal{C}_{\sigma^*(K)})}.$$

Then, for any $\sigma \in \mathfrak{S}_K$, we have:

$$\forall \alpha \in (0,1), \ \ \mathrm{MV}_{s_{\mathcal{P},\sigma^*}}(\alpha) \leq \mathrm{MV}_{s_{\mathcal{P},\sigma}}(\alpha).$$

We point out that σ^* corresponds to a permutation σ which makes $\mathrm{MV}_{\mathcal{P},\sigma}$ convex on $[0, F(\cup_{k \leq K} \mathcal{C}_k)]$.

Approximation of the optimal MV curve. Let $\epsilon \in (0,1)$ and $\Delta: \ \alpha_0 = 0 < \alpha_1 < \cdots < \alpha_K = 1 - \epsilon < \alpha_{K+1} = 1$ be a subdivision of the unit interval. The MV curve of the piecewise constant scoring function

$$s_{\Delta}^*(x) \ = \ \sum_{k=1}^{K} (K - k + 1) \cdot \mathbb{I}\{x \in \Omega_{\alpha_k}^* \setminus \Omega_{\alpha_{k-1}}^*\}$$

is the piecewise linear interpolant of MV^* on the interval $[0, 1-\epsilon]$ related to the meshgrid Δ. Hence, if MV^* is of class $\mathcal{C}^2$ on $[0, 1-\epsilon]$ and such that $\sup_{\alpha \in [0,1-\epsilon]} |\mathrm{MV}^{*\prime\prime}(\alpha)| \leq \kappa < +\infty$, we classically have the error estimate, see de Boor [2001]: $\forall \alpha \in [0, 1-\epsilon]$,

$$\mathrm{MV}_{s_{\Delta}^*}(\alpha) - \mathrm{MV}^*(\alpha) \leq \frac{1}{8} \kappa \cdot m_{\Delta}^2, \tag{2}$$

with $m_{\Delta} = \max_{0 \leq k \leq K} (\alpha_{i+1} - \alpha_i)^2$. When $\lambda(suppF) < +\infty$, this is still true for $\epsilon = 0$.

3.4 Empirical Performance

In practice, MV curves are unknown, just like X's probability distribution, and must be estimated based on the observed sample $X_1, \ldots, X_n$. Replacing the mass of level sets by its empirical counterpart in Definition 2 leads to define the notion of *empirical* MV *curve*. We set, for all $t \geq 0$,

$$\widehat{\alpha}_s(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{s(X_i) > t\}.$$

Notice that $\widehat{\alpha}_s$ takes its values in the set $\{k/n: \ k = 0, \ldots, n\}$ and set $\widehat{\widehat{\alpha}}_s = (1/n) \sum_{i=1}^{n} \mathbb{I}\{s(X_i) > 0\}$.

Definition 7. (Empirical MV curve) *Let $s \in \mathcal{S}$. By definition, the empirical* MV *curve of s is the graph of the (piecewise constant) function*

$$\widehat{\mathrm{MV}}_s : \alpha \in [0, \widehat{\widehat{\alpha}}_s] \mapsto \lambda_s \circ \widehat{\alpha}_s^{-1}(\alpha).$$

By convention, set $\widehat{\mathrm{MV}}_s(\alpha) = +\infty$ for $\alpha \in [\widehat{\widehat{\alpha}}_s, 1]$.

The empirical MV curve can be viewed as the restriction to $(0, \widehat{\widehat{\alpha}}_s)$ of the plot of a mapping denoted by $\widehat{\mathrm{MV}}_s: \ \alpha \in (0,1) \mapsto \widehat{\mathrm{MV}}_s(\alpha)$, where $\widehat{\mathrm{MV}}_s(\alpha) = +\infty$ on $(\widehat{\alpha}_s(s(X_{\sigma_s(n_s)}), 1)$.

Remark 1. (Alternative definition) *One could also consider, as statistical version of the* MV *curve, the broken line connecting the knots corresponding to jumps of the curve $\widehat{\mathrm{MV}}_s$. Its asymptotic properties can be straightforwardly derived from those of $\widehat{\mathrm{MV}}_s$, see Theorem 8 below.*

The theorem below reveals that, under mild assumptions, the empirical MV curve is a consistent and asymptotically Gaussian estimate of the MV curve, uniformly over any subinterval of $[0, 1[$. It involves the assumptions listed below. Let $\epsilon \in (0,1)$ be fixed.

