Information Theoretic Novelty Detection

Maurizio Filippone, Guido Sanguinetti
Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street Sheffield, S1 4DP - United Kingdom.
email - filippone@dcs.shef.ac.uk, g.sanguinetti@dcs.shef.ac.uk

February 2009

Technical Report CS-09-02

Abstract

We present a novel approach to online change detection problems based on estimating the expected information content of a new data point. We show that in the Gaussian and multivariate Gaussian cases an approximate expression can be derived which does not depend on the estimated statistics of the training distribution, but only on the size of the training set. This allows an accurate estimate of the expected false positive rate even when the size of the training set is small. We analyze the connections with statistical testing in the case of the Gaussian, and we propose an approximation scheme to the case of the mixture of Gaussian.

1 Introduction

Novelty detection is a fundamental task in machine learning which plays a prominent role in many application domains, from fault detection [2, 7, 9] to monitoring medical conditions [15, 19]. In its classic formulation, novelty detection is the identification of data that deviate from the norm using only knowledge from normality. In order to do this, 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 the user 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 [13, 4], extreme value statistic [16], support vector methods [17, 6] and non-parametric Bayesian approaches [20] (for a good review of statistical approaches for novelty detection see *e.g.* [3, 12]). 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 kernel density estimation, for example, it is difficult to select the bandwidth of the kernel in order to model data properly avoiding the problem of overfitting. On the other hand, parametric approaches usually rely on strong distributional assumptions thus limiting their applicative domain. For some distributions, however, it is possible to achieve important theoretical results on the tests for novelty detection. Unfortunately, the class of distributions having this property is very limited; for a relevant set of distributions there are no statistical tests with such theoretical guarantees.

In this report, 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. We show that the KL divergence in the univariate and multivariate Gaussian cases does not depend on the estimated statistics of the training distribution; we also emphasize the connections with statistical testing in such cases. In particular, the resulting method is able to control the false positive rate even when the training set is small. Finally, we extend the information theoretic approach to the case of the mixture of univariate and multivariate Gaussian. 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 [10]. 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 Gaussian.

The report is organized as follows: in Section 2 we report some basic concepts on statistical

testing, in Section 3 we present our method, in Section 4 we show the results on some simulated data, and we report the conclusions in Section 5.

2 Statistical Testing

Let $X = \{x_1, \ldots, x_n\}$ be a set of points $x_i \in \mathbb{R}$ generated from an unknown *probability density function (pdf)* $p(x)$. In data modeling, it is usual to try to approximate the pdf generating the data. This is a crucial point in many application domains, and several approaches have been proposed. We can broadly divide such approaches into parametric and non-parametric. In this report, we will focus on parametric density estimation as a way to approach the novelty detection problem. In particular, we parameterize the pdf $p(x)$ generating the points x_i by means of a set of parameters $\mathbf{w}$. In other words, we assume $x_i \sim p(x_i|\mathbf{w})$. In data modeling, we are interested in inferring the parameters $\mathbf{w}$ on the basis of the observed data set X.

From a frequentist point of view, we can consider the likelihood of having observed the particular set X given the assumed model, and we can maximize it with respect to the parameters $\mathbf{w}$. Formally, under the i.i.d. assumption for the points x_i, the likelihood reads:

$$L = p(X|\mathbf{w}) = \prod_{i=1}^{n} p(x_i|\mathbf{w})$$

The density estimation problem, then, reduces to the maximization of L with respect to $\mathbf{w}$:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \{p(X|\mathbf{w})\}$$

This is the so called *Maximum Likelihood (ML) approach*. In such approach, the predictive probability distribution results in $p(x|\hat{\mathbf{w}})$. More sophisticated approaches could take into account the variability of the parameters given the fact that their estimate is based on a limited set of points; we will discuss this issue in more detail in the next section.

