Constraining Anomaly Detection with Anomaly-Free Regions

Maximilian Toller Know Center Graz, Austria mtoller@know-center.at Hussain Hussain Graz University of Technology Graz, Austria hussain@tugraz.at Roman Kern Graz University of Technology Graz, Austria rkern@tugraz.at Bernhard C. Geiger Know Center Graz, Austria geiger@ieee.org

Abstract—We propose the novel concept of anomaly-free regions (AFR) to improve anomaly detection. An AFR is a region in the data space for which it is known that there are no anomalies inside it, e.g., via domain knowledge. This region can contain any number of normal data points and can be anywhere in the data space. AFRs have the key advantage that they constrain the estimation of the distribution of non-anomalies: The estimated probability mass inside the AFR must be consistent with the number of normal data points inside the AFR. Based on this insight, we provide a solid theoretical foundation and a reference implementation of anomaly detection using AFRs. Our empirical results confirm that anomaly detection constrained via AFRs improves upon unconstrained anomaly detection. Specifically, we show that, when equipped with an estimated AFR, an efficient algorithm based on random guessing becomes a strong baseline that several widely-used methods struggle to overcome. On a dataset with a ground-truth AFR available, the current state of the art is outperformed.

Index Terms—anomaly detection, maximum likelihood estimation, constrained optimization

I. INTRODUCTION

Anomaly detection (AD) is a common data mining task where the goal is to distinguish deviating data, i.e., anomalies, from normal data. Many common AD methods are based on the so called concentration assumption [1]. The concentration assumption implies that normal data points form dense clusters, where normal data points have low distances to other normal data points. Anomalies are then typically found in less dense regions, or have a higher distance to other points. The concentration assumption is made by a number of prominent AD methods (e.g. LOF [2], SVDD [3], Isolation Forest [4], and others [2]–[9]) and has demonstrated its usefulness in a number of scenarios of practical relevance. However, in cases where additional knowledge about the nature of the dataset is available, such knowledge cannot be effectively integrated in the assumption.

a) Concept.: We propose a novel concept, namely Anomaly Free Regions (AFR), which is similar to the concentration assumption. An AFR is a region in the data space for which it is known that there are no anomalies inside, e.g., via domain knowledge. This region can contain any number of normal data points and can be anywhere in the data space. The AFR does not need to coincide with a high-density region—it could be a region that contains no data points at all. Hence, using AFRs is appropriate in cases where domain knowledge allows one to derive such regions and certain parts of the data space do not exhibit anomalous behavior. Prior works that include such domain knowledge are related to mechanical engineering [10], [11], mobile data traffic monitoring [12], building energy consumption [13], and digital twins [14].

b) Novelty.: Several typical concepts in AD literature are comparable to AFRs, such as density level sets [1] and inlier clusters [15]. AFRs have two distinct properties: a) they are not required to contain normal data points; b) they are a subset of the data space that has zero probability of containing anomalies. As such, AFRs encode prior information about the anomaly class, which makes them comparable to semi-supervised AD methods (e.g. [16], [17]) and Positive-Unlabeled (PU) learning [9], [18]. However, in contrast to semi-supervised methods, AFRs do not require explicit labels for either normal or abnormal data points.

c) Illustration.: AFRs are especially useful for probabilistic AD methods since they constrain the estimation of the distribution of non-anomalies: The estimated probability mass inside the AFR must be consistent with the number of normal data points inside the AFR. For example, if the AFR is empty, we would expect very little probability mass inside it. Conversely, if the AFR contains many non-anomalies, we would expect the distribution of non-anomalies to have much of its probability mass inside the AFR. Figure 1 provides an illustration of this benefit.

d) Mathematical foundation.: Due to their mathematical nature, we can analyse AFRs in great detail. Thus, we build a solid foundation for a novel, *constrained* form of AD and construct a concrete reference algorithm. This algorithm provides a sound basis for further methods based on AFRs.

e) Real-world usefulness.: In settings where no AFR is known, an AFR can be estimated. We show the proposed algorithm behaves solidly in cases where no AFR is known and is competitive with state of the art AD methods.

f) Contributions.: Our contributions include the following:

The Know-Center is funded within COMET Competence Centers for Excellent Technologies under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry for Digital and of Economic Affairs, the Austrian Research Promotion Agency (FFG), and by the province of Styria. The COMET programme is managed by the FFG.

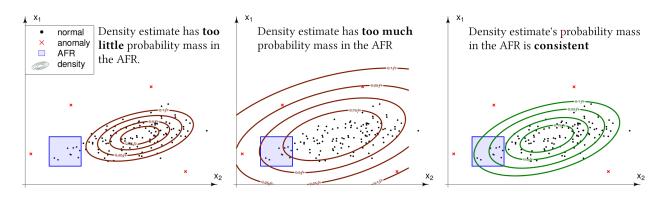


Fig. 1. Constraining density estimation with an AFR. For a fixed toy dataset and a fixed AFR, three different density estimates for normal data points are compared. The *left* and *middle* density estimates are inconsistent and give a too low likelihood of observing the (known normal) data points inside the AFR. The *right* density estimate makes use of the available knowledge (AFR) and gives a reasonable likelihood for the data points inside the AFR.

- Proposal of the novel AFR concept
- Theoretical foundation for constrained anomaly detection
- Proposition of an algorithm for constrained density estimation and application to anomaly detection
- Demonstration of the behaviour on several benchmark datasets where no AFRs are known or may not exist.
- Demonstration of very good performance on a benchmark dataset where an AFR is known.

II. RELATED WORK

There are two main branches of literature that describe theory and methods which relate to constrained AD via AFRs.

- Identification of two-component mixture models where one component is partially known and the other is unknown.
- 2) Semi-supervised AD, one-class classification and PU learning when positive class labels represents anomalies.

The first branch stems from computational statistics and information theory. It is primarily concerned with distributions/mixtures of the form

$$g(x) = (1-p)f_N(x,\boldsymbol{\theta}) + pf_{\infty}(x) \tag{1}$$

where f_N is the "normal" distribution, θ are its parameters, f_{∞} is the unknown (anomaly) distribution, and p is the mixing parameter. Works within this branch typically start their analysis with (a variant of) Equation (1) and then split according to the assumptions they make. These distinguishing assumptions are symmetry of f_{∞} [19], membership of f_N in the Gaussian family and linearly constrained f_{∞} [20], Gaussian zero-mean f_N [21], and that the true form of f_N and its parameters are given [22]-[24]. These works are related to our work because we also start our theoretical investigation with Equation (1). Our work is different from the cited works since we constrain f_{∞} via AFRs. As we will see later, this implies a consistency constraint via the Law of Large Numbers (LLN) that further distinguishes our setup. In the context of AD, AFRs are easier to justify than the assumptions made in related work, e.g., symmetry of f_{∞} .

The second branch of related work stems from machine learning and data science. Semi-supervised AD, and PU learning are a form of classification that uses labeled and unlabeled data to delimit a region where most data from the negative class can be found. In the context of AD, this translates to learning to delimit the region where normal data live using labeled anomalies and unlabeled data. Recently, many relevant semi-supervised and PU-learning schemes were proposed for AD. These are *deep* [16], use *fast random projections* [25], use *metric-learning* [18], and are *specialized on tabular data* [17]. Many other related approaches are summarized in the survey by Ruff et al. [1]. Semi-supervised AD methods are related to our work since they also leverage available prior information to facilitate the detection task. However, unlike these methods, our work does not focus on generalizing from labels and instead uses an AFR to encode prior information about the anomaly class.

III. THEORY

The starting point of our analysis is the two-component mixture described in (1). The probability density function (PDF) f_N generates "normal" data, whereas PDF f_∞ generates anomalies, and $0 \le p \le 1$ is the mixing parameter. The " ∞ " symbol is used in f_∞ to convey the intuition that there are arbitrarily many different anomaly types/classes. In the following, it is assumed that an AFR exists, which is specified in Definition 1.

Definition 1 (Anomaly-free Region). Let $R \subset \mathbb{R}$ and let f_{∞} be the unknown anomaly PDF in (1). Subset R is an AFR with respect to f_{∞} if it holds that

$$\int_{R} f_{\infty}(x) dx = 0.$$
 (2)

In other words, an AFR is a region in which there is zero probability that anomalies occur. Note that this does not impose a restriction on the region where normal data can occur.

