# Statistical Measures of Distance

ARTUR JAROSZEWICZ and MEGAN ROYTMAN, UCLA

ABSTRACT

Measures of statistical distance are widely used in techniques such as clustering and classification, when we wish to identify objects that are in some sense similar to each other. The choice of distance between data points is an important one, and there are a large number of measures to choose from. In this paper we present some of the most commonly used measures of distance and compare their usefulness when analyzing data sets with different properties. They include general metrics such as the Minkowski, which includes the classic Euclidian, the Chebyshev, the Manhattan, and the Hamming distance. We present the Mahalanobis metric, which is similar to the Euclidian but corrects for strong structure in the data. In addition, we present the concept of correlation as a distance measure, covering the properties of the Pearson and Spearman correlations.

## 1. INTRODUCTION

The concept of statistical distance is useful when we want to identify objects that are in some sense close to each other. Such measures of distance are useful in classic statistical techniques such as clustering and classification, where we attempt to categorize data into distinct classes. There are a host of measures available for quantifying the distance between data points, but how does one choose among all these measures? Issues that can affect a metric's appropriateness in a given situation include things like the distributions of the data, the presence of outliers, and the linearity of the relationship we intend to model. However, more generally, we must consider the similarities between our data points that we wish to emphasize, whether it be there absolute magnitudes or perhaps some higher level patterns we hope to capture.

The purpose of this paper is to review some of the most important distance measures that are currently used in the field and consider some of the relationships between them. In addition, we provide examples of applications for each metric and why the particular choice of distance is appropriate for the problem at hand.

## 2. CLASSES OF DISTANCE MEASURES

All measures of distance should in some way indicate the strength of the relationship between two data points. However, different classes of measures satisfy different properties, and some are more stringent than others.

The most restrictive definition of a distance measure is the metric. A metric $d : X \times X \to \mathbb{R}$ must satisfy the following four conditions:

(1) $d(x, y) \geq 0$ (separation or non-negativity axiom).
(2) $d(x, y) = 0$ if and only if $x = y$ (identity of indiscernibles, or coincidence axiom).
(3) $d(x, y) = d(y, x)$ (symmetry).
(4) $d(x, z) \leq d(x, y) + d(y, z)$ (subadditivity or triangle inequality).

Although these metrics are most commonly used to find distance between two points in a vector space, there are other generalized metrics used. For instance, *pseudometrics* can violate the positive definiteness property, *quasi metrics* can violate the symmetry property, and *semi metrics* can violate the triangle inequality.

Another general class of distance measures is the similarity function. A similarity function must satisfy the following three properties:

(1) $0 \leq s(x, y) \leq 1$.
(2) $s(x, y) = s(y, x)$.
(3) $s(x, x) = 1$.

In contrast with the distance metric, the higher the value of the similarity function, the more related the two data points are.

## 3. MINKOWSKI METRIC

The Minkowski metric, also known as the *L-p metric*, is a generalized version of several widely used metrics, including the Euclidean metric, Manhattan, and Chebyshev metrics. Because it is parametrized, it can be used for a wide variety of applications and situations. It is defined as

$$d(\vec{X}, \vec{Y}) = \left( \sum_{i=1}^{n} (X_i - Y_i)^p \right)^{\frac{1}{p}}, \tag{1}$$

where $p$ is a parameter describing the relation between different dimensions. This parameter can be tuned from the data or chosen *a priori* depending on knowledge about the structure of the data. We will be exploring the more common choices of $p$, and discussing the merits and problems associated with each. Throughout these descriptions, we will be motivated by the problem of clustering genes based on their *expression values*. Briefly, if we consider $n$ total samples under the same conditions $X_1, ..., X_n$, and observe the expression of each of $m$ genes for each sample, we can construct a *gene expression matrix* as follows:

| | Samples | | | | |
|---|---|---|---|---|---|
| Gene | $X_1$ | $X_2$ | $X_3$ | $\dots$ | $X_n$ |
| Gene $A$ | $X_{1A}$ | $X_{2A}$ | $X_{3A}$ | $\dots$ | $X_{nA}$ |
| Gene $B$ | $X_{1B}$ | $X_{2B}$ | $X_{3B}$ | $\dots$ | $X_{nB}$ |
| Gene $C$ | $X_{1C}$ | $X_{2C}$ | $X_{3C}$ | $\dots$ | $X_{nC}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| Gene $m$ | $X_{1m}$ | $X_{2m}$ | $X_{3m}$ | $\dots$ | $X_{nm}$ |