$\mathbf{A_3}$ The r.v. $s(X)$ is bounded, *i.e.* $||s(X)||_{\infty} < +\infty$, and has a continuous distribution with differentiable density $f_s(t)$ such that:

$$\forall \alpha \in [0, 1-\epsilon], \ \ f_s(\alpha_s^{-1}(\alpha)) > 0$$

and, for some $\gamma > 0$,

$$\sup_{\alpha \in [0, 1-\epsilon]} \frac{d \ \log(f_s \circ \alpha_s)}{d\alpha}(\alpha) \leq \gamma < +\infty.$$

$\mathbf{A_4}$ The mapping λ_s is of class $\mathcal{C}^2$.

Theorem 8. *Let $0 < \epsilon \leq 1$ and $s \in \mathcal{S}$. Assume that Assumptions $\mathbf{A_3} - \mathbf{A_4}$ are fulfilled. The following assertions hold true.*

(i) (Consistency) *With probability one, we have uniformly over $[0, 1-\epsilon]$:*

$$\lim_{n \to +\infty} \widehat{\mathrm{MV}}_s(\alpha) = \mathrm{MV}_s(\alpha).$$

(ii) (Strong approximation) *There exists a sequence of Brownian bridges $(W^{(n)}(\alpha))_{\alpha \in (0,1)}$ such that we almost-surely have, uniformly over the compact interval $[0, 1-\epsilon]$: as $n \to \infty$,*

$$\sqrt{n} \left(\widehat{\mathrm{MV}}_s(\alpha) - \mathrm{MV}_s(\alpha) \right) = Z^{(n)}(\alpha) + o(v_n),$$

with, for all $n \geq 1$,

$$v_n = \frac{(\log \log n)^{\rho_1(\gamma)} \log^{\rho_2(\gamma)} n}{\sqrt{n}},$$

where

$$(\rho_1(\gamma), \rho_2(\gamma)) = \begin{cases} (0,1) & \text{if} \quad \gamma < 1 \\ (0,2) & \text{if} \quad \gamma = 1 \\ (\gamma, \gamma - 1 + \eta) & \text{if} \quad \gamma > 1 \end{cases},$$

the parameter $\eta > 0$ being arbitrary, and

$$Z^{(n)}(\alpha) = \frac{\lambda_s'(\alpha_s^{-1}(\alpha))}{f_s(\alpha_s^{-1}(\alpha))} W^{(n)}(\alpha), \text{ for } \alpha \in (0,1).$$

See the Appendix for the technical proof, which relies on standard strong approximation results for the quantile process, see Csorgo and Revesz [1981]. Assertion (ii) means that the fluctuation process $\{\sqrt{n}(\widehat{\mathrm{MV}}_s(\alpha) - \mathrm{MV}_s(\alpha))\}_{\alpha \in [0,1-\epsilon]}$ converges in the space of càd-làg function on $[0, 1-\epsilon]$ equipped with the sup norm, to the law of a Gaussian stochastic process $\{Z^{(1)}(\alpha)\}_{\alpha \in [0,1-\epsilon]}$.

Remark 2. (Asymptotic normality) *It results from Assertion (ii) in Theorem 8 that, for any $\alpha \in (0,1)$, the pointwise estimator $\widehat{\mathrm{MV}}_s(\alpha)$ is asymptotically Gaussian under Assumptions $\mathbf{A_3} - \mathbf{A_4}$. For all $\alpha \in (0,1)$, we have the convergence in distribution: $\sqrt{n}\{\widehat{\mathrm{MV}}_s(\alpha) - \mathrm{MV}_s(\alpha)\} \Rightarrow \mathcal{N}(0, \sigma_s^2)$, as $n \to +\infty$, with $\sigma_s^2 = \alpha(1-\alpha)(\lambda_s'(\alpha_s^{-1}(\alpha))/f_s(\alpha_s^{-1}(\alpha)))^2$.*

