Distance Between the Closest and Farthest Neighbors

Étienne Pepin

January 23, 2025

In this document I prove (theorem 1) that as the number of dimensions increases, the distance between the nearest and farthest neighbor of an isotropic distribution tends toward a constant that is dependent on the number of samples in the distribution and not the number of dimensions. This theorem is based on the fact that iid distributions have a distance distribution that tends towards a normal distribution (2).

Theorem 1. *Let $\mathbf{Y}_d = [Y_1, \ldots, Y_d]$ be an iid random vector. Then, if $X = ||\mathbf{Y}_d||$ and $d \to \infty$, then*

$$\mathrm{E}[X_{(n:n)} - X_{(1:n)}] = c \tag{1}$$

, where c is a constant independent of d for a given Y_i and n.

Proof. From theorem 2, $\lim_{d\to\infty} X \xrightarrow{dist} N(\sqrt{\mu d}, \frac{\sigma^2}{4\mu})$. Let $A \sim N(\sqrt{\mu d}, \frac{\sigma^2}{4\mu})$ and $B \sim N(0, \frac{\sigma^2}{4\mu})$. Then $A = B + \sqrt{\mu d}$ and $A_{(i:n)} = B_{(i:n)} + \sqrt{\mu d}$. Now we calculate the expectation of the difference between farthest and closest neighbor

$$\begin{aligned}\mathrm{E}[A_{(n:n)} - A_{(1:n)}] &= E[(B_{(n:n)} + \sqrt{\mu d}) - (B_{(1:n)} - \sqrt{\mu d})] \\ &= E[B_{(n:n)} - B_{(1:n)}]\end{aligned}$$

The quantity $E[B_{(n:n)} - B_{(1:n)}]$ is the expectation of the range (David and Nagaraja 2003, p. 1) and can be calculated numerically easily for any distribution and n value. Since it is independent of d, when X can be approximated by a normal distribution then the expectation of the range of the approximation is only dependent on σ^2, μ and n and not the number of dimensions. □

Because the distance distribution converges in distribution to a normal distribution with a variance that is independent of the dimension, the difference between the farthest and closest neighbor also converges, even if the mean of the distribution increases.

This brings context to Beyer et al. 1999 where they looked at the ratio of distances between the closest and farthest neighbor. If $\sqrt{\mu d} + B_{(1)}$ represent the closest neighbor, then the ratio of distances when d tends toward infinity can be written in terms of order statistics:

$$\lim_{d\to\infty} \frac{\mathrm{E}[\sqrt{\mu d} + B_{(1)}]}{\mathrm{E}[\sqrt{\mu d} + B_{(n)}]} = 1 \tag{2}$$

since

$$\lim_{d\to\infty} \sqrt{\mu d} = \infty \quad \text{and} \quad \mathrm{E}[B_{(1)}], \mathrm{E}[B_{(n)}] \in \mathbb{R}$$

This confirms the findings of Beyer et al. 1999 for the normal distribution, but also adds context. They mentioned that "as dimensionality increases, the distance to the nearest neighbor approaches the distance to the farthest neighbor". The ratio of distances tends toward 1, but the expected distance between closest and farthest neighbor can stay the same, as is the case for any multivariate distribution with iid distributions along all axes. This also confirms findings in Aggarwal, Hinneburg, and Keim 2001 where they found that the distance between farthest and nearest neighbor tends toward a constant for the Euclidean norm.

With the order statistics model we can also predict the expected range between the farthest and closest neighbor for iid d-dimensional variables, as long as the mean and variance along any axis are known. Figure 1 shows the expected distance between the closest and farthest neighbor using the normal approximation $N(\sqrt{d\mu_{Y^2}}, \frac{\sigma^2_{Y^2}}{4\mu_{Y^2}})$. The distance increases with the number of samples and is stable in higher dimensions.

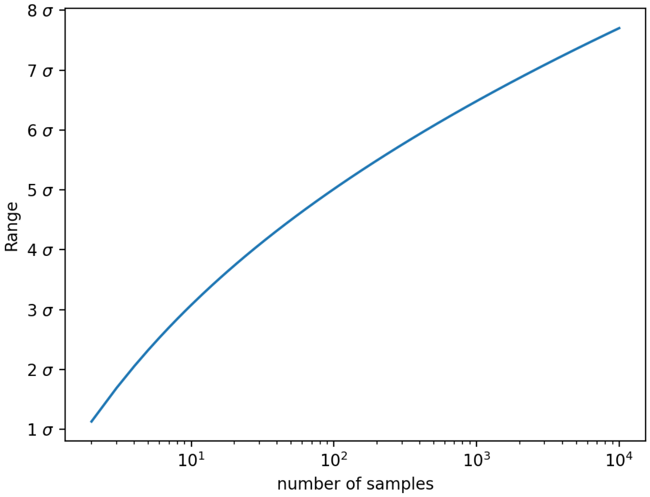

Figure 1: Expected distance between the closest and farthest neighbor in high dimensions using the normal approximation with $\sigma = \sqrt{\frac{\sigma^2_{Y^2}}{4\mu_{Y^2}}}$

The following theorem approximates Euclidean distance distributions as a normal distribution as the number of dimensions tends toward infinity. Angiulli 2018 provided the distribution for the squared Euclidean distance, here we derive the Euclidean distance.