From a Bayesian perspective, instead, the observed data are used to update our prior belief about the values assumed by the parameters $\mathbf{w}$. In this framework, we encode our prior belief in a prior probability distribution $p(\mathbf{w})$ over $\mathbf{w}$. As in the ML approach, we can write the likelihood of having observed X given the parameters $\mathbf{w}$:

$$L = p(X|\mathbf{w}) = \prod_{i=1}^{n} p(x_i|\mathbf{w})$$

Now, we can use Bayes' theorem to compute the probability density function of the parameters after having observed X. Since this pdf is computed after having observed X, it is usually called posterior distribution, and results in:

$$p(\mathbf{w}|X) = \frac{1}{Z} p(X|\mathbf{w}) p(\mathbf{w})$$

where Z is a normalization constant ensuring that the posterior is a proper pdf, i.e.:

$$Z = \int p(X|\mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

The Bayesian analysis offers an interesting view of the data modeling problem. In particular, it yields a probabilistic measure on how likely are the values of the parameters, given that we observed X. At this point, the predictive probability for a new point x is $p(x|\mathbf{w})$, and we can attempt to marginalize out the parameters, thus obtaining:

$$p(x|X) = \int p(x|\mathbf{w})p(\mathbf{w}|X)d\mathbf{w}$$

It is important to remark that in many cases Bayesian data analysis is not straightforward. Often, the direct computation of some of the integrals involved in such analysis is not trivial. Approximation schemes have been proposed to overcome this limitation, but often come to a computational cost [5].

Let's now consider the problem of novelty detection, where we want to decide whether a new test point, that we will denote as x_*, is an outlier or not. After the density estimation part, we can use the predictive pdf as a starting point for the test. Intuitively, regions of the space where the predictive probability is low would indicate a low probability of observing data points there. Therefore, it is reasonable to identify such regions on the basis of some statistical means. In the case of univariate distributions, we can identify these regions by using the quantiles, corresponding to specific rejection levels, of the predictive distribution. In other words, it is possible to set a specific amount of the area on the tails of the predictive distribution that we want to consider as a "rejection region". Corresponding to this area, we can identify the regions of the real line where the total probability of a point coming from the predictive distribution is equal to the selected amount. We notice here that we select the regions on the basis of the predictive pdf that has been built from X, that is a finite-size set. This would give a deviation from the results that we expect. We also notice that, even if we know the true predictive pdf, setting a rejection rate corresponds to setting the rate of normal points that we are willing to misclassify as outliers. Such situation gives rise to false positives, and occurs when we flag normal pattern as outliers. In this scenario, setting a specific rejection rate corresponds to setting the false alarm rate that we are willing to tolerate.

It is important to remark that the problem of finding rejection regions can be difficult. For some univariate pdfs belonging to the exponential family, the computation of rejection regions is feasible, since there is a closed form for computing the quantiles. In the case of some important univariate pdfs such as mixture models, however, the identification of rejection regions is not straightforward. Also, and most important, for almost all the multivariate distributions, there is no closed form for the computation of rejection regions based on quantiles. In all these cases, a way to identify these regions could be based on sampling techniques. This means that we can sample from the predictive pdf, estimated from data, and compute the quantiles of the histogram of the sampled points. When a new test point has to be analyzed, its pdf value is compared against the threshold obtained from the computed quantiles. It is worth noting, however, that in multivariate cases the amount of data required to sample the data space can be large. In some cases, it is possible to overcome this limitation by transforming the test for novelties from multivariate to univariate, as we will see shortly in the case of the multivariate Gaussian.

2.1 Univariate Gaussian

Let's assume that data are generated from a univariate Gaussian with mean m and variance s^2, i.e., $p(x) = \mathcal{N}(x|m, s^2)$. In a frequentist setting, we can use the ML approach to compute the statistics

$\hat{m}$ and $\hat{s}^2$, that result 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 in not a good choice, since the values of the statistics can strongly deviate from the true values. In particular, the sample mean $\hat{m} \sim \mathcal{N}(m, \frac{s^2}{n})$, and the sample variance $\hat{s}^2 \sim \frac{s^2}{n}\chi^2_{(n-1)}$. 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]:

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

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 report, 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. It is possible to come to similar conclusions in the case of a Bayesian analysis too. The F distribution results from the square of a Student-t distribution, that is obtained by marginalizing out the parameters from the predictive pdf in the Bayesian inference for the Gaussian [5].

