Estimating the Contamination Factor's Distribution in Unsupervised Anomaly Detection

Lorenzo Perini [1] **Paul-Christian Bürkner** [2] **Arto Klami** [3]

Abstract

Anomaly detection methods identify examples that do not follow the expected behaviour, typically in an unsupervised fashion, by assigning real-valued anomaly scores to the examples based on various heuristics. These scores need to be transformed into actual predictions by thresholding so that the proportion of examples marked as anomalies equals the expected proportion of anomalies, called contamination factor. Unfortunately, there are no good methods for estimating the contamination factor itself. We address this need from a Bayesian perspective, introducing a method for estimating the posterior distribution of the contamination factor for a given unlabeled dataset. We leverage several anomaly detectors to capture the basic notion of anomalousness and estimate the contamination using a specific mixture formulation. Empirically on 22 datasets, we show that the estimated distribution is well-calibrated and that setting the threshold using the posterior mean improves the detectors' performance over several alternative methods.

1. Introduction

Anomaly detection aims at automatically identifying samples that do not conform to the normal behaviour, according to some notion of normality (see e.g., Chandola et al. (2009)). Anomalies are often indicative of critical events such as intrusions in web networks (Malaiya et al., 2018), failures in petroleum extraction (Martí et al., 2015), or breakdowns in wind and gas turbines (Zaher et al., 2009; Yan & Yu, 2019). Such events have an associated high cost and detecting them avoids wasting time and resources.

[1]DTAI lab & Leuven.AI, Department of Computer Science, KU Leuven, Belgium [2]Cluster of Excellence SimTech, University of Stuttgart, Germany [3]Department of Computer Science, University of Helsinki, Finland. Correspondence to: Lorenzo Perini <lorenzo.perini@kuleuven.be>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Typically, anomaly detection is tackled from an unsupervised perspective (Maxion & Tan, 2000; Goldstein & Uchida, 2016; Zong et al., 2018; Perini et al., 2020b; Han et al., 2022) because labeled samples, especially anomalies, may be expensive and difficult to acquire (e.g., you do not want to voluntarily break the equipment simply to observe anomalous behaviours), or simply rare (e.g., you may need to inspect many samples before finding an anomalous one). Unsupervised anomaly detectors exploit data-driven heuristic assumptions (e.g., anomalies are far away from normals) to assign a real-valued score to each sample denoting how anomalous it is. Using such anomaly scores enables ranking the samples from most to least anomalous.

Converting the anomaly scores into discrete predictions would practically allow the user to flag the anomalies. Commonly, one sets a decision threshold and labels samples with higher scores as anomalous and samples with lower scores as normal. However, setting the threshold is a challenging task as it cannot be tuned (e.g., by maximizing the model performance) due to the absence of labels. One approach is to set the threshold such that the proportion of scores above it matches the dataset's *contamination factor* γ, i.e. the expected proportion of anomalies. If the ranking is correct (that is, all anomalies are ranked before any normal instance) then thresholding with exactly the correct γ correctly identifies all anomalies. However, in most of the real-world scenarios the contamination factor is unknown.

Estimating the contamination factor γ is challenging. Existing works provide an estimate by using either some normal labels (Perini et al., 2020a) or domain knowledge (Perini et al., 2022). Alternatively, one can directly threshold the scores through statistical threshold estimators, and derive γ as the proportion of scores higher than the threshold. For instance, the Modified Thompson Tau test thresholder (MTT) finds the threshold through the modified Thompson Tau test (Rengasamy et al., 2021), while the Inter-Quartile Region thresholder (IQR) uses the third quartile plus 1.5 times the inter-quartile region (Bardet & Dimby, 2017). In Section 4 we provide a comprehensive list of estimators.

Transforming the scores into predictions using an incorrect estimate of the contamination factor (or, equivalently, an incorrect threshold) deteriorates the anomaly detector's

performance (Fourure et al., 2021; Emmott et al., 2015) and reduces the trust in the detection system. If such an estimate was coupled with a measure of uncertainty, one could take into account this uncertainty to improve decisions. Although existing methods propose Bayesian anomaly detectors (Shen & Cooper, 2010; Roberts et al., 2019; Hou et al., 2022; Heard et al., 2010), none of them study how to transform scores into hard predictions.

Therefore, we are the first to study the estimation of the contamination factor from a Bayesian perspective. We propose γGMM, the first algorithm for estimating the contamination factor's (posterior) distribution in unlabeled anomaly detection setups. First, we use a set of unsupervised anomaly detectors to assign anomaly scores for all samples and use these scores as a new representation of the data. Second, we fit a Bayesian Gaussian Mixture model with a Dirichlet Process prior (DPGMM) (Ferguson, 1973; Rasmussen, 1999) in this new space. If we knew which components contain the anomalies, we could derive the contamination factor's posterior distribution as the distribution of the sum of such components' weights. Because we do not know this, as a third step γGMM estimates the probability that the k most extreme components are jointly anomalous, and uses this information to construct the desired posterior. The method explained in detail in Section 3.

In summary, we make four contributions. First, we adopt a Bayesian perspective and introduce the problem of estimating the contamination factor's posterior distribution. Second, we propose an algorithm that is able to sample from this posterior. Third, we demonstrate experimentally that the implied uncertainty-aware predictions are well calibrated and that taking the posterior mean as point estimate of γ outperforms several other algorithms in common benchmarks. Finally, we show that using the posterior mean as a threshold improves the actual anomaly detection accuracy.

2. Preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $X\colon \Omega \to \mathbb{R}^d$ a random variable, from which a dataset $D = \{X_1, \ldots, X_N\}$ of N random examples is drawn. Assume that X has a distribution of the form $P = (1-\gamma) \cdot P_1 + \gamma \cdot P_2$, where P_1 and P_2 are the distributions on $\mathbb{R}^d$ corresponding to normal examples and anomalies, respectively, and $\gamma \in [0, 1]$ is the *contamination factor*, i.e. the proportion of anomalies. An (unsupervised) *anomaly detector* is a measurable function $f\colon \mathbb{R}^d \to \mathbb{R}$ that assigns real-valued anomaly scores $f(X)$ to the examples. Such anomaly scores follow the rule that *the higher the score, the more anomalous the example*.

A Gaussian mixture model (GMM) with K components (see e.g. Roberts et al. (1998)) is a generative model defined by a distribution on a space $\mathbb{R}^M$ such that $p(s) = \sum_{k=1}^{K} \pi_k \mathcal{N}(s|\mu_k, \Sigma_k)$ for $s \in \mathbb{R}^M$, where $\mathcal{N}(s|\mu_k, \Sigma_k)$ denotes the Gaussian distribution with mean vector μ_k and covariance matrix $\Sigma_k \in \mathbb{R}^{M \times M}$, and π_k are the mixing proportions such that $\sum_{k=1}^{K} \pi_k = 1$. For finite mixtures, we typically have a Dirichlet prior over $\pi = [\pi_1, \ldots, \pi_K]$, but Dirichlet Process (DP) priors allow treating also the number of components as unknown (Görür & Rasmussen, 2010). For both cases, we need approximate inference to estimate the posterior of the model parameters.

