THE (GRAPHICAL) REPRESENTATION OF PROXIMITY INFORMATION: DESCRIPTIVE DATA ANALYSIS RUN AMUCK

Lawrence Hubert

The University of Illinois

The MATLAB M-files used in these examples (other than the one taken from the Statistics Toolbox for multidimensional scaling) are available at:

http://cda.psych.uiuc.edu/srpm_mfiles

I might begin with a little story. I was at a conference recently and was listening to a very opaque and rather confused paper presentation by one of my quantitative colleagues. After the paper was finished, an obviously more substantively-oriented listener sitting next to me, commented that he and his colleagues generally had the ability to share their research interests in presentations. He wondered why quantitative people often appeared only able to inflict theirs on others.

Well, hopefully I'll share more than I inflict today — but if you like, we could take a vote when I'm done.

To begin, it seems that for quite a long time now, most of my research interests have centered around data analysis strategies for what we might generically call proximity data or proximity information. Proximity merely refers to a collection of numerical information we have about the relationship between whatever objects or things we are interested in.

Also, I have pursued a path of making MATLAB M-files (open-source and freely available) to implement everything being developed. So, all the analyses we will discuss here are repeatable easily by the listener (with access to MATLAB or to a free clone like OCTAVE; plus in one case, the MATLAB Statistics Toolbox). The place to get all these M-files was noted on the title slide.

OUTLINE

A) A very brief review of what proximities are and where they may come from —

B) The general problem of attempting to represent (at least in an approximate manner) the information contained in proximity data by some formal structure intended to help explain the patterning of the original data —

C) Review five such formal structures in the context of a specific example (that deals with proximities between facial expressions of emotion):

1) Multidimensional scaling (representation by distances from placements in both Euclidean and city-block spaces)

2) Hierarchical clustering (representation through a sequence of nested partitions)

3) Additive-tree analysis (representation by distances from a placement in a specific tree-like graphical structure)

4) Unidimensional scaling, both Linear (representation by distances from a placement along a line) and Circular (representation by distances from a placement around a closed continuum)

5) Imposition of an anti-Robinson pattern (representation by fitting a certain gradient structure to the data matrix)

D) Possible extensions (and for which analysis routines are presently available in the way of M-files; for the present, however, we will not review these extensions):

1) The use of multiple (additive) structures for representation

2) Alternative formal representation structures, including those that incorporate a double-centering operation (through a centroid component)

3) Proximity data defined between two distinct sets of objects (two-mode as opposed to one-mode)

4) The inclusion of proximity data transformations (e.g., monotonic, or more constrained forms, such as convex, concave, or some of both)

Given some set of objects of interest — people, stimuli, variables, situations, and the like — the phrase "proximity data" merely refers to numerically specified information about the relation between each pair of objects.

Usually organized as an $n \times n$ matrix, say, $\mathbf{P} = \{p_{ij}\}$, where p_{ij} denotes the relationship between objects i and j (which is symmetric, so, $p_{ij} = p_{ji}$).

Where do proximities come from?

A) Direct measures (or judgements) of proximity —

pair comparisons: subjects judge the similarity of object pairs (stimuli), possibly according to some specific type of similarity

same/different judgements: percentage of "same" judgements for pairs of stimuli, or latencies for "same" judgements sorting: subjects partition objects into groups according to similarity, with proximity then defined by object-pair co-occurrence frequencies

joint occurrence frequencies: proportion of times that object pairs co-occur together (over sites, conditions, and so on)

interaction measures: degrees of communication or flow between objects

confusion measures: frequencies with which objects are confused with one another

B) Indirect (or derived) measures —

Given initial object by variable (attribute) data:

profile dissimilarity (distance measures) between objects

correlation or other measures of association between variables

GENERAL REPRESENTATION TASK

Given a proximity matrix $\mathbf{P}$, find some matrix, say $\mathbf{P}^*$, that is:

(a) "close" to $\mathbf{P}$ (captures [a major amount of] the information present in $\mathbf{P}$

(b) the entries in $\mathbf{P}^*$ have some particularly convenient structure that can be represented formally by some (graphical) mechanism

Thus, to explain what is going on in $\mathbf{P}$, we use $\mathbf{P}^*$, and the graphical mechanism it induces.