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 $\mathbf{y} \sim \mathcal{N}(\mathbf{m}, aS)$, and $A \sim W_\nu(S)$ ($a \in \mathbb{R}$, S a positive semidefinite covariance matrix, ν the number of degrees of freedom of the Wishart distribution W):

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

Given that we know how the sample mean and the sample covariance are distributed, we see that the distribution of $\hat{z}^2$ is:

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

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 Kullback-Leibler Divergence for Novelty Detection

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. Using this parametric assumption, one can easily obtain an estimate $\hat{\mathbf{w}}$ of the parameters from the training set via maximum likelihood (ML) 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

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

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 [11]. In that approach, the KL divergence measures the dissimilarity between prior and posterior in Bayesian inference. In some sense it shares some similarities with the method we are proposing here, but we intend to use such approach in novelty detection applications. Also, in Ref. [8] there is a connection between Neyman-Pearson lemma and the KL divergence. In statistics, the Neyman-Pearson lemma gives an important result on the power of a statistical test, through the likelihood ratio analysis, where two hypotheses have to be tested. In novelty detection, such tests are not straightforward, since usually we infer normality from a data set, and we don't know the model of what constitutes the alternative hipothesis.

3.1 Univariate Gaussian distribution

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

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

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 a specific false positive rate. 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.

3.2 Multivariate Gaussian distribution

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 the Appendix 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. Mimicking the procedure followed in the univariate case, we obtain that the KL divergence between the training distribution and the updated distribution is

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

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 Mixture of Gaussian distributions

In this section, we consider the mixture of c Gaussian distributions as the pdf generating the data. Let's 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 π_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 Gaussian. 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 summerized here as:

1. computing a ML solution by the *Expectation Maximization* (EM) algorithm on X;
2. updating the parameters doing 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 [5]. 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.

(Step 2). The update of the parameters when we add a new test point x_* would require to start from the ML solution obtained by EM, 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 $X \cup \{x_*\}$. An interesting online method for the update of the parameters can be found in Ref. [18]. 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 $p(x|\hat{\mathbf{w}})$ is 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:

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

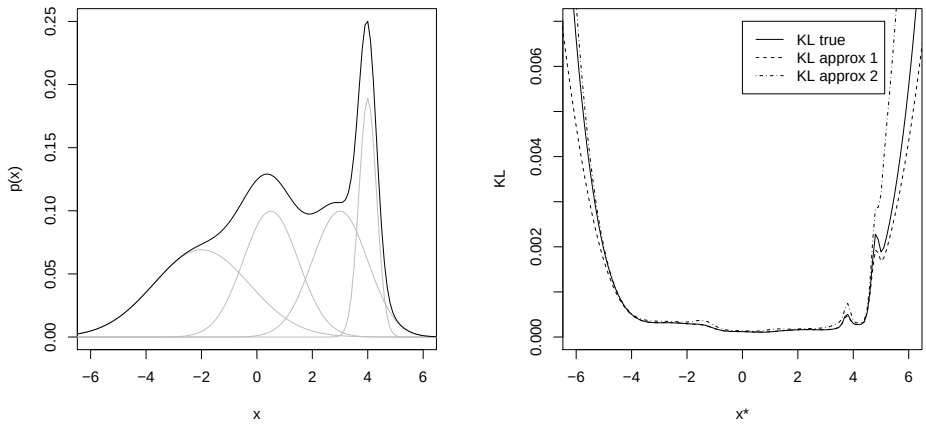

Figure 1: Left: Mixture distribution composed by four Gaussian. Right: Quality of the approximation of the KL divergence with respect to the values of x_*.

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 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, we show the pdf generating the data composed by four 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 on the plot show the quality of the approximations. The solid line represents the true value of $\mathrm{KL}[p(x|\hat{\mathbf{w}}) \| p(x|\hat{\mathbf{w}}^*)]$ computed by integrating numerically Eq. 10. The dashed line represents the approximated value of the KL divergence after expanding the logarithm up 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.

