Canonical Correlation Analysis I

Mengqian LU

Introduction

- Linear Regression: $Y = \beta \mathbf{X}$
  - Maximize R^2 to find a linear combination of $\mathbf{X}$ that explains the most variance in Y
  - $\mathbf{X}$ can be observed variables, or PCs from PCA, or Factors from FA etc.
  - Use PCs or Factors, we have reduced dimensionality and independence of predictors, which solves issues of collinearity. (Conventional FA with orthogonal rotation)
- What if now Y → $\mathbf{Y}$? Multiple dependent variables
  - We need a way of measuring the linear relationship between two sets of variables → Canonical Correlation Analysis
  - An extension of multiple regression, to find linear combinations of both sets, that maximizes their correlations
  - Extraction of coefficients is similar to the process of finding PCs

Canonical Correlation Analysis

- Canonical Correlation Analysis extends the analysis on "Y~**X**" to 2+ continuous IVs with 2+ continuous DVs
- Key objective is to answer "How are the best linear combinations of predictors related to the best linear combinations of the DVs?"
  1. Linear combination of predictors
  2. Linear combination of predictants
  3. The relationship between the two linear combinations
  4. … more (coming up)

PCA finds the best combination in terms of maximum variance

- CCA is a multidimensional exploratory statistical method in the same vein as PCA, however CCA aims at the exploration of sample correlations between two sets of quantitative variables observed on the same experimental units, while PCA focus on dimension reduction of one data set through linear combination of variables.

When to use?

- Two sets of variables are related – at exploratory stage
- Your modeling assumptions rely on the theory that the predictor set as a whole inform the predictant set as a whole
- One set of two or more variables related longitudinally across two time points
- Examples: (1) An psychologist measure a number of aptitude variables and a number of achievement variables on some students and study the relation between "aptitude" and "achievement"; (2) An researcher study the relation between p measurements related to the yield of plants (e.g., height, dry weight, number of leaves) at each of n sites in a region and q variables related to the weather conditions at these sites. (e.g., average daily rainfall, humidity, hours of sunshine)

Formulation

- One pair of bases – the ones correspond to the largest CC
  - Consider the linear combinations $x = \boldsymbol{x}^T\widehat{w}_x$ and $y = \mathbf{y}^T\widehat{w}_y$ of the two variables respectively, this means we need to maximize:

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\hat{\mathbf{w}}_x^T\mathbf{x}\mathbf{y}^T\hat{\mathbf{w}}_y]}{\sqrt{E[\hat{\mathbf{w}}_x^T\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}}_x]E[\hat{\mathbf{w}}_y^T\mathbf{y}\mathbf{y}^T\hat{\mathbf{w}}_y]}}$$

$$= \frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y}}.$$

  - The maximum of ρ w.r.t w_x and w_y is the maximum canonical correlation. The subsequent CCs are uncorrelated for different solutions, i.e.,

$$\begin{cases} E[x_ix_j] &= E[\mathbf{w}_{xi}^T\mathbf{x}\mathbf{x}^T\mathbf{w}_{xj}] = \mathbf{w}_{xi}^T\mathbf{C}_{xx}\mathbf{w}_{xj} = 0 \\ E[y_iy_j] &= E[\mathbf{w}_{yi}^T\mathbf{y}\mathbf{y}^T\mathbf{w}_{yj}] = \mathbf{w}_{yi}^T\mathbf{C}_{yy}\mathbf{w}_{yj} = 0 \\ E[x_iy_j] &= E[\mathbf{w}_{xi}^T\mathbf{x}\mathbf{y}^T\mathbf{w}_{yj}] = \mathbf{w}_{xi}^T\mathbf{C}_{xy}\mathbf{w}_{yj} = 0 \end{cases} \quad \text{for} \quad i \neq j.$$

  - The projections onto w_x and w_y, i.e. x and y, are the canonical variates

Layers of Analysis

1. Correlation between pairs of canonical variates
2. Loadings between predictors (IVs) and their canonical variates
3. Loadings between predictants (DVs) and their canonical variates
4. Adequacy
5. Communalities
6. Redundancy – between IVs and their canonical variates or between DVs and their canonical variates

Formulation (Cont'd)

- Let's create some single variable F that represent ***x*** and another single variable G for **y**, and F and G are linear combinations of X and Y, respectively: $F = f_1X_1 + f_2X_2 + \ldots + f_pX_p$; $G = g_1Y_1 + g_2Y_2 + \ldots + g_pY_p$.
- The first canonical correlation is the maximum correlation coefficient between F and G, for all F and G
- The correlation of our interest are now those between the canonical variates (linear combinations of **X** and **Y**)
- Terms for CCA: variables → (pairs of) canonical variates (both predictors and dependent variables) → canonical correlation
  - The maximum number of canonical variate pairs equals to the number of variables in the smaller set
  - There could be many canonical correlations, arranged in descending order, most of the time, only the first one or two are of the interest.