The criterion of "closeness" we use is consistently least-squares; thus we seek to minimize

$$\sum_{i,j}(p_{ij} - p_{ij}^*)^2 \ ,$$

by the choice of $\mathbf{P}^*$. Also, as a measure of "fit" adequacy we use the variance-accounted-for (vaf) criterion of

$$\text{vaf} = 1 - \frac{\sum_{i,j}(p_{ij} - p_{ij}^*)^2}{\sum_{i,j}(p_{ij} - \bar{p})^2} \ ,$$

where $\bar{p}$ is the mean of the off-diagonal entries in $\mathbf{P}$.

Analogies abound — e.g., to explain some single dependent measure, choose a multiple regression equation based on a collection of independent variables; interpret what is "going on" from the weights on the independent variables. "Closeness" here is also least-squares.

An Example for Illustration:

To make all of this a little more concrete, it might be useful if I introduce an example at this point, that I can carry through during the rest of the talk.

Some very early research dating back to the 1930s on the psychology of expression was occupied with the question of whether subjects could correctly identify an intended emotional message from a person's facial expression. It was found that the errors in the rather extensive number of misinterpretations were not random, and the emotion perceived was usually "psychologically similar" to the emotion expressed by the sender. A number of individuals, particularly Schlosberg, attempted to develop a theory of the differentiability of facial expressions, concluded that three perceptual "dimensions" were needed for a meaningful classification: pleasant/unpleasant, attention/rejection, tension/sleep.

Over a variety of different studies, subjects could fairly reliably classify facial expressions within this system. Although subjects may be able to classify facial expressions according to the Schlosberg scales when made explicit, it is still uncertain whether judges that are uninstructed use these particular “dimensions” in making judgements about facial expression — or possibly would use others.

Scene (Lightfoot series)	PU	AR	TS
1) [7] Grief at death of mother	3.8	4.2	4.1
2) [13] Savoring a coke	5.9	5.4	4.8
3) [15] Very pleasant surprise	8.8	7.8	7.1
4) [16] Maternal love — baby in arms	7.0	5.9	4.0
5)[20] Physical exhaustion	3.3	2.5	3.1
6) [28] Something wrong with plane	3.5	6.1	6.8
7) [29] Anger seeing dog beaten	2.1	8.0	8.2
8) [30] Pulling hard on seat of chair	6.7	4.2	6.6
9) [32] Unexpectedly meets old boyfriend	7.4	6.8	5.9
10)[36] Revulsion	2.9	3.0	5.1
11) [37] Extreme pain	2.2	2.2	6.4
12) [51] Knows plane will crash	1.1	8.6	8.9
13) [56] Light sleep	4.1	1.3	1.0

Note: the entries in brackets, [ ], indicate the original scene number in the Lightfoot series.

PU: pleasant/unpleasant;

AR: attention/rejection;

TS: tension/sleep

(Schlosberg Scale Values are from Engen, Levy, & Schlosberg, 1958, *JEP*, 454–458 — empirical averages on a 9-point scale over a group of subjects)

Correlations:

PU vs. AR: .18

PU vs. TS: $-.15$

AR vs. TS: .75

MULTIDIMENSIONAL SCALING

Given the proximity matrix $\mathbf{P}$, find a new matrix $\mathbf{P}^*$ that is "close" to $\mathbf{P}$ and the entries in $\mathbf{P}^*$ are (a linear transformation of) Euclidean or city-block distances (between the objects [faces] placed in some K-dimensional space).

For now, we use the MATLAB Statistical Toolbox routine, `mdscale.m`, with `Criterion` set to `metricstress` for the Euclidean scaling; for the city-block alternative, we use my routine, `biscalqa.m`.

Formally, if $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ denote the coordinates in two dimensions, then the distances between objects i and j are

city-block: $d_{ij} = |x_j - x_i| + |y_j - y_i|$

euclidean: $d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$

HIERARCHICAL CLUSTERING

Given the proximity matrix $\mathbf{P}$, find a new matrix $\mathbf{P}^*$ that is "close" to $\mathbf{P}$ (in a least squares sense), *and* the entries in $\mathbf{P}^*$ satisfy the ultrametric property: for any three objects [faces] i, j, and k, among the three corresponding entries in $\mathbf{P}^*$, p^*_{ij}, p^*_{ik}, and p^*_{jk}, the largest two must be equal.

