(2005)

To appear

in Greenacre & Blasius (Ed)

Multiple Correspondence Analysis

and Related Methods

Chapman & Hall

Chapter 5

The geometric analysis of structured individuals × variables tables

Henry Rouanet

5.1 Introduction

By *Geometric Data Analysis* (GDA) — a name suggested by Patrick Suppes in 1996 — is meant 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, 1982). 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 modelling* (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 modelling). 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, p. 289). This tradition extends the geometric approach far beyond the scaling of categorical data — a fact well perceived by Greenacre (1981, p. 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 the line of the book by Le Roux & Rouanet (2004), 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 numerical vs categorized matters technically, but is not essential methodologically. As Benzécri (2003, p.7) states: "One should not say: 'Continuous numerical magnitude' $\simeq$ 'quantitative data' vs 'finite number of categories' $\simeq$ '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 (§2). Then I introduce structuring factors and Structured Data Analysis (§3). Last, I describe two analyses of structured individuals × variables tables, embedding ANOVA techniques into the geometric framework (§4 and §5).

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: i) PCA preexisted as an established multivariate analysis procedure. ii) As Benzécri (1992, p. 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 a $n \times p$ table with numerical entries. The following excerpt by Kendall & Stuart (1976, p. 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 a n–dimensional space)." In what follows, these spaces will be called the "space of individuals" and the "space of variables", respectively. Now in conventional multivariate analysis — see, for example, Kendall & Stuart (1976), 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 & 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 on the difference between their description profiles, possibly allowing for different weights on variables; see e.g. Rouanet & Le Roux (1993), Le Roux & Rouanet (2004, p. 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 $J = \sum_q J_q$) 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 & Rouanet, 2004). Once the distance between individuals is defined, the cloud of individuals is determined.

Steps 2 and 3. They are the same as in PCA.

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 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 may attempt to put them as passive elements, while managing to preserve the structural regularities of MCA; see the paper by Benali & Escofier (1987), reproduced in Escofier (2003), and the method of *specific* MCA in Le Roux (1999), Le Roux & Chiche (2004), Le Roux & Rouanet (2004, chap. 5).

Remark 2. There is a *fundamental property* relating the two clouds. Consider the subcloud of the individuals that have chosen category j, hence the mean point of this subcloud (*category mean–point*); let $\overline{f}_s$ denote its s–th principal coordinate (in the cloud of individuals), and g_s the s–th principal coordinate of point j in the cloud of categories; then one has: $\overline{f}_s = \gamma_s \, g_s$, where γ_s denotes the s–th singular value of the CA of the $I \times J$ table. This fundamental property follows from transition formulas; see Lebart et al (1984, p. 94), Benzécri (1992, p. 410), Le Roux & Rouanet (1998, p.

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

For each table to analyze, the same strategy can be applied, with the following phases (phrased in terms of MCA, to be adapted for PCA).

Phase 1. Construction of the individuals × *variables table*

• Elementary statistical analyses and coding of data.

• Choice of active and supplementary individuals, of active and supplementary variables, and of structuring factors (see §3).

Phase 2. Interpretation of axes

• Determine the eigenvalues, the principal coordinates and the contributions of categories to axes, and 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 calculate the contributions of deviations (cf. Le Roux & 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; 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 & Rouanet (2004), and data sets at the web site `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 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 & Saint–Martin (1978) and emphasized again by Bourdieu (2001, p.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:

i) *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 (2002).

ii) *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 various issues such as political or environmental? 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 (2005).

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 aka 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* (§4), and for the variables in the Education Progam for Gifted Youth study, in short EPGY *study* (§5).

In conventional statistics, there are techniques for handling structuring factors, such as analysis of variance (ANOVA) — including MANOVA extensions — and regression; yet, these techniques are not typically used in the framework of GDA. By *structured data analysis* we mean the integration of such techniques into GDA, while preserving the GDA construction: see Le Roux & Rouanet (2004, chap 6).

5.3.2 From experimental to observational data

In the experimental paradigm, there is a clear distinction between *experimental factors*, or independent variables, and *dependent variables.* Statistical analysis aims at studying the *effects* of experimental factors on dependent variables. When there are several dependent variables, a GDA can be performed on them, and the resulting space takes on the status of a "geometric dependent variable".

