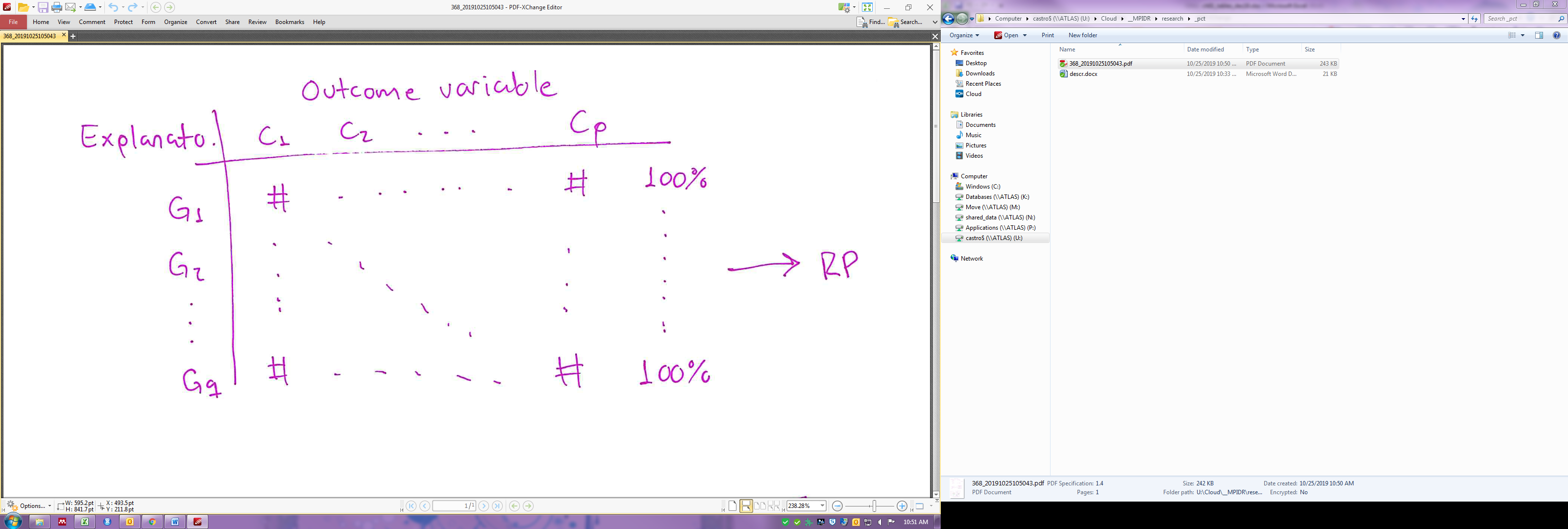
**Plotting contingency tables**

Andres Castro

The idea is to generate a small series (ideally only one) of two dimensional plots that allow the researcher to see the main associations between outcome categories and explanatory categories with two conditions: (1) the plot should display all potential comparisons without the need of a reference category, this include the representation of the marginal distribution of the outcome variable as the center of the plot, (2) the plot should provide information on both the strength of the associations and the level of uncertainty associated to them (error due to sampling). Because the plot would necessary be a reduced version of the actual model, the plot would only display main associations and the degree of accuracy regarding uncertainty will be lower than the accuracy of numeric standard errors.

Consider an hypothetical situation where there is an outcome variable of *p* categories (C1, C2, …, Cp) and an explanatory variable with *q* categories (G1, G2, …, Gq). First, one should get a table of predicted probabilities for all the groups (Gi) of the explanatory variable using the fitted model. In other words, one needs to obtain conditional distributions along with their standard errors. These conditional distribution are organized in a table where rows are the categories of the explanatory variables and the columns are the categories of the outcome variable (see table below). Technically speaking this table is called the row-profile table of a (predicted) contingency table between the outcome and the explanatory variable.