The mixture in (1) can be thought of as combining two distributions via a (biased) coin flip—with probability p the coin lands on "heads" and we draw an anomaly from f_{∞} , and

with probability 1 - p the coin lands on "tails" and we draw a normal data point from f_N . Let B be the Bernoulli random variable that describes this coin flip. By making B explicit, we can rewrite (1) as

$$g(x,B) = ((1-p)f_N(x,\theta))^{1-B} (pf_{\infty}(x))^B.$$
 (3)

Next, the mixture in (3) needs to be connected with a dataset. Let $\boldsymbol{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n$ be a dataset obtained by drawing *n* data points independently from $g(\cdot|B)$ according to a sequence of iid Bernoulli variables $\boldsymbol{B} = [B_1, \ldots, B_n]$. Because of independence, the likelihood of the parameters given dataset \boldsymbol{x} is given by

$$\mathcal{L}(\boldsymbol{\theta}, p, \boldsymbol{B} | \boldsymbol{x}) = \prod_{t=1}^{n} \left((1-p) f_N(x_t, \boldsymbol{\theta}) \right)^{1-B_t} \left(p f_\infty(x_t) \right)^{B_t}.$$
(4)

For each x_t in x, the corresponding B_t indicates if x_t was drawn from f_N or f_{∞} , i.e., if x_t is normal or an anomaly. Hence, the underlying anomaly detection problem is solved if we figure out B. However, it is likely impossible to optimize B without knowing anything about f_{∞} . To resolve this dilemma, one can replace B with an estimate \hat{B} . This way, one can rewrite the likelihood in (4) as

$$\mathcal{L}(\boldsymbol{\theta}, p | \hat{\boldsymbol{B}}, \boldsymbol{x}) = \prod_{t=1}^{n} \left((1-p) f_N(x_t, \boldsymbol{\theta}) \right)^{1-\hat{B}_t} \left(p f_\infty(x_t) \right)^{\hat{B}_t}$$
(5)

which is the same as (4) except that B is replaced with its estimate \hat{B} , and that this estimate is given. Notice that in likelihood function (5) the anomaly distribution f_{∞} has no learnable parameters, which has the advantage that the maximum of \mathcal{L} only depends on p, θ .

A. Constrained Likelihood

Directly optimizing the likelihood in (5) is not advisable since the result may violate two important constraints:

- 1) Valid parameter constraint: We must ensure that only permitted values of θ and p are considered during the optimization, i.e., θ must be a real-valued vector/matrix and $0 \le p \le 1$.
- Consistency constraint: We must ensure that θ and p are such that the probability measure of g inside R is consistent with the number of data we observe inside R due to the LLN.

In contrast to the first constraint, the second is less intuitive. To further motivate the necessity of the consistency constraint, note that the probability of observing data outside R is given by

$$\mathbb{P}(x \notin R) = 1 - \int_{R} g(x) = (1-p) \left(1 - \int_{R} f_{N}(x, \theta)\right) + p$$
$$= 1 - (1-p) \int_{R} f_{N}(x, \theta) = P.$$
(6)

If x is drawn from g, then due to the LLN P can be estimated via

$$P = P(\boldsymbol{\theta}, p) \approx \hat{P} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}_{\overline{R}}(x_t)$$
(7)

where $\mathbf{1}_{\overline{R}}(\cdot)$ is the indicator function of the complement of R. The consistency constraint has the purpose to ensure that $\boldsymbol{\theta}$ and p are chosen such that the approximation in (7) holds.

Due to the high relevance of P for AFR-based anomaly detection, it is important to have a reliable estimator for P. Depending on R, the true P may be very close to its extreme values 0 and 1. Since the estimator \hat{P} in (7) is biased at these extremes, it is better to use the Wilson score method

$$P \in \overline{P} \pm w$$
$$\overline{P} = \frac{1}{1 + \frac{z^2}{n}} \left(\hat{P} + \frac{z^2}{2n} \right)$$
(8)

$$w = \frac{z}{1 + \frac{z^2}{n}} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n} + \frac{z^2}{4n^2}}$$
(9)

where z is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution for significance level α .

Adding the two discussed constraints to likelihood (5) gives the following optimization problem:

maximize
$$\mathcal{L}(\boldsymbol{\theta}, p | \boldsymbol{B}, \boldsymbol{x}, R)$$

subject to $0 \le p \le 1, \boldsymbol{\theta} \in \mathbb{R}^{q}$
 $\bar{P} - w \le 1 - (1 - p) \int_{R} f_{N}(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} \le \bar{P} + w$
(10)

Solving this optimization problem gives the constrained maximum likelihood estimates of parameters θ and anomaly probability p.

B. Maximum Likelihood Estimators for p and θ

As we will see, mixture parameter the maximum likelihood estimate (MLE) of p has the same general form regardless of the underlying normal distribution f_N , while the MLE of θ is specific to f_N . In the following, we abbreviate $I = \int_B f_N(x, \theta) dx$ to keep the notation simple.

1) Karush-Kuhn-Tucker conditions: We can rewrite the constrained optimization problem (10) as a system of

(in)equations using the Karush-Kuhn-Tucker conditions:

$$L = L(\boldsymbol{\theta}, p | \boldsymbol{B}, \boldsymbol{x})$$

=
$$\frac{\mathcal{L}(\boldsymbol{\theta}, p | \hat{\boldsymbol{B}}, \boldsymbol{x}) - \lambda_1 (p^2 - p) - \lambda_2 ((1 - (1 - p)I - \bar{P})^2 - w^2)$$
(11a)

$$0 = \frac{\partial \mathcal{L}}{\partial p} - \lambda_1 (2p - 1) - 2\lambda_2 I (1 - (1 - p)I - \bar{P}) \quad (11b)$$

$$0 = \frac{\partial \mathcal{L}}{\partial \theta^{(1)}} + 2\lambda_2 (1-p) \left(1 - (1-p)I - P\right) \frac{\partial}{\partial \theta^{(1)}} I$$
(11c)

$$0 = \frac{\partial \mathcal{L}}{\partial \theta^{(2)}} + 2\lambda_2 (1-p) \left(1 - (1-p)I - \bar{P}\right) \frac{\partial}{\partial \theta^{(2)}} I$$
(11d)

$$0 = \frac{\partial \mathcal{L}}{\partial \theta^{(q)}} + 2\lambda_2(1-p) \left(1 - (1-p)I - \bar{P}\right) \frac{\partial}{\partial \theta^{(q)}} I(11e)$$

$$0 = \lambda_1 \left(p^2 - p \right) \qquad (compl. slack.) \qquad (111)$$

$$0 = \lambda_2 \left(\left(1 - (1 - p)I - \bar{P} \right)^2 - w^2 \right) \quad \text{(compl. slack.)} \quad (11g)$$

$$0 \ge p^2 - p \tag{11h}$$

$$w^2 \ge \left(1 - (1 - p)I - \bar{P}\right)^2$$
 (11i)

$$0 \le \lambda_1, \lambda_2 \tag{11j}$$

Because of complementary slackness [26], we must have one of the following four cases:

- 1) $\lambda_1 = 0$ and $\lambda_2 = 0$: The standard unconstrained MLEs for θ and p satisfy both constraints.
- 2) $\lambda_1 \neq 0$ and $\lambda_2 = 0$: The valid parameter constraint requires p to be either p = 0 or p = 1, whereas θ can be estimated with the standard MLE since θ does not appear in (11f).
- 3) $\lambda_1 = 0$ and $\lambda_2 \neq 0$: The consistency constraint requires p, θ such that $(1 (1 p)I \bar{P})^2 = w^2$
- 4) $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$: Both constraints are active. The implications are the same as for the previous case with the additional (simplifying) condition that either p = 0 or p = 1.

The third case is arguably the most interesting since it frequently occurs for bad estimates \hat{B} , and since it is hard to solve.

