Learning Hyperparameters for Unsupervised Anomaly Detection

Albert Thomas[1,2] ALBERT.THOMAS@TELECOM-PARISTECH.FR
Stéphan Clémençon[1] STEPHAN.CLEMENCON@TELECOM-PARISTECH.FR
Vincent Feuillard[2] VINCENT.FEUILLARD@AIRBUS.COM
Alexandre Gramfort[1] ALEXANDRE.GRAMFORT@TELECOM-PARISTECH.FR
[1]LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
[2]Airbus Group Innovations, 12 rue Pasteur, 92150, Suresnes, France

Abstract

While all unsupervised anomaly detection algorithms require the specification of parameters, no labelled data is available to assess their performance, making parameter tuning a challenge. This work presents a two step solution to this issue. The first step consists in using empirical estimates of Mass Volume (MV) curves. The area under the MV curves is computed on left-out data to select the best hyperparameters while preventing overfitting. The second step uses the concept of model ensembling. Using several random splits of the data, the variance of the estimates is reduced by averaging the solutions obtained for each split. Results demonstrate that the proposed approach offers a viable solution to the important practical problem of model selection, and that it allows to boost performance of anomaly detection algorithms on both simulated and real data sets.

1. Introduction

An anomaly, also referred to as an outlier or a novelty, corresponds to "an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980). In most applications, these observations are either to be removed or to be further investigated as potentially holding useful information. Unsupervised anomaly detection aims at finding such observations in an unlabelled data set. Applications include data cleaning, machine fault detection, network intrusion detection, fraud detection and medical diagnosis (Lin et al., 2005; Fujimaki et al., 2005).

Presented at ICML2016 Anomaly Detection Workshop, New York, NY, USA, 2016.

From a statistical point of view, one considers n observations $X_1, \ldots, X_n \in \mathbb{R}^d$, $d \geq 1$, assumed to be i.i.d. realizations of an unknown probability distribution P. The anomalies are the rare events, located in the low density regions such as the tail of the distribution. Anomaly detection consists therefore in estimating a high density region. The problem is to estimate a minimum volume set (Einmahl & Mason, 1992; Polonik, 1997).

Observations located in the minimum volume set will be considered as normal and observations located outside as anomalies. Given $\alpha \in (0, 1)$, a minimum volume set is a solution of the following optimization problem:

$$\min_{\Omega \in \mathcal{B}(\mathbb{R}^d)} \lambda(\Omega) \quad \text{such that } P(\Omega) \geq \alpha \ , \tag{1}$$

where $\mathcal{B}(\mathbb{R}^d)$ is the set of all measurable subsets of $\mathbb{R}^d$ and λ the Lebesgue measure on $\mathbb{R}^d$. Assuming that P has a density h with respect to the Lebesgue measure, one can show (Polonik, 1997; Nunez-Garcia et al., 2003) that if h is bounded and has no flat parts (i.e., for all $\rho > 0$, $\lambda(\{x, h(x) = \rho\}) = 0$), then the optimization problem (1) has a unique solution Ω_α^* such that $P(\Omega_\alpha^*) = \alpha$. Furthermore there exists $\rho_\alpha > 0$ such that $\Omega_\alpha^* = \{x, h(x) \geq \rho_\alpha\}$.

A natural approach to estimate a minimum volume set is therefore to resort to a so-called *plug-in approach* where one first estimates the density $h(x)$ and then thresholds it at an offset ρ such that the estimated set has an empirical mass α. Other approaches include partition based methods (Scott & Nowak, 2006), kernel based algorithms (One-Class SVM (OCSVM) (Schölkopf et al., 2001) and Support Vector Data Description (SVDD) (Tax & Duin, 2004)) and K-NN based methods (Sricharan & Hero, 2011; Qian & Saligrama, 2012).

Results show that all aforementioned algorithms can perform well in some settings as long as their hyperparameters are well tuned. Hyperparameter tuning is therefore a challenge of critical importance to increase the impact of unsupervised anomaly detection methods in various appli-

cations. The challenge originates from the fact that one cannot resort to common metrics used in supervised learning. This model selection issue is rarely discussed in papers while representing the major limitation of such approaches in a real life context and industrial applications.

