Canonical Correlation

Psy 524

Andrew Ainsworth

Matrices

Summaries and reconfiguration

Trace

- sum of diagonal elements

$$B = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 6 & 1 & 0 \\ 2 & 8 & 7 \\ 2 & 4 & 5 \end{pmatrix}$$

$$Trace = 6 + 8 + 5 = 19$$

Trace

- If the matrix is an SSCP matrix then the trace is the sum-of-squares
- If the matrix is the variance/covariance matrix than the trace is simply the sum of variances.
- If it is a correlation matrix the trace is just the number of variables.

Determinant

$$D = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

$$|D| = a_{11}a_{22} - a_{12}a_{21}$$

- this is considered the generalized variance of a matrix. Usually signified by | | (e.g. |A|)
- For a 2 X 2 matrix the determinate is simply the product of the main diagonal – the product of the other diagonal

Determinant

- For any other matrices the calculations are best left to computer
- If a determinate of a matrix equals 0 than that matrix cannot inverted, since the inversion process requires division by the determinate. What is a common cause of determinates equaling zero?

Eigenvalues and Eigenvectors

- this is a way of rearranging and consolidating the variance in a matrix.

$$\underset{MxM}{D}\ \underset{Mx1}{V} = \underset{1x1}{\lambda}\ \underset{Mx1}{V}$$

$$D = any\ square\ matrix$$

$$V\ =\ Eigenvector$$

$$\lambda =\ Eigenvalue$$

Eigenvalues and Eigenvectors

- Think of it as taking a matrix and allowing it to be represented by a scalar and a vector (actually a few scalars and vectors, because there is usually more than one solution).

Eigenvalues and Eigenvectors

- Another way to look at this is:

$$(D - \lambda I)V = 0$$

$$\left[\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right]\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$\left[\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right]\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

Eigenvalues and Eigenvectors

If v1 and v2 equal zero the above statement is true, but boring. A non-boring solution comes when the determinate of the leftmost matrix is 0.

$$(a-\lambda)(d-\lambda)-bc=0$$

$$\lambda^2-(a+d)\lambda+ad-bc=0$$

Generalize it to $x\lambda^2-y\lambda+z=0$

To solve for λ apply:

$$\lambda=\frac{-y\pm\sqrt{y^2-4xy}}{2x}$$

Eigenvalues and Eigenvectors

$$D = \begin{bmatrix} 5 & 1 \\ 4 & 2 \end{bmatrix}$$

$$\lambda^2 - (5+2)\lambda + 5*2 - 1*4 = 0$$

$$\lambda^2 - 7\lambda + 6 = 0$$

$$\lambda = \frac{-7 + \sqrt{7^2 - 4*1*6}}{2*1} = 6$$

$$\lambda = \frac{-7 - \sqrt{7^2 - 4*1*6}}{2*1} = 1$$

$$\lambda_1 = 6, \lambda_2 = 1$$

Eigenvalues and Eigenvectors

- Using the first eigenvalue we solve for its corresponding eigenvector

$$\begin{bmatrix} 5-6 & 1 \\ 4 & 2-6 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} -1 & 1 \\ 4 & -4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

This gives you two equations:

$$-1v_1 + 1v_2 = 0$$

$$4v_1 - 4v_2 = 0$$

$$V_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Eigenvalues and Eigenvectors

- Using the second eigenvalue we solve for its corresponding eigenvector

$$\begin{bmatrix} 5-1 & 1 \\ 4 & 2-1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} 4 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

This gives you two equations:

$$4v_1 + 1v_2 = 0$$

$$4v_1 + 1v_2 = 0$$

$$V_2 = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$$

Eigenvalues and Eigenvectors

- Let's show that the original equation holds

$$\begin{bmatrix} 5 & 1 \\ 4 & 2 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \end{bmatrix} \text{ and } 6 * \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$

$$\begin{bmatrix} 5 & 1 \\ 4 & 2 \end{bmatrix} * \begin{bmatrix} -1 \\ 4 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix} \text{ and } 1 * \begin{bmatrix} -1 \\ 4 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$$

Canonical Correlation

Canonical Correlation

- measuring the relationship between two separate sets of variables.
- This is also considered multivariate multiple regression (MMR)

Canonical Correlation

- Often called Set correlation
  - Set 1 $\left(y_1, \ldots, y_p\right)$
  - Set 2 $\left(x_1, \ldots, x_q\right)$
    - p doesn't have to equal q
- Number of cases required ≈ 10 per variable in the social sciences where typical reliability is .80, if higher reliability than less subjects per variable.