In this review we will take the approach of clustering genes over the samples. We begin our discussion with the $L^2$, or Euclidean, metric.

### 3.1 Euclidean Metric

If we set $p = 2$ in the Minkowski metric, we obtain the Euclidean, or $L^2$, metric, e.g.,

$$d(A, B) = \sqrt{\sum_{i=1}^{n} (X_{iA} - X_{iB})^2}.$$

We consider two genes to be "close" if they are similarly expressed across all samples in the group, and further away if they are expressed differently across the samples. In Euclidean or rectangular geometrically, this is the canonical metric; it is derived from the Pythagorean Theorem, which is used to calculate distances in our common understanding of dimensions. It is both translation- and rotation-invariant, and easily lends itself to optimization. For example, when trying to minimize distance between some genes and a cluster center, it is enough to minimize the terms underneath the square root. Because each of the $n$ terms is quadratic, the derivative is linear, which is easily solvable. Although the Euclidean metric is arguably the most ubiquitous metric, it may not make the most intuitive sense in certain spaces.

Although the Euclidean metric is rotation invariant, it may not make sense to speak of rotation in a high-dimensional space such as gene expression levels. Additionally, observing a large distance across a single sample necessarily *decreases* the additional contribution to total distance from other samples. For example, observing a distance of 100 in one dimension and 20 in another is scarcely larger than the distance in just the first dimension; in fact, the total distance in this case is $\sqrt{100^2 + 20^2} \approx 102$.

### 3.2 Generalized $L^p$ Metrics for $p > 2$ and the Chebyshev Metric

As we take values of $p$ larger than 2 for the Minkowski metric, distances become more and more heavily weighted by singular dimensions with large deviations. In optimization problems, this has the effect of making results more dense, as opposed to $p < 1$ metrics, which yield sparser results.

At the extreme, if we look at the Minkowski metric as $p$ approaches infinity, we obtain the Chebyshev metric:

$$\lim_{p \to +\infty} \left( \sum_{i=1}^{n} (X_A - X_B)^p \right)^{\frac{1}{p}} = \max_{i=1}^{n} |X_{iA} - X_{iB}| .$$

It is equal to the maximum absolute deviation over all dimensions, hence being known as the *chessboard metric*, named so for the number of moves it takes a king to move to another square, which is the maximum of the number of squares moved horizontally and vertically. In our example of calculating distances between genes, the distance would be equal to the maximum difference across samples. Under such conditions, a deviation of 100 within any sample would yield the same total distance as observing a deviation of 100 within all samples. This could be used for a process of elimination method: if we are confident that the genes within a cluster should be consistently expressed across all samples, we can use the Chebyshev metric to eliminate situations in which one observation has a much larger deviation than others. For the same reasons, the Chebyshev metric allows for *denser* solutions, e.g., for a maximum distance of 5 within a cluster, any and all deviations across all observations are unrestricted, and can be between 0 and 5 with no added penalty.

There has been much argument to the usefulness of $L^p$ metrics where $p \geq 2$ in high dimensional data. According to Beyer, et. al. [1], the ratio of $L^p$ (for $p \geq 2$) distances between the closest neighbor and the furthest neighbor to a given point approaches 1 as the dimensionality of the data grows large. What this implies in practice is that there ceases to be any separation between what it means to be "close" or "far" in any high-dimensional data, and thus that distances of this sort become meaningless.