In this paper, the aim is to propose a strategy to learn the hyperparameters of any unsupervised anomaly detection methods in order to optimize their performance and increase their impact on applications. This work develops our previous work on the calibration of the OSCVM for minimum volume set estimation (Thomas et al., 2015).

2. Preliminaries and Mass Volume Curve

Unsupervised anomaly detection algorithms based on minimum volume set estimation can be viewed as follows. A *scoring function* $\hat{s} : x \in \mathbb{R}^d \mapsto \mathbb{R}$ is learnt on the data set such that the smaller $\hat{s}(x)$ the more abnormal is the observation x. This scoring function is then thresholded at an offset ρ such that the set $\widehat{\Omega}_\alpha = \{x, \hat{s}(x) \geq \rho\}$ is an estimation of the minimum volume set Ω^*_α. The offset ρ is either returned by the algorithm (e.g. for the OCSVM), or computed such that $P_n(\hat{s}(X) \geq \rho) = \alpha$, where for all measurable set Ω, $P_n(\Omega) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_\Omega(X_i)$ denotes the empirical probability distribution based on the samples $X_1, \ldots, X_n$.

We assume that we are given a generic anomaly detection algorithm $\mathcal{A}$ that performs the following task. Given a hyperparameter $\theta \in \Theta$ and a data set $S_n = \{X_1, \ldots, X_n\}$ the anomaly detection algorithm $\mathcal{A}$ learns a scoring function $\hat{s}_\theta$:

$$\begin{array}{rccc} \mathcal{A}: & \mathcal{S}_n \times \Theta & \to & \mathbb{R}^{\mathbb{R}^d} \\ & (S_n, \theta) & \mapsto & \hat{s}_\theta \ . \end{array}$$

To assess the performance of the scoring function $\hat{s}_\theta$ we use the Mass Volume (MV) curve which was rigorously introduced by Clémençon & Jakubowicz (2013) and studied as a criterion for M-estimation of scoring functions.

Definition 1 (Mass Volume curve (Clémençon & Jakubowicz, 2013))**.** *The Mass Volume curve MV_s of a scoring function $s : \mathbb{R}^d \to \mathbb{R}$ is defined as the parametric curve $t \in \mathbb{R} \mapsto (\alpha_s(t), \lambda_s(t))$ where $\alpha_s(t) = \mathbb{P}(s(X) \geq t)$ and $\lambda_s(t) = \lambda(\{x, s(x) \geq t\})$. If α_s has no flat parts, the Mass Volume curve MV_s can also be defined as the function*

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

where $\alpha_s^{-1}(\alpha) = \inf\{t \in \mathbb{R}, \alpha_s(t) \leq \alpha\}$. By convention, points of the curve corresponding to possible jumps are connected by line segments.

As shown in (Clémençon & Jakubowicz, 2013), if $\mathcal{S}^*$ denotes the set of scoring functions that are increasing transformations of the density h, then for all scoring function s and for all $s^* \in \mathcal{S}^*$,

$$\forall \alpha \in [0,1), \quad \mathrm{MV}_{s^*}(\alpha) \leq \mathrm{MV}_s(\alpha) \ . \tag{2}$$

Denoting MV^* the MV curve of the density h, the closer is MV_s to MV^*, the better is the scoring function s.

As the probability distribution P is unknown, Clémençon & Jakubowicz (2013) also define the empirical MV curve of a scoring function. Given some observations $X_1, \ldots, X_n$, the empirical MV curve of a scoring function s is defined as

$$\widehat{\mathrm{MV}_s} : \alpha \in [0,1) \mapsto \lambda_s(\widehat{\alpha}_s^{-1}(\alpha)) \tag{3}$$

where $\widehat{\alpha}_s(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{x, s(x) \geq t\}}(X_i)$.

To select the parameter θ and prevent overfitting we consider the approach described in the following section.

