A Modified Energy Statistic for Unsupervised Anomaly Detection

Rupam Mukherjee

GE Research, Bangalore, Karnataka, 560066, India
rupam.mukherjee@ge.com, rupam.mukherjee@gmail.com

Abstract

For prognostics in industrial applications, the degree of anomaly of a test point from a baseline cluster is estimated using a statistical distance metric. Among different statistical distance metrics, *energy distance* is an interesting concept based on Newton's Law of Gravitation, promising simpler computation than classical distance metrics. In this paper, we review the state of the art formulations of energy distance and point out several reasons why they are not directly applicable to the anomaly-detection problem. Thereby, we propose a new energy-based metric called the $\mathcal{P}$-statistic which addresses these issues, is applicable to anomaly detection and retains the computational simplicity of the *energy distance*. We also demonstrate its effectiveness on a real-life data-set.

1. Introduction

Prognostics is a critical requirement in many industrial fields today owing to the potential of cost savings and operational efficiency entitlement through elimination of unscheduled failures and shut-downs. One of the many important modules used in a typical Prognostics and Health Management (PHM) application is a health indicator module for the system under consideration which not only estimates a health metric but also tracks the evolution of this metric with time. The health indicator is estimated from a collection of sensor readings at different timestamps. Generally, one of many statistical distances of a test-point from a baseline cluster may be used as this health indicator. The data-points can be multi-dimensional resulting from multiple sensor outputs that can be tapped from the system under monitoring. This statistical distance or some function of it may be used as the *health metric* or the *health indicator*.

Time trending of the health metric as a function of historical and future forecast usage patterns give us a visibility into the Remaining Useful Life (RUL) which is the ultimate target of PHM. However, even before hitting the ultimate target of RUL, doing *anomaly detection* as an intermediate step has value. Anomaly is flagged when the distance metric exceeds a threshold and an alert is generated, thereby preventing sudden unplanned failures. Although as is captured in (Goldstein & Uchida, 2016), anomaly may be both *supervised* and *unsupervised* in nature depending on the availability of labelled ground-truth data, in this paper *anomaly detection* is considered to be the *unsupervised* version which is more convenient and more practical in many industrial systems.

Choice of the statistical distance formulation has an important impact on the effectiveness of RUL estimation. In literature, statistical distances are described to be of two types as follows.

1. **Divergence measures:** These estimate the distance (or, similarity) between probability distributions. Some common divergence measures are *Kullback-Leibler divergence*, *Jensen-Shannon divergence* and *Hellinger distance*.

2. **Distance measures:** These measures estimate the distance between a single point and a distribution by comparing it with a sample drawn from the distribution. The most well-known measure in this category is *Mahalanobis distance* (Mahalanobis, 1936). A few other distance measures are *Bhattacharya distance* (Bhattacharyya, 1943) and the *energy distance*.

For the industrial anomaly detection problem, it is mostly the second category which is more significant because most of the times, we end up comparing a single point with a baseline cluster.

In this paper, we are interested in the class of distances called *energy distance*. These are interesting because they are based on the notion of Newton's law of gravitational energy and considers statistical observations as celestial objects having gravitational pull between each other. A distance metric called $\mathcal{E}$-statistic may be written to represent the *energy distance* between distributions.The $\mathcal{E}$-statistic can be used to test the statistical hypothesis of equality of two distributions. This concept was proposed and developed in (Székely, Rizzo, et al., 2004; Székely & Rizzo, 2013; Szekely & Rizzo, 2017) where it was shown the $\mathcal{E}$-statistic is more general and powerful than many classical statistics.

https://doi.org/10.36001/IJPHM.2021.v12i1.1323

Application of energy distances to single and multi sample goodness of fit have been explored in (Rizzo, 2002b, 2002a; Székely et al., 2004; Baringhaus & Franz, 2004; Székely & Rizzo, 2005; Rizzo, 2009; Yang, 2012). Several other applications have been shown in (Szekely, Rizzo, et al., 2005; Székely & Rizzo, 2009; Feuerverger, 1993; Matteson & James, 2014; Kim, Marzban, Percival, & Stuetzle, 2009).

In this paper, we are interested in exploring it further because of its ability to work with Euclidean distances between data-points rather than with the data-points themselves. This ability leads to reduced computational complexity compared to the Mahalanobis distance and makes the $\mathcal{E}$-statistic potentially advantageous for memory and computation challenged systems.

However, despite the aforementioned advantages, we found some problems with the $\mathcal{E}$-statistic when applied to *anomaly detection*. These problems arise from the fact that the $\mathcal{E}$-statistic has been developed primarily for comparing equality between two distributions whereas for anomaly detection we need to compare a single test point with a distribution (baseline). In this paper, we analyse these problems and propose a new metric called the *modified energy distance* ($\mathcal{P}$-statistic) based on the notion of Newton's law which addresses these problems and is more suitable to be used in an anomaly detection problem.

Here, we would like to mention that detailed comparison of the proposed metric against all other classical distance metrics is a task we would attempt in future work. In this work, we focus on establishing the improvements over the $\mathcal{E}$-statistic.

This paper is laid out as follows. In Section 2, we briefly review the $\mathcal{E}$-statistic and mention the reasons why it is computationally simpler than other techniques. In Section 3, we explain the issues which make the $\mathcal{E}$-statistic less suitable for anomaly detection and in Section 4, we propose the $\mathcal{P}$-statistic and through the use of synthetic data, show that it addresses these issues. In Sections 5.1 and 5.2, we derive a method for estimating probability from the $\mathcal{P}$-statistic using a chosen parametric distribution. In Section 5.4, we show how the performance of this new metric compares in terms of discrimination performance against that of the Mahalanobis Distance, a classical multi-dimensional distance metric. In Section 6, we demonstrate a simple application of the $\mathcal{P}$-statistic on real-life data. In Section 7, we show how the training time of the proposed metric compares against that of Mahalanobis Distance for an incremental baseline update approach. Finally, in Section 8, we summarize the observations.

2. REVIEW OF ENERGY DISTANCE AND ITS COMPUTATIONAL SIMPLICITY

Let X and Y be two independent real-valued random variables with probability density functions f_X and f_Y respectively. Let $X_S = \{X_1, X_2, ..., X_{n_1}\}$ be a random sample of size n_1 drawn from the density function f_X. Similarly, $Y_S = \{Y_1, Y_2, ..., Y_{n_2}\}$ is a random sample of size n_2 drawn from the density function f_Y. The energy distance $\mathcal{E}(X, Y)$ between the distributions f_X and f_Y has been defined in (Székely et al., 2004; Székely, 1989, 2002). It is a population statistic for the pair of random variables X and Y.

Now, a sample statistic $\mathcal{E}_{n_1 n_2}$ may be used as an estimate for the population statistic $\mathcal{E}(X, Y)$. It may be calculated from the pair of samples (X_S, Y_S) as defined in (Székely et al., 2004; Székely, 1989, 2002) and has the following form.

$$\mathcal{E}_{n_1 n_2}(X_S, Y_1, Y_2, ...Y_{n_2}) = \frac{n_1 n_2}{n_1 + n_2}\left[\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2} ||X_i - Y_m||_2 - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}||X_i - X_j||_2 - \frac{1}{n_2^2}\times \sum_{l=1}^{n_2}\sum_{m=1}^{n_2}||Y_l - Y_m||_2\right]. \quad (1)$$

For a d-dimensional space, each observation X_i and Y_j is a $d \times 1$ array. Mathematically, the entire sample may be represented as a $d \times n_1$ matrix for X_S and $d \times n_2$ matrix for Y_S.

In industrial systems, usually after a baseline data-set is accumulated, future test-points are compared against it and the baseline itself is not frequently replaced except when there is a need to re-establish the baseline. Thus, although the baseline cluster varies depending on which time instant it was acquired from sensor readings and hence has a random nature, it is considered a constant in the anomaly analysis once it is captured and saved. Hence, in subsequent analysis, it is considered invariant. Going forward, in this paper, we will be referring to the sample X_S as the *baseline cluster* (or *baseline*, for simplicity).

For the anomaly detection problem, we want to compare a single point against a baseline cluster having many members. Hence, in (1), we assume that Y_S has sample-size 1 and contains only one member Y_1. We discard the notations n_1 and n_2 as $n_2 = 1$. We represent n_1 by n going forward as the subscript in n is no longer needed. For sufficiently large sample size for X_S, $n_1 \approx n_1 + 1$. Also, we use y in place of Y_1 in future analysis since the subscript is not crucial for clarity in representing a single member set Y_S.