Remark 3. (Confidence regions) *In practice, it is difficult to construct confidence regions for* MV *curves from simulated brownian bridges, based on the approximation above. Following in the footsteps of Silverman and Young [1987], it is recommended instead to implement a smoothed bootstrap procedure. The asymptotic validity of such a resampling method immediately derives from the strong approximation result previously stated, by standard coupling arguments.*

4 A M-ESTIMATION APPROACH

The anomaly scoring problem consists in building a scoring function $s(x)$, based on the training set $X_1, \ldots, X_n$, such that MV_s is as close as possible to the optimum MV^*. Due to the functional nature of the criterion performance, there are many ways of measuring how close the MV curve of a scoring function candidate and the optimal one are. The L_p-distances, for $1 \leq p \leq +\infty$, provide a relevant collection of risk measures. Let $\epsilon \in (0,1)$ be fixed (take $\epsilon = 0$ if $\lambda(suppF) < +\infty$) and consider the losses related to the sup-norm and that related to the L_1-distance:

$$\begin{aligned} d_1(s,f) &= \int_0^{1-\epsilon} |\mathrm{MV}_s(\alpha) - \mathrm{MV}^*(\alpha)| d\alpha, \\ d_\infty(s,f) &= \sup_{\alpha \in [0,1-\epsilon]} \{\mathrm{MV}_s(\alpha) - \mathrm{MV}^*(\alpha)\}. \end{aligned}$$

Observe that, by virtue of Proposition 3, the "excess-risk" decomposition applies in the L_1 case and the learning problem can be directly tackled through standard M-estimation arguments:

$$d_1(s,f) = \int_0^{1-\epsilon} \mathrm{MV}_s(\alpha) d\alpha - \int_0^{1-\epsilon} \mathrm{MV}^*(\alpha) d\alpha.$$

Hence, possible learning techniques could be based on the minimization, over a set $\mathcal{S}_0 \subset \mathcal{S}$ of candidates, of empirical counterparts of the area under the MV curve, such as $\int_0^{1-\epsilon} \widehat{\mathrm{MV}}_s(\alpha) d\alpha$. In contrast, the approach cannot be straightforwardly extended to the sup-norm situation. A possible strategy would be to combine M-estimation with approximation methods, so as to "discretize" the optimization task. This would lead to replace the unknown curve MV^* by an approximant, a piecewise linear interpolant $\widetilde{\mathrm{MV}}^*$ related to a subdivision $\Delta : \ 0 < \alpha_1 < \cdots < \alpha_K = 1 - \epsilon$ say and decompose the L_∞-risk as

$$\begin{aligned} d_\infty(s,f) \leq & \sup_{\alpha \in [0,1-\epsilon]} \{\mathrm{MV}_{s_\Delta^*}(\alpha) - \mathrm{MV}^*(\alpha)\} + \\ & \sup_{\alpha \in [0,1-\epsilon]} \{\mathrm{MV}_{s_\Delta^*}(\alpha) - \mathrm{MV}_s(\alpha)\}, \end{aligned}$$

the first term on the right hand side of the bound above being viewed as the *bias* of the statistical method. If one restricts optimization to the set of piecewise constant scoring functions taking $K+1$ values, the problem thus boils down to recovering the bilevel sets $\mathcal{R}_k^* = \Omega_{\alpha_k}^* \setminus \Omega_{\alpha_{k-1}}^*$ for $k = 1, \ldots, K$. This simple observation paves the way for designing scoring strategies relying on the estimation of a finite number of minimum volume sets, just like the approach described in the next section.

5 A PARTITIONING ALGORITHM

Now that the anomaly scoring problem has been rigorously formulated, we propose a data-based partitioning method to solve it and establish learning rates for the latter. For simplicity, we assume that $suppF$ is a compact subset of $\mathbb{R}^d$, the unit cube $[0,1]^d$ say.