**Table 1 – General structure of a row-profile (RP) table**

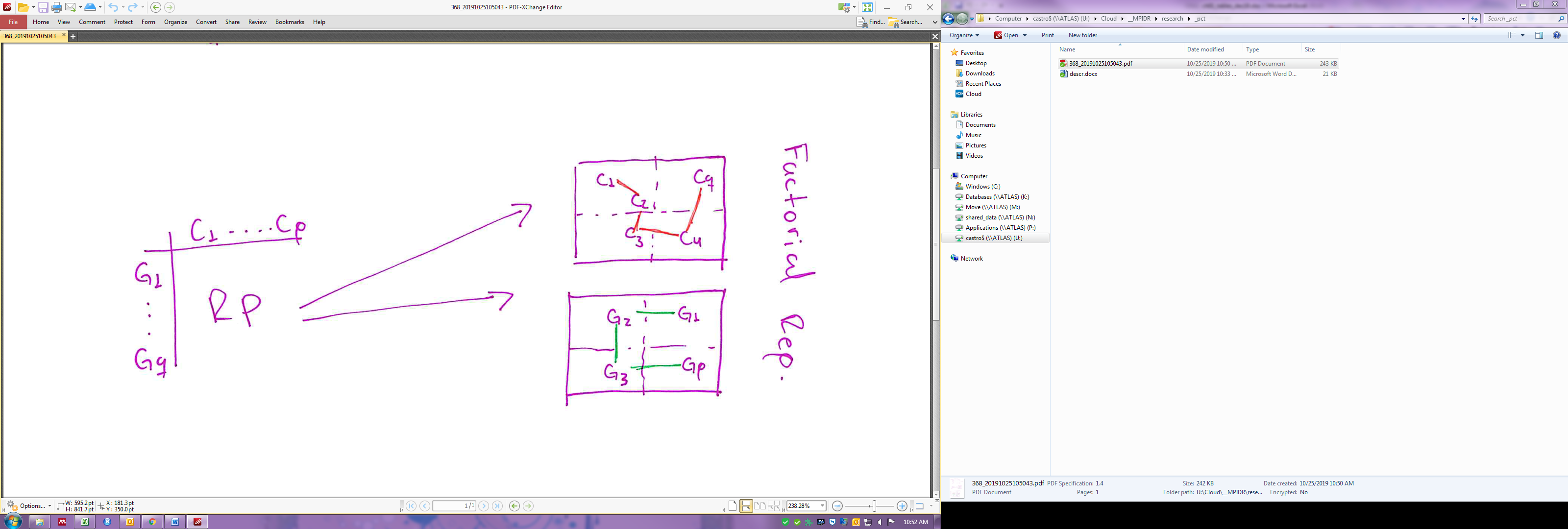


Further, one can run a Principal Component Analysis (PCA) on this row-profile table. PCA is geometric data analysis technique that summarizes relationships between rows and columns and allows for the joint representations of them in an hierarchically organized Euclidean space. Basically, PCA decomposes the variance of a given table into orthogonal axes. These axes (often called factorial dimensions) are hierarchical organized in terms of the proportion of variance they account for. The first axis comprises the largest proportion. The number of axes equals the number of categories of the outcome variable. Hence, each row is represented as a *p*-dimensional point, and each column as a *q*-dimensional point. These two set of coordinates have mathematical properties that allow for a joint representation of them.

Depending on the distribution of the % of explained variances across the factorial dimensions, the researcher must choose how many factorial axes are needed to produce an adequate representation of the row-profile table. There is no criteria for making this decisions, but typically 2 to 3 dimensions are enough to depict the main patterns in table of the size that social scientists typically produce (say, *p* and *q* < 10). Note that if only two dimensions are needed, then the table will be represented in a single plot where the x-axis is the first factorial dimension, and the y-axis the second. If three dimensions are kept, three plot can be produced (first dimensions vs. second dimension, and first dimension vs. third dimension, and second dimension vs. third dimension).

