

Measures of Association and Distance

All data modeling is about inferring relationships. Relationships may be between objects. For example, a novel protein is related to a known protein that is a kinase and therefore is also likely to be a kinase. [There is a slight of hand in the previous sentence. The fact that a protein is related to another protein does not necessarily mean that we can transfer the function of one protein to the other (so-called, "annotation transfer"). That all depends on the meaning of this "relationship"—a topic, we will extensively investigate later.] Relationships may be between variables. For example, the expression level of gene A may be related to the expression level of gene B and therefore the two genes may be controlled by the same upstream gene C. Relationships may be between more complex objects that represent data. For example, the genealogical tree graph for gene A may be related to genealogical tree graph for gene B to a degree that depends on ancestral recombination events between the two loci.

In fact, the idea of relationships can be extended very generally. We may have objects/variables defined in one domain that have relationship to objects/variables defined in another domain. For example, the relationship between the three dimensional shapes of proteins (domain of coordinate configurations) and their binding constant with DNA (domain of statistical reaction constants). Or, the relationship between the transcriptome of a cell (domain of RNA counts) and a set of environmental conditions like pH, temperature, tissue coordinates (domain of environmental parameters). Data models try to uncover such relationships, ideally leading to a more detailed understanding such as a physical-chemical-dynamical model of why the relationship exists. Often the inference of the relationships themselves is the model. (Sometimes the relationship is interpreted as a causal model; e.g., "change in the pH causes the change in gene expression".) Or, the establishment or the observation of the relationships hint at some more complex or higher order relational organizations. For example, the relationship between pH and gene expression leads to models of environmental sensing and transcription regulation.

To begin to construct such models, we usually start with some marginal characterization of relationships. Suppose we wanted to infer a genealogical tree. We might follow the following steps:

0. Input a matrix of multiple-aligned sequences
1. Compute a measure of distance between each position of a pair of aligned DNA sequences.
2. Summarize the distance measure over all positions.
3. Compute such summary for every possible pairs of alignment.
4. Estimate a tree graph representation of the relationships that imply the pairwise distance relationships.

Each of the above steps involves computing some simpler measure of association/distance that is built up into a more complex picture of relationships. In this unit, I will discuss some of these simpler measures of relationship. In subsequent units, we will discuss higher-order inferences.

Given two objects, it is more natural to talk about their distances (as realized on some measured variables) while given two variables it may be more natural to talk about their association (as manifested on some observations). Both distances and association measures can be thought of as complementary measures so I will not try to make any fundamental distinction between them. I should also note that often we are given data in the following "data matrix" form:

	v1	v2	v3	v4
obj 1				
obj 2				
obj 3				
obj 4				

where the rows are objects and the columns are variables whose values were measured on the objects. Let's call this the data matrix **R**. As might be expected from the idea that we can always transpose this matrix we can regard objects and variables as complementary structures. In fact, I could have said the previous sentence as "columns are variables and rows are objects denoting the actualized values of the variables". The rows of R might be called "points" while columns might be called "coordinates", or rows might be called "strings" and columns be called "sites", etc. I will just use the generic term "objects" and "variables", unless I am talking about specific examples. In a more abstract framework, we can consider variables as a function[1] over the objects. For example, "eye color" is a function that assigns the values "blue, black, red, green…" to each individual (our objects). Conversely, we can also think of objects as functions over the variables where to each variable value we can assign a value {0, 1} dependent on whether an object has that value (so-called indicator functions). It suffices to say that we should be mindful of the dual nature of what we call objects and variables.

Nominal Values

We start with the consideration where the values in the cells of **R** are what we call "nominal" values. This means that each element $r_{ij} \in \mathbf{R}$ can take values from some set, say {red, blue, black}, where all we know is that "red" is different from "blue", but not how much different or even whether "red" < "blue". If the values can be ordered, then we call them "ordinal" values, and if we can order the difference of the values then we call them "metric" values. The simplest case for nominal values is the binary set, say {female, male}, which we will just generically denote by the values {0, 1}. We might compute a measure of distance between ith and jth objects by summing the number of variables that are different:

$$d(R_i, R_j) = \sum_{k=1}^{q} I\{r_{ik} \neq r_{jk}\} \quad \text{(Eq 14.1)}$$

where I{X} is an indicator function taking the value 1 when X is true and 0 when it is false. I will also use the subscript notation R_i to denote the ith row (object) of the data matrix **R** and the superscript notation R^i to denote the ith column (variable). We can compute analogous quantity for the variables to measure distance between two variables by:

$$d(R^i, R^j) = \sum_{k=1}^{n} I\{r_{ki} \neq r_{kj}\} \quad \text{(Eq 14.2)}$$

Since these are variables, it might be more pleasing to compute the association measure by normalizing by the number of observations and subtracting from 1 to make this a similarity measure:

$$r(R^i, R^j) = 1 - \frac{1}{n}\sum_{k=1}^{n} I\{r_{ki} \neq r_{kj}\} \quad \text{(Eq 14.3)}$$

If the nominal values are binary so that the data matrix looks like this:

	R^1	R^2	R^3	R^4	R^5
R_1	0	1	0	1	1
R_2	0	1	1	0	1

We can also think of the binary values as denoting inclusion and exclusion of the set of variables (or objects). So that we might rewrite:

$R_1 = \{R^2, R^4, R^5\}$ and $R_2 = \{R^2, R^3, R^5\}$

[1] Here, I mean function in the mathematical sense, say like $f(x) = x^2$.

to which we can apply a general notion called Jaccard similarity coefficient:

$J(R_1, R_2) = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|}$ (Eq 14.4)

This can also be made into a distance measure, Jaccard distance:

$D_J(R_1, R_2) = 1 - \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|}$ (Eq 14.5)

Jaccard's coefficient is purposefully defined over abstract collections described as sets. But the above data matrix with binary values can be more concretely characterized by counting the number of shared 1's and 0's. Let f_{00} be the number of variables where R_1 and R_2 share 0's, f_{01} is where R_1 is 0 and R_2 is 1, f_{10} is the converse, and f_{11} is where they share 1's. Then Jaccard's coefficient and distance are:

$J(R_1, R_2) = \frac{f_{11}}{f_{00}+f_{10}+f_{01}+f_{11}}$ $D_J(R_1, R_2) = 1 - \frac{f_{11}}{f_{00}+f_{10}+f_{01}+f_{11}}$ (Eqs 14.6 and 14.7)

The above motivates us to consider a variation called the Simple Matching Coefficient:

$SM(R_1, R_2) = \frac{f_{00}+f_{11}}{f_{00}+f_{10}+f_{01}+f_{11}}$ (Eq 14.8)

which can also be made into a distance by subtracting from 1. Note that this is equivalent to Eq 14.3.

Before we go on, we should note that the intuitive idea of a distance suggests that whatever quantity, *d(i, j)*, that we call a distance should have some characteristics that befit the standard notion of distance. We should have:

d(i, i) = 0, and *d(i, j) > 0* if $i \neq j$. (Eq 14.9)

That is, distance of an object to itself should be zero and the measure should have positive values otherwise. In general, distance measures should also be symmetric so that:

d(i, j) = d(j, i). (Eq 14.10)

And, well-behaved distance measures should satisfy the triangle inequality:

$d(i, j) + d(j, k) \geq d(i, k)$ (Eq 14.11)

The triangle inequality basically says that if you go from object *i* to *k* through another object *j*, that path should be at least as great as direct path from *i* to *k*. Later it will turnout many of the conditions of Eq 14.9—14.11 are not satisfied by measures that we will still call distances. But, the above conditions hold for the standard intuitive notion of distance.

Back to the future.

Jaccard's coefficient and the simple matching coefficient differ only in how they treat the shared value of zeros. Which should we use when? Suppose the interpretation of the zero is in fact "absence". For example, say each variable is a gene and 0 means "no expression while 1 means "expressed". Now

supposed we have 15,000 genes. Two objects (say cells) might differ in 10 genes where each gene is expressed in only one cell and not the other. It also turns out that only 10 genes are expressed in either cells and they share 14,990 non-expressed genes. Is it reasonable to say that the two cells have the association 14,990/15,000 = 0.9993? What if two other cells are co-expressing 9 out of the 10 genes so their simple matching coefficient is 14,999/15,000 = 0.99993? Would we notice the difference? On the other hand, if we were to compute the Jaccard coefficient we would have 0/10 = 0 for the first case and 9/10 = 0.9 for the second case. Which of these two calculations are reasonable for describing the relationship of the two cells and their transcriptomes? The answer, of course, depends on the biological question. Most cells express large number of genes, so if two cells turned off 14,990 genes, no matter what happens with the "on" genes, it may be that we should consider these cells very similar. On the other hand, maybe we have a special sub-population of cells that we know tends to shut off most of the genes in the genome, then maybe we really care about the fine distinction made by those genes that are on.