5.1 Histogram Scoring Rules

Suppose that we are given a partition of the space $\mathcal{X} = [0,1]^d$ formed of $(\Lambda = J^d)$ cubes of side length $1/J$, $J \geq 1$: $\mathcal{C}_1, \ldots, \mathcal{C}_\Lambda$. Consider a subdivision $\Delta : \alpha_0 = 0 < \alpha_1 < \cdots < \alpha_K < \alpha_{K+1} = 1$ of $[0,1]$. We set: $\widehat{\alpha}(\mathcal{C}) = (1/n) \sum_{i=1}^n \mathbb{I}\{X_i \in \mathcal{C}\}$ for any subset $\mathcal{C} \subset \mathcal{X}$. Let $\phi_n \in (0,1)$ be some tolerance parameter, the algorithm is implemented in three steps as follows.

Algorithm

1. Sort the cubes by increasing order of magnitude of the empirical mass:

$$\widehat{\alpha}\left(\mathcal{C}_{(1)}\right) \geq \cdots \geq \widehat{\alpha}\left(\mathcal{C}_{(\Lambda)}\right).$$

2. For $k = 1$ to $K-1$,
   (a) Compute
   $$j_k = \inf\left\{ j \geq 1 : \ \sum_{l=1}^{j} \widehat{\alpha}\left(\mathcal{C}_{(l)}\right) \geq \alpha_k - \phi_n \right\}$$
   (b) Bind the cubes together, so as to form
   $$\Omega_{J,k} = \bigcup_{j=1}^{j_k} \mathcal{C}_{(j)}.$$
3. Compute the piecewise constant scoring function:
   $$\widehat{s}_{J,\Delta}(x) = \sum_{k=1}^{K} (K-k+1) \cdot \mathbb{I}\{x \in \mathcal{R}_{J,k}\}, \quad (3)$$
   where $\mathcal{R}_{J,k} = \Omega_{J,k} \setminus \Omega_{J,k-1}$ for $1 \leq k \leq K$, with $\Omega_{J,0} = \emptyset$ by convention.

Incidentally, we point out that the empirical MV curve of the scoring function produced by the algorithm above is always convex, just like the target MV^* (see Proposition 4). Borrowing standard concepts of the *finite element method*, consider the "hat functions": $\psi_k(\cdot) = \psi(\cdot; (\alpha_{k-1}, \alpha_k)) - \psi(\cdot; (\alpha_k, \alpha_{k+1}))$, for $1 \leq k < K$, with $\psi(\alpha, (\alpha', \alpha'')) = (\alpha - \alpha')/(\alpha'' - \alpha') \cdot \mathbb{I}\{\alpha \in [\alpha', \alpha'']\}$ for $\alpha' < \alpha''$, and set $\psi_K = \psi(\cdot; (\alpha_K, 1))$. We may then write:

$$\widehat{\mathrm{MV}}_{\widehat{s}_{J,\Delta}}(\alpha) = \sum_{k=1}^{K} \frac{j_k}{J^d} \cdot \psi_k(\alpha).$$

One may also easily check that this curve is dominated everywhere on $[0,1]$ by the MV curve of any scoring function taking at most $K+1$ different values and constant on each cube $\mathcal{C}_{(j)}$: the function $\widehat{s}_{J,\Delta}(x)$ is thus the *empirical risk minimizer* over the corresponding class of scoring rules in the strong L_∞-sense. Of course, the success of such an approach crucially depends on the accuracy of the (linear) approximation scheme the algorithm tries to mimic and on the capacity of binded cubes of side length $1/J$ to approximate well bilevel sets of the underlying density $f(x)$.

5.2 Rate Bound Analysis

We now investigate to which extent the procedure described above yields a scoring rule whose MV curve is close to the optimal one, in the sup norm sense. As a first go, the theorem below describes the accuracy of empirical minimum volume sets, on which the scoring function (3) is based. Except it incorporates the impact of the bias, the result is a straightforward application of Theorem 3 in Scott and Nowak [2006].

