Abstract

This work describes a general method for the detection of different types of structures in a set of elements that belong to a certain ambient space. The general idea is that such structures express themselves as unusual concentrations of subsets of such elements around these structures, as opposed to being randomly scattered throughout the space.

1 Preliminaries

Notation

- Dimensions are lower capital letters such as m, n, p. Random variables are capital letters such as X, Y, Z.
- Vectors are lower case bold, e.g., $\mathbf{x}$.
- Matrices are upper-case bold as in $\mathbf{A}$.
- Spaces are in double bold, such as $\mathbb{R}, \mathbb{C}, \mathbb{X}$.
- Sets are in calligraphic, such as $\mathcal{O}$; their cardinality is written as $|\mathcal{O}|$
- The probability of a given event ω is $\Pr\{\omega\}$.
- $\mathbb{E}_X[f(X)]$ is the expectation of the function $f(X)$ w.r.t. the distribution of X.
- The indicator function is denoted as $\mathbf{1}(A = 1)$
- For a vector $\mathbf{x}$, the i-th element is denoted as $x[i]$
- For elements in a set, the j-th element of a set is denoted as x_j
- Combining both definitions above, the i-th element of the j-th vector in a set of vectors is denoted by $\mathbf{x}_j[i]$.

Conventions Letters such as i, j, k, l, r are reserved for indexes. The letter t indicates iteration number. Letters such as m, n, p are for dimensions. a, b, c are usually constants. d is used for distances, e for errors, f, g, h for functions. u, v, w, x, y, z usually denote realizations of correponding random variables U, V, W, X, Y, Z. The letters x, y, z can also denote 2D/3D spatial coordinates. Greek letters $\alpha, \beta, \gamma, \tau$ are usually system hyperparameters (such as a threshold τ), whereas μ, θ, σ are commonly used for PMF/PDF parameters.

2 Problem Setting

The task is to detect meaningful structures hidden among the set of observed elements. We begin by restricting such structures to a parametric family Q. We begin by considering Q to be the set of affine subspaces of a given dimension m on an ambient space of dimension n.

We represent affine subspaces as a triplet $(\mathbf{u}, \mathbf{V}, \mathbf{W})$ where $\mathbf{u}$ is the distance of the set to the origin, $\mathbf{V}$ is a matrix whose columns generate the affine space, and $\mathbf{U}$ is the orthogonal complement of $\mathbf{V}$.

A generic point in the ambient space is represented by $\mathbf{x} \in \mathbb{R}^n$. The set of observed data points is

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}.$$

. Alternatively, we might represent this dataset as a matrix $\mathbf{X} \in \mathbb{R}^{n \times N}$ whose columns are the data points.

The notion of *concentration* arond a model q implies a metric and a distance. We denote the distance from a point $\mathbf{x}$ to any model q is $d(o, m) \in \mathbb{R}^+$ as $d(\mathbf{x}, q)$. For affine sets, the distance is defined as the norm of the distance between $\mathbf{x}$ and its projection $\Pi_q(\mathbf{x})$ onto q.

Any given point is considered to belong to some model q or either the *background*. Thus, we need to provide a model for the background as well (this would be the *null hypothesis* in classical statistics). The main contribution of this work is to consider *local* background models that are defined relative to a region around a given candidate model q. The reason for this will become clear later.

3 General detection framework

The general detection pipeline involves the following steps:

- Define the model family and the background model
- Find a set of candidate models $\mathcal{Q}$
- For each model $q \in \mathcal{Q}$, define *local background model* $\hat{q}$ representing the abscence of evidence for that model; this is in contrast to other works, which consider a global background model, such as a uniform distribution over a compact subset of the ambient space.
- keep those models which pass the *significance test* in a set $\mathcal{S}^+$

- Generate a set of *refined models*, $\mathcal{S}^*$, by removing those models in $\mathcal{S}$ that are *redundant* w.r.t. some criterion.

The above pipeline is quite standard in many detection frameworks. The novel aspect of this work lies in performing a *significance test* w.r.t. a *local* background model. This in turn calls for novel ways of defining such tests. This is where most of our work will be devoted to.