Whether shared absence of something should be considered as part of evidence for relationship is an age-old problem in biology. If two species both lack wings, does that indicate closeness? What about lack of a membrane bound nucleus? A more recent manifestation of this problem is where people compare microbiomes and determine relatedness of communities while counting shared absence of some species. Because this is an important issue systematicists have invented the terms "sympleisiomorphy" to denote traits shared by two organisms as an ancestral trait (which are often "absence") and "synapomorphy" to denote derived (and different from the ancestor) shared traits. I should note that these two types of relatedness cannot be resolved from looking at pairs of objects and require a third reference object. In fact, as we try to infer more complex models, it will be often the case that pairwise relationships are insufficient amount of information.

The previous example is a KEY POINT of this unit and also subsequent units. ***What measures of association or distance is appropriate really depends on the question, the biology, and the problem that is being addressed.*** For the most part, there really isn't a formulaic answer to what is the most appropriate quantity to compute. One important thing that helps think about right kinds of measures is considering the possible configurations of the background domain of inference. For example, if we are considering cells that generally have most of their genes expressed then shared absence of expression of 99% of the genes is an important hint at a relationship. If we were considering cells that are generally quiescent then it is not an important indicator.

The fact that we don't have set answers for what to do is generally true for all kinds of numerical characterization of data. Don't let people tell you that you always have to use statistic A instead of statistic B. It all depends. Those who insist on a particular rule without considering the problem are revealing their ignorance. In fact, sometime using the "wrong procedure" can be more efficient for some problems. The crux is to have enough knowledge to know when this is the case.

Going back to the binary data matrix, the Jaccard coefficient, simple matching coefficient and possibly others (e.g., so-called Sorensen index, $\frac{2|A\cap B|}{|A|+|B|}$) are functions of the counts of shared 0's, 0 and 1's, 1 and 0's, and shared 1's. We can make the counts for objects R_1 and R_2 into a table like this:

	$R_1 = 0$	$R_1 = 1$
$R_2 = 0$	f_{00}	f_{10}
$R_2 = 1$	f_{01}	f_{11}

Or, if we were measuring relationship between variables, R^1 and R^2:

	$R^1 = 0$	$R^1 = 1$
$R^2 = 0$	f_{00}	f_{10}
$R^2 = 1$	f_{01}	f_{11}

In fact, the variable R^1 might be say an indicator variable for a class of objects, say $R^1 = 0$ means normal while $R^1 = 1$ means "diseased". So, the above table could be a depiction of relationship between a variable, R^2, and classes of objects (indicated by R^1). For nominal values, the above table is called a **contingency table**, usually with the idea that there are two variables, each with p and q different nominal states, measured on n objects, The Jaccard's index and simple matching coefficient basically measures the size of the diagonal elements compared to either the total or another marginal count. This suggests several other possible measures such as the determinant of the above table, $D = f_{00}f_{11} - f_{10}f_{01}$. Or, we can measure the standard chi-square statistic:

$$\chi^2 = \frac{(f_{00}-f_{.0}f_{0.}/n)^2}{f_{.0}f_{0.}/n} + \frac{(f_{10}-f_{.0}f_{1.}/n)^2}{f_{.0}f_{1.}/n} + \frac{(f_{01}-f_{.1}f_{0.}/n)^2}{f_{.1}f_{0.}/n} + \frac{(f_{11}-f_{.1}f_{1.}/n)^2}{f_{.1}f_{1.}/n} \quad \text{(Eq 14.12)}$$

where $f_{i.} = \sum_{j=1}^{p} f_{ij}$ and $f_{.j} = \sum_{i=1}^{q} f_{ij}$; i.e., the marginal sum of ith row or jth column and n is the number of all cells of the contingency table. After some algebra it will turn out that for this 2-by-2 table,

$$\chi^2 = \frac{nD}{f_{.o}f_{.1}f_{0.}f_{1.}} \quad \text{(Eq 14.13)}$$

where D is the aforementioned determinant. Again, it is not clear whether any of these measures are better than others. The determinant has some nice geometric interpretation while the chi-square statistic, which something like a "normalized determinant" can be converted into a probability value by some assumptions about the distribution of the counts.