Theorem 9. *Suppose that Assumptions* $\mathbf{A_1} - \mathbf{A_2}$ *are fulfilled. Let* $(\alpha, \delta) \in (0,1)^2$. *Fix* $J \geq 1$. *Assume that the boundary of* Ω_α^* *is of finite perimeter* $per(\partial\Omega_\alpha^*) < +\infty$. *Take* $\phi_n(\delta) = \sqrt{(2^{Jd}\log(2) + \log(2/\delta))/(2n)}$ *and* $j = \inf\{l \in \{1, \ldots, J\} : \ \sum_{l=1}^{j} \widehat{\alpha}(\mathcal{C}_{(l)}) \geq \alpha - \phi_n\}$ *and set* $\widehat{\Omega}_\alpha = \bigcup_{l \leq j} \mathcal{C}_{(l)}$. *Then, for some constant* $c < +\infty$, *with probability at least* $1-\delta$: $\forall n \geq 1$,

$$\mathbb{P}\{X \in \widehat{\Omega}_\alpha\} \geq \alpha - 2\phi_n \text{ and } \lambda(\widehat{\Omega}_\alpha) \leq \lambda^*(\alpha) + c\frac{per(\partial\Omega_\alpha^*)}{J^d}.$$

We point out that, if $f(x)$ is of bounded variation, $\partial\Omega_\alpha^*$ is classically of finite perimeter for all $\alpha \in (0,1)$, *cf* Evans and Gariepy [1992].

Theorem 10. *Suppose that Assumptions* $\mathbf{A_1} - \mathbf{A_2}$ *are satisfied. Assume that there exist finite constants* M, $0 < \theta_0 < \theta_1$, κ *and* C *such that, for all* $t \geq 0$, $per(\partial\Omega_{f,t}) \leq M < +\infty$ *and* MV^* *is of class* $\mathcal{C}^2$ *with* $\mathrm{MV}^{*\prime\prime} \leq \kappa$, $\theta_0 \leq \mathrm{MV}^{*\prime} \leq \theta_1$ *and* $\Delta = \Delta_n$ *is such that* $m_\Delta \leq C/K$. *Then, we have with probability at least* $1-\delta$: $\forall \alpha \in [0,1]$,

$$\mathrm{MV}_{\widehat{s}_{J,\Delta}}(\alpha) - \mathrm{MV}^*(\alpha) \leq c_1 \log(1/\delta) \times \left\{ \frac{1}{J^d} + \frac{1}{K^2} + \sqrt{\frac{2^{Jd} + \log K}{n}} \right\},$$

where c_1 *is a constant depending on* M, c, θ_0, θ_1 *and* κ *solely.*

Of course, much faster rates can be derived under more restrictive smoothness assumptions for the boundaries $\partial\Omega_t$, see Mammen and Tsybakov [1995]. Observe also that, as a byproduct, one gets an empirical estimate of the optimal MV curve, $\widehat{\mathrm{MV}}_{\widehat{s}_{J,\Delta}}$, whose sup norm distance to MV^* can be easily shown to be bounded as in the theorem above (up to a multiplicative constant).

Before summarizing our findings and sketching lines of further research, a few remarks are in order.

Remark 4. (Plug-in) *A possible strategy would be to estimate first the unknown density function* $f(x)$ *by means of (non-) parametric techniques and next use the resulting estimator as a scoring function. Beyond the computational difficulties one would be confronted to for large or even moderate values of the dimension, we point out that the goal pursued in this paper is by nature very different from density estimation: the local properties of the density on a given cell* $\mathcal{C}_j$ *are useless here, only the ordering of the cells is of importance.*

Remark 5. (Extensions) *In the procedure studied above, the subdivision* Δ *is known in advance. A natural extension of the method would be to consider a flexible grid of the unit interval, which could be possibly selected in an adaptive fashion, depending on the local properties of the curve* MV^*.

6 CONCLUSION

Motivated by a wide variety of applications, we have formulated the issue of learning how to rank observations in the same order as that induced by the density function, which we called *anomaly scoring* here. For this problem, much less ambitious than estimation of the local values taken by the density, a functional performance criterion, the MV curve namely, is proposed. Its statistical estimation has been investigated from an asymptotic perspective and we have provided a partition-based strategy to build a scoring function with statistical guarantees in terms of rate of convergence for the sup norm in the MV space. Its analysis suggests a number of novel and important issues for statistical learning, such as extension of the approach promoted here to the case where the support of the distribution of interest is of infinite Lebesgue measure.

