Correspondence Analysis

MIRI / MVA - Lab #5 Report

Authors: Cedric Bhihe <cedric.bhihe@gmail.com>
Santi Calvo <s.calvo93@gmail.com>

Delivery: before 2018.05.04 – 23:55

1. Read again the PCA_quetaltecaen data

As in our previous Lab·#2 report, we read the "¿Qué tal te caen?" data set. . The initial data set is available as `Data/PCA_quetaltecaen.csv`. We imported it as the data-frame object, `qttc`, by means of our R script available at `Lab5/lab5-script_ca.R`. We treat qttc as a contingency table where columns are the 8 modalities of the "cultural identity" categorical variable and rows are the 8 modalities of the "geographical" categorical variable.

Exploring that data, we perform a χ^2 test in order to test whether a significant association (dependence) exists between row and column modalities of the categorical variables. There are (8-1) * (8-1) = 49 degrees of freedom, while the null hypothesis, H_0, is: "there is no significant association between row and column modalities of the categorical variables," or in other words "the joint distribution of the cell counts in the 2-D contingency table is the product of the row and column marginals".

We obtain a 2 sided χ^2 test statistics of ~7.26, which falls outside the interval[i] of ~[31.6 , 70.2] at 5% overall risk level 2.5% on each side). This allow us to reject H_0 at the risk 5% of erring. Taking into account the warning issued by the chi-squared R method, our preliminary conclusion is that there is a significant association of row and column categorical variables.

Ballon and mosaic plots in Fig.1 below illustrate the relative importance of each cell in the count table by scaling the blue dot size (left) or the bar's surface (right) with each cell's count number, and showing that each cell's χ^2 residuals remains small. Cells' residuals are generally negative off diagonal and always positive on diagonal, thereby suggesting in a very anticipated way that the diagonal may be slightly overloaded.

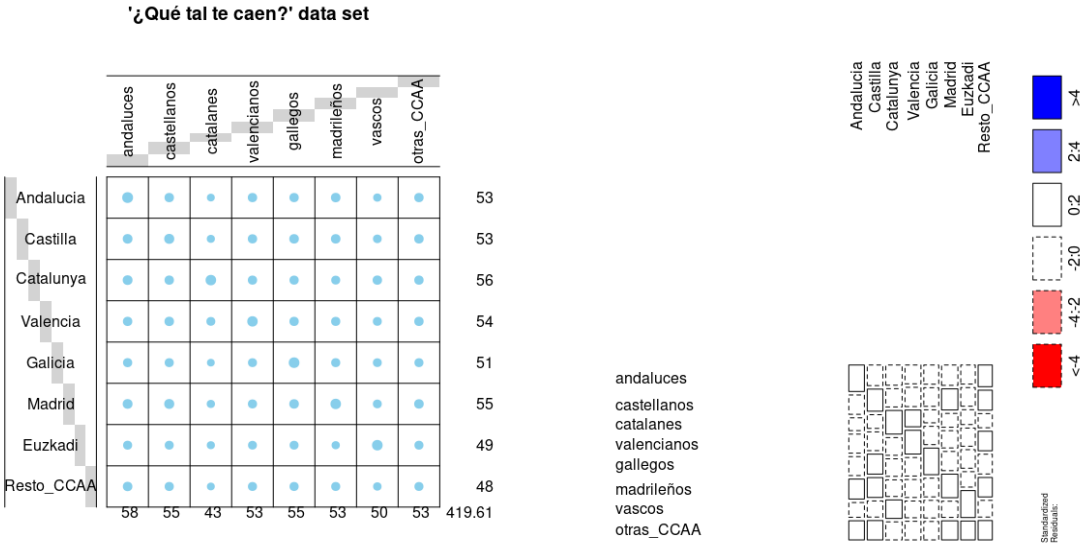

***Fig. 1**: The balloon plot (left) shows that the diagonal is slightly preferred with generally larger dots than for off diagonal terms. The mosaic plot (right) shows that observed and expected cell's frequencies are close (standardized residuals are colored white).*

2. Perform a CA on the data. How many significant dimensions are there ? Interpret the first factorial plan

We conducted a manual centering of the table frequencies to obtain X_{ctd}. "Cultural identity" frequency marginals show incidentally that "Catalanes" (column modality) have the lowest perceived similarity of all populations in Spain ($f_{\cdot j} \approx 0.1024$ for j=3) followed in that by "Vascos" ($f_{\cdot j} \approx 0.1188$ for j=7) .