Remarkably, in the multivariate case, following the same derivation, we obtain the same expression in Eq. 12. The details for the derivation of the multivariate case can be found in the Appendix. 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:

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

3.3.1 Monte Carlo simulation to obtain rejection thresholds on the KL divergence

In this section, we study how we can apply efficiently the proposed method in practice. We 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. Then, we can compare it with the value of the KL divergence computed on the basis of the test point. In order to compute the quantile of the KL divergence, 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 components 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. Of course, we need to use the ML estimate of the proportions, and we are assuming that in the mixture case the ML estimates of the mean and the covariance are distributed as in the unimodal case. This last assumption is not true in general; for example, when the components of the mixture are strongly overlapped, the distribution of the mean of the components does not follow a Gaussian distribution. In the experimental part we will see how these facts affect the performances.

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. Focusing on two components only, 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 Gaussian.

---

1. set the false positive rate ρ, the number of components c, and the number of points ν for the Monte Carlo simulation;
2. run the EM algorithm on X with c components;
3. compute 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 θ_ρ 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 $\hat{\mathbf{v}}_i$:

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

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 $(\mathbf{x}_* - \hat{\mathbf{m}}_i)$ 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 from 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 γ 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 Gaussian.

4 Experimental Validation

We report some results on synthetic generated data. 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 from the same generating distribution. We set the false alarm 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 times and for different cardinalities of X. Finally, we report the mean and the standard deviation of the false alarm rate over different repetitions and different values of n.

4.1 Gaussian

For the univariate Gaussian case, we generated the data using $p(x) = \mathcal{N}(x|m, s^2)$, with $m = 2.3$ and $s^2 = 1.4$. We generate the test set with cardinality $r = 10^6$, and for each value of n, we repeat the test $l = 200$ times. 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 alarm rate. We set the value of the rejection rate to 3%, and it is what we get on average with the proposed method.

In the multivariate Gaussian case, the pdf generating the data is $p(x) = \mathcal{N}(\mathbf{x}|\mathbf{m}, S)$; we set the parameters to $\mathbf{m} = (1.1, 3.2)$ and $S = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$. We generate the test set with cardinality $r = 10^5$, and for each value of n, we repeat the test over $l = 200$ repetitions. 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, where we set the false positive rate to 3%. Again, the proposed method gives us the expected false positive rate on average.

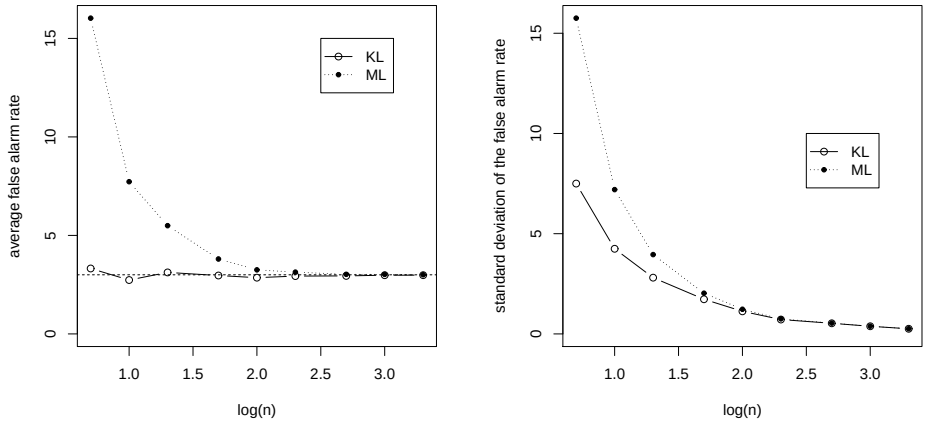

Figure 2: Comparison of the average (left) and standard deviation (right) of the false positive rate over 200 repetitions for the KL and the ML methods in the univariate Gaussian case.

