When the observations, related to the phenomenon to be analyzed, are univariate, measurements are generally considered as "abnormal" when they are either too high or else to small with respect to central measures such as the mean or the median. Anomaly detection then relies on tail analysis of the variable of interest. Heavy-tail modelling and extreme value theory then permit to reduce the problem to the statistical estimation of a few parameters (e.g. the shape parameter of the extremal distribution). The farther in the tails the observations, the likelier they are considered as anomalies. In practice, this boils down to comparing the observed numerical values to the quantiles of the ()semi-)parametric distribution previously fitted, just like in a standard testing procedure.

Extension to the multivariate setup is far from straightforward. There is no natural order on $\mathbb{R}^d$ when $d > 2$. In the multivariate framework, many anomaly/novelty detection techniques often correspond to ad-hoc procedures, consisting on the computation of a real-valued statistic, supposedly with a capacity to discriminate between the normal and abnormal regimes of the system under study, e.g. [**?**], and involving no statistical learning stage (i.e. no optimization of an empirical criterion, accounting for the performance of the possible functionals used for the discrimination task). Such approaches can also be connected with the literature dedicated to *statistical depth functions* in nonparametric statistics and operations research, see [9] 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.

Extreme value theory. In a multidimensional framework, extreme value analysis is much less easy than in dimension 1, simply because the class of (max-stable) distributions is of infinite dimension and the (nonparametric) estimation of the angular/spectral measure is a challenging problem. The corresponding framework is asymptotic and is related to extremal observations (possibly out-of-sample), see [4]. As shall be seen, the angle developed in this thesis is very different.

Minimum volume sets. The notion of minimum volume sets has been introduced in the seminal contribution [3] 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

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

where the minimum is taken over all measurable subsets Ω of 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 $X \setminus \Omega^*_\alpha$. In the continuous setting, it can be shown that there exists a threshold value $t^*_\alpha \overset{def}{=} Q(f, \alpha) \geq 0$ 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:

$$\forall \alpha \in (0,1), \quad \lambda^*(\alpha) \stackrel{def}{=} \lambda(\Omega^*_\alpha).$$

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 X, supposed rich enough to include all density level sets (or reasonable approximants of the latter). In [3], 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 [6], 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. Except histogram-based rules, dyadic tree partitioning techniques (see [6]) and SVM-one class ([8]), few algorithms, implementing an efficient and fast search of subset solutions to the constrained optimization problem mentioned above are documented in the machine-learning literature. In particular, no approach involving aggregation or randomisation have been proposed to tackle this problem, in high dimension in particular. An objective of the thesis is to develop novel practical methods to build flexible solutions, with the capacity to capture the geometry of level sets of densities in high dimension and propose theoretical concepts (complexity measures) to investigate their generalization ability.

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 [7] or [5] for instance. Such an approach however yields significant computational issues even for moderate values of the dimension, inherent to the curse of dimensionality phenomenon.

Anomaly Ranking/Scoring. In [2], a natural criterion to evaluate the accuracy of decision rules in regard to anomaly scoring has been proposed.

Definition 1. (TRUE MASS-VOLUME CURVE) *Let $s \in S$. 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*

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

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 X, the ordering provided by s_1 is better than that induced by s_2 when

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

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

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

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

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

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

where Δ denotes the symmetric difference.

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

Proposition 3. (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 $\partial\Omega^*_\alpha = \{x \in X : f(x) = Q^*(\alpha)\}$,* 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 $\partial\Omega^*_\alpha$.

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 2, 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 $S_0 \subset S$ 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

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

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 preliminary approaches described in [2] and [1]. It is the major purpose of this thesis to develop machine-learning algorithms in this framework. The PhD candidate will also provide a sound statistical theory, grounding the procedures proposed, as well as strong empirical evidence of their performance on artificial and real datasets. Applications of the methodology elaborated to aeronautics, retail banking and maritime traffic will be considered in particular.

References

[1] S. Clémençon. Mass volume curves and anomaly ranking. *Submitted for Publication*, 2013.

[2] S. Cl émençon and J. Jakubowicz. Scoring anomalies:a m-estimation approach. 2013.

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

[4] S. Resnick. *Heavy-Tail Phenomena*. Springer, New York, 2007.

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

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

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

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

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