Module 19 Multivariate Analysis for Genetic data
Session 07: Canonical Correlation Analysis

Jan Graffelman[1,2]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

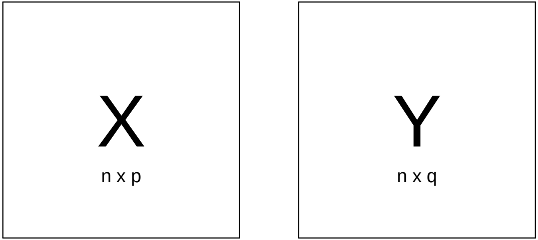
28th Summer Institute in Statistical Genetics (SISG 2023)

Contents

Some History

- Hotelling, H. (1935) The most predictable criterion, *Journal of Educational Psychology* 26 pp. 139-142.
- Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, 28 pp. 321-377.

Objective



- Study the relationship between two sets of variables **X** and **Y**
- Find linear combinations of **X** and **Y** that have maximal correlation

Notes

Canonical correlation analysis is important from a theoretical point of view, since it:

- underlies multiple linear regression ($q = 1$)
- underlies discriminant analysis (**Y** categorical)
- underlies correspondence analysis (**X** and **Y** categorical)
- underlies .... many multivariate methods!

Some notation

- $\mathbf{x}$ a p-variate random vector
- $\mathbf{y}$ a q-variate random vector
- $\mathbf{\Sigma}_{xx}$ within set correlation matrix of x variates
- $\mathbf{\Sigma}_{yy}$ within set correlation matrix of y variates
- $\mathbf{\Sigma}_{xy}$ between set correlation matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy} \end{bmatrix},$$

- $\mathbf{u} = \mathbf{a}'\mathbf{x}$ first canonical x variate
- $\mathbf{v} = \mathbf{b}'\mathbf{y}$ first canonical y variate

Maximization problem

We have:

$$V(\mathbf{u}) = \mathbf{a}'\mathbf{\Sigma}_{xx}\mathbf{a} \quad V(\mathbf{v}) = \mathbf{b}'\mathbf{\Sigma}_{yy}\mathbf{b} \qquad Cov(\mathbf{u}, \mathbf{v}, =)\,\mathbf{a}'\mathbf{\Sigma}_{xy}\mathbf{b}$$

Maximize

$$\rho(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{a}'\mathbf{\Sigma}_{xy}\mathbf{b}}{\sqrt{\mathbf{a}'\mathbf{\Sigma}_{xx}\mathbf{a}}\sqrt{\mathbf{b}'\mathbf{\Sigma}_{yy}\mathbf{b}}}$$

Equivalently

$$\max_{\mathbf{a},\mathbf{b}} \; \mathbf{a}'\mathbf{\Sigma}_{xy}\mathbf{b} \text{ subject to } \mathbf{a}'\mathbf{\Sigma}_{xx}\mathbf{a} = 1 \quad \mathbf{b}'\mathbf{\Sigma}_{yy}\mathbf{b} = 1$$

Solution as eigenvalue problem

The coefficients **a** and **b** are obtained as eigenvectors from the equations

$$\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}\mathbf{a} = \lambda_1\mathbf{a}$$

$$\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}\mathbf{b} = \lambda_1\mathbf{b}$$

- Eigenvalues are squared canonical correlations
- $min(p, q)$ succesive uncorrelated canonical variates can be extracted
- The covariance matrices $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{xy}$ and $\mathbf{\Sigma}_{yy}$ are estimated by the respective sample covariance or correlation matrices.

Solution as a singular value decomposition

$$\mathbf{K} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2} = \tilde{\mathbf{A}}\mathbf{D}\tilde{\mathbf{B}}'$$

- $\mathbf{D}$ diagonal matrix with canonical correlations
- $\mathbf{A} = \mathbf{R}_{xx}^{-1/2}\tilde{\mathbf{A}}$ matrix of canonical weights for x variables, $\mathbf{A}'\mathbf{R}_{xx}\mathbf{A} = \mathbf{I}$.
- $\mathbf{B} = \mathbf{R}_{yy}^{-1/2}\tilde{\mathbf{B}}$ matrix of canonical weights for y variables, $\mathbf{B}'\mathbf{R}_{yy}\mathbf{B} = \mathbf{I}$.
- Canonical x variates $\mathbf{U} = \mathbf{X}_s\mathbf{A}$
- Canonical y variates $\mathbf{V} = \mathbf{Y}_s\mathbf{B}$