As an example, let us take an $n$-dimensional $L^2$ sphere of radius 1, and place points uniformly within it. As we take higher and higher $n$, more and more points are found at the boundary of the sphere. This can be shown with a simple example: given a 10-dimensional unit sphere, only about a third of the volume is found within a radius of $.9$ of the center. In fact, we see as we approach greater and greater dimensionality, the ratio of the volume found at the boundary of an $L^2$ sphere to the volume of the whole sphere approaches 1, supporting the idea that the closest neighbor and furthest neighbor become equidistant from a single point. This motivates us to use $L^p$ metrics with $p \leq 1$, and preferably a fraction [2].

### 3.3 Manhattan Metric

If we set $p = 1$ in the Minkowski metric, we obtain the equation

$$d(A, B) = \sum_{i=1}^{n} |X_{iA} - X_{iB}|,$$

where the absolute value is necessary for the condition of non-negativity. Intuitively, the Manhattan metric is the sum of all the differences across all dimensions. As compared to the Euclidean metric, the Manhattan metric is more robust to outliers. There are no dependencies between different observations, i.e., varying the deviation in a single dimension while holding others constant has a linear effect on the total distance. This is particularly useful when we would like our distance to scale linearly across all dimensions. However, due to the fact that each term is linear, there are not always stable or unique solutions to optimization problems such as finding an optimal cluster center. Also, due to a lack of analytical solutions, it is inefficient to find optima.

### 3.4 Generalized $L^p$ "Metrics" for $p < 1$ and Hamming Distance

If we use a $p$ less than one, we have the exact opposite properties as a $p > 2$ metric. Namely, distances become more heavily weighted not by singular dimensions with large deviations, but by deviations along multiple dimensions. Strictly speaking, such distances are not true metrics, as they do not satisfy the triangle inequality or subadditivity property. This does not prohibit us from using it in practice, however. As with all other $L^p$ metrics except for $L^2$, it is difficult to optimize. Unlike the Chebyshev and similar metrics, however, solutions tend to be sparse. We would use such a distance measure in situations where we would like to penalize consistent changes across multiple observations or could allow for a small number of observations to be largely deviated.

An extreme case of this sort of metric is as we take the limit as $p$ approaches 0. In this case,

$$\lim_{p \to 0} \left( \sum_{i=1}^{n} (X_A - X_B)^p \right)^{\frac{1}{p}} = \sum_{i=1}^{n} \mathbf{1}(X_{iA} \neq X_{iB}),$$

where **1** is the indicator function. This is known as *Hamming Distance*. Though basically useless in application such as gene expression, it is quite useful in applications like string comparison, where $X_{iA}$ is defined to be the character in the $i^{th}$ position of string $A$.

## 4. MAHALANOBIS METRIC

The Mahalanobis metric, like the Minkowski, is a generalization of the Euclidean metric. It is defined as

$$d(\vec{X}_1, \vec{X}_2) = \sqrt{(\vec{X}_1 - \vec{X}_2)^\top \Sigma^{-1} (\vec{X}_1 - \vec{X}_2)},$$

where $\Sigma$ is the covariance matrix of vectors $\vec{X}_1$ and $\vec{X}_2$:

$$\Sigma = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \dots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \dots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \dots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

This metric is particularly useful for data drawn from populations with strong structure. What we mean by this is that many dimensions are highly correlated or dependent. A very good motivating example is that of *Genome Wide Association Studies*, or *GWAS* for short. GWAS is the examination of many variants throughout the genome with the purpose of finding ones that are associated with a given phenotype. In GWAS, the genotypes, or genetic states, of many individuals are measured. At any given locus $i$ in the genome, $X_i$ represents the binary vector of individuals' states: either not having a mutation (0) or having one (1). Because of the way offsprings' genomes are inherited from parents, locations closer to each other on the genome are more likely to be inherited together. This creates structure in populations, as loci that are closer to each other are more strongly correlated with each other. Due to this correlation between neighboring mutations, and due to some mutations being more prevalent in one population than another, we often obtain spurious results from statistical testing.