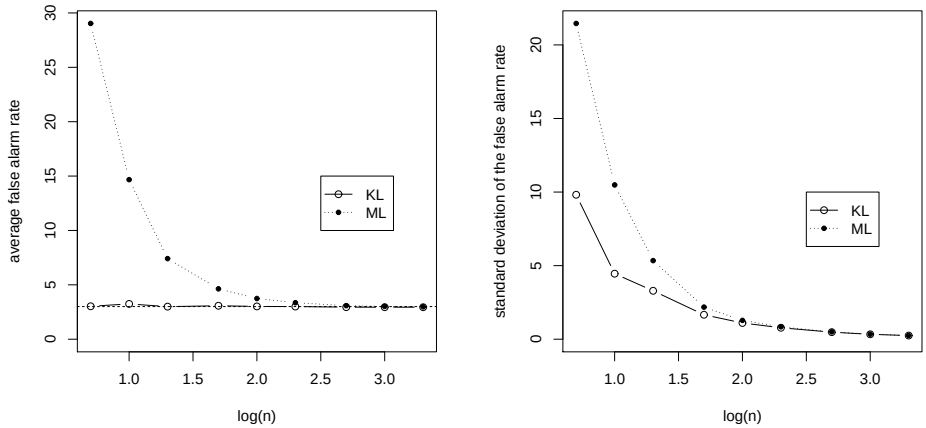

Figure 3: Comparison of the average (left) and standard deviation (right) of the false positive rate over 200 repetitions for the KL and the ML methods in the multivariate Gaussian case.

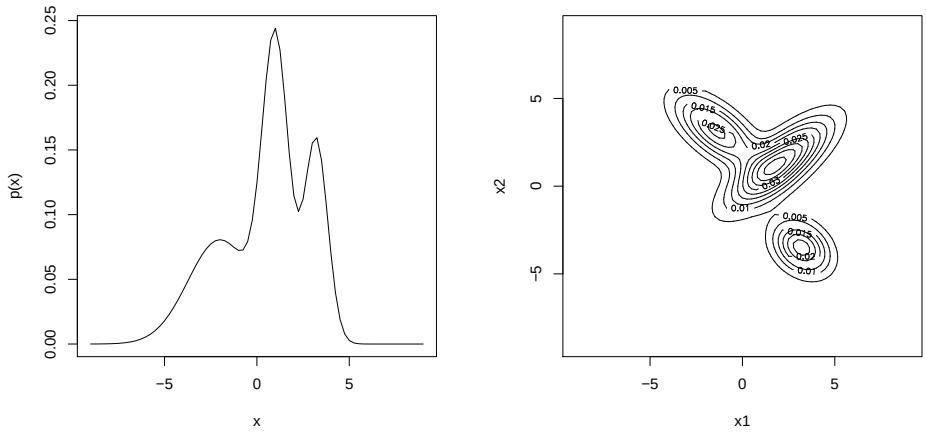

Figure 4: Pdfs of the univariate and multivariate mixture of Gaussian used in the experiments comparing KL and ML for novelty detection.

4.2 Mixture of Gaussian

We report here the results obtained in the mixture of univariate Gaussian. 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. We generate the test set with cardinality $r = 5 \cdot 10^4$, and for each value of n, we repeat the test over $l = 100$ repetitions. 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 where we set the false positive rate to 1%.

In the mixture of multivariate Gaussian 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 $\pi_1 = 0.5$, $\pi_2 = 0.2$, $\pi_3 = 0.3$, $\mathbf{m}_1 = (2.0, 1.3)$, $\mathbf{m}_2 = (3.2, -3.5)$, $\mathbf{m}_3 = (-1.40, 3.15)$, $S_1 = \begin{pmatrix} 3.0 & 2.1 \\ 2.1 & 2.5 \end{pmatrix}$, $S_2 = \begin{pmatrix} 1.0 & -0.3 \\ -0.3 & 1.0 \end{pmatrix}$, $S_3 = \begin{pmatrix} 2.2 & -1.3 \\ -1.3 & 1.8 \end{pmatrix}$. The pdf is plotted in Fig. 4 (right plot). We generate the test set with cardinality $r = 5 \cdot 10^4$, and for each value of n we repeat the test over $l = 100$ repetitions. 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; we set the false positive rate to 1%.