On the other hand, finding a proper set of candidate models is a very difficult problem. Clearly, the family $\mathbb{Q}$ is infinite and generally uncountable. Even if we restrict ourselves only to those models that can be determined by subsets of objects (e.g., all line segments defined by all pairs of points in $\mathcal{X}$), the number of candidates can be huge. Thus, for now and most of this work, we will assume $\mathcal{Q}$ to be given to us.

3.1 Background model

For an affine model of dimension m on ambient space n, we consider points whose distance to q is no larger than a given scatter parameter s. Formally speaking, such points belong to the direct product of a sphere of dimension $n-m$ and the affine subsspace q; this is the local background model $\hat{q} = \{a+b : a \in q, b \in \mathbb{R}^{n-m}, \|b\|_2 \leq s\}$.

We further assume that background points are distributed uniformly in $\hat{q}$. Although this does not define a proper distribution, the distribution of the *distance* to such points to q is well defined. Indeed, any point $\mathbf{x} \in \hat{q}$ can be written as $\mathbf{x} = \mathbf{u} + \mathbf{W}\mathbf{a} + \mathbf{b}\mathbf{V}$ with $\|\mathbf{a}\| \leq s$. Now, the distance of $\mathbf{x}$ to q is given by

$$\left\|\mathbf{W}^T(\mathbf{x}-\mathbf{u})\right\|_2 = \left\|\mathbf{W}^T(\mathbf{W}\mathbf{a}+\mathbf{V}\mathbf{b})\right\| = \|\mathbf{a}\|_2 ,$$

and $\mathbf{a}$ is supported on the closed ball in $\mathbb{R}^{n-m}$, $\mathcal{B}_s^{n-m} = \{\mathbf{a} : \|\mathbf{a}\|_2 \leq s\}$. Now, in turn, this uniform distribution on the sphere defines a distribution on the *distance* δ of a point $\mathbf{x}$ in $\hat{q}$ to q. If we denote by S_δ^{n-m} the volume of the ball $\mathcal{B}_\delta^{n-m}$, then the induced density is

$$f(\delta \leq z) = \frac{\mathcal{B}_z^{n-m}}{\mathcal{B}_s^{n-m}},$$

which, up to a constant is simply a truncated power law:

$$\Pr\{\delta \leq z\} = \begin{cases} \kappa\delta^{n-m} & \text{if } \delta \leq s \\ 1 & \text{otherwise} \end{cases}$$

Figure 2 on this.

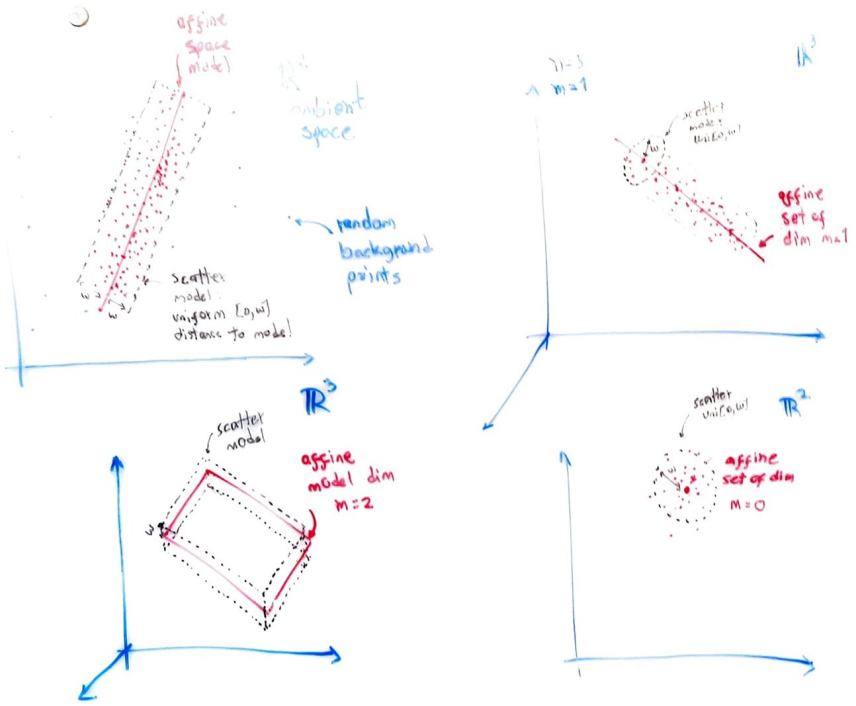

Figure 1: Examples of some affine sets (in red), ambient spaces, and local background (scatter) models (in black). Notice that single points are considered zero-dimensional affine sets.