3. Methodology

We tackle the problem: **Given** an unlabeled dataset D and a set of M unsupervised anomaly detectors; **Estimate** a (posterior) distribution of the contamination factor γ.

Learning from an unlabeled dataset has three key challenges. First, the absence of labels forces us to make relatively strong assumptions. Second, the anomaly detectors rely on different heuristics that may or may not hold, and their performance can hence vary significantly across datasets. Third, we need to be careful in introducing user-specified hyperparameters, because setting them properly may be as hard as directly specifying the contamination factor.

In this paper, we propose γGMM, a novel Bayesian approach that estimates the contamination factor's posterior distribution in four steps, which are illustrated in Figure 1:
Step 1. Because anomalies may not follow any particular pattern in covariate space, γGMM maps the covariates $X \in \mathbb{R}^d$ into an M dimensional anomaly space, where the dimensions correspond to the anomaly scores assigned by the M unsupervised anomaly detectors. Within each dimension of such a space, the evident pattern is that "the higher the more anomalous".
Step 2. We model the data points in the new space $\mathbb{R}^M$ using a Dirichlet Process Gaussian Mixture Model (DPGMM) (Neal, 1992; Rasmussen, 1999). We assume that each of the (potentially many) mixture components contains either only normals or only anomalies. If we knew which components contained anomalies, we could then easily derive γ's posterior as the sum of the mixing proportions π of the anomalous components. However, such information is not available in our setting.
Step 3. Thus, we order the components in decreasing order, and we estimate the probability of the largest k components being anomalous. This poses three challenges: (a) how to represent each M-dimensional component by a single value to sort them from the most to the least anomalous, (b) how to compute the probability that the kth component is anomalous given that the $(k-1)$th is such, (c) how to derive the target probability that k components are jointly anomalous.
Step 4. γGMM estimates the contamination factor's posterior by exploiting such a joint probability and the components' mixing proportions posterior.

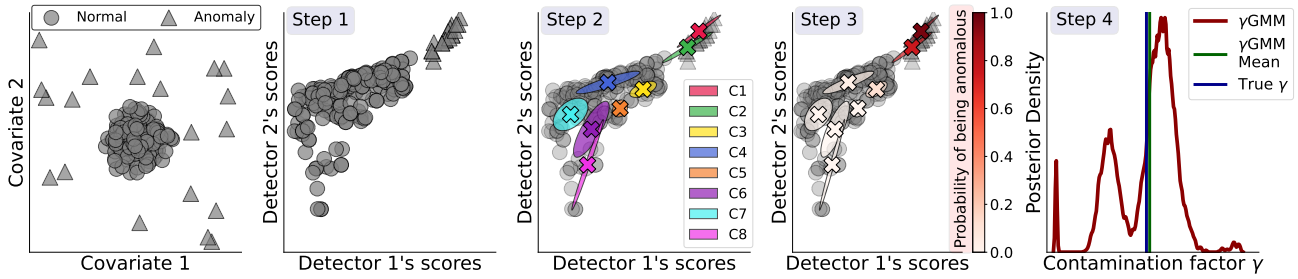

Figure 1. Illustration of the γGMM's four steps on a 2D toy dataset (left plot): we 1) map the 2D dataset into an $M = 2$ dimensional anomaly space, 2) fit a DPGMM model on it, 3) compute the components' probability of being anomalous (conditional, in the plot), and 4) derive $\gamma|S$'s posterior. γGMM's mean is an accurate point estimate for the true value γ^*.

In the following, we describe these steps in detail.

3.1. Representing Data Using Anomaly Scores

Learning from an unlabeled anomaly detection dataset has two major challenges. First, anomalies are rare and sparse events, which makes it hard to use common unsupervised methods like clustering (Breunig et al., 2000). Second, making assumptions on the unlabeled data is challenging due to the absence of specific patterns in the anomalies, which makes it hard to choose a specific anomaly detector.

Therefore, we use a set of M anomaly detectors to map the d-dimensional input space into an M-dimensional score space $\mathbb{R}^M$, such that a sample x gets a score s:

$$\mathbb{R}^d \ni x \rightarrow [f_1(x), f_2(x), \ldots, f_M(x)] = s \in \mathbb{R}^M.$$

This has two main effects: (1) it introduces an interpretable space where the evident pattern is that, within each dimension, higher scores are more likely to be anomalous, and (2) it accounts for multiple inductive biases by using multiple arbitrary anomaly detectors.

To make the dimensions comparable, we (independently for each dimension) map the scores $s \in S$ to $\log(s - \min(S) + 0.01)$, where the log is used to shorten heavy right tails, and normalize them to have zero mean and unit variance.

3.2. Modeling the Density with DPGMM

We use mixture models as basis for quantifying the distribution of the contamination factor, relying on their ability to model the proportions of samples using the mixture weights. For flexible modeling, we use the DPGMM

$$\begin{aligned} s_i &\sim \mathcal{N}(\tilde{\mu}_i, \tilde{\Sigma}_i) \qquad i = 1, \ldots, N \\ (\tilde{\mu}_i, \tilde{\Sigma}_i) &\sim G \\ G &\sim DP(G_0, \alpha) \\ G_0 &= \mathcal{NIW}(M, \lambda, V, u) \end{aligned}$$

where G is a random distribution of the mean vectors μ_i and covariance matrices Σ_i, drawn from a DP with base distribution G_0. We use the explicit representation $G = \sum_{k=1}^{\infty} \pi_k \delta_{(\mu_k, \Sigma_k)}(\tilde{\mu}_i, \tilde{\Sigma}_i)$, where $\delta_{(\mu_k, \Sigma_k)}$ is the delta distribution at (μ_k, Σ_k) and π_k follow the stick-breaking distribution. We set G_0 as Normal Inverse Wishart (Nydick, 2012) with parameters M, λ, V, u common to all components. We use variational inference (VI; see e.g. Blei et al. (2017) for details) for approximating the posterior as VI is computationally efficient and sufficiently accurate for our purposes. Alternative methods (e.g., Markov Chain Monte Carlo (Brooks et al., 2011)) could also be used but were not considered worth the additional computational effort here.

Choice of DPGMM. DPGMM has two key properties that justify its use over other flexible density models. First, we choose Gaussian distributions over more robust heavy-tailed distributions because isolated samples are likely candidates for outliers, and encouraging the model to represent them using the heavy tails would be counter-productive. Second, the rich-get-richer property of DPs is desirable because we expect some very large components of normals but want to allow arbitrarily small clusters of anomalies. Moreover, the DP formulation allows us to refrain from specifying the number of components K. After fitting the model, we only consider the components with at least one observation assigned to them and propagate all the remaining density uniformly over the active components. Thus, for the following steps we can still proceed as if the model was a finite mixture with π following a Dirichlet distribution.

