# An intrinsic dimension estiamtor

Koorosh Sadri

## 1 Introduction

Consider a set of data points of length $D$ (points in $R^D$) which lie on a curved unknown manifold of lower dimension $d$. It is clear that the dimension $d$ is preserved if you simply describe your data points in any other space of dimension $d' > d$. For example a line is always a line whether described in a plane or a 3-dimensional space or even in a 4-dimensional space-time as a worldline. This allows us to attach $d$, (the intrinsic dimension of our data points) to the data set as a property of the data set itself.

What follows is a method to estimate the intrinsic dimension given (noisy) data points.

### 1.1 Different length scales of the problem

There are three signififcant length scales in the problem; The first is the radius of curvature of the unknown manifold[1] which we shall denote by $R$. The second is the mean distance between nearest data points, denoted here by $\lambda$ and the last length scale is the amplitude[2] of the noises attached to data points; here denoted by $\varepsilon$.

As an example of a manifold to demonstrate the three lengthscales, consider a piece of paper with little sand particles (to play the role of our data-points) glued on it. In this picture, the mean distance between adjacent sand particles is $\lambda$ and the size of the sand particles is the noise amplitude $\varepsilon$ since larger sand particles deviate more from the paper sheet. $R$ then is simply the curvature of our plane.

Following the example to a extreme case, let's consider the described paper crumpled down to a paper ball. In this case it is clear that the data-points (sand particles) don't exhibit any two dimensional characteristics and are randomly

[1] This, ofcourse differs from point to point and direction to direction, as will get clear the smallest Radii will be the most concerning one. You can think of $R$ to be $\inf\{R\}$ really.

[2] It is common and possible for the noise to have different amplitudes in different directions. The next section will discuss the matter in more details.

distributed through the 3D space. This tells us that estimating the intrinsic dimension is only possible when the conditions

$$R \gg \lambda \qquad R \gg \varepsilon$$

are simultaneously met. Hence we shall limit our considerations only to situations where the above two conditions stand. Note that the conditions are met only when sufficient low noise ($\varepsilon \ll R$) data points are present.

Three different regimes of interest can be recognized based on the ratio between $\varepsilon$ and $\lambda$.

**(i)** $\varepsilon \ll \lambda$ low noise regime
**(ii)** $\varepsilon \sim \lambda$ medial regime
**(iii)** $\varepsilon \gg \lambda$ Many data regime
We will study the three different regimes in future sections.

### 1.2 Local isotropy

different components of our data points may have different variations due to both

### 1.3 Homogeneity assumption

## 2 Low noise regime

Let's begin by the simplest case where not nuisant data points (or at least $\varepsilon \ll \lambda$) are given. We'll find an expression for the distribution of $\ell$ ,the distance to the nearest data-point, by assuming a dimensionality of $d$ for the manifold. The $d$ that matches best to the data points will be reported as the estimated dimension.

### 2.1 Toss till tails

Solving the following simple problem will prove useful in future sections.

Consider tossing an unfair coin (probability $p$ for 'tails') till getting the first 'tails'. The number of tosses before quitting the game is a random variable we denote by $N$. To end the game after $N$ tosses you need to get $N-1$ successive 'head's and a final 'tail'. The distribution for $N$ is hence given by

$$P[N] = (1-p)^{N-1}p \qquad N > 0$$

### 2.2 $\ell$ distribution

Close enough to a data point (distances of order $\lambda$) the manifold can be regarded as a $d$ (still unknown) dimensional flat (since $R \gg \lambda$) space. The other data points are distributed randomly and uniformly here and around. The problem in this section is to find a distribution for the distance of the nearest data point.

Let's first calculate the probability of finding a data point in an infinitesimal $d$-volume of $\delta V$, assuming a uniform data per volume density $n$. Dividing a volume $V = N\delta V$ to $N$ identical pieces each of volume $\delta V$ the number of points inside the orignial volume will obey a Poisson distribution of parameter $\mu$. We have

$$\mu = \lim_{N\to\infty} \mathrm{P}\{\text{a data point exists in a volume}\frac{V}{N}\}N = \langle \#\text{ of points inside } V\rangle = nV$$

Which results in a simple expression for the probability of interest as

$$\mathrm{P}\{\text{a data point exists in a small volume } \delta V\} \approx n\delta V$$

Now for the nearest point to be further than a distance $\ell$ apart, there should be no data points inside a sphere of radius $\ell$ whose volume we denote by $V_d(\ell)$. Dividing the volume into small pieces we get

$$\mathrm{P}\{\text{no data points in volume } V_d(\ell)\} = \lim_{N\to\infty}\Big(1 - \frac{nV_d(\ell)}{N}\Big)^N = e^{-nV_d(\ell)}$$

Taking minus the derivative with respect to $\ell$ the distribution for $\ell$ becomes

$$f_d(\ell) = n\frac{dV_d}{d\ell}e^{-nV_d(\ell)}$$

The volume can be calculated to be

$$V_d(\ell) = \frac{\pi^{\frac{d}{2}}\ell^d}{\Gamma\Big(\frac{d}{2}+1\Big)}$$

And the distribution can be written with proper $s$

$$f_d(\ell) = \frac{d}{s}\Big(\frac{\ell}{s}\Big)^{d-1}e^{-\left(\frac{\ell}{s}\right)^d} \tag{1}$$

## 2.3 $\ell$ distribution in terms of $\lambda$

Trying to rewrite eq. (1) in terms of the mean distance $\lambda$, let's first find a relation between $\lambda$ and $s$

$$\begin{aligned}\lambda &= \int_0^\infty \frac{d}{s}\ell\Big(\frac{\ell}{s}\Big)^{d-1}e^{-\left(\frac{\ell}{s}\right)^d}d\ell \\ &= s\Gamma\Big(1+\frac{1}{d}\Big)\end{aligned}$$

Substituting $s$ we get

$$\frac{1}{\lambda}d(\ell/\lambda)^{d-1}\Gamma^d(1+1/d)\exp\{-[(\ell/\lambda)\Gamma(1+1/d)]^d\} \tag{2}$$

## 2.4 Method

From the data present, $\lambda$ can be calculated carefully to be $\langle \ell \rangle$ leaving only one parameter of the distribution free namely the dimension $d$. Running an M.L.E. method to estimate $d$ first we need to find the maximum to the function

$$\log d + (d-1)\langle \log \ell \rangle + d \log \Gamma(1+1/d) - \langle \ell^d \rangle \Big( \frac{\Gamma(1+1/d)}{\lambda} \Big)^d$$

## 2.5 examples

**Cantor set.** generating a million points in the cantor set, taking 10000 of them as a sample to calculate expected values. we get

$$d = 0.6273$$