APPENDIX - TECHNICAL PROOFS

Proof of Proposition 4

Assume that convexity does not hold for MV^*. It means that there exist α and ϵ in $(0,1)$ such that $\mathrm{MV}^*(\alpha)-\mathrm{MV}^*(\alpha-\epsilon) > \mathrm{MV}^*(\alpha+\epsilon)-\mathrm{MV}^*(\alpha)$. This is in contradiction with Proposition 3. Indeed, consider then the scoring function $\widetilde{f}(x)$ equal to $f(x)$ on $\Omega^*_{\alpha-\epsilon} \cup (\mathcal{X} \setminus \Omega^*_{\alpha+\epsilon})$, to

$$\frac{Q^*(\alpha+\epsilon)-Q^*(\alpha)}{Q^*(\alpha)-Q^*(\alpha-\epsilon)}f(x)+\frac{Q^*(\alpha-\epsilon)Q^*(\alpha+\epsilon)-Q^*(\alpha^2)}{Q^*(\alpha-\epsilon)-Q^*(\alpha)}$$

on $\Omega^*_\alpha \setminus \Omega^*_{\alpha-\epsilon}$ and to

$$\frac{Q^*(\alpha)-Q^*(\alpha-\epsilon)}{Q^*(\alpha+\epsilon)-Q^*(\alpha)}f(x)+\frac{Q^*(\alpha^2)-Q^*(\alpha-\epsilon)Q^*(\alpha+\epsilon)}{Q^*(\alpha)-Q^*(\alpha+\epsilon)}$$

on $\Omega^*_{\alpha+\epsilon} \setminus \Omega^*_\alpha$. One may then check easily that Eq. (1) is not fulfilled at α for $s=\widetilde{f}$. The formula for $\mathrm{MV}^{*\prime}$ is a particular case of that given in Proposition 5.

Proof of Proposition 5

Applying the co-area formula (see [Federer, 1969, p.249, th.3.2.12]) for Lipschitz functions to $g(x) = (1/|\nabla f(x)|)\mathbb{I}\{x : |\nabla f(x)| > 0, s(x) \geq t\}$ yields $\lambda_s(t) = \int_t^{+\infty} \mathrm{d}u \int_{s^{-1}(u)} (1/|\nabla f(y)|)\mu(\mathrm{d}y)$, and thus $\lambda_s'(t) = -\int_{s^{-1}(\{t\})} 1/|\nabla f(y)|\mu(dy)$. And the desired formula follows from the composite differentiation rule.

Proof of Theorem 8

By virtue of Theorem 3 in Csorgo and Revesz [1978], under Assumption $\mathbf{A_3}$, there exists a sequence of independent Brownian bridges $\{W^{(n)}(\alpha)\}_{\alpha\in(0,1)}$ such that, we a.s. have:

$$\sqrt{n}\left\{\widehat{\alpha}_s^{-1}(\alpha)-\alpha_s^{-1}(\alpha)\right\} = \frac{W^{(n)}(\alpha)}{f_s(\alpha_s^{-1}(\alpha))} + o\left(v_n\right),$$

uniformly over $[0,1-\epsilon]$, as $n\to+\infty$. Now the desired results can be immediately derived from the Law of Iterated Logarithm for the Brownian Bridge, combined with a Taylor expansion of λ_s.

Proof of Theorem 9

Observe that it results from Proposition 9.7 in Mallat [1990] that there exists a constant $c<+\infty$ such that:

$$\min_{\Omega\in\mathcal{G}_J,\ \mathbb{P}\{X\in\Omega\}\geq\alpha} \lambda(\Omega)-\lambda(\Omega^*_\alpha) \leq c\times per(\partial\Omega^*_\alpha)\times J^{-d},$$

where $\mathcal{G}_J$ denotes the collection of subsets obtained by union of cubes $\mathcal{C}_j$. When combined with Proposition 3 in Scott and Nowak [2006], this yields the result.