3.3. Estimating the Components' Anomalousness

We assume that each mixture component either contains only anomalous or only normal samples. All unsupervised methods rely on some assumption on nearby samples sharing latent characteristics, and this cluster assumption is a natural and weak assumption. If we knew which components contain anomalies, we could directly derive the posterior of

the contamination factor γ as the sum of the mixing proportions π_k of those components. This is naturally not the case, but we need to estimate it in an unsupervised fashion.

More formally, we estimate the probability that k (out K) components are anomalous such that we can derive γ's posterior by averaging over all the values $0 \leq k \leq K$. We do this in three steps. Initially, we sort the components of score vectors in decreasing order (by degree of anomalousness), which comes natural from the representation we made in Step 1 (Sec. 3.1). Then, our insight is that the kth component can be anomalous only if the $(k-1)$th is such. This points to the estimation of conditional probabilities, i.e., the probability of c_k = "**the kth component is anomalous**" given c_{k-1}. Finally, the probability that exactly the first k components are anomalous can be obtained using basic rules of probability theory.

Assigning an ordering to the components. As initial step for computing the joint probability, we need to design a decreasing ordering map for the components based on their anomalousness. We do this in a manner that accounts for the uncertainty of the components' parameters to rank high the components that can be reliably identified as anomalous: we want the means to be high but the variance low, to avoid the risk that also samples with low anomaly scores could belong to the component.

We construct the overall ranking using dimension-specific scores because our normalization cannot remove all statistical differences between the different detectors. Formally, let $r\colon \mathbb{R}^M \times \mathbb{R}^{M\times M} \to \mathbb{R}$ be the function of the mean vector μ_k and the covariance matrix Σ_k that assigns a real value representing the component k's anomalousness. We set r as

$$r\left(\mu_k^{(z)}, \Sigma_k^{(z)}\right) = \frac{1}{M}\sum_{j=1}^{M} \frac{\mu_k^{j\,(z)}}{1+\sqrt{\Sigma_k^{j,j\,(z)}}}, \tag{1}$$

where $\mu_k^{(z)}$ and $\Sigma_k^{(z)}$ are samples from the parameters' posterior distributions of the kth component. We obtain a representative value of the whole component by taking the expected value of r, i.e. through $\mathbb{E}[r(\mu_k, \Sigma_k)]$. Equation (1) intentionally does not consider inter-dimension correlations, as it remains unclear to us how those should ideally be included and what benefits it would actually provide.

We add 1 to the component's standard deviation for two reasons. First, if a component contains samples with almost the same covariate values, the standard deviation would be close to 0 and the ratio would explode towards infinity, masking any effect of the mean. Second, adding 1 is reasonable because it is equal to the theoretical upper bound of the components' variances, as they are normalized (Sec. 3.1).

Without loss of generality, from now on we assume that the components' index k is ordered based on their representative value such that the kth component has a higher value (i.e., more anomalous) than the $(k+1)$th component.

Estimating the probability that the kth component is anomalous. Because the components are sorted by anomalousness, our key insight is that *the kth component can be anomalous only if the $(k-1)$th is anomalous*. Formally,

$$\mathbb{P}(c_k|\, c_{k-1}) > 0 \;\;\&\;\; \mathbb{P}(c_k|\, \bar{c}_{k-1}) = 0 \quad (1 < k \leq K)$$

where $\bar{c}_{k-1}$ means "not c_{k-1}". Moreover, we assume $\mathbb{P}(c_1) \in (0,1)$. That is, we allow for the data to not have anomalies (< 1) but exclude certain knowledge of no anomalies (> 0). This is a sensible assumption because, if one knew for sure that no anomalies are in the data, then we trivially have $\gamma = 0$, whereas we still need to allow for the data to be free of anomalies if evidence suggests so.

We estimate the conditional probability as

$$\mathbb{P}(c_k|c_{k-1}) = \frac{1}{1+e^{(\tau+\delta\cdot r(\mu_k,\Sigma_k))}}, \tag{2}$$

where τ and δ are the two hyperparameters of the sigmoid function, which will be carefully discussed in Section 3.4. Note that the principle itself is not restricted to this particular choice of functional form. One could apply any transformation that maps to $[0,1]$, but the detailed derivations of the parameters would naturally be different.

Deriving the components' joint probability. Given the conditional probability $\mathbb{P}(c_k|\, c_{k-1})$, the joint probability follows from simple steps. Taking inspiration from the sequential ordinal models (Bürkner & Vuorre, 2019), our insight is that exactly k components are jointly anomalous if and only if each of them is conditionally anomalous and the $(k+1)$th is not anomalous. We indicate this as $C^* = k$. Essentially,

$$\begin{aligned}\mathbb{P}(C^*=k) &\coloneqq \mathbb{P}(c_1,\ldots,c_k,\bar{c}_{k+1},\ldots,\bar{c}_K)\\ &= \mathbb{P}(c_1)\prod_{t=1}^{k-1}\mathbb{P}(c_{t+1}|c_t)(1-\mathbb{P}(c_{k+1}|c_k))\end{aligned} \tag{3}$$

for any $k \leq K$, where $\mathbb{P}(c_{K+1}\,|c_K) = 0$ by convention.

3.4. Estimating the Contamination Factor's Distribution

Given the joint probability that the first k components are anomalous (for $k \leq K$), the contamination factor γ's posterior distribution can be obtained as

$$p(\gamma|S) = \sum_{k=1}^{K} p(C^*=k)\cdot p\left(\sum_{j=1}^{k}\pi_j \middle| S\right) \tag{4}$$

where $p(\sum_{j=1}^{k}\pi_j|S)$ is the posterior distribution of the sum of the first k components' mixing proportions, $p(C^* =$

k) are densities WRT the counting measure. Note that $p(\sum_{j=1}^{k} \pi_j | S) = \text{BETA}(\sum_{j=1}^{k} \alpha_j, \sum_{j=k+1}^{K} \alpha_j)$, if $p(\pi_1, \ldots, \pi_K | S) = \text{DIR}(\alpha_1, \ldots, \alpha_K)$ (Lin, 2016).

Setting the sigmoid's hyperparameters τ and δ. Introducing new hyperparameters when the task is to estimate the contamination factor γ's posterior is risky because setting their value may be as difficult as directly providing a point estimate of γ. Our key insight is that we can obtain τ and δ by asking the user two simple questions: (a) How likely is that no anomalies are in the data? (b) How likely is that a large amount of anomalies occurred, say, more than $t = 15\%$ of the data? Both of these values are supposed to be low. Let's call p_0 and p_{high} the two answers. Formally,

$$p_0 = 1 - \mathbb{P}(c_1) = 1 - \frac{1}{1 + e^{(\tau + \delta \cdot r(\tilde{\mu}_1, \tilde{\Sigma}_1))}}$$

$$p_{\text{high}} = \mathbb{P}(\gamma \geq t | S) = \sum_{k=1}^{K} \mathbb{P}(C^* = k) \cdot \mathbb{P}\left(\sum_{j=1}^{k} \pi_j \geq t | S \right)$$

One can use a numerical solver for non-linear equations with linear constraints (e.g., the least square optimizer implemented in SKLEARN) to find the values of τ and δ that satisfy such constraints. The problem has a unique solution whenever $p_{\text{high}} \geq \mathbb{P}(\pi_1 \geq t | S)$. This holds almost always in our experimental cases, but, in case such a constraint cannot be satisfied, we keep running again the variational inference method (with different starting points) for the DPGMM until the constraint on p_{high} holds. If this cannot happen or does not happen within 100 iterations, we reject the possibility of too high contamination factors and just set it to 0. In the experiments (Q5), we show that changing the p_0 and p_{high} does not have a large impact on γ's posterior.