From (1), we write a simplified form $\mathcal{E}_n$ for $\mathcal{E}_{n_1 n_2}$ as

$$\begin{aligned}\mathcal{E}_n(X_S, Y_1) &= \mathcal{E}_n(X_S, y) \\ &\approx \frac{2}{n}\sum_{i=1}^{n} ||X_i - y||_2 \\ &\quad -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||X_i - X_j||_2 \,. \end{aligned} \tag{2}$$

Reference (Székely et al., 2004) mentions that for $\mathcal{E}_n(X_S, y)$ to be a valid distance metric, there must exist a c_α for every $\alpha \in (0, 1)$ such that

$$P(Z > c_\alpha) = \alpha. \tag{3}$$

Here, Z is a random variable. A random sample z drawn from the distribution of Z is a function of the baseline X_S and a random sample x drawn from the distribution f_X defined earlier. It takes a form

$$z = \mathcal{E}_n(X_S, x). \tag{4}$$

As mentioned in Section 1 and in (Székely et al., 2004; Székely & Rizzo, 2013; Szekely & Rizzo, 2017), the $\mathcal{E}$-statistic is simpler and more general than classical distance metrics. While writing this paper, on comparing (1) with the classical distance metrics described in (*Statistical distance*, n.d.), the advantages which specifically interested us are

1. The Euclidean distances in (1) can be computed in parallel and hence the computation time of $\mathcal{E}$-statistic does not scale with data size if implemented in a parallel manner.
2. Covariance matrices are not required in (1). For high-dimensional data, this can take up a significant amount of computation and memory
3. Matrix inversion is not needed. In many classical methods, inversion of covariance matrices is required. Again, for high dimensional data, this can involve significantly heavy computation.

3. SOME GAPS IDENTIFIED IN $\mathcal{E}$-STATISTIC

In this section, we analyse the general requirements from any anomaly detection metric and point out some weaknesses in the $\mathcal{E}$-statistic which prevent it from satisfying some of these conditions. Going forward, these weaknesses form the basis of the proposed innovation in this paper.

3.1. Requirements from an Acceptable Metric for Anomaly Detection

In anomaly detection, we usually obtain a single test sample y and compare it against the baseline (say X_S) which is a cluster of samples drawn from the distribution f_X. Although there is an element of randomness in the baseline cluster based on when that data set was captured in time, but once they are collected for any machinery, they are usually not changed during the comparison or anomaly detection phase. Thus, for all practical purposes, they are same as a set of multi-dimensional real-valued points.

From the way anomaly detection is usually implemented in industrial systems, we intuitively desire the following conditions from any acceptable anomaly measure (say $\mathcal{E}_n^*$) which may be used in a practical anomaly detection system.

The desired characteristics with respect to cluster size and distance from centroid along with their mathematical expressions are as follows.

1. **Relationship with cluster size**

   (a) For the test-point y situated at a given distance from the centroid of the baseline cluster and outside the baseline, it should appear less anomalous if it is closer to the outer boundary of the baseline cluster.

   Given the saved baseline X_S, we define a random variable R such that

   $$R = ||X - \overline{X_S}||_2 \text{ and} \tag{5}$$

   R has a probability density function f_R whose model parameters may be estimated from the sample X_S. Here, $\overline{X_S}$ is the centroid of the baseline cluster which is nothing but the sample mean.

   For the test-point y, let $r = ||y - \overline{X_S}||_2$. Since, f_R is a function of the baseline cluster X_S and from (5) the domain of R is the set of Euclidean distances of all possible samples of X from $\overline{X_S}$, the probability density of any point with distance r from $\overline{X_S}$ may be expressed as $f_R(r, X_S)$.

   This problem may now be cast in the form of a hypothesis test as follows.

   Null hypothesis(H_0): Random samples drawn from f_R will be more extreme than r. y is flagged as an anomaly with respect to X_S if the null hypothesis is rejected.

   Rejection criterion: The null hypothesis is rejected if

   $$\begin{aligned} p_R(X_S) &= P(\left|\left|X - \overline{X_S}\right|\right|_2 > r) = P(R > r) \\ &= \int_r^\infty f_R(x, X_S)dx < p_{\text{th}} \end{aligned} \tag{6}$$

   which is a pre-determined threshold.

   Let there be a second cluster $X_S^{'}(\beta)$ which is formed by scaling the baseline with respect to its centroid such that its kth element $X_S^{'}(\beta)_k$ may be written as

$$X_S^{'}(\beta)_k = \beta(X_k - \overline{X_S}) + \overline{X_S} \text{ where} \quad (7)$$

$\beta > 0$. If $\beta > 1$, the number of points, relative orientation of points and cluster shape are maintained unchanged between X_S and $X_S^{'}(\beta)$ with the second cluster occupying a larger spatial volume. Hence, the test-point will appear closer to the outer boundary of $X_S^{'}(\beta)$ than to that of X_S. In this scenario, if $\mathcal{E}_n^*$ is a faithful indicator of anomaly, it should appear to be less in the former case. This behaviour may be expressed mathematically as

$$\mathcal{E}_n^*(X_S^{'}(\beta), y) < \mathcal{E}_n^*(X_S, y) \text{ if } \beta > 1. \quad (8)$$

Now, we saw that a larger radius of baseline is supposed to make y appear less anomalous. Also, larger the anomaly, less should be the value of the integral in (6) since larger anomaly would mean less net probability of having points more extreme. Hence,

$$p_r(X_S^{'}(\beta)) > p_r(X_S) \text{ if } \beta > 1. \quad (9)$$

(b) For an infinitesimally small baseline cluster, any test-point would appear to have a very large anomaly, irrespective of the value of the Euclidean distance from the centroid. This is because the distance always appears large relative to the cluster size.

Hence,

$$\lim_{\beta \to 0} \mathcal{E}_n^*(X_S^{'}(\beta), y) = \infty \quad (10)$$

With the anomaly metric increasing asymptotically, we will have an asymptotic reduction of the integral value in (6).

2. **Relationship with distance from cluster centroid**

(a) If the test-point is further away from the centroid of the baseline cluster, it should appear more anomalous and hence, the anomaly metric should increase. Thus, if there are two test-points y_1 and y_2,

$$\mathcal{E}_n^*(X_S, y_1) > \mathcal{E}_n^*(X_S, y_2) \text{ if} \quad (11)$$
$$\left|\left|y_1 - \overline{X_S}\right|\right|_2 > \left|\left|y_2 - \overline{X_S}\right|\right|_2. \quad (12)$$

(b) If the test-point is at the center, the anomaly metric should be close to zero. Hence,

$$\mathcal{E}_n^*(X_S, y) \to 0 \text{ if } \left|\left|y - \overline{X_S}\right|\right|_2 \to 0. \quad (13)$$

It should be noted that (7) and (11) indicate that the ideal *anomaly metric* should follow a *monotonic* behaviour with respect to the baseline size and also distance of the test-point from the baseline cluster. These ideal conditions hold good for data of any dimension.

3.2. Synthetic Dataset

In order to examine how $\mathcal{E}_n$ performs with respect to the desired characteristics stated in Section 3.1, we consider a d-dimensional random variable X which is distributed as a uniform ball. We sample from this distribution to create a synthetic data-set following the method described in (Harman & Lacko, 2010).

X can be written as

$$X = r_b \left(Z_1 / \left|\left|Z_1\right|\right|_2\right) Z_2^{1/d} \text{ where} \quad (14)$$

Z_1 is sampled from a multi-variate uncorrelated standard normal distribution, $Z_2 \sim U(0,1)$ and r_b is the desired radius of the ball.

For the purpose of this study, we restrict the dimension of X in (14) to 2 to keep the analysis and visualization simple. If the two dimensions of X are written as $X^{(1)}$ and $X^{(2)}$, they may be expressed in a simplified form as a function of two random variables R and θ in the following way.

$$X^{(1)} = R\cos\theta \text{ and } X^{(2)} = R\sin\theta \text{ where}$$
$$R \sim (0, r_b)^{1/2} \text{ and } \theta \sim U(0, 2\pi) \text{ and} \quad (15)$$

r_b is the chosen radius of the example cluster. It may be shown from (15) that $\forall\ (\theta, R)$, the probability density function $f_{\theta,R} = 1/(\pi r_b^2)$. Thus, the distribution is uniform in nature. We also consider a fixed test-observation with location $\theta = 0$ and $R = 1$.

3.3. Shortcomings of $\mathcal{E}$-statistic

We sweep r_b in (14) from 0 to 5 thereby progressively getting a different-sized baseline and thereby changing Δ in (7). This implies different probabilities of the fixed test-point belonging to this cluster. We then evaluate the simplified $\mathcal{E}_n$ in (2).

Figures 1 and 2 show the baselines and test-points for two different values of r_b along with computed values of $\mathcal{E}_n$ for both the cases. Figure 3 shows a continuous plot of how $\mathcal{E}_n$ varies with r_b.

From these plots, the $\mathcal{E}$-statistic can be shown to violate the desired characteristics mentioned in Section 3.1 in the following manner.

1. **Limiting condition behaviour:** The $\mathcal{E}$-statistic violates the limiting conditions in the following manners.

