Multi-criteria Anomaly Detection using Pareto Depth Analysis: Supplementary Material

Ko-Jen Hsiao, Kevin S. Xu, Jeff Calder, and Alfred O. Hero III
University of Michigan, Ann Arbor, MI, USA 48109
{coolmark,xukevin,jcalder,hero}@umich.edu

1 Proofs of Theorems 1 and 2

Before presenting the proofs of Theorems 1 and 2 we need a preliminary result.

Lemma 1. *For any $n \geq 1$ and $A \subset \mathbb{R}^d$ measurable, we have*

$$E|\mathcal{F}_A| = n \int_A f(x) \left(1 - \int_{y \preceq x} f(y)dy\right)^{n-1} dx. \tag{1}$$

Proof. Since $Y_1, \ldots, Y_n$ are i.i.d, we have $E|\mathcal{F}_A| = nP(Y_1 \in \mathcal{F})$. Conditioning on Y_1 we obtain $E|\mathcal{F}_A| = n \int_{\mathbb{R}^d} f(x)P(Y_1 \in \mathcal{F} \,|\, Y_1 = x)dx$. The proof is completed by noting that

$$P(Y_1 \in \mathcal{F} \,|\, Y_1 = x) = \begin{cases} \left(1 - \int_{y \preceq x} f(y)dy\right)^{n-1}, & x \in A, \\ 0, & x \notin A. \end{cases}$$ □

Proof of Theorem 1. By selecting $h > 0$ smaller, if necessary, we can write (1) as

$$E|\mathcal{F}_{T_h}| = \int_T \int_0^h nf(x) \left(1 - \int_{y \preceq x} f dy\right)^{n-1} (1 + O(t))dtdz, \tag{2}$$

where $x = z + t\nu(z)$ for $z \in T$. Since $\partial\Omega$ is smooth, we can approximate T near z by a hyperplane with normal $\nu(z)$. By the assumption that $\{y \in \overline{\Omega} : y \preceq x\} = \{x\}$ we can make $h > 0$ smaller, if neceessary, so that $\{y \in \Omega : y \preceq x\}$ is approximately a simplex with side lengths $t/\nu_i(z)$. Hence

$$\begin{aligned} \int_{y \preceq x} f(y)dy &= (f(z) + O(t/\delta)) \int_{y \preceq x} dy \\ &= \frac{f(z)t^d}{d!\nu_1(z) \cdots \nu_d(z)} + O\left(\frac{t^{d+1}}{\delta^{d+1}}\right). \end{aligned}$$

Substituting this into (2), we have

$$E|\mathcal{F}_{T_h}| = \int_T \int_0^h n(f(z) + O(t)) \left(1 - \frac{f(z)t^d}{d!\nu_1(z) \cdots \nu_d(z)} + O(t^{d+1}/\delta^{d+1})\right)^{n-1} dtdz. \tag{3}$$

We can now do an asymptotic analysis of the inner integral which is a special case of the general equation

$$A_n := \int_0^h t^\lambda (1 - at^d + O(bt^{d+1}))^{n-1} dt, \;\; \lambda \in [0, 1], a, b > 0.$$

Making the change of variables $-s = (n-1)\ln(1 - at^d + O(bt^{d+1}))$ and simplifying, we obtain

$$A_n = \frac{1}{(a(n-1))^{\frac{1+\lambda}{d}}} \int_0^{P(n-1)} \left(\frac{1}{d} s^{\frac{1+\lambda}{d}-1} + \frac{b}{(n-1)^{\frac{1}{d}}} O(s^{\frac{2+\lambda}{d}-1})\right) e^{-s} ds,$$

where

$$P = -\ln(1 - ah^d + bO(h^{d+1})).$$

We can, of course, choose h small enough so that P is finite and positive. Recalling the definition of the Gamma function, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$, we see that

$$A_n = \frac{\Gamma\left(\frac{1+\lambda}{d}\right)}{d(an)^{\frac{1+\lambda}{d}}} + O\left(\frac{b}{n^{\frac{2+\lambda}{d}}}\right).$$

