Overview

- Cluster analysis discussion
- Baayen 5
- Reading Qs
- Install.Packages():
  - languageR
  - ape
- Out to a worked example...
Reading questions

- [JR] Principal Components Analysis. “is PCA useful for sociolinguistic/sociophonetic data, and if so, has it been used?”


- [MO] EFCA. Varimax vs. Promax. “...The varimax rotation builds on the assumption that the rotated factors are uncorrelated. It is preferentially used when we are interested primarily in the generalizability of the results. The promax rotation allows the factors to be correlated, and tends to be selected when the primary concern is to obtain a factor model that provides a close fit to the data.” I generally understand this, but I think it might help me to discuss when we might be interested in generalizability versus a closer fit to the data.”

- [ABW] p. 128, in describing choice not to include more factors, indicates rationale is that p > 0.05. Why? *greater than?*

- [NK] Support Vector Machines. “…support vectors are chosen to be the data points that are exactly the 'margin distance' away from the hyperplane. It seems to me that it would be very difficult or unlikely to find a plane of any dimensionality such that more than two points in the data set are exactly the same distance from it. Does Baayen mean that the support vectors are within the margin distance, or am I thinking about this wrong?”
Clustering vs. Classification

- **Multivariate data**: datasets with more than two vectors.
- **Rows** = cases
- **Columns** = factors (properties)
- Two approaches for finding structure in multivariate datasets:
  - classification: groupings are known; supervised
  - clustering: groupings are unknown; unsupervised
Goals: Clustering

• Supervised methods are used in hypothesis-testing. We know (or sampled!) groups, we want to test ideas about between- or within-group differences, patterns of use, etc.

• Unsupervised methods are used in hypothesis-generation. We have a set of linguistic forms, entities, etc. and want to know: 1) what properties make speaker-listeners judge them to be similar, different, groupable, etc.), or 2) what accounts for the patterns subjects find in our data.

• Assists in pattern-detection in noisy data
Goals, cont.

• **Data reduction:** We have characterized (measured) the data on a number of variables. Which of these are important for perception of difference? for grammaticality? etc.? Which are not?

• We reduce an n-dimensional space to a reduced-dimensional space
Cautions

• Subjectivity always a part of interpreting clustering output

• Clustering will be sensitive to the level of measurement of the data (if measures are too coarse, or too few variables are used to reflect properties of the data, patterns less likely to emerge)
Choosing the # of Factors

• Theory-driven determination

• Rule of thumb is to begin with 1 component fewer than the number of factors represented in the dataset

• a component is important if it accounts for at least 5% of the variance

• Use external validation: when a partition has captured the structure in the data, this partition should be stable with regard to perturbation of the data.
Choosing a Similarity Measure

- Distance or Proximity (e.g., Euclidean distances) - interval or ratio scaled data
- Matching type - nominal data (described in terms of attributes yielding a profile)
Clustering Techniques

- Principal Components Analysis (PCA)
- Exploratory Factor Analysis (EFCA)
- Multidimensional Scaling (MDS)
- Correspondence Analysis
- Hierarchical Cluster Analysis (HCA)
- Basic idea: there is structure in some dataset; we want to know how many dimensions are SUFFICIENT for describing this structure. Reduce our space to include fewest dimensions necessary for describing the data, rotate our space to redefine each datapoint in terms of these dimensions, ordered by how much variability they account for.

- Data: Numerical vectors (use subsetting to omit other vector types)
PCA cont.

- **Single-linkage**: (Distance b/w 2 clusters is the shortest distance between a point in the cluster and an outside, closeby point or second cluster). Find 2 datapoints with smallest distance. Cluster. A third point joins the already-formed cluster of two if the minimum distance to any of the members of the cluster is smaller than the distance between the two closest unclustered points. Otherwise, the two closest unclustered points are placed in a cluster. The process continues until all points end up in one cluster.

- **Complete-linkage**: Again, cluster points with smallest distance. A third point joins the already formed cluster if the maximum distance to any of the members of the cluster is smaller than the distance between the two closest unclustered points. In other words, the distance between two clusters is the longest distance from a point in the first cluster to a point in the second cluster.

- **Average-linkage**: the distance between two clusters is the average distance from points in the first cluster to points in the second cluster.
• Ward’s method: (ANOVA logic) Find two points with the minimum within groups sum of squares. Points continue to be joined to the first cluster or to other points depending on which combination minimizes the error sum of squares from the group centroid. This method is also known as a k-means approach. The k-means method assigns the case to the closest centroid. The approach may take two forms, the most common of which is this:

K-means
1. All observations are considered as one set. The group is split based on the one variable which makes the greatest contribution to within-group sum of squares. This point becomes the centroid of its own group.

2. Group centroids are computed and distances b/w points and all group centroids are computed. The point that would best improve the objective is re-assigned. Repeat until no further improvement is reached.

3. The group with the largest within-groups sum of squares is selected for splitting. Steps 2 and 3 are then repeated until the desired number of clusters is identified.

source: Qualtronics.com
Method (PCA)

1. Select matrix or dataframe with numerical vectors only
2. Create a principal components object
3. Explore the components created by the algorithm
4. Determine which components are relevant to our analysis
5. Output includes: stdev, proportions of variance, cumulative proportions
6. Visualize (splom(), plot(), plclust(), pltree())
EFCA  

- Extension of PCA

- Basic idea: In PCA, you partition the total variation space among the components. This means the variance explained by a PC is given by that PC’s variance divided by the summed variances of all the PC. In EFCA, we include an error term to allow for noise in the data.

- Data: Numerical vectors (use subsetting to omit other vector types)
Method (EFCA)

1. Select matrix or dataframe with numerical vectors only
2. Create a principal components object, choosing rotation method (varimax, promax)
3. Explore the components created by the algorithm
4. Determine which components are relevant
5. Output includes: uniquenesses, loadings, tests of significance
6. Visualize (splom(), plot(), plclust(), pltree())
Correspondence Analysis  
corres.fnc()

- Special case of MDS
- Basic idea: we have datapoints, and know they may be characterized as distances from other datapoints on one or more measures
- Data: Tables with counts
Method (CorA)

1. Generate the square matrix for both rows and cols:
   
   step 1: create a matrix for column-column distances

   step 2: create a second matrix for row-row distances

2. Prepare a scatterplot of distances

3. Output: Eigenvalue rates, tables displaying how the row distances relate to the column distances

7. Visualize (splom(), plot(), plclust(),
HCA \texttt{hclust()}, \texttt{diana()}

- Extension of PCA
- Basic idea: Family including 2 types of clustering methods. 1) \textit{agglomerative clustering}: begin with single points, group by similarities, then group similar clusters into increasingly larger clusters, 2) \textit{divisive clustering}: begin with a cluster of all datapoints; successively partition into smaller clusters
- Data: Numerical vectors (use subsetting to omit other vector types)
Method (HFCA)

1. Tabulate data (collect properties of the observations; measures or counts)

2. Compute a similarity/dissimilarity matrix on the basis of a user-defined similarity/dissimilarity metric

3. Compute a cluster structure on the basis of a user-defined clustering rule

4. Represent the cluster structure (dendrogram, etc.)

5. Post-hoc tests for significance may be applied.