Or less intuitively, we have the trio inequality: $p_{ij} \leq \max\{p_{ik}, p_{jk}\}$.

The MATLAB routine used is called `ultrafnd.m`

ADDITIVE TREE ANALYSIS

Given the proximity matrix $\mathbf{P}$, find a new matrix $\mathbf{P}^*$ that is "close" to $\mathbf{P}$ *and* the entries in $\mathbf{P}^*$ satisfy the additive-tree property: for any four objects [faces], i, j, k, and h, among the three sums, $p^*_{ij} + p^*_{kh}$, $p^*_{ik} + p^*_{jh}$, and $p^*_{ih} + p^*_{jk}$, the two largest must be equal.

Or, to keep the musical motif going, we have the less intuitive quartet inequality: $p^*_{ij} + p^*_{kh} \leq$ max$\{p^*_{ij} + p^*_{kh},\ p^*_{ij} + p^*_{kh}\}$.

Generally, in representing an additive tree graphically, each branch represents the common feature of those "below" (the progeny); the branch length indicates the importance of this group of common features (all within a Tverskian notion of common and distinctive features).

The MATLAB routine used is called `atreefnd.m`

An alternative view of additive trees represents them (very non-uniquely) as an ultrametric plus a centroid “metric” matrix. A (symmetric) centroid matrix, say, $\mathbf{C} = \{c_{ij}\}$, has main-diagonal entries, $c_{ii} = 0$, for $1 \leq i \leq n$, and off-diagonal entries $(i \neq j)$, $c_{ij} = g_i + g_j$, for some set of values, $g_1, \ldots, g_n$. Because some $g_1, \ldots, g_n$ may be negative (and lead to negative entries in $\mathbf{C}$), we put the word “metric” in quotes.

A (closed-form) least-squares approximation to $\mathbf{P}$ by a centroid “metric”, in effect double-centers the residual matrix (so row and column sums are zero).

UNIDIMENSIONAL SCALING

Linear (LUS):

Find a set of coordinates, $x_1, \ldots, x_n$, to minimize

$$\sum_{i<j} (p_{ij} - \{|x_j - x_i| - c\})^2 \ ,$$

where c is an additional additive constant to be estimated; here c could be considered part of the model bing fitted, as we suggest above; or alternatively, we could interpret the proximities as being translated, i.e., $p_{ij} + c$ is fit by $|x_j - x_i|$.

Once an appropriate order is obtained, the coordinate estimation is immediate. (The MATLAB routines we use are called `order.m` (to generate an appropriate object order), and `linfitac.m` (to estimate c and the coordinates based on the found order.)

Circular (CUS):

Find a set of coordinates, $x_1, \ldots, x_n$, and an $(n+1)$st value, x_0 (the circumference of a circular structure), $x_0 \geq |x_j - x_i|$ for all $1 \leq i \neq j \leq n$, minimizing

$$\sum_{i<j}(p_{ij} - [\min\{|x_j - x_i|, x_0 - |x_j - x_i|\} - c])^2 ,$$

where c is again an additive constant to be estimated. (The MATLAB routine we use is called `unicirac.m`.)

SCALING A MATRIX TO BE IN ANTI-ROBINSON FORM

Given the proximity matrix $\mathbf{P}$, find a new matrix $\mathbf{P}^*$ that is "close" to $\mathbf{P}$ *and* the entries in $\mathbf{P}^*$ satisfy the anti-Robinson property: there is some reordering of the rows and columns of $\mathbf{P}^*$ so that the entries within each row and within each column never decrease in moving away from the main diagonal.

In other words, we have a regular gradient present both within the rows and within the columns.

The $n(n-1)/2$ subsets defined by the choice of endpoints for an interval in the given order used to demonstrate the anti-Robinson form, along with their subset diameters (as measures of salience), can be used to "explain" the gradient (at least hopefully).