# MS-E2112 Multivariate Statistical Analysis (5cr)
# Lecture 8: Canonical Correlation Analysis

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

# Contents

Canonical Correlation Analysis

Testing Independence

Scoring and Predicting

Words of Warning

References

Canonical Correlation Analysis
Testing Independence
Scoring and Predicting
Words of Warning
References

# Canonical Correlation Analysis

# Canonical Correlation Analysis



Canonical Correlation Analysis
Testing Independence
Scoring and Predicting
Words of Warning
References

Canonical correlation analysis involves partition of variables into two vectors $x$ and $y$. The aim is to find linear combinations $\alpha^T x$ and $\beta^T y$ that have the largest possible correlation.

# Canonical Correlation Analysis



Canonical Correlation Analysis
Testing Independence
Scoring and Predicting
Words of Warning
References

Let $x$ be a $p$-variate random vector and let $y$ be a $q$-variate random vector. The object in canonical correlation analysis is to find linear combinations

$$u_k = \alpha_k^T x$$

and

$$v_k = \beta_k^T y$$

that maximizes the correlation $|corr(u_k, v_k)|$ between $u_k$ and $v_k$ subject to

$$var(u_k) = var(v_k) = 1,$$

and

$$corr(u_k, u_t) = 0, corr(v_k, v_t) = 0, \ t < k.$$

# Correlation



Quick reminder:

$$corr(w_1, w_2) = \frac{E[(w_1 - \mu_{w_1})(w_2 - \mu_{w_2})]}{\sigma_{w_1}\sigma_{w_2}}.$$

# Canonical Correlation Analysis



The vectors $\alpha_k$ and $\beta_k$ are called the $k$th canonical vectors and

$$\rho_k = |corr(u_k, v_k)|$$

are called canonical correlations.

# Canonical Correlation Analysis



Whereas principal component analysis considers interrelationships within a set of variables, canonical correlation analysis considers relationships between two groups of variables.

# Canonical Correlation Analysis, Examples

- Exercise — health.
- Open book exams — closed book exams.
- Job satisfaction — performance.

# Canonical Correlation Analysis, Regression Analysis



Canonical Correlation Analysis
Testing Independence
Scoring and Predicting
Words of Warning
References

Canonical correlation analysis can be seen as an extension of multivariate regression analysis. However, note that in canonical correlation analysis there is no assumption of causal asymmetry - $x$ and $y$ are treated symmetrically!

# Canonical Correlation Analysis, Solution

Let $z = (x^T, y^T)^T$, and let

$$cov(z) = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Define

$$M_1 = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

and

$$M_2 = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

# Canonical Correlation Analysis, Solution



Now, the canonical vectors $\alpha_k$ are the eigenvectors of $M_1$ ($\alpha_k$ corresponds to the $k$th largest eigenvalue), the canonical vectors $\beta_k$ are the eigenvectors of $M_2$, and $\rho_k^2$ are the eigenvalues of the matrix $M_1$ (and of $M_2$ as well). The proof of this solution can be found from pages 283-284 of [1].

# Canonical Correlation Analysis, Solution



Note that the eigenvectors $\alpha_k$ and $\beta_k$ do not have length$= 1$!
Requirements

$$var(u_k) = var(\alpha_k^T x) = 1$$

and

$$var(v_k) = var(\beta_k^T y) = 1$$

define the lengths of the eigenvectors.

# Canonical Correlation Analysis, Solution

If the covariance matrices $\Sigma_{11}$ and $\Sigma_{22}$ are not full rank, similar results may be obtained using generalized inverses. One may also consider dimension reduction as a first step.

# Canonical Correlation Analysis, Sample Version



Sample estimates $\hat{\alpha}_k$, $\hat{\beta}_k$ and $\hat{\rho}_k$ of $\alpha_k$, $\beta_k$ and $\rho_k$, respectively, are obtained by using sample covariance matrices calculated from the samples $x_1, x_2, ..., x_n$, $y_1, y_2, ..., y_n$ and $z_1, z_2, ..., z_n$.

# Canonical Correlation Analysis, Standardization



Canonical Correlation Analysis
Testing Independence
Scoring and Predicting
Words of Warning
References

As in PCA, also in canonical correlation analysis, the data is sometimes standardized first.

# Testing Independence

# Testing Independence



Assume that $z = (x^T, y^T)^T \sim N_{p+q}(\mu, \Sigma)$. Consider testing

$H_0$ : $x$ and $y$ are independent,

against

$H_1$ : $x$ and $y$ are not independent.

# Testing Independence



Let $m = min\{p, q\}$, and let

$$T = -(n - \frac{1}{2}(p + q + 3)) \ln(\prod_{k=1}^{m}(1 - \hat{\rho}_k^2)).$$

Now, under $H_0$, and under the assumption of multivariate normality, the test statistic $T$ is asymptotically distributed as $\chi^2(pq)$.

# Testing Partial Independence



Assume that $z = (x^T, y^T)^T \sim N_{p+q}(\mu, \Sigma)$. Consider testing

$H_0$ : Only $s$ of the canonical correlation coefficients are nonzero,

against

$H_1$ : The number of nonzero canonical correlation coefficients is larger than $s$.

# Testing Partial Independence



Let $m = min\{p, q\}$, and let

$$T_s = -(n - \frac{1}{2}(p + q + 3)) \ln( \prod_{k=s+1}^{m} (1 - \hat{\rho}_k^2)).$$

Now, under $H_0$, and under the assumption of multivariate normality, the test statistic $T$ is asymptotically distributed as $\chi^2((p - s)(q - s))$.