5 Conclusion

In this report 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

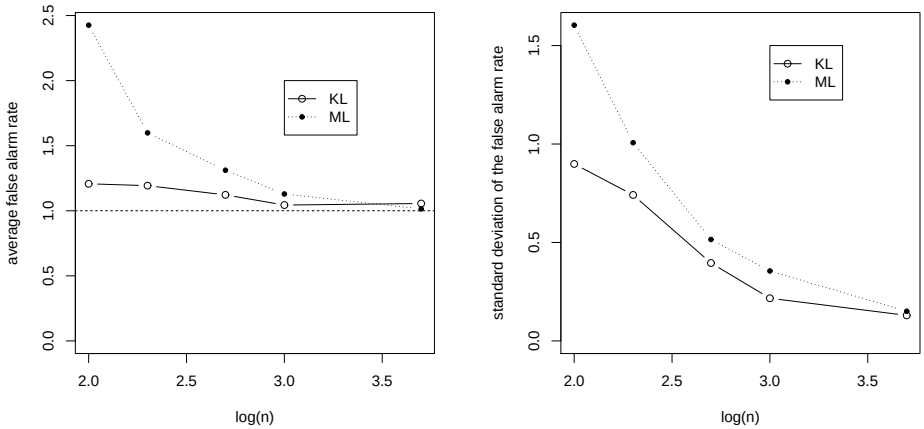

Figure 5: Comparison of the average (left) and standard deviation (right) of the false positive rate over 100 repetitions for the KL and the ML methods in the univariate mixture Gaussian case.

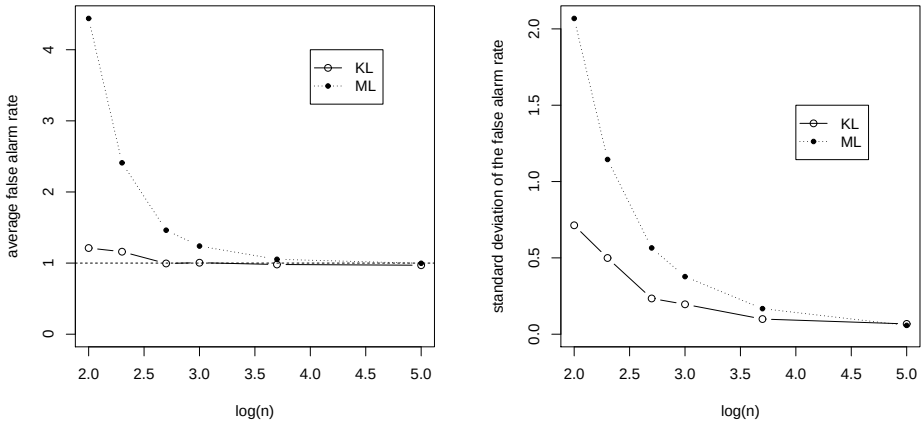

Figure 6: Comparison of the average (left) and standard deviation (right) of the false positive rate over 100 repetitions for the KL and the ML methods in the univariate mixture Gaussian case.

data point given that we had a training set of a certain size n. The rationale for doing this is that this approach 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 results 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 Gaussian 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 of 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. Our method is quite accurate in setting thresholds which capture the desired fraction of the mass of the unknown data distribution. However, the flipside 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.

Although we present the extension to the mixture of Gaussian, there are pdfs that cannot be modelled adequately. In these cases, non-parametric methods such as [6] are preferable 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. Finally, it would be interesting to compare our method with the mixture of Student-t distributions proposed in Ref. [14].

A Appendix

A.1 Multivariate Gaussian

Here we report the algebraic manipulations involved in the derivation of the KL divergence in the multivariate Gaussian case. The KL divergence between the multivariate Gaussian distribution $\mathcal{N}(\hat{\mathbf{m}}, \hat{S})$ and the updated version $\mathcal{N}(\hat{\mathbf{m}}_*, \hat{S}_*)$ is:

$$\mathrm{KL} = \frac{1}{2}\left[\log\left(\det \hat{S}_* \hat{S}^{-1}\right) + \mathrm{tr}\left(\hat{S}_*^{-1}\hat{S}\right) + (\hat{\mathbf{m}} - \hat{\mathbf{m}}_*)^{\mathrm{T}} \hat{S}_*^{-1}(\hat{\mathbf{m}} - \hat{\mathbf{m}}_*) - d\right] \tag{15}$$

We recall that we introduced:

