Contents

Preliminaries
0: Introduction
1: Graphical Displays
2: Reduction of Dimensionality
3: Multidimensional Scaling Techniques
4: Discriminant Analysis
5: Multivariate Regression Analysis
6: Canonical Correlation Analysis
7: Partial Least Squares
8: Statistical Analysis of Multivariate Data
9: Statistical Discriminant Analysis

Relationships Between Variables

- **PCA:–**
  - *unsupervised* technique for investigating structure in a single set of observations
    - only *indirect* information on variables
    - e.g. interpretation of loadings of PCs
      - (contrast between *size* & *shape* variables in BM Mummy Pots example)

- **LDA**
  - *supervised* technique for investigating relationship between one categorical variable & others
    - interest in ***predicting*** the category from others
    - also interest in ***relationship*** between categories and other variables
      - which variables distinguish between groups?
      - examine discriminant functions (*crimcoords*) loadings
- **Note:–**
  - for LDA must have n > p
  - but for PCA can have n > p or n < p
    - depends whether matrix inversion is involved
      - (correlation/covariance)

- Prediction or relationship?
  - n > p or n < p?
- Compare univariate regression of Y on X
  - (Y a random variable, X a fixed variate)
  - $E[y_i] = \alpha + \beta x_i$
    - used for modelling the ***dependence*** of Y upon X
    - ⇒ used for predicting value of y for a [new] value x
    - If both X and Y are random variables then use the same model ***conditional*** on value of X
    - model extends to p X-variables (***provided n > p***) (& is essentially identical to LDA if Y is categorical)
  - Model for predicting **X** from **Y** is different
  - If interest is in relationship between then we examine the ***correlation*** between X & Y

- Prediction or relationship?
  - n > p or n < p?
- Collection of techniques available
  - Multivariate Regression
    - **predicting** a multivariate a multivariate random variable Y from a multivariate variable X
      - Y ***dependent***, X ***independent***
        - requires "n > p"
  - Canonical Correlation
      - **relationship** between two multivariate R.V.s X & Y
        - requires "n > p"
  - Partial Least Squares
    - prediction or relationship, does not require "n > p"

- **Multivariate Regression**
  - Essentially identical to univariate regression
    - model $E[Y_i] = X\beta$,
      $\hat{\beta} = (X'X)^{-1}X'Y, \quad \Sigma = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n - q - 1)$
    - Confidence regions for β obtained from
      $(\beta - \hat{\beta})'\hat{\Sigma}^{-1}(\beta - \hat{\beta}) \sim \chi^2$ under H_0: $\beta = 0$
      - can get individual coefficients from univariate regressions
      - but not the simultaneous confidence regions as easily
    - (see lecture notes for an example)

Canonical Correlation

- Two variables X and Y of dimensions p & q
  - aim is to find a ***linear*** combinations of the X & Y variables, a′X & b′Y, which have maximum correlation
  - maximize $\rho_{XY} = a'\Sigma_{XY}b/\sqrt{\{a'\Sigma_{XX}ab'\Sigma_{YY}b\}}$
    - solve by The Same Procedure as Last Time
    - converting to an eigenanalysis problem
      - (constraining the numerator = 1)
  - obtain sets of *canonical variates*
    - used to display data
    - examination of loadings allows interpretation of X & Y relationship
      - identical to *crimcoords* if Y is categorical

Example:

- X = (head length, head width [of 1st sons])
- Y = (head length, head width [of 2nd sons])
  - a_1 = (0.70, 0.72)′, b_1 = (0.74, 0.67)′ , is first canonical variate with correlation ρ_1 = 0.79
    - both of these reflect ***size***
  - a_2 = (0.71, −0.71)′, b_1 = (0.71, −0.71)′ is 2nd canonical variate with correlation ρ_2 = 0.05
    - both of these reflect ***shape***
- Interpretation:–
  - greatest similarity between brothers is in head size but there is little relationship between their head shapes

- No general reason why canonical variates in CCA should reflect *size*, even if variables are all linear measurements of dimension
  - unlike PCA and the Perron-Frobenius theorem
- see notes for more complex example relating measures of depression & health to socio-demographic measures
- other common uses: questionnaire analysis
  - evaluation of a product and measures about people
    - socio-demographics:- age, income, postcode grading
    - opinions of hair-spray:- lustre, bounce, glossiness, ….
  - allows insight into *market segmentation* revealed by plot on canonical variates
    - cf plots on PCs revealing groupings

Partial Least Squares

» (for n < p and n << p problems)

- Comments (see also lecture notes)
  - key idea is to replace correlation matrices by covariance matrices so ‘avoiding’ inversion of singular variance matrices
    - again look for linear combinations of variables
    - derivation is algorithmic / iterative
      - (not a solution of an eigenanalysis problem???)
  - implementation easy in R
    - using **`plsgenomics`** library and routines **`pls.regression(.)`** & **`pls.lda(.)`**
  - see recent MSc dissertations by Maria Taboada, Rich Jacques, Lu Zou, James Bentham, ….

Other techniques for n < p discrimination

- need some method of preliminary dimensionality reduction
- take first k principal components (k < n)
  - rationale:-
    - some PCs may be associated with components of variance attributable to differences between groups
  - seems to be useful if p is ‘medium range’ (<< ~100)
  - doesn’t work well if n << p (e.g. n ~ 50⁻, p ~ 1000⁺)

Other techniques for n < p discrimination

- instead of ranking PCs by variances on the PCs
  - i.e. by the eigenvalues

  rank them by some other criterion reflecting group differences such as $a_i'Ba_i/\lambda_i$
    - B the ***between groups*** covariance matrix
  - this (and also ordinary PCs) are particular cases of the **Karhunen-Loève** transformations
  - these give interpretation of loadings etc, i.e. some form of ***analysis*** of differences
- see also techniques of ***feature extraction***
  - in Pattern Recognition areas of Machine Intelligence literature
    - typically no *analysis* by examination of loadings