Index of 7-L (revised)[1]

[1] Revision I: 2 typos corrected;
Revision II: results for full CCA for butterflies revised, and path diagram added.

Practical Information

Course schedule change:

- move presentations one week to Nov 21st,
- keep session on Nov 14th (not next week!) for course wrap-up, including second quiz (coming up sometime next week, hopefully),
- any further topics/presentations, anyone?

Today's lecture:

- follow-up on cluster analysis: more examples from Manly,
  - data examples: dogs, employment,

  and some practical considerations,
- canonical correlation analysis (Manly: Chapter 9(10); TF: Chapter 12),
  - data example: butterflies,
- multidimensional scaling → session on Nov 14th.

Linkage Methods in Cluster Analysis

Simple applied overview (from Minitab help):

Linkage methods

Used with Cluster Observations and Cluster Variables, they determine how the distance between two clusters is defined. Choosing one over another may not make an appreciable difference with your data. However, because the goal of cluster amalgamation is somewhat subjective, you may find different methods are more or less appropriate with your particular situation and data.

Method	Distance between two clusters is...	Reasons to use this method
Single	The minimum distance between an item in one cluster and an item in the other cluster. Also called the "nearest neighbor" method	Best suited for observations or variables that are clearly separated. When they lie close together, single linkage tends to identify long chain-like clusters that can have a relatively large distance separating items at either end of the chain.
Average	The mean distance between an item in one cluster and an item in the other cluster.	Whereas the single or complete linkage methods group clusters based upon single pair distances, average linkage uses a more central measure of location.
Centroid	The distance between the cluster centroids or means.	Like average linkage, this method is another averaging technique.
Complete	The maximum distance between an item in one cluster and an item in the other cluster. Also called "furthest neighbor" method.	Ensures that all items in a cluster are within a maximum distance and tends to produce clusters with similar diameters. The results can be sensitive to outliers.
Median	The median distance between an item in one cluster and an item in the other cluster.	Similar to average or centroid method, though it reduces the effect of outliers.
McQuitty's	The average of the distances of the soon to be joined clusters to that other cluster. For example, if clusters 1 and 3 are to be joined into a new cluster, say 1*, then the distance from 1* to cluster 4 is the average of the distances from 1 to 4 and 3 to 4. Also called "weighted average linkage."	Here, distance depends on a combination of clusters rather than individual items in the clusters. Similar to average linkage, but the size of the clusters are assumed equal, so the pairwise distances are weighted accordingly.
Ward's	A function of the linkage criteria: the sum of squared deviations from points to centroids, minimizing the within-cluster sum of squares.	Tends to produce clusters with similar numbers of items, but it is sensitive to outliers.

CA: Practical Considerations

Should we standardize variables?[2]

+ removes scale effects, and makes variables comparable,

− makes distances more difficult to interpret.

The dendrogram: what's on the y-axis?

i) the *distance* (or *dissimilarity*) between clusters (according to chosen distance measure and linkage method),

- horizontal lines $\sim$ clusters joined at that distance,
- no reconstruction of individual distances, only cluster distances,[3]
- large vertical distances $\sim$ relatively strong separation between clusters; short distances $\sim$ similar groups/clusters,

ii) the *similarity* between clusters, defined for a pair (i, j) of clusters at distance $d(i, j)$, as (Minitab):

$$s(i, j) = 1 - \frac{d(i,j)}{d_{\max}}, \quad \text{expressed in } \%,$$

where $d_{\max}$ is the maximal distance among all observations (not clusters).[4]

What about variable types?

- Euclidean distance most natural for quantitative data,
- many other distance measures exist (e.g. binary/ordinal data).

[2] One may also consider to standardize observations if all variables are measured on the same scale..

[3] Some related diagrams (e.g., phylogram, phylogenetic tree) represent actual distances.

[4] Note that this normalization does not make real sense for all linkages (e.g., Ward) and may produce negative similarities.

CA Examples: Dogs, Employment

Summary (dogs): 6 jaw measurements for each of 7 dog-related species; interest is in closeness between species; we use Euclidean distance for standardized variables (5L–3):[5]

- smallest distance between prehistoric and modern dog $\rightarrow$ first cluster to be formed (any linkage),
- clusters similar across linkages, but not cluster distances!,
- similarities relative to $d_{\max} = 7.12$.

Summary (employment): 9 percentages on 30 countries; interest in exploring how clusters relate to country grouping (EU, EFTA, Eastern, Other); we consider 3rd ed. data and use standardized variables:

- smallest distance between Denmark and Sweden (makes sense!), most separated country is Albania[6],
- different clusters produced by linkage methods; single (nearest neighbour) may not be the best one,
- sample results for complete linkage, similarity $\geq 50\%$:
  - single clusters: **1)** Albania, **2)** Gibraltar,
  - small clusters: **3)** Czech/Slovac & Hungary; **4)** Greece, Poland & Turkey; **5)** Bulgaria, Russia, Romania, Yugoslavia,
  - all remaining (19) countries in one cluster.

[5] Note that distances in Manly's book are slightly off, possibly to round-off errors.
[6] Largest individual distance is $d_{\max} = 9.82$ (Albania, Gibraltar).