Sampling from γ's posterior. Our estimate of the contamination factor's posterior $p(\gamma | S)$ does not have a simple closed form. However, we can sample from the distribution using a simple process. The DPGMM inference determines an approximation for $p(\pi, \mu, \Sigma | S)$ and all the quantities required for Equations (2), (3), (4) can be computed based on samples from the approximation. Formally, we derive a sample from $p(\gamma | S)$ in four steps by repeating the next operations for all $k \leq K$. First, we draw a sample $\pi_k^{(z)}, \mu_k^{(z)}, \Sigma_k^{(z)}$ from π_k (Dirichlet), μ_k (Normal), Σ_k (Inverse Wishart). Second, we transform $\pi_k^{(z)}$ by taking the cumulative sum and obtain a sample $\sum_{j=1}^{k} \pi_j^{(z)}$. Third, we pass $\mu_k^{(z)}$ and $\Sigma_k^{(z)}$ through the sigmoid function (2) to get the conditional probabilities $\mathbb{P}(c_k \mid c_{k-1})$, and transform them into the exact joint probabilities $\mathbb{P}(C^* = k)$ using the equation 3. Finally, we multiply the samples following Formula 4 and obtain a sample $\gamma^{(z)}$ from $p(\gamma | S)$.

Additional technical details. Because our method uses the variational inference approximation, we run it 10 times and concatenate the samples to reduce the risk of biased distributions due to local minima. Moreover, after sorting the components, we set $\mathbb{P}(c_k | c_{k-1}) = 0$ for all $k > K' = \arg\max\{k\colon\ \mathbb{E}[\sum_{j=1}^{k} \pi_j] < 0.25\}$. This has the effect of setting an upper bound of 0.25 to the contamination factor γ. Because anomalies must be rare, we realistically assume that it is not possible to have more than 25% of them. Although "0.25" could be considered a hyperparameter, this value has virtually no impact on the experimental results. Moreover, note that $\mathbb{E}[\pi_1] \geq 0.25$ cannot occur, as otherwise we could not set the hyperparameters p_0 and p_{high}.

4. Experiments

We empirically evaluate two aspects of our method: (a) whether it accurately estimates the contamination factor's posterior, and (b) how thresholding the scores using our method affects the anomaly detectors' performance. To this end, we address the following five experimental questions:

Q1. Is the posterior estimate sharp and well-calibrated?

Q2. How does γGMM compare to threshold estimators?

Q3. Does a better point estimate of γ improve the anomaly detector performance?

Q4. What is the impact of the number of detectors M?

Q5. How sensitive the method is to p_0 and p_{high}?

4.1. Experimental Setup

Methods. We compare the sample mean of γGMM[1] with 21 threshold estimators that we cluster into 9 groups:
1. Kernel-based. FGD (Qi et al., 2021) and AUCP (Ren et al., 2018) both use the kernel density estimator to estimate the score density; FGD exploits the inflection points of the density's first derivative, while AUCP uses the percentage of the total kernel density estimator's AUC to set the threshold;
2. Curve-based. EB (Friendly et al., 2013) creates elliptical boundaries by generating pseudo-random eccentricities, while WIND (Jacobson et al., 2013) is based on the topological winding number with respect to the origin;
3. Normality-based. ZSCORE (Bagdonavičius & Petkevičius, 2020) exploits the Z-scores, DSN (Amagata et al., 2021) measures the distance shift from a normal distribution, and CHAU (Bol'shev & Ubaidullaeva, 1975) follows the Chauvenet's criterion before using the Z-score;
4. Regression-based. CLF and REGR (Aggarwal, 2017) are two regression models that separate the anomalies based on

[1]Code and online Supplement are available at: https://github.com/Lorenzo-Perini/GammaGMM

the y-intercept value;

5. Filter-based. FILTER (Hashemi et al., 2019), and HIST (Thanammal et al., 2014) use the wiener filter and the Otsu's method to filter out the anomalous scores;

6. Statistical test-based. GESD (Alrawashdeh, 2021), MCST (Coin, 2008) and MTT (Rengasamy et al., 2021) are based on, respectively, the generalized extreme studentized, the Shapiro-Wilk, and the modified Thompson Tau statistical tests;

7. Statistical moment-based. BOOT (Martin & Roberts, 2006) derives the confidence interval through the two-sided bias-corrected and accelerated bootstrap; KARCH (Afsari, 2011) and MAD (Archana & Pawar, 2015) are based on means and standard deviations, i.e., the Karcher mean plus one standard deviation, and the mean plus the median absolute deviation over the standard deviation;

8. Quantile-based. IQR (Bardet & Dimby, 2017) and QMCD (Iouchtchenko et al., 2019) set the threshold based on quantiles, i.e., respectively, the third quartile Q_3 plus 1.5 times the inter-quartile region $|Q_3 - Q_1|$, and the quantile of one minus the Quasi-Monte Carlo discrepancy;

9. Transformation-based. MOLL (Keyzer & Sonneveld, 1997) smooths the scores through the Friedrichs' mollifier, while YJ (Raymaekers & Rousseeuw, 2021) applies the Yeo-Johnson monotonic transformations.

We apply each threshold estimator to the univariate anomaly scores of each detector at a time. *We average the contamination factors over the M detectors and use it as the final point estimate for each dataset.*

Data. We carry out our study on 20 commonly used benchmark datasets and additionally 2 (proprietary) real tasks. The benchmark datasets contain semantically useful anomalies widely used in the literature (Campos et al., 2016). The datasets vary in size, number of features, and true contamination factor. The online Supplement provides further details. For the real tasks, our experiments focus on preventing blade icing in wind turbines. We use two public wind turbine datasets, where sensors collect various measurements (e.g., wind speed, power energy, etc.) every 7 seconds for either 8 weeks (turbine 15) or 4 weeks (turbine 21). Following (Zhang et al., 2018), we construct feature-vectors by taking the average over the time segment of one minute.