(a) It has been shown in Appendix A (59) that the metric $\mathcal{E}_n$ proposed in (2) has the following property.

$$\lim_{\beta \to 0} \mathcal{E}_n(X_S^{'}(\beta), y) = 2\left|\left|\overline{X_S} - y\right|\right|_2 \neq \infty. \quad (16)$$

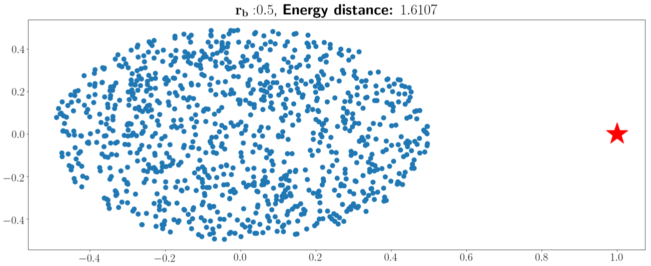

Figure 1. Baseline (blue) and test-point (red) with $r_b = 0.5$ and $\mathcal{E}_n = 1.6107$

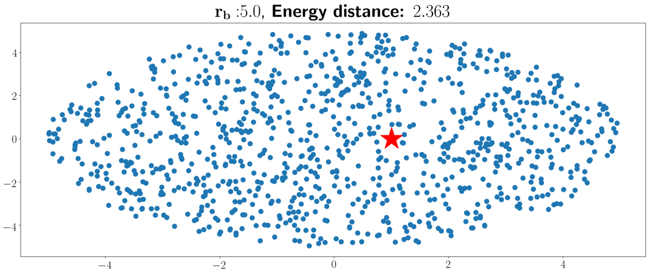

Figure 2. Baseline (blue) and test-point (red) with $r_b = 5.0$ and $\mathcal{E}_n = 2.363$

This violates the condition required for the ideal metric $\mathcal{E}_n^*$ stated in (10). Thus, $\mathcal{E}_n$ in its present form cannot function as the ideal metric.

From Figure 3, it is clear that when the baseline cluster becomes progressively smaller and close to 0, the variation of $\mathcal{E}_n$ is not asymptotic to infinity but hits a finite value equal to the distance of the test-point from the baseline centroid.

(b) Appendix A (62) also proves the following.

$$\lim_{\beta\to\infty} \mathcal{E}_n(X_S^{'}(\beta), y) = \infty. \tag{17}$$

From Figure 3, it is clear that when baseline cluster increases in radius, the value of $\mathcal{E}_n$ does not always reduce tending towards 0. It goes through an inflection point beyond which it increases instead of reducing and reaches significantly high values instead of very small ones. For the particular case in Figure 2, the $\mathcal{E}_n$ value of 2.363 is not appropriate. It should be very close to zero. Also, the $\mathcal{E}_n$ should be higher in Figure 3 compared to Figure 2. However, the opposite is actually observed.

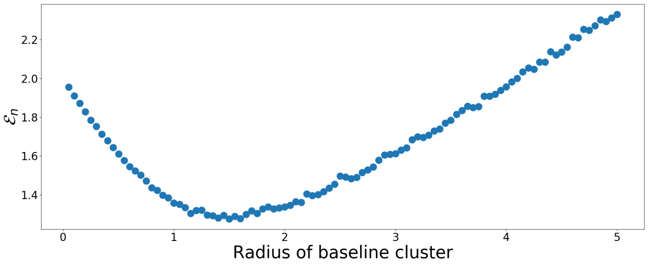

Figure 3. Variation of $\mathcal{E}_n$ with baseline cluster radius (r_b), given a fixed test-point.

(c) As stated above, in Section 3.1, $\mathcal{E}_n^* = 0$ *if and only if* $y = X_c$ for a viable distance-metric for anomaly detection. But this condition is violated by the $\mathcal{E}$-statistic in its present form because in (1), $\mathcal{E}_{n_1 n_2} = 0$ *if and only if* the distributions X and Y are identical i.e. $X = Y$. However the single test-point y considered here forms a single member distribution which cannot be identical to X under any circumstance.

2. **Monotonic behaviour:** From Figure 3, it is seen that the value of $\mathcal{E}_n$ goes through an inflection point as r varies instead of varying monotonically as required in Section 3.1

The issues with $\mathcal{E}$-statistic raised in this section have been further addressed and a modified algorithm suggested in Section 4. Since the target in this paper is anomaly detection, we have also extended the analysis to reach a closed form analytical expression for the probability density function (PDF) based on which the p-value of the test-point may be estimated.

4. MODIFIED ENERGY DISTANCE

The reason why energy distance is computationally simple is that it works with distributions of Euclidean distances whereas other distance metrics first fit a multivariate distribution and then work our probabilities from that. The $\mathcal{E}$-statistic is one of the first energy-based metric proposed for anomaly detection. However, we saw in Section 3.3 that it has a few weaknesses. In this section, we borrow the concept of working with Euclidean distances from a study of the $\mathcal{E}$-statistic and propose an alternative and slightly modified metric based on them.

The potential energy of a configuration consisting of two particles having masses m_1 and m_2 respectively and having positions X_1 and X_2 respectively is the work that needs to be done to move one particle from infinity to its current position against the gravitational force exerted by the other particle already in place. It can be written as

$$PE(X_1, X_2) = -\frac{Gm_1m_2}{||X_1 - X_2||_2}. \tag{18}$$

We define the *incremental potential energy* of y to be the work that needs to be done to move the test-point y from infinity to its present position against the collective force exerted by the baseline sample cluster X_S. The *incremental potential energy* may be written as

$$\begin{aligned} \Delta P(X_S, y) &= \sum_{k=1}^{n} PE(X_k, y) \\ &= -\sum_{k=1}^{n} \frac{C}{||X_k - y||_2}, \end{aligned} \tag{19}$$

where C is a positive constant if we assume that each data-point in the baseline sample cluster X_S as well as the test-point y represent particles of equal mass.

The net potential energy for assembling the baseline cluster X_S may be written as

$$\Delta P(X_S) = \sum_{j=2}^{n} \sum_{k<j} PE(X_k, X_j) \tag{20}$$

Since from (18), potential energy is inversely proportional to distance, we may assume a *modified energy distance* E_n between X_S and y such that

$$\begin{aligned} E_n(X_S, y) &= -1/\Delta P(X_S, y) \\ &= 1 \Big/ \sum_{k=1}^{n} \frac{C}{||X_k - y||_2} \text{ from (19).} \end{aligned} \tag{21}$$

Also, the effective *modified energy distance* of the entire cluster X_S with respect to itself may be written as

$$\begin{aligned} E_n(X_S) &= -1/\Delta P(X_S) \\ &= 1 \Big/ \sum_{j=2}^{n} \sum_{k<j} \frac{C}{||X_k - X_j||_2}. \end{aligned} \tag{22}$$

Considering the fact that the degree of anomaly should factor in the total energy content of the baseline cluster itself, we define a *potential energy statistic*, the $\mathcal{P}$-statistic based on the *modified energy distance* formulated in (21) as

$$\begin{aligned} \mathcal{P}_n(X_S, y) &= \frac{E_n(X_S, y)}{E_n(X_S)} \\ &= \sum_{j=2}^{n} \sum_{k<j} \frac{1}{||X_k - X_j||_2} \Big/ \sum_{k=1}^{n} \frac{1}{||X_k - y||_2} \end{aligned} \tag{23}$$

One concern with (23) might be numerical stability as y overlaps with any particular X_k and the denominator in (23) goes to ∞. In order to overcome this, we propose introducing a saturation term τ as

$$\mathcal{P}_n(X_S, y) = \sum_{j=2}^{n} \sum_{k<j} \frac{1}{||X_k - X_j||_2} \Big/ \sum_{k=1}^{n} \frac{1}{||X_k - y + \tau||_2} \tag{24}$$

If the kth dimension of τ be τ^k, $\tau^1 = \epsilon$ and $\tau^j = 0 \; \forall \; j > 1$. Here, ϵ is an infinitesimal positive number on the real line. The closer ϵ is to 0, the better is the match between (23) and (24). A guideline may be to use

$$0 < \epsilon < 1\text{e}{-04} \frac{1}{n} \sum_{k=1}^{n} \left|\left|X_k - \overline{X_S}\right|\right|_2. \tag{25}$$

The reader may note that this is a guideline and one may use smaller values of ϵ if they want. However, larger values of ϵ may not be advisable. We would like the reader to know that a rigorous evaluation of the impact of the choice of this value on the results needs to be conducted as future work but at present, we do not see this as a big risk to this method.

