Module 19 Multivariate Analysis for Genetic data
Session 06: Correspondence analysis

Jan Graffelman

jan.graffelman@upc.edu

[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

- Benzécri, J.P. (1973) *Analyse des Données*, Dunod, Paris.
- Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press.

Objective

- Study the relationship between two (or more) categorical variables.
- Provide a picture of a contingency table.

Example data set: STR data

- Microsatellites consist of short sequences (e.g. ATT) that repeat a certain number of times (e.g. ATTATTATTATT).
- A small (2-6) number of base pairs is repeated.
- Individuals vary in the number of repeats they have.
- Produces count data, with a limited number of outcomes.
- Microsatellites have many alleles.

Microsatellites or STRs (Short Tandem Repeat)

- STRs can be coded in different ways:
  - reporting the number of repeats an individual has on each chromosome.
  - reporting the total size of the repeating sequences as the number of base pairs on each chromosome.
- Example:
  - a tri-nucleotide STR: ATT.
  - an individual has the DNA sequences (ATTATTATT,ATTATTCAA)
  - can be coded as (3/2) (repeats)
  - (9/6) (total size)
- In the statistical analysis mostly treated as categorical.

A glance at a STR database

	Id	STR1	STR2	STR3	STR4	STR5	STR6	STR7	STR8	STR9	...
1	794	129	264	142	156	157	171	205	183	196	...
2	794	155	292	146	156	166	179	205	187	196	...
3	795	145	288	138	168	157	171	205	195	196	...
4	795	150	292	142	172	166	175	210	203	196	...
5	796	155	292	138	156	157	167	205	183	184	...
6	796	155	300	142	156	169	171	205	199	196	...
7	797	150	264	142	156	157	171	205	187	196	...
8	797	155	292	146	176	163	175	205	187	196	...
9	798	150	292	138	156	157	171	205	183	187	...
10	798	155	300	146	160	166	171	205	207	190	...
11	799	155	296	146	152	157	167	205	179	196	...
12	799	155	296	146	176	157	171	210	183	196	...
13	800	145	264	138	156	157	163	205	187	190	...
14	800	160	296	146	156	157	171	210	199	196	...
15	801	155	264	142	156	157	175	205	183	196	...
16	801	155	292	146	184	166	179	209	199	199	...
17	802	145	292	138	176	157	159	193	183	187	...
18	802	155	296	142	180	166	171	201	187	187	...
19	803	155	280	142	172	166	163	205	183	196	...
20	803	155	300	142	176	169	175	213	187	196	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Example data set: NIST microsatellites

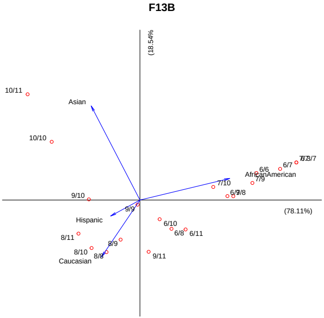
F13B	African American	Asian	Caucasian	Hispanic
10/10	5	54	58	48
10/11	0	1	1	0
6.3/7	1	0	0	0
6/10	31	1	22	27
6/11	1	0	1	0
6/6	43	1	4	4
6/7	50	0	1	3
6/8	23	0	22	8
6/9	57	0	15	11
7/10	16	1	5	3
7/7	7	0	0	0
7/8	10	0	3	1
7/9	24	0	3	3
8/10	10	5	64	32
8/11	0	0	0	1
8/8	6	1	24	9
8/9	19	2	41	28
9/10	21	26	73	49
9/11	1	0	2	0
9/9	17	5	22	9

- $n = 1036$ individuals of four ancestries.
- strbase.nist.gov
- Exact test p-value $= 0.0005$

There is association, but what is the nature of this association?