Now turning to observational data, let us consider an educational study, for instance, where for each student, variables on various subject matters are used as active variables to "create distance" between students and so to construct an educational space. In addition to these variables, structuring factors, such as identification characteristics (gender, age, etc.) may have been recorded; their relevance is reflected in a question such as: "How are boys and girls scattered in the educational space?". Carrying over the experimental language, one may speak of the "effect of gender"; or one may prefer the more neutral language of *prediction*, hence the question: knowing the gender of a student ("predictor variable"), predict the position of this student in the space ("geometric variable to be predicted"). As another question of interest, suppose the results of students at some final exam are available. Taking this variable as a structuring factor on the set of individuals, one may ask: knowing the position of a student in the space ("geometric predictor"), predict the success of this student at the exam. The geometric space is now the predictor, and the structuring factor the variable to be predicted.

5.3.3 Supplementary variables vs structuring factors

As a matter of fact, there is a technique in GDA that handles structured data, namely that of *supplementary variables*; see Benzécri (1992, p. 662), Cazes (1982), Lebart et al (1984). Users of GDA have recognized for a long time that introducing supplementary variables amounts to doing regressions, and they have widely used this technique, both in a predictive and explanatory perspective. For instance, Bourdieu (1979), to build the lifestyle space of *La Distinction*, puts the age, father's profession, education level, and income as supplementary variables, to demonstrate that differences in lifestyle can be explained by those status variables.

The limitations of the technique of supplementary variables become apparent, however, when it is realized that in the space of individuals, considering a supplementary category amounts to confining attention to the mean point of a subcloud (fundamental property of MCA), ignoring the dispersion of the subcloud. Taking again *La Distinction*, this concern led Bourdieu, in his study of the upper class, to regard the fractions of this class — "the most powerful explanatory factor", as he puts it — as what we call a structuring factor in the cloud of individuals. In Diagrams 11 and 12 of *La Distinction*, the subclouds corresponding to the fractions of class are stylized as contours (for further discussion, see Rouanet et al, 2000).

To sum up, the extremely useful methodology of supplementary variables appears as a first step toward structured data analysis; a similar case can be made for the method of contributions of points and deviations developed in Le Roux & Rouanet (1998).

5.3.4 Breakdown of variances

In experimental data, the relationships between factors take the form of a *factorial design*, involving relations such as nesting and crossing of factors. In observational data, even in the absence of prior design, similar relationships between structuring factors may also be defined, as discussed in Le Roux & Rouanet (1998). As in genuine experimental data, the nesting and crossing relations generate effects of factors of the following types: main effects, between–effects, within–effects and interaction effects. In contrast, in observational data, the crossing relation is usually not orthogonal (as opposed to experimental data, where orthogonality is often ensured by design), that is, structuring factors are usually *correlated*. As a consequence, within–effects may differ from main effects. For instance, if social categories and education levels are correlated factors, the effect of the education factor within social categories may be smaller, or greater, or even reversed with respect to the main (overall) effect of the education factor ("structural effect").

Given a partition of a cloud of individuals, the mean points of the classes of the partition (category mean–points) define a derived cloud, whose variance is called the between–variance of the partition. The weighted average of the variances of subclouds is called the within–variance of the partition. The overall variance of the cloud decomposes itself additively in between–variance plus within–variance. Given a set of sources of variation and of principal axes, the *double breakdown of variances* consists in calculating the parts of variance of each source on each principal axis, in the line of Le Roux & Rouanet (1984). This useful technique will be used repeatedly in the applications presented in §4 and §5.

5.3.5 Concentration ellipses

Useful geometric summaries of clouds in a principal plane are provided by *ellipses of inertia*, in the first place *concentration ellipses*: see Cramér, (1946, p. 284). The concentration ellipse of a cloud is the ellipse of inertia such that a uniform distribution over the interior of the ellipse has the same variance as the cloud; this property leads to the ellipse with a half–axis along the s-th principal direction equal to $2\,\gamma_s$ (i.e. twice the corresponding singular value). For a normally–shaped cloud, the concentration ellipse contains about 86% of the points of the cloud. Concentration ellipses are especially useful for studying families of subclouds induced by a structuring factor or a clustering procedure; see e.g. the EPGY *study* in §5.

Remark. In statistical inference, under appropriate statistical modelling, the family of inertia ellipses also provides *confidence ellipses* for the true mean points of clouds. For large n, the half–axes of the $1-\alpha$ confidence ellipse are $\sqrt{\chi^2_\alpha/n}\,\gamma_s$ (χ^2 with 2 d.f.); for instance, for $\alpha = .05$, they are equal to $\sqrt{5.991/n}\,\gamma_s$, and the confidence ellipse can be obtained by shrinking the concentration ellipse by a factor equal to $1.22/\sqrt{n}$. Thus the same basic geometric construction yields descriptive summaries for subclouds and inductive summaries for mean points. For an application, see the *political space* study in Le Roux & Rouanet (2004, p.383 and 388).