2) *MLE of p*: Parameter *p* is the mixture weight of PDF *g* and thus is equal to the probability of observing an anomaly. The MLE of *p* has four possible solution candidates that depend on the cases described above. In case 1, we have the well-known result $p = \frac{1}{n} \sum_{t=1}^{n} \hat{B}_t$, which is the standard unconstrained MLE; In cases 2 and 4, *p* must either be p = 0 or p = 1 to satisfy the first constraint. Case 3 has the following solution:

Theorem 1. For $\lambda_1 = 0$ and $\lambda_2 \neq 0$, the equation system (11) has the following solution for p

$$p = \frac{1}{n - \Omega} \sum_{t=1}^{n} \hat{B}_t \tag{12}$$

where $\Omega < n - \sum_{t=1}^{n} \hat{B}_t$ is the density surplus gradient, which is given by

$$\Omega = \frac{\frac{\partial}{\partial \theta^{(1)}} \log \mathcal{L}}{\frac{\partial}{\partial \theta^{(1)}} \log I} I = \dots = \frac{\frac{\partial}{\partial \theta^{(q)}} \log \mathcal{L}}{\frac{\partial}{\partial \theta^{(q)}} \log I} I$$
(13)

and where (either of the) derivatives of the right-hand side are evaluated at an MLE of θ .

Thus, the constrained solution for p is similar to the unconstrained MLE, with the only difference being the subtraction of Ω in the denominator. Density surplus gradient Ω is 0 in the unconstrained case; positive if the p_{MLE} is too "small"; and negative if p_{MLE} is too "large" to fulfil the consistency constraint.

3) MLE of θ under the Gaussian assumption: Since f_N can be arbitrary, θ has a different maximum likelihood solution for most distributions. In the following, we discuss the MLE for Gaussian f_N with $\theta = [\mu, \sigma^2]$ under the assumption that AFR R is an interval [a, b]. Other distributions are discussed in the Appendix.

Theorem 2. For $\lambda_2 \neq 0$, Gaussian $f_N = \mathcal{N}(\mu, \sigma^2)$ and R = [a, b], equation system (11) has the following solution for σ^2

$$\sigma^2 = \frac{\beta m}{m + \beta W_r(\gamma)} \tag{14}$$

with

x

$$\bar{x} = \frac{1}{n - \sum_{t=1}^{n} \hat{B}_t} \sum_{t=1}^{n} (1 - \hat{B}_t) x_t$$
(15a)

$${}^{2} = \frac{1}{n - \sum_{t=1}^{n} \hat{B}_{t}} \sum_{t=1}^{n} (1 - \hat{B}_{t}) x_{t}^{2}$$
(15b)

$$m = \frac{1}{2} \left((a - \mu)^2 - (b - \mu)^2 \right)$$
(15c)

$$\alpha = x^2 - \mu \bar{x} + (\mu - \bar{x})a \tag{15d}$$

$$\beta = x^2 - \mu \bar{x} + (\mu - \bar{x})b \tag{15e}$$

$$\gamma = \left(\frac{\alpha - \beta}{\beta^2}\right) m e^{-\frac{m}{\beta}} \tag{15f}$$

and where $W_r(\cdot)$ is the r-Lambert function [27] with $r = -\frac{\alpha}{\beta}e^{-\frac{m}{\beta}}$.

Theorem 2 gives an analytic expression for variance σ^2 if the consistency constraint is active—if it is not, σ^2 is given by the standard Gaussian variance estimator. Since the *r*-Lambert function has up to three real branches, (14) gives up to three solution candidates for σ^2 . To determine the correct solution, we can leverage (13): Only one of the three branches will yield a real-valued σ such that $\Omega_{\mu} = \Omega_{\sigma^2}$.

There is no analytic expression for mean μ because this would require an analytic expression for $\int_a^b e^{-x^2} dx$, which does not exist. However, a careful study of equation system (11) reveals a straightforward scheme for accurately approximating μ : Let $s_B = \sum_{t=1}^n \hat{B}_t$, let $\Phi(\cdot)$ be the standard Gaussian cumulative distribution function, and let $\Phi_{\theta}(x) =$

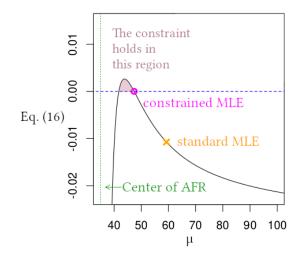


Fig. 2. Illustration of the estimation procedure for μ . Since Eq. (16) is quasi-concave, it is easy to find the constrained MLE, which is located at the function's root that is closer to the unconstrained MLE. The center of the AFR is at $\frac{a+b}{2}$, and there we have $\frac{\partial L}{\partial \mu} = 0$, so Eq. (16) is undefined at $\frac{a+b}{2}$.

 $\Phi\left(\frac{(x-\mu)(m+\beta W_r(\gamma))}{\beta m}\right)$. Theorems 1 and 2 allow us to write the constraint equation (11g) as a function of μ .

$$h(\mu|\boldsymbol{x}, \hat{\boldsymbol{B}}, R) = w^2 \cdot \left(1 - \left(1 - \frac{s_B}{n - \Omega_\mu}\right) \left(\Phi_\theta(b) - \Phi_\theta(a)\right) - \bar{P}\right)$$
(16)

This function is real-valued for $\mu \in (-\infty; \frac{a+b}{2})$ if $\bar{x} < \frac{a+b}{2}$, or else for $\mu \in (\frac{a+b}{2}; \infty)$. Importantly, (16) is quasi-concave in its real domain. Hence, it is easy to find its root(s) using standard optimization procedures, e.g., numerical gradient or binary search. In the case where (16) has two roots, one can determine the better solution using the log-likelihood function ℓ . An illustration is depicted in Figure 2.

IV. IMPLEMENTATION

After presenting the theoretical foundation of AFRs and constrained AD, we now focus on practical aspects. There are many different ways to implement AFR-based AD. Hypothetically speaking, most existing detectors could be equipped with an AFR by leveraging the detector's output to construct estimate \hat{B} , e.g., via Otsu's thresholding method [28], and plug that into Equation (16). However, for complicated detectors, e.g., based on deep learning, it is unclear how the AFR affects the detector, so we instead propose a more direct scheme.

A. Algorithm

In general, constrained AD requires four inputs: i) dataset x; ii) AFR R; iii) anomaly label estimate \hat{B} , iv) significance level α .¹ Estimate \hat{B} can be arbitrary, i.e., it does not have to be a good estimate. We propose to select several \hat{B} randomly, e.g. $n_B = 5$, and compute the average deviation from the mode of f_N over all \hat{B} . Thus, we can provide n_B instead of

a concrete \hat{B} as input. The necessary steps to achieve this are listed in Algorithm 1.

Algorithm 1 Constrained Anomaly MLE (CAMLE)
1: Input $\boldsymbol{x}, R, n_B, \alpha$
2: Use α to compute \overline{P}, w as in Equations (7), (8), (9)
3: Score _t $\leftarrow 0 \forall t \in 1, \dots, n$
4: for $j \in 1, \ldots, n_B$ do
5: Randomly initialize \hat{B}
$6: \qquad \hat{B}_t \leftarrow 0 \forall x_t \in R$
7: $p, \boldsymbol{\theta} \leftarrow \text{standard MLE}(\boldsymbol{x}, \hat{\boldsymbol{B}})$
8: if $(1 - (1 - p) \int_R f_N(x, \theta) dx - \bar{P})^2 > w^2$ then
9: $p, \theta \leftarrow \text{constrained MLE}(\boldsymbol{x}, \hat{\boldsymbol{B}}, R, \bar{P}, w)$
10: end if
11: Score _t \leftarrow Score _t $+ \max_{\nu} f_N(\nu, \theta) - f_N(x, \theta) \forall t$
12: end for
13: $\operatorname{Score}_t \leftarrow \frac{1}{n_B} \operatorname{Score}_t \forall t$
14: Output Score

Conceptually, Algorithm 1 (CAMLE) is comparable to bootstrap aggregation. It draws n_B subsets of data outside the AFR, joins them with data inside the AFR, estimates the parameters, and finally computes an average over all subsets. ²CAMLE has an asymptotic runtime of $\mathcal{O}(n_B \cdot n)$. In most realistic settings, n_B can be set to a small constant, which leads to linear runtime in total.