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

and that the updated versions of the mean and the covariance matrix are:

$$\hat{\mathbf{m}}_* = \frac{n}{n+1}\hat{\mathbf{m}} + \frac{1}{n+1}\mathbf{x}_* \tag{16}$$

$$\hat{S}_* = \frac{n}{n+1}\hat{S} + \frac{n}{(n+1)^2}A \tag{17}$$

Let's analyze each term of Eq. 15. The first term is:

$$\begin{aligned} \det\left[\left(\hat{S}_*\hat{S}^{-1}\right)\right] &= \det\left[\frac{n}{(n+1)^2}\left((n+1)I + A\hat{S}^{-1}\right)\right] \\ &= \left(\frac{n}{(n+1)^2}\right)^d \det\left[(n+1)I + A\hat{S}^{-1}\right] \end{aligned} \tag{18}$$

The determinant of a matrix is the product of its eigenvalues. The eigenvalues of $(n+1)I + A\hat{S}^{-1}$ are the eigenvalues of $A\hat{S}^{-1}$ plus $(n+1)$. We notice that A has rank equal to one; this implies that $A\hat{S}^{-1}$ has only one non-zero eigenvalue, that we call $\hat{z}^2$. Recalling that the eigenvalues and the eigenvectors of a generic operator B satisfy $B\mathbf{v} = \lambda\mathbf{v}$, we have:

$$A\hat{S}^{-1}\tilde{\mathbf{x}}_* = \tilde{\mathbf{x}}_*\tilde{\mathbf{x}}_*^{\mathrm{T}}\hat{S}^{-1}\tilde{\mathbf{x}}_* = \tilde{\mathbf{x}}_*\hat{z}^2$$

The variable $\hat{z}^2$ is the eigenvalue of $A\hat{S}^{-1}$ associated to the eigenvector $\tilde{\mathbf{x}}_*$; it is also the quadratic form $\tilde{\mathbf{x}}_*^{\mathrm{T}}\hat{S}^{-1}\tilde{\mathbf{x}}_*$. Finally, the first term of Eq. 15 is:

$$\log\left(\frac{\hat{z}^2}{n+1} + 1\right) + d\log\left(\frac{n}{n+1}\right) \tag{19}$$

The second term can be rewritten using the properties of the trace:

$$\begin{aligned} \mathrm{tr}\left(\hat{S}_*^{-1}\hat{S}\right) &= \mathrm{tr}\left(\hat{S}\hat{S}_*^{-1}\right) \\ &= \mathrm{tr}\left[\left(\hat{S}_*\hat{S}^{-1}\right)^{-1}\right] \\ &= \frac{(n+1)^2}{n}\mathrm{tr}\left[\left((n+1)I + A\hat{S}^{-1}\right)^{-1}\right] \end{aligned} \tag{20}$$

The trace of a generic matrix B is the sum of its eigenvalues. It is easy to prove that:

$$B\mathbf{x} = \lambda\mathbf{x} \qquad \Rightarrow \qquad B^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}$$

We have already analyzed the eigenvalues of $(n+1)I + A\hat{S}^{-1}$; the second term of Eq. 15 will then result in:

$$\left(\frac{n+1}{n}\right)\left(d - 1 + \frac{1}{\frac{\hat{z}^2}{n+1} + 1}\right) \tag{21}$$

The third term can be rewritten in the following way:

$$(\hat{\mathbf{m}} - \hat{\mathbf{m}}_*)^{\mathrm{T}}\hat{S}_*(\hat{\mathbf{m}} - \hat{\mathbf{m}}_*) = \frac{1}{n}\tilde{\mathbf{x}}_*^{\mathrm{T}}\hat{S}^{-1}\left[(n+1)I + A\hat{S}^{-1}\right]^{-1}\tilde{\mathbf{x}}_* \tag{22}$$

We notice that:

$$\left[(n+1)I + A\hat{S}^{-1}\right]^{-1}\tilde{\mathbf{x}}_* = \frac{\tilde{\mathbf{x}}_*}{\hat{z}^2 + n + 1}$$

Recalling that $\hat{z}^2 = \tilde{\mathbf{x}}_*^{\mathrm{T}} \hat{S}^{-1} \tilde{\mathbf{x}}_*$, the third term of Eq. 15 results in:

$$\frac{\hat{z}^2}{n(\hat{z}^2 + n + 1)} \tag{23}$$

Introducing $y = 1/n$, the expression of the KL divergence will be:

$$\mathrm{KL}(y, \hat{z}) = \frac{1}{2}\left[\log\left(1 + y + y\hat{z}^2\right) - (d+1)\log\left(1+y\right) - 1 + \frac{1+y}{1+y+y\hat{z}^2} + yd\right] \tag{24}$$

