# Information Theoretic Novelty Detection[☆]

Maurizio Filippone[*,a], Guido Sanguinetti[a,b]

[a] *Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street Sheffield, S1 4DP - United Kingdom.*
[b] *Department of Chemical and Process Engineering, University of Sheffield, Mappin Street Sheffield S1 3JD - United Kingdom.*

**Abstract**

We present a novel approach to online change detection problems when the training sample size is small. The proposed approach is based on estimating the expected information content of a new data point and allows an accurate control of the false positive rate even for small data sets. In the case of the Gaussian distribution, our approach is analytically tractable and closely related to classical statistical tests. We then propose an approximation scheme to extend our approach to the case of the mixture of Gaussians. We evaluate extensively our approach on synthetic data and on three real benchmark data sets. The experimental validation shows that our method maintains a good overall accuracy, but significantly improves the control over the false positive rate.

*Key words:* Novelty Detection, Information Theory, Mixture of Gaussians, Density Estimation

## 1. Introduction

Novelty detection is an important task in machine learning which plays a prominent role in many application domains, from fault detection [2, 8, 13]

[☆]This work is independent research commissioned by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Health for Health Research, or the Department of Health.

[*]Corresponding Author
*Email addresses:* filippone@dcs.shef.ac.uk (Maurizio Filippone), g.sanguinetti@dcs.shef.ac.uk (Guido Sanguinetti)

to monitoring medical conditions [20, 25]. In its classic formulation, novelty detection is the identification of data that deviate from the norm using only knowledge from normality. The goal of novelty detection is twofold: first of all, one needs to be as accurate as possible in detecting samples which do deviate from normality (*true positives*). At the same time, one needs to be able to predict how many normal examples will be erroneously flagged as positives (*false positives*). Given that by definition true positives are rare, the ability to control the false positive rate is by no means ancillary to the accuracy in predicting true positives. In order to achieve these goals, one needs to estimate some characteristics of the distribution of the normal data. This could be a full estimation of the statistics of the training set distribution or, usually, some weaker form of this, such as estimating quantiles of the distribution. This then allows to fix a threshold for acceptance of new data while having a degree of control over the number of false alarms raised.

Several approaches have been considered to tackle this problem including neural networks [19, 5], extreme value statistic [21], support vector methods [22, 7], frequentist [24] and Bayesian [26] non-parametric approaches (for a good review of statistical approaches for novelty detection see *e.g.* [4, 18]). In general, for reasonably sized training sets, most of these approaches achieve very good performance both in terms of accurately identifying novelty and in terms of containing the number of false positives. However, there are many applications where the size of the training set is very small, or, equivalently, where the user requirements prohibit a long training phase for the system. Typical examples might include monitoring the health conditions of patients or the lifestyle of elderly people, where a system which only requires a short training would clearly be advantageous.

The problem encountered when deploying techniques based on density estimation when the training set is extremely small is that, in general, there is little control over the inherent variability introduced by the sparsity of the training data. In practice, this means that the estimated quantiles can differ substantially from the true quantiles of the distribution. This may not have a dramatic effect on the efficacy of the method in detecting true positives, but it usually results in serious difficulties in accurately predicting the false positive rate.

In this paper, we recast the novelty detection problem in the framework of *information theory*. Our approach focuses on computing the distribution of the *information content* carried by a new data point. Namely, given a distributional assumption for the training data, we compute the *Kullback-*

*Leibler* (KL) divergence between the density estimated with the training set and the density estimated with the training set *and* the test point. The basic idea is that the information content of a new point should naturally depend on the size of the training set, and hence account for the small-sample variability introduced in any estimates. We show that the KL divergence in the univariate and multivariate Gaussian cases does not depend on the estimated statistics of the training distribution, rather it automatically incorporates the finite sample variability in the detection criterion. Consequently, the resulting method is able to control the false positive rate even when the training set is small. We then extend the information theoretic approach to the case of the mixture of univariate and multivariate Gaussians. Such extension allows to deal with more general forms of generating distributions, thus leading to a method that can be employed in a wider range of applications [14]. We present an approximation scheme for the computation of the KL divergence between the density estimated with the training set and the density estimated with the training set and the test point. From there, we introduce an efficient Monte Carlo scheme yielding an algorithm able to control the false positive rate in the case of the mixture of Gaussians.

The paper is organized as follows: in Section 2 we review some basic concepts on statistical testing, in Section 3 we present our method, in Section 4 we show the results on some real and simulated data, and we conclude in Section 5 by discussing the merits and limitations of our approach.

## 2. Statistical Testing

In this section we briefly review some basic concepts from the theory of statistical testing. In particular, we introduce the classic $F$-test for the Gaussian and multivariate Gaussian distribution, which will play an important role in the theoretical underpinning of our approach to novelty detection.

### *2.1. Testing a Univariate Gaussian*

Let $X = \{x_1, \ldots, x_n\}$ be the training set, comprised of i.i.d. points from a univariate Gaussian with mean $m$ and variance $s^2$, i.e., $p(x_i) = \mathcal{N}(x_i|m, s^2)$. A simple but widely used approach is to compute the statistics $\hat{m}$ and $\hat{s}^2$ by maximum likelihood (ML), which results in:

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{m})^2$$

These quantities are the sample mean and the sample variance, and are the ML estimates for the mean and the variance. In this framework, we could consider $p(x|\hat{m}, \hat{s}^2)$ as the predictive pdf, and we could select the rejection regions on its basis. When the set $X$ has low cardinality, however, this is not a good choice, since the values of the statistics can strongly deviate from the true values. In particular, the sample mean is known to be normally distributed as $\hat{m} \sim \mathcal{N}(m, \frac{s^2}{n})$, while the distribution of the sample variance is a chi-square with degrees of freedom equal to the number of training points minus one, $\hat{s}^2 \sim \frac{s^2}{n}\chi^2_{(n-1)}$. We can try to exploit these facts to obtain a better way to identify the rejection regions.

Let's study the distribution of the following random variable:

$$\hat{z}^2 = \frac{(x_* - \hat{m})^2}{\hat{s}^2}$$

Assuming that $x_*$ comes from the same distribution as the training points, we can study the distribution of $\hat{z}^2$. Since $\hat{m} \sim \mathcal{N}(m, \frac{s^2}{n})$, $x_* \sim \mathcal{N}(m, s^2)$, and $x_*$ and $\hat{m}$ are independent, the difference:

$$(x_* - \hat{m}) \sim \mathcal{N}\left(0, s^2\left(1 + \frac{1}{n}\right)\right)$$

Its square is a random variable distributed as:

$$(x_* - \hat{m})^2 \sim s^2\left(1 + \frac{1}{n}\right)\chi^2_{(1)}$$

We also know that:

$$n\hat{s}^2 \sim s^2\chi^2_{(n-1)}$$

Then, dividing the last two expressions, we see that [1]:

$$\hat{z}^2 \sim \left(\frac{n+1}{n-1}\right)F_{(1,n-1)}$$

where $F$ is the *F-distribution*, the distribution of the ratio of two chi-square variates. We can use this result to test if a new point is an outlier by checking that the $\hat{z}^2$ variable, computed for a test point, is below the selected quantile of the $F$ distribution. We note that this test is valid for all the Gaussian distributions, since it does not depend on the estimated statistics. In the

rest of this paper, we will refer to this novelty detection method as the $F$-test. We emphasize here that under the null hypothesis that $x_*$ comes from the same $\mathcal{N}(m, s^2)$, the distribution of $\hat{z}^2$ is known exactly. Therefore, the $F$-test is the optimal test in the sense that allows to set precisely the rejection regions corresponding to the selected false positive rate.