Theorem 2. *Let $\mathbf{Y}_d$ be any multivariate random distribution such that $\mathbf{Y}_d = [Y_1, Y_2, \ldots, Y_d], y_i \in \mathbb{R}$, $\mathrm{E}[\mathbf{Y}] = \mathbf{0}$ and all Y_i are iid. Then the distance distribution $X_d = ||\mathbf{Y}_d||$ converges in distribution to a normal distribution such that*

$$\lim_{d\to\infty} X_d \xrightarrow{dist} N(\sqrt{d\mu_{Y^2}}, \frac{\sigma^2_{Y^2}}{4\mu_{Y^2}}) \tag{3}$$

where μ_{Y^2} is the expectation of Y_i^2 and $\sigma^2_{Y^2}$ is the variance of Y_i^2.

Proof. We start by proving it for X_d^2 and then for X_d. X_d^2 tends in distribution toward a normal distribution following the central limit theorem (CLT) such that when d tends toward infinity:

$$\frac{\sqrt{d}\left(\frac{\sum Y_i^2}{d} - \mu_{Y^2}\right)}{\sigma_{Y^2}} \xrightarrow{dist} \mathcal{N}(0,1)$$

$$\sum_{i=1}^{d} Y_i^2 - \mu_{Y^2} d \xrightarrow{dist} \mathcal{N}(0, d\sigma^2_{Y^2})$$

$$\sum_{i=1}^{d} Y_i^2 \xrightarrow{dist} \mathcal{N}(d\mu_{Y^2}, d\sigma^2_{Y^2})$$

For example, when $Y_i \sim \mathcal{N}(0,1)$, X_d^2 is a chi-squared distribution with Y_i^2 having $\mu_{Y^2} = 1$ and $\sigma^2_{Y^2} = 2$, which makes $X_d^2 \sim \mathcal{N}(d, 2d)$ when d is large.

The proof for the square of the distance makes use of the delta method, which states that if there is a sequence of random variables X_d satisfying

$$\sqrt{d}[X_d - \theta] \xrightarrow{dist} \mathcal{N}(0, \beta^2)$$

, then

$$\sqrt{d}[g(X_d) - g(\theta)] \xrightarrow{dist} \mathcal{N}(0, \beta^2[g'(\theta)]^2)$$

for any function g satisfying the property that $g'(\theta)$ exists and is non-zero valued. Continuing from the CLT demonstration:

$$\sqrt{d}\left(\frac{\sum Y_i^2}{\sqrt{d}} - \mu_{Y^2}\sqrt{d}\right) \xrightarrow{dist} \mathcal{N}(0, \sigma^2_{Y^2} d)$$

let $g(x) = \sqrt{x}$, $\theta = \mu_{Y^2}\sqrt{d}$ and $\beta^2 = \sigma^2_{Y^2} d$ with $[g'(\theta)]^2 = \frac{1}{4\mu_{Y^2}\sqrt{d}}$, then:

$$\begin{aligned} \sqrt{d}\left(\frac{\sqrt{\sum Y_i^2}}{d^{1/4}} - \sqrt{\mu}_{Y^2} d^{1/4}\right) &\xrightarrow{dist} \mathcal{N}(0, \frac{\sigma^2_{Y^2}\sqrt{d}}{4\mu_{Y^2}}) \\ \sqrt{\sum_{i=1}^{d} Y_i^2} - \sqrt{\mu_{Y^2} d} &\xrightarrow{dist} \mathcal{N}(0, \frac{\sigma^2_{Y^2}\sqrt{d}}{4\mu_{Y^2}})\frac{1}{d^{1/4}} \\ \sqrt{\sum_{i=1}^{d} Y_i^2} &\xrightarrow{dist} \mathcal{N}(\sqrt{d\mu_{Y^2}}, \frac{\sigma^2_{Y^2}}{4\mu_{Y^2}}) \end{aligned}$$

Which proves the theorem. □

References

Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim (2001). “On the Surprising Behavior of Distance Metrics in High Dimensional Space”. en. In: *Database Theory — ICDT 2001*. Ed. by Gerhard Goos et al. Vol. 1973. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 420–434. ISBN: 978-3-540-41456-8 978-3-540-44503-6.

Angiulli, Fabrizio (Apr. 2018). “On the behavior of intrinsically high-dimensional spaces: Distances, direct and reverse nearest neighbors, and hubness”. In: *Journal of Machine Learning Research* 18, pp. 1–60.

Beyer, Kevin et al. (1999). “When Is “Nearest Neighbor” Meaningful?” en. In: *Database Theory — ICDT’99*. Ed. by Catriel Beeri and Peter Buneman. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 217–235. ISBN: 978-3-540-49257-3.

David, H. A. and H. N. Nagaraja (2003). *Order statistics*. 3rd ed. Hoboken, N.J: John Wiley. ISBN: 978-0-471-38926-2.