We now examine the limiting conditions or $\beta \to 0$ and $\beta \to \infty$ which have been discussed with respect to $\mathcal{E}_n$ in (16) and (17) respectively. Using the definition of β as given in (7),

$$\begin{aligned} \lim_{\beta \to 0} \mathcal{P}_n(X_{S1}, y) &= \lim_{\beta \to 0} \left(\sum_{j=2}^{n} \sum_{k<j} \frac{1}{||\beta(X_j - X_k)||_2} \right) \Bigg/ \\ &\quad \left(\sum_{k} \frac{1}{||\beta X_k - \beta \overline{X_S} + \overline{X_S} - y||_2} \right) \\ &= \infty. \end{aligned} \tag{26}$$

Using the same data-set and sweep parameters as in Figure 3, Figure 4 plots $\mathcal{P}_n$ vs. radius of baseline cluster for a fixed test-point. In it, we notice the asymptotic behaviour of $\mathcal{P}_n$ near infinitesimal baseline radius i.e., $\beta = 0$ as implied by (26).

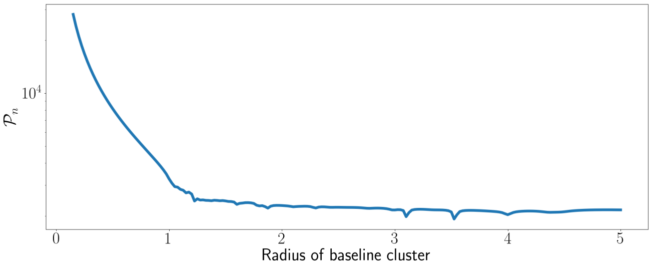

Figure 4. Semi-log plot for variation of $\mathcal{P}_n$ with baseline cluster size, given a fixed test-point.

Since y is an n-dimensional point, $\mathcal{P}_n$ is a function of the distance r and several other variables related to the other degrees of freedom. Thus, functionally, $\mathcal{P}_n$ may be written as

$$\mathcal{P}_n(X_S, y) = \mathcal{P}_n(X_S, r, D) \text{ where} \tag{27}$$

D represents the set of remaining degrees of freedom other than r. While D represents the radial directional vector in the high dimensional space with respect to $\overline{X_S}$, r indicates position along that vector.

We now want to study the impact of r on $\mathcal{P}_n$ for any given D.

Assuming X_S to be a fixed baseline against which test-points are compared, we can consider D and X_S to be constants. Hence, we have a simplified functional relationship as follows.

$$\mathcal{P}_n(r) = \mathcal{P}_n(X_S, r, D). \tag{28}$$

Thus, if we consider the space as a high dimensional ball, we are studying the variation of $\mathcal{P}_n$ for points at different radial distances along a given radial vector direction.

In Appendix B, we prove the following regarding the geometric properties for the function $\mathcal{P}_n()$.

$$\exists \text{ a finite } r^* \text{ such that } \forall\, r > r^*,\ \frac{d\mathcal{P}_n}{dr} > 0, \tag{29}$$

$$\lim_{r\to\infty} \frac{d\mathcal{P}_n}{dr} = \text{constant from (71)}, \tag{30}$$

$$\mathcal{P}_n(r) \text{ is finite } \forall\, r < \infty, \text{ from (72) and} \tag{31}$$

$$\mathcal{P}_n(r) \geq 0\ \forall\, r \text{ from (73).} \tag{32}$$

Now, we examine the impact of increasing the spatial volume of the baseline cluster. Following the proposition in (7), we examine the impact of increasing β.

From (23), using the definition of β in (7),

$$\begin{aligned} \mathcal{P}_n(X_{S1}, y) &= \mathcal{P}_n\left(X_S, \frac{y}{\beta} + \overline{X_S}\left(1 - \frac{1}{\beta}\right)\right) \\ \Rightarrow \lim_{\beta\to\infty} \mathcal{P}_n(X_{S1}, y) &= \mathcal{P}_n(X_S, \overline{X_S}) < \infty \end{aligned} \tag{33}$$

if all the dimensions of y are finite.

From (33), for large spatial volume of the baseline i.e., for large values of β, $\mathcal{P}_n(r)$ will be finite unlike in (17) and in Figure 3 where the $\mathcal{E}$-statistic is seen to increase with increasing β after a certain value and tends to infinity. Also, in Figure 4, we see that as the baseline cluster becomes progressively larger thereby indicating a reduction in anomaly level, we see a reduction of $\mathcal{P}_n$ to a value of 0 instead of the inflection seen in Figure 3.

The variation of $\mathcal{P}_n$ with distance of test-point from baseline centroid, for a few values of the baseline radius is shown in Figure 5. It serves as a demonstration of the fact that as per (29), after a small initial value of r, $\mathcal{P}_n$ increases monotonically with increase of distance of the test-point from the baseline centroid. This is consistent with the fact that farther the test point, the more anomalous it is.

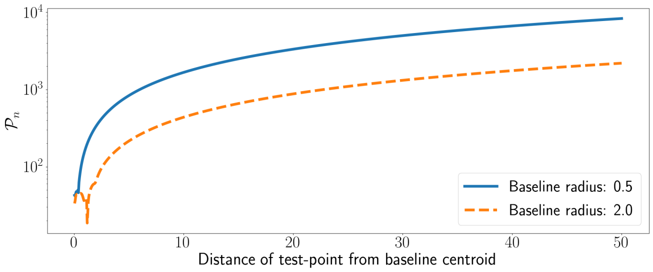

Figure 5. Semi-log plot for variation of $\mathcal{P}_n$ with distance from centroid, for different baseline sizes.

We also observe from Figure 5 that the rate of increase of $\mathcal{P}_n$ is less when the baseline cluster size is larger. This is because for any fixed test-point, it will appear less anomalous when the baseline cluster has a larger radius.

All of these observations are consistent with $\mathcal{P}_n$ being a valid anomaly statistic. All the discrepancies observed in Figure 3 are addressed and $\mathcal{P}_n$ satisfies the requirements stated in Section 3.1. The observations are further validated in Figures 6 and 7 where the $\mathcal{P}$-statistic is used to evaluate the cases earlier analysed in Figures 3 and 2. They show a much lower $\mathcal{P}$-statistic value for the case in Figure 7 where the test-point is very close to the centroid $\overline{X}$.

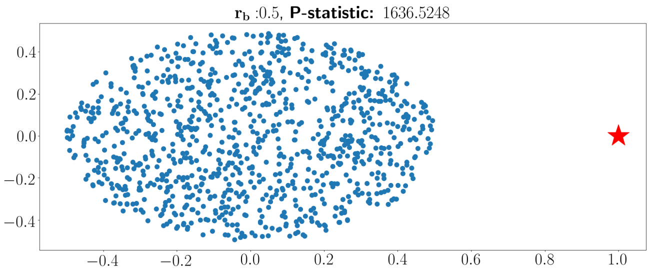

Figure 6. Baseline (blue) and test-point (red) with $\underline{r_b = 0.5}$ and $\underline{\mathcal{P}_n = 1636.5248}$

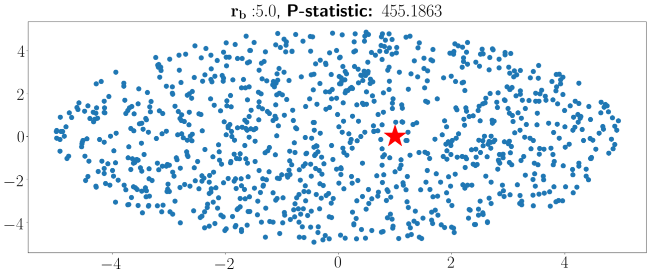

Figure 7. Baseline (blue) and test-point (red) with $\underline{r_b = 5.0}$ and $\underline{\mathcal{P}_n = 455.1863}$

The observations in this section establish that the variation of $\mathcal{P}$-statistic with baseline size and test-point distance is in accordance with what is expected intuitively from an anomaly metric. However, to fully establish its usefulness as a valid anomaly metric, we should be able to compute a *probability density function* (PDF) from this metric.

5. COMPUTATION OF PROBABILITY

5.1. Existence of a Valid Probability Measure using $\mathcal{P}_n$

Let r represent the distance of the test-point from the centroid of the baseline cluster. Also, in the space of real numbers, let

R represent a single-dimensional random variable having a density function f_R from which the r-values are assumed to be sampled. R has a range defined by

$$\text{Range}(R) = [0, \infty). \tag{34}$$

Since R is a random variable, $Z = \mathcal{P}_n(R)$ is also a random variable as the function of a random variable is itself a random variable. Let f_Z be the density function for Z and $z \in$ Range(Z).

In Appendix B, it has been shown that $\mathcal{P}_n()$ is a continuous function. Also, from (29), (30) and (32),

$$\text{Range(Z)} = [0, \infty) \tag{35}$$

Thus, the domain and the co-domain of $\mathcal{P}_n()$ are both $[0, \infty)$.

As mentioned in (Székely et al., 2004) and (3), for $\mathcal{P}_n(r)$ to be a valid metric for anomaly detection,

$$P(Z > c_\alpha) = \alpha, \; \forall \, c_\alpha \in \text{Range}(Z) \text{ and} \tag{36}$$

$$\alpha \in (0, 1), \tag{37}$$

α being unique for every choice of c_α. (36) and (37) are characteristics necessary for the existence of the cumulative density function (CDF) corresponding to the probability density function f_Z.