Result of Correspondence Analysis

F13B	African American	Asian	Caucasian	Hispanic
10/10	5	54	58	48
10/11	0	1	1	0
6.3/7	1	0	0	0
6/10	31	1	22	27
6/11	1	0	1	0
6/6	43	1	4	4
6/7	50	0	1	3
6/8	23	0	22	8
6/9	57	0	15	11
7/10	16	1	5	3
7/7	7	0	0	0
7/8	10	0	3	1
7/9	24	0	3	3
8/10	10	5	64	32
8/11	0	0	0	1
8/8	6	1	24	9
8/9	19	2	41	28
9/10	21	26	73	49
9/11	1	0	2	0
9/9	17	5	22	9



Some notation

- For the sake of illustration, we here disregard the Hispanic sample
- $\mathbf{N}$ the $I \times J$ contingency table.
- $\mathbf{P} = \mathbf{N}/n$ with $n = \mathbf{1}'\mathbf{N}\mathbf{1}$, and thus $\mathbf{1}'\mathbf{P}\mathbf{1} = 1$.
- $\mathbf{P}$ a matrix of probabilities (the correspondence matrix).

	African American	Asian	Caucasian	**r**
10/10	0.006	0.068	0.072	0.146
10/11	0.000	0.001	0.001	0.002
6.3/7	0.001	0.000	0.000	0.001
6/10	0.039	0.001	0.028	0.068
6/11	0.001	0.000	0.001	0.002
6/6	0.054	0.001	0.005	0.060
6/7	0.062	0.000	0.001	0.064
6/8	0.029	0.000	0.028	0.056
6/9	0.071	0.000	0.019	0.090
7/10	0.020	0.001	0.006	0.028
7/7	0.009	0.000	0.000	0.009
7/8	0.012	0.000	0.004	0.016
7/9	0.030	0.000	0.004	0.034
8/10	0.012	0.006	0.080	0.099
8/8	0.007	0.001	0.030	0.039
8/9	0.024	0.002	0.051	0.077
9/10	0.026	0.032	0.091	0.150
9/11	0.001	0.000	0.002	0.004
9/9	0.021	0.006	0.028	0.055
c	0.427	0.121	0.451	1.000

- Row masses

$$r_i = \sum_{j=1}^{J} p_{ij} \qquad \mathbf{r} = \mathbf{P}\mathbf{1} \qquad \mathbf{D}_r = diag(\mathbf{r})$$

- Column masses

$$c_j = \sum_{i=1}^{I} p_{ij} \qquad \mathbf{c} = \mathbf{P}'\mathbf{1} \qquad \mathbf{D}_c = diag(\mathbf{c})$$

Profiles

- A profile is a vector of non-negative elements that sum 1.
- The contingency table can be converted into a matrix of profiles.

	Row profiles		
	African American	Asian	Caucasian
10/10	0.043	0.462	0.496
10/11	0.000	0.500	0.500
6.3/7	1.000	0.000	0.000
6/10	0.574	0.019	0.407
6/11	0.500	0.000	0.500
6/6	0.896	0.021	0.083
6/7	0.980	0.000	0.020
6/8	0.511	0.000	0.489
6/9	0.792	0.000	0.208
7/10	0.727	0.045	0.227
7/7	1.000	0.000	0.000
7/8	0.769	0.000	0.231
7/9	0.889	0.000	0.111
8/10	0.127	0.063	0.810
8/8	0.194	0.032	0.774
8/9	0.306	0.032	0.661
9/10	0.175	0.217	0.608
9/11	0.333	0.000	0.667
9/9	0.386	0.114	0.500

	Column profiles		
	AfricanAmerican	Asian	Caucasian
10/10	0.015	0.557	0.161
10/11	0.000	0.010	0.003
6.3/7	0.003	0.000	0.000
6/10	0.091	0.010	0.061
6/11	0.003	0.000	0.003
6/6	0.126	0.010	0.011
6/7	0.146	0.000	0.003
6/8	0.067	0.000	0.061
6/9	0.167	0.000	0.042
7/10	0.047	0.010	0.014
7/7	0.020	0.000	0.000
7/8	0.029	0.000	0.008
7/9	0.070	0.000	0.008
8/10	0.029	0.052	0.177
8/8	0.018	0.010	0.066
8/9	0.056	0.021	0.114
9/10	0.061	0.268	0.202
9/11	0.003	0.000	0.006
9/9	0.050	0.052	0.061