Evaluation metrics. We use three evaluation metrics to assess the performance of the methods. Contrary to all the threshold estimators, our method estimates the posterior of γ. Therefore, we measure the **probabilistic calibration** of γGMM's posterior using a QQ-plot with the x-axis representing the expected probabilities and on the y-axis the empirical frequencies. That is, for $v \in [0, 0.5]$,

$$\text{Expected Prob.} = \mathbb{P}\left(\gamma^* \in [q(0.5 - v), q(0.5 + v)]\right) = 2v$$

$$\text{Empirical Freq.} = \frac{|\{\gamma \in [q(0.5 - v), q(0.5 + v)]\}|}{\#\text{experiments}},$$

where $q(u)$ is the quantile at the value u of our distribution, for $u \in [0, 1]$, and γ^* refers to the true dataset's contamination factor. For evaluating the point estimate of the methods, we use the **mean absolute error** (MAE) between the method's point estimate and the true value. Finally, we measure the impact of thresholding the scores using the methods' point estimate through the F_1 score (Goutte & Gaussier, 2005), as common metrics like the Area Under the ROC curve and the Average Precision are not affected by different thresholds. Specifically, for $m = 1, \ldots, M$, we measure the **relative deterioration of the F_1 score**:

$$F_1 \text{ deterioration} = \frac{F_1(f_m, D, \gamma^*) - F_1(f_m, D, \hat{\gamma})}{F_1(f_m, D, \hat{\gamma})}$$

where we compute the F_1 score on the dataset D using the anomaly detector f_m, and either the true value γ^* or an estimate $\hat{\gamma}$ to threshold the scores. The F_1 deterioration of a method is (mostly) negative, and the higher the better.

Setup. In the experiments we assume a transductive setting (Campos et al., 2016; Scott & Blanchard, 2008; Toron et al., 2022), where a dataset D is used both for training and testing. This is the typical setting of anomaly detection (Breunig et al., 2000; Schölkopf et al., 2001; Angiulli & Pizzuti, 2002; Liu et al., 2012) because the absence of labels and patterns (for the anomaly class) avoids overfitting issues.

For each dataset, we proceed as follows: (i) use a set of M anomaly detectors to assign the anomaly scores S to each observation in the dataset D; (ii) map each anomaly score $s \in S$ to $\log(s - \min(S) + 0.01)$ and normalize them to have mean equal to 0 and standard deviation equal to 1; (iii) either use our method to estimate the contamination factor's posterior and extract the posterior mean as point estimate $\hat{\gamma}$, or use one of the threshold estimators to directly obtain a point estimate $\hat{\gamma}$ of the contamination factor (see methods paragraph above); (iv) evaluate the point estimates using the mean absolute error (MAE) between such estimate and the true value γ^*; (v) use the contamination factor's point estimate to threshold the anomaly scores of each of the M anomaly detectors f_m (individually); (vi) finally, we measure the F_1 score and compute the relative deterioration.

Hyperparameters, anomaly detectors and priors. Our method introduces two new hyperparameters: p_0 and p_{high}. We both of them set to 0.01 as default value because extremely high contamination, as well as no anomalies, are unlikely events. We will experimentally check the impact of these two hyperparameters in Q5.

We use 10 anomaly detectors with different inductive biases (Soenen et al., 2021): KNN (Angiulli & Pizzuti, 2002) assumes that the anomalies are far away from normals, IFOREST (Liu et al., 2012) assumes that the anomalies are easier to isolate, LOF (Breunig et al., 2000) exploits the examples' density, OCSVM (Green & Richardson, 2001) encapsulates the data into a multi-dimensional hypersphere, AE (Chen et al., 2018) and VAE (Kingma & Welling, 2013) use the reconstruction error as anomaly score function in a, respectively, deterministic and probabilistic perspective, LSCP (Zhao et al., 2019a) is an ensemble method that selects competent detectors locally, HBOS (Goldstein & Dengel, 2012) calculates the degree of anomalousness by building histograms, LODA (Pevnỳ, 2016) is an ensemble of weak detectors that build histograms on randomly generated projected spaces, and COPOD (Li et al., 2020) is a copula based method. All these methods are implemented in the python library PyOD (Zhao et al., 2019b).

The threshold estimators are implemented in PYTHRESH[2] with default hyperparameters. Finally, the DPGMM is implemented in SKLEARN: we use the Stick-breking representation (Dunson & Park, 2008), with 100 as upper bound of K. We set the means' prior to 0, and the covariance matrices' prior to identities of appropriate dimension. We opt for such (in our context) weakly-informative priors because sensible prior knowledge of the DPGMM hyperparameters is hard to come by in practice.

4.2. Experimental Results

Q1. Does our method estimate a sharp and well-calibrated posterior of γ? Figure 2 shows the contamination factor γ's posterior estimated by our method on the 22 datasets. In several cases (e.g., WPBC, Cardio, SpamBase, Wilt and T21), the distribution looks accurate as γ's true value (blue line) is close to the posterior mean (i.e., the expected value, the green line). On the contrary, some datasets (e.g., Arrhythmia, Shuttle, KDDCup99, Parkinson, Glass) obtain less accurate distributions: although γ's true value sometimes falls on low-density regions (Arrhythmia, Shuttle), in many cases it would be quite likely to sample the true value from our posterior (KDDCup99, Parkinson, Glass), which makes the density still quite reliable.

Figure 3 shows the calibration plot. The posterior is well-calibrated as it is very close to the dashed black line indicating a perfectly calibrated distribution. The empirical frequencies deviate from the real probabilities by less than 5% (dark shadow grey) in more than 76% of the cases, while never deviating by more than 10% (light shadow grey).

Q2. How does γGMM compare to the threshold estimators? We take γGMM's posterior mean as our best point estimate of γ and compare such value to the point estimates obtained from the threshold estimators. Figure 4 illustrates the ordered MAE (mean $\pm$ std.) between the methods' estimate and the true γ. On average, γGMM obtains a MAE of 0.026 that is 20% lower than the best runner-up MTT and 27% lower than the third best method QMCD (MAE of 0.033 and 0.036). For each experiment, we rank the methods from the best (position 1, lowest MAE) to the worst (position 22, greatest MAE). Our method has the best average rank (2.13 ± 1.04). Moreover, γGMM ranks first 8 times ($\approx 36\%$ of the cases), and for 13 times ($\approx 60\%$ of the cases) it is in the top two. The next best method, MTT, ranks first in 6 cases with an average rank of 2.30 ± 1.10.

Q3. Does a better contamination improve the anomaly detectors' performance? We use γGMM's posterior mean as a point estimate to measure the F_1 score of the anomaly detectors because sampling from the distribution would not imply a fair comparison against the other methods that can only provide a point estimate. Moreover, anomaly detectors that fail to rank the samples accurately perform poorly even when using the correct γ. Since our focus is studying the effect of γ, for each dataset D, we compare F_1 scores only over the detectors that achieve the greatest F_1 score using the true contamination factor γ^*, i.e. $\arg\max_{f_m} \{F_1(f_m, D, \gamma^*)\}$. The online Supplement contains the list of detectors used for each experiment.

Figure 5 shows the average ($\pm$ std.) deterioration for each of the methods. On average, γGMM has the best F_1 deterioration (-0.117 ± 0.228) that is around 10% better than the runner-up QMCD (-0.131 ± 0.238), and 58% better than the next best KARCH (-0.279 ± 0.248). For 25% of the cases we get higher F_1 score with γGMM than when using the true γ^*. This is due to the (still incorrect) ranks made by the detectors, which achieve better performance with slightly incorrect contamination factors. The online Supplement provides further details on how the methods perform in terms of false alarms and false negatives.