We can easily think up of several other possible measures. For example, a classic measure is the odds ratio:

$$(f_{00}/f_{01})/(f_{10}/f_{11}) \quad \text{(Eq 14.14)}$$

The first part is the odds of the value "0" vs "1" in the column variable, given that the row variable is "0". The second part is the odds of the value "0" vs "1" in the column variable, given that the row variable is "1". The ratio of the two is the "odds ratio". Of course, if the row variable has no association with the column variable then we would expect the ratio to be about 1.0.

All of the measures we discussed so far can be extended to multi-valued nominal values such as {A, C, G, T}. A relevant example is if the data matrix, R, is a multiple alignment of DNA string and each column variable is an alignment site, each row is a string (say from different organisms), and the possible values are {A, C, G, T}. For ith and jth string, can characterize each site as being same in value or different. It follows that we might compute a measure of distance between ith and jth string by summing the number of columns (sites) that are different:

$$d(i,j) = \sum_{k=1}^{q} I\{r_{ik} \neq r_{jk}\} \qquad \text{(Eq 14.15)}$$

where I{X} is an indicator function taking the value 1 when X is true and 0 when it is false. The distance measure defined in Eq 14.15 is usually called Hamming distance, which is related to the Jaccard coefficient. We can also make a contingency table with multiple rows and columns and compute the

determinant or the chi-square statistic, etc. We can also substitute I{X} some scoring matrix *w(x,y)* like we did for sequence alignment and sum of the scores. We will see later that we can treat variables as a realization of a stochastic process like Jukes-Cantor model we previously discussed and then compute a distance measure that corresponds to the expectation of the stochastic process.

Before I go on, let's consider the contingency table for two nominal variables with four states {A, B, C, D} analogous to the tables above:

	R^1 = A	R^1 = B	R^1 = C	R^1 = D
R^2 = A	**100**	**0**	35	21
R^2 = B	**0**	**100**	35	32
R^2 = C	28	29	38	27
R^2 = D	26	33	27	31

If we look at the above table, we can see that there is a strong association between the states A and B (red bold letters), but not with rest of the states. If we compute some kind of summary relationship between R^1 and R^2, say by computing a chi-square statistic, the resulting numbers may or may not suggest that the two variables are related. Whatever the number is, it will most likely not indicate that A and B are especially related.

This above data matrix illustrates two problems. First, while for any paired relationships, we can come up with reasonable measures of distance or association. However, when there are many pairs of comparison, it is hard to come up with a summary that everybody can agree on. For example, given a single locus, computing the genetic distance between two populations is straightforward. If we have multiple loci, there are many different measures of genetic distance proposed by many people that all try to capture different aspects genetic relationships. Second, the problem above show an important and difficult to solve problem: relationship between objects or variables may only exist for a subset of cases. For example, suppose we have various different cells, say fibroblasts, cardiomyocytes, neurons, and embryonic stem cells, and we have measured the transcriptome so we have say 10,000 variables. It is very likely that some of the genes may have one kind of relationship within some of the cells while either the same genes or a different set of genes may have a different kind of relationship within other cells. However, in combination over all cells and all genes, we may not see a clear relationship. Finding relationships within unknown arbitrary subsets of cases (i.e., subsets of objects and variables) is a both conceptually and computationally difficult problem. We will see a few examples later.