Biplots in Canonical correlation analysis

$$\mathbf{K} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2} = \tilde{\mathbf{A}}\mathbf{D}\tilde{\mathbf{B}}'$$

$$\mathbf{F} = \mathbf{R}_{xx}\mathbf{A}\mathbf{D} \quad \mathbf{G} = \mathbf{R}_{yy}\mathbf{B}$$

Joint plot of first two columns of **F** and **G** is a biplot of the between set correlation matrix.

$$\begin{aligned}\mathbf{F}\mathbf{G}' &= \mathbf{R}_{xx}\mathbf{A}\mathbf{D}(\mathbf{R}_{yy}\mathbf{B})' = \mathbf{R}_{xx}\mathbf{A}\mathbf{D}\mathbf{B}'\mathbf{R}_{yy} = \mathbf{R}_{xx}^{1/2}\tilde{\mathbf{A}}\mathbf{D}\tilde{\mathbf{B}}'\mathbf{R}_{yy}^{1/2} \\ &= \mathbf{R}_{xx}^{1/2}\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2}\mathbf{R}_{yy}^{1/2} = \mathbf{R}_{xy}\end{aligned}$$

Like in PCA, there are different scalings

$$\begin{aligned}\mathbf{F} &= \mathbf{R}_{xx}\mathbf{A}\mathbf{D} \qquad \mathbf{G} = \mathbf{R}_{yy}\mathbf{B} \\ \mathbf{F} &= \mathbf{R}_{xx}\mathbf{A} \qquad \mathbf{G} = \mathbf{R}_{yy}\mathbf{B}\mathbf{D} \\ \mathbf{F} &= \mathbf{R}_{xx}\mathbf{A}\mathbf{D}^{1/2} \qquad \mathbf{G} = \mathbf{R}_{yy}\mathbf{B}\mathbf{D}^{1/2}\end{aligned}$$

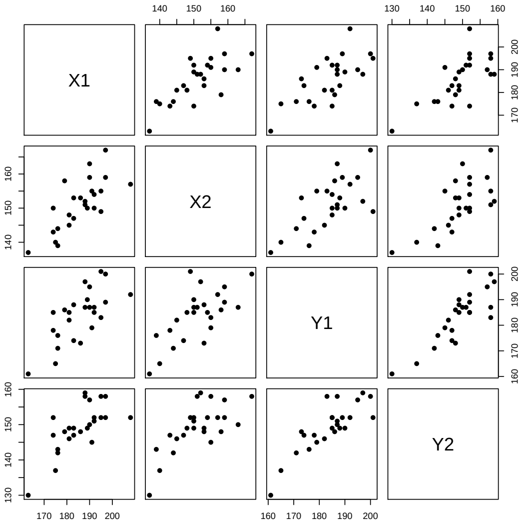
Textbook example: Fret's (1921) data

- 25 families
- x_1 head length of first son
- y_1 head length of second son
- x_2 head breadth of first son
- y_2 head breadth of second son

Family	x_1	x_2	y_1	y_2
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150

```
library(calibrate); data(heads);
```

Scatterplot matrix



Correlation matrix

	X_1	X_2	Y_1	Y_2
X_1	1.00	0.73	0.71	0.70
X_2	0.73	1.00	0.69	0.71
Y_1	0.71	0.69	1.00	0.84
Y_2	0.70	0.71	0.84	1.00

Canonical variates

$$U_1 = 0.5522x_1 + 0.5215x_2$$
$$V_1 = 0.5044y_1 + 0.5383y_2$$

$$U_2 = -1.3664x_1 + 1.3784x_2$$
$$V_2 = -1.7686y_1 + 1.7586y_2$$

Standard numerical output

	$r_1 = 0.7885$		$r_2 = 0.0537$	
	U_1		U_2	
x_1	0.5522	(0.9353)	-1.3664	(-0.3539)
x_2	0.5215	(0.9272)	1.3784	(0.3747)
Var. Expl.	0.8672		0.1328	
	V_1		V_2	
y_1	0.5044	(0.9562)	-1.7686	(-0.2927)
y_2	0.5383	(0.9616)	1.7586	(0.2743)
Var. Expl.	0.9195		0.0805	