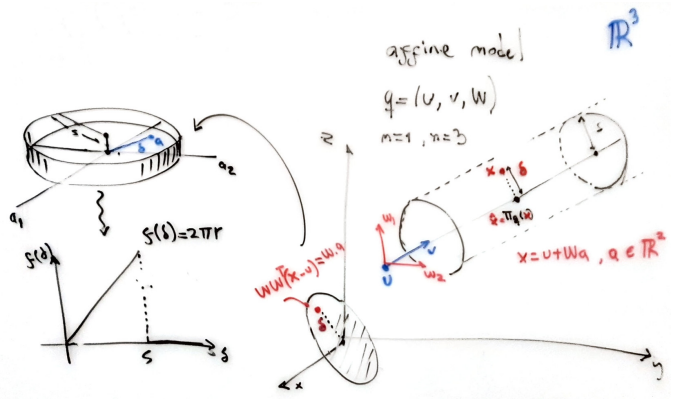

Figure 2: Some intuition behind the background model. The example q here is a one-dimensional affine set (a line) on a 3D Euclidean space. In this setting we have $\mathbf{u} \in \mathbb{R}^3$, $\mathbf{V} \in \mathbb{R}^{3\times 1}$ and $\mathbf{W} \in \mathbb{R}^{2\times 3}$. Given points of the form $\mathbf{x} = \mathbf{u} + \mathbf{W}\mathbf{a} + \mathbf{V}\mathbf{b}$, the density of $\mathbf{a} \in \mathbb{R}^2$ is uniform on the 2D disk or radius s, which in turn generates a uniform distribution on the distance of points in the disk to the origin whose density is linear function truncated at s, in this case it is simply $f(\delta) = 2\delta/s^2$; in general, $f(\delta) \propto \delta^{n-m-1}$.

3.2 Significance test

A given model q will be considered significant if the points in $\hat{q}$ are unusually close to q to be uniformly distributed. We can analyze this in terms of the empirical vs. expected distributions of the distances δ. The expected distribution is the one given by the background model, that is, $F(\delta) = \kappa\delta^{n-m}$. The empirical distribution is given by:

$$\hat{F}(\delta) = \frac{\sum_{\mathbf{x}\in\hat{q}} \mathbf{1}\left(d(\mathbf{x}, q) \leq \delta\right)}{\sum_{\mathbf{x}\in\hat{q}} 1}$$

If the points are *closer* to what one would expect, then the empirical CDF $\hat{F}(\delta)$ should be larger than the background CDF $F(\delta)$ for small δ, and smaller for larger δ. Figure 3 depicts this scenario.

Ideally, we would like to condense the above criterion in terms of a p-value, that is, of the probability that the discrepancy between the observed and background distributions differ in some way, and determine a significance threshold, e.g., to declare a detection if $p < 0.05$.

A standard method for such task is the Kolmogorov-Smirnov test. This test is defined in terms of the maximum absolute pointwise difference be-

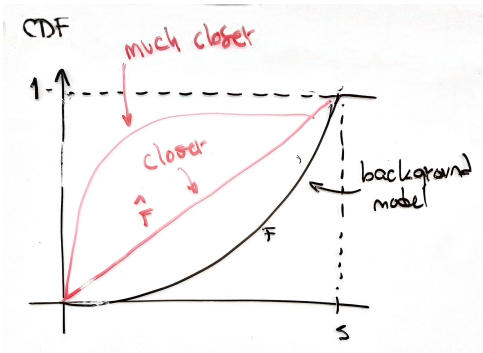

Figure 3: In black we see the background CDF for the distance δ for the case where $n - m = 2$ (a 2D disk) and a couple of imagined empirical CDFs which would indicate that the concentration of points near the model q is unusually high.

tween F and $\hat{F}$. Although this does not distinguish between negative or positive differences, and closer or farther values, it still serves to some extent and is a well established procedure, and thus constitutes our first approach to the problem.