Using the Mahalanobis metric in this problem, we first calculate the covariance matrix, or the covariance between every pair of loci. Multiplying the distance between vectors by inverse covariance matrix $\Sigma^{-1}$ has the effect of *decorrelating* every pair of vectors, assuming each vector is Gaussian. Once we do this, each vector is independent of every other, making the calculating of association statistics substantially easier. It is more difficult to intuit what *distance* means in this case, but it can be thought of the *unlikeness* of any two individuals not with regards to the population they come from, but due to random fluctuation. If we consider a set of $k$ variants with population frequencies of 50% in two populations, we can have the scenario where two individuals from different populations happen to have many of the same variants by chance and may be closer than two individuals from the same population who happen to have many different variants. In practice, the Mahalanobis metric is more commonly used for finding similarities in DNA and protein sequences in order to predict evolutionary changes [3] and protein function by way of similarity to other proteins classes [4].

A special case of the Mahalanobis metric is when the covariance matrix $\Sigma$ is the identity matrix $I$. In this scenario, the correlation between every pair of vectors is zero, and the distance between any

two such vectors is simply equal to its inner product, or it's Euclidean distance:

$$d(\vec{X}_1, \vec{X}_2) = \sqrt{(\vec{X}_1 - \vec{X}_2)^\top I(\vec{X}_1 - \vec{X}_2)} = \sqrt{\sum_{i=1}^{n} (\vec{X}_{1i} - \vec{X}_{2i})^2}.$$

Another special case is where the covariance matrix is diagonal, but not the identity matrix. This is equivalent to the *weighted* Euclidean distance, where each dimension, though independent, has an weight $k$, $0 \leq k \leq \infty$, attributed to it which describes the importance of the dimension.

## 5. CORRELATION

The correlation coefficient is useful when looking at higher-level associations between variables. Oftentimes we wish to measure similarity between two higher-dimensional data points not by the absolute differences in their magnitudes, but by their corresponding patterns of fluctuation. A common example of such an application is gene expression analysis. When attempting to cluster genes into functional categories, it is more interesting to know whether the expression of two genes has corresponding variation over time, rather than requiring similar magnitudes of expression at any given time point.

The two correlation coefficients we will cover here are the Pearson and Spearman coefficients.

### 5.1 Pearson Correlation Coefficient

The Pearson product-moment correlation coefficient, or *r*, is the most commonly used measure of correlation. It is a means of measuring the strength of the association between two variables which have a linear relationship. The possible range of the Pearson correlation coefficient is -1 to 1. If the two variables indicate a strong positive relationship, *r* will have a value close to 1. A strong negative relationship will produce a value close to -1. A value close to 0 will thus signify a weak relationship between the two variables.

The Pearson correlation coefficient *r* has the following definition: Consider two variables *X* and *Y*, each having *n* values $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$. The Pearson correlation *r* is

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}, \tag{2}$$

where $\bar{X}$ and $\bar{Y}$ are the mean of $X$ and $Y$, respectively.

For an intuitive explanation of this formula, let us consider the meanings of the numerator and denominator. The denominator contains the sum of squares of the deviations of each $X$ and $Y$ value from their means. Their summations are what we use to calculate the variance and standard deviations of these variables. However, they differ from the variance in that they are not divided by the total sample size. These summations are thus termed *variations* rather than variances. Dividing by the square root of the product of the variations in $X$ and $Y$ leads to the scaling of the correlation coefficient to have limits of -1 and +1 [5].

The numerator of this expression represents the *covariation* of $X$ and $Y$. It considers the difference

of each consecutive data point of $X$ and $Y$ from its mean, and multiplies the two values together. Intuitively, when these differences both have high values in a single direction, the covariation of $X$ and $Y$ is high. When they have low values in a single direction, the covariation is low, but positive. When they have values in opposite directions, the covariation becomes negative. Thus the formula for the Pearson correlation represents an aggregate covariation between $X$ and $Y$, corrected for the individual variations such that the coefficient always falls between -1 and +1 [5].

This correlation can be converted to a measure of distance when its value is subtracted from one. It is often used, as previously mentioned, to cluster genes together based on the similarity of their expression over time, for example.