3. Algorithm

3.1. Selection of the hyperparameters

To assess the performance of the scoring function $\hat{s}_\theta$ over the interval of probabilities $I = [\alpha_1, \alpha_2]$ we use the L_1 distance between the optimal MV curve MV^* and $\mathrm{MV}_{\hat{s}_\theta}$ over the interval I : $\|\mathrm{MV}_{\hat{s}_\theta} - \mathrm{MV}^*\|_{L_1} = \int_I |\mathrm{MV}_{\hat{s}_\theta}(\alpha) - \mathrm{MV}^*(\alpha)| d\alpha$. We thus want to choose the parameter θ^* minimizing this distance:

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} \|\mathrm{MV}_{\hat{s}_\theta} - \mathrm{MV}^*\|_{L_1} = \operatorname*{argmin}_{\theta \in \Theta} \|\mathrm{MV}_{\hat{s}_\theta}\|_{L_1} \ , \tag{4}$$

where the second equality comes from (2). This is equivalent to choosing the parameter θ^* minimizing the area under $\mathrm{MV}_{\hat{s}_\theta}$ over I. We use the notation $\mathrm{AMV}_I(s)$ to denote the area under the MV curve of a scoring function s over an interval $I = [\alpha_1, \alpha_2]$:

$$\mathrm{AMV}_I(s) = \int_{\alpha_1}^{\alpha_2} \mathrm{MV}_s(\alpha) d\alpha \ .$$

As the true MV curve is unknown, we consider the empirical MV curve (3). To prevent overfitting we use the following approach. The available data set is split into a training set S_{train} and a test set S_{test}. For $\theta \in \Theta$, the scoring function $\hat{s}_\theta$ is estimated on S_{train} thanks to the anomaly detection algorithm $\mathcal{A}$, and its performance is assessed using the area under the empirical MV curve on $[\alpha_1, \alpha_2]$ estimated from S_{test}.

Let $\mathrm{MV}^{test}_{\hat{s}_\theta}$ denote the empirical Mass Volume curve estimated on S_{test} using the scoring function $\hat{s}_\theta$ estimated on S_{train}. Instead of (4) we thus choose the hyperparameter $\theta^\dagger$ minimizing $\mathrm{AMV}^{test}_I(\hat{s}_\theta)$ over Θ, where $\mathrm{AMV}^{test}_I(\hat{s}_\theta) = \int_{\alpha_1}^{\alpha_2} \mathrm{MV}^{test}_{\hat{s}_\theta}(\alpha) d\alpha$. This leads to the hyperparameter selection method presented in Algorithm 1.

Algorithm 1 Automatic hyperparameters selection

Inputs: algorithm $\mathcal{A}$, data set S_n, hyperparameters subspace Θ, interval $[\alpha_1, \alpha_2]$, discretization parameter n_α
Randomly split S_n into a training set S_{train} and a test set S_{test}
for each hyperparameter $\theta \in \Theta$ **do**
  $\hat{s}_\theta = \mathcal{A}(S_{train}, \theta)$
  for β in $\left\{\alpha_1 + j\frac{\alpha_2-\alpha_1}{n_\alpha-1}, j \in \{0, \dots, n_\alpha - 1\}\right\}$ **do**
    Compute ρ_β^{test}, the empirical $1-\beta$ quantile of $\hat{s}_\theta$ from S_{test}, and $\lambda_{\hat{s}_\theta}(\rho_\beta^{test})$
  end for
  Compute $\mathrm{AMV}_I^{test}(\hat{s}_\theta)$
end for
Return $\theta^\dagger = \mathrm{argmin}_{\theta\in\Theta} \mathrm{AMV}_I^{test}(\hat{s}_\theta)$ and $\hat{s}_{\theta^\dagger}$

3.2. Volume computation

Running Algorithm 1 requires computing $\lambda(\{x, \hat{s}_\theta(x) \geq \rho_\beta^{test}\})$. Computing the volume of general sets in a high dimensional space is known to be a hard task. This problem of integral estimation by sampling is an active topic of research in the MCMC literature (Lovász & Vempala, 2006). We resort to Monte Carlo integration from a generation of uniform samples in the hypercube enclosing the data and we limit our experiments to small dimensions. Improvement on the computation of the volumes of minimum volume sets in order to scale better with the dimension will be part of future work.

3.3. Aggregation

The algorithm presented in the previous subsection to select the best hyperparameter depends on the random split of the data set into a training set and a test set. For small data set, this leads to instability as for each different random split, a different hyperparameter is selected and hence a different scoring function $\hat{s}_{\theta^\dagger}$ is obtained.