*2.2. Multivariate Gaussian*

We can extend the $F$-test to the multivariate case. Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the training set, where each data point $\mathbf{x}_i \in \mathbb{R}^d$ is drawn from a multivariate $d$-dimensional Gaussian $\mathcal{N}(\mathbf{x}|\mathbf{m}, S)$ with mean $\mathbf{m}$ and covariance matrix $S$. The multivariate counterpart of the $\hat{z}^2$ score is the following:

$$\hat{z}^2 = (\mathbf{x}_* - \hat{\mathbf{m}})^{\mathrm{T}} \hat{S}^{-1} (\mathbf{x}_* - \hat{\mathbf{m}}) \tag{1}$$

In order to find the distribution of this score, we use a result from statistics [1] stating that given $\mathbf{y} \sim \mathcal{N}(0, aS)$, and $A \sim W_\nu(S)$ ($a \in \mathbb{R}$, $S$ a positive semidefinite covariance matrix, $\nu$ the number of degrees of freedom of the Wishart distribution $W$):

$$\mathbf{y}^{\mathrm{T}} A^{-1} \mathbf{y} \sim \frac{ad}{\nu - d + 1} F_{(d, \nu - d + 1)}$$

Since the sample mean $\hat{\mathbf{m}} \sim \mathcal{N}(\mathbf{x}|\mathbf{m}, S/n)$, and $n$ times the sample covariance $n\hat{S} \sim W_{(n-1)}(S)$, we obtain:

$$\hat{z}^2 \sim (n+1) \frac{d}{n-d} F_{(d, n-d)} \tag{2}$$

It is important to notice that this is a univariate test for a multivariate pdf, that leads to a considerable computational advantage. We analyze directly $\hat{z}^2$ for a test point, comparing it with the quantiles of the $F$ distributed random variable. Again, this test is independent from the estimated statistics of the pdf.

## 3. Information Theoretic Measure for Novelty Detection

In this section, we describe the main theoretical material underpinning our approach to the novelty detection problem. Let $X = \{x_1, \ldots, x_n\}$ be the training set and let $x_*$ be a new point to test for anomaly (the generalization to test sets with multiple points is trivial). We assume that the training

set is generated from a probability distribution $p(x|\mathbf{w})$ of known functional form but unknown parameters. As stated before, one could use this parametric assumption to obtain ML estimates $\hat{\mathbf{w}}$ of the parameters and then fix a threshold for acceptance of the new point $x_*$ based on the estimated statistics. We will refer to this method as the ML method; notice that the ML statistics summarize all the information of the training set, and do not explicitly depend on its size.

We instead propose to update the ML estimates of the parameters using the test point and then compute the similarity of the two distributions obtained. The rationale for doing this is that the new estimates $\hat{\mathbf{w}}_*$ will depend explicitly on the old estimates, on the test point *and* on the size of the training set. Intuitively, for small training sets we will expect the information content of a new point, even drawn from the same distribution, to be quite high. As a measure of (dis)similarity between the two distributions we will use the Kullback-Leibler (KL) divergence (see, *e.g.*, [6] Section 1.6)

$$\mathrm{KL}\left[p(x)\|q(x)\right] = \int p(x) \log \left[\frac{p(x)}{q(x)}\right] dx. \tag{3}$$

It can be easily shown that the KL divergence is always non-negative and is zero if and only if the two distributions coincide. However, it is not a metric, since is not symmetric and does not obey to the triangular inequality.

We note here that a similar approach has been proposed for the computational theory of surprise [17], where the KL divergence was used to measure the dissimilarity between prior and posterior distribution in Bayesian inference.

### *3.1. Novelty Detection for Univariate Gaussians*

Let's assume the training data is generated by a Gaussian probability distribution $p(x|m, s^2) = \mathcal{N}(x|m, s^2)$ with mean $m$ and variance $s^2$, and let $\hat{m}$ and $\hat{s}^2$ be their respective ML estimates. When a new point $x_*$ is drawn, the estimates of the mean and the variance can be updated in the following way:

$$\hat{m}_* = \frac{n}{n+1}\hat{m} + \frac{1}{n+1}x_* \qquad \hat{s}_*^2 = \frac{n}{n+1}\hat{s}^2 + \frac{n}{(n+1)^2}(x_* - \hat{m})^2 \tag{4}$$

To evaluate the amount of information carried by the new point $x_*$, we compute the KL divergence between the probability distribution $\mathcal{N}(x|\hat{m}, \hat{s}^2)$ and

the updated one $\mathcal{N}(x|\hat{m}_*, \hat{s}_*^2)$. The KL divergence between two normal distributions with parameters $(m_1, s_1^2)$ and $(m_2, s_2^2)$ is readily obtained from Eq. (3) as [15]

$$\mathrm{KL}(m_1, m_2, s_1^2, s_2^2) = \frac{1}{2}\left[\log\left(\frac{s_2^2}{s_1^2}\right) - 1 + \frac{s_1^2}{s_2^2} + \frac{(m_1 - m_2)^2}{s_2^2}\right]. \tag{5}$$

Inserting the update rule (4) one obtains

$$\mathrm{KL}(y, \hat{z}^2) = \frac{1}{2}\left[\log\left(1 + y + y\hat{z}^2\right) - 2\log\left(1 + y\right) - 1 + \frac{1+y}{1+y+y\hat{z}^2} + y\right] \tag{6}$$

where we have defined $y = \frac{1}{n}$ and $\hat{z}^2 = \frac{(x_* - \hat{m})^2}{\hat{s}^2}$ for brevity. This result is remarkable. From the analysis of the former section, we have seen that $\hat{z}^2$ is distributed as an $F$ variable, depending only on the cardinality of $X$. This means that the information content of a new data point coming from the true distribution, in the univariate Gaussian case, does not depend on the estimated statistics. Therefore, under the null hypothesis that $x_*$ comes from the same distribution as the training set, we know exactly the distribution of $\mathrm{KL}(y, \hat{z}^2)$. Obtaining the quantiles of this distribution will enable us to set thresholds at specific false positive rates. In fact, this procedure corresponds to the test of $\hat{z}^2$ directly, in the sense that they lead to the same results. This is the link between statistical testing and the information theoretic approach in the univariate Gaussian case. In particular, the proposed approach leads to a method able to control the false alarm rate to the desired level, since, as in the $F$-test, it takes into account the variability of the estimated statistics.

We note here that Eguchi and Copas [10] established a connection between Neyman-Pearson lemma and the KL divergence; however, they did not, to our knowledge, analyze the novelty detection scenario where no alternative hypothesis is provided.

### *3.2. Novelty Detection for Multivariate Gaussians*

These results can be extended to the multivariate Gaussian case in $d$ dimensions $\mathcal{N}(\mathbf{x}|\mathbf{m}, S)$. However, the calculations involved, while totally analogous to the univariate case, are somewhat intricate; for presentation's sake we will only outline them here, referring the reader to Ref. [11] for the details. Using the same notation as in the previous section, let

$$\tilde{\mathbf{x}}_* = \mathbf{x}_* - \hat{\mathbf{m}} \qquad\qquad A = \tilde{\mathbf{x}}_* \tilde{\mathbf{x}}_*^{\mathrm{T}}$$

When a new point $\mathbf{x}_*$ is drawn, the updated versions of the mean and the covariance matrix are

$$\hat{\mathbf{m}}_* = \frac{n}{n+1}\hat{\mathbf{m}} + \frac{1}{n+1}\mathbf{x}_* \qquad \hat{S}_* = \frac{n}{n+1}\hat{S} + \frac{n}{(n+1)^2}A \tag{7}$$

where $\hat{\mathbf{m}}$ and $\hat{S}$ are the ML estimates of the mean and covariance of the distribution obtained from the training set. The KL divergence between the training distribution and the updated distribution is computed as the KL divergence between two multivariate Gaussians [15]

$$\mathrm{KL} = \frac{1}{2}\left[\log\left(\det \hat{S}_*\hat{S}^{-1}\right) + \mathrm{tr}\left(\hat{S}_*^{-1}\hat{S}\right) + (\hat{\mathbf{m}} - \hat{\mathbf{m}}_*)^{\mathrm{T}}\hat{S}_*^{-1}(\hat{\mathbf{m}} - \hat{\mathbf{m}}_*) - d\right]. \tag{8}$$

Inserting Eq. 7 into Eq. 8, and using repeatedly the properties of traces and determinants, we obtain, after some algebra, the following expression for the KL divergence

$$\mathrm{KL}(y, \hat{z}) = \frac{1}{2}\left[\log\left(1 + y + y\hat{z}^2\right) - (d+1)\log(1+y) - 1 + \frac{1+y}{1+y+y\hat{z}^2} + yd\right]. \tag{9}$$

Here again we have introduced $y = \frac{1}{n}$ and $\hat{z}^2 = \tilde{\mathbf{x}}_*^{\mathrm{T}}\hat{S}^{-1}\tilde{\mathbf{x}}_*$.