Goodness-of-fit of between-set correlation matrix

	1	2
eigenvalue	0.6217	0.0029
fraction	0.9954	0.0046
cumulative	0.9954	1.0000

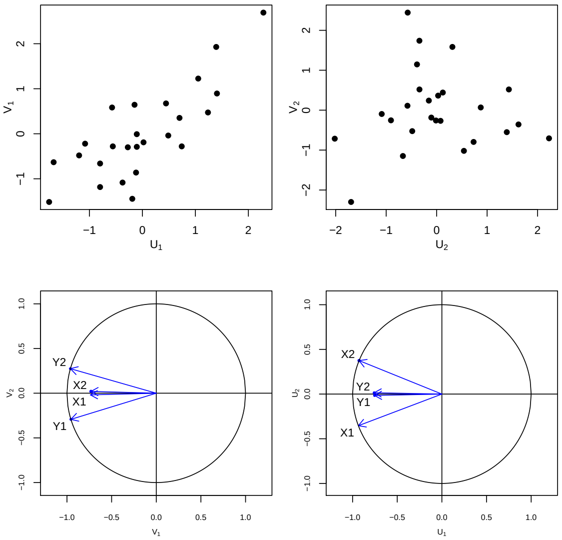
Correlation matrix of the canonical variates

	U_1	U_2	V_1	V_2
U_1	1.00	0.00	0.79	0.00
U_2	0.00	1.00	0.00	0.05
V_1	0.79	0.00	1.00	0.00
V_2	0.00	0.05	0.00	1.00

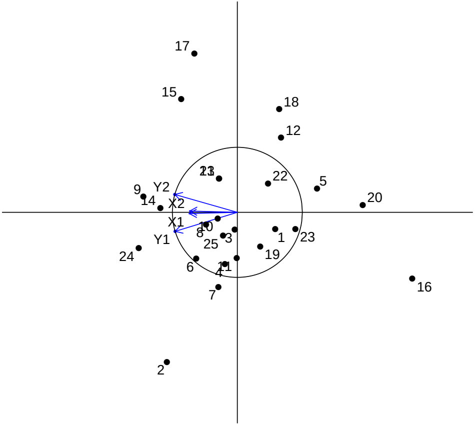
Shared and explained variance

- Shared variance between X and Y variables: squared canonical correlations
- Adequacy coefficients
  - Variance in X set explained by canonical x-variates
  - Variance in Y set explained by canonical y-variates
- Redundancy coefficients
  - Variance in X set explained by canonical y-variates
  - Variance in Y set explained by canonical x-variates

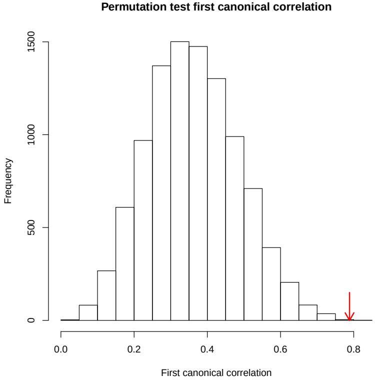
Classical plots and Biplots

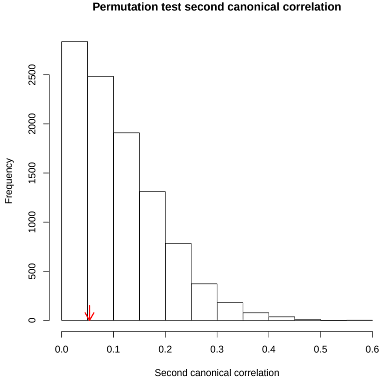


Representing cases

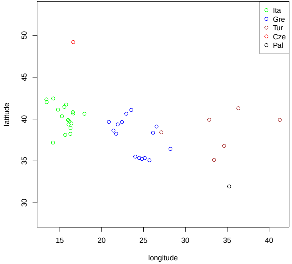


Significance of the canonical correlations





Genetic example



- 41 samples, 16 STRs
- Do allele frequencies depend on geographic coordinates?
- We use D10S1248

Messina F, et al. (2016) Spatially Explicit Models to Investigate Geographic Patterns in the Distribution of Forensic STRs: Application to the North-Eastern Mediterranean. *PLoS ONE* 11(11): e0167065.