3.3 Analysis scale

A critical parameter of the above procedure is the assumed maximum distance s of points to a model q; we also call this parameter the *scatter scale* or *analysis scale*, as it is clearly related to the scale of the objects we want to recognize. This is a very sensitive parameter: different scales will often yield different results. Moreover, for sufficiently large s, given that data points are bounded, a detection will always be produced for any candidate object q! Thus, conjuring a method for setting this scale parameter automatically is an important aspect of our work. This is still work in progress. Meanwhile, we will continue with s assumed to be known.

3.4 Multiple tests

In general, the set of candidate models $\mathcal{Q}$ will contain several elements. Therefore, in order to reduce the number of false detections (NFA for abbreviation) we need take this multiplicity into account. The standard approach is to use a Bonferroni correction, that is, to divide the detection threshold by the number of elements in $\mathcal{Q}$.

4 Technical details

Here we describe two things: first, the detection algorithm; second, the simmulations.

4.1 Detection algorithm

Inputs:

1. Dataset $\mathcal{X}$ consisting of N samples of dimension n
2. Set of candidate models $\mathcal{Q}$ given as triplets $(\mathbf{u}, \mathbf{V}, \mathbf{W})$ so that points in the affine set are generated as $\mathbf{u} + \mathbf{V}\mathbf{b}$ and $\mathbf{W}$ is the orthogonal complement of $\mathbf{V}$.
3. Analysis scale s
4. Significance level α

Algorithm: for each candidate model q in $\mathcal{Q}$ (the number of elements in $\mathcal{Q}$ is N_T

1. Select all points in $\mathcal{X}$ whose distance is less than or equal than s, $\mathcal{X}' = \{\mathbf{x} \in \mathcal{X} : d(\mathbf{x}, q) \leq s)\}$.
2. Compute empirical distribution of distances in $\mathcal{X}'$
3. Compute the Kolmogorov-Smirnov score using the variant that only considers the largest *positive* difference beteween the empirical and background distribution
4. Declare a detection if the KS score is below α/N_T

4.2 Simmulation

Inputs:

1. Total number of points N
2. Proportion of points that go to all the models, ρ
3. Number of ground truth models, M
4. Ambient space dimension n

5. Affine space dimension $m < n$
6. Distribution $g()$ used to displace model points from their exact position (e.g., Laplacian)
7. Scatter ratio σ: scale of displacement distribution

Algorithm:

1. Draw $(1 - \rho)N$ background points so that their coordinates are uniformly distributed between 0 and R
2. Repeat M times:
   (a) Draw $m+1$ points $\{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_m\}$ in the same way as the background
   (b) Set $\mathbf{u} = \mathbf{y}_0$
   (c) Set $\mathbf{y}'_i = \mathbf{y}_i - \mathbf{u}$
   (d) Construct an orthogonal basis $\mathbf{V}$ for the affine set using $\{\mathbf{y}'_1, \ldots, \mathbf{y}'_m\}$
   (e) Construct $\mathbf{W}$, a basis for the orthogonal complement of $\mathbf{V}$
   (f) Add $q = (\mathbf{u}, \mathbf{V}, \mathbf{W})$ to the set of ground truth models
   (g) Draw $K = \rho/M$ random $\mathbb{R}^m$ points $\{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$ so that $b_i \sim \text{Uni}(0, R)$
   (h) Draw $K = \rho/M$ random $\mathbb{R}^{n-m}$ points $\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$ so that $\|a_i\|_2 \sim g(\delta)$ and the direction of $\mathbf{a}$ is uniformly distributed. This is easy to approximate by first sampling from a Gaussian distribution and then rescaling the points to norm δ where δ follows the desired distribution
   (i) Construct model points $\mathbf{x}_i = \mathbf{u} + \mathbf{W}\mathbf{a}_i + \mathbf{V}\mathbf{b}_i$
   (j) Add model points to $\mathcal{X}$

Background points are generated so that their coordinates follow a uniform distribution between 0 and R, where R is a parameter.

5 Results