### *3.3. Novelty Detection for Mixtures of Gaussians*

As we have seen, the information theoretic approach for novelty detection is analytically tractable in the case of univariate and multivariate Gaussian distributions. The result is a method which is closely related to the classical $F$-test, which is known to be the optimal test for Gaussian distributions. However, the Gaussian assumption may be often overly restrictive, implying a limited applicability of the method. In this subsection, we present an extension of our approach to the case of the mixture of Gaussians. As any distribution can be in principle approximated as a mixture of Gaussians, this guarantees a much greater flexibility and applicability for the method. The price to pay for this greater flexibility is that the computation of the KL divergence is no longer analytically feasible. We will however introduce an approximation scheme which allows a fairly accurate closed-form solution.

Let us consider a mixture of $c$ Gaussian distributions as the pdf generating the data. For clarity's sake we will start with a mixture of univariate distributions; at the end of this section, we will report the extension to the

multivariate case. We collect a data set $X = \{x_1, \ldots, x_n\}$, with $x_i \sim p(x|\mathbf{w})$, where:

$$p(x|\mathbf{w}) = p(x|\pi_1, \ldots, \pi_c, m_1, \ldots, m_c, s_1^2, \ldots, s_c^2) = \sum_{k=1}^{c} \pi_k \mathcal{N}(x|m_k, s_k^2)$$

In order to apply our method, we would need the ML estimate $\hat{\mathbf{w}}$ of the parameter vector $\mathbf{w}$ that comprises the proportions $\pi_k$, the means $m_k$, and the variances $s_k^2$ of the $c$ components. Then, we would need an expression for $\hat{\mathbf{w}}^*$ that is the updated version of the parameters obtained by computing the ML estimates on the augmented set $X \cup \{x_*\}$. Finally, we would compute the KL divergence between the two pdf parameterized by $\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}^*$.

In order to do that, we need to overcome some computational problems. First, it is known that there is no closed form solution for the ML problem of the mixture of Gaussians. Second, there is no closed form for the KL divergence between two mixture distributions. We propose a way to overcome these limitations by resorting to some approximations; the main steps are summarized here as:

1. computing a ML solution to the model on $X$;
2. updating the parameters performing one EM step on $X \cup \{x_*\}$;
3. writing the first order approximation of $p(x|\hat{\mathbf{w}}^*)$ based on $p(x|\hat{\mathbf{w}})$ and the updated parameters;
4. expand the logarithm in the expression of the KL divergence up to the second order and compute the expectation over $p(x|\hat{\mathbf{w}})$.

(Step 1). To obtain a ML solution for the mixture, we can employ the *Expectation Maximization* (EM) algorithm [6]. The EM algorithm on the set $X$ yields a maximum likelihood estimation $\hat{\mathbf{w}}$ of the parameters via an iterative scheme that maximizes the likelihood. We note here that using EM to solve the model on $X$, while convenient, is not necessary for our method.

(Step 2). The update of the parameters when we add a new test point $x_*$ would require to start from the ML solution, and restart the iteration on the augmented set $X \cup \{x_*\}$. Since this update of the parameters is not in closed form, we consider a single iteration of the EM algorithm on the augmented set $X \cup \{x_*\}$. Under the assumption that augmenting the set leads to a small change in the ML statistics, we can assume that a single EM step might already give a good estimate of the correct statistics. An interesting online method for the update of the parameters can be found in Ref. [23].

The E step will not affect the responsibilities of the points belonging to $X$, while the responsibilities of $x_*$ will be computed as:

$$u_{*k} = \frac{\hat{\pi}_k \mathcal{N}(x_*|\hat{m}_k, \hat{s}_k^2)}{\sum_{r=1}^{c} \hat{\pi}_r \mathcal{N}(x_*|\hat{m}_r, \hat{s}_r^2)} = \frac{\frac{\hat{\pi}_k}{\hat{s}_k} \exp\left(-\frac{(x_*-\hat{m}_k)^2}{2\hat{s}_k^2}\right)}{\sum_r \frac{\hat{\pi}_r}{\hat{s}_r} \exp\left(-\frac{(x_*-\hat{m}_r)^2}{2\hat{s}_r^2}\right)}$$

In the M step the parameters $\hat{\mathbf{w}}$ will be updated into $\hat{\mathbf{w}}^*$. Introducing $n_k = \sum_j u_{jk}$ as the cardinality of the $k$-th component, the update equations for the M step are the following:

$$\hat{m}_k^* = \frac{\hat{m}_k n_k + x_* u_{*k}}{n_k + u_{*k}}$$

$$\hat{s_k^*}^2 = \frac{n_k}{n_k + u_{*k}} \hat{s}_k^2 + \frac{n_k u_{*k}}{(n_k + u_{*k})^2} (x_* - \hat{m}_k)^2$$

$$\hat{\pi}_k^* = \frac{n}{n+1} \hat{\pi}_k + \frac{u_{*k}}{n+1}$$

Let's rewrite the former equations in an incremental form:

$$\hat{m}_k^* = \hat{m}_k + \Delta\hat{m}_k = \hat{m}_k + \frac{(x_* - \hat{m}_k) u_{*k}}{n_k + u_{*k}}$$

$$\hat{s_k^*}^2 = \hat{s}_k^2 + \Delta\hat{s}_k^2 = \hat{s}_k^2 + \frac{n_k u_{*k}}{(n_k + u_{*k})^2} (x_* - \hat{m}_k)^2 - \frac{\hat{s}_k^2 u_{*k}}{(n_k + u_{*k})}$$

$$\hat{\pi}_k^* = \hat{\pi}_k + \Delta\hat{\pi}_k = \hat{\pi}_k + \frac{u_{*k} - \hat{\pi}_k}{n+1}$$

(Step 3). We are interested in the KL divergence between $p(x|\hat{\mathbf{w}})$ and $p(x|\hat{\mathbf{w}}^*)$:

$$\mathrm{KL}[p(x|\hat{\mathbf{w}}) \| p(x|\hat{\mathbf{w}}^*)] = \int p(x|\hat{\mathbf{w}}) \log \left[\frac{p(x|\hat{\mathbf{w}})}{p(x|\hat{\mathbf{w}}^*)}\right] dx \tag{10}$$

We compute a first order approximation of $p(x|\hat{\mathbf{w}}^*) = \sum_{k=1}^{c} \hat{\pi}_k^* \mathcal{N}(x|\hat{m}_k^*, \hat{s_k^*}^2)$. First of all:

$$\hat{\pi}_k^* \mathcal{N}(x|\hat{m}_k^*, \hat{s_k^*}^2) \simeq \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2) + \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2) \Delta\psi_k$$

where:

$$\Delta\psi_k = \left[\Delta\hat{m}_k \frac{(x - \hat{m}_k)}{\hat{s}_k^2} + \frac{\Delta\hat{s}_k^2}{2} \left(\frac{(x - \hat{m}_k)^2}{\hat{s}_k^4} - \frac{1}{\hat{s}_k^2}\right) + \frac{\Delta\hat{\pi}_k}{\hat{\pi}_k}\right]$$

Thus:

$$p(x|\hat{\mathbf{w}}^*) \simeq p(x|\hat{\mathbf{w}}) + \sum_k \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2)\Delta\psi_k$$

Now let's compute the KL divergence $\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)]$. First of all:

$$\log\left(\frac{p(x|\hat{\mathbf{w}})}{p(x|\hat{\mathbf{w}}^*)}\right) = -\log\left(1 + \frac{\sum_k \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2)\Delta\psi_k}{p(x|\hat{\mathbf{w}})}\right)$$

(Step 4). Expanding the logarithm up to the second order, we obtain:

$$-\log\left(1 + \frac{\sum_k \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2)\Delta\psi_k}{p(x|\hat{\mathbf{w}})}\right) \simeq -\frac{\sum_k \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2)\Delta\psi_k}{p(x|\hat{\mathbf{w}})} + \\ + \frac{1}{2}\frac{\left[\sum_k \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2)\Delta\psi_k\right]^2}{[p(x|\hat{\mathbf{w}})]^2}$$

The expectation of the first term of the former equation over $p(x|\hat{\mathbf{w}})$ turns out to be zero, leading to the following result:

$$\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)] \simeq \frac{1}{2}\int \frac{\left[\sum_k \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2)\Delta\psi_k\right]^2}{p(x|\hat{\mathbf{w}})} dx \tag{11}$$