Profiles

- Row (column) profiles are obtained by summing the elements of a row (column) in **P** and dividing by the total.
- $\mathbf{R} = \mathbf{D}_r{}^{-1}\mathbf{P}$ row profiles $\quad \mathbf{C} = \mathbf{D}_c{}^{-1}\mathbf{P}'$ column profiles
- Row and column masses turn out be weighted averages of the profiles

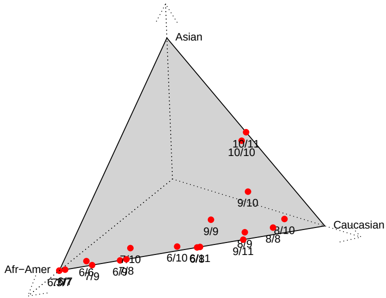
$$\mathbf{r}'\mathbf{D}_r{}^{-1}\mathbf{P} = \mathbf{1}'\mathbf{P} = \mathbf{c}' \qquad \mathbf{c}'\mathbf{D}_c{}^{-1}\mathbf{P}' = \mathbf{1}'\mathbf{P}' = \mathbf{r}'$$

Profiles and average profile

	Row profiles		
	African American	Asian	Caucasian
10/10	0.043	0.462	0.496
10/11	0.000	0.500	0.500
6.3/7	1.000	0.000	0.000
6/10	0.574	0.019	0.407
6/11	0.500	0.000	0.500
6/6	0.896	0.021	0.083
6/7	0.980	0.000	0.020
6/8	0.511	0.000	0.489
6/9	0.792	0.000	0.208
7/10	0.727	0.045	0.227
7/7	1.000	0.000	0.000
7/8	0.769	0.000	0.231
7/9	0.889	0.000	0.111
8/10	0.127	0.063	0.810
8/8	0.194	0.032	0.774
8/9	0.306	0.032	0.661
9/10	0.175	0.217	0.608
9/11	0.333	0.000	0.667
9/9	0.386	0.114	0.500
c	0.427	0.121	0.451

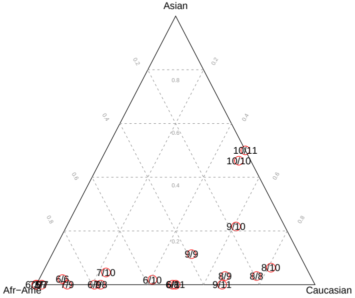
	Column profiles			
	African American	Asian	Caucasian	**r**
10/10	0.015	0.557	0.161	0.146
10/11	0.000	0.010	0.003	0.002
6.3/7	0.003	0.000	0.000	0.001
6/10	0.091	0.010	0.061	0.068
6/11	0.003	0.000	0.003	0.002
6/6	0.126	0.010	0.011	0.060
6/7	0.146	0.000	0.003	0.064
6/8	0.067	0.000	0.061	0.056
6/9	0.167	0.000	0.042	0.090
7/10	0.047	0.010	0.014	0.028
7/7	0.020	0.000	0.000	0.009
7/8	0.029	0.000	0.008	0.016
7/9	0.070	0.000	0.008	0.034
8/10	0.029	0.052	0.177	0.099
8/8	0.018	0.010	0.066	0.039
8/9	0.056	0.021	0.114	0.077
9/10	0.061	0.268	0.202	0.150
9/11	0.003	0.000	0.006	0.004
9/9	0.050	0.052	0.061	0.055

Row profiles in three dimensions



Row profiles in two dimensions

- The profile matrix for the data under consideration has rank 2.
- In this case, the profiles can be represented in a ternary plot.



