**Canonical Correlation**

1. Description
   1. PCA vs CanCor
      1. PCA: derive a linear combination of variables that explains the maximum variance within a single set of variables
      2. CanCor: derive two linear functions that explain the maximum correlation between two sets of variables
   2. condenses the number of variables with minimum loss of information and reducing redundancy
   3. reveal the common structure of two multivariate datasets
   4. multivariate analysis of correlation
   5. Canonical correlation is the general form of any type of regression
2. Data
   1. two sets of variables, each set is consistently all continuous, categorical, or count variables
   2. N = number of samples (entities)
   3. P = number of variables in group A
   4. M = number of variables in group B
   5. each sample entity measured for all variables
   6. sample size
      1. N > P + M at least
      2. N > 100 + (P + M)2
      3. N > 10(P + M) + 100
      4. N > 3(P + M)
   7. P does not have to equal M
   8. don’t use data that are unimportant
3. Assumptions
   1. multivariate normality: not required but relationships are more realistic and interpretable if you have it
   2. data matrix must be non-singular (no variable is an exact linear combination of another variable in the dataset) and not exhibit multicollinearity. Solutions: PC scores or remove the “offending” variable(s)
   3. independent random sample
   4. outliers can be a problem
   5. linearity: it’s a good idea
4. Central results
   1. canonical correlations
   2. canonical coefficients
   3. Wilke’s lambda (U value): test the significance of the first canonical correlation coefficient