You have **2** free member-only stories left this month. Sign up for Medium and get an extra one

Introducing Distance Correlation, a Superior Correlation Metric.

A modern-day metric that addresses the number one problem of Pearson's correlation

Terence Shin Feb 12 · 4 min read ★

Photo by Coffee Geek on Unsplash

Table of Content

3. Mathematics Behind Distance Correlation

4. Implementing Distance Correlation in Python

. . .

Introduction

I think we can agree that one of the most commonly used measures in business is correlation, more specifically, Pearson's correlation.

To recap, correlation measures the **linear** relationship between two variables, and that in itself is already a problem because there are MANY relationships that are not linear.

And so, for the sake of an example, you might conclude that the relationship between variable X and revenue is not correlated, when it in fact is correlated, just not linearly.

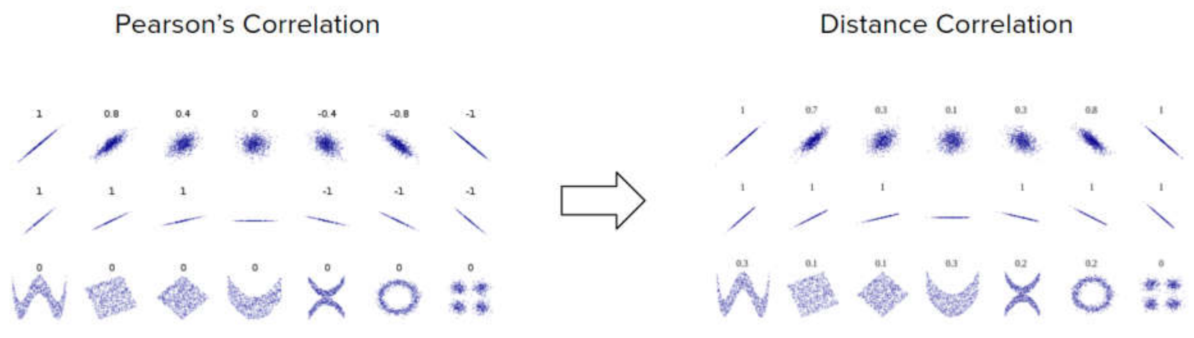
And this is where **distance correlation** comes in!

. . .

What is Distance Correlation?

Distance correlation is a measure of association strength between non-linear random variables. It goes beyond Pearson's correlation because it can spot more than linear associations and it can work multi-dimensionally. Distance correlation ranges from 0 to 1, where 0 implies independence between X & Y and 1 implies that the linear subspaces of X & Y are equal.

The image below shows how distance correlation measurements compare to Pearson's correlation.



The formula for distance correlation as follows:

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)\,dVar(Y)}}$$

Distance correlation formula

Distance correlation is not the correlation between the distances themselves, but it is a correlation between the scalar products which the "double centered" matrices are composed of.

If that didn't make sense to you, let's dive deeper into the math.

. . .

Mathematics behind distance correlation

Let (Xk, Yk), k = 1, 2, …, n be a statistical sample from a pair of two random variables, X & Y.

First, we compute the n by n distance matrices (aj, k) and (bj, k) containing all pairwise distances.

$$a_{j,k} = \left\|X_j - X_k\right\|, \; j,k = 1,2,\dots,n$$
$$b_{j,k} = \left\|Y_j - Y_k\right\|, \; j,k = 1,2,\dots,n$$

Then we take the double centered distances.

$$A_{j,k} = a_{j,k} - \bar{a}_{j\bullet} - \bar{a}_{\bullet k} + \bar{a}_{\bullet\bullet}$$
$$B_{j,k} = b_{j,k} - \bar{b}_{j\bullet} - \bar{b}_{\bullet k} + \bar{b}_{\bullet\bullet}$$

centered matrix).