Q4. What is the impact of M on γ's posterior? In the previous experiments, we used $M = 10$ detectors. We evaluate the effect of M by running all the experiments 10 times with (different) randomly chosen detectors for $M = 3, 5, 7$. Figure 6 shows that the calibration suffers if using fewer detectors, but already $M = 5$ let the method work fairly well. The variance of the results (over repeated experiments) also increases for lower M.

Q5. Impact of the hyperparameters p_0 and p_{high}. We evaluate the impact of p_0 and p_{high} by running the experiments with smaller and larger values than 0.01: we vary, one at a time, p_0, $p_{\text{high}} \in [0.0001, 0.001, 0.05, 0.1]$ and keep the other set as default. Figure 7 shows the QQ-plot for p_0

[2]Link: https://github.com/KulikDM/pythresh.

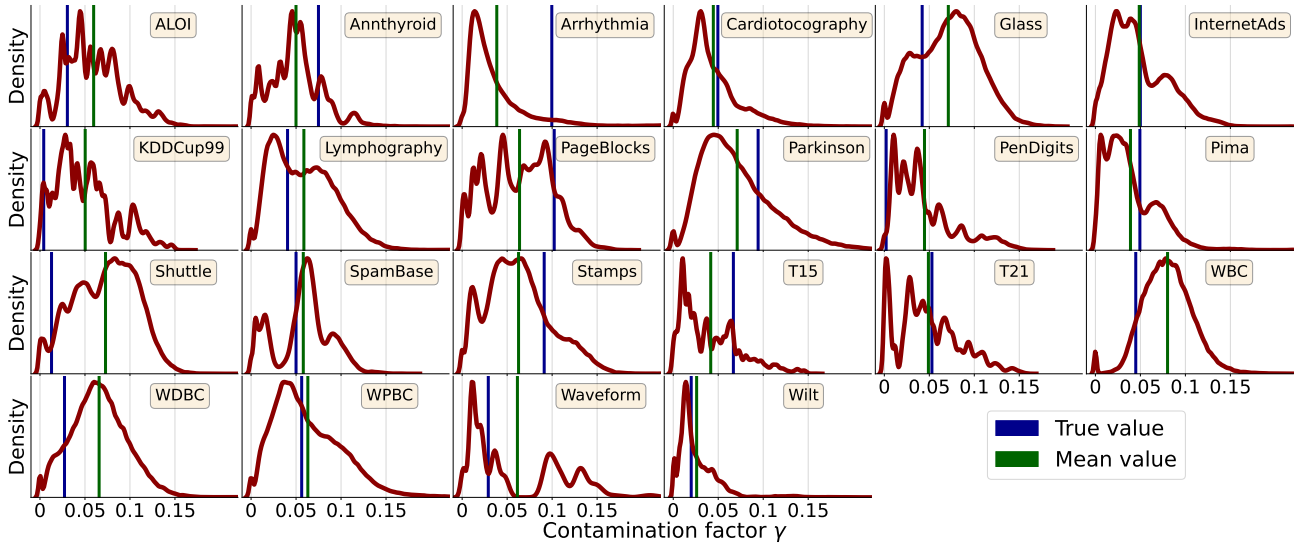

Figure 2. Illustration of how γGMM estimates γ's posterior distribution (red) on all the 22 datasets. The blue vertical line indicates the true contamination factor, while the green line is the posterior's mean.

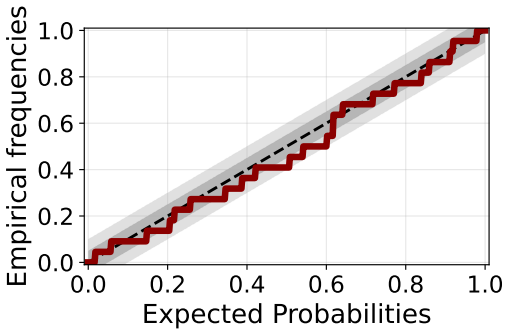

Figure 3. QQ-plot of γGMM's distribution estimate. The black dashed line illustrates the perfect calibration, while shades indicate a deviation of 5% (dark) and 10% (light) from the black line.

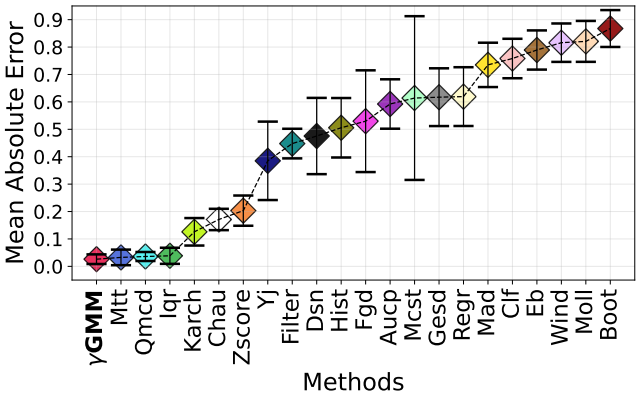

Figure 4. Average MAE ($\pm$ std.) of γGMM's sample mean compared to the other methods. Our method has the lowest (better) average, which is 20% lower than the runner-up.

(left) and p_{high} (right). In both cases, smaller hyperparameters lead to slightly under-estimated expected probabilities. Overall, our method is robust to different values of p_0, while p_{high} affects the calibration slightly more. Comparing the resulting 8 variants of γGMM in terms of MAE, we conclude that the posterior means produce similar values to our default setting, obtaining an MAE that varies from 0.252 ($p_{\text{high}} = 0.001$, the best) to 0.32 ($p_0 = 0.0001$, the worst).

5. Conclusion

The literature on anomaly detection has focused on unsupervised algorithms, but largely ignored practical challenges in their application. The algorithms are evaluated on performance metrics focusing on the ranking of the samples (e.g., AUC), and the ultimate choice of detecting the actual anomalies by thresholding the predictions is left to the practitioners. They lack good means for thresholding and thus often resort to using labels for such goal. This largely defeats the point of using unsupervised methods.

We presented the first practical method for estimating the posterior distribution of the contamination factor γ in a completely unsupervised manner. We empirically demonstrated on 22 datasets that our mean estimates effectively solve the question of where to threshold the predictions. We outper-

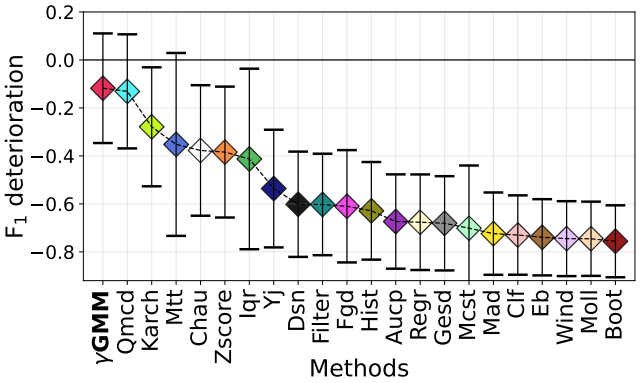

Figure 5. F_1 deterioration (mean $\pm$ std) for each method, where the higher the better. γGMM ranks as best method, obtaining $\approx 10\%$ higher average than QMCD.