B. Estimation of AFR

In some settings, a valid AFR is not available and needs to be estimated. Depending on the setting, there are different ways to achieve this. We discuss the following three scenarios.

1) Labels available for normal and anomaly class: In some settings a few labeled data points of both classes are available. Here, a valid AFR can be approximated as any region that contains no labeled anomalies and no unlabeled data points. We recommend selecting a region that contains several labeled normal data points, since this will typically improve the quality of the constrained AD solution due to the estimation of anomaly proportion p being more accurate.

2) Labels available only for normal class: If labeled normal and unlabeled data points are available, but no labeled anomalies, then the AFR can be estimated via a region that exclusively contains labeled normal data. Here it is reasonable to prefer regions that are remote from unlabeled data, if possible.

3) No labels available: In unsupervised settings, no labels are available. Here, there are two options: a) Select an empty region as AFR; b) Include unlabeled data in the AFR. Option a) is safer since it avoids the possibility of falsely including an anomaly in the AFR. However, option b) has higher statistical efficiency and should be preferred if one can make the mild assumption that central, high-density regions of the data space are unlikely to contain anomalies. This is also the approach that we use in the benchmark tests in the following section.

¹Note that α is non-critical, since it only affects the speed of the constrained estimation, but not the accuracy.

V. EXPERIMENTS

The theoretical findings and the algorithm presented in this article motivate us to study pertinent and testable research questions.

- **RQ1. Simulation**: How does the estimation performance of the constrained MLE compare against the unconstrained MLE in a controlled setting?
- **RQ2.1.** Unsupervised benchmark test: How strong is CAMLE's detection performance on popular benchmark datasets where no AFR is known and has to be estimated?
- **RQ2.2**. **Unsupervised sensitivity analysis**: How does the AFR estimation affect CAMLE's detection performance?
 - **RQ3.** Semi-supervised benchmark test: How does CAMLE's detection performance compare against other semi-supervised methods on a benchmark dataset with a valid AFR?

For each research question, a separate experiment is conducted. In the following, motivation, setup, data, baselines and results of each experiment are described separately. A combined discussion of all research questions comes afterward. All data, descriptions, runtimes and source code can be privately accessed online.²

A. RQ1: Simulation

1) Motivation: The constrained estimators presented above are optimal in the AFR-constrained AD setting, but the merit of the constraint itself needs to be assessed separately. The AFR encodes information about anomaly class f_{∞} , yet it is not obvious that this information is useful for estimating the parameters of normal class f_N . Moreover, from a theoretical viewpoint, it is interesting to study scenarios in which the classical MLE is sub-optimal.

2) Setup: We start with a controlled experiment on data drawn from a known distribution. Hence, we use the following simulation procedure: We select a Gaussian distribution for f_N and a 50:50 mixture of two uniform distributions for f_{∞} . We uniformly draw 100 different ground-truth parameter combinations from the following ranges: $\mu \in [-5;5], \sigma \in$ $[0.1; 2], p \in [0.05; 0.95]$. The AFR is then fixed to $[\mu 0.98\sigma; \mu + 0.99\sigma$ and the limits of the f_{∞} mixture are set to $[\mu - 10\sigma; a]$ and $[b; \mu + 10\sigma]$, respectively. Then, for each ground truth combination (μ, σ, p, a, b) , we independently draw 100 datasets consisting of 1000 samples, and 10 different "guessed" label sets $\hat{B} \sim \text{Bernoulli}(p)$ for each dataset. On each resulting dataset, the unconstrained MLE and the constrained MLE of each parameter are first computed using the ground truth labels B, and then using the 10 "gussed" label sets B, respectively. The estimates are compared against the ground truth via the median absolute deviation (MAD), and the average over all datasets is computed. We prefer the MAD over the classical mean-squared error since in rare cases the estimation procedure for the r-Lambert function converges to the wrong branch due to floating point inaccuracy.

3) Results: The results of the simulation are listed in Table I. When the true labels were provided, both constrained and unconstrained MLE performed similarly and achieved small errors. When guessed labels were provided, CAMLE achieved smaller errors than unconstrained MLE for μ and σ , but yielded larger errors for the mixture fraction p.

In conclusion, this experiment suggests that constrained MLE improves upon unconstrained MLE in a suitable controlled setting where all assumptions are fulfilled.

TABLE I
MEDIAN ABSOLUTE DEVIATION BETWEEN GROUND TRUTH PARAMETERS
AND ESTIMATES. A NEGATIVE ERROR DIFFERENCE INDICATES THAT THE
Unconstrained MLE has lower error, whereas a positive error
DIFFERENCE INDICATES THAT THE CONSTRAINED MLE IS BETTER.

		True labels	8	G	uessed lab	els
	μ	σ	p	μ	σ	p
MLE	0.0409	0.0284	0.0134	0.0510	1.6435	0.0426
CAMLE	0.0410	0.0284	0.0134	0.0176	0.6897	0.1314
Err. Diff.	-0.001	0.000	0.000	0.0334	0.9538	-0.0888

B. RQ2.1: Unsupervised Benchmark Test

1) Motivation: Having seen that constrained MLE is promising in a controlled setting, it is natural to study its empirical performance and to see how it compares against widely-used AD methods. In data mining research, it is conventional to conduct such a study on a large variety of diverse AD benchmark datasets. Due to this diversity, we cannot expect that CAMLE's assumptions will be satisfied during the experiment, so the theoretical guarantees will not hold. However, it is interesting to observe how CAMLE behaves in practice–especially when no AFR is known and has to be estimated.

2) Datasets: We use two different groups of datasets, namely development datasets and evaluation datasets. The development datasets are used for preliminary analyses of the approaches, e.g., does the detector encounter an error, how many initial \hat{B} should be sampled, and so on. Evaluation datasets are used for the quantitative comparison. The rationale for using two different groups is to avoid data dredging (see [29]). We aim to ensure that the performance of our method is not specific to the datasets on which we conducted method development.

We use the following publicly available datasets as *development datasets*: GLASS [30]; HEART [31]; HEP [31]; ION [30]; LYMPH [32]; PARK [31]; PIMA [31]; STAMPS [33]; SHUTTLE [34]; WPBC [31]. We select the following datasets as *evaluation datasets*: ANNTHYROID [35]; CARDIO [36]; CARDIOTOCOGRAPHY [36]; LETTER [37]; SATIMAGE; VOW-ELS [38]; WAVEFORM [39]; WILT [40]; YEAST [41]. Further information about these datasets such as number of samples and description of normal and anomaly classes, as well as several other details can be found in [42].

²This article's supplementary material is available at https://github.com/ mtoller/afr

TABLE II ED BENCHMARK TEST. DEDICTED VALUES IN DOWS STARTING WITH DATASET NAME

Method class		Heuristic	al		Probabil	istic	Deep	learning
Data / Measure	LOF	IForest	COPOD	EM	MLE	CAMLE	AutoEnc	DeepSVDD
Annthyroid	0.68	0.82	0.78	0.68	0.68	0.96	0.67	0.72
Cardio	0.51	0.93	0.92	0.91	0.92	0.71	0.95	0.31
Cardiotocography	0.55	0.69	0.66	0.64	0.70	0.68	0.75	0.53
Letter	0.89	0.63	0.56	0.52	0.53	0.56	0.52	0.53
Satimage	0.62	0.99	0.97	0.95	0.95	0.95	0.98	0.39
Vowels	0.87	0.77	0.50	0.55	0.56	0.59	0.60	0.40
Waveform	0.61	0.73	0.73	0.63	0.61	0.52	0.64	0.55
Wilt	0.61	0.45	0.34	0.33	0.35	0.39	0.34	0.47
Yeast	0.47	0.40	0.38	0.40	0.42	0.44	0.40	0.49
ø Rank (per method class)	2.11	1.56	2.33	2.56	1.89	1.56	1.44	1.56
ø Rank (overall)	4.11	2.56	4.56	6.11	4.67	4.11	4.22	5.67
Total runtime (seconds)	0.53	8.81	0.74	0.17	0.70	6.05	435.27	62.28

3) Compared Detectors: The following heuristic detectors from the AD literature are compared (presented in chronolog-ical order):