# Independence Testing

If the normality assumption $z = (x^T, y^T)^T \sim N_{p+q}(\mu, \Sigma)$ does not hold, the $p$-values of the above mentioned test statistics can be approximated using permutations.

# Scoring and Predicting

# Scoring and Predicting



Let $X$ and $Y$ denote the $n \times p$ and $n \times q$ data matrices for $n$ individuals, and let $\hat{\alpha}_k$ and $\hat{\beta}_k$ denote the $k$th (sample) canonical vectors. Then the $n \times 1$ vectors

$$\eta_k = X\hat{\alpha}_k$$

and

$$\phi_k = Y\hat{\beta}_k$$

denote the scores of the $n$ individuals on the $k$th canonical correlation variables.

# Scoring and Predicting



If the $x$ and $y$ variables are interpreted as the "predictor" and "predicted" variables, respectively, then the $\eta_k$ score vector can be used to predict the $\phi_k$ score vector by using least square regression:

$$(\tilde{\phi_k})_i = \hat{\rho}_k((\eta_k)_i - \hat{\alpha}_k^T \bar{x}) + \hat{\beta}_k^T \bar{y}.$$

The canonical correlation $\hat{\rho}_k$ estimates the proportion of the variance of $\phi_k$ that is explained by the regression on $x$.

# Example

Example: closed book exams — open book exams.

# Example



Marks in open-book (O) and closed-book (C) exams:

| i | Mechanics (C) | Vectors (C) | Algebra (O) | Analysis (O) | Statistics (O) |
|---|---|---|---|---|---|
| 1 | 77 | 82 | 67 | 67 | 81 |
| 2 | 63 | 78 | 80 | 70 | 81 |
| 3 | 75 | 73 | 71 | 66 | 81 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 100 | 46 | 52 | 53 | 41 | 40 |

Source: K. V. Mardia, J. T. Tent, J. M. Bibby, Multivariate analysis, Academic Press, London, 2003 (reprint of 1979).

# Example



Means:

| Variable | Mean |
|---|---|
| $x_1$ | 38.9545 |
| $x_2$ | 50.5909 |
| $y_1$ | 50.6023 |
| $y_2$ | 46.6818 |
| $y_3$ | 42.3068 |

# Example

Covariance matrix

$$\Sigma = \begin{array}{c|cc|ccc|}
 & \multicolumn{2}{c}{\Sigma_{11}} & \multicolumn{3}{c}{\Sigma_{12}} \\
\hline
 & 302.3 & 125.8 & 100.4 & 105.1 & 116.1 \\
 & & 170.9 & 84.2 & 93.6 & 97.9 \\
\hline
 & & & 111.6 & 110.8 & 120.5 \\
 & & & & 217.9 & 153.8 \\
 & & & & & 294.4 \\
\hline
 & \multicolumn{2}{c}{\Sigma_{21}} & \multicolumn{3}{c}{\Sigma_{22}}
\end{array}$$

# Example

Calculate the eigenvectors

$$M_1 = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \Rightarrow \hat{\alpha}_k$$

and

$$M_2 = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \Rightarrow \hat{\beta}_k.$$

Here

$$\hat{\alpha}_1 = \begin{bmatrix} 0.0260 \\ 0.0518 \end{bmatrix}$$

and

$$\hat{\beta}_1 = \begin{bmatrix} 0.0824 \\ 0.0081 \\ 0.0035 \end{bmatrix}.$$

# Example

$$u_1 = 0.0260x_1 + 0.0518x_2$$

and

$$v_1 = 0.0824y_1 + 0.0081y_2 + 0.0035y_3.$$

The highest correlation occurs between an average of $x_1$ and $x_2$ weighted on $x_2$ and an average of $y1$, $y2$ and $y_3$, heavily weighted on $y_1$

The canonical correlations

$$\rho_1 = 0.6630$$

and

$$\rho_2 = 0.0412.$$

# Example



Predicting

$$\left(\tilde{\phi}_k\right)_i = \hat{\rho}_k \left((\eta_k)_i - \hat{\alpha}_k^T \bar{x}\right) + \hat{\beta}_k^T \bar{y}.$$

Here

$$\left(\tilde{\phi}_1\right)_i = 0.6630\left((\eta_1)_i - (0.0260 * 38.9545 + 0.0518 * 50.5909)\right)$$
$$+(0.0824 \cdot 50.6023 + 0.0081 \cdot 46.6818 + 0.0035 \cdot 42.3068)$$

$$\approx 0.6630(\eta_1)_i + 2.2905$$

$$\approx 0.6630(0.0260(x_1)_i + 0.0518(x_2)_i) + 2.2905$$

$$\approx 0.0172(x_1)_i + 0.0343(x_2)_i + 2.2905.$$

Note that this almost predicts $y_1$.

# Words of Warning



Canonical Correlation Analysis
Testing Independence
Scoring and Predicting
Words of Warning
References

# Some Words of Warning



Canonical Correlation Analysis
Testing Independence
Scoring and Predicting
Words of Warning
References

- The procedure maximizes the correlation between the linear combination of variables — it can be more than difficult to interpret the results.
- Correlation does not automatically imply causality.

# Next Week

Next week we will talk about discriminant analysis and classification.

# References

# References I



K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

# References II



- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.