Ordinal values

Ordinal values have states that we can order; e.g., "white" < "sky blue" < "blue". Some times we can create partial orders but not complete orders. For example, two people might have an ancestor descendent relationship (e.g., mother > son). An ordering is complete if for every pair of states/objects we can state the order while it is a partial order if some pairs have no comparison. No comparison is different from "equal". We can allow the ordering relationship to have "white" = "sky blue". Given a set of measurements, e.g., "white", "sky blue", "blue", "Prussian blue", "black", and an ordering such that we can order any pair of values, we can use the pairwise ordering to rank the set and replace the values with ranks: {white = 1, sky blue = 2, blue = 3, Prussian blue = 4, black = 5}. If we have a data matrix, R, as before with ordinal values, we can replace it with ranks—except that the ranks will have to correspond to either ranks of variables for each object or ranks of object for each variable For example, we could for each cell rank the transcriptome by their relative expression or for each gene we can rank the cells by relative expression for that gene across each cell. For the most part we cannot rank both rows and columns simultaneously.

Given two sets of ranks, say ranks of expressed genes for two cells, we can compute a measure of association of the two cells by the expression ranks. A commonly used measure is Spearman's rho, which is computing Pearson's correlation coefficient directly on the ranks with one caveat that if we allow for ties in the rankings then the average of the ranks are used for the computation. I discuss Pearson's correlation coefficient below so I will defer discussion of this measure—I just note that we use exactly the same formula except that the *x* and *y* values will be ranks rather than metric numbers. However, one thing to note is that Spearman's rho considers the relative magnitude of the rank difference to be meaningful. That is, the difference between rank 1 and rank 2 is smaller than that between rank 1 and 10. This may seem reasonable, on the other hand, we have to remember the whole reason to consider ordinal variables is not for ranks but because we want to be considering only relative orders of values.

An alternative measure of rank association is Kendall's tau, which mainly tries to see if there is consistency in the relative ranks. From the data matrix R, suppose we have ranked variables R^1 and R^2 where the rankings are for each variable across the objects. Consider now the ith object and jth object and the values for the two variables: (r_{i1}, r_{i2}) and (r_{j1}, r_{j2}). Since both variables are ranked, we can ask if the rankings are consistent. That is, the two sets of ranks are **concordant** if $[r_{i1} < r_{j1}$ AND $r_{i2} < r_{j2}]$ or $[r_{i1} > r_{j1}$ AND $r_{i2} > r_{j2}]$ otherwise they are **discordant.** We now consider all n(n-1)/2 possible pairs of objects (excluding self-comparison) and count the number of concordant pairs of ranks and discordant pairs of ranks. Kendall's tau is:

$$tau = \frac{n_c - n_d}{n(n-1)/2} \quad \text{(Eq 14.16)}$$

That is, the average number of difference in the concordant and discordant pairs of ranks. I note that by its definition, Kendall's tau can be computed without using ranks. We just need to do binary comparisons. In this measure, when examining the quantity $[r_{i1} < r_{j1}$ AND $r_{i2} < r_{jl}]$ we do not care if the first pair of ranks is much more different than the second pair of ranks; just that they are consistent in direction. When we have ties in the ranking in either the first comparison $(r_{i1} < r_{j1})$ or the second $(r_{i2} < r_{jl})$, we leave them out of the counting and adjust the denominator. How the denominator is adjusted for ties can have different effects. In Kendall's tau-b, we do the following:

$$\sqrt{\left(\frac{n(n-1)}{2} - ties\ in\ first\right)\left(\frac{n(n-1)}{2} - ties\ in\ second\right)}$$

where "ties in first" means the cases where $(r_{i1} < r_{j1})$ comparison was tied and "ties in second" means the cases where $(r_{i2} < r_{jl})$ comparison was tied. We subtract each, multiply the result and take the square root---this is called the geometric mean of two quantities. On the other hand, a quantity called Goodman and Kruskall's gamma is:

$$gamma = \frac{n_c - n_d}{n_c + n_d} \quad \text{(Eq 14.17)}$$

That is, the denominator is simply the sum of the concordant and discordant comparisons, leaving out any of the ties. As always, the choice of the measure, Pearson's, Kendall's, Goodman and Kruskal's, all depend on the kind of relation one is trying to characterize. Finally, all of the measures of association can be made into measures of distance by appropriate transformation. For example,

dist = 1 - |Pearson's rho|.

or

dist = 1 − (Pearson's rho)²

(which one is better?)