- LOF: Local outlier factor—a local neighborhood and density-based anomaly detector [2]. Five nearest neighbors and Minkowski distance are used.
- IForest: Isolation Forest, which uses random hyperrectangles to detect isolated data [4]. 1000 random trees are used in each computation.
- 3) *COPOD*: Copula-based outlier detection—a heuristic method to determine data extremeness via tail probabilities [43]. This method has no hyperparameters.³

The following probabilistic approaches are included in the comparison:

- 1) *EM*: Expectation maximization with a Gaussian and a uniform component, and with discrete labels. This particular setup of expectation maximization is conceptually similar to an unconstrained version of our method where the labels are updated in a greedy classical MLE sense.
- 2) *MLE*: The same as Alg. 1, except that it always returns the unconstrained solution. Scores are computed for each dimension and then aggregated via the average over all dimensions. $n_B = 5$ random initializations are used. The AFR [a; b] is inferred from the 0.24 and 0.75 quantiles, respectively. Significance level α is set to 0.05.
- 3) *CAMLE*: Our method as presented in Alg. 1. Scores are computed for each dimension and then aggregated via the average over all dimensions. $n_B = 5$ random initializations are used. The AFR [a; b] is inferred from the 0.24 and 0.75 quantiles, respectively. Significance level α is set to 0.05.

For MLE and CAMLE, we choose the lower boundary of the AFR to be the 0.24 quantile instead of the more common 0.25 quantile to decrease the chance that \bar{x} coincides with $\frac{a+b}{2}$, which has no constrained solution.

The following deep-learning AD methods are compared:

- 1) *AutoEnc*: An autoencoder network [44] whose reconstruction error is used as anomaly score. The hidden layers have $1\rightarrow 64\rightarrow 32\rightarrow 32\rightarrow 64\rightarrow 1$ neurons and use tanh as activation function, and the network was trained using Adam with a dropout rate of 20%.
- Deep-SVDD: SVDD [3] applied to the latent representation of an autoencoder [45]. The hidden and output layers use ReLU and sigmoid activation functions, respectively, and the network was trained using Adam with a dropout rate of 20%.

4) Evaluation Setup: The predictions of every detector are collected from every evaluation dataset. These predictions are then compared with the true anomaly labels using the area under curve of the receiver-operator characteristic (AUC-ROC), which is an estimate of the probability that a detector will rank a randomly chosen anomaly higher than a randomly chosen non-anomaly. It is computed via AUC-ROC = $\int_{-\infty}^{\infty} \text{TPR}(\tau)\text{FPR}'(\tau)d\tau$ where TPR is the true positive rate $\frac{\text{TP}}{\text{TP}+\text{FP}}$, FPR the false positive rate $\frac{\text{FP}}{\text{FP}+\text{TN}}$, and τ is a threshold that separates the scores of anomalies and non-anomalies. We also track each method's total execution runtime to assess the practical efficiency of all implementations.

5) Results: The results of the conducted experiment are shown in Table II. The best average rank among heuristic methods was achieved by IForest; among probabilistic methods by CAMLE; among deep-learning methods by AutoEnc. The best average rank in total was achieved by IForest, with CAMLE and LOF sharing the second place. The unconstrained MLE achieved a worse average rank than the constrained MLE. The fastest method was EM, followed by LOF. The deep-learning based methods AutoEnc and DeepSVDD had the longest and second longest execution runtime, respectively.

Effectively, these results highlight that an AFR-based AD method is capable of outperforming several widely-used AD methods on selected benchmark datasets even when no AFR is known and has to be estimated.

C. RQ2.2: Unsupervised Sensitivity Analysis

1) Motivation: It remains unclear how sensitive CAMLE is to changes in the AFR estimation procedure. In particular, it is

RESULTS OF THE UNSUPERVISED BENCHMARK TEST. DEPICTED VALUES IN ROWS STARTING WITH DATASET NAMES ARE AUC-ROC SCORES.

³Despite its sound theoretical basis, we classify COPOD as heuristic since the algorithm described in [43] does not use a copula.

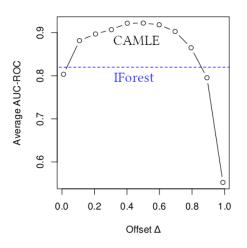


Fig. 3. Results of the sensitivity analysis. CAMLE was computed with 1100 different AFRs per dimension of dataset ANNTHYROID. CAMLE retaints the highest scores among all compared detectors for AFR lengths $0.108 \leq \Delta < 0.892$.

interesting whether CAMLE can retain its high performance across many different AFRs, or if only a small number of specific AFRs produce high results.

2) Setup: For this analysis, we use the ANNTHYROID dataset. For other datasets, please refer to the Appendix. We initialize 1100 different AFRs (per dimension of the dataset) using the following procedure:

- 1) Eleven equidistant offsets Δ are selected from the range [0.01;0.99], i.e, 0.010, 0.108,...,0.892, 0.990.
- 2) For each Δ , 100 equidistant a, b pairs are selected as empirical quantiles, ranging from 0.0001 to 0.999.

We aggregate CAMLE's predictions scores on ANNTHYROID for each Δ using the average over all a, b pairs.

3) Results: The results of this analysis are depicted in Figure 3. CAMLE retained a score higher than the second highest-scoring method (IForest) for $0.108 \leq \Delta < 0.892$. Outside this region, CAMLE's performance dropped below an AUC-ROC of 0.82.

These results suggest that CAMLE is not sensitive to the AFR estimation procedure and that its performance does not drop significantly.

D. RQ3: Semi-Supervised Benchmark Test

1) Motivation: If a valid AFR is given, this is a different setup and needs to be investigated separately. The simulation study alone is not sufficient since it does not cover CAMLE's empirical performance, and RQ2 did not cover the case that an AFR is given.

2) Dataset: To the best of our knowledge, there is no publicly available benchmark dataset with a known AFR. Hence, we collect such a dataset, and refer to it as OFFICE. We collected OFFICE from an office worker's time sheets and interviewed the worker to determine anomaly labels. The worker's personal details as well as all company information were removed to preserve anonymity. OFFICE consists of 365 data points—1 per day for a full year, including weekends with 0 expected working time. Each data point is the number of minutes by which the actual working time deviates from this day's expected working time. AFR R is ± 29 minutes since deviations shorter than 30 minutes are considered normal by the management regardless of what caused them. There are 298 normal data inside R, 54 normal data outside R, and 13 labeled anomalies. We publish this dataset along with this article.

3) Setup: We compare CAMLE with semi-supervised methods PReNet [16] and Overlap [8]. For both semisupervised methods, we re-use the original implementations. To analyse different setups, PReNet and Overlap receive 10%, 20% and 50% of OFFICE as labeled training data, respectively. CAMLE receives AFR=[a; b] = [-29; 29] as input. We compute the AUC-ROC score for each setting.

4) *Results:* The results of this experiment are depicted in Table III. CAMLE obtains higher scores than PReNet and Overlap at all training sizes. The scores of PReNEt and Overlap increase as the training size increases.

These results suggest that an AFR-based method can outperform semi-supervised methods if a valid AFR is provided.

TABLE III
AUC-ROC SCORES ON OFFICE. AS THE TRAINING SIZE INCREASES,
SEMI-SUPERVISED METHODS PRENET AND OVERLAP ACHIEVE BETTER
SCORES. CAMLE ACHIEVES THE HIGHEST SCORES WITH JUST AFR $[a; b]$
PROVIDED.

Method	PReNet		Overlap	CAMLE
Train. size	10% 20%	50% 10%	20% 50%	AFR
AUC-ROC	0.66 0.92	0.96 0.66	0.66 0.72	0.97

VI. DISCUSSION

A. Implications

1) RQ1: The simulation study confirms that the constrained MLE is superior over the unconstrained MLE at estimating f_N 's parameters μ , σ when a valid AFR is known and $\hat{B} \neq B$. It is evident that the constrained estimation of p in this setting must be inferior to the unconstrained estimation to compensate for the other improved estimates. Otherwise we could obtain a solution with higher likelihood, given the AFR, which is not possible since the classical MLE already maximizes the likelihood.