Proof of Theorem 10

For $1\leq k\leq K$, let $\widetilde{\alpha}_k = \mathbb{P}\{X\in\Omega_{J,k}\}$, $\widetilde{\alpha}_0=0$ and $\widetilde{\alpha}_{K+1}=1$ and consider the "hat functions" $\widetilde{\psi}_k(\cdot)=\psi(\cdot;(\widetilde{\alpha}_{k-1},\widetilde{\alpha}_k))-\psi(\cdot;(\widetilde{\alpha}_k,\widetilde{\alpha}_{k+1}))$ and $\widetilde{\psi}_{K+1}(\cdot)=\psi(\cdot,(\widetilde{\alpha}_K,1))$. We have $\mathrm{MV}_{\widehat{s}_{J,\Delta}}(\alpha) = \sum_{k=1}^K \lambda(\Omega_{J,k})\widetilde{\psi}_k(\alpha)$ and we may write:

$$\begin{aligned}\mathrm{MV}_{\widehat{s}_{J,\Delta}}(\alpha)-\mathrm{MV}^*(\alpha) &= \sum_{k=1}^{K+1}\mathrm{MV}^*(\widetilde{\alpha}_k)\widetilde{\psi}_k(\alpha)-\mathrm{MV}^*(\alpha)\\ &\quad+\sum_{k=1}^{K+1}\left(\lambda(\Omega_{J,k})-\mathrm{MV}^*(\widetilde{\alpha}_k)\right)\widetilde{\psi}_k(\alpha). \quad (4)\end{aligned}$$

Therefore, it results from Theorem 9 combined with the union bound that, with probability at least $1-\delta$, we have the following bounds: $\forall k\in\{1,\ldots,K\}$,

$$\widetilde{\alpha}_k \geq \alpha_k - 2\phi_n(\delta/K) \text{ and } \lambda(\Omega_{J,k}) \leq \mathrm{MV}^*(\alpha_k)+\frac{cM}{J^d},$$

which also implies that $\widetilde{\alpha}_k \leq \alpha_k + cM/(\theta_0 J^d)$. Hence, for $k\leq K$, we have $(\widetilde{\alpha}_{k+1}-\widetilde{\alpha}_k)^2 \leq 3((C/K)^2+2(\max\{2\phi_n(\delta/K),\ (cM)/(\theta_0J^d)\})^2)$ and thus $|\sum_{k=1}^{K+1}\mathrm{MV}^*(\widetilde{\alpha}_k)\widetilde{\psi}_k(\alpha)-\mathrm{MV}^*(\alpha)|$ is bounded by

$$\frac{\kappa}{8}\left(\left(\frac{C}{K}\right)^2+2\left(\max\left\{2\phi_n(\delta/K),\frac{cM}{\theta_0J^d}\right\}\right)^2\right), \quad (5)$$

using Eq. (2). In addition, for $1\leq k\leq K$, the quantity $|\lambda(\Omega_{J,k})-\mathrm{MV}^*(\widetilde{\alpha}_k)|$ is bounded by

$$\begin{aligned}\frac{cM}{J^d}+|\mathrm{MV}^*(\alpha_k)-\mathrm{MV}^*(\widetilde{\alpha}_k)| &\leq \frac{cM}{J^d}\\ &\quad+\theta_1\max\left\{2\phi_n(\delta/K),\ (cM)/(\theta_0J^d)\right\}, \quad (6)\end{aligned}$$

by virtue of the finite increment theorem. Now, bounds (4), (5) and (6) combined with the specific choices made for K and J yield the desired rate bound.

References

M. Csorgo and P. Revesz. Quantile process approximations. *The Annals of Statistics*, 6(4):882–894, 1978.

M. Csorgo and P. Revesz. *Strong Approximation Theorems in Probability and Statistics*. Academic Press, 1981.

C. de Boor. *A practical guide to splines*. Springer, 2001.

L.C. Evans and R.F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

H. Federer. *Geometric Measure Theory*. Springer, 1969.

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

S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1990.

E. Mammen and A. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. *The Annals of Statistics*, 23(2):502–524, 1995.

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

P. Rigollet and R. Vert. Fast rates for plug-in estimators of density level sets. *Bernoulli*, 14(4):1154–1178, 2009.

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

B. Silverman and G. Young. The bootstrap: to smooth or not to smooth. *Biometrika*, 7(4):469–479, 1987.

A.B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–960, 1997.

B.Y. Zuo and R. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000.