The data

	Sampling.location	Country.Island	Long.E	Lat.N	N	A9	A11	A12	A13	A14	A15	A16	A17	A18
1	Brno	Czech Rep. (Moravia)	16.61	49.20	49	0.00	0.00	0.04	0.24	0.31	0.24	0.10	0.05	0.01
2	L'Aquila	Italy (Abruzzo)	13.40	42.35	32	0.00	0.02	0.03	0.25	0.33	0.09	0.17	0.11	0.00
3	Avezzano	Italy (Abruzzo)	13.43	42.03	28	0.00	0.00	0.02	0.25	0.32	0.23	0.14	0.04	0.00
4	Pescara	Italy (Abruzzo)	14.22	42.46	18	0.00	0.06	0.08	0.33	0.28	0.14	0.11	0.00	0.00
5	Benevento	Italy (Campania)	14.78	41.13	45	0.00	0.01	0.06	0.22	0.34	0.20	0.12	0.04	0.00
6	Foggia	Italy (Apulia)	15.54	41.46	26	0.00	0.00	0.04	0.31	0.31	0.12	0.14	0.10	0.00
7	Gargano promontory	Italy (Apulia)	15.75	41.73	30	0.00	0.00	0.03	0.22	0.28	0.27	0.17	0.03	0.00
8	Cilento promontory	Italy (Campania)	15.25	40.33	44	0.00	0.00	0.03	0.26	0.29	0.24	0.12	0.04	0.00
9	Matera	Italy (Basilicata)	16.60	40.67	34	0.00	0.00	0.04	0.26	0.29	0.19	0.16	0.04	0.00
10	Altamura	Italy (Apulia)	16.55	40.83	20	0.00	0.00	0.00	0.25	0.32	0.25	0.17	0.00	0.00
11	Brindisi	Italy (Apulia)	17.94	40.63	93	0.00	0.00	0.02	0.28	0.34	0.18	0.10	0.06	0.01
12	Lungro	Italy (Calabria)	16.12	39.74	24	0.00	0.02	0.00	0.25	0.23	0.27	0.21	0.02	0.00
13	Acri	Italy (Calabria)	16.38	39.49	30	0.00	0.00	0.03	0.20	0.43	0.22	0.05	0.07	0.00
14	Mormanno	Italy (Calabria)	15.99	39.89	36	0.00	0.00	0.00	0.24	0.29	0.26	0.12	0.08	0.00
15	Paola	Italy (Calabria)	16.04	39.36	38	0.00	0.00	0.01	0.34	0.35	0.13	0.09	0.05	0.01
16	Lamezia	Italy (Calabria)	16.28	38.93	17	0.00	0.00	0.00	0.38	0.12	0.38	0.06	0.06	0.00
17	Locri	Italy (Calabria)	16.26	38.23	38	0.00	0.01	0.01	0.14	0.38	0.22	0.17	0.04	0.01
18	Reggio Calabria	Italy (Calabria)	15.65	38.11	41	0.00	0.04	0.02	0.26	0.29	0.17	0.17	0.05	0.00
19	Butera	Italy (Sicily)	14.18	37.19	24	0.00	0.00	0.04	0.25	0.40	0.21	0.08	0.02	0.00
20	Ioannina	Greece (Epirus)	20.85	39.66	25	0.00	0.00	0.00	0.18	0.26	0.30	0.14	0.10	0.02
21	Agrinion	Greece (Aetolia-Acarnania)	21.41	38.62	22	0.00	0.00	0.00	0.32	0.29	0.20	0.14	0.04	0.00
22	Patrai	Greece (Akhaia)	21.73	38.25	62	0.00	0.01	0.02	0.18	0.33	0.22	0.23	0.02	0.00
23	Kardhitsa	Greece (Thessaly)	21.92	39.36	17	0.00	0.00	0.03	0.21	0.38	0.09	0.23	0.03	0.03
24	Larisa	Greece (Thessaly)	22.42	39.64	17	0.00	0.00	0.03	0.21	0.29	0.26	0.18	0.00	0.03
25	Thessaloniki	Greece (Central Macedonia)	22.94	40.64	15	0.00	0.00	0.03	0.17	0.43	0.30	0.07	0.00	0.00
26	Serrai	Greece (Central Macedonia)	23.54	41.09	27	0.00	0.00	0.04	0.17	0.43	0.24	0.09	0.04	0.00
27	Mitilini	Greece (Lesvos)	26.56	39.11	18	0.00	0.00	0.03	0.17	0.44	0.22	0.06	0.06	0.03
28	Khios	Greece (Khios)	26.14	38.37	30	0.00	0.00	0.00	0.35	0.37	0.15	0.10	0.03	0.00
29	Rhodos	Greece (Rhodos)	28.22	36.44	50	0.00	0.00	0.01	0.27	0.33	0.17	0.14	0.07	0.01
30	Crete (unspecified)	Greece (Crete)	24.81	35.24	65	0.00	0.01	0.04	0.20	0.33	0.20	0.16	0.06	0.00
31	Khania	Greece (Crete)	24.02	35.51	45	0.00	0.03	0.06	0.17	0.37	0.18	0.14	0.06	0.00
32	Rethmnon	Greece (Crete)	24.48	35.36	67	0.00	0.00	0.04	0.20	0.42	0.13	0.16	0.04	0.00
33	Iraklion	Greece (Crete)	25.14	35.34	48	0.00	0.02	0.03	0.18	0.41	0.16	0.17	0.03	0.01
34	Lasithi plateau	Greece (Crete)	25.71	35.08	38	0.00	0.00	0.07	0.14	0.43	0.18	0.10	0.07	0.00
35	Western Mediterranean coast	Turkey	27.13	38.42	54	0.00	0.00	0.03	0.20	0.34	0.30	0.09	0.04	0.00
36	Central Anatolia	Turkey	32.85	39.92	94	0.00	0.00	0.04	0.27	0.25	0.24	0.17	0.03	0.00
37	Black Sea coast	Turkey	36.33	41.29	42	0.00	0.00	0.06	0.23	0.29	0.23	0.14	0.06	0.00
38	Eastern Anatolia	Turkey	41.28	39.91	33	0.00	0.00	0.01	0.20	0.27	0.20	0.27	0.04	0.00
39	Eastern Mediterranean coast	Turkey	34.63	36.80	27	0.00	0.00	0.06	0.26	0.26	0.30	0.09	0.04	0.00
40	Cyprus	Turkey (Cyprus)	33.43	35.13	46	0.00	0.01	0.02	0.26	0.35	0.24	0.08	0.04	0.00
41	Palestine (CEPH)	Palestine	35.23	31.95	50	0.01	0.01	0.05	0.19	0.29	0.25	0.13	0.07	0.00