We can rewrite the former equation:

$$\frac{\left[\sum_k \hat{\pi}_k \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2)\Delta\psi_k\right]^2}{p(x|\hat{\mathbf{w}})} = \sum_r \sum_j \hat{\pi}_r \mathcal{N}(x|\hat{m}_r, \hat{s}_r^2) u_j \Delta\psi_r \Delta\psi_j$$

where $u_j$ is the responsibility function of $x$ for the class $j$. Here we make a further approximation by neglecting the terms when $j \neq r$. In other words, we replace $u_j \mathcal{N}(x|\hat{m}_r, \hat{s}_r^2)$ with zero when $j \neq r$, and with $\mathcal{N}(x|\hat{m}_r, \hat{s}_r^2)$ when $j = r$. This leads to:

$$\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)] \simeq \frac{1}{2}\sum_k \hat{\pi}_k \int \mathcal{N}(x|\hat{m}_k, \hat{s}_k^2)\Delta\psi_k^2 dx$$

The integral can be computed easily, yielding:

$$\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)] \simeq \frac{1}{2}\sum_k \hat{\pi}_k \left[\frac{\Delta\hat{m}_k^2}{\hat{s}_k^2} + \frac{(\Delta\hat{s}_k^2)^2}{2\hat{s}_k^4} + \frac{\Delta\hat{\pi}_k^2}{\hat{\pi}_k^2}\right]$$

This is a closed form approximation of the KL divergence, and is a function of $x_*$. We define:

$$\hat{z}_k^2 = \frac{(x_* - \hat{m}_k)^2}{\hat{s}_k^2} \qquad n_k^* = n_k + u_{*k}$$

We can rewrite the former equation by substituting the expressions of the updates of the parameters and rearranging the terms, thus obtaining:

$$\begin{aligned} \mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)] \simeq \frac{1}{4}\sum_k \hat{\pi}_k \frac{u_{*k}^2}{(n_k^*)^4}\left[n_k^2\hat{z}_k^4 + 2u_{*k}(n_k^*)\hat{z}_k^2 + (n_k^*)^2\right] \\ + \frac{1}{2}\sum_k \frac{(u_{*k} - \hat{\pi}_k)^2}{\hat{\pi}_k(n+1)^2} \end{aligned} \tag{12}$$

We can also rewrite $u_{*k}$ in terms of $\hat{z}_k^2$:

$$u_{*k} = \frac{\frac{\hat{\pi}_k}{\hat{s}_k}\exp\left(-\frac{\hat{z}_k^2}{2}\right)}{\sum_r \frac{\hat{\pi}_r}{\hat{s}_r}\exp\left(-\frac{\hat{z}_r^2}{2}\right)}$$

that depends from the proportions and the variances. Therefore, the approximated expression of the KL divergence depends on the $\hat{z}_k^2$ variables as well as the ML estimates of the proportions $\hat{\pi}_k$ and the variances $\hat{s}_k^2$:

$$\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)] \simeq f(\hat{z}_1, \ldots, \hat{z}_c, \hat{\pi}_1, \ldots, \hat{\pi}_c, \hat{s}_1^2, \ldots, \hat{s}_c^2)$$

Unfortunately, the dependence from the ML estimates of the proportions and the variances remains, and will result in a deviation from the expected results.

We evaluate the quality of the proposed approximation by means of an illustrative example. In Fig. 1 (left panel), we show the pdf generating the data composed by three Gaussian. From that, we sample 100 data points and we plot the KL divergence obtained by adding a test point corresponding to the value on the $x_*$-axis. The curves in the right panel show the quality of the approximations. The solid line represents the exact value of $\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)]$ computed by integrating numerically Eq. 10 ($\hat{\mathbf{w}}^*$ has been computed performing a single EM step). The dashed line represents the approximated value of the KL divergence after expanding the logarithm up

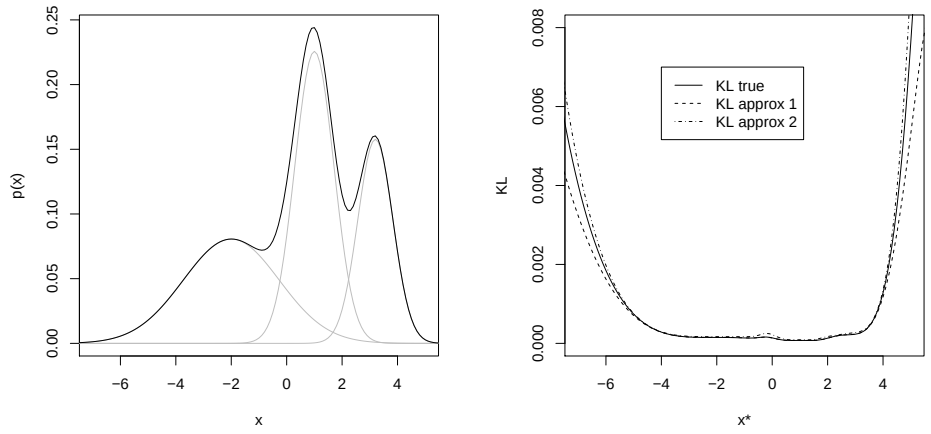


Figure 1: Left: Mixture distribution composed by three Gaussians. Right: Quality of the approximation of the KL divergence with respect to the values of $x_*$.

to the second order (Eq. 11); again, the integral has been computed numerically. Finally, the dashed-dotted line represents the approximated value of the KL divergence in Eq. 12, that is what we use as the final approximation. In the multivariate case, following the same derivation, we obtain the same expression as in Eq. 12. The details for the derivation of the multivariate case can be found in Ref. [11]. In this case, the variables $\hat{z}_k^2$ read:

$$\hat{z}_k^2 = (\mathbf{x}_* - \hat{\mathbf{m}}_k)^{\mathrm{T}} \hat{S}_k^{-1} (\mathbf{x}_* - \hat{\mathbf{m}}_k)$$

and the responsibilities for the test point:

$$u_{*k} = \frac{\hat{\pi}_k |\hat{S}_k|^{-1/2} \exp\left(-\frac{\hat{z}_k^2}{2}\right)}{\sum_r \hat{\pi}_r |\hat{S}_r|^{-1/2} \exp\left(-\frac{\hat{z}_r^2}{2}\right)}$$

We note here that other approximation schemes of the KL divergence between mixtures have been proposed [15]. Our contribution is to devise an asymptotic expansion for the online update case.

*3.3.1. Sampling rejection thresholds*

Assessing whether a new point is novel or normal usually involves the definition of a rejection region based on the training data. This can be a difficult task for complex multivariate distributions such as a mixture of multivariate Gaussians. One of the major advantages of the proposed approach is that

the problem is reduced to estimating rejection regions for the distribution on the KL divergence, which is intrinsically a one dimensional problem. Recall that the KL divergence results in a function of this form:

$$\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)] \simeq f(\hat{z}_1, \ldots, \hat{z}_c, \hat{\pi}_1, \ldots, \hat{\pi}_c, \hat{S}_1, \ldots, \hat{S}_c)$$

We are interested in evaluating the quantile of this random variable corresponding to a specific rejection rate. To do this, we can employ a Monte Carlo simulation. In principle we could sample directly $\mathbf{x}_*$, if we knew the true pdf generating the data, and compute the quantiles of the resulting histogram of the KL divergence. This would mean sampling from each component of the mixture the right proportions of points with the true mean and variances. Since we only have an estimate of such parameters, it is better to sample directly the values of $\hat{z}_k^2$ with the estimated proportions for all the components. This would allow to include the uncertainty on the estimated statistics in the method, given that we would sample $\hat{z}_k^2$ from an $F$ distribution instead that $\mathbf{x}_*$ from the predictive pdf. We use the ML estimate of the proportions, and we assume that in the mixture case the ML estimates of the mean and the covariance are distributed as in the unimodal case. We will see in the experimental part that these simplifying assumptions do not lead to a drop in performance in practice.