Metric values

Metric values are the usual kind of numbers with which we are familiar. More precisely, not only can we order the values, but we can order the difference of the values. You have probably already seen many measures of relationship that can be computed on metric variables. The most common is the Euclidean distance measure:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \quad \text{(Eq 14.18)}$$

which we can apply to either the rows or the columns of the data matrix R. (That is, *x* and *y* represent either rows or columns of R.) The Euclidean distance is motivated by the Pythagorean theorem—which means the coordinate system for each x_j involves orthogonal axes. We will return to the distance measures in vector space later. It turns out that there are some important and elegant ideas.

For association measures, the most commonly used measure is the Pearson's correlation coefficient given by:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{(Eq 14.19)}$$

We can write both Eq 14.18 and 14.19 more compactly using the vector/matrix algebra notations. Suppose we look at a column of the data matrix for the variable *x*, then the values of the column represent the object values. We can write this as a vector[2] $\vec{x} = (x_1, \ldots, x_n)'$. Remember from the section on vectors and matrices that a vector is an object by itself without needing a list of numbers to represent it. When we say that we can represent the vector by an ordered set of numbers like $\vec{x} = (x_1, \ldots, x_n)'$, we are saying that there is a coordinate system for the vector $\vec{x}$. Conversely, if we have an ordered set of numbers like from a data matrix, we can view this as a set of coordinates for a vector with respect to some coordinate system. If we take a variable and a column of object values for this variable, we might call this a "variable vector" with its coordinates being object values. We can also be looking at a row of the data matrix and then consider this as an "object vector" with its coordinates being variable values. Whether a vector represents a variable or whether it represents an object depends on the context. I will note the difference when it matters.

Recall the Euclidean norm for vectors (using the notation < , > for inner products of vectors):

$$|\vec{x}| = \sqrt{<\vec{x},\vec{x}>} = \sqrt{\sum x_i^2}$$

[2] The notation (…)' is meant to represent the transpose of the row of values. This just follows the convention that a vector is usually given as a column of numbers. Since it is not easy to print a column of numbers "in-line" with plain text, it is conventional to add a transpose operator to the in-line notation to remind the reader that this is a column of numbers.

which denotes the length of the vector assuming an orthogonal coordinate system and the Pythagorian theorem. You can see that the distance formula in Eq 14.18 is in fact, the norm of the difference of the two vectors:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} = \sqrt{< \vec{x} - \vec{y}, \vec{x} - \vec{y} >} \quad \text{(Eq 14.20)}$$

This fact, that distances between two objects, can be expressed in terms of inner products of vectors will become very important later on.

For the correlation measure Eq 14.19, let's assume that the vectors $\vec{x} = (x_1, \ldots, x_n)'$ and $\vec{y} = (y_1, \ldots, y_n)'$ are mean-centered, such that the mean of the coordinates are zero. This means that we can ignore the subtraction by mean in Eq 14.19. We have:

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} = \frac{<\vec{x},\vec{y}>}{|\vec{x}||\vec{y}|} = < \frac{\vec{x}}{|\vec{x}|}, \frac{\vec{y}}{|\vec{y}|} > \quad \text{(Eq 14.21)}$$

The last part of Eq 14.21 says that the correlation coefficient is the inner product of two unit vectors in the x-direction and y-direction. From basic geometry we find that the inner product of two unit vectors equal the cosine of the angle between the two vectors, so we get:

$$r = < \frac{\vec{x}}{|\vec{x}|}, \frac{\vec{y}}{|\vec{y}|} > = cos\theta \quad \text{(Eq 14.22)}$$

and sometimes we define the "angular distance" as:

$$\theta = cos^{-1} r \quad \text{(Eq 14.23)}$$

Besides showing that the standard correlation coefficient measures the angle between two vectors, each of which represent the object measurements for a variable, Eq 14.22 shows an important idea: the inner product in vector space defines both distances (through the norm) and directions (through the angles). Distances and directions are the basic ingredients of geometry. Therefore, this single notion of an inner product establishes the basic geometry of vector space as represented by coordinate systems.