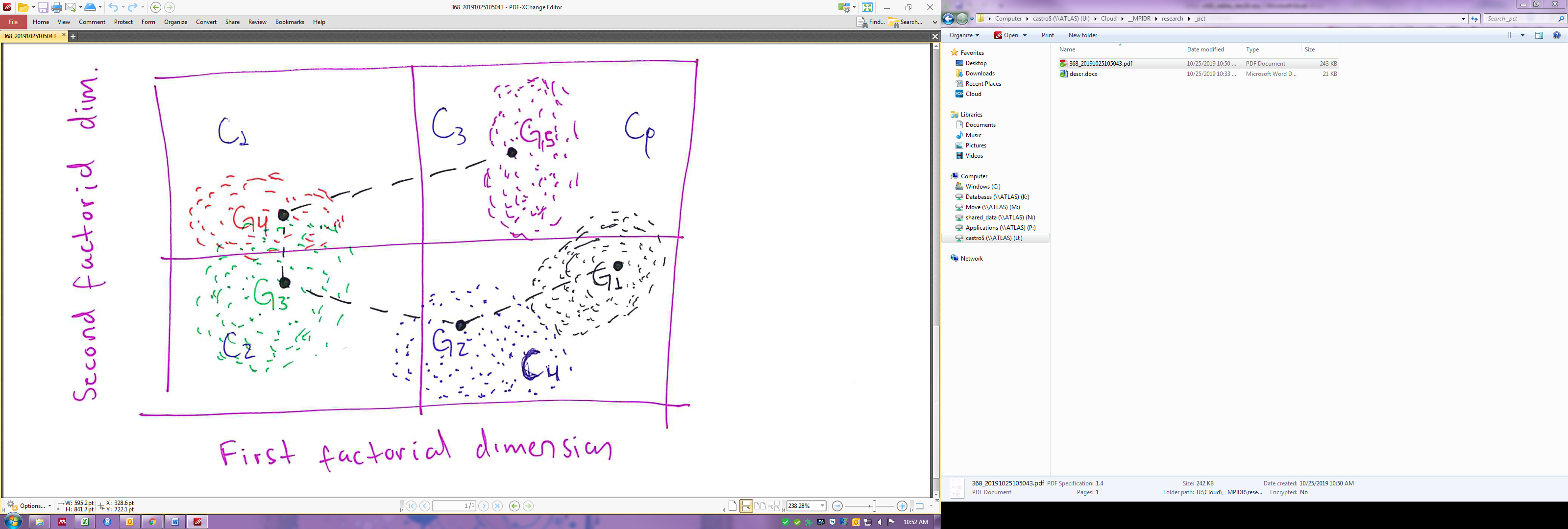
Factorial dimensions have two important characteristics when they are computed via PCA. First, the origin of the space, i.e. the (0,0, …, 0) coordinate represents the marginal distribution of the outcome variable. This means that deviations from the center can be interpreted as discrepancy with respect to the overall row-profile. Second, proximity and distance among categories can be directly interpreted as positive and negative association respectively. If two categories are close to one another, this means they are positively correlated, if two categories are distal, this implies they are negatively related. The Figure below summarizes the main idea of this approach, namely, converting a contingency table into factorial plots.

**Figure 1 – Relationship between row-profile tables and factorial representations**



In addition, one can use the fitted model for replicating the row-profile table by sampling conditional distributions (row-profiles) several times, say 1,000 times. All these table can be appended in a single master table of 1,000 x *q* rows and *p* columns, lets called that big table **X**. One can run a PCA on **X** and create bi-dimensional plots as described above. Instead of points for the categories of explanatory variables one would get clouds of points. These clouds represent the space of potential outcomes, i.e. potential associations. Overlapping could signal correlations, while the absence of intersection between clouds would signal strong negative relationships. The figure below presents an hypothetical outcome of this approach. I added lines to connect categories of the explanatory variables, assuming they are ordinal factors.

**Figure 2 – Hypothetical cloud of points**



It may be worth recalling the advantage of this approach.

1. Joint representation of explanatory and outcome variable categories
2. Absence of reference category
3. Center of plot represent the mean profile
4. Interpretation: proximity among categories of independent variable implies similarity across groups. Proximity between outcomes and explanatory categories implies positive correlations
5. Ordinal variables can be connected using lines to highlights patterns

This work proposes a methodology to produce bi-dimensional plots that summarize complex association patterns between categorical outcomes and explanatory factors. The methodology is very useful when the number of categories is large (say above 3 for outcomes and explanatory variables) and when the explanatory variables are ordinal, i.e. when explanatory categories are hierarchically organized (e.g. primary education, secondary education, etc.). These contexts are very common in social science research, especially when researches combine cluster analysis and multinomial models.

The combination of these two methods is very useful for contexts where researchers want to explore the associations between a categorical outcome, typically the cluster solution, and categorical explanatory variables including their interactions. Both cluster and multinomial analysis are well-designed for measuring associations but each of them has one limitation that can be overcome with plots.