A.2 Multivariate Mixture of Gaussian

In the multivariate case, the update of the parameters after one M step of the EM algorithm are the following:

$$\hat{\mathbf{m}}_k^* = \frac{\hat{\mathbf{m}}_k n_k + \mathbf{x}_* u_{*k}}{n_k + u_{*k}}$$

$$\hat{S}_k^* = \frac{n_k}{n_k + u_{*k}}\hat{S}_k + \frac{n_k u_{*k}}{(n_k + u_{*k})^2}(\mathbf{x}_* - \hat{\mathbf{m}}_k)(\mathbf{x}_* - \hat{\mathbf{m}}_k)^{\mathrm{T}}$$

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

Their incremental version reads:

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

$$\hat{S}_k^* = \hat{S}_k + \Delta\hat{S}_k = \hat{S}_k + \frac{n_k u_{*k}}{(n_k + u_{*k})^2}(\mathbf{x}_* - \hat{\mathbf{m}}_k)(\mathbf{x}_* - \hat{\mathbf{m}}_k)^{\mathrm{T}} - \frac{\hat{S}_k u_{*k}}{(n_k + u_{*k})}$$

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

The first order approximation of $p(x|\hat{\mathbf{w}}^*) = \sum_{k=1}^{c} \hat{\pi}_k^* \mathcal{N}(x|\hat{\mathbf{m}}_k^*, \hat{S}_k^*)$ can be computed by following the idea used in the univariate case. In this case we have to deal with matrix and vector derivatives:

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

where:

$$\Delta\psi_k = \left[(\mathbf{x} - \hat{\mathbf{m}}_k)^{\mathrm{T}} \hat{S}_k^{-1} \Delta\hat{\mathbf{m}}_k + \frac{1}{2}\left((\mathbf{x} - \hat{\mathbf{m}}_k)^{\mathrm{T}} \hat{S}_k^{-1} \Delta\hat{S}_k \hat{S}_k^{-1}(\mathbf{x} - \hat{\mathbf{m}}_k) - \mathrm{tr}(\hat{S}_k^{-1}\Delta\hat{S}_k)\right) + \frac{\Delta\hat{\pi}_k}{\hat{\pi}_k}\right]$$

Thus:

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

To compute an approximation of the KL divergence $\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)]$ we start from:

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

Expanding the logarithm up to the second order, we obtain:

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

The expectation of the first term of the former equation over $p(x|\hat{\mathbf{w}})$ is 0, leading to the following result:

$$\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)] \simeq \frac{1}{2}\int \frac{\left[\sum_k \hat{\pi}_k \mathcal{N}(x|\hat{\mathbf{m}}_k,\hat{S}_k)\Delta\psi_k\right]^2}{p(x|\hat{\mathbf{w}})}dx \tag{25}$$

The former equation can be rewritten in the following way:

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

where u_j is the responsibility function of x for the class j. We can use again the approximation where we neglect the terms when $j \neq r$. In other words, we approximate $u_j\mathcal{N}(x|\hat{m}_r,\hat{s}_r^2)$ with zero when $j \neq r$, and with $\mathcal{N}(x|\hat{m}_r,\hat{s}_r^2)$ when $j = r$. This leads to:

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

The integral yields:

$$\mathrm{KL}[p(x|\hat{\mathbf{w}})\|p(x|\hat{\mathbf{w}}^*)] \simeq \frac{1}{4}\sum_k \hat{\pi}_k \frac{u_{*k}^2}{(n_k^*)^4}\left[n_k^2\hat{z}_k^4 + 2u_{*k}(n_k^*)\hat{z}_k^2 + (n_k^*)^2\right] + \frac{1}{2}\sum_k \frac{(u_{*k}-\hat{\pi}_k)^2}{\hat{\pi}_k(n+1)^2} \tag{26}$$

Again we have defined:

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

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

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

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

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

[14] D. Peel and G. J. Mclachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.