We can now extend the notion of a norm and distance in various ways. A very common set of norms is called p-norm, defined as:

$$|\vec{x}|_p = (\sum |x_i|^p)^{1/p} \quad \text{(Eq 14.24)}$$

which naturally defines a distance, which for historical reasons we call L_p-distance.

$$L_p(\vec{x},\vec{y}) = (\sum |x_i - y_i|^p)^{1/p} \quad \text{(Eq 14.25)}$$

Euclidean distance above is $p = 2$. When $p = 1$ the distance is sometimes called Manhattan distance, because the distance is computed by summing up each coordinates like what one would have to do when moving along city blocks. There is a special case of $p = \infty$, in which case we define the distance as the maximum of the coordinate differences:

$$L_\infty(\vec{x},\vec{y}) = \max_i |x_i - y_i|$$

At this point, there is an important point to make. In the topic of geometry, a very fundamental notion is defining distances between points. We start out with a collection of points, usually given as a set, say **S**. Then we define a notion of distance between pairs of points, say by the L_2 distance. The combination of set of points and a notion of distance creates a **space** and we might talk about the "L_2 space". In fact, if we think of normal objects such as points on earth, we can measure distances between two points without any notion of coordinates—this is what people used to do before GPS and all that. Now given a space, we can associate with points in the space a set of numbers that we call coordinates. The notion of coordinates can become pretty elaborate so for the moment think of it like the usual coordinate system that you are familiar with. The coordinates are just a systematic index of numbers to label each point. (I will return to the discussion of coordinates later.) Coordinates allow us to use algebra to compute things like distances and areas, etc. The main point here is that in the study of geometry, the notion of a space comprised of a set of points and distances is prior to the notion of a coordinate system. Having said this, now consider empirical datasets. In rare cases, we have direct measurement of distances between objects of interest. For example, we might have direct measurement of distances between two proteins by their degree of reactivity with an antibody. Or, we might have some micrograph image of a cell and directly measure things like diameter. But, except for these rare cases, we usually have a set of measurements for biological objects; say, whole transcriptome RNA levels for a cell. That is, we have a set of numbers that we have measured on the biological objects, which we think characterizes the relationships between the biological objects. That is, we've measured coordinate points first and we are trying to figure out the "biological space". Suppose we regard these measurements as coordinates of a vector space. Then, we are ***embedding*** the biological objects in a vector space comprised of a coordinate system of the measurements. The key point is this:

The biological objects may have a natural geometry that is not the geometry of their vector space embedding.

Let me explain this a bit more. If I ask you to think of earth, you will probably think of a globe like a basketball. This imagery is actually the view of earth as an object embedded in a 3-dimensional vector space. However, if you are on the surface of earth the relevant geometry is not the ball embedded in 3-D but how the points on the surface behave as we move around. We will soon find that it has funny properties like you can continue moving in one direction and eventually you come back to the same point and that the shortest paths only makes sense if we start thinking of the geometry as being curved. This is what we mean by the natural geometry of the object versus how it is represented as embedded in some other space (like n-dimensional vector space).

To continue with a biological example, suppose we made transcriptome measurements of neurons comprised of 15,000 different genes. Then, we will have a 15,000 dimensional vector space. However, suppose that when we look at the data, it turns out that all genes always have exactly the same expression level (this is just a hypothetical scenario). This means that every cell falls on a diagonal line from the origin, implying that the real geometry is that of a one-dimensional line. Of course, this is a completely unrealistic scenario but the point is that the set of transcriptome states occupied by a neuron is probably some subset of all the possible points in a 15,000 dimensional vector space. And, a key modeling goal is to identify exactly what subspace is relevant to cells and perhaps a notion of natural geometry of this subspace. Measurements on the biological objects place them in a standard p-dimensional vector space. What we want to know is the natural geometry of the objects[3]. We presume that this natural geometry will reflect the natural biological processes. For example, if two genes are always expressed at the same level,

[3] Here, I am talking about geometry in a loose sense as relationship of biological states for observed biological objects (e.g., cells). In practice, it is not clear whether such relationships can be characterized using the rigid notion of geometry that has been formalized in mathematics up to now.

then the natural geometry is a line and it probably arises because the underlying biological regulation couples the transcription of the two genes.