In the next sections, structured individuals × variables tables are presented and analyzed in two research studies: 1) *basketball* (sport); 2) *EPGY* (education). PCA was used in the first study, MCA in the second one.

Other extensive analyses of structured individuals × variables tables, following the same overall strategy, will be found in the *Racism* study by Bonnet et al (1990), in the *political space* study by Chiche et al (2000), and in the *Norwegian field of power* study by Hjellbrekke et al (2005). The paper by Rouanet et al (2002) shows how regression techniques can be embedded into GDA.

5.4 The basketball study

In the study by Wolff et al (1998), the judgments of basketball experts were recorded in the activity of selecting high–level potential players. First, an experiment was set up, where video sequences were constructed covering typical game situations; eight young players (structuring factor P) were recorded in all sequences. These sequences were submitted to nine experts (structuring factor E); each expert expressed for each player free verbal judgments about the potentialities of this player. The compound factor $I = P \times E$, made of the $8 \times 9 = 72$ (player, expert) combinations, defines the set of statistical units, on which the expert judgments were recorded. Then, a content analysis of the 72 records was carried out, leading to construct 11 judgment variables about the following four aspects of performance, relating to upper body (four variables), lower body (four variables), global judgment (one variable), play strategy (two variables: attack and defense), respectively. The basic data set can be downloaded from the site `http://math-info.univ-paris5.fr/~rouanet`. Hereafter, we show how ANOVA techniques can be embedded into PCA.

Weights were allocated to the 11 standardized variables following expert's advice, namely 7, 3.5, 2.5 and 3 for the four aspects, respectively; hence a total weight of 16. Then, a weighted PCA was performed on the 72×11 table. The variance of the whole cloud is equal to 16 (sum of weights), with 11 non-zero eigenvalues, hence the average $\overline{\lambda} = 16/11 = 1.45$. Two eigenvalues ($\lambda_1 = 9.30$ and $\lambda_2 = 2.10$) were found exceeding average. Accordingly, two axes were interpreted: axis 1 was found to be related to "dexterity", and axis 2 to "strategy" (attack and defense).

Then from the basic cloud of 72 individual points, the cloud of the eight mean points of players (indexed by factor P) and the cloud of the nine mean points of experts (factor E) were derived, and the additive cloud, i.e. the fitted cloud without interaction was constructed. Table 5.1 shows the double breakdown of variances, for the first two axes, according to the three sources of variation: main effect of P, main effect of E, and interaction effect; it also shows the variance of the additive cloud ($P + E$).

The table shows the large individual differences among players, the over-

	Variances	
	Axis1	Axis2
$I = P \times E$	9.298	2.104
$Main\ P$ (Players)	8.913	1.731
$Main\ E$ (Experts)	0.051	0.124
$Interaction$	0.335	0.250
$P + E$	8.964	1.855

Table 5.1: *Basketball.* Double breakdown of variances according to players (P), experts (E), interaction, and additive cloud ($P + E$).

all homogeneity of experts, and the small interaction between players and experts. Figure 5.1 shows the basic cloud, structured by the players mean points, and Figure 5.2 the fitted additive cloud. In Figure 5.1, the observed interaction effects between players and experts are reflected in the various locations and distances of expert points with respect to player mean points. For instance, the point of expert #6 is on the right of $p1$ but on the left of $p6$, expert #4 is on the bottom side for most players, not for player $p8$, etc. On Figure 5.2, the pattern of experts points is the same for all players; the lines joining expert points (numbered arbitrarily from 1 to 9) exhibit the parallelism property of the additive cloud.

Remark. Structural effect and interaction effect are two different things. In the basketball study, the two factors players and experts are orthogonal, therefore there is no structural effect, that is, for each axis, the sum of the two variances of the main effects P and E is exactly the variance of the additive cloud; on the other hand, the observed interaction effect between the two factors is not exactly zero.

5.5 The EPGY study

The Education Program for Gifted Youth (EPGY) at Stanford University is a continuing project dedicated to developing and offering multimedia computer–based distance–learning courses in a large variety of subjects; for instance, in Mathematics, EPGY offers a complete sequence from kindergarten through advanced–undergraduate (see Tock & Suppes, 2002).