2) RQ2: The unsupervised benchmark test suggests that CAMLE is superior over its unconstrained variant MLE on the majority of the evaluation datasets. Moreover, CAMLE is capable of outperforming several baseline methods from the literature, even if CAMLE's underlying assumptions are not fulfilled by the benchmark datasets (e.g. no AFR known, non-Gaussianity, high-dimensional datasets). The sensitivity analysis confirms that a reasonable AFR can be estimated if none is given, which is useful for practical data mining. 3) RQ3:: The semi-supervised benchmark test suggests that AFRs are more efficient than training labels at encoding information about the anomaly class. The semi-supervised competitors PReNet and Overlap seem to require a large number of training labels to become competitive with CAMLE. It seems likely that these methods would eventually outperform CAMLE if the training size is increased beyond 50%. However, in a real-world AD task, it is not realistic that so many labeled training instances are available.

B. Limitations

The results of our experiments are limited to the small selection of included datasets, as well as to the considered parameter ranges. It was assumed that all data within a dataset are independent and identically distributed. Dependent data types such as *time series*, *graphs*, *spatial*, *text*, and *images* are not covered. The aggregation-based extension of Alg. 1 to multi-dimensional datasets used for RQ2 assumes that there is no dependency between dimensions. This assumption is violated on many datasets.

VII. CONCLUSION AND OUTLOOK

We introduced the novel concept of anomaly-free regions (AFR) to improve anomaly detection. AFRs and the associated constraints are a fundamental addition to classical anomaly detection. AFRs allow one to encode available information about the anomaly class, and can be soundly estimated if no such information is available. The resulting constrained form of anomaly detection improves upon the classical unconstrained anomaly detection, and hence is a promising direction for a new class of detectors, e.g., constrained one-class classifiers.

Despite this paper's focus on anomaly detection, we believe the theoretical contributions will have an impact on other research fields in the future. For example, our theoretical contributions can be applied in classification, as the AFRconcept is equivalent to a region that does not contain one (out of several) classes. The same generalization can be made for other tasks such as clustering and novelty detection.

REFERENCES

- [1] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021.
- [2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.
- [3] D. M. Tax and R. P. Duin, "Support vector data description," Machine learning, vol. 54, no. 1, pp. 45–66, 2004.
- [4] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in 2008 eighth ieee international conference on data mining. IEEE, 2008, pp. 413–422.
- [5] M. B. Toller, B. C. Geiger, and R. Kern, "Cluster purging: Efficient outlier detection based on rate-distortion theory," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [6] H. Zhang, L. Cao, P. VanNostrand, S. Madden, and E. A. Rundensteiner, "Elite: Robust deep anomaly detection with meta gradient," in *Proceed*ings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 2174–2182.
- [7] G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 353–362.

- [8] M. Jiang, S. Han, and H. Huang, "Anomaly detection with score distribution discrimination," arXiv preprint arXiv:2306.14403, 2023.
- [9] L. Perini, V. Vercruyssen, and J. Davis, "Learning from positive and unlabeled multi-instance bags in anomaly detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 1897–1906.
- [10] Y. Wei, Y. Li, M. Xu, and W. Huang, "A review of early fault diagnosis approaches and their applications in rotating machinery," *Entropy*, vol. 21, no. 4, p. 409, 2019.
- [11] Y. Sha, S. Gou, J. Faber, B. Liu, W. Li, S. Schramm, H. Stoecker, T. Steckenreiter, D. Vnucec, N. Wetzstein *et al.*, "Regional-local adversarially learned one-class classifier anomalous sound detection in global long-term space," in *Proceedings of the 28th ACM SIGKDD Conference* on Knowledge Discovery and Data Mining, 2022, pp. 3858–3868.
- [12] A. H. Ahmed, M. A. Riegler, S. A. Hicks, and A. Elmokashfi, "Rcad: Real-time collaborative anomaly detection system for mobile broadband networks," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2682–2691.
- [13] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," *Applied Energy*, vol. 287, p. 116601, 2021.
- [14] H. Huang, L. Yang, Y. Wang, X. Xu, and Y. Lu, "Digital twin-driven online anomaly detection for an automation system based on edge intelligence," *Journal of Manufacturing Systems*, vol. 59, pp. 138–150, 2021.
- [15] A. Degirmenci and O. Karal, "Efficient density and cluster based incremental outlier detection in data streams," *Information Sciences*, vol. 607, pp. 901–920, 2022.
- [16] G. Pang, C. Shen, H. Jin, and A. van den Hengel, "Deep weaklysupervised anomaly detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 1795–1807.
- [17] C.-H. Chang, J. Yoon, S. Ö. Arik, M. Udell, and T. Pfister, "Dataefficient and interpretable tabular anomaly detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 190–201.
- [18] H. Ju, D. Lee, J. Hwang, J. Namkung, and H. Yu, "Pumad: Pu metric learning for anomaly detection," *Information Sciences*, vol. 523, pp. 167–183, 2020.
- [19] Y. Ma and W. Yao, "Flexible estimation of a semiparametric twocomponent mixture model with one parametric component," *Electronic Journal of Statistics*, vol. 9, pp. 447–474, 2015.
- [20] D. Al Mohamad and A. Boumahdaf, "Semiparametric two-component mixture models when one component is defined through linear constraints," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 795–830, 2017.
- [21] S. Song, D. L. Nicolae, and J. Song, "Estimating the mixing proportion in a semiparametric mixture model," *Computational statistics & data analysis*, vol. 54, no. 10, pp. 2276–2283, 2010.
- [22] Z. Shen, M. Levine, and Z. Shang, "An mm algorithm for estimation of a two component semiparametric density mixture with a known component," *Electronic Journal of Statistics*, vol. 12, no. 1, pp. 1181– 1209, 2018.
- [23] B. Shokirov, "On a problem connected with mixture parameter estimation," *Informacni Bulleten Ceske statisticke spolecnosti*, vol. 22, pp. 95–101, 2010.
- [24] R. K. Patra and B. Sen, "Estimation of a two-component mixture model with applications to multiple testing," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 4, pp. 869–893, 2016.
- [25] A. Bhattacharya, S. Varambally, A. Bagchi, and S. Bedathur, "Fast oneclass classification using class boundary-preserving random projections," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 66–74.
- [26] S. M. Robinson, "A quadratically-convergent algorithm for general nonlinear programming problems," *Mathematical programming*, vol. 3, pp. 145–156, 1972.
- [27] I. Mező and Á. Baricz, "On the generalization of the lambert w function," *Transactions of the American Mathematical Society*, vol. 369, no. 11, pp. 7917–7934, 2017.
- [28] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

- [29] C. Andrade, "Harking, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices," *The Journal of clinical psychiatry*, vol. 82, no. 1, p. 25941, 2021.
 [30] F. Keller, E. Muller, and K. Bohm, "Hics: High contrast subspaces
- [30] F. Keller, E. Muller, and K. Bohm, "Hics: High contrast subspaces for density-based outlier ranking," in 2012 IEEE 28th international conference on data engineering. IEEE, 2012, pp. 1037–1048.
- [31] D. Dua and C. Graff, "Uci machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
- [32] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, pp. 157–166.
- [33] B. Micenková, J. van Beusekom, and F. Shafait, "Stamp verification for automated document authentication," in *Computational Forensics: 5th International Workshop, IWCF 2012, Tsukuba, Japan, November 11,* 2012 and 6th International Workshop, IWCF 2014, Stockholm, Sweden, August 24, 2014, Revised Selected Papers. Springer, 2015, pp. 117–129.
- [34] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13. Springer, 2009, pp. 813–822.
- [35] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.
- [36] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de Sa, and L. Pereira-Leite, "Sisporto 2.0: a program for automated analysis of cardiotocograms," *Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311–318, 2000.
- [37] P. W. Frey and D. J. Slate, "Letter recognition using holland-style adaptive classifiers," *Machine learning*, vol. 6, pp. 161–182, 1991.
- [38] M. Kudo, J. Toyama, and M. Shimbo, "Multidimensional curve classification using passing-through regions," *Pattern Recognition Letters*, vol. 20, no. 11-13, pp. 1103–1111, 1999.
- [39] W.-Y. Loh, "Classification and regression trees," Wiley interdisciplinary reviews: data mining and knowledge discovery, vol. 1, no. 1, pp. 14–23, 2011.
- [40] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data mining and knowledge discovery*, vol. 30, pp. 891–927, 2016.
- [41] P. Horton and K. Nakai, "A probabilistic classification system for predicting the cellular localization sites of proteins." in *Ismb*, vol. 4, 1996, pp. 109–115.
- [42] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "Adbench: Anomaly detection benchmark," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32142–32159, 2022.
- [43] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "Copod: copulabased outlier detection," in 2020 IEEE International Conference on Data Mining (ICDM). IEEE, 2020, pp. 1118–1123.
- [44] Y. Le Cun and F. Fogelman-Soulié, "Modèles connexionnistes de l'apprentissage," *Intellectica*, vol. 2, no. 1, pp. 114–143, 1987.
- [45] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

APPENDIX

A. Proof of Theorem 1

Let us restate Theorem 1.