From (29) and (31), $\exists \, r^{**} \geq r^*$ such that

$$\mathcal{P}_n(r) > \mathcal{P}_n(r^{**}), \; \forall \, r > r^{**} \text{ where} \tag{38}$$

$$\begin{aligned} \mathcal{P}_n(r^{**}) &= \max\left(\{\mathcal{P}_n(r) | r \in [0, r^{**}]\}\right), \\ &= c_{\max} \text{ (say), a finite quantity.} \end{aligned} \tag{39}$$

These relations are true because r and r^{**} are all positive numbers. Also, since probability is a non-negative quantity, it may be shown that

$$\min\left(\{\mathcal{P}_n(r) | r \in [0, r^{**}]\}\right) \geq 0. \tag{40}$$

We now, analyse the behaviour of $\mathcal{P}_n(r)$ in two zones, i.e., $c_\alpha \leq c_{\max}$ and $c_\alpha > c_{\max}$. These are analysed as follows.

1. **Case 1:** $c_\alpha \leq c_{\max}$.

   For any $c_\alpha \in [0, c_{\max}]$,

   $$\mathcal{P}_n^{-1}(c_\alpha) = X_\alpha = \{r | \mathcal{P}_n(r) = c_\alpha\}. \tag{41}$$

   From (29), $\mathcal{P}_n(r)$ need not be strictly increasing or strictly decreasing in nature if $r \leq r^{**}$. Hence the set X_α may have more than one member as defined in (41). Since $c_\alpha < c_{\max}$, (39) tells us that all members of the set X_α in (41) lie within the interval $[0, r^{**}]$.

   Let (X, Σ) and (Y, T) be measurable spaces such that X is a set comprising of all samples of R between 0 and r^{**}. X and Y are equipped with σ-algebras Σ and T respectively. Also, let $\mathcal{P}_n$ be a function from X to Y. Since $\mathcal{P}_n(r)$ is strictly increasing for $r > r^{**}$, $\max(X_\alpha)$ defined in (41) cannot be greater than r^{**} with $c_\alpha \leq c_{\max}$. Also, by definition $r \geq 0$.

   Hence, if we consider any open set of $\mathcal{C}$ consisting of values from $[0, c_{\max}]$, from (41), its inverse values $\mathcal{P}_n^{-1}(\mathcal{C})$ will map to the region $[0, r^{**}]$. None of it would lie outside of this. Mathematically, this may be written as

   $$\begin{aligned} & \mathcal{P}_n^{-1}(\mathcal{C}) = \{x | \mathcal{P}_n(x) = \mathcal{C}\} \; \in \; \Sigma, \text{ for } \mathcal{C} \in T \\ \Rightarrow \;\; & \mathcal{P}_n : X \to Y \text{ is a measurable function.} \end{aligned} \tag{42}$$

   Thus, $\mathcal{P}_n(R)$ is a valid choice for a random variable. We can show that $\forall \, c_\alpha$, $\exists$ a unique values of α_1

   $$P(\mathcal{P}_n(R) \; \in \; [c_\alpha, c_{\max}] = \alpha_1 < P(R \; \in \; [0, r^{**}]). \tag{43}$$

   Also, from (29), $\mathcal{P}_n^{-1}()$ is single-valued in for $r > r^{**}$ and hence,

   $$P(\mathcal{P}_n(R) > c_r) = P(R > \mathcal{P}_n^{-1}(c_r)) = P(R > r) \text{ (say)}. \tag{44}$$

   From (43) and (44), for $r \; \in \; [0, r^{**}]$,

   $$\begin{aligned} P(\mathcal{P}_n(R) > c_\alpha) &= P(\mathcal{P}_n(R) \; \in \; [c_\alpha, c_{\max}]) \\ &\quad + P(\mathcal{P}_n(R) > c_{\max}) \\ &= \alpha_1 + P\left(R > \mathcal{P}_n^{-1}(c_{\max})\right) \\ &= \alpha_1 + P(R > r^{**}) \\ \Rightarrow P(\mathcal{P}_n(R) > c_\alpha) &= \alpha \; \in \; (0, 1) \text{ (say).} \end{aligned} \tag{45}$$

   In (45), α is uniquely defined for any chosen value of c_α, since α_1 is unique as shown in (43).

2. **Case 2:** $c_\alpha > c_{\max}$.

   Since $\mathcal{P}_n(r)$ is monotonically increasing for $r > r^{**}$, $\mathcal{P}_n^{-1}()$ is well-defined and single-valued. Hence, for any c_α,

   $$\begin{aligned} P(\mathcal{P}_n(R) > c_\alpha) &= P(R > \mathcal{P}_n^{-1}(c_\alpha)) \\ &= P(R > r) \text{ (say).} \end{aligned} \tag{46}$$

   $$\Rightarrow P(\mathcal{P}_n(R) > c_\alpha) = \alpha \; \in \; (0, 1), \text{ (say).} \tag{47}$$

   It must be noted that (46) could be written only because $\mathcal{P}_n^{-1}()$ is well-defined in the chosen regime which is $r > r^{**}$. In (47), α is unique for every chosen value of c_α since the CDF of R is well-defined.

From (45) and (47), it is seen that (36) is validated over the entire range of the random variable $Z = \mathcal{P}_n(R)$. Thus, $\mathcal{P}_n()$ is a valid metric for anomaly detection.

5.2. Calculating p-value

In order to calculate the p-value, we need to first choose a suitable form for the PDF $f(x)$ that best represents the data being analysed. Different parametric distributions may be chosen for the PDF $f(x)$. Since $\mathcal{P}_n(x) > 0 \,\forall\, x$ from (32), we assume that it is sampled from a *standard half-normal distribution* with mean $\mu = 0$ and standard deviation σ. We can write a standard half-normal variant of $\mathcal{P}_n$ as

$$p_n = \mathcal{P}_n/\sigma \text{ and } p_n = |Z| \text{ where } Z \sim N(0,1). \tag{48}$$

Here, σ can be estimated as

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{P}_n(X_S - \{X_i\}, X_i)\right)^2}. \tag{49}$$

Here, it must be noted that since X_S is defined as a *set* of random samples, $X_S - \{X_i\}$ represents the *set subtraction* operation resulting in the complementary set obtained by eliminating X_i from X_S.

The p-value for the test-point y is

$$\begin{aligned} p(y) &= 1 - 2\int_0^{p_n} e^{-t^2/2}dt \\ &= 1 - \text{erf}\left(\frac{p_n}{\sqrt{2}}\right). \end{aligned} \tag{50}$$

An efficient computation of $p(y)$ would need a compact approximation for the *error function* **erf(x)**. For this purpose, we use the *Bürmann series* expansion which converges quickly for real values of x (Schöpf & Supancic, 2014). As explained in (Dixon, 1901; Schöpf & Supancic, 2014), the error function may be approximated as

$$\begin{aligned} \text{erf}(x) \approx\ & \frac{2}{\sqrt{\pi}}\text{sgn}(x)\sqrt{1-e^{-x^2}} \times \\ & \left(\frac{\sqrt{\pi}}{2} + \frac{31}{200}e^{-x^2} - \frac{341}{8000}e^{-2x^2}\right). \end{aligned} \tag{51}$$

From (50) and (51), $p(y)$ may be approximated as

$$p(y) \approx 1 - \frac{2}{\sqrt{\pi}}\sqrt{1-A}\left(\frac{\sqrt{\pi}}{2} + \frac{31}{200}A - \frac{341}{8000}A^2\right), \tag{52}$$

where

$$A = \exp\left(-\frac{p_n^2}{2}\right). \tag{53}$$

We henceforth use the compact expression of $p(y)$ in (52) in all calculations going forward.

Figure 8 shows how the $p(y)$ varies with distance from centroid and also the baseline sizes using the same test-cases as in Figure 5. We see that for any given test-point, the baseline cluster having a bigger radius will imply higher p-value and thus less chance of rejecting the null hypothesis. This is because the larger cluster would make the test-point appear less anomalous.

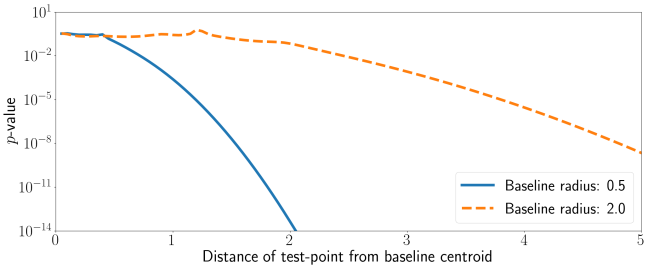

Figure 8. Semi-log plot for variation of $p(y)$ calculated in (50) with distance from the baseline centroid, for different baseline sizes.

