Supplementary material - Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere

Albert Thomas, Stephan Clémençon, Alexandre Gramfort, Anne Sabourin

1 Proof of Theorem 1

Throughout the proof, we assume that the marginal c.d.f. F_j, $j = 1, \ldots, d$, are known. In other words, we analyze a surrogate of the spherical MV-set algorithm, where $\widehat{\mathbf{V}}$ may be taken as $\mathbf{V}$ itself. Recall the definition of the finite distance angular measure $\Phi_t(A) = t\mathbb{P}(\mathbf{V} \in t\mathcal{C}_A)$, $A \subset \mathbb{S}_{d-1}$, where $\mathcal{C}_A = \{tx, x \in A, t \geq 1\}$ is the truncated cone generated by A. Notice $\Phi(A) = \lim_{t\to\infty} \Phi_t(A)$ and the underlying regular variation assumption may be recast as $\mathbb{P}(r(\mathbf{V}) \geq t, \theta(\mathbf{V}) \in A) \approx t^{-1}\Phi(A) \approx t^{-1}\Phi_t(A)$ when t is large. Observe the following error decomposition

$$\sup_{B\in\mathcal{G}} \left|\widehat{\Phi}_{n,k} - \Phi\right|(B) \leq \sup_{B\in\mathcal{G}} \left|\widehat{\Phi}_{n,k}(B) - \Phi_{n/k}\right|(B) + \underbrace{\sup_{B\in\mathcal{G}} \left|\Phi_{n/k} - \Phi\right|(B)}_{\mathrm{bias}(n/k,F,\Phi)} . \quad (1)$$

In this paper we do not attempt to control the bias term. This may be done under additional assumptions usually referred to as 'second order assumptions' (see *e.g.* [1], Chapter 3). Our analysis focuses on the first term in the right-hand side of (1). The following result is crucial for our purposes.

Lemma 1 (a normalized VC-inequality for low probability, finite classes)**.** *Let $\mathcal{A}$ be a class of sets of finite cardinality $|\mathcal{A}|$ within a sample space E and let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be an i.i.d. sample of random E-valued variables distributed as $\mathbf{Y}$. Define the probability of hitting a member of the class, $p = \mathbb{P}(\mathbf{Y} \in \cup_{A\in\mathcal{A}} A)$. Consider the relative Rademacher average*

$$\mathcal{R}_{n,p} = \mathbb{E} \sup_{A\in\mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A} \right| .$$

Then $\mathcal{R}_{n,p}$ satisfies

$$\mathcal{R}_{n,p} \leq \sqrt{\frac{2\log(|\mathcal{A}|)}{np}} , \quad (2)$$

and for any $\delta \geq e^{-np}$, with probability $1-\delta$,

$$\frac{1}{p}\sup_{A\in\mathcal{A}}\left|\mathbb{P}(\mathbf{Y}\in A)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{Y}_i\in A}\right| \leq \quad 2\mathcal{R}_{n,p} \;+\; 3\sqrt{\frac{1}{np}\log\frac{1}{\delta}}$$
$$\leq \frac{1}{\sqrt{np}}\left[2\sqrt{2\log|\mathcal{A}|}+3\sqrt{\log(1/\delta)}\right]. \quad (3)$$

Proof. The bound (2) in the statement is a variation on the result established in [5], Lemma 14, which states that, for a VC class of dimension $V_{\mathcal{A}}$, $\mathcal{R}_{n,p} \leq C\sqrt{V_{\mathcal{A}}/(np)}$, where C is a universal constant. To obtain the stated inequality and remove the unknown constant, we replace the upper bound on the standard Rademacher average $\mathcal{R}_K \leq C\sqrt{V_{\mathcal{A}}/K}$ used in their proof by the usual bound (see *e.g.* [2]) combined with the fact that, for finite classes, the shattering coefficient satisfies $S_{\mathcal{A}} \leq |\mathcal{A}|$ (see [4], Theorem 13.6), *i.e.*

$$\mathcal{R}_K \leq \sqrt{\frac{2\log S_{\mathcal{A}}(K)}{K}} \leq \sqrt{\frac{2\log|\mathcal{A}|}{K}},$$

where K is the number of $\mathbf{Y}_i$ which hit the class $\mathcal{A}$. The rest of the proof of the above mentioned lemma is unchanged, which yields (2). The high probability bound (3) is then obtained as a direct consequence of Theorem 10 in the cited reference. □