Taking advantage of ensemble methods known to improve the performance of unstable estimators (Dietterich, 2000), we consider B random splits of the data set into a training set and a test set. For each random split $b \in \{1, \dots, B\}$ the best hyperparameter $\theta_b^\dagger \in \Theta$ is selected. The final scoring function $\hat{S}$ is the average of all the scoring functions obtained during the procedure: $\widehat{S}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{s}_{\theta_b^\dagger}^b(x)$.

4. Experiments

In this section we compare the performance of multiple anomaly detection algorithms using either a priori fixed hyperparameters, or hyperparameters learnt using the approach developed above. When using fixed parameters we rely on paramater values used by the original authors or on theoretical results. We assess the performance of the algorithms without using a labelled data set. We are in the unsupervised anomaly detection setting quite common in industrial applications.

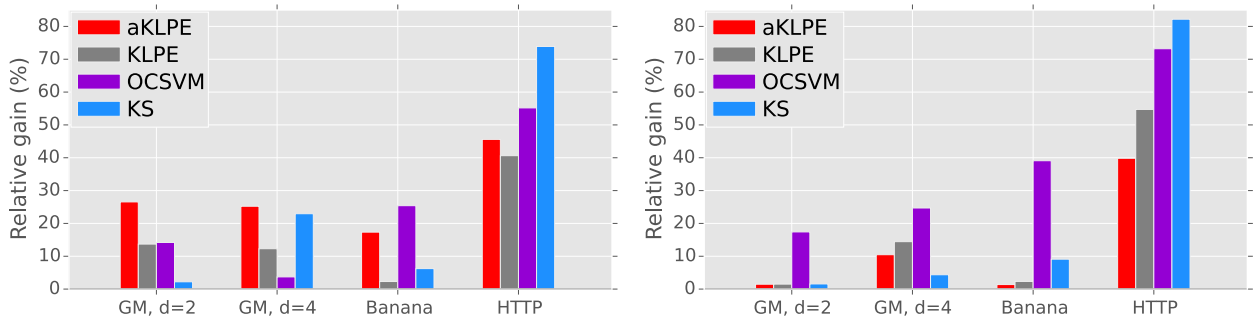
Table 1. Data sets name, the dimension d, the class used and the number of samples n_{perf} used for performance evaluation. NA: Not Applicable.

DATA SET	d	CLASS	n_{perf}
GAUSSIAN MIXTURE	2	NA	10,000
GAUSSIAN MIXTURE	4	NA	10,000
HTTP	3	NORMAL	56,016
BANANA	2	$+1$	1,876

We consider the scoring functions given by several anomaly detection algorithms: kernel smoothing, OCSVM and K-NN based anomaly detection. While the anomaly detection algorithm Isolation Forest (Liu et al., 2008) also returns a scoring function, to our knowledge there is no reference in the literature that makes a connection with minimum volume set estimation (see supplementary materials for empirical results including Isolation Forest). For kernel smoothing and OCSVM we use the implementation of Scikit-Learn (Pedregosa et al., 2011). The OCSVM uses the underlying LIBSVM library (Chang & Lin, 2011). The K-NN based anomaly detection methods were reimplemented using the efficient implementation of K-NN provided by Scikit-Learn.

For our approach 80% of the data set is used as the training set and the other 20% as the test set. For the paramater selection step, the area under the MV curve is computed over the interval of probabilities $[0.7, 0.99]$ and we use $n_\alpha = 20$. To compute the volumes with Monte Carlo integration we draw 10,000 samples from the uniform distribution over the hypercube enclosing the data. For all the experiments we choose $B = 50$ as it turned out to be large enough to stabilize all estimates.