5.3. Algorithm Steps and Computation Requirement

We can now summarize the steps involved in the computation of $p_d(y)$ in Algorithm 1.

Algorithm 1: Computation of p-value from $\mathcal{P}_n$

```
Input: Baseline cluster $X$
Output: $p(y)$
Data: Test-point $y$
Initialization: $flag = 0$
/* If baseline has not been characterized,
   enter baselining phase                          */
if $flag == 0$ then
    $X_c = 1/n \sum_{k=1}^{n} X_k$
    $A = 0$ for $j = 2 : n$ do
        for $k < j$ do
            $A+ = 1/\,||X_k - X_j||_2$
    $\alpha = 0$
    for $i = 1 : n$ do
        $B = 0$
        for $j = 1 : n$ do
            if $j \neq i$ then
                $B+ = 1/\,||X_j - Xi||_2$
        $\mathcal{P}_n(X_i) = A/B - 1$
        $\alpha+ = (\mathcal{P}_n(X_i))^2$
    Estimate $\hat{\sigma} = \sqrt{\alpha/n}$.
    Save $\hat{\sigma}$.
    $flag = 1$
else
    /* Estimate $p$-value for test point           */
    From (23), calculate $\mathcal{P}_n(y)$ for the test-point $y$.
    $p_n = \mathcal{P}_n/\hat{\sigma}$            // Transform $\mathcal{P}_n$ to $p_n$.
    Calculate $p(y)$ from (50).
```

5.4. Comparison of $\mathcal{P}$-statistic with Mahalanobis Distance

Since in this paper we are trying to formulate and establish a new statistical metric for anomaly detection, it would be interesting to see how it performs with respect to the Mahalanobis Distance (MD) which is one of the established classical techniques in this domain. The questions we are trying to answer at this stage are the following.

1. Is there a *one-to-one* correspondence between MD and p_n ? Or, does every value of p_n corresponds to one and only one value of MD ?
2. Is it possible to estimate MD from p_n for different dimensions of the data-points and different physical sizes of the baseline clusters ? By answering this question, we are trying to figure out if tribal knowledge about what constitutes anomaly in terms of MD for a particular application can be translated to the domain of the $\mathcal{P}$-statistic.

Let us represent the Mahalanobis Distance (MD) (Mahalanobis, 1936) of test-point y with respect to baseline cluster sampled from the random variable X as $M(y, X)$. Thus,

$$M(y, X) = \sqrt{(y - \overline{X})^T S^{-1} (y - \overline{X})} \text{ where} \tag{54}$$

the d dimensions of y and X are represented along the rows and S is the covariance matrix of the cluster sampled from X. We now examine the relationship between p_n and $M(y, X)$ in an empirical fashion by varying different parameters of the synthetic data-set designed in (14).

Similar to Figure 8, we sweep the distance of test-point from baseline centroid and calculate $M(y, X)$ and p_n for each position. Finally, the MD values corresponding to each value of p_n are plotted for comparison in Figure 9 for two different values of baseline radius. Also, we assume different values of dimension d in (14) and for each value of d, we define a baseline cluster X and test-point y as in (14). For each d and a given baseline radius, Figure 10 shows the relationship between p_n and $M(y, X)$.

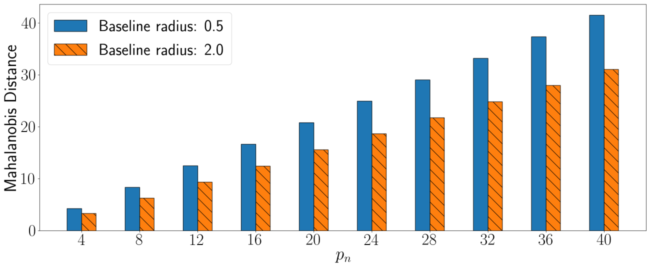

Figure 9. Relationship between p_n and $M(Y, X)$ for different values of baseline radius.

From Figures 9 and 10, p_n is linearly related to the Mahalanobis Distance and the relationship holds good for different dimensions of the problem-space. Thus, any threshold in $M(y, X)$ can be translated to a corresponding threshold in p_n by utilizing the linearity of the curves in Figure 10. This enhances the confidence that an existing threshold-based alerting strategy that depends on values of Mahalanobis Distance may be translated to an equivalent $\mathcal{P}$-statistic-based one. Hence, the p_n-value in (48) and the corresponding p-value calculated in (50) can be used effectively for anomaly detection.

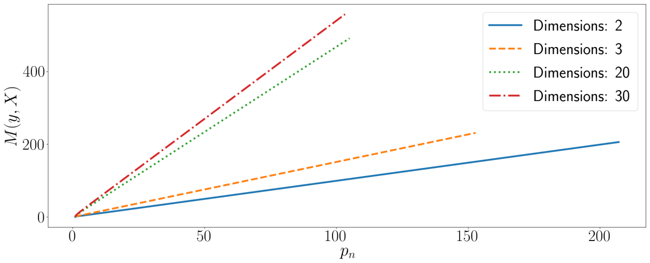

Figure 10. Relationship between p_n and $M(Y, X)$ for different values of dimension d and a given baseline radius of 0.5.

6. PERFORMANCE ON REAL-LIFE DATA

For verifying the performance of the $\mathcal{P}$-statistic, we take the **Breast Cancer Data-set** available publicly in the **UCI Machine Learning Repository** (Dua & Graff, 2017). We use the version embedded inside Python's **scikit-learn** package and load it using the native python command. For details on how to load and use the data-set, please check scikit-learn documentation (*Scikit-learn breast cancer data-set*, n.d.).

6.1. Data Description

This data-set consists of features computed from a set of fine needle aspirate (FNA) images of breast masses from a set of patients along with the ground-truth of the diagnosis i.e. malignant or benign. The characteristics of cell nuclei in the image are described in this data-set. There are a total of 30 features per image with more detailed descriptions available in (Dua & Graff, 2017). The feature matrix is shaped 357×30 and in the label vector, 0 represents malignant and 1 represents benign.

In order to visualize this high dimensional data, we use t-distributed Stochastic Neighbourhood (tSNE) (Maaten & Hinton, 2008) to calculate a 2D embedding of the data so that it can be plotted as a 2D scatter plot and examined manually for proximity between data-points and clusters. Before applying tSNE, each of the feature columns are normalized to extend from 0 to 1. Figure 11 represents the two clusters of data available in this data-set. The subsequent analysis has now been performed on this 2-D embedded version of this data-set so that results can be visually related to the cluster neighbourhood behaviour.

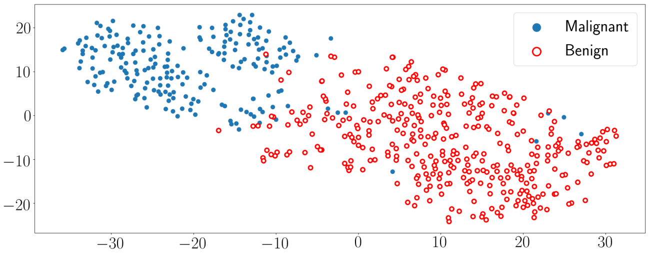

Figure 11. tSNE embedding of the malignant and benign clusters in the breast cancer data-set from UCI ML repository

6.2. Analysis with $\mathcal{P}$-statistic

We now consider 40 % of the benign cluster as the training baseline and the remaining points in the benign and malignant groups as test set. We now use the steps summarized in Algorithm 1 to estimate the p-values for all the data-points in both benign and malignant clusters in the test set with respect to the training baseline. Figures 12 and 13 show the $\mathcal{P}$-statistic and Mahalanobis Distance respectively, for all these points. From these, we see that the capability for discrimination between the baseline and anomalous data are similar for both the approaches.

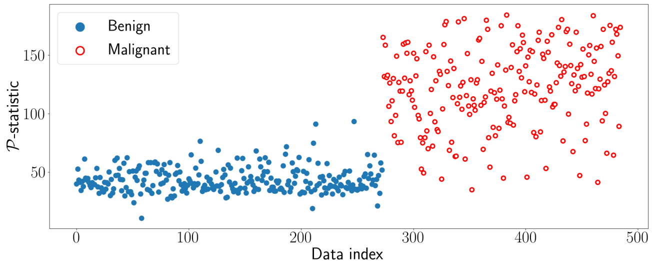

Figure 12. $\mathcal{P}$-statistic for benign and malignant data-points.

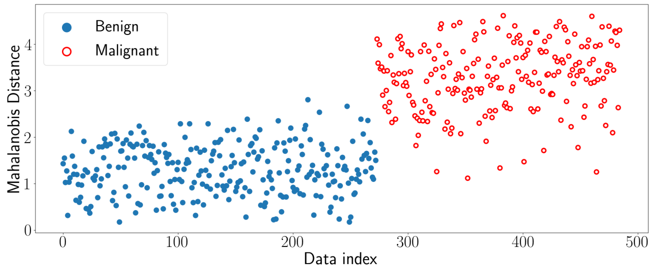

Figure 13. Mahalanobis Distance for benign and malignant data-points.