Manually calculated centering is performed by incorporating the implicit χ^2 metric effect into the row-profile cloud representation *(lines 87 to 116 of the R-script)*. From this a diagonalization can ensue to reveal eigenvectors as well as eigen-

i The χ^2 statistics is the sum over all frequencies of the difference squared between observed and expected frequencies, divided by the expected frequencies (assuming independence).

values and the cumulative variance representativeness (Table 1) associated with each corresponding PC.

The number of significant dimensions is 3, at a level of representation of total inertia of at least 80% (~83% in our case). The diagonalization yields only 7 principal directions as the dimensionality of the table is reduced by 1 due to the imposed double constraint: $\sum_j f_{.j} = \sum_i f_{i.} = 1$. Those numerical results are confirmed a PCA analysis (not represented here but implemented in the R-script) and the correspondence analysis (CA) presented hereafter.

Eigenvalues	Cumulative variance (%)
7.7541e-03	44.83
4.4953e-03	70.82
2.1099e-03	83.01
1.5065e-03	91.72
1.2303e-03	98.83
1.9718e-04	99.97
4.7384e-06	100.00

***Table 1**: Eingenvalues and cumulative percentual representation of inertia.*

We proceed with a CA using FactoMineR, ensuring that rows are weighed with rowsums, $f_{i.}$.

Fig. 2 shows a factor map in the first factorial plane (left) and a screeplot (right). In the former **row (blue)** and **column (red)** profiles are printed together with distinct colors for easier differentiation. Distances between same-color points are distances in the χ^2 sense. A red point (column profile) is a barycenter for the blue points (row profiles) weighted by said column, and vice versa. The reader should be warned against the temptation of interpreting row (blue) and column (red) profiles' proximity on the PC1-2 plane in the χ^2 distance sense. Differently colored points may appear close, but no viable conclusion can be drawn from that fact. For that reason we also plot separate projections for row and column profiles in Fig.3.