Dimensionality

- The column rank of the matrix of row profiles is at most $J - 1$
- The row rank of the matrix of column profiles is at most $I - 1$
- "The" rank of the CA solution is $min(I - 1, J - 1)$

χ^2 statistic

	African American	Asian	Caucasian
10/10	5	54	58
	50.02	14.19	52.80
10/11	0	1	1
	0.85	0.24	0.90
6.3/7	1	0	0
	0.43	0.12	0.45
6/10	31	1	22
	23.09	6.55	24.37
6/11	1	0	1
	0.85	0.24	0.90
6/6	43	1	4
	20.52	5.82	21.66
6/7	50	0	1
	21.80	6.18	23.01
6/8	23	0	22
	19.24	5.46	20.31
6/9	57	0	15
	30.78	8.73	32.49
7/10	16	1	5
	9.40	2.67	9.93
7/7	7	0	0
	2.99	0.85	3.16
7/8	10	0	3
	5.56	1.58	5.87
7/9	24	0	3
	11.54	3.27	12.18
8/10	10	5	64
	33.77	9.58	35.65
8/8	6	1	24
	13.25	3.76	13.99
8/9	19	2	41
	26.50	7.52	27.98
9/10	21	26	73
	51.30	14.55	54.15
9/11	1	0	2
	1.28	0.36	1.35
9/9	17	5	22
	18.81	5.33	19.86

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(5 - 50.02)^2}{50.02} + \cdots + \frac{(54 - 14.19)^2}{14.19} = 467.95$$

Profiles and χ^2 statistic

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i,j} \frac{(np_{ij} - nr_i c_j)^2}{nr_i c_j} = n \sum_{i,j} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

$$\frac{\chi^2}{n} = \sum_{i,j} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = \sum_{i,j} r_i^2 \frac{(\frac{p_{ij}}{r_i} - c_j)^2}{r_i c_j} = \sum_{i,j} r_i \frac{(\frac{p_{ij}}{r_i} - c_j)^2}{c_j} = \sum_i r_i \sum_j \frac{(\frac{p_{ij}}{r_i} - c_j)^2}{c_j}$$

Likewise, for column profiles

$$\frac{\chi^2}{n} = \sum_j c_j \sum_i \frac{(\frac{p_{ij}}{c_j} - r_i)^2}{r_i}$$

- The quantity $\frac{\chi^2}{n}$ is known as the total inertia of the contingency table.
- Note that $\sum_j (\frac{p_{ij}}{r_i} - c_j)^2$ is squared Euclidean distance between profile i and average row profile
- Note that $\sum_j \frac{1}{c_j} (\frac{p_{ij}}{r_i} - c_j)^2$ is weighted squared Euclidean distance between profile i and average row profile (called χ^2 distance)
- Inertia is a weighted average of weigthed squared Euclidean distances.
- Inertia is a measure of spread of the profiles w.r.t. their average.

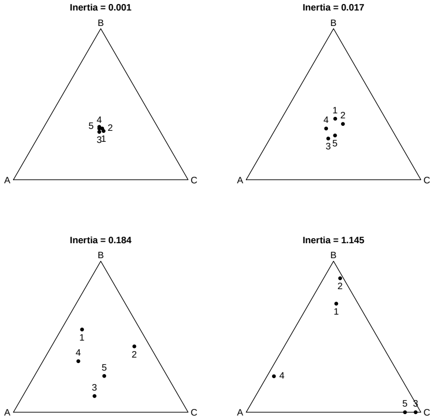
The geometrical interpretation of Inertia

	A	B	C
1	20	20	22
2	19	20	20
3	21	19	20
4	20	20	19
5	20	21	19
	100	100	100

	A	B	C
1	15	21	16
2	17	24	24
3	26	18	22
4	22	20	17
5	20	17	21
	100	100	100

	A	B	C
1	14	23	5
2	7	34	37
3	27	6	23
4	34	25	15
5	18	12	20
	100	100	100

	A	B	C
1	4	23	5
2	1	47	5
3	2	0	65
4	91	30	5
5	2	0	20
	100	100	100



Limiting situations

- Perfect independence: minimal inertia $= 0, \chi^2 = 0$.
- Perfect association: maximal inertia $= min(I - 1, J - 1)$.