The experiments are run on four different datasets, two synthetic and two standard datasets previously used in the literature in the context of anomaly scoring (Root et al., 2015) (see Table 1). The Banana and the HTTP data sets are standardized, i.e., component wise center and scale to unit variance. The scoring functions are estimated using n samples. For the synthetic data, we consider two Gaussian mixtures (GM) with distribution of the form $w_1\mathcal{N}(\mu_1, \Sigma_1) + w_2\mathcal{N}(\mu_2, \Sigma_2)$. The first Gaussian mixture is a two-dimensional Gaussian mixture with $w_1 = 0.2$, $w_2 = 0.8$, $\mu_1 = (5, 0)^T$, $\mu_2 = (-5, 0)^T$ and Σ_1 and Σ_2 diagonal matrices with diagonal $(1, 9)$ and $(9, 1)$ respectively. The second Gaussian mixture is a four-dimensional Gaussian mixture with the same weights values w_1 and w_2



(a) $n = 100$ (b) $n = 500$

Figure 1. Relative gain of our approach compared to a priori fixed hyperparameters

and $\mu_1 = (5, 0, 0, 0)^T$, $\mu_2 = (-5, 0, 0, 0)^T$ and Σ_1 and Σ_2 diagonal matrices with diagonal $(1, 1, 9, 9)$ and $(9, 9, 1, 1)$ respectively.

4.1. Scoring functions and hyperparameters

Kernel smoothing (KS). An estimation $\hat{h}$ of the true underlying density h can be seen as a scoring function. Here we consider a multivariate kernel density estimator where the kernel is a Gaussian kernel defined for $\sigma > 0$ by $k_\sigma(x, x') = \exp\left(-\frac{\|x-x\prime\|^2}{2\sigma^2}\right)$. The hyperparameter to tune is the kernel bandwidth σ. We consider 50 values of σ equally spaced between 0.01 and 5. We compare our approach with the Scott's rule (Scott, 1992) for the bandwidth selection.

OCSVM. We here consider the OCSVM applied with the Gaussian kernel k_σ. The scoring function $\hat{s}$ obtained is given by $\hat{s}(x) = \frac{1}{|S_{train}|}\sum_{x_i \in S_{train}} \gamma_i k_\sigma(x, x_i)$, where $0 \leq \gamma_i \leq \frac{1}{\nu|S_{train}|}$, ν being a parameter specified by the user. The two hyperparameters that need to be learnt are the Gaussian kernel bandwidth σ and $\nu \in (0, 1]$. Here we fix $\nu = 0.4$ and only tuned the kernel bandwidth (see (Thomas et al., 2015) for the reason behind the value of ν). We consider 50 values of σ equally spaced between 0.01 and 5. We compare our approach with the approach suggested in (Muñoz & Moguerza, 2004) that consists in choosing σ such that $2\sigma^2 = \max_{1 \leq i,j \leq p} \|x_i - x_j\|^2$.

K-NN based anomaly detection (KLPE and aKLPE). In our experiments we consider two versions of anomaly detection based on K-NN, $K \in \mathbb{N}^*$. Let $\Delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ be a distance function on $\mathbb{R}^d$. We denote by $x_{i_1}, \ldots, x_{i_K}$, the K closest points to $x_i \in \mathbb{R}^d$ in $S_{train} \backslash \{x_i\}$ with respect to the distance Δ.

In the first version, known as KLPE (Sricharan & Hero, 2011), the scoring function $\hat{s}_1(x)$ is given for all $x_i \in \mathbb{R}^d$ by $\hat{s}_1(x_i) = -\max_{j \in \{i_1, \ldots, i_K\}} \Delta(x_i, x_j)$. In the second version, known as aKLPE (Qian & Saligrama, 2012) the scoring function $\hat{s}_2(x)$ is given for all $x_i \in \mathbb{R}^d$ by $\hat{s}_2(x_i) = -\frac{1}{K}\sum_{j \in \{i_1, \ldots, i_K\}} \Delta(x_i, x_j)$.

The minus sign in the formula of $\hat{s}_1$ and $\hat{s}_2$ is here to respect the convention that the scoring function at the point x should be small for abnormal obversations. The parameter to tune is the number of neighbors K. We consider values of K in $\{2K + 1, 1 \leq K \leq 50\}$. Only odd values are considered to speed up the tuning. Following (Zhao & Saligrama, 2009) we chose in our comparisons $K = n^{0.4}$ for KLPE and following (Root et al., 2015) $K = 20$ for aKLPE.