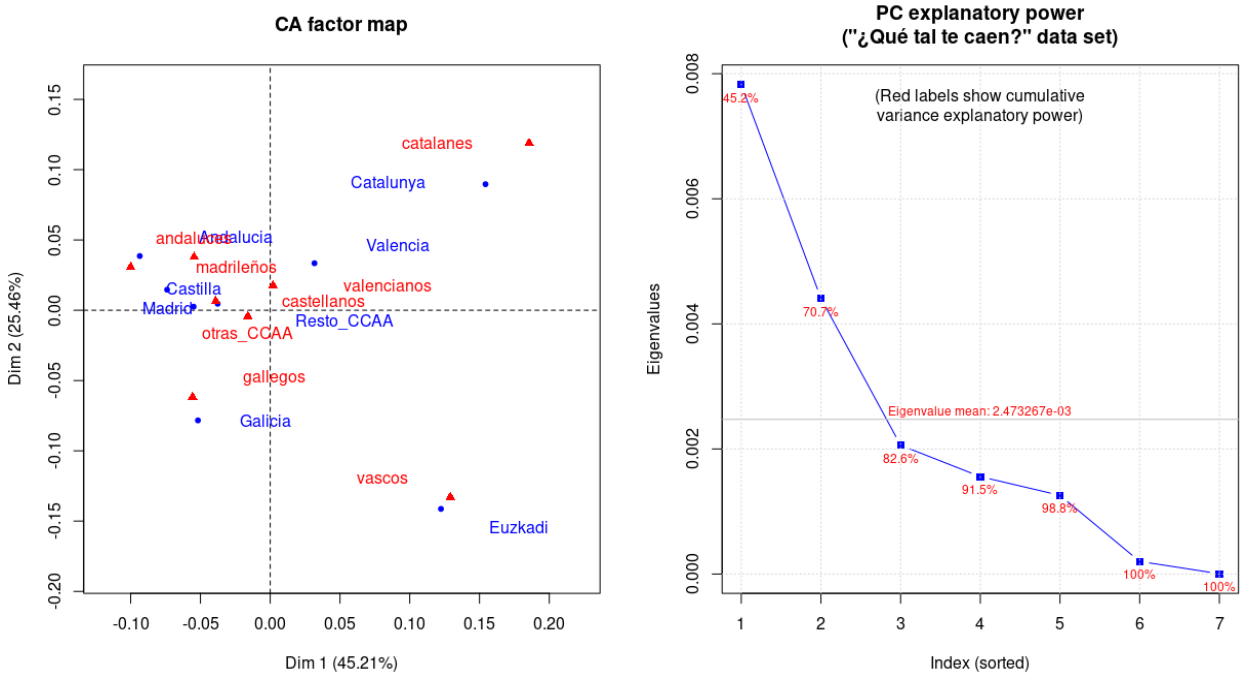

***Fig. 2**: CA factor map and screeplot for the data set "¿Qué tal te caen?".*

We observe that Catalonia, the Basque country and Galicia on one hand and their populations on the other hand are separated from the rest of regions and populations. Citing from our previous Lab#3 report, the most striking aspect is how Basques and Catalans are perceived as removed and different from the rest of Spanish regions. They are also the most distant from one another along PC2, contrasting with their proximity along PC1. To a lesser degree, Galicians are also perceived as being "apart" from the rest of the pack. As is the case for the Basque region and Catalonia, Galicia is a region characterized by a strong cultural identity and a specific language. By contrast the rest of regions all share castillan as their *prima lingua*.

As also pointed out earlier, we need to have access to the semantics of the questionnaire submitted to the populations of those various regions to begin to refine that interpretation.

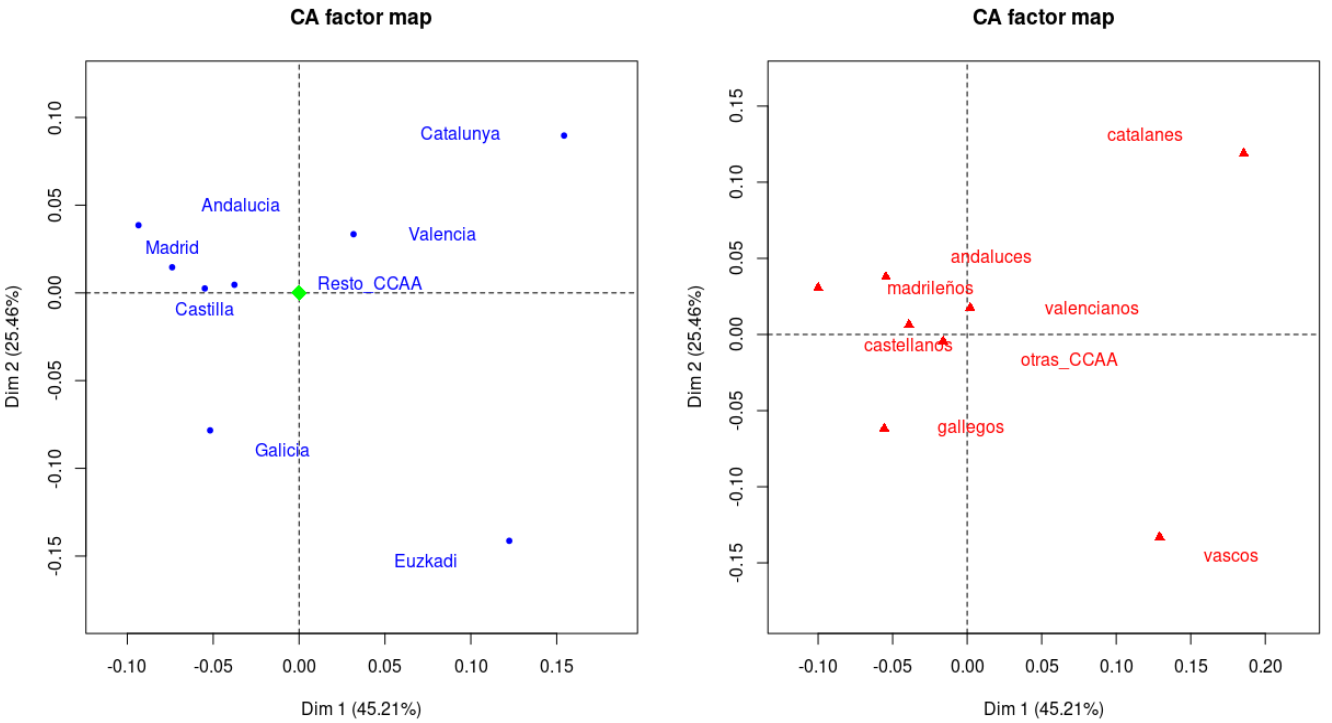

***Fig. 3**: Row profile cloud (left) and column profiles cloud (right). Both show a pronounced Gutmann effect where Catalunya and Euzkadi on one side and "catalanes" and "vascos" on the other are on opposite sides of crescent shaped scatter-plots.*

The distance between profile points and the origin measures the quality of the profile points on the factor map. Points situated away from the origin are generally well represented on the factor map.