What we need to do is then trying to obtain a sampling scheme where all the quantities are expressed in terms of $\hat{z}_k^2$. For the sake of presentation, let's focus on two of the components of the mixture that we will denote as $i$ and $j$; also, let's consider the multivariate case. When we sample points from the $i$-th component, we need to compute their contribution to the KL divergence with respect to the other components of the mixture. Sampling a value of $\mathbf{x}_*$ would correspond to set the value of both $\hat{z}_i^2$ and $\hat{z}_j^2$, since:

$$\hat{z}_k^2 = (\mathbf{x}_* - \hat{\mathbf{m}}_k)^{\mathrm{T}} \hat{S}_k^{-1} (\mathbf{x}_* - \hat{\mathbf{m}}_k) = \hat{\mathbf{z}}_k^{\mathrm{T}} \hat{\mathbf{z}}_k \quad \forall k = 1, \ldots, c$$

where we introduced $\hat{\mathbf{z}}_k = \hat{S}_k^{-\frac{1}{2}} (\mathbf{x}_* - \hat{\mathbf{m}}_k)$. The last expression shows also that sampling directly $\hat{z}_i^2$, for example, corresponds to selecting values of $\mathbf{x}_*$ that correspond to values of $\hat{z}_j^2$. Rewriting the expression for $\hat{z}_j^2$, indeed:

$$\hat{z}_j^2 = \|\hat{\mathbf{z}}_j\|^2 = \|\hat{S}_j^{-\frac{1}{2}} (\mathbf{x}_* - \hat{\mathbf{m}}_j)\|^2 = \|\hat{S}_j^{-\frac{1}{2}} (\mathbf{x}_* - \hat{\mathbf{m}}_i + \hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)\|^2$$

$$\hat{z}_j^2 = \|\hat{S}_j^{-\frac{1}{2}} \hat{S}_i^{\frac{1}{2}} \hat{S}_i^{-\frac{1}{2}} (\mathbf{x}_* - \hat{\mathbf{m}}_i)\|^2 + \|\hat{S}_j^{-\frac{1}{2}} (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)\|^2 + 2(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^{\mathrm{T}} \hat{S}_j^{-1} \hat{S}_i^{\frac{1}{2}} \hat{S}_i^{-\frac{1}{2}} (\mathbf{x}_* - \hat{\mathbf{m}}_i)$$

Table 1: Pseudo-code of the information theoretic novelty detection method for the mixture of Gaussians.

---

1. set the false positive rate $\rho$, the number of components $c$, and the number of points $\nu$ for the Monte Carlo simulation;
2. run the EM algorithm on $X$ with $c$ components;
3. compute $\mathrm{KL}^*$ that is the KL value with the test point $\mathbf{x}_*$ using Eq. 12;
4. **procedure** Monte_Carlo:
   (a) **for** $(i = 1, \ldots, c)$ **repeat**
      i. generate $\nu\hat{\pi}_i$ values of $\hat{z}_i^2$ using Eq. 2
      ii. generate $\nu\hat{\pi}_i$ vectors $\hat{\mathbf{v}}_i$ using Eq. 14
      iii. Compute $\hat{z}_j^2$ using Eq. 13 $\forall j \neq i$
   (b) compute the distribution of KL using Eq. 12
   (c) **return** the value $\theta_\rho$ corresponding to the $(100 - \rho)$-th quantile of the distribution of KL
5. **if** $(\mathrm{KL}^* > \theta_\rho)$ **then** flag $\mathbf{x}_*$ as outlier
6. **else** flag $\mathbf{x}_*$ as normal

---

$$\hat{z}_j^2 = \|\hat{S}_j^{-\frac{1}{2}}\hat{S}_i^{\frac{1}{2}}\hat{\mathbf{z}}_i\|^2 + \|\hat{S}_j^{-\frac{1}{2}}(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)\|^2 + 2(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^{\mathrm{T}}\hat{S}_j^{-1}\hat{S}_i^{\frac{1}{2}}\hat{\mathbf{z}}_i$$

Now, let $\hat{\mathbf{z}}_i = \|\hat{\mathbf{z}}_i\|\hat{\mathbf{v}}_i = \sqrt{\hat{z}_i^2}\hat{\mathbf{v}}_i$, where we introduced the unit norm vector $\hat{\mathbf{v}}_i$:

$$\hat{z}_j^2 = \hat{z}_i^2\|\hat{S}_j^{-\frac{1}{2}}\hat{S}_i^{\frac{1}{2}}\hat{\mathbf{v}}_i\|^2 + \|\hat{S}_j^{-\frac{1}{2}}(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)\|^2 + 2\sqrt{\hat{z}_i^2}(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^{\mathrm{T}}\hat{S}_j^{-1}\hat{S}_i^{\frac{1}{2}}\hat{\mathbf{v}}_i \quad (13)$$

The last equation shows that when we sample a particular value of $\hat{z}_i^2$, we implicitly set the value of $\hat{z}_j^2$ with some uncertainty given by the direction of $\hat{\mathbf{v}}_i$. This effect can be easily seen on the univariate case. If we sample $\hat{z}_i^2$, we lose the information on the side of the Gaussian $x_*$ is coming from. When we compute the corresponding $\hat{z}_j^2$, we need to recover the information about the sign of $(x_* - \hat{m}_i)$ to compute it correctly. The situation is even worse in the case of multivariate distributions, where $\hat{\mathbf{v}}_i$ can assume all the possible directions in a $d$-dimensional space.

In our case, we want to generate, using a Monte Carlo simulation, several values of $\hat{z}_i^2$, and use them compute the corresponding $\hat{z}_j^2$. Using the last equation for $\hat{z}_j^2$, we see that we can generate independently $\hat{z}_i^2$ and $\hat{\mathbf{v}}_i$.

Since we want $\hat{\mathbf{v}}_i$ to be uniformly distributed on the unit $d$-dimensional hypersphere, we can generate them in this way:

$$\hat{\mathbf{v}}_i = \frac{1}{\gamma}\left(\mathcal{N}(m=0, s^2=1), \ldots, \mathcal{N}(m=0, s^2=1)\right) \tag{14}$$

where $\gamma$ is the normalization term needed to ensure that $\hat{\mathbf{v}}_i$ has norm 1.

Given $\hat{z}_i^2$, it is then possible to compute $\hat{z}_j^2$ for all $j \neq i$ by generating independently $c$ standardized normal Gaussian variables (for $\hat{\mathbf{v}}_i$) and an $F$ distributed variable (for $\hat{z}_i^2$). We can repeat this procedure for all the values of $i$ running from 1 to $c$, and compute the values of the random variable $\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)]$ in Eq. 12. We can finally compute the quantiles of the histogram of the sampled values, thus obtaining a threshold on the value of the KL divergence corresponding to a specific rejection rate. When a test point has to be analyzed, we compute its information content using the expression of the KL divergence in Eq. 12 and we compare it with the threshold to assess whether it is novel or not. This scheme takes explicitly into account the variability of the mean and the covariance of the components of the mixture, accommodating their intrinsic variability. Table 1 shows the steps composing the information theoretic novelty detection method for the mixture of Gaussians.

## 4. Experimental Validation

In this section, we evaluate the performance of our approach on both synthetic and real data. We report first some results on synthetic generated data. In this case, the parametric form of the true density is assumed known (Gaussian or mixture of $c$ Gaussians); we therefore compare our approach with the parametric ML approach. The procedure that we use to evaluate the performances of the proposed method is the following. We generate a training set $X$ of cardinality $n$ and a test set of cardinality $r = 10^6$ from the same generating distribution. We set the false positive rate that we are willing to tolerate, and we run the novelty detection algorithms using $X$ for the training stage. At this point, we compute the false positive rate on the test, namely the percentage of test points that we flag as outliers. We repeat this procedure $l = 200$ times for the unimodal and $l = 100$ for the mixture distributions, and for different cardinalities of $X$. Finally, we report the mean and the standard deviation of the false positive rate over different repetitions and different values of $n$. For the sake of presentation, in all the experiments

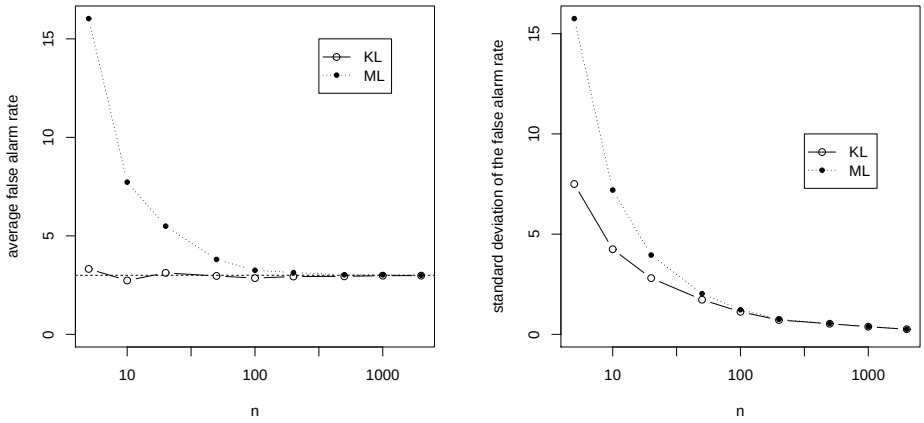


Figure 2: Comparison of the average (left) and standard deviation (right) of the false positive rate over 200 repetitions for the KL and the ML methods on the univariate Gaussian case.

we set the rejection levels to 3%. In Ref. [11] it is possible to find results using different threshold levels.