Theorem 1. For $\lambda_1 = 0$ and $\lambda_2 \neq 0$, the equation system (11) has the following solution for p

$$p = \frac{1}{n - \Omega} \sum_{t=1}^{n} \hat{B}_t \tag{12}$$

where $\Omega < n - \sum_{t=1}^{n} \hat{B}_t$ is the density surplus gradient, which is given by

$$\Omega = \frac{\frac{\partial}{\partial \theta^{(1)}} \log \mathcal{L}}{\frac{\partial}{\partial \theta^{(1)}} \log I} I = \dots = \frac{\frac{\partial}{\partial \theta^{(q)}} \log \mathcal{L}}{\frac{\partial}{\partial \theta^{(q)}} \log I} I$$
(13)

and where (either of the) derivatives of the right-hand side are evaluated at an MLE of θ .

Proof. Since $\lambda_2 \neq 0$ we can solve (11c) for λ_2 which gives

$$\lambda_2 = -\frac{\partial \mathcal{L}}{\partial \theta^{(1)}} \Big/ \left(2(1-p)\left(1-(1-p)I-\bar{P}\right) \frac{\partial}{\partial \theta^{(1)}}I \right).$$
(17)

Plugging (17) into (11b) gives

$$\frac{\partial \mathcal{L}}{\partial p} + 2I \left(1 - (1 - p)I - \bar{P} \right) \left(\frac{\partial \mathcal{L}}{\partial \theta^{(1)}} \right) \Big/$$
$$\left(2(1 - p) \left(1 - (1 - p)I - \bar{P} \right) \frac{\partial}{\partial \theta^{(1)}} I \right) = 0$$

since $\lambda_1 = 0$. This can be further simplified to

$$\frac{\partial \mathcal{L}}{\partial p} + \frac{\partial \mathcal{L}}{\partial \theta^{(1)}} \frac{1}{1 - p} \frac{I}{\frac{\partial}{\partial \theta^{(1)}} I} = 0.$$
(18)

Now the rule $\frac{d}{dx}f(x) = f(x)\frac{d}{dx}\log f(x)$ and the shorthand notation $\log \mathcal{L} = \ell$ allows to combine (13) and (18) as

$$\frac{\partial \ell}{\partial p} + \frac{1}{1-p} \frac{\partial \ell}{\partial \theta^{(1)}} \frac{1}{\frac{\partial}{\partial \theta^{(1)}} \log I} = \frac{\partial \ell}{\partial p} + \frac{\Omega}{1-p} = 0.$$

Deriving the log likelihood with respect to p gives

$$\frac{\sum_{t=1}^{n} \hat{B}_{t}}{p} - \frac{n - \sum_{t=1}^{n} \hat{B}_{t}}{1 - p} + \frac{\Omega}{1 - p} = 0$$

which can be solved for p to obtain (12). Note that $\theta^{(1)}$ was chosen arbitrarily, and that the same p must be obtained for all $\theta^{(j)}$, $j \in 1, ..., q$. Hence, (13) must also hold.

B. Proof of Theorem 2

Let us restate Theorem 2.

Theorem 2. For $\lambda_2 \neq 0$, Gaussian $f_N = \mathcal{N}(\mu, \sigma^2)$ and R = [a, b], equation system (11) has the following solution for σ^2

$$\sigma^2 = \frac{\beta m}{m + \beta W_r(\gamma)} \tag{14}$$

Proof. In the following we abbreviate $e_a = e^{-\frac{(a-\mu)^2}{2\sigma^2}}$, $e_b = e^{-\frac{(b-\mu)^2}{2\sigma^2}}$, $C_a = a - \mu$ and $C_b = b - \mu$. We start by stating the derivatives of ℓ and I with respect to μ and σ :

$$\frac{\partial \ell}{\partial \mu} = \sum_{t=1}^{n} (1 - \hat{B}_t) \frac{x_t - \mu}{\sigma^2}$$

$$\frac{\partial \ell}{\partial \sigma} = \sum_{t=1}^{n} (1 - \hat{B}_t) \frac{(x_t - \mu)^2 - \sigma^2}{\sigma^3}$$

$$\frac{\partial I}{\partial \mu} = \int_a^b \frac{\partial}{\partial \mu} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma}} (e_a - e_b)$$

$$\frac{\partial I}{\partial \sigma} = \int_a^b \frac{\partial}{\partial \sigma} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} (C_a e_a - C_b e_b)$$
(19)

Plugging (19) into (13) and dividing both sides by I gives

$$\sum_{t=1}^{n} (1 - \hat{B}_t) \frac{x_t - \mu}{\sigma^2} \Big/ \left(\frac{1}{\sqrt{2\pi\sigma}} \left(e_a - e_b \right) \right)$$
$$= \sum_{t=1}^{n} (1 - \hat{B}_t) \frac{(x_t - \mu)^2 - \sigma^2}{\sigma^3} \Big/ \left(\frac{1}{\sqrt{2\pi\sigma^2}} \left(C_a e_a - C_b e_b \right) \right).$$

By further abbreviating $s_B = \sum_{t=1}^n \hat{B}_t$, $s_x = \sum_{t=1}^n (1 - \hat{B}_t)x_t$, $s_{x^2} = \sum_{t=1}^n (1 - \hat{B}_t)x_t^2$ and eliminating duplicate terms, this statement can be simplified to

$$\frac{s_x - (n - s_B)\mu}{e_a - e_b} = \frac{s_{x^2} - 2\mu s_x + (n - s_B)(\mu^2 - \sigma^2)}{C_a e_a - C_b e_b}$$
$$\frac{C_a e_a - C_b e_b}{e_a - e_b} = \frac{s_{x^2} - 2\mu s_x + (n - s_B)(\mu^2 - \sigma^2)}{s_x - (n - s_B)\mu}.$$
 (20)

The left-hand side of (20) can be rewritten as

$$\frac{C_a e_a - C_b e_b}{e_a - e_b} = \frac{a e_a - b e_b}{e_a - e_b} - \mu = \mathcal{E} - \mu.$$

Inserting this into (20) and multiplying with the right-hand side denominator gives

$$(\mathcal{E} - \mu) \left(s_x - (n - s_B) \mu \right) = s_{x^2} - 2\mu s_x + (n - s_B)(\mu^2 - \sigma^2).$$
(21)

If we study (21) closely, it becomes apparent that dividing both sides by $(n - s_B)$ converts several terms to the averages defined in (15a) and (15b)

$$(\mathcal{E} - \mu)(\bar{x} - \mu) = \bar{x^2} - 2\mu\bar{x} + \mu^2 - \sigma^2$$

$$\sigma^2 = \bar{x^2} - \mu\bar{x} + (\mu - \bar{x})\mathcal{E}.$$
 (22)

Note that (22) coincides with the unconstrained Gaussian MLE if $\mu = \bar{x}$ or $\mathcal{E} = 0$. While (22) is easy to understand, it is not a solution for σ^2 since the term \mathcal{E} contains σ^2 . In the following steps, we rearrange (22) to the standard form for the *r*-Lampert function given in [27].