Larger tables

- The profiles of genotypes over three populations data can be represented exactly in two-dimensional space
- Profiles of $I \times J$ contingency table can be represented exactly in $min(I-1, J-1)$ dimensional space.
- We search for an approximation of the profiles in one, two or at most three dimensions.
- The approximation is obtained by a (weighted) singular value decomposition.
- The criterion is to miminize errors in the approximation of the profiles, which is equivalent to maximizing the inertia of the profiles in a k dimensional subspace.
- The optimal solution can be obtained in several ways, here we use the svd of the standardized residuals of the contingency table.

Solution as a singular value decomposition

- $\mathbf{E} = \mathbf{P} - \mathbf{rc}'$ deviations from independence
- $\mathbf{D}_r^{-1/2}\mathbf{E}\mathbf{D}_c^{-1/2} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_c^{-1/2}$ "standardized residuals"

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_c^{-1/2} = \mathbf{UDV}'$$

Standard coordinates

$$\mathbf{F}_s = \mathbf{D}_r^{-1/2}\mathbf{U} \qquad \mathbf{G}_s = \mathbf{D}_c^{-1/2}\mathbf{V}$$

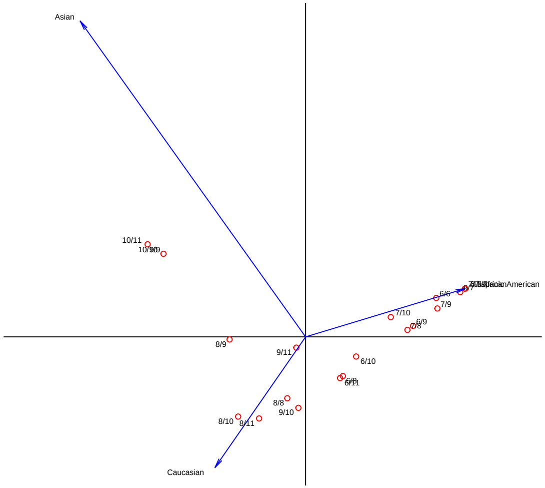
Principal coordinates

$$\mathbf{F}_p = \mathbf{D}_r^{-1/2}\mathbf{UD} = \mathbf{F}_s\mathbf{D} \qquad \mathbf{G}_p = \mathbf{D}_c^{-1/2}\mathbf{VD} = \mathbf{G}_s\mathbf{D}$$

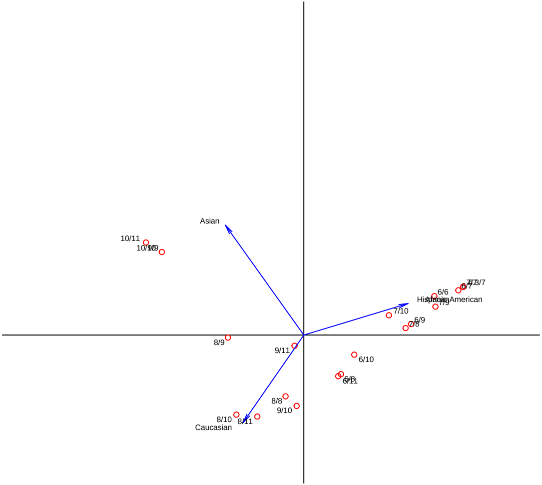
Graphical output of Correspondence analysis