Note that we are keeping track of $O(b)$ terms because $b = O(1/\delta^{d+1})$ may become large at different points of T, whereas $O(1/a)$ is uniformly bounded independent of δ along T. Applying this to (3) with

$$a = \frac{f(z)}{d!\nu_1(z)\cdots\nu_d(z)}, \text{ and } b = \delta^{-(d+1)},$$

completes the proof. □

Proof of Theorem 2. Since $Y_1, \ldots, Y_n$ are i.i.d., we have $E|\mathcal{L}| = nP(Y_1 \in \mathcal{L})$. For $(x, y) \in [0, 1]^2$ let $D_{x,y}$ be the event that $Y_1 = (x, y)$ and $(x, y) \in \mathcal{F}$. Conditioning on $D_{x,y}$ we have

$$\begin{aligned} E|\mathcal{L}| &= n\int_0^1\int_0^1 (1-xy)^{n-1}P((x,y) \in \mathcal{L} \mid D_{x,y})\,dxdy \\ &= n\int_0^{\frac{1}{2}}\int_0^{\frac{1}{2}} (1-xy)^{n-1}P((x,y) \in \mathcal{L} \mid D_{x,y})\,dxdy + O(1). \end{aligned} \tag{4}$$

Define

$$A = \left\{(u, v) \in [0, 1]^2 \mid 0 < u < x,\ y < v < 2y - \frac{uy}{x}\right\},$$

and

$$B = \left\{(u, v) \in [0, 1]^2 \mid x < u < 1,\ 0 < v < 2y - \frac{uy}{x}\right\}.$$

Let E be the event that A and B each contain at least one sample from $Y_2, \ldots, Y_n$. If E occurs, then (x, y) is in the interior of the convex hull of $\mathcal{F}$ and hence $(x, y) \notin \mathcal{L}$. Let F denote the event that none of the samples from $Y_2, \ldots, Y_n$ fall in $A \cup B$. If F occurs, then we clearly have $(x, y) \in \mathcal{L}$. It follows that

$$P(F \mid D_{x,y}) \le P((x, y) \in \mathcal{L} \mid D_{x,y}) \le P(E^c \mid D_{x,y}).$$

Conditioned on $D_{x,y}$, the samples $Y_2, \ldots, Y_n$ remain independent. The conditional density function of each remaining sample is $f_{Y_i \mid D_{x,y}}(u, v) = \frac{1}{1-xy}$. Let E_A (resp. E_B) denote the event that no samples from $Y_2, \ldots, Y_n$ are drawn from A (resp. B). Then $E^c = E_A \cup E_B$ and $F = E_A \cap E_B$. Noting that $|A| = |B| = \frac{1}{2}xy$, we see that

$$\begin{aligned} P(E^c \mid D_{x,y}) &= P(E_A \mid D_{x,y}) + P(E_B \mid D_{x,y}) - P(E_A \cap E_B \mid D_{x,y}) \\ &= 2\left(1 - \frac{xy}{2(1-xy)}\right)^{n-1} - \left(1 - \frac{xy}{1-xy}\right)^{n-1}, \end{aligned}$$

and

$$P(F \mid D_{x,y}) = P(E_A \cap E_B \mid D_{x,y}) = \left(1 - \frac{xy}{1-xy}\right)^{n-1}.$$

Substituting this into (4), we obtain

$$E|\mathcal{L}| \le n\int_0^{\frac{1}{2}}\int_0^{\frac{1}{2}} 2\left(1 - \frac{3}{2}xy\right)^{n-1} - (1-2xy)^{n-1}\,dxdy,$$

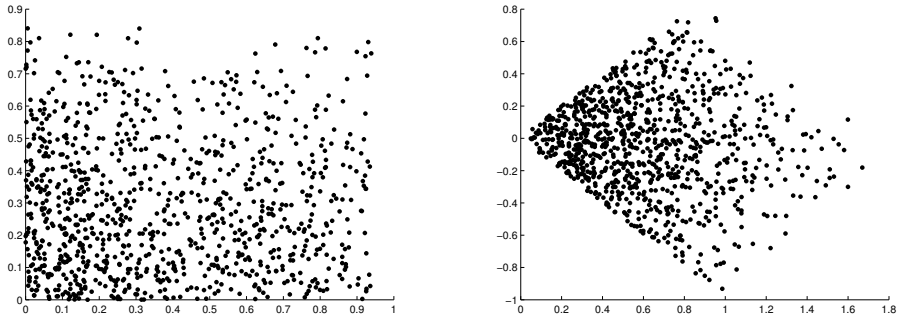
and

$$E|\mathcal{L}| \ge n\int_0^{\frac{1}{2}}\int_0^{\frac{1}{2}} (1-2xy)^{n-1}\,dxdy.$$