### 5.2 Spearman Correlation Coefficient

When a variable is measured at the ordinal level, or on an arbitrary numerical scale, we can either treat it as if it were on an interval scale or use a correlation coefficient designed for ordinal variables. The Spearman correlation coefficient, $r_s$, is often used in the latter case [5].

When using the Spearman coefficient, the values of the variables are ranked. The ranks of the variables are determined by ordering the vectors $X$ and $Y$ from lowest to highest, or highest to lowest. For each data point, the variable takes on the value of its rank rather than its true value on the scale. The formula for the Spearman coefficient then follows that of the Pearson coefficient,

$$r_s = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}, \tag{3}$$

where $x$ and $y$ now designate ranks rather than actual values.

One important property for the data to exhibit when using Spearman's correlation is monotonicity. Spearman's coefficient is often used rather than Pearson's when the data are not linearly related, but the data should still follow a monotonic relationship. That is, the data should exhibit one of the following patterns: (1) as the value of one variable increases, the value of the other variable also increases; or (2) as the value of one variable increases, the value of the other variable decreases. This requirement follows naturally when we consider what happens when we apply a ranking function to monotonic data - the ranked version becomes linear, in fact. This is why we can successfully use an analogous form of the Pearson correlation on ranked data to produce the Spearman coefficient.

Another interesting consideration when using the Spearman correlation is that it is more robust (or less sensitive) to outliers. An extreme value in the data will produce an extreme rank, but this rank will not deviate from the rest of the data points as the true value would in the Pearson method.

## 6. DISCUSSION

There is an abundance of distance measures that we can use in statistical analysis, and the choice between them should not be arbitrary. Each measure is designed for a specific purpose, and it is useful to understand the properties each one best models. In addition, it is important to think about the relations between our data that we most hope to capture.

As seen with the Minkowski metric, for example, variations on the parameters of a metric can lead to

very interesting behaviors. When we wish to model our Minkowski distance as the maximum deviation in one dimension, we choose a $p$ that is higher than 2. Conversely, when we wish for the accumulation of smaller, sparser deviations to produce higher distances, we should choose a $p$ lower than 1.
As we saw with correlation, the nature of our data becomes very important in the choice between Pearson and Spearman. Linear relationships between continuously measured numerical data are tend to be very well modeled by the Pearson correlation. However, when our relationships are not linear or our data comes in a nominal form, we may think about using the Spearman correlation instead. And more generally, the choice between using absolute measures of distance such as Minkowski and Mahalanobis and using more shape-based measures such as correlation is often a very crucial one.

The steps we must take before choosing a distance measure are thus two-fold: we must think carefully about both our data and our measure. What are the relationships we wish to focus on in our data? And which distance measure best accounts for the particular properties our data exhibits? These are the questions we should strive to answer when choosing a distance measure for statistical analysis.

## 7. DIVISION OF WORK

Artur Jaroszewicz wrote sections 3 and 4 on the Minkowski and Mahalanobis metrics. Megan Roytman wrote the Abstract, Introduction, the Correlation section, and the Discussion. The division of labor was approximately 50/50.

## REFERENCES

K. Beyer, J. Golstein, R. Ramakrishnan, U. Shaft, "When is Nearest Neighbors Meaningful", *ICDT Conference Proceedings*, 1999.

C. Aggarwal, A. Hinneburg, D. Klein, "On the surprising Behavior of Distance Metrics in High Dimensional Space", *Database Theory, ICDT 200, 8th International Conderence, London, UK* January 4 - 6, 2001.

H. Suzuki, M Sota, CJ Brown, EM Top, *Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes*. Nucleic Acids Res. 2008 Dec;36(22):e147.

W. M. Liu, K. C. Chou, *Prediction of protein structural classes by modified mahalanobis discriminant algorithm*. J Protein Chem. 1998 April; 17(3): 209?217.

P. Gingrich, *Introductory Statistics for the Social Sciences*. Regina: U of Regina, 1993.