In order to compare the discrimination capability between these two metrics analytically, we plot the *true positive rate* vs. *false positive rate* with varying detection thresholds (*ROC curve*) in Figure 14. When calculating this ROC curve the *benign* and *malignant* labels are considered *negative* and *positive* respectively. We can see that the curves are nearly overlapping indicating that the $\mathcal{P}$-statistic is as effective as the Mahalanobis Distance in classification tasks as visible from this data.

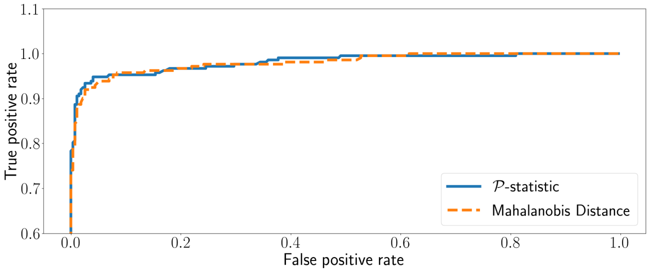

Figure 14. ROC curve for $\mathcal{P}$-statistic and Mahalanobis Distance

7. SOME OBSERVATIONS ON COMPUTATION TIME

7.1. Test Setup and Limitations in Field

Since one of the premises on which the choice of energy-based distance was made was that of computation simplicity, we analyse the computation requirement for *training* and *inference* separately in this section for both $\mathcal{P}$-statistic and MD. We notice from (24) and (54) that the bulk of the computation happens during training time when model parameters are being computed from the baseline cluster X_S. However, the inference phase has less computation to perform and is likely to have similar computation time for both $\mathcal{P}$-statistic and MD. We henceforth focus our attention on the training time.

Since many field deployed Data Acquisition (DAQ) systems like edge devices and protection relays are likely to be deployed without a live internet connection, having the training done in a cloud-based infrastructure and the trained model transferred to these devices may not be feasible. Hence, any algorithm that would reduce training computation requirement would be advantageous as these devices are usually challenged in terms of computation power.

It must be noted that the analysis in this section is conducted on a CPU-based windows machine without parallel processing. It has 16 GB RAM and an $i5$ processor from Intel. No GPU is available. This is a valid test scenario because field-deployed industrial DAQ systems are not likely to have GPUs making massive parallel processing difficult to implement.

7.2. Computation Time Analysis

A practical industrial system cannot wait till the entire baseline of a desired size is accumulated before producing results as it might take a long time for an industrial machine to capture baseline samples from all possible operating conditions. After each new data-point is received, the baseline cluster parameters are expected to be computed incrementally based on the calculations of the previous stage and comparison run on the new baseline thus obtained.

The block of computation that needs to happen during baseline computation phase is all that which can be done on the baseline X_S alone and without the test-point y being available.

From (24), the training computation for $\mathcal{P}$-statistic calculation is limited to calculating the numerator N (say) as

$$N = \sum_{j=2}^{n} \sum_{k<j} \frac{1}{||X_k - X_j||_2}. \tag{55}$$

For the jth update step, N_j may be written incrementally as

$$\begin{aligned} N_j &= N_{j-1} + \sum_{k<j} \frac{1}{||X_j - X_k||_2} \text{ with} \\ N_0 &= 0. \end{aligned} \tag{56}$$

For any such step, the computation time is written as $T_{\mathcal{P}}$.

Similarly, for computing the MD value in (54), the training computation block may be defined by the following components.

1. $\overline{\mathbf{X}}$: Centroid of baseline cluster.
2. $\mathbf{S^{-1}}$: Inverse of the covariance matrix for the input dataset.

From the above steps, only the centroid can be computed incrementally. Since it is not very straightforward to write the covariance matrix in an incremental fashion, we have to recompute the covariance matrix after each step. The inversion needs to be repeated any way. Let the computation time be T_{MD} for each of these update steps.

We consider a baseline distribution as one defined in Figure 6. Only now, the dimension d is considered more than 2 and a variable. We assume that the data-points arrive serially and compare computation times $T_{\mathcal{P}}$ and T_{MD} for a single update step with a new data-point given an already existing baseline size. Figure 15 shows how these two compare against each other for varying baseline sizes and values of the dimension d.

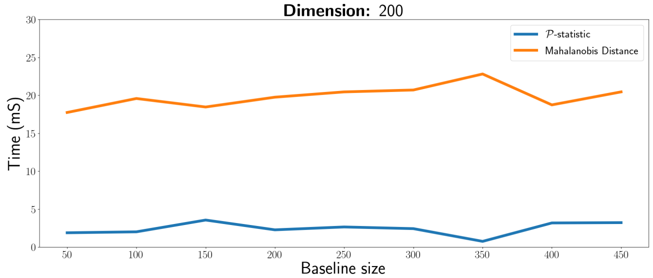

Figure 15. Time taken (ms) for a single incremental update step when using $\mathcal{P}$-statistic and MD, plotted for different already existing baseline sizes

For any baseline size along the x axis in Figure 15, the variability in CPU loading is accounted for by computing the times as an average of those calculated for 20 random choices of the baseline cluster from the data-set defined above. The CPU time has been calculated by using *time_ns()* function from the *time* package in Python before and after each code block during execution and taking the difference to be the computation time for that segment.

For a given existing baseline size, the ratio of the two timings is now plotted against the dimension in Figure 16.

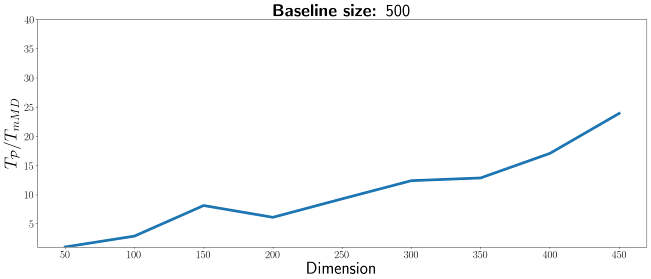

Figure 16. Comparison of $T_{\mathcal{P}}$ and T_{MD} for different data dimensions for a given baseline size.

From Figures 15 and 16, we have the following observations

1. Time taken for incremental baseline computation for MD is more than that of the $\mathcal{P}$-statistic after each update step.
2. MD calculation time scales up much more with data dimension than the $\mathcal{P}$-statistic computation. Thus, the advantage offered by $\mathcal{P}$-statistic is more pronounced for data of high dimensions.

This demonstrates the computation advantage which is claimed by energy-based distances over Mahalanobis Distance.

8. CONCLUSION

In this paper, we have reviewed the *energy distance*, a distance metric proposed in literature and shown to have simpler computation than most classical distance metric. We have shown that this metric possesses several shortcomings which prevent it from being applied in a practical anomaly detection application. We have proposed a modified metric called $\mathcal{P}$-statistic which works on distributions of Euclidean distances just like the *energy distance*. Using a synthetic data-set, we show that the above shortcomings are overcome by this modified metric. We have taken an example real-life data, the UCI breast-cancer data and shown that the $\mathcal{P}$-statistic and the Mahalanobis Distance have similar performance as a discrimination metric, as shown by an almost overlapping ROC curve. We have also demonstrated that computation times for Mahalanobis Distance calculation are higher than those for $\mathcal{P}$-statistic when computed in an incremental fashion with each new data arrival. Specially, we have seen that $\mathcal{P}$-statistic offers pronounced advantage with data having high dimension. Thus, we conclude that it is feasible to use this metric for anomaly detection without losing discrimination performance, at the same time utilizing the simpler computation that *energy distance* based methods offer.

REFERENCES

Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of multivariate analysis*, *88*(1), 190–206.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*(35), 99-109.

Dixon, A. (1901). On burmann's theorem. *Proceedings of the London Mathematical Society*, *1*(1), 151–153.

Dua, D., & Graff, C. (2017). *UCI machine learning repository.* Retrieved from `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\%28Diagnostic\%29`

Feuerverger, A. (1993). A consistent test for bivariate dependence. *International Statistical Review/Revue Internationale de Statistique*, 419–433.

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, *11*(4), e0152173.

Harman, R., & Lacko, V. (2010). On decompositional algorithms for uniform sampling from n-spheres and n-balls. *Journal of Multivariate Analysis*, *101*(10), 2297–2304.

Kim, A. Y., Marzban, C., Percival, D. B., & Stuetzle, W. (2009). Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Processing*, *89*(12), 2529–2536.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Mahalanobis, P. C. (1936). On the generalized distance in statistics..

Matteson, D. S., & James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, *109*(505), 334–345.

Rizzo, M. L. (2002a). *A new rotation invariant goodness-of-fit test* (Unpublished doctoral dissertation). Bowling Green University.