3. For the same dataset, compute the contribution of each cell to the total inertia. Compute the percentage of inertia due to the diagonal cells.

	Andaluces	Castellanos	Catalanes	Valencianos	Gallegos	Madrileños	Vascos	Otras_CCAA
Andalucia	1.303498e-03	4.739381e-06	3.659717e-04	3.084034e-08	3.368955e-05	4.433239e-05	4.663950e-04	4.361343e-06
Castilla	1.550959e-06	2.516498e-04	1.580075e-04	4.374119e-06	8.937748e-07	3.411794e-05	1.430193e-04	1.139230e-05
Catalunya	1.590274e-05	8.625272e-05	3.846138e-03	7.931462e-05	2.308695e-04	2.581110e-04	1.564309e-05	3.095999e-05
Valencia	1.656530e-04	7.303155e-06	1.012375e-04	9.868932e-04	7.539885e-05	4.565323e-05	5.430294e-05	2.403469e-06
Galicia	9.871097e-05	4.394589e-05	2.970143e-04	2.715242e-05	1.859812e-03	3.756521e-06	2.327578e-06	7.829120e-06
Madrid	8.210152e-06	8.902189e-05	1.520766e-04	1.483750e-04	1.710377e-05	9.646440e-04	1.392706e-04	1.782190e-06
Euzkadi	1.120936e-04	4.212389e-05	1.602183e-05	3.308056e-05	7.111767e-05	5.204745e-04	3.611128e-03	3.542396e-08
Resto_CCAA	2.905985e-05	2.782769e-06	8.099904e-05	8.175150e-06	4.822098e-07	9.236540e-06	6.037135e-05	9.757530e-06

The total inertia is: 1.7298 10^{-2} and diagonal cells account for ~74.2 % of that. We conclude that there exists an overloaded diagonal effect known as the Gutmann effect.

Our earlier observation following our χ^2 test of independence for rows and columns was that we could reject the notion that columns and rows were independent. Independence would have been consistent with the clouds of row profiles (blue dots) and of column profiles (red dots) being concentrated around the centroid (green dot in Fig. 3 for row profiles) and the inertia to be zero or very close, in keeping with H_0: $f_{ij}[s,t] = f_{i.}[s] \times f_{.j}[t]$ for all s and t.

4. Nullify the influence of overloaded diagonal elements (in terms of inertia). Heed the fact that each imputation modifies the marginals frequencies.

We apply a convergence criterion of $\varepsilon=1e^{-4}$ on the total inertia. Our algorithm is easily grasped as follows:

1) From count table, initialize total count N, relative frequencies tables (RFT), f_{ij} and inertia matrix
2) calculate row weights $f_{i.}$ and column weights $f_{.j}$
3) Impute new diagonal terms: $N * f_{i.} * f_{.j}$ in count table

4) Recompute new total count `N = sum(count table)`
5) Recompute inertia matrix and total inertia
6) if | `new_tot_inertia - old_tot_inertia` | `>=` ε go to step (2).

The resulting new total inertial contribution by cells is:

	Andaluces	Castellanos	Catalanes	Valencianos	Gallegos	Madrileños	Vascos	Otras_CCAA
Andalucia	1.795704e-08	4.568475e-08	9.469690e-05	2.536410e-05	1.017592e-06	1.475727e-04	1.726201e-04	5.922702e-06
Castilla	5.946106e-07	6.579982e-08	3.518200e-05	5.603027e-07	1.746085e-05	5.467831e-05	4.100794e-05	1.535041e-08
Catalunya	2.639605e-05	1.669288e-05	1.653618e-07	1.013686e-06	3.312910e-05	7.329674e-05	2.958001e-04	5.932073e-06
Valencia	7.215353e-05	2.212422e-06	4.196512e-04	9.052688e-10	9.544417e-06	8.452728e-06	9.379084e-07	5.999876e-06
Galicia	1.039368e-05	1.031151e-05	4.773939e-05	7.840742e-07	4.972199e-08	1.850870e-05	8.356879e-05	1.854005e-06
Madrid	3.348218e-06	1.176160e-04	1.110114e-05	7.391455e-05	2.566611e-06	7.398639e-10	1.388856e-05	1.247870e-07
Euzkadi	2.084750e-06	3.349735e-07	8.835364e-05	7.483029e-06	2.984638e-06	2.346470e-04	1.499002e-07	1.796651e-05
Resto_CCAA	2.768955e-05	3.304666e-06	1.876364e-05	4.598647e-06	3.933453e-08	5.458740e-06	1.832184e-05	1.808862e-07

The new total inertia value is about 8 times smaller at 2.3963 10^{-3}. and diagonal cells now account for ~0.03% of that. We effectively got rid of the diagonal overloading effect.