Canonical Correlation Analysis: the Idea

CCA is a mathematical[7] procedure to represent the correlation between two sets of (correlated) variables: $(X_1, \ldots, X_p)$ and $(Y_1, \ldots, Y_q)$ (where $q \leq p$) by two new sets of variables $(U_1, \ldots, U_q)$ and $(V_1, \ldots, V_q)$ such that:

- each U_i is a linear combination of the X_j's, and each V_i is a linear combination of the Y_j's,
- the variables $U_1, \ldots, U_q$ are independent (uncorrelated/orthogonal), and the same is true for the variables $V_1, \ldots, V_q$,
- the variables U_1 and V_1 have maximal correlation; the variables U_2 and V_2 have maximal correlation subject to their independence on U_1 and V_1, respectively; and so forth.

First impressions/interpretations:

* the pairs (U_1, V_1), (U_2, V_2),...,(U_q, V_q) are unique (up to sign change) when standardized; these are called *canonical variates*, and $\sim$ independent "directions" among the (X_j) and (Y_j) with maximal correlation, and
  - may have useful subject-matter interpretations,
  - may represent useful data reductions,
* some similarity with PCA, but here we maximize correlation instead of variance,
* an extension of multiple regression to multiple Y_j's, because with only Y_1 the predictions $X\hat{\beta}$ have maximal correlation with Y_1,
* although suggested by the notation, there is no direction $X \to Y$ (and of course no claim of causality).

[7] CCA is not in itself based on a statistical model, but inference about the number of signficant components is.

CCA Example: Butterflies

Data: 4 environmental and 6 genetic variables (Pgi gene frequencies) for 16 colonies of a butterfly in California and Oregon; interest is in describing relation between the genetic and environmental variables,

- complete collinearity among genetic variables ⇒ one variable must be dropped (genetic type 1.30),
- gene types 0.40 and 0.60 combined (due to low frequencies),

⇒ 4 X-variables (environmental) and 4 Y-variables (genetic).

(For illustration only) Multiple linear regression for `freq_pgi08`:

- $R^2 = 0.415$ equals the (first) canonical correlation squared,
- $F = 1.95$ test for overall significance equals test for (first) canonical variate.

Results of full CCA:[8]

- canon. corr.: $0.862, 0.450, 0.386, 0.089$, no overall significance: ⇒ focus here on first component,

U_1	altitude	precip	temp_max	temp_min
stand. coef.	0.124	0.293	-0.468	-0.260
loading/corr.	0.922	0.771	-0.898	-0.919

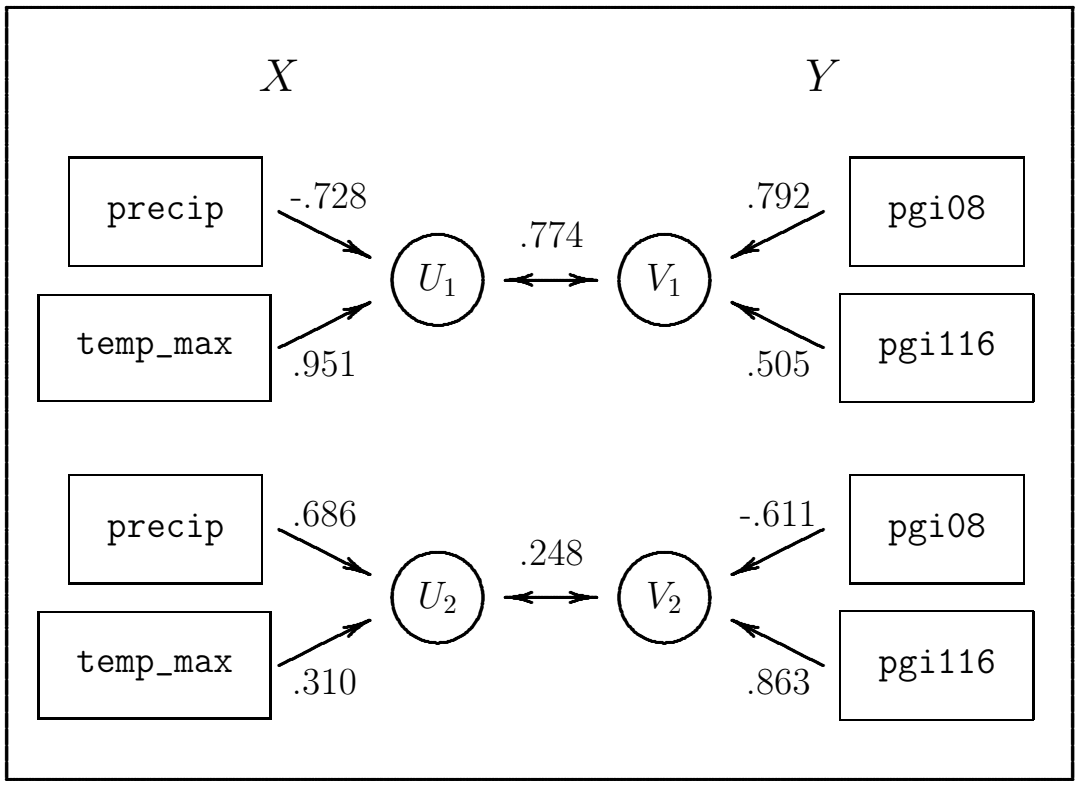
V_1	pgi046	pgi08	pgi10	pgi116
stand. coef.	-0.548	-0.422	0.089	-0.826
loading/corr.	-0.384	-0.740	0.961	-0.475

- some disagreement between patterns in coef. and loadings/correlations: $U_1 \sim$ contrast altitude & precip vs. temperatures, $V_1 \sim$ mostly a contrast between `freq_pgi10` and others,
- plot of scores for V_1 vs. U_1 shows one outlying observation.