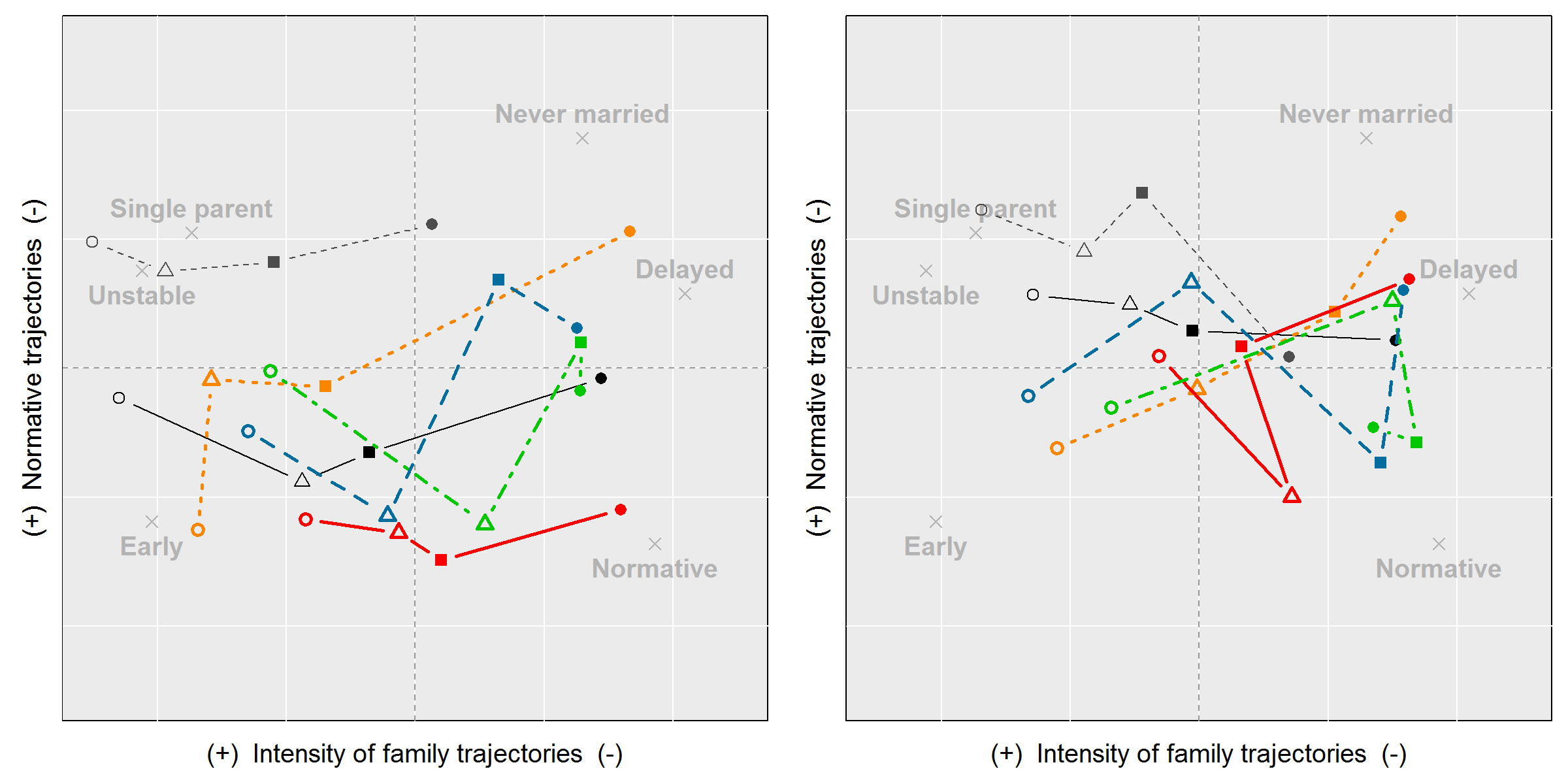
* **Cluster analysis:** often, adequate cluster solution involved a large number of clusters (e.g. 5 to 8). A categorical variable with 5-8 categories is difficult to describe, and researchers often focus on one or two salient groups. This approach is particularly prone to miss overarching patterns.
* **Multinomial models:** the interpretation of summary tables from a multinomial model is cumbersome because it requires the selection of a reference category for the dependent and the independent variables. This situation gets more complicated if researchers are interested in interactions because the number of categories increases and the reference category is, itself, an interaction.

**One application**

Table below presents the conditional distribution of women across a six-category typology of family trajectories. The explanatory categories corresponds to a combination of women’ place of birth (US or abroad, age at migration, and race/ethnicity). Each rows adds to 100 and the number in parenthesis correspond to the standard error. The last row contains the marginal distribution of the family typology, termed here family profile.



Patterns can be read from the previous table, but a plot will provide a faster solution for this.



**Educational attainment:** ○ Lowest ∆ Low *■* Med. ● High

**Age at migration:** ——NH-white — —Other • • • **Before age 18**

**——19 to 24 — - —25 to 30 — —After age 30**

Notes: the mean profiles are (Women-Men): *Never married* (9.1-14.0), *Delayed* (17.5-22.9), *Normative* (28.6-29.6), *Unstable* (9.2-9.9), *Single parent* (8.1-5.8), *Early* (27.6-17.8).