### *4.1. Gaussian*

When the data generating distribution is a Gaussian, we have shown that our test reduces to a form of the $F$-test, and therefore has theoretical guarantees of optimality. As a sanity check, however, we present some results showing that the proposed approach does indeed substantially improve on the ML estimation. For the univariate Gaussian case, we generated the data using $p(x) = \mathcal{N}(x|m, s^2)$, with $m = 2.3$ and $s^2 = 1.4$. The ML approach here consists in the test of $\hat{z}^2$ assuming that $\hat{m}$ and $\hat{s}^2$ are the true statistics of the generating distribution; in other words, we assume that $\hat{z}^2 \sim \chi_{(1)}$. In Fig. 2 we can see the comparison between the proposed method (that is equivalent to the $F$-test) and the ML approach on the average and the standard deviation of the false positive rate. While the ML approach grossly underestimates the false positive rates for small training sets, the KL approach on average achieves the desired false positive rate for training set as small as only 5 points. From Fig. 2 (right panel), we notice also that the empirical standard deviation on the false positive rate is significantly smaller in the KL case than the ML case, giving a greater reliability as well as higher accuracy in setting the false positive rate.

In the multivariate Gaussian case, the pdf generating the data is $p(x) =$

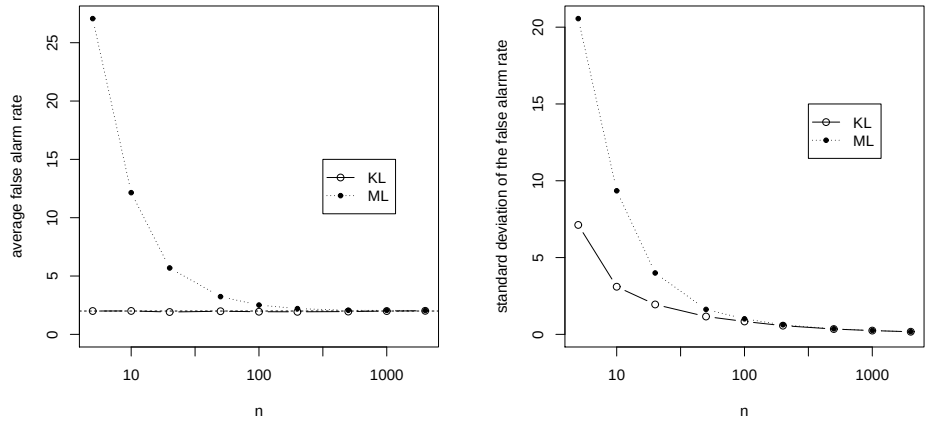


Figure 3: Comparison of the average (left) and standard deviation (right) of the false positive rate over 200 repetitions for the KL and the ML methods in the multivariate Gaussian case.

$\mathcal{N}(\mathbf{x}|\mathbf{m}, S)$; we set the parameters to $\mathbf{m} = (1.1, 3.2)$ and $S = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$. The ML approach here consists in the test of $\hat{z}^2$ assuming that $\hat{\mathbf{m}}$ and $\hat{S}$ are the true statistics of the generating distribution; in other words, we assume that $\hat{z}^2 \sim \chi_{(d)}$. In Fig. 3 we can see the comparison between the proposed method and the ML approach. Again, the proposed method gives us the expected false positive rate on average.

*4.2. Mixture of Gaussians*

We report here the results obtained in the mixture of univariate Gaussians. The generating distribution $p(x|\mathbf{w}) = \sum_{k=1}^{c} \pi_k \mathcal{N}(x|m_k, s_k^2)$ has $c = 3$ components, and the parameters are $\pi_1 = 0.35$, $\pi_2 = 0.40$, $\pi_3 = 0.25$, $m_1 = -2.0$, $m_2 = 1.0$, $m_3 = 3.2$, $s_1^2 = 3.0$, $s_2^2 = 0.5$, $s_3^2 = 0.4$. The pdf is plotted in Fig. 4 (left plot). We repeat the same procedure carried out in the unimodal case. The ML approach consists in computing the quantile of $p(x|\hat{\mathbf{w}})$, corresponding to the selected false positive rate, based on the training set $X$; then, the test points are flagged as novel when $p(x_*|\hat{\mathbf{w}})$ is below that threshold. In Fig. 5, we can see the comparison between the proposed method and the ML approach.

In the mixture of multivariate Gaussians case, the generating distribution is $p(x|\mathbf{w}) = \sum_{k=1}^{c} \pi_k \mathcal{N}(x|\mathbf{m}_k, S_k)$ with $c = 3$ components; the parameters are $\pi_1 = 0.5$, $\pi_2 = 0.2$, $\pi_3 = 0.3$, $\mathbf{m}_1 = (2.0, 1.3)$, $\mathbf{m}_2 = (3.2, -3.5)$,

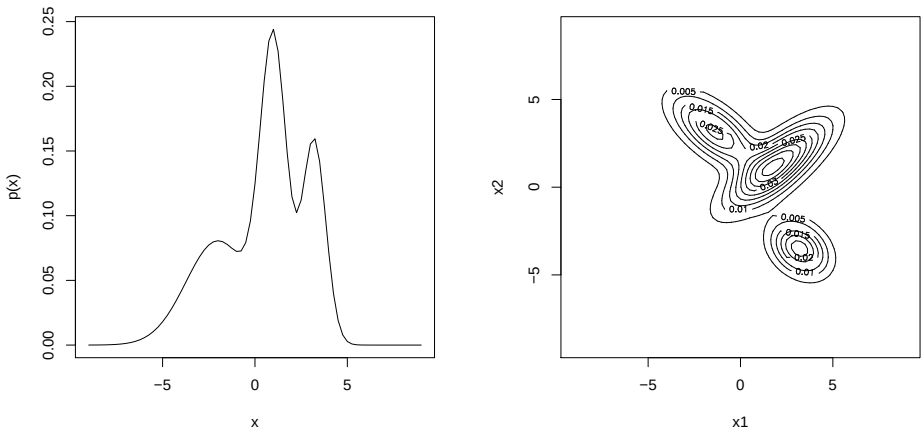


Figure 4: Pdfs of the univariate and multivariate mixture of Gaussian used in the experiments comparing KL and ML for novelty detection.

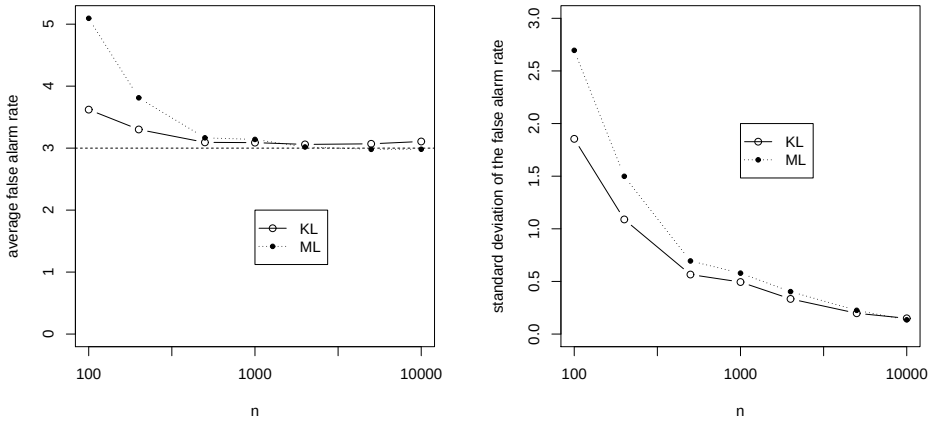


Figure 5: Comparison of the average (left) and standard deviation (right) of the false positive rate over 100 repetitions for the KL and the ML methods on the univariate mixture Gaussian case.