4.2. Results

To evaluate the benefit of our approach compared to not using it, we consider a relative gain performance criterion G. The relative gain performance $G_{\mathcal{A}}(s_t, s_f)$ for an anomaly detection algorithm $\mathcal{A}$ giving a scoring function s_t with our approach and a scoring function s_f with a priori fixed hyperparameters is defined as

$$G_{\mathcal{A}}(s_t, s_f) = \frac{\text{AMV}_I(s_f) - \text{AMV}_I(s_t)}{\text{AMV}_I(s_f)}$$

where $I = [0.9, 0.99]$ as we are interested in the performance of the scoring functions in the low density regions.

The MV curves are computed on left out data and the volumes are estimated using 100,000 samples drawn from the uniform distribution over the hypercube enclosing the data. If $G_{\mathcal{A}} > 0$, then the scoring function s_t performs better in terms of area under the MV curves. The value of $G_{\mathcal{A}}$ gives us a quantified measure of how much it performs better. Reasons for using $G_{\mathcal{A}}$ over the ROC-AUC are given in supplementary materials.

Figure 1 shows the values of $G_{\mathcal{A}}$ for each data set and each algorithm $\mathcal{A}$ (see supplementary materials for more empirical results). Our approach always performs better than the ones with a priori fixed hyperparameters.

References

Chang, C-C. and Lin, C-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Clémençon, S. and Jakubowicz, J. Scoring anomalies: a M-estimation formulation. In *AISTATS '13: Sixteenth international conference on Artificial Intelligence and Statistics*, volume 31, pp. 659–667, 2013.

Dietterich, T. G. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pp. 1–15. Springer-Verlag, 2000.

Einmahl, J. H. J. and Mason, D. M. Generalized quantile processes. *The Annals of Statistics*, 20:1062–1078, 1992.

Fujimaki, R., Yairi, T., and Machida, K. An approach to spacecraft anomaly detection problem using kernel feature space. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pp. 401–410, New York, NY, USA, 2005. ACM.

Hawkins, D. M. *Identification of outliers*. Monographs on applied probability and statistics. Chapman and Hall, London, 1980.

Lin, J., Keogh, E., Fu, A., and Van Herle, H. Approximations to magic: finding unusual medical time series. In *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on*, pp. 329–334, 2005.

Liu, F.T., Ting, K.M., and Zhou, Z-H. Isolation forest. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pp. 413–422, 2008.

Lovász, L. and Vempala, S. Simulated annealing in convex bodies and an $O(n^4)$ volume algorithm. *Journal of Computer and System Sciences*, 72(2):392 – 417, 2006.

Muñoz, A. and Moguerza, J. M. One-class support vector machines and density estimation: The precise relation. In *Progress in Pattern Recognition, Image Analysis and Applications: 9th Iberoamerican Congress on Pattern Recognition, CIARP 2004*, pp. 216–223. Springer Berlin Heidelberg, 2004.

Nunez-Garcia, J., Kutalik, Z., Cho, K-H., and Wolkenhauer, O. Level sets and minimum volume sets of probability density functions. *Approximate reasoning*, 34: 25–47, 2003.

Pedregosa, F., Varoquaux, G, Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dufour, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

Qian, J. and Saligrama, V. New statistic in p-value estimation for anomaly detection. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pp. 393–396, 2012.

Root, J., Qian, J., and Saligrama, V. Learning efficient anomaly detectors from K-NN graphs. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015*, 2015.

Schölkopf, B., Platt, J., Shawe-Taylor, J. Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 2001.

Scott, C. D. and Nowak, R. D. Learning minimum volume sets. *Journal of Machine Learning Research*, 7:665–704, 2006.

Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

Sricharan, K. and Hero, A. O. Efficient anomaly detection using bipartite k-nn graphs. In *Advances in Neural Information Processing Systems 24*, pp. 478–486. 2011.

Tax, D. M. J. and Duin, R. P. W. Support vector data description. *Machine Learning*, 54:45–66, 2004.

Thomas, A., Feuillard, V., and Gramfort, A. Calibration of one-class SVM for MV set estimation. In *2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, pp. 1–9, 2015.

Zhao, M. and Saligrama, V. Anomaly detection with score functions based on nearest neighbor graphs. In *Advances in Neural Information Processing Systems 22*, pp. 2250–2258. 2009.