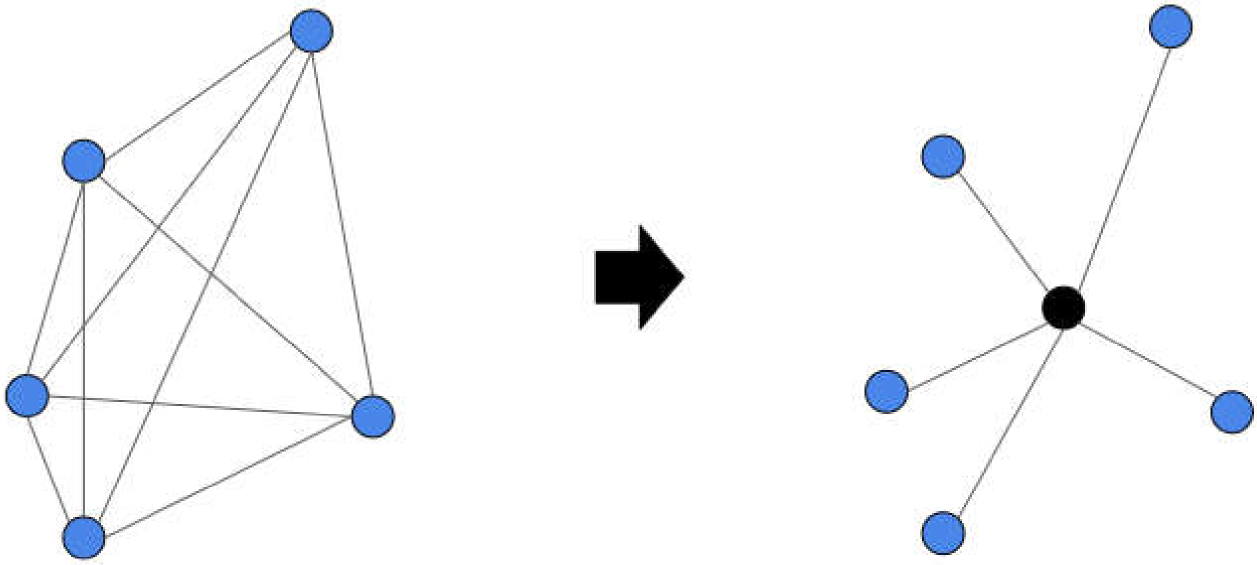


Image created by Author

Why do we do this?

The reason that we do this is for the following reason. Any sort of covariance is the cross-product of moments. Since distances aren't moments, we have to compute them into moments. To compute these moments, you have to calculate the deviations from the mean first, which is what double centering achieves.

Lastly, we compute the arithmetic average of the products A and B to get the squared **sample distance covariance**:

$$dCov_n^2(X,Y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{j,k}B_{j,k}$$

Distance covariance formula

The **distance variance** is simply the distance covariance of two identical variables. It is the square root of the following:

$$dVar_n^2(X) = dCov_n^2(X,X) = \frac{1}{n^2}\sum_{k,l} A_{k,l}^2$$

. . .

Implementing Distance Correlation in Python

Convinced that this is the metric for you? You're in luck because there's a library for distance correlation, making it super easy to implement.

Here's an example code snippet:

```
import dcor

def distance_correlation(a,b):
    return dcor.distance_correlation(a,b)
```

With this function, you can easily calculate the distance correlation of two samples, a and b.

. . .

Thanks for Reading!

I hope you found this interesting! Personally, I've found this extremely useful in my day-to-day, and I hope you find it useful too.

There are definitely pros and cons to this metric and I would love to hear your thoughts. What do think about a correlation metric that can detect non-linear relationships but is bounded by a range only between 0 and 1?

As always, I wish you the best in your learning endeavors!

Not sure what to read next? I've picked another article for you:

10 Statistical Concepts You Should Know For Data Science Interviews

Study smart, not hard.

towardsdatascience.com

and another one!

21 Tips for Every Data Scientist for 2021

#19. Learning how to set expectations will make a big difference in how "successful" you are in your career.

towardsdatascience.com

Terence Shin

- *If you enjoyed this, follow me on Medium for more*
- *Interested in collaborating? Let's connect on LinkedIn*
- *Sign up for my email list here!*

Sign up for The Variable

By Towards Data Science

Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. Take a look.

Your email

Get this newsletter

By signing up, you will create a Medium account if you don't already have one. Review our Privacy Policy for more information about our privacy practices.

Data Science Statistics Machine Learning Education Artificial Intelligence

Get started

Open in app

Download on the App Store

GET IT ON Google Play