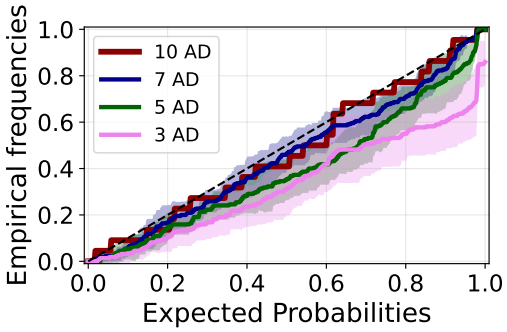

Figure 6. QQ-plot comparing the calibration curves of γGMM when a different number M of detectors is used. The colored shades report the uncertainty obtained by randomly sampling the detectors from a set of 10 detectors. The plot shows that the higher the number of detectors, the more calibrated the distribution.

form all 21 comparison methods and show that the gap in detection accuracy between our estimate and the ground truth (available for these benchmark datasets) is small.

Besides solving the practical question of thresholding the predictions, we seek to persuade the anomaly detection community of the usefulness of a fully probabilistic solution for the problem. Especially in unsupervised settings, it would be completely unreasonable to expect the contamination factor could be identified exactly, but rather we need to characterize its uncertainty. However, we are not aware of any previous works even attempting this. As shown in Fig. 2, the posterior distribution of γ may not only be wide but also multi-modal. Communicating these aspects to the practitioner is critical so that they can e.g. use additional domain knowledge to interpret the alternatives. We showed that our estimates have near-perfect calibration over the broad range

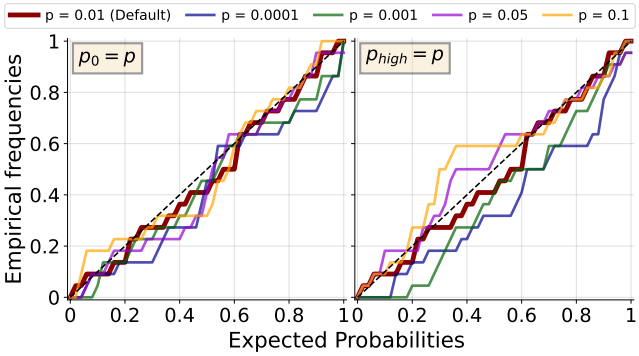

Figure 7. QQ-plot showing how calibrated γGMM's posterior mean would be if we varied p_0 (left) and p_{high} (right). While p_0 does not have a large impact on the method, the empirical frequencies slightly under (over) estimate the expected probabilities for low (high) values of p_{high}.

of datasets and hence can be relied on in practical use.

On first impression, the success of our method in solving this challenging and seemingly ill-posed problem may seem surprising. However, it can be attributed to a careful choice of strong inductive biases built into the underlying probabilistic model. We argue that all of the following elements are necessary, each substantially contributing to the overall success: (i) representing the data in the space of anomaly detector scores defines a meaning for the dimensions and allows borrowing inductive biases of arbitrary detector algorithms, (ii) the mixture model encodes a natural clustering assumption for both the normal samples and the anomalies, (iii) the ordering used for determining the final distribution incorporates both the location and shape of the mixture components in a carefully balanced manner, and (iv) the transformation from the ordering to probabilities is robustly parameterized via just two intuitive hyperparameters, enabling use of the same defaults for all cases.

Acknowledgments

This work was done during LP's research visit to the University of Helsinki, funded by the Gustave Boël - Sofina Fellowship (grant V407821N). Moreover, this work is supported by (LP) the FWO-Vlaanderen aspirant grant 1166222N and by the Flemish government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program, (PB) the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016, (AK) the Academy of Finland (grants 313125 and 336019), the Flagship program Finnish Center for Artificial Intelligence (FCAI), and the Finnish-American Research and Innovation Accelerator (FARIA).

References

Afsari, B. Riemannian l^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

Aggarwal, C. C. An introduction to outlier analysis. In *Outlier analysis*, pp. 1–34. Springer, 2017.

Alrawashdeh, M. J. An adjusted Grubbs' and Generalized Extreme Studentized Deviation. *Demonstratio Mathematica*, 54(1):548–557, 2021.

Amagata, D., Onizuka, M., and Hara, T. Fast and exact outlier detection in metric spaces: a proximity graph-based approach. In *Proceedings of the 2021 International Conference on Management of Data*, pp. 36–48, 2021.

Angiulli, F. and Pizzuti, C. Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*, pp. 15–27. Springer, 2002.

Archana, N. and Pawar, S. Periodicity Detection of Outlier Sequences using Constraint Based Pattern Tree with MAD. *International Journal of Advanced Studies in Computers, Science and Engineering*, 4(6):34, 2015.

Bagdonavičius, V. and Petkevičius, L. Multiple outlier detection tests for parametric models. *Mathematics*, 8 (12):2156, 2020.

Bardet, J.-M. and Dimby, S.-F. A new non-parametric detector of univariate outliers for distributions with unbounded support. *Extremes*, 20(4):751–775, 2017.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational Inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Bol'shev, L. and Ubaidullaeva, M. Chauvenet's test in the classical theory of errors. *Theory of Probability & Its Applications*, 19(4):683–692, 1975.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

Bürkner, P.-C. and Vuorre, M. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101, 2019.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., Assent, I., and Houle, M. E. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*, 30(4):891–927, 2016.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly Detection: A survey. *ACM computing surveys (CSUR)*, 41(3): 1–58, 2009.

Chen, Z., Yeo, C. K., Lee, B. S., and Lau, C. T. Autoencoder-based network Anomaly Detection. In *2018 Wireless telecommunications symposium (WTS)*, pp. 1–5. IEEE, 2018.

Coin, D. Testing normality in the presence of outliers. *Statistical Methods and Applications*, 17(1):3–12, 2008.

Dunson, D. B. and Park, J.-H. Kernel Stick-Breaking processes. *Biometrika*, 95(2):307–323, 2008.

Emmott, A., Das, S., Dietterich, T., Fern, A., and Wong, W.-K. A meta-analysis of the Anomaly Detection problem. *arXiv preprint arXiv:1503.01158*, 2015.

Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230, 1973.

Fourure, D., Javaid, M. U., Posocco, N., and Tihon, S. Anomaly Detection: how to artificially increase your f_1-score with a biased evaluation protocol. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 3–18. Springer, 2021.

Friendly, M., Monette, G., and Fox, J. Elliptical insights: understanding statistical methods through elliptical geometry. *Statistical Science*, 28(1):1–39, 2013.

Goldstein, M. and Dengel, A. Histogram-based outlier score (HBOS): A fast unsupervised Anomaly Detection algorithm. *KI-2012: poster and demo track*, 9, 2012.

Goldstein, M. and Uchida, S. A comparative evaluation of unsupervised Anomaly Detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.

Görür, D. and Rasmussen, E. C. Dirichlet Process Gaussian Mixture Models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

Goutte, C. and Gaussier, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*, pp. 345–359. Springer, 2005.