5. Perform a new CA for the dataset with modified diagonal. Interpret the results.

Fig. 4 summarizes the result of the new CA based on the new count table corrected for the Gutmann effect. This time the 3 first dimensions account for close to 90% of all inertia while the first 2 account only for 78%. Altogether the first factorial plane represent inertia better than previously (78% now, instead of 71% before).

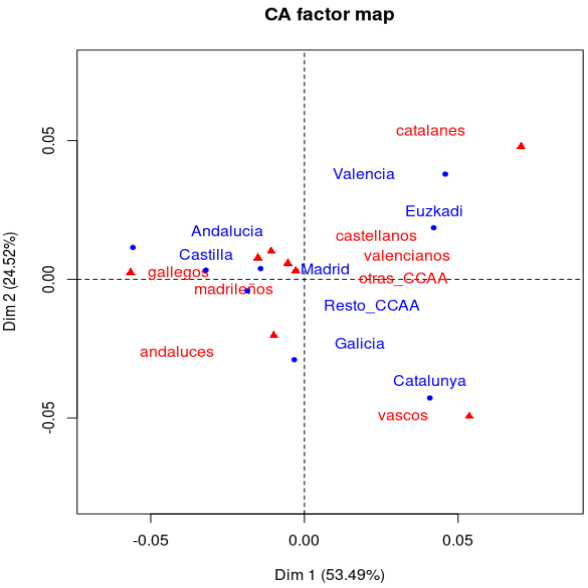

***Fig. 4 (right)**: Observe a regrouping of the scatter plot projections around the centroid, consistent with smaller inertia.*

This is best represented by the bar plot (Fig. 5) of the sum of squared cosines in the first factorial plane before and after the Gutmann effect correction, in this case for row profiles. The closer a bar gets to 1, the better represented the corresponding variable modality is by its projection in the first factorial plane. Table 2 (below right) exhibits squared cosines for row profiles along the 3 first PCs. Adding the first 2 columns of Table 2 yields the height of the red bar in Figure 5.

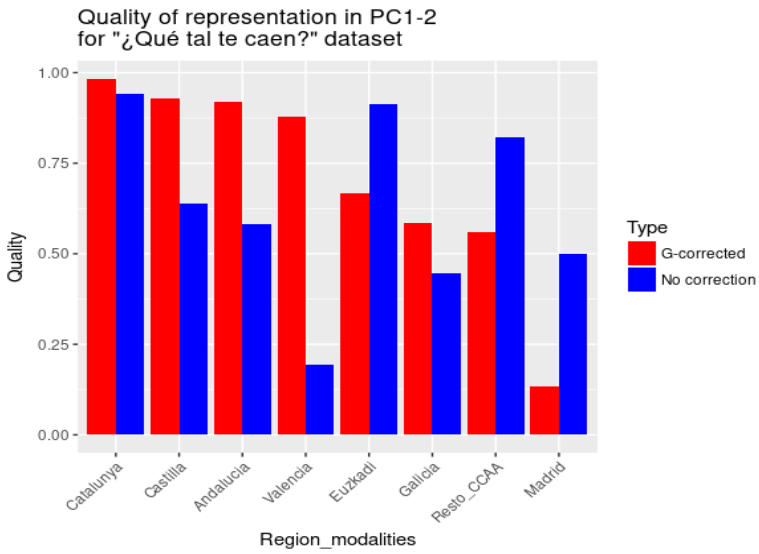

***Fig. 5**: Quality of representation of row profiles in the PC1-2 factorial plane, before (blue) and after (red) the Gutmann effect correction.*

	PC1	PC2	PC3
Andalucia	0.883	0.037	0.046
Castilla	0.918	0.010	0.004
Catalunya	0.468	0.515	0.000
Valencia	0.522	0.357	0.069
Galicia	0.007	0.578	0.216
Madrid	0.124	0.009	0.700
Euzkadi	0.558	0.109	0.066
Resto_CCAA	0.533	0.027	0.002

***Table 2**: cos^2 along the first 3 PCs for row profiles, after correcting the Gutmann effect.*