We simplify \mathcal{E} using (15c):

$$\mathcal{E} = \frac{ae_a - be_b}{e_a - e_b} = b + \frac{(a - b)e_a}{e_a - e_b} = b + \frac{a - b}{1 - e_b/e_a} = b + \frac{a - b}{1 - e_b^{\frac{\pi}{\sigma}}}$$
(23)

Now, let $z = \frac{1}{\sigma^2}$ and plugging (15e) and (23) into (22) gives

$$\frac{1}{z} = \beta + (\mu - \bar{x}) \frac{a - b}{1 - e^{mz}}.$$

Multiplying with the denominators and plugging in (15d) gives

$$1 - e^{mz} = z \left(\alpha - \beta e^{mz}\right)$$

$$1 - \alpha z = e^{mz} \left(1 - \beta z\right)$$

$$\frac{\alpha}{\beta} \left(z - \frac{1}{\alpha}\right) = e^{mz} \left(z - \frac{1}{\beta}\right)$$

$$\frac{\alpha}{\beta} = e^{mz} \left(z - \frac{1}{\beta}\right) / \left(z - \frac{1}{\alpha}\right) \qquad (24)$$

Equation (24) matches exactly the form of [27, Theorem 3] and thus we have

$$z = \frac{1}{\beta} + \frac{1}{m} W_{-\frac{\alpha}{\beta} e^{-\frac{m}{\beta}}} \left(m \left(\frac{\alpha - \beta}{\beta^2} \right) e^{-\frac{m}{\beta}} \right).$$

Lastly, we replace z by $\frac{1}{\sigma^2}$, and multiply with βm , and the proof is complete.

In the main matter, we focused our theoretical exposition of the estimation of f_N 's parameters on the Gaussian distribution. Here, we describe the (somewhat disappointing) solution for uniform f_N .

Proposition 1. For $\lambda_1 \neq 0, \lambda_2 \neq 0$, Uniform $f_N = U(k, l)$ and R = [a, b], equation system (11) has no constrained solution. The constraint is always fulfilled.

Proof. We conduct a proof by contradiction. Assume that $0 s.t. <math>\Omega_k = \Omega_l$ as per Thm 1. Uniform f_N is given by

$$\mathcal{E}_{N}(x,k,l) = \begin{cases} rac{1}{l-k} & k \leq x \leq l \\ 0 & else \end{cases}$$

Assume without loss of generality that $k \leq x \leq l$. The quantities in Ω_l are given by

$$\frac{\partial \ell}{\partial l} = \frac{n - s_B}{l - k}$$
$$I = \frac{b - a}{l - k}$$
$$\frac{\partial I}{\partial l} = -\frac{b - a}{(l - k)^2}$$

and hence, we have

$$\Omega_l = \frac{\frac{\partial \ell}{\partial l}}{\frac{\partial I}{\partial l}}I = s_B - n.$$

Plugging this into (12) gives

1

$$p = \frac{s_B}{n - \Omega_l} = \frac{s_B}{n - n + s_B} = 1$$

which violates the assumption and completes the desired contradiction. $\hfill \Box$

For the sake of completeness, we report the results of all methods studied in RQ2.1 on the development datasets. We did not report these results in the main matter since we used these datasets to test various setups of our method. However, we believe that the results might nevertheless be interesting to some readers. The overall ranking of methods on the development datasets is slightly, but not considerable different from the ranking on the evaluation datasets. The most notable difference is the improved performance of COPOD. This was expected, since the development datasets include several datasets that were studied in the original COPOD publication [43].

For the sake of completeness, we conduct the same experiment as in RQ2.2 for all other evaluation datasets. The results of this experiment are depicted in Figure 4. In this figure, the dashed blue line indicates CAMLE's performance on this dataset using the default parametrization as reported in Table II. These results indicate high stability of the method with respect to the estimated AFR. Please note the small variance of the results.

TABLE IV

RESULTS FOR RQ2 ON THE DEVELOPMENT DATASET. DEPICTED VALUES IN ROWS STARTING WITH DATASET NAMES ARE AUC-ROC SCORES. WE DID NOT REPORT THESE RESULTS IN THE MAIN MATTER SINCE WE USED THESE DATASETS TO TEST VARIOUS SETUPS OF OUR METHOD. THE OVERALL RANKING OF METHODS ON THE DEVELOPMENT DATASETS IS SLIGHTLY, BUT NOT CONSIDERABLE DIFFERENT FROM THE RANKING ON THE EVALUATION DATASETS. THE MOST NOTABLE DIFFERENCE IS THE IMPROVED PERFORMANCE OF COPOD. THIS WAS EXPECTED, SINCE THE DEVELOPMENT DATASETS INCLUDE SEVERAL DATASETS THAT WERE STUDIED IN THE ORIGINAL COPOD PUBLICATION [43].

Method class		Heuristic	ristical		Probabilistic			Deep learning	
Data / Measure	LOF	IForest	COPOD	EM	MLE	CAMLE	AutoEnc	DeepSVDD	
Glass	0.64	0.79	0.76	0.75	0.72	0.76	0.73	0.68	
Heart	0.55	0.60	0.65	0.73	0.73	0.74	0.44	0.53	
Нер	0.49	0.69	0.80	0.37	0.36	0.40	0.76	0.56	
Ion	0.90	0.85	0.79	0.77	0.76	0.77	0.83	0.61	
Lymph	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.83	
Park	0.47	0.50	0.51	0.48	0.42	0.42	0.25	0.61	
Pima	0.59	0.67	0.65	0.59	0.56	0.56	0.63	0.46	
Stamps	0.51	0.90	0.93	0.93	0.92	0.94	0.90	0.55	
Shuttle	0.64	0.87	0.82	0.90	0.91	0.74	0.94	0.67	
WPBC	0.46	0.50	0.52	0.49	0.48	0.51	0.44	0.51	
ø Rank (per method class)	2.8	1.6	1.5	1.6	2.5	1.8	1.3	1.7	
ø Rank (overall)	5.9	3.0	2.5	4.0	5.2	4.5	4.6	5.9	
Total runtime (seconds)	0.01	0.74	0.00	0.01	0.02	0.46	12.85	2.84	

A visualization of the newly collected Office dataset is depicted in Figure 5. An anonymous worker's daily work time deviation has natural fluctuation due to flexible work time. Some deviations are anomalous, e.g., because the worker forget to log out when leaving the company. The region [-29; 29] contains no anomalies and is a valid AFR derived from domain knowledge.

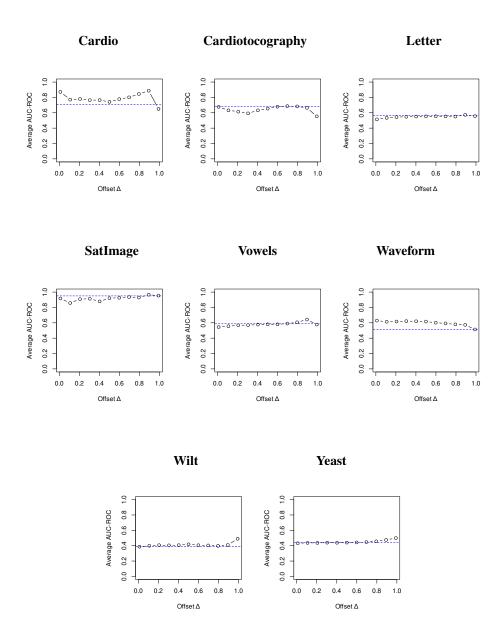


Fig. 4. Sensitivity analysis for all other evaluation datasets. The dashed blue line indicates CAMLE's performance on this dataset using the default parametrization as reported in Table II. These results indicate high stability of the method with respect to the estimated AFR. Please note the small variance of the results.

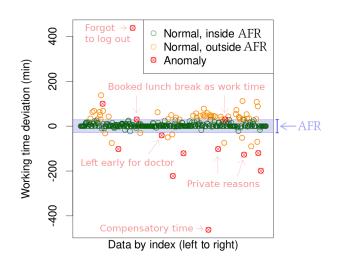


Fig. 5. Visualization of the OFFICE dataset. An anonymous worker's daily work time deviation has natural fluctuation due to flexible work time. Some deviations are anomalous, e.g., because the worker forget to log out when leaving the company. The region [-29; 29] contains no anomalies and is a valid AFR derived from domain knowledge.