Proposition 1. *Under the assumptions of Theorem 1, and with the notations of Lemma 1, with probability at least $1-\delta$,*

$$\sup_{B\in\mathcal{G}}\left|\Phi_{n/k}(B)-\widehat{\Phi}_{n,k}(B)\right| \leq \sqrt{\frac{d}{k}}\left[2\sqrt{2\log(2)dJ^{d-1}}+3\sqrt{\log(1/\delta)}\right]. \quad (4)$$

Proof. Consider the class of sets $\mathcal{A}=\{\frac{n}{k}\mathcal{C}_B, B\in\mathcal{G}\}$ in the sample space $E=\mathbb{R}_+^d$, and take $\mathbf{Y}, \mathbf{Y}_i = \mathbf{V}, \mathbf{V}_i, 1\leq i\leq n$ in Lemma 1. Then the probability of the class union is $p \leq dk/n$, and inequality (3) implies

$$\begin{aligned}\sup_{B\in\mathcal{G}}\left|\Phi_{n/k}(B)-\widehat{\Phi}_{n,k}(B)\right| &= \sup_{B\in\mathcal{G}}\left|\frac{n}{k}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{V}_i\in\frac{n}{k}\mathcal{C}_B}-\frac{n}{k}\mathbb{P}(\mathbf{V}\in\frac{n}{k}\mathcal{C}_B)\right| \\ &= \frac{n}{k}\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\mathbb{1}_{\mathbf{V}_i\in A}-\mathbb{P}(\mathbf{V}\in A)\right| \\ &\leq \frac{n}{k}\frac{p}{\sqrt{np}}\left[2\sqrt{2\log|\mathcal{A}|}+3\sqrt{\log(1/\delta)}\right] \\ &\leq \sqrt{\frac{d}{k}}\left[2\sqrt{2\log|\mathcal{A}|}+3\sqrt{\log(1/\delta)}\right]\end{aligned}$$

which is the desired inequality, considering that $|\mathcal{A}| = |\mathcal{G}| = 2^{dJ^{d-1}}$. □

Following the proof of theorem 3 in [6], we obtain the desired result.

2 Model selection

Let $\widehat{\Omega}_\alpha^J$ denote the solution of the problem (9) over $\mathcal{G} = \mathcal{G}_J$. One should pick the value

$$\widehat{J} = \underset{J_{\min} \leq J \leq J_{\max}}{\arg\min} \left\{ \lambda_d(\widehat{\Omega}_\alpha^J) + \psi_k(\delta, J) \right\}, \tag{5}$$

minimizing thus a complexity-penalized version of the volume, in order to estimate the MV-set over $\mathbf{G} = \cup_{J=1}^{J_{max}} \mathcal{G}_J$ as accurately as possible without overfitting the data. An oracle inequality showing that the chosen set $\widehat{\Omega}_\alpha^{\widehat{J}}$ corresponds to an optimal trade-off between excess volume and missing mass can be straightforwardly derived from the analysis carried out in subsection 3.2, just like in Theorem 11 in [6].

We first introduce the risk of a set Ω defined as

$$R(\Omega) = (\lambda_d(\Omega) - \lambda_d(\Omega_\alpha^*))_+ + \left(\alpha - \Phi_{n/k}(\Omega)\right)_+$$

where, for any $u \in \mathbb{R}^d$, we denote by $u_+ = \max(u, 0)$ its positive part. From Theorem 1, we have with probability at least $1 - \delta$,

$$R(\widehat{\Omega}_\alpha) \leq \left(\inf_{\Omega \in \mathcal{G}_\alpha} \lambda_d(\Omega) - \lambda_d(\Omega_\alpha^*) \right) + 2\psi_k(\delta)\,.$$

The first term can be viewed as the approximation error and the second term as a bound on the stochastic error. The goal of model selection through complexity penalization is to find the model $\mathcal{G}_J$ achieving the optimal trade-off between these two errors.

As in [6], we thus consider the solution $\widehat{\Omega}_\alpha$ of the following optimization problem that can be viewed as the structural risk minimization version of (7)

$$\widehat{\Omega}_\alpha \in \underset{\Omega \in \mathbf{G}}{\arg\min} \left\{ \lambda_d(\Omega) + 2\psi_k(2^{-J}\delta, J), \widehat{\Phi}_{n,k}(\Omega) \geq \alpha - \psi_k(2^{-J}\delta, J) \right\} \tag{6}$$

where for each $J \in \{1, \ldots, J_{max}\}$ and each $\delta \in (0, 1)$, $\psi_k(\delta, J)$ denotes a penalty for the model $\mathcal{G}_J$.