Green, P. J. and Richardson, S. Modelling heterogeneity with and without the Dirichlet Process. *Scandinavian journal of statistics*, 28(2):355–375, 2001.

Han, S., Hu, X., Huang, H., Jiang, M., and Zhao, Y. ADBench: Anomaly Detection Benchmark. *arXiv preprint arXiv:2206.09426*, 2022.

Hashemi, N., German, E. V., Ramirez, J. P., and Ruths, J. Filtering approaches for dealing with noise in Anomaly Detection. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 5356–5361. IEEE, 2019.

Heard, N. A., Weston, D. J., Platanioti, K., and Hand, D. J. Bayesian Anomaly Detection methods for social networks. *The Annals of Applied Statistics*, 4, 2010.

Hou, Y., He, R., Dong, J., Yang, Y., and Ma, W. IoT Anomaly Detection Based on Autoencoder and Bayesian Gaussian Mixture Model. *Electronics*, 11(20):3287, 2022.

Iouchtchenko, D., Raymond, N., Roy, P.-N., and Nooijen, M. Deterministic and quasi-random sampling of optimized Gaussian Mixture distributions for Vibronic Monte Carlo. *arXiv preprint arXiv:1912.11594*, 2019.

Jacobson, A., Kavan, L., and Sorkine-Hornung, O. Robust inside-outside segmentation using generalized winding numbers. *ACM Transactions on Graphics (TOG)*, 32(4): 1–12, 2013.

Keyzer, M. A. and Sonneveld, B. Using the mollifier method to characterize datasets and models: the case of the universal soil loss equation. *ITC Journal*, 3(4):263–272, 1997.

Kingma, D. P. and Welling, M. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Li, Z., Zhao, Y., Botta, N., Ionescu, C., and Hu, X. COPOD: copula-based outlier detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1118–1123. IEEE, 2020.

Lin, J. On the Dirichlet distribution. *Department of Mathematics and Statistics, Queens University*, 2016.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation-based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.

Malaiya, R. K., Kwon, D., Kim, J., Suh, S. C., Kim, H., and Kim, I. An empirical evaluation of Deep Learning for Network Anomaly Detection. In *2018 International Conference on Computing, Networking and Communications (ICNC)*, pp. 893–898. IEEE, 2018.

Martí, L., Sanchez-Pi, N., Molina, J. M., and Garcia, A. C. B. Anomaly Detection based on sensor data in petroleum industry applications. *Sensors*, 15(2):2774–2797, 2015.

Martin, M. A. and Roberts, S. An evaluation of bootstrap methods for outlier detection in least squares regression. *Journal of Applied Statistics*, 33(7):703–720, 2006.

Maxion, R. A. and Tan, K. M. Benchmarking anomaly-based detection systems. In *Proceeding International Conference on Dependable Systems and Networks. DSN 2000*, pp. 623–630. IEEE, 2000.

Neal, R. M. Bayesian Mixture Modeling. In *Maximum Entropy and Bayesian Methods*, pp. 197–211. Springer, 1992.

Nydick, S. W. The Wishart and inverse Wishart distributions. *Electronic Journal of Statistics*, 6(1-19), 2012.

Perini, L., Vercruyssen, V., and Davis, J. Class prior estimation in active positive and unlabeled learning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI 2020)*, pp. 2915–2921. IJCAI-PRICAI, 2020a.

Perini, L., Vercruyssen, V., and Davis, J. Quantifying the confidence of anomaly detectors in their example-wise predictions. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 227–243. Springer, 2020b.

Perini, L., Vercruyssen, V., and Davis, J. Transferring the Contamination Factor between Anomaly Detection Domains by Shape Similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4128–4136, 2022.

Pevnỳ, T. LODA: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, 2016.

Qi, Z., Jiang, D., and Chen, X. Iterative gradient descent for outlier detection. *International Journal of Wavelets, Multiresolution and Information Processing*, 19(04):2150004, 2021.

Rasmussen, C. The infinite Gaussian Mixture Model. *Advances in neural information processing systems*, 12, 1999.

Raymaekers, J. and Rousseeuw, P. J. Transforming variables to central normality. *Machine Learning*, pp. 1–23, 2021.

Ren, K., Yang, H., Zhao, Y., Chen, W., Xue, M., Miao, H., Huang, S., and Liu, J. A robust AUC maximization framework with simultaneous outlier detection and feature selection for positive-unlabeled classification. *IEEE*

transactions on neural networks and learning systems, 30 (10):3072–3083, 2018.

Rengasamy, D., Rothwell, B. C., and Figueredo, G. P. Towards a more reliable interpretation of machine learning outputs for safety-critical systems using feature importance fusion. *Applied Sciences*, 11(24):11854, 2021.

Roberts, E., Bassett, B. A., and Lochner, M. Bayesian Anomaly Detection and Classification. *arXiv preprint arXiv:1902.08627*, 2019.

Roberts, S. J., Husmeier, D., Rezek, I., and Penny, W. Bayesian approaches to Gaussian Mixture Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471, 2001.

Scott, C. and Blanchard, G. Transductive Anomaly Detection. Technical report, Tech. Rep., 2008, http://www. eecs. umich. edu/cscott, 2008.

Shen, Y. and Cooper, G. A new prior for Bayesian Anomaly Detection. *Methods of Information in Medicine*, 49(01): 44–53, 2010.

Soenen, J., Van Wolputte, E., Perini, L., Vercruyssen, V., Meert, W., Davis, J., and Blockeel, H. The effect of hyperparameter tuning on the comparative evaluation of unsupervised Anomaly Detection methods. In *Proceedings of the KDD*, volume 21, pp. 1–9, 2021.

Thanammal, K., Vijayalakshmi, R., Arumugaperumal, S., and Jayasudha, J. Effective Histogram Thresholding Techniques for Natural Images Using Segmentation. *Journal of Image and Graphics*, 2(2):113–116, 2014.

Toron, N., Mourão-Miranda, J., and Shawe-Taylor, J. Transductgan: a Transductive Adversarial Model for Novelty Detection. *arXiv e-prints*, pp. arXiv–2203, 2022.

Yan, W. and Yu, L. On accurate and reliable Anomaly Detection for gas turbine combustors: A deep learning approach. *arXiv preprint arXiv:1908.09238*, 2019.

Zaher, A., McArthur, S., Infield, D., and Patel, Y. Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, 12(6):574–593, 2009.

Zhang, L., Liu, K., Wang, Y., and Omariba, Z. B. Ice detection model of wind turbine blades based on Random Forest classifier. *Energies*, 11(10):2548, 2018.

Zhao, Y., Nasrullah, Z., Hryniewicki, M. K., and Li, Z. LSCP: Locally selective combination in parallel outlier ensembles. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 585–593. SIAM, 2019a.

Zhao, Y., Nasrullah, Z., and Li, Z. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research*, 20:1–7, 2019b.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep Autoencoding Gaussian Mixture Model for unsupervised Anomaly Detection. In *International conference on learning representations*, 2018.