A short calculation (change variables to $u = anxy$ and $v = x$) shows that

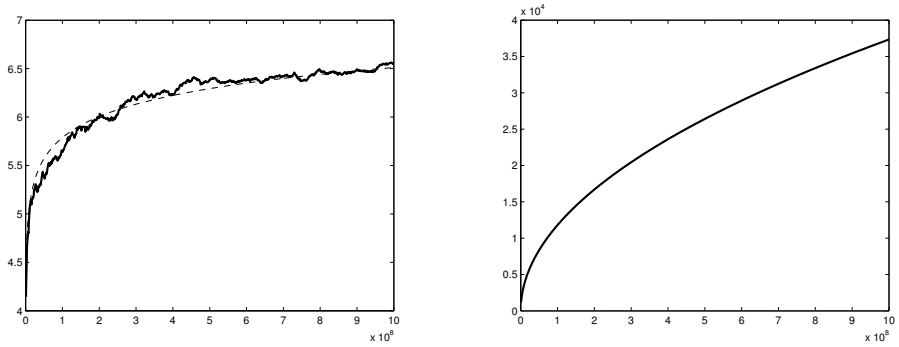
$$\int_0^{\frac{1}{2}}\int_0^{\frac{1}{2}} n(1-axy)^{n-1}dxdy = \frac{1}{a}\ln n + O(1).$$

Applying this result to the bounds above completes the proof. □



(a) Criteria $|\Delta x|$,$|\Delta y|$ (b) Criteria $|\Delta x| + |\Delta y|$,$|\Delta x| - |\Delta y|$

Figure 1: 990 dyads constructed with two different sets of criteria from 45 samples uniformly distributed in $[0, 1]^2$.



(a) Criteria $|\Delta x|$,$|\Delta y|$ (b) Criteria $|\Delta x| + |\Delta y|$,$|\Delta x| - |\Delta y|$

Figure 2: Sample means for $E|\mathcal{F} \setminus \mathcal{L}|$ versus n. We can see the expected logarithmic and half-power growth in (a) and (b) respectively. The dotted lines indicate the best fit curves described in this section. In (b), the best fit curve is too closely aligned with the experimental data to be visible.

2 Experimental support for Theorems 1 and 2

Independence of $Y_1, \ldots, Y_n$ is built into the assumptions of Theorems 1 and 2, but it is clear that dyads (as constructed in Section 4 of the main paper) are not independent. Each dyad $D_{i,j}$ represents a connection between two independent samples X_i and X_j. For a given dyad $D_{i,j}$, there are $2(N-2)$ corresponding dyads involving X_i or X_j and these are clearly not independent from $D_{i,j}$. However, all other dyads are independent from $D_{i,j}$. So while there are $O(N^2)$ dyads, each dyad is independent from all other dyads except for a set of size $O(N)$. Since Theorems 1 and 2 deal with asymptotic results, this suggests they should hold for the dyads even though they are not i.i.d. In this section we present some experimental results that support this non-rigorous statement.

We first drew samples uniformly in $[0, 1]^2$ and computed the dyads corresponding to the two criteria $|\Delta x|$ and $|\Delta y|$, which denote the absolute differences between the x and y coordinates, respectively. The domain of the resulting dyads is again the box $[0, 1]^2$, as shown in Figure 1(a), so this experiment tests Theorem 2. In this case, Theorem 2 suggests that $\mathcal{F} \setminus \mathcal{L}$ should grow logarithmically. Figure 2(a) shows the sample means versus number of dyads and a best fit logarithmic curve of the form $y = \alpha \ln n$, where $n = \binom{N}{2}$ denotes the number of dyads. A linear regression on $y/\ln n$ versus $\ln n$ gave $\alpha = 0.3142$ which falls in the range specified by Theorem 2.

We next looked to find criteria that induce domains other than boxes in order to test Theorem 1. A somewhat contrived example involves the criteria $|\Delta x| + |\Delta y|$ and $|\Delta x| - |\Delta y|$, which, when

applied to uniformly sampled data on $[0,1]^2$, yields dyads sampled on a diamond domain, as shown in Figure 1(b). In this case, Theorem 1 suggests that $\mathcal{F} \setminus \mathcal{L}$ should grow as $\sqrt{n}$. Figure 2(b) shows the sample means versus number of dyads and a best fit curve of the form $y = \alpha n^{\beta}$. A linear regression on $\ln y$ versus $\ln n$ gave $\alpha = 1.1642$ and $\beta = 0.5007$. Although this example may not be practical, it is simply meant to illustrate the applicability of Theorem 1 for non-independent samples. In each experiment, we varied the number of dyads between 10^6 to 10^9 in increments of 10^6 and computed the size of $\mathcal{F} \setminus \mathcal{L}$ after each increment. We ran each experiment $1,000$ times to compute the sample means shown in Figure 2.