- Joint plot of the rows of $\mathbf{F}_s$ and $\mathbf{G}_p$ (biplot of scaled row profiles)
- Joint plot of the rows of $\mathbf{F}_p$ and $\mathbf{G}_s$ (biplot of scaled column profiles)
- Joint plot of the rows of $\mathbf{D}_r{}^{-1/2}\mathbf{U}\mathbf{D}^{\frac{1}{2}}$ and $\mathbf{D}_c{}^{-1/2}\mathbf{V}\mathbf{D}^{\frac{1}{2}}$ ("symmetric" biplot)
- Joint plot of the rows of $\mathbf{F}_p$ and $\mathbf{G}_p$ (not a biplot)
- note that $\mathbf{F}_s\mathbf{G}_p{}' = \mathbf{D}_r{}^{-1/2}\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{D}_c{}^{-1/2} = (\mathbf{D}_r{}^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{D}_c{}^{-1}$
- and that $\mathbf{G}_s\mathbf{F}_p{}' = \mathbf{D}_c{}^{-1/2}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{D}_r{}^{-1/2} = (\mathbf{D}_c{}^{-1}\mathbf{P}' - \mathbf{1}\mathbf{r}')\mathbf{D}_r{}^{-1}$
- for convenience, the biplot vectors can be scaled, by the premultiplication of $\mathbf{G}_s$ by $\mathbf{D}_c{}^{1/2}$, or of $\mathbf{F}_s$ by $\mathbf{D}_r{}^{1/2}$ (the "contribution" biplot). This often pulls in outlying categories with small weight.

Biplot of the genotype data (row profiles)



Contribution biplot of the genotype data (row profiles)



Inertia decomposition STR F13B data

	1	2
Eigenvalue	0.458	0.127
Proportion	0.783	0.217
Cumulative	0.783	1.000

Note that

$$\frac{\chi^2}{n} = \frac{467.952}{800} = 0.585 = 0.458 + 0.127$$

Transition relationships (barycentric relationships)

From previous results

- $\mathbf{F}_p = \mathbf{D}_r{}^{-1}\mathbf{P}\mathbf{G}_s$
- $\mathbf{G}_p = \mathbf{D}_c{}^{-1}\mathbf{P}'\mathbf{F}_s$
- Principal coordinates of the rows are weighted averages of standard coordinates of the columns
- Useful for calculating coordinates of supplementary points

Supplementary points

- Supplementary points or inactive points are rows (columns) of the data matrix, usually collected under different conditions, that do not intervene in the computation of the solution.
- However, their representation in a biplot, posterior to the analysis, can be helpful for interpretation.
- Supplementary points can be situated in CA biplots by expressing them as profiles and using the transition relationships.

Contributions to Inertia

- In PCA we have seen that the total variance of data matrix can be decomposed into contributions made by dimensions (principal components), by variables, and finally by individual observations.
- In CA, a similar decomposition is possible, where the total inertia of a contingency table can be decomposed into contributions made by dimensions (principal axis), by the rows of the table, the columns of the table, and finally, the individual cells of a table.
- Such a decomposition is useful for spotting influential points in the analysis.

Contributions to Inertia

- we had

$$\frac{\chi^2}{n} = \sum_i r_i \sum_j \frac{(\frac{p_{ij}}{r_i} - c_j)^2}{c_j} = \sum_j c_j \sum_i \frac{(\frac{p_{ij}}{c_j} - r_i)^2}{r_i}$$

- each row (and column) make a contribution to the total inertia, these are called **row** and **column inertias**.
- note that

$$\frac{\chi^2}{n} = \sum_{i,j} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = \mathrm{tr}(\mathbf{D}_r{}^{-1}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_c{}^{-1}(\mathbf{P} - \mathbf{rc})') = \mathrm{tr}(\mathbf{D}^2)$$

- squared singular values (eigenvalues) are called principal inertias and constitute the contribution of each dimension in the solution to the total
- the inertias of each row (column) can be decomposed into contributions made by the principal axis. This allows one to judge how much of the inertia of each row (column) is accounted for by each axis, and to compute goodness-of-fit statistics for each point.

References

- Benzécri, J. P. (1973) *Analyse des Données*, Dunod, Paris.
- Greenacre, M. J. (1993), *Correspondence Analysis in Practice*, Academic Press.