To discover natural relationships between biological objects (and therefore natural geometry), we want to take into account the biases, constraints, and dynamics arising from the biological processes. Suppose we measure the expression level of two genes over many different cells. And, we find that for the first gene, the values range from 50-60 while for the second gene the values range from 0 to 10,000. How should we measure distances between cells? If we use the embedded vector space geometry and Euclidean distances, then we should just take the differences for each gene, square them, sum them, and then take the square root. But, it doesn't take much to see that the second gene will dominate the distance value; the cells have a bias in their expressed values. In the terminology I used above, the natural geometry of the cells seem to indicate that we should rescale the coordinates to reflect the biological dispersion—the dispersion that is giving us hints about the natural geometry. One simple thing to do is rescale by the standard deviation of each gene's expression. (Why standard deviation instead of, say, the range?) What if we had two types of cells? Then, maybe we should rescale by some kind of pooled measure of standard deviation. This is more or less what we do when we compute a t-test statistic. We can think of a t-statistic as a distance measure that is scaled by the standard deviation to reflect the natural geometry of the objects. A generalized measure like this is called the Mahalanobis distance. The idea is to scale each coordinate by the standard deviation and in addition, if there is any correlation, de-correlate by rotating the axes. If the observed points follow a multivariate normal distribution with a particular covariance structure, Mahalanobis distance basically inverts the covariance structure such that the pattern of dispersion is uncorrelated and has constant variance in all directions. The idea is that if the natural geometry is such that the point scatter looks like multi-variate normal with non-zero covariances between variables, we should compute distances that reflect this natural geometry.

I should expand on the last point a bit. How can we try to learn what is the natural geometry? The data we have is the observed set of points (objects) like cells in the transcriptome vector space. We might argue that if the transcriptome vector space was the natural geometry, the cells should be evenly spread out through out the space. (In practice we would need to define "evenly" more carefully.) If the observed objects are not spread out evenly in the embedded space then maybe whatever the natural space is, we argue that our observed objects should be evenly spread in that space. Therefore, in the embedded vector space geometry, the pattern of spread of the observed points should give us information about the natural geometry. If we were to sample 3D coordinates of trees on earth with respect to some local astronomical coordinate frame, we should see a scatter that indicates a globe, even if it is clumped (by dry land). Then we might argue that if we have the right notions of distance (e.g., great circle distance) the scatter of the observed points should be evenly distributed according to this distance, reflecting the natural geometry. This is the idea behind measures of distance like the Mahalanobis distance. The problem is, of course, that it may be very difficult to see the dispersal pattern, we might have sampled poorly and the points (objects) might be concentrated in one or another, and there might not be a uniform geometry that applies to all observations.

Another important point is that the natural geometry might have constraints—for example, forbidden regions of the vector space where no biological point can exist because it is incompatible with biological function (e.g., all gene expression value are zero). The most obvious everyday example is streets. If we have two points on a map, we should not measure distances between the points by their Euclidean coordinates. We should be measuring path length over connected streets. A common representation of constraints is by using graphs to represent possible paths. Then, distances are measured by path lengths on the graph. Dijkstra's algorithm is an example of how we can compute path lengths over a graph. It is also possible that distances should be directional. For example, the genetic distance from a mother to a daughter might have a different notion than from a daughter to a mother. Incorporating these kinds of constraints into a coherent notion of distances is a pretty difficult task.

I close this section with the idea that we can measure distances not only between points (objects) but between sets of objects. For example, we might want to measure distances between two microbiomes. A commonly used measure in these cases is called the Hausdorff distance. Let the two sets be X and Y; say, two different microbiomes. For each object x in X find the object in Y that has minimal distance to x. For example, x might be E. coli, then we would might compute the evolutionary distance of E. coli to every microorganism in community Y and find the minimal distance. Given such a minimal distance, we repeat this procedure for all objects in X; say, all microorganisms in community X. Now take the maximum of such minimum distances (this is why Hausdorff distance is sometimes called max-min distance). We repeat the process starting with the set Y. That is, for each point in Y find the minimal point in X and then find the maximum of such distances. The larger of the two max-min distances is the Hausdorff distance. One of the main points of computing distance in this manner is that it keeps the distance consistent when there is overlap between the two sets. For example, suppose both X and Y have E. coli but no other species are in common. If we were to find, say minimal distance between X and Y, we might come out with a measure of zero because of the shared E. coli, but this would not make sense since the two sets are clearly different.