APPENDIX 2. NOTE ON MULTIPLE CORRESPONDENCE ANALYSIS (MCA)

To introduce Multiple Correspondence Analysis (MCA), it is convenient to adopt the *language of questionnaire*. The basic data set for MCA is an Individuals × Questions table, where the questions are *categorized variables*, i.e. variables with a finite number of categories, or *modalities.* MCA applies directly to a questionnaire in "standard format", that is, when for each question, each individual chooses one and only one response modality. For questionnaires not in standard format, a preliminary phase of coding is necessary. The modalities may be qualitative, or result from the splitting of quantitative variables. Denoting I the set of n individuals and Q the set of questions, the basic data table analyzed by MCA is thus an $I \times Q$ table, with in cell (i, q) the modality of question q chosen by individual i. As a method of Geometric Data Analysis (GDA), MCA provides a geometric model of data, that is, it basically represents the set of individuals by a *cloud of points*, for which principal directions are sought. Methodologically, MCA appears to be the counterpart of Principal Component Analysis (PCA) for categorized variables. A detailed presentation of MCA together with case studies are found in Le Roux & Rouanet (2004, chapters 5 & 9).

Historical note. The references Guttman (1941) and Burt (1950) are precursor papers of MCA as an optimal scaling procedure of categorized variables. In the early seventies, MCA emerged as an extension of CA to a table of Individuals × Categorized variables after disjunctive coding: see Benzécri (1982). The phrase "Analyse des Correspondances Multiples" appears for the first time in the paper by Lebart (1975), devoted to MCA as a method in its own right. In the late seventies, MCA became a major method for the analysis of questionnaires; it has been constantly used in Bourdieu's sociological school at least since Bourdieu & Saint-Martin (1978).

I. Principles of MCA

Distance, cloud of individuals and cloud of modalities

The distance between two individuals is determined by their responses to the questions to which they give different answers. Suppose that for question q, individual i chooses modality k and individual i' a modality k' different from k; let n_k and $n_{k'}$ be the numbers of individuals who have chosen modalities k and k' respectively; the part of distance between individuals i and i' due to question q is defined by the formula $d^2_q(i,i') = 1/f_k + 1/f_{k'}$, where $f_k = n_k/n$ and $f_{k'} = n_{k'}/n$ (relative frequencies of k and k'). Then the overall distance $d(i,i')$ between i and i' is defined by the formula $d^2(i,i') = 1/Q \sum_{q \in Q} d^2_q(i,i')$. If Q is the number of questions and K the overall number of modalities, the distances between individuals determine the *cloud of individuals*, consisting of n points in a space with (at most) $K - Q$ dimensions .

The *cloud of modalities* follows, consisting of K points; if $n_{kk'}$ denotes the number of individuals who have chosen both modalities k and k', the distance $d(k,k')$ is given by the formula $d^2(k,k') = (n_k + n_{k'} - 2n_{kk'})/(n_k n_{k'}/n)$. (The numerator in the formula is the number of individuals who have chosen k or k' but not both; the denominator is the familiar theoretical frequency).

Both clouds have the same number of dimensions and the same overall variance.

Principal axes, eigenvalues and contributions

A cloud can be fitted by a one-dimensional cloud, by projecting it orthogonally onto a straight line. A line such that the variance of the projected cloud is maximal (unique in general) is called *the first principal axis* of the cloud, and the variance of the projected cloud is called the variance of the first axis, or first *eigenvalue,* denoted λ_1; in this sense the first principal axis

provides the best one-dimensional fit of the cloud. By looking for the best fit of the cloud by a two-dimensional cloud (plane), by a three-dimensional cloud, etc., one defines the sequence of principal axes, with decreasing eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$ The principal axes of the cloud of individuals are in a one-one correspondence with those of the cloud of modalities, and they have the same variances axis by axis (eigenvalues).
Contributions are the main *aid to interpretation* in Geometric Data Analysis. The proportion of variance of the axis due to a point is called the contribution of the point to the variance of axis. If y^k denotes the abscissa of modality k of weight f_k on the axis of variance λ, the contribution of modality k to axis is $\text{Ctr}_k = (f_k/Q)(y^k)^2/\lambda$.
Contributions add up by grouping; which allows calculating contributions of questions, and contributions of headings. For a standard MCA, the contribution of question q to the cloud is : $\text{Ctr}_q = (K_q - 1)/ (K - Q)$, K_q denoting the number of modalities of question q.

II. Steps of analysis

Choosing active questions and encoding modalities

The first and crucial step of MCA is the choice of *active questions*, that is, the questions that create distances between individuals. The next step is the encoding of the modalities of active questions. To achieve homogeneity among questions, the numbers of modalities within questions should not differ too much across questions. When the questions pertain to several themes or headings, the contributions of the various headings to the total variance should be kept to the same order of magnitude.
The smaller the frequencies of modalities of active questions, the more they create distance. This property tends to enhance the importance of infrequent modalities, which is a desirable property – up to a certain point. Rare modalities (say, of frequencies less that 5%) need to be pooled with others whenever feasible, or alternatively be put as "passive" ones (Specific MCA). Moreover, there may be modalities of active variables that one would like to discard (e.g. nonresponses, "junk modalities") while preserving the structural properties of MCA; then, in the *specific* MCA devised for this purpose, they can be put as *passive modalities* (Le Roux & Rouanet, 2004, p. 203).
Some variables can be introduced into the analysis without participating to the determination of axes; they are called *supplementary variables*, that is, their principal coordinates are computed, and their modalities plotted in the diagrams. In the same way, one can put some individuals as supplementary ones.

Interpreting axes

MCA software produces the following basic output: Eigenvalues λ_1, $\lambda_2 \ldots$; principal coordinates of modalities and of individuals; contributions (Ctr) of modalities and of individuals.
To appreciate the relative importance of axes, and retain an appropriate subspace for interpretation, the use of *modified rates* is recommended (Benzécri, 1992, p. 412, Le Roux & Rouanet, 2004, p. 209).
The interpretation of axes will be conducted in the cloud of modalities and based on the modalities whose contributions to axis exceed some threshold, such as the average contribution.

Exploring the cloud of individuals

Consider some modality k; the subset of individuals that have chosen that modality determines a subcloud of the cloud of individuals, whose mean point will be called the

modality mean-point denoted $\bar{k}$. For each axis, the coordinate of point $\bar{k}$ is equal to $\sqrt{\lambda}\, y^k$, where y^k is the coordinate of modality k in the space of modalities. This is a fundamental property of MCA that relates the two clouds of individuals and of modalities and that is preserved in specific MCA.

Putting a variable as a *structuring factor* allows studying not only the associated mean points, but also the subclouds induced by the variable.

Geometric summaries of subclouds in a principal plane are provided by the family of inertia ellipses, among which *concentration ellipses* are especially convenient (Le Roux & Rouanet, 2004, p. 97-99). The length of each half-axis of the concentration ellipse is twice the standard deviation of the subcloud along this direction; for a normally-shaped cloud, the concentration ellipse contains about 86% of the points of the cloud.