3 Implementation of PDA anomaly detector

Pseudocode for the PDA anomaly detector was presented as Algorithm 1 in Section 4.2 of the main paper. The training phase involves creating $\binom{N}{2}$ dyads corresponding to all pairs of training samples. Computing all pairwise dissimilarities in each criterion requires $O(mKN^2)$ floating-point operations (flops), where m denotes the number of dimensions involved in computing a dissimilarity. The Pareto fronts are constructed by non-dominated sorting. In Section 3.1 we present a fast algorithm for non-dominated sorting in two criteria; for more than two criteria, we use the non-dominated sort of Deb et al. [1] that constructs all of the Pareto fronts using $O(KN^4)$ comparisons in the worst case.

The testing phase involves creating dyads between the test sample and the k_l nearest training samples in criterion l, which requires $O(mKN)$ flops. For each dyad D_i^{new}, we need to calculate the depth e_i. This involves comparing the test dyad with training dyads on multiple fronts until we find a training dyad that is dominated by the test dyad. e_i is the front that this training dyad is a part of. Using a binary search to select the front and another binary search to select the training dyads within the front to compare to, we need to make $O(K \log^2 N)$ comparisons (in the worst case) to compute e_i. The anomaly score is computed by taking the mean of the s e_i's corresponding to the test sample; the score is then compared against a threshold σ to determine whether the sample is anomalous. As mentioned in the main paper, both the training and testing phases scale linearly with the number of criteria K.

3.1 Fast non-dominated sorting for two criteria

We present here a fast algorithm for non-dominated sorting in two criteria. The standard algorithm of Deb et al. [1] takes $O(n^2)$ time and requires $O(n^2)$ memory, where $n = \binom{N}{2}$ is the number of dyads. In our experience, the memory requirement is the largest obstacle to applying Pareto methods to large data sets. Our algorithm runs in $O(n^{3/2})$ time on average and requires $O(n)$ memory. It is based on the following observation: if the data set is sorted in ascending order in the first criterion, then the first point is Pareto-optimal, and each subsequent Pareto-optimal point can be found by searching for the next point in the sorted list that is not dominated by the most recent addition to the Pareto front. For two criteria, there are on average $O(\sqrt{n})$ Pareto fronts, and finding each front with this algorithm requires visiting at most n points, hence the $O(n^{3/2})$ average complexity. The worst case complexity is $O(n^2)$ occurring when each Pareto front consists of a single point. Pseudocode for the algorithm is shown in Algorithm 1. It has recently come to our attention that an $O(n \ln n)$ algorithm exists for the canonical anti-chain partition problem [3], which is equivalent to non-dominated sorting in two criteria, and can also be used to quickly construct the Pareto fronts.

3.2 Selection of parameters

The parameters to be selected in PDA are $k_1, \ldots, k_K$, which denote the number of nearest neighbors in each criterion. We connect each test sample X to a training sample X_j if X_j is one of the k_i nearest neighbors of X in terms of the dissimilarity measure defined by criterion i. We now discuss how these parameters $k_1, \ldots, k_K$ can be selected. For simplicity, first assume that there is only one criterion, so that a single parameter k is to be selected. PDA is able to detect an anomaly if the distribution of its dyads with respect to the Pareto fronts differs from that of a nominal sample. Specifically the mean of the depths of the dyads (the e_i's) corresponding to an anomalous sample must be higher than that of a nominal sample. If k is chosen too small, this may not be the case, especially if there are training samples present near an anomalous sample, in which case, the dyads

Algorithm 1 Fast non-dominated sorting.

Require: Arrays X and Y of length n (the values of the two criteria)
Sort X and Y according to X in ascending order
while X and Y are nonempty **do**
  Add $(X(1), Y(1))$ to current Pareto front
  $y \leftarrow Y(1)$
  for $i = 2 \rightarrow \text{length}(X)$ **do**
    if $Y(i) \leq y$ **then**
      Add $(X(i), Y(i))$ to current Pareto front
      $y \leftarrow Y(i)$
  Remove current Pareto front from X,Y

corresponding to the anomalous sample may reside near shallow fronts much like a nominal sample. On the other hand, if k is chosen too large, many dyads may correspond to connections to training samples that are far away, even if the test sample is nominal, which also makes the mean depths of nominal and anomalous samples more similar.

