HUDM5124 Spring 2020

KEY – Assn. 1 Exercises on metric spaces and distance:

In the posted notes for Lecture 1, the following axioms are given. If you are interested, further discussion of the axioms is given in the two posted extracts from the book *What is Distance?* by Yuri Shreider.

These axioms are constraints that must be satisfied in order for a set of pairwise measures on a set to be considered as distances in a metric space. We will use the following set of axioms (note this is not the minimal axiomatization, but it is clear). The distances must satisfy:

a. symmetry ($d_{xy} = d_{yx}$) *(the distance of x to y must equal the distance of y to x)*

b. positivity ($d_{xy} > 0$) *(the distance between any two distinct points must be positive)*

c. minimality ($d_{xx} = 0$) *(the distance of a point to itself is = 0)*

d. the triangle inequality ($d_{xy} + d_{yz} \geq d_{xz}$) *(for the distances among any three points, the sum of any two of these distances must be equal to or greater than the third)*

In order for a geometric space (or any other kind of distance model) to make sense as a model for proximity data, the data ought to “look like” distances, i.e. the dissimilarities ought to satisfy the metric axioms as well.

Please answer the following questions.

DISTANCES IN AN ACTUAL SPACE.

1. Consider five points (labeled A-E) on a line. Discuss if each of the axioms above is satisfied for the actual interpoint distances.

Symmetry is of course satisfied for the distance between points on a line, as are positivity and minimality. For triples of points on a line, the triangle inequality is satisfied as an equality: given ordered points *a*, *b* and *c* on a line, $d_{ab} + d_{bc} = d_{ac}$.

PROXIMITY DATA. In the first course session, I have posted three example proximity data sets: the Rothkopf Morse code data (confusions among pairs of Morse code signals), Ekman’s data on similarity ratings of color chips, and a set of ratings of the dissimilarity among emotion terms that I gather some years ago.

2. For the emotions dissimilarities, discuss which of these axioms seem to be satisfied and which do not.

Emotion data:

Symmetry is satisfied for these data – they were collected in such a way as to ensure that, with each pair of stimuli given in only one order (or they were previously symmetrized as a first step in the data analysis).

Positivity is satisfied; the instructions to subjects were to rate dissimilarity on a 1-9 scale, which ensures that all dissimilarities are positive.

Minimality cannot be falsified, since we do not have information on the diagonal entries of the matrix.

Triangle Inequality: It's laborious to check every triple of objects by hand – better to do it with computer code (which we will do later). But checking a few triples,

d(happy,calm)=4.05 d(happy,elated)=2.44 d(calm,elated)=7.83
d(H,C) + d(H,E) = 6.49, which is less than d(C,E) = 7.83
So this first triple checked violates the T.I. (so we can stop)

d(happy,calm)=4.05 d(happy,amused)=3.35 d(calm,amused)=6.52
d(H,C) + d(H,A) = 7.40, which is greater than d(C,A) = 6.52
So this triple does NOT violate the T.I.

Etc.

3. For the Rothkopf Morse code confusions, discuss which of these axioms seem to be satisfied and which do not. Note that confusions data may be considered as similarities. Thus, you will have to "translate" the distance axioms as appropriate for similarity data; one way to do this is to assume that your similarities will be translated into dissimilarities via a linear transformation: e.g., $d_{xy} = 100 – s_{xy}$. BTW, the task that was used to gather these data was a "same-different" task, i.e. a pair of Morse code signals was presented (very rapidly) to the subject, and he or she had to say if the two tones were the same signal or different signals. The table shows the percentage of "same" responses (which is correct for the diagonal elements, and incorrect (a confusion) for the off-diagonal elements.

Often in an MDS, we might convert the similarities to "distances" by a linear transform before proceeding with the analysis. For certain types of MDS or cluster analysis, we might then check the metric axioms for the transformed data.

But here we want to see if the original similarities resemble (inverse) distances.

Rothkopf data:

Symmetry is NOT satisfied for these data. For example, s(D,A)=8 but s(A,D)=13; s(I,A)=64 but s(A,I)=46; s(Z,Y)=42 but s(Y,Z)=23; etc.

Positivity can't really be falsified for similarities that are assumed only to be related to "distances" by a linear transformation.

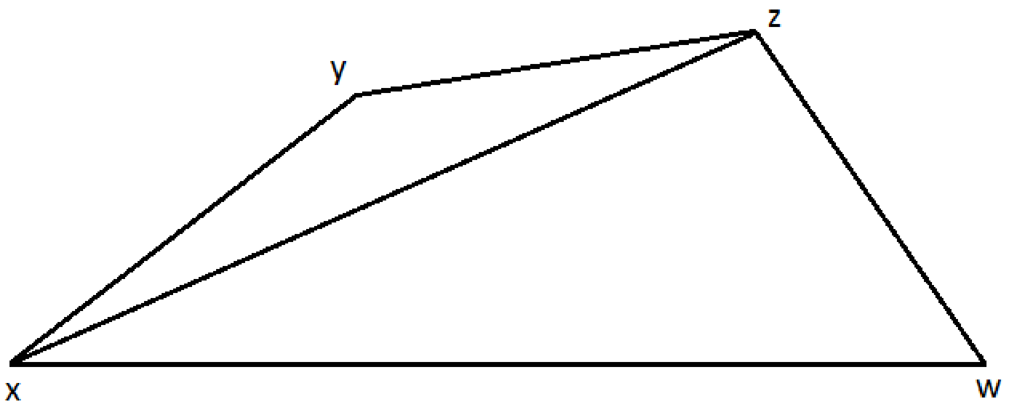
Minimality too cannot be falsified; however positivity and minimality taken together imply: d(x,y) > 0 = d(x,x) → s(x,y) < s(x,x) This seems to be satisfied for all rows & columns.

Triangle Inequality: When back-translated into a condition for similarities (by a linear transform), it's not clear exactly what this axiom would imply. But cases where s(x,y) and s(y,z) are very large while s(x,z) is very small would be suspicious or "fishy". One such fishy triple is: S(B,X) = 84, s(X,Q) = 61 but s(B,Q) = 25. Fishy cases are not very common.

OPTIONAL EXERCISES (more mathematical):

4. Prove that the triangle inequality generalizes to more than three points; i.e. that for points x,y,z,w, d(x,w) ≤ d(x,y) + d(y,z) + d(z,w)

Consider (w.l.o.g) the following configuration of points:



If the triangle inequality holds, then it follows that:
d(x,z) ≤ d(x,y) + d(y,z)
and d(x,w) ≤ d(x,z) + d(z,w).
Substituting the first inequality into the second, it is trivial to show that
d(x,w) ≤ d(x,y) + d(y,z) + d(z,w)

5.. Imagine a flat landscape with circular lakes of varying sizes scattered through it. You cannot travel through the lakes, only over land. Would the metric axioms hold for distances in this space? You can offer formal proofs, or merely arguments.

If we define "distance" to mean the minimal-distance path between any two points, then yes, the first three axioms must hold, AND the triangle inequality must hold. For if we were to find a d(x,z) that is greater than d(x,y) + d(y,z) for some y, we could simply choose an alternative (shorter) path, namely traveling between x and z by traveling first to point y then from there to z. Thus, d(x,z) can never be greater than d(x,y) + d(y,z).

(see picture below for a related case)