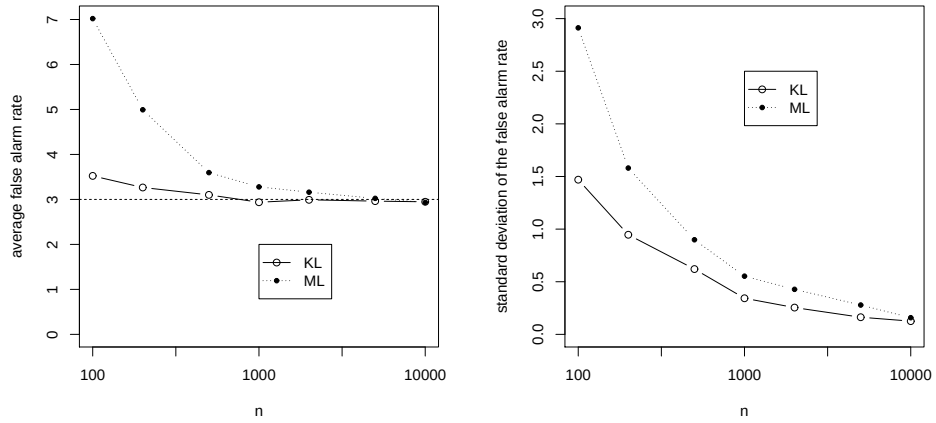


Figure 6: Comparison of the average (left) and standard deviation (right) of the false positive rate over 100 repetitions for the KL and the ML methods on the multivariate mixture Gaussian case.

$\mathbf{m}_3 = (-1.40, 3.15)$, $S_1 = \begin{pmatrix} 3.0 & 2.1 \\ 2.1 & 2.5 \end{pmatrix}$, $S_2 = \begin{pmatrix} 1.0 & -0.3 \\ -0.3 & 1.0 \end{pmatrix}$, $S_3 = \begin{pmatrix} 2.2 & -1.3 \\ -1.3 & 1.8 \end{pmatrix}$. The pdf is plotted in Fig. 4 (right plot). The ML approach consists in computing the quantiles of $p(\mathbf{x}|\hat{\mathbf{w}})$ based on the training set $X$; then, the test points are flagged as novel when $p(\mathbf{x}_*|\hat{\mathbf{w}})$ is below the threshold corresponding to the selected false positive rate. In Fig. 6 we can see the comparison between the proposed method and the ML approach.

*4.3. Real data sets*

In this section, we report some results on Iris [12] and Yeast [16] data sets from the UCI repository [3], and on the Biomed data set [9]. Since all these data sets are labeled, we use patterns from some of the classes to compose the normal class, and we use patterns from the remaining classes to compose the class of the outliers. In all these experiments, we analyze the performances of our method in terms of scores computed on the confusion matrix between the real and the predicted class, i.e. normal or outlier. We repeat the procedure for different subsamples of the normal class.

We compare our method with the ML approach and the *Kernel Density Estimation* (KDE) one [18]. In the ML approach we set a threshold on the probability of the test points by computing the quantiles of the probabilities of the training points. In the KDE approach, we use a Gaussian kernel with

standard deviation $\sigma$ estimated from data in the following way:

$$\sigma = \gamma \bar{D}$$

where $\bar{D}$ is the mean distance between the patterns in $X$, and $\gamma$ is a scaling factor. The value of $\gamma$ is chosen on the basis of the performances on the test data. Because of the limited training set size, we decided to set the threshold on the probability of the test points by computing the quantiles of the values of the probability of $10^4$ points generated from the pdf estimated using the KDE method.

*4.3.1. Iris*

The data set contains three classes of iris plants ("setosa", "versicolor", and "virginica") of 50 patterns each. The "setosa" class is linearly separable from the other two that are overlapped. The features are four: sepal length, sepal width, petal length, and petal width. We use the union of the classes "versicolor" and "virginica" as the normal class, and the class "setosa" as the one containing outliers. In our approach we run the EM algorithm with the number of components $c = 2$. In the KDE method, we set $\gamma = 1/7$ for setting the value of the kernel width $\sigma$. The results are shown in Fig. 7. The KL approach performs clearly much better than the ML approach in controlling the false positive rate. The improvement over the non-parametric KDE approach is smaller but still significant for less than 60 training points, where KDE is at risk of overfitting the data. At the same time, the performance in terms of overall classification accuracy is almost identical for KL and KDE, and slightly better than the ML approach.

*4.3.2. Yeast*

The Yeast data set is composed of 1484 patterns divided in ten classes. Each class represents the cellular localization site of proteins; each protein is described by a set of 8 attributes. We decided to consider the union of the classes "CYT", "ME3", "MIT", and "NUC" as the normal one, and the rest as the one containing outliers. In this way, the normal class is composed of 1299 patterns, and the outliers class of 185 patterns. Every time we sample $n$ points from the normal class, we perform PCA in order to project data onto a lower dimensional space of dimension 2; in this new subspace we perform novelty detection. This preprocessing is needed to reduce the number of free parameters; alternatively, one could select other strategies such as imposing a diagonal covariance structure for each class. In our approach we run the

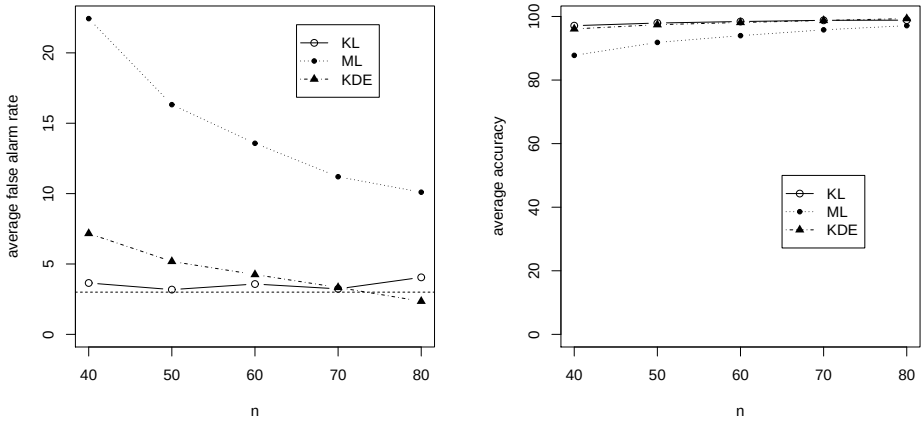


Figure 7: Comparison of the average false positive rate (left) and accuracy (right) over 100 repetitions for the KL, ML, and KDE methods on the Iris data set.

EM algorithm with the number of components $c = 2$, as we noticed that the projected data could be fit by two Gaussian. In the KDE method, we set $\gamma = 1/5$ for setting the value of the kernel width $\sigma$. The results are shown in Fig. 8. As in the Iris data set, we can observe a very similar performance in terms of overall accuracy, but a significant improvement in achieving the desired false positive rate for small training sets.

*4.3.3. Biomed*

The Biomed data set contains measurements of blood samples of people involved in a study for the identification of a rare genetic disorder. The patients are represented by four measurements on their blood samples and are grouped in two classes, one for the non-carriers and one for the carriers. We removed from the data set the patterns having missing values; the composition of the data set results then in 127 patterns for the normal class and 67 for the outliers class. We use the class of the non-carriers as the class of the normal patterns. In our approach we run the EM algorithm with the number of components $c = 2$. In the KDE method, we set $\gamma = 1/4$ for setting the value of the kernel width $\sigma$. The results are shown in Fig. 9. Again, performance accuracy is comparable for the three methods, but only the KL approach achieves the desired false positive rate.