Rizzo, M. L. (2002b). A test of homogeneity for two multivariate populations. *Proceedings of the American Statistical Association, Physical and Engineering Sciences Section.*

Rizzo, M. L. (2009). New goodness-of-fit tests for pareto distributions. *ASTIN Bulletin: The Journal of the IAA*, *39*(2), 691–715.

Schöpf, H., & Supancic, P. (2014). On Bürmann‘s theorem and its application to problems of linear and nonlinear heat transfer and diffusion. *The Mathematica Journal*, *16*(11).

Scikit-learn breast cancer data-set. (n.d.). Retrieved from `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html`

Statistical distance. (n.d.). Retrieved from `https://en.wikipedia.org/wiki/Statistical\_distance`

Székely, G. J. (1989). Potential and kinetic energy in statistics. LBudapest Institue of Technology.

Székely, G. J. (2002). *E-statistics: energy of statistical samples* (Tech. Rep. No. 02-16). Bowling Green State University, Dep. Math. stat.

Székely, G. J., & Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, *93*(1), 58–80.

Székely, G. J., & Rizzo, M. L. (2009). Brownian distance covariance. *The annals of applied statistics*, 1236–1265.

Székely, G. J., & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, *143*(8), 1249–1272.

Szekely, G. J., & Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, *4*, 447–479.

Székely, G. J., Rizzo, M. L., et al. (2004). Testing for equal distributions in high dimension. *InterStat*, *5*(16.10), 1249–1272.

Szekely, G. J., Rizzo, M. L., et al. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, *22*(2), 151–184.

Yang, G. (2012). *The energy goodness-of-fit test for univariate stable distributions* (Unpublished doctoral dissertation). Bowling Green State University.

A. LIMITING CONDITIONS OF $\mathcal{E}$-STATISTIC

From (2) and (7),

$$\mathcal{E}_n(X'_S(\beta), y) = \frac{2}{n}\sum_{i=1}^{n} ||\beta(X_i - \overline{X_S}) + \overline{X_S} - y||_2 - \frac{\beta\gamma}{n^2} \text{ where} \tag{57}$$

$$\gamma = \sum_{i=1}^{n}\sum_{j=1}^{n} ||X_i - X_j||_2 < \infty. \tag{58}$$

$$\Rightarrow \lim_{\beta \to 0} \mathcal{E}_n(X'_S(\beta), y) = 2||\overline{X_S} - y||_2 < \infty. \tag{59}$$

Also, from (57),

$$\lim_{\beta\to\infty}\mathcal{E}_n = 2\left(\lim_{\beta\to\infty}\frac{\beta}{n}\right)\left(\sum_{i=1}^{n}||X_i-\overline{X_S}||_2-\frac{\gamma}{2n}\right) \tag{60}$$

Now,

$$\begin{aligned}
||X_i-\overline{X_S}||_2 &= \frac{1}{n}\left|\left|nX_i-\sum_{j=1}^{n}X_j\right|\right|_2\\
&= \frac{1}{n}\left|\left|\sum_{j=1}^{n}(X_i-X_j)\right|\right|_2\\
&> \frac{1}{n}\sum_{j=1}^{n}||X_i-X_j||_2\\
&> \sum_{j=1}^{n}\frac{1}{2n}||X_i-X_j||_2.\\
\Rightarrow\sum_{i=1}^{n}||X_i-\overline{X_S}||_2 &> \frac{\gamma}{2n}\text{ from (58).}\\
\Rightarrow)\left(\sum_{i=1}^{n}||X_i-\overline{X_S}||_2-\frac{\gamma}{2n}\right) &> 0 \qquad (61)
\end{aligned}$$

From (60) and (61),

$$\lim_{\beta\to\infty}\mathcal{E}_n(X_S^{'}(\beta),y)=+\infty. \tag{62}$$

Here, the reader should note that it was necessary to prove the relationship in (61) to establish whether the above-mentioned limit was $+\infty$ or $-\infty$. If the term on the *LHS* in (61) was negative, the limit in (62) would be $-\infty$ and not $+\infty$.

B. SOME GEOMETRICAL PROPERTIES OF $\mathcal{P}$-STATISTIC

Let $y_{(d)}$ be the value of the dth coordinate of the test-point y in the D-dimensional space. We assume that y is situated at a distance r from the origin. Since we are studying the impact of r alone on the values of $\mathcal{P}_n$ for any given radial direction, we may assume, without any loss of generality, that the coordinate system is oriented in such a manner that

$$y_{(d)}=\begin{cases} r, & \text{for } d=1,\\ 0, & \text{for } d=[2,3,...,D].\end{cases} \tag{63}$$

From (24), $\mathcal{P}_n(r)$ may be written as

$$\begin{aligned}
\mathcal{P}_n(r) &= \alpha\bigg/\sum_{k=1}^{n}\frac{1}{s_k}\text{ where} \qquad (64)\\
\alpha &= \sum_{j=2}^{n}\sum_{k<j}\frac{1}{||X_k-X_j||_2}\text{ and}\\
s_k &= ||X_k-y+\tau||_2\\
&= \sqrt{(X_{k(1)}-r+\epsilon)^2+\sum_{d=2}^{D}X_{k(d)}^2}\\
&= \sqrt{||X_k||_2^2+(r-\epsilon)^2-2(r-\epsilon)X_{k(1)}}. \qquad (65)
\end{aligned}$$

Here, $X_{k(d)}$ is the value of the dth coordinate of the point X_k which is D-dimensional in nature.

From (64) and (65),

$$\begin{aligned}
\frac{d\mathcal{P}_n}{dr} &= -\alpha\left(\sum_k\frac{1}{s_k}\right)^{-2}\sum_k\frac{d}{dr}\left(\frac{1}{s_k}\right)\\
&= \alpha\left(\sum_k\frac{1}{s_k}\right)^{-2}\sum_k\frac{1}{s_k^2}\frac{ds_k}{dr}\\
&= \alpha\left(\sum_k\frac{1}{s_k}\right)^{-2}\sum_k\frac{r-X_{k(1)}-\epsilon}{s_k^3}. \qquad (66)
\end{aligned}$$

From (66), there exists $r^*>X_{k(1)}+\epsilon\;\forall\;k$ such that

$$\frac{d\mathcal{P}_n}{dr}>0\;\forall\;r>r^*\text{ since }\alpha>0\text{ and }s_k>0\;\forall\;k. \tag{67}$$

Also, from (66),

$$\begin{aligned}
\lim_{r\to\infty}\frac{d\mathcal{P}_n}{dr} &= \lim_{r\to\infty}\alpha\sum_k\frac{\dfrac{r-X_{k(1)}-\epsilon}{s_k^3}}{\left(\dfrac{1}{s_k}+\displaystyle\sum_{j\neq k}\frac{1}{s_j}\right)^2}\\
&= \alpha\sum_k\frac{\left(1-\displaystyle\lim_{r\to\infty}\frac{X_{k(1)}+\epsilon}{r}\right)\bigg/\left(\displaystyle\lim_{r\to\infty}\frac{s_k}{r}\right)}{\left(1+\displaystyle\sum_{j\neq k}\lim_{r\to\infty}\frac{s_k}{s_j}\right)^2} \qquad (68)
\end{aligned}$$

Now, from (65),

$$\begin{aligned}
\lim_{r\to\infty}\frac{s_k}{r} &= \lim_{r\to\infty}\sqrt{\delta^2-\frac{2\delta X_{k(1)}}{r}+\left(\frac{||X_k||_2}{r}\right)^2}\\
&= \lim_{r\to\infty}\delta,\text{ where }\delta=1-\epsilon/r.\\
&= 1. \qquad (69)
\end{aligned}$$

Also, from (65),

$$\begin{aligned}\lim_{r\to\infty}\frac{s_k}{s_j} &= \lim_{r\to\infty}\sqrt{\frac{||X_k||_2^2+(r-\epsilon)^2-2(r-\epsilon)X_{k(1)}}{||X_j||_2^2+(r-\epsilon)^2-2(r-\epsilon)X_{j(1)}}}\\ &= 1. \end{aligned} \tag{70}$$

From (68), (69) and (70),

$$\begin{aligned}\lim_{r\to\infty}\frac{d\mathcal{P}_n}{dr} &= \alpha\sum_k 1\Bigg/\left(1+\sum_{j\neq k}1\right)^2\\ &= \alpha/N,\ \text{a constant,}\end{aligned} \tag{71}$$

since α is a constant. Here N is assumed to be the number of elements in the cluster X_S.

A few other points about the nature of $\mathcal{P}_n(r)$ that need to be mentioned here are

$$\forall\, r<\infty,\ \mathcal{P}_n(r)<\infty \text{ since } \alpha<\infty \text{ and } s_k<\infty. \tag{72}$$

and

$$\mathcal{P}_n \geq 0\,\forall\, r \text{ since } \alpha>0 \text{ and } s_k>0\,\forall\, k. \tag{73}$$