5.1 Introduction

Geometric data analysis (GDA)—a name suggested by Patrick Suppes in 1996—is the approach to multivariate statistics initiated by J.-P. Benzécri in the 1960s, known in French-speaking literature as *Analyse des Données* (Benzécri et al. 1973; Benzécri 1982b). Beyond the "leading case" of correspondence analysis (CA), GDA includes principal component analysis (PCA), recast as a GDA method, and multiple correspondence analysis (MCA), an outgrowth of CA. The three key ideas of GDA are *geometric modeling* (constructing Euclidean clouds), *formal approach* (abstract linear algebra; GDA is properly the formal geometric approach to multivariate statistics), and *inductive philosophy* (descriptive analysis comes prior to probabilistic modeling). In applications, there are the two principles of *homogeneity* and *exhaustiveness*.

To sum up, GDA is Benzécri's tradition of multivariate statistics, with the spectral theorem as the basic mathematical tool: "All in all, doing a data analysis, in good mathematics, is simply searching eigenvectors; all the science (or the art) of it is just to find the right matrix to diagonalize." (Benzécri et al. 1973: 289). This tradition extends the geometric approach far beyond the scaling of categorical data, a fact well perceived by Greenacre (1981: 122): "The geometric approach of the French school gives a much broader view of correspondence analysis, widening its field of application to other types of data matrices apart from contingency tables."

The present chapter, in line with the book by Le Roux and Rouanet (2004a), is rooted in Benzécri's tradition. It is devoted to individuals × variables tables—a basic data set in many research studies—by either PCA (numerical variables) or MCA (categorized ones). The distinction between numerical vs. categorized matters technically, but it is not essential methodologically. As Benzécri (2003: 7) states: "One should not say: 'Continuous numerical magnitude' ≃'quantitative data' vs. 'finite number of categories' ≃ 'qualitative data.' Indeed, at the level of a statistical individual, a numerical datum is not to be taken as a rule with its full accuracy but according to its meaningfulness; and from this point of view, there is no difference in nature between age and (say) profession."

The chapter is organized as follows. I describe PCA and MCA as GDA methods in Section 5.2. Then I introduce structuring factors and structured data analysis in Section 5.3. In Sections 5.4 and 5.5, I describe two analyses of structured individuals × variables tables, embedding ANOVA (analysis of variance) techniques into the geometric framework. Finally, concluding comments are offered in Section 5.6.

5.2 PCA and MCA as geometric methods

5.2.1 PCA: from multivariate analysis to GDA

To highlight geometric data analysis, PCA is a case in point on two counts: (a) PCA preexisted as an established multivariate analysis procedure and (b) as Benzécri (1992: 57) points out: "Unlike correspondence analysis, the various methods derived from principal component analysis assign clearly *asymmetrical roles* to the individuals and the variables."

Letting n denote the number of individuals and p the number of variables, the data table analyzed by PCA is an $n \times p$ table with numerical entries. The following excerpt by Kendall and Stuart (1973: 276) nicely describes the two spaces involved: "We may set up a Euclidean space of p dimensions, one for each variable, and regard each sample set … as determining a point in it, so that our sample consists of a swarm of n points; or we may set up a space of n dimensions, one for each observation, and consider each variable in it, so that the variation is described by p vectors (lying in a p-dimensional space embedded in an n-dimensional space)." In the following discussion, these spaces will be called the "space of individuals" and the "space of variables," respectively. In conventional multivariate analysis—see, for example, Kendall and Stuart (1973) and Anderson (1958)—the space of variables is the basic one; principal variables are sought as linear combinations of initial variables having the largest variances under specified constraints. On the other hand, in PCA recast as a GDA method, the basic space is that of individuals (see, for example, Lebart and Fénelon 1971).

In PCA as a GDA method, the steps of PCA are the following:

Step 1: The distance $d(i,i')$ between individuals i and i' is defined by a quadratic form of the difference between their description profiles, possibly allowing for different weights on variables; see, for example, Rouanet and Le Roux (1993) and Le Roux and Rouanet (2004a: 131).

Step 2: The principal axes of the cloud are determined (by orthogonal least squares), and a principal subspace is retained.

Step 3: The principal cloud of individuals is studied geometrically, exhibiting approximate distances between individuals.

Step 4: The geometric representation of variables follows, exhibiting approximate correlations between variables. Drawing the *circle of correlations* has become a tradition in PCA as a GDA method.

5.2.2 MCA: a GDA method

Categorized variables are variables defined by (or encoded into) a finite set of categories; the paradigm of the individuals × categorized variables table is the $I \times Q$ table of a questionnaire in standard format, where for each question q there is a set J_q of response categories—also called *modalities*—and each individual i chooses for each question q one and only one category in the set J_q. To apply the algorithm of CA to such tables, a preliminary coding is necessary. When each question q has two categories, one of them being distinguished as "presence of property q," CA can immediately be applied after *logical coding*: "0" (absence) vs. "1" (presence); in this procedure there is no symmetry between presence and absence. The concern for symmetry—often a methodologically desirable requirement—naturally led to the coding where each categorized variable q is replaced by J_q indicator variables, that is, (0,1) variables (also known as "dummy variables"), hence (letting $J = \sum_q J_q$) producing an $I \times J$ indicator matrix to which the basic CA algorithm is applied. In this procedure, all individuals are given equal weights. In the early 1970s in France, this variant of CA gradually became a standard for analyzing questionnaires. The phrase "analyse des correspondances multiples" appears for the first time in the paper by Lebart (1975), which is devoted to MCA as a method in its own right. Special MCA software was soon developed and published (see Lebart et al. 1977).

The steps for MCA parallel the ones for PCA described above.

Step 1: Given two individuals i and i' and a question q, if both individuals choose the same response category, the part of distance due to question q is zero; if individual i chooses category j and individual i' category $j' \neq j$, the part of (squared) distance due to question q is $d_q^2(i,i') = \frac{1}{f_j} + \frac{1}{f_{j'}}$, where f_j and $f_{j'}$ are the proportions of individuals choosing j and j', respectively. The overall distance $d(i,i')$ is then defined by $d^2(i,i') = \frac{1}{Q}\sum_q d_q^2(i,i')$(see Le Roux and Rouanet 2004a). Once the distance between individuals is defined, the cloud of individuals is determined.

Steps 2 and 3: These steps are the same as in PCA (above).

Step 4: The *cloud of categories* consists of J *category points*.

Remark 1

(i) Only disagreements create distance between individuals. (ii) The smaller the frequencies of disagreement categories, the greater is the

distance between individuals. Property (i) is essential; property (ii), which enhances infrequent categories, is desirable up to a certain point. Very infrequent categories of active questions need to be pooled with others; alternatively, one can attempt to put them as passive elements while managing to preserve the structural regularities of MCA; see the paper by Benali and Escofier (1987), reproduced in Escofier (2003), and the method of *specific* MCA in Le Roux (1999), Le Roux and Chiche (2004), and Le Roux and Rouanet (2004a: chap. 5).

Remark 2

There is a *fundamental property* relating the two clouds. Consider the subcloud of the individuals that have chosen category j and, hence, the mean point of this subcloud (*category mean point*); let $\bar{f}_s$ denote its sth principal coordinate (in the cloud of individuals) and g_s the sth principal coordinate of point j in the cloud of categories; then one has: $\bar{f}_s = \gamma_s \, g_s$, where γ_s denotes the sth singular value of the CA of the $I \times J$ table. This fundamental property follows from transition formulas; see Lebart et al. (1984: 94), Benzécri (1992: 410), and Le Roux and Rouanet (1998: 204). As a consequence, the derived cloud of category mean points is in a one-to-one correspondence with the cloud of category points, obtained by shrinkages by scale factors γ_s along the principal axes $s = 1, 2, \ldots, S$.

5.2.3 A strategy for analyzing individuals × variables tables

The same strategy can be applied to each table to be analyzed. The strategy is outlined in the following three phases (phrased in terms of MCA, to be adapted for PCA).

Phase 1: Construction of the individuals × variables table
- Conduct elementary statistical analyses and coding of data.
- Choose active and supplementary individuals, active and supplementary variables, and structuring factors (see Section 5.3).

Phase 2: Interpretation of axes
- Determine the eigenvalues, the principal coordinates, and the contributions of categories to axes, and then decide about how many axes to interpret.
- Interpret each of the retained axes by looking at important questions and important categories, using the contributions of categories.

- Draw diagrams in the cloud of categories, showing for each axis the most important categories, and then calculate the contributions of deviations (see Le Roux and Rouanet 1998).

Phase 3: Exploring the cloud of individuals

- Explore the cloud of individuals, in connection with the questions of interest.
- Proceed to a Euclidean classification of individuals, and then interpret this classification in the framework of the geometric space.

Each step of the strategy may be more or less elaborate, according to the questions of interest. As worked-out examples, see the "culture example" in Le Roux and Rouanet (2004a) and the data sets that are available on the Web site at http://www.math-info.univ-paris5.fr/~lerb.

5.2.4 Using GDA in survey research

In research studies, GDA (PCA or MCA) can be used to construct geometric models of individuals × variables tables. A typical instance is the analysis of questionnaires, when the set of questions is sufficiently broad and at the same time diversified enough to cover several themes of interest (among which some balance is managed), so as to lead to meaningful multidimensional representations.

In the social sciences, the work of Bourdieu and his school is exemplary of the "elective affinities" between the spatial conception of social space and geometric representations, described by Bourdieu and Saint-Martin (1978) and emphasized again by Bourdieu (2001: 70): "Those who know the principles of MCA will grasp the affinities between MCA and the thinking in terms of field."

For Bourdieu, MCA provides a representation of the two complementary faces of social space, namely the space of categories—in Bourdieu's words, the space of *properties*—and the space of individuals. Representing the two spaces has become a tradition in Bourdieu's sociology. In this connection, the point is made by Rouanet et al. (2000) that doing correspondence analyses is not enough to do "analyses à la Bourdieu," and that the following principles should be kept in mind:

1. *Representing individuals*: The interest of representing the cloud of individuals is obvious enough when the individuals are "known persons"; it is less apparent when individuals are anonymous, as in opinion surveys. When, however, there are factors structuring the individuals (education, age, etc.), the interest of depicting the individuals not as

an undifferentiated collection of points, but structured into subclouds, is soon realized and naturally leads to analyzing subclouds of individuals. As an example, see the study of the electorates in the French political space by Chiche et al. (2000).

2. *Uniting theory and methodology*: Once social spaces are constructed, the geometric model of data can lead to an *explanatory use* of GDA, bringing answers to the following two kinds of questions: How can individual positions in the social space be explained by structuring factors? How can individual positions, in turn, explain the position-takings of individuals about political or environmental issues, among others? As examples, see the studies of the French publishing space by Bourdieu (1999), of the field of French economists by Lebaron (2000, 2001), and of Norwegian society by Rosenlund (2000) and Hjellbrekke et al. (in press).

5.3 Structured data analysis

5.3.1 Structuring factors

The geometric analysis of an individuals × variables table brings out the relations between individuals and variables, but it does not take into account the structures with which the basic sets themselves may be equipped. By *structuring factors*, we mean descriptors of the two basic sets that do not serve to define the distance of the geometric space; and by *structured data*, we designate data tables whose basic sets are equipped with structuring factors. Clearly, structured data constitute the rule rather than the exception, leading to questions of interest that may be central to the study of the geometric model of data. Indeed, the set of statistical individuals, also known as units, may have to be built from basic structuring factors, a typical example being the subjects × treatments design, for which a statistical unit is defined as a pair (subject, treatment). Similarly, the set of variables may have to be built from basic structuring factors. I will exemplify such constructions for the individuals in the basketball study (see Section 5.4), and for the variables in the Education Program for Gifted Youth (EPGY) study (see Section 5.5).

In conventional statistics, there are techniques for handling structuring factors, such as analysis of variance (ANOVA)—including multivariate (MANOVA) extensions—and regression; yet, these techniques are not typically used in the framework of GDA. By *structured data analysis*