We propose to use the properties of k-nearest neighbor graphs (k-NNGs) constructed on the training samples to select the number of training samples to connect to each test sample. We construct symmetric k-NNGs, i.e. we connect samples i and j if i is one of the k nearest neighbors of j or j is one of the k nearest neighbors of i. We begin with $k = 1$ and increase k until the k-NNG of the training samples is connected, i.e. there is only a single connected component. By forcing the k-NNG to be connected, we ensure that there are no isolated regions of training samples. Such isolated regions could possibly lead to dyads corresponding to anomalous samples residing near shallow fronts like nominal samples, which is undesirable. By keeping k small while retaining a connected k-NNG, we are trying to avoid the problem of having too many dyads so that even a nominal sample may have many dyads located near deep fronts. This method of choosing k to retain connectivity has been used as a heuristic in other unsupervised learning problems, such as spectral clustering [2]. Note that by requiring the k-NNG to be connected, we are implicitly assuming that the training samples consist of a single class or multiple classes that are in close proximity. If the training samples contain multiple well-separated classes, such an approach may not work well.

Now let's return to the situation PDA was designed for, with K different criteria. For each criterion i, we construct a k_i-NNG using the corresponding dissimilarity measure and increase k_i until the k_i-NNG is connected. We then connect each test sample to $s = \sum_{i=1}^{K} k_i$ training samples. Note that we are choosing each k_i independent of the other criteria, which is probably not an optimal approach. In principle, an approach that chooses the k_i's jointly could perform better; however, such an approach would add to the complexity. We choose *separate* k_i's for each criterion, which we find is necessary to obtain good performance when different dissimilarities have varying scales and properties. There are, however, pathological examples where the independent approach could choose k_i's poorly, such as the well-known example of two moons. These examples typically involve multiple well-separated classes, which may be problematic as previously mentioned. How to choose the k_i's when the training samples contain multiple well-separated classes is beyond the scope of this paper and is an area for future work. We find the proposed heuristic to work well in practice, including for both examples presented in the main paper.

4 Additional discussion on pedestrian trajectories experiment

Figure 3 shows some abnormal trajectories and nominal trajectories detected using PDA. Recall that the two criteria used are walking speed and trajectory shape. Anomalous trajectories could have anomalous speeds or shapes (or both), so some anomalous trajectories in Figure 3 may not look anomalous by shape alone. We find that the heuristic proposed in Section 3.2 for choosing the k_i's performs quite well in this experiment, as shown in Figure 4. Specifically, the AUC obtained when using the parameters chosen by the proposed heuristic is very close to the AUC obtained when using the optimal parameters, which are not known in advance. As discussed in Section 5.2 of the main paper, it is also higher than the AUCs of all of the single-criterion anomaly detection methods, even under the best choice of weights.

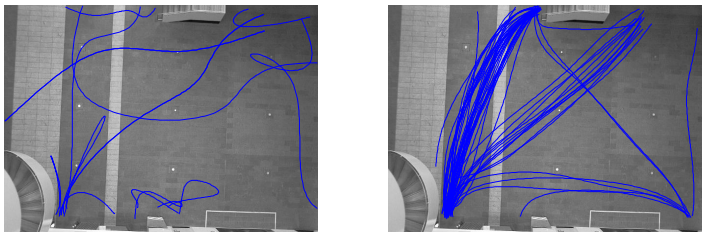


Figure 3: *Left:* Some abnormal trajectories detected by PDA method. *Right:* Trajectories with relatively low anomaly scores.

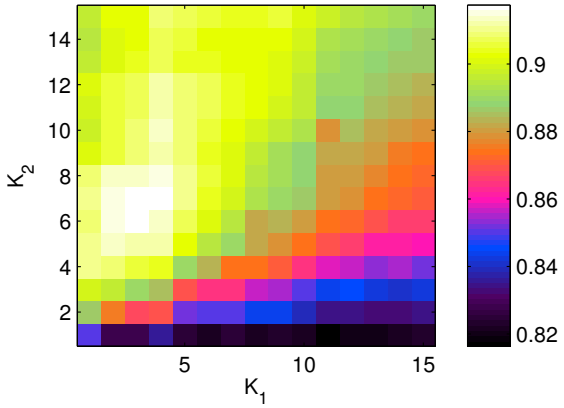


Figure 4: AUCs for different choices of $[k_1, k_2]$. The automatically selected parameters $[k_1 = 3, k_2 = 6]$ are very close to the optimal parameters $[k_1 = 4, k_2 = 7]$.

References

[1] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*.

[2] U. von Luxburg (2007). A tutorial on spectral clustering. *Statistics and Computing* **17**(4):395–416.

[3] S. Felsner and L. Wernisch (1999). Maximum k-chains in planar point sets: Combinatorial structure and algorithms. *SIAM Journal on Computing*, **28**(1):192–209.