The analysis

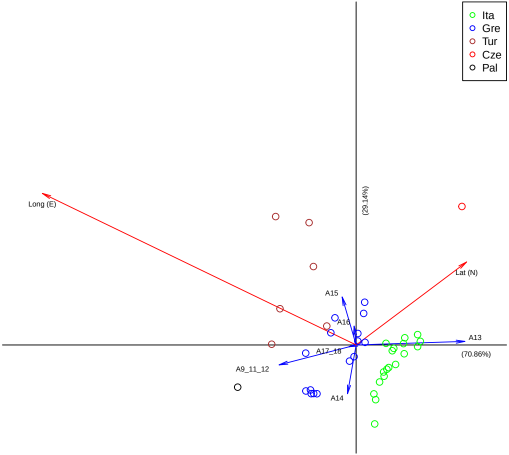
- Canonical analysis with $\mathbf{X}$ = geographical coordinates, $\mathbf{Y}$ = allele frequencies.
- We treat the allele frequencies as compositional, and do the clr transformation.
- We need to map samples to the biplot.
- We need to accomodate for singularity of the covariance matrix of the clr transformed data.

- Graffelman, J. (2005) Enriched biplots for canonical correlation analysis. *Journal of Applied Statistics* 32(2) pp. 173–188.
- Graffelman, J., Pawlowsky-Glahn, V., Egozcue, J.J. Buccianti, A. (2018) Exploration of geochemical data with compositional canonical biplots. *Journal of Geochemical Exploration* 194 pp. 120–133.

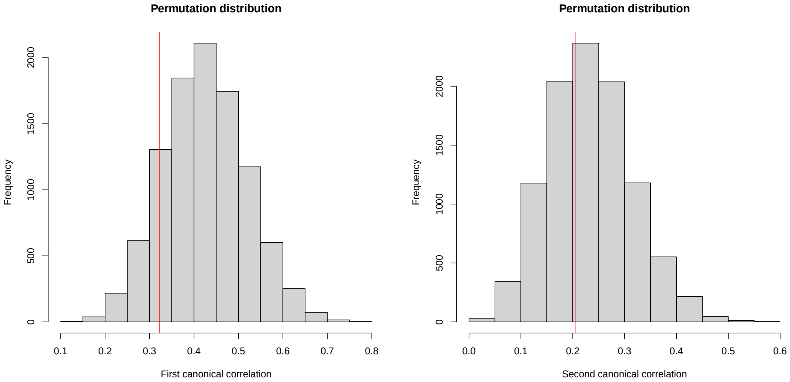
Goodness-of-fit

	1	2
correlation	0.322	0.206
eigenvalue	0.103	0.043
fraction	0.709	0.291
cumulative	0.709	1.000

The canonical biplot



The permutation test



permutation p-values 0.8631 and 0.6116

References

- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall, Chapter 10.
- Mardia, K.V. et al. (1979) Multivariate Analysis. Academic press. Chapter 10.