Following the steps of the proof of Theorem 11 in [6] we have the following oracle inequality.

Theorem 1. *Assume that assumptions* $\mathbf{A_1} - \mathbf{A_2}$ *are fulfilled by the angular probability measure* $\Phi_{n/k}$. *Let* $\delta \in (0, 1)$ *and* $\mathcal{G}_{J,\alpha} = \mathcal{G}_J \cap \mathcal{G}_\alpha$. *With probability at least* $1 - \delta$, *we have*

$$R(\widehat{\Omega}_\alpha) \leq \left(1 + \frac{1}{K^{-1}_{\Phi_{n/k}}(1-\alpha)} \right) \inf_{1 \leq J \leq J_{max}} \min_{\Omega \in \mathcal{G}_{J,\alpha}} \left\{ \lambda_d(\Omega) - \lambda_d(\Omega_\alpha^*) + 2\psi_k(2^{-J}\delta, J) \right\}.$$

3 On the bias of the MV-set estimation procedure

As the class $\mathcal{G}_J$ contains sets that consists in union of hypercubes, the box-counting class of angular distributions (see [6]), which contains all angular distributions such that the boundary $\partial\Omega_\alpha^*$ has Lipschitz regularity, seems appropriate to control the bias introduced by the MV-set estimation procedure. Let $c_1 > 0$ and $N_J(\partial\Omega_\alpha^*)$ denote the number of hypercubes that intersect $\partial\Omega_\alpha^*$. The box counting class $\mathcal{D}_{box}$ is the set of all distributions such that

$\mathbf{A_3}$ $N_J(\partial\Omega^*_\alpha) \leq c_1 J^{d-2}$.

As explained in [7], assumption $\mathbf{A_3}$ is more convenient than only assuming that $\partial\Omega^*_\alpha$ has Lipschitz boundary. It allows boundaries with arbitrary orientation and multiple connected components whereas the Lipschitz boundary assumption restricts the boundary to have a functional form. For further discussion on the connection between assumption $\mathbf{A_3}$ and Lipschitz regularity of the boundary $\partial\Omega^*_\alpha$, see Lemma 3 in [7].

Remark 1. *Other approaches are possible to control the bias of the MV-set estimation procedure. For instance, one can assume that $\partial\Omega^*_\alpha$ has a finite perimeter [3].*

Let $J \in \{1, \ldots, J_{max}\}$, $\delta \in (0,1)$ and Ω^J_α denote a solution of $\min_{\Omega \in \mathcal{G}_{J,\alpha}} \lambda_d(\Omega)$. If the angular distribution $\Phi_{n/k} \in \mathcal{D}_{box}$ then,

$$\min_{\Omega \in \mathcal{G}_{J,\alpha}} \lambda_d(\Omega) - \lambda_d(\Omega^*_\alpha) \leq \lambda_d(\Omega^J_\alpha \Delta \Omega^*_\alpha) \leq N_J(\partial\Omega^*_\alpha)\frac{1}{J}^{d-1} \leq \frac{c_1}{J}$$

where the second inequality comes from the fact that the two sets can only differs on the hypercubes that intersect $\partial\Omega^*_\alpha$.

4 Setting the tolerance parameter to 0 for the numerical experiments

The tolerance parameter $\psi_k(\delta)$ is set to 0 in the numerical experiments of the paper. As remarked by one of the reviewers this loses the connection between Theorem 1 and the experiments. However by corollary 12 in [6], one can solve the empirical minimum volume set optimization problem without the tolerance parameter in the mass constraint and obtain a theoretical result similar to the one of Theorem 1.

References

[1] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics, 2005.

[2] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[3] S. Clémençon and J. Jakubowicz. Scoring anomalies: a M-estimation formulation. In *Proceedings of the 16-th International Conference on Artificial Intelligence and Statistics, Scottsdale, USA*, 2013.

[4] L Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[5] N. Goix, A. Sabourin, and S. Clémençon. Learning the dependence structure of rare events: a nonasymptotic study. In *Proceedings of the International Conference on Learning Theory, COLT'15*, 2015.

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

[7] C. Scott and R. D. Nowak. Minimax-Optimal Classification With Dyadic Decision Trees. *Information Theory, IEEE Transactions on*, 52(4):1335–1353, 2006.