[8] Results agree roughly with Manly; same results in Stata and SAS.

CCA Path Diagram: Butterflies

Representation of loadings for simple CCA with 2 envir. variables (`precip`,`temp_max`) and 2 genetic variables (`freq_pgi08`,`freq_pgi116`):



* eigenvalues: 0.60 ($= .774^2$) and 0.06 ($= .248^2$), also interpretable as proportions of overlapping variance between X and Y for the two pairs (U_1, V_1) and (U_2, V_2),
* among X, the canon. var. U_1 explains $(.728^2 + .951^2)/2 = .717$ and U_2 explains $(.686^2 + .310^2)/2 = .283$ of the variance,
* among Y, the canon. var. V_1 explains $(.792^2 + .505^2)/2 = .441$ and V_2 explains $(.611^2 + .863^2)/2 = .559$ of the variance,
* redundancy: among Y, the canon. var. U_1 explains $.717 \times .774^2 = .43$ and U_2 explains $.283 \times .248^2 = .02$ of the variance;
  similarly, among X, the canon. var. V_1 explains $.441 \times .774^2 = .26$ and V_2 explains $.559 \times .248^2 = .003$ of the variance.

CCA: the Steps

Given a full covariance matrix $\mathbf{S}$ $(p+q) \times (p+q)$ for the combined set $(X_1, \ldots, X_p, Y_1, \ldots, Y_q)$ partitioned (split) into submatrices,[9]

$$\mathbf{S} = \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^t & \mathbf{B} \end{pmatrix},$$

- let $\mathbf{R} = \mathbf{B}^{-1}\mathbf{C}^t\mathbf{A}^{-1}\mathbf{C}$ $(q \times q)$,
- the *eigenvalues* of $\mathbf{R}$ (proportions of variance explained) are the squared canonical correlations $r_1^2, \ldots, r_q^2$ (where $r_i = \mathrm{Corr}(U_i, V_i)$),
- the ith eigenvector (say $b^{(i)}$) gives the coefficients for V_i on $(Y_1, \ldots, Y_q)$: $V_i = b_1^{(i)}Y_1 + \ldots + b_q^{(i)}Y_q$,
- U_i coefficients: $U_i = a_1^{(i)}X_1 + \ldots + a_p^{(i)}X_p$ with $a^{(i)} = \mathbf{A}^{-1}\mathbf{C}\, b^{(i)}$,
- assuming MVN for $(X_1, \ldots, X_p, Y_1, \ldots, Y_q)$, tests can be computed for all canon. corr./components.[10]

Interpretation of components — two suggestions (described here for the first component pair (U_1, V_1)), look at:

- loadings of X_j's on U_1 and of Y_j's on V_1,[11]
- correlations between X_j's and U_1 and between Y_j's and V_1,

— both may be problematic if X_j's or Y_j's are highly collinear (Manly).

Rotation of components is possible: Stata offers varimax rotation.

[9] Here $\mathbf{A}$ $(p \times p)$ and $\mathbf{B}$ $(q \times q)$ are the covariance matrices for X_j's and Y_j's, respectively, and $\mathbf{C}$ $(p \times q)$ is the matrix of correlations between X_j's and Y_j's.

[10] Manly mentions Bartlett's (χ^2-) test for all components; Stata uses F-tests. Manly also dismisses tests for some components only as " not reliable", but Stata will gladly compute them ...

[11] Squared loadings averaged across the X_j's (or Y_j's) give the proportion of variance explained by corresponding variate; further multiplication with the squared correlations give the *redundancies*: proportions of variance explained for the other set of variables; details in TF.

CCA: Limitations and Practical Issues[12]

Many issues similar to PCA — with some adaptations:

- normality is required for validity of test statistics, but CCA can be meaningfully applied to any variables for which variance and correlation make sense,
- assumed linearity enters in two ways: correlations measure linear relationships only, and by the linearly constructed variates,
- generally quite sensitive to minor changes in the data.

Interpretation of components and relations not straightforward or automatic:

- the canonical variates maximize correlation, but not necessarily interpretability; rotations are less common (and documented) than in factor analysis,
- results depend on *both sets* of variables: changing one set of variables will affect both sets of canonical variates — may be undesired or at least requires caution,
- highly collinear variables (in one or both sets) may make the canonical variates difficult to interpret (Manly); therefore the correlations may be better for interpretations.

[12] Based largely on TF: Section 12.3.2.