Canonical Correlation

- In general, CanCorr is a method that basically does multiple regression on both sides of the equation

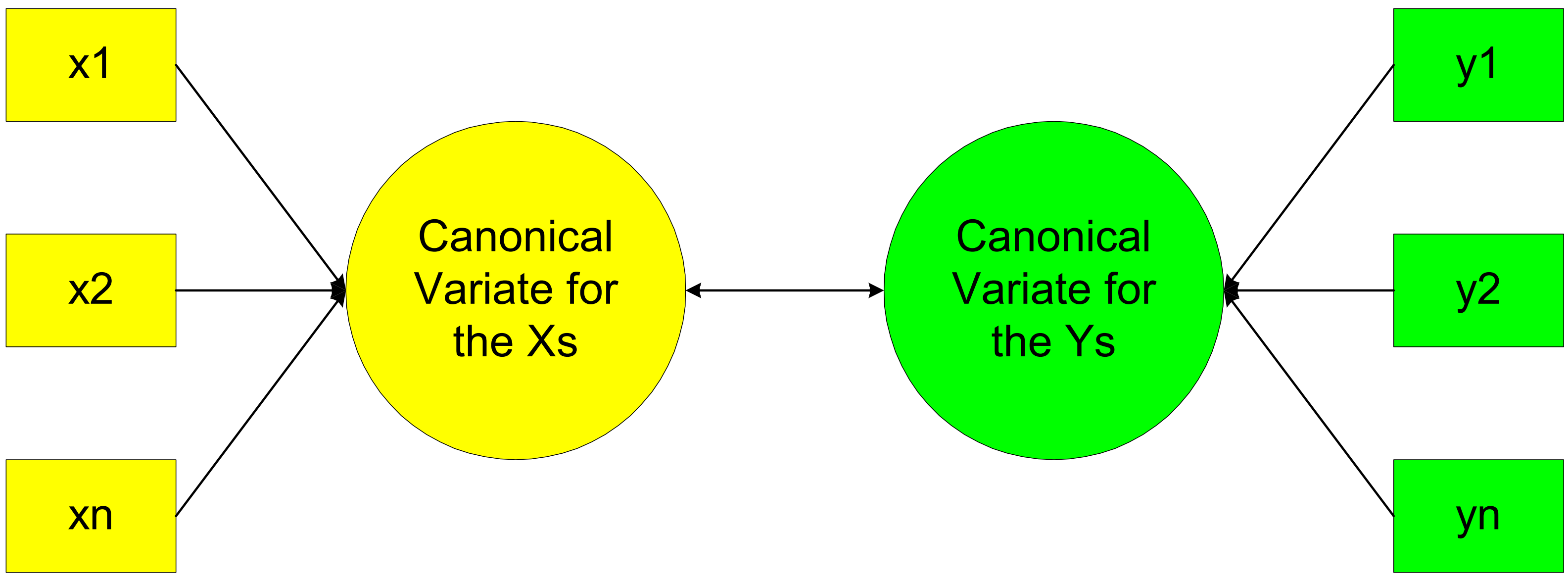
$$\psi_1 y_1 + \psi_2 y_2 + \psi_n y_n = \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n$$

- this isn't really what happens but you can think of this way in general.

Canonical Correlation

- A better way to think about it:
  - Creating some single variable that represents the Xs and another single variable that represents the Ys.
  - This could be by merely creating composites (e.g. sum or mean)
  - Or by creating linear combinations of variables based on shared variance

Canonical Correlation



- Make a note that the arrows are coming from the measured variables to the canonical variates.

Canonical Correlation

- In multiple regression the linear combinations of Xs we use to predict y is really a single canonical variate.

Jargon

- Variables
- Canonical Variates – linear combinations of variables
  - One CanV on the X side
  - One CanV on the Y side
- Canonical Variate Pair - The two CanVs taken together make up the pair of variates

Background

- Canonical Correlation is one of the most general multivariate forms – multiple regression, discriminate function analysis and MANOVA are all special cases of CanCorr

- Since it is essentially a correlational method it is considered mostly as a descriptive technique.

Background

- The number of canonical variate pairs you can have is equal to the number of variables in the smaller set.
- When you have many variables on both sides of the equation you end up with many canonical correlates. Because they are arranged in descending order, in most cases the first couple will be legitimate and the rest just garbage.

Questions

- How strongly does a set of variables relate to another set of variables? How strong is the canonical correlation?
- How strongly does a variable relate to its own canonical variate?
- How strongly does a variable relate to the other set's canonical variate?

Assumptions

- Multicollinearity/Singularity
  Check Set 1 and Set 2 separately
  - Run correlations and use the collinearity diagnostics function in regular multiple regression
- Outliers – Check for both univariate and multivariate outliers on both set 1 and set 2 separately

Assumptions

- Normality
  - Univariate – univariate normality is not explicitly required for MMR
  - Multivariate – multivariate normality is required and there is not way to test for except establishing univariate normality on all variables, even though this is still no guarantee.

Assumptions

- Linearity – linear relationship assumed for all variables in each set and also between sets
- Homoskedasticity – needs to be checked for all pairs of variables within and between sets.