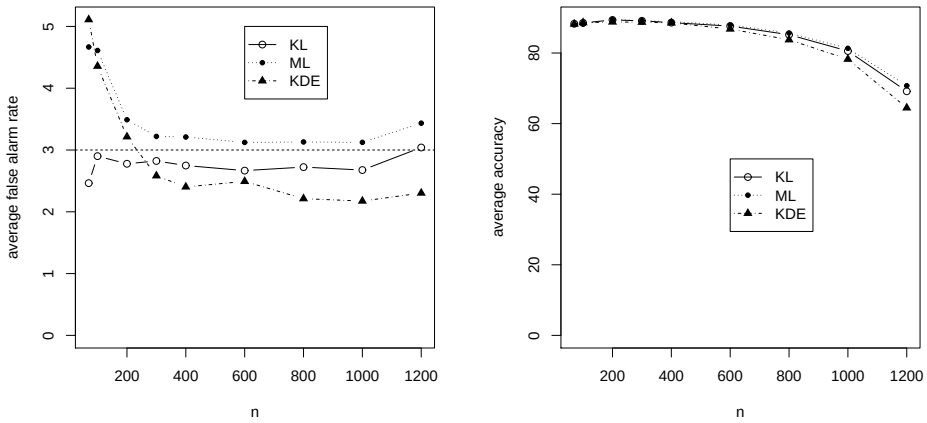


Figure 8: Comparison of the average false positive rate (left) and accuracy (right) over 100 repetitions for the KL, ML, and KDE methods on the Yeast data set.

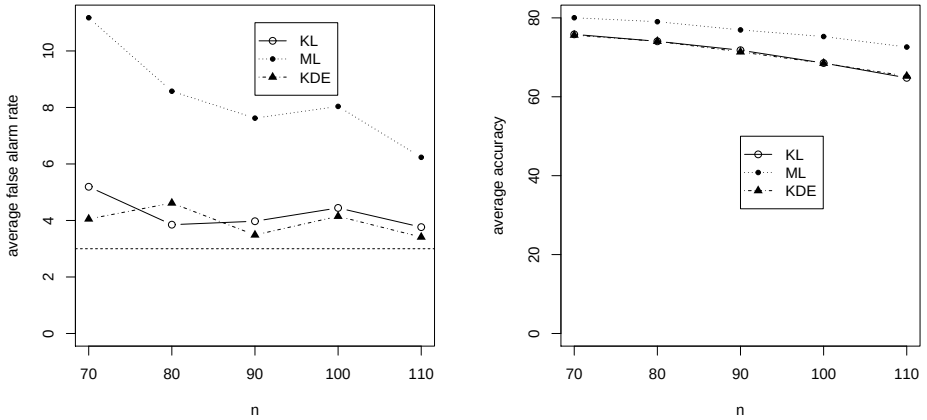


Figure 9: Comparison of the average false positive rate (left) and accuracy (right) over 100 repetitions for the KL, ML, and KDE methods on the Biomed data set.

## 5. Conclusion

In this paper we propose a novel method to control the false positives rate in novelty detection problems. The method is based on estimating the distribution of the information content of a new data point given a training set of a size $n$. The rationale for doing this is that it explicitly takes into account the size of the training set in setting a threshold for novelty. Remarkably, we show that in the univariate and multivariate Gaussian cases the distribution of this information content does not depend on the statistics of the data distribution. This result leads to a natural connection with statistical testing. In particular, the novelty detection test is able to control the false positive rate even when the training set size is small.

We also propose an extension of our approach to the mixture of Gaussians case. The information theoretic approach allows us to use an approximation scheme for the computation of the information content of a test point, yielding a novelty detection method controlling the false positive rate.

While we believe the method to be novel and potentially useful, it is important to stress once more its limitations as well. The question we set ourselves to answer is to produce a novelty detection method that gives good guarantees of achieving a certain rate of false positives when the training set is small. As we have seen, the ML method with small training set tends to give thresholds that typically capture less probability mass than expected, resulting in false positives rates well beyond the expected, while non-parametric methods such as KDE suffer from overfitting and difficulties in setting hyperparameters. Our method is quite accurate in setting thresholds which capture the desired fraction of the mass of the unknown data distribution. However, the flip-side of this is that, if the distribution of the anomalous points is not far from the training distribution, it will inevitably let more abnormal cases slip through the net. This might result in a lower accuracy in detecting true positives. In some applications this behavior is not desirable, since the cost of letting an abnormal case go undetected is much higher than the cost of raising a false alarm.

The fact that our method performs well in the mixture of Gaussians case is encouraging, since this gives the chance to model a broader class of distributions. The method illustrated here relies on using the EM algorithm, and for high dimensional data or when the distribution involves several components it may become computationally demanding or unfeasible for small training sets. In these cases, non-parametric methods such as [7] are prefer-

able and may result in better performance. It should be pointed out though that non-parametric methods generally depend on the value of hyperparameters which can be problematic to determine. It is clearly a very interesting question whether it is possible to extend this information theoretic approach to non-parametric methods.

## References


[1] T. W. Anderson and T. W. Anderson. *An Introduction to Multivariate Statistical Analysis, 2nd Edition*. Wiley-Interscience, 2 edition, September 1984.

[2] C. Archer, T. K. Leen, and A. M. Baptista. Parameterized novelty detectors for environmental sensor monitoring. In *Advances in Neural Information Processing Systems 16, NIPS 2003*, December 2003.

[3] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007.

[4] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley Series in Probability & Statistics. Wiley, April 1994.

[5] C. M. Bishop. Novelty detection and neural network validation. *IEE Proceedings on Vision, Image and Signal processing*, 141(4):217–222, 1994.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[7] C. Campbell and K. P. Bennett. A linear programming approach to novelty detection. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13, NIPS 2000*, pages 395–401, 2000.

[8] D. A. Clifton, N. Mcgrogan, L. Tarassenko, D. King, S. King, and P. Anuzis. Bayesian extreme value statistics for novelty detection in gas-turbine engines. In *Proceedings of IEEE Aerospace*, 2008.

[9] L. H. Cox, M. M. Johnson, and K. Kafadar. Exposition of statistical graphics technology. In *ASA Proceedings of the Statistical Computation Section*, 1982.

[10] S. Eguchi and J. Copas. Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of Multivariate Analysis*, 97(9):2034–2040, 2006.

[11] M. Filippone and G. Sanguinetti. Information theoretic novelty detection. Technical Report CS-09-02, Department of Computer Science, University of Sheffield, February 2009.

[12] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugenics*, 7:179–188, 1936.

[13] P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis. Support vector novelty detection applied to jet engine vibration spectra. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13, NIPS 2000*, pages 946–952. MIT Press, 2000.

[14] C. He, M. Girolami, and G. Ross. Employing optimized combinations of one-class classifiers for automated currency validation. *Pattern Recognition*, 37(6):1085–1096, 2004.

[15] J. R. Hershey and P. A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.*, volume 4, pages IV–317–IV–320, 2007.

[16] P. Horton and K. Nakai. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 109–115. AAAI Press, 1996.

[17] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems 18, NIPS 2005*, December 2005.

[18] M. Markou and S. Singh. Novelty detection: a review - part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[19] D. Martinez. Neural tree density estimation for novelty detection. *IEEE Transactions on Neural Networks*, 9(2):330–338, Mar 1998.

[20] J. A. Quinn and C. K. I. Williams. Known unknowns: Novelty detection in condition monitoring. In J. Martí, J. M. Benedí, A. M. Mendonça, and J. Serrat, editors, *Pattern Recognition and Image Analysis, Third Iberian Conference, IbPRIA 2007*, volume 4477 of *Lecture Notes in Computer Science*, pages 1–6. Springer, June 2007.

[21] S. J. Roberts. Novelty detection using extreme value statistics. *IEE Proceedings on Vision, Image and Signal Processing*, 146(3):124–129, 1999.

[22] B. Schölkopf, R. C. Williamson, A. J. Smola, J. S. Taylor, and J. C. Platt. Support vector method for novelty detection. In S. A. Solla, T. K. Leen, K. R. Müller, S. A. Solla, T. K. Leen, and K. R. Müller, editors, *Advances in Neural Information Processing Systems 12, NIPS 1999*, pages 582–588. The MIT Press, 1999.

[23] Y. Singer and M. K. Warmuth. Batch and on-line parameter estimation of gaussian mixtures based on the joint entropy. In M. J. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11, NIPS 1998*, pages 578–584. The MIT Press, 1998.

[24] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Fourth International Conference on Artificial Neural Networks*, pages 442–447, 1995.

[25] C. K. I. Williams, J. A. Quinn, and N. Mcintosh. Factorial switching kalman filters for condition monitoring in neonatal intensive care. In *Advances in Neural Information Processing Systems 18, NIPS 2005*, December 2005.

[26] J. Zhang, Z. Ghahramani, and Y. Yang. A probabilistic model for online document clustering with application to novelty detection. In *Advances in Neural Information Processing Systems 17, NIPS 2004*, December 2004.