This case study, conducted by Brigitte Le Roux and Henry Rouanet in cooperation with Patrick Suppes, deals with the detailed performances of 533 students in the third grade in the course of Mathematics, with its five topics organized as *strands*, namely Integers, Fractions, Geometry, Logic and Measurement, and for each strand, performance indicators of three types, namely error rates, latencies (for correct answers) and numbers of exercises (to master the concepts of the strand). The overall objective was to construct a geometric space of data exhibiting the organization of individual differences among gifted students. A specific question of interest was to investigate the trade–off between errors and latencies. In this respect, the body of existing knowledge about "ordinary students" appears to be of limited relevance, so the case study is really exploratory. The detailed study can be found in Le Roux & Rouanet (2004, chapter 9) and on the Web–site `http://epgy.stanford.edu/research/GeometricDataAnalysis.pdf`.

5.5.1 Data and coding

In such a study, in order to analyze the individuals × variables table, the set of 533 students is naturally taken as the set of individuals, whereas the set of variables is built from the two structuring factors: the set S of the five strands and the set T of the three types of performance. Crossing the two factors yields the compound factor $S \times T$, which defines the set of $5 \times 3 = 15$ variables.

In the first and indispensable phase of elementary analyses and coding of variables, we examine in detail the variables and their distributions. For error rates, the distributions differ among the strands; they are strongly asymmetric for Integers, Fractions and Measurement, and more bell-shaped for Geometry and Logic. Latencies differ widely among strands. The number of exercises is a discrete variable. To cope with this heterogeneity, we choose MCA for constructing a geometric model of data. The coding of variables into small numbers of categories (2, 3 or 4) will aim to achieve as much homogeneity as possible, as required to define a distance between individuals.

For *error rates*, we start with a coding in three categories from 1 (low error rate) through 3 (high error rate), for each strand, hence 15 categories for the five strands; now with this coding, category 3 has a frequency less than 1% for Integers and Fractions, therefore we pool this category with category 2, resulting in 13 categories. For *latencies*, we take, for each strand, a 4–category coding, from 1 (short) through 4 (long); hence $4 \times 5 = 20$ categories. For *numbers of exercises*, we code in two categories for Integers, Fractions, and Measurement, and in three categories for Geometry and Logic, from 1 (small number) through 3 (large number); hence 12 categories. All in all, we get $13 + 20 + 12 = 45$ categories for 15 variables.

5.5.2 MCA and first interpretations

The basic results of MCA are the following: (i) the eigenvalues; (ii) the principal coordinates and the contributions of the 45 categories to axes (they are given in Le Roux & Rouanet, 2004, p. 400; see also the web–site); (iii) the principal coordinates of the 533 individuals; (iv) the geometric representations of the two clouds (categories and individuals).

Eigenvalues and modified rates. There are $J - Q = 45 - 15 = 30$ eigenvalues, and the sum of eigenvalues $(J - Q)/Q$ is equal to 2. How many axes to interpret? Letting $\lambda_m = 1/Q$ (mean eigenvalue, here .067) and $\lambda' = (\lambda - \lambda_m)^2$, we have calculated modified rates by Benzécri's formula $\lambda'/\sum \lambda'$, the sum being taken over the eigenvalues greater than λ_m (Benzécri, 1992, p.412). The modified rates indicate how the cloud deviates from a *spherical cloud* (with all eigenvalues equal to the mean eigenvalue). As an alternative to Benzécri's formula — which Greenacre (1993) claims to be too optimistic — we have also calculated modified rates using Greenacre's formula. See

Table 5.2. At any rate, one single axis is not sufficient. In what follows we will concentrate on the interpretation of the first two axes.

	Axis 1	Axis 2
Eigenvalues (λ)	.3061	.2184
Raw rates of inertia	15.3%	10.9%
Benzécri's modified rates	63.1%	25.4%
Grenacre's modified rates	55.5%	21.8%

Table 5.2: *EPGY*. First two axes. Eigenvalues; raw rates and modified rates

From the basic table of the contributions of the 45 categories to axes (not reproduced here), the contributions of the 15 variables can be obtained by adding up their separate contributions (Table 5.3). The more compact contributions of the three types of performance can be similarly derived, as well as the contributions of the five strands. See Table 5.3 (contributions greater than average are in bold).

	Ctr	Axis 1	Axis 2
	Integers	**.083**	.043
	Fraction	.051	.034
Error Rate	Geometry	**.104**	.035
	Logic	**.096**	.053
	Measuremt	**.098**	.018
Error rate	(Total)	**.433**	.184
	Integers	.048	**.159**
	Fraction	.043	**.157**
Latency	Geometry	.059	**.123**
	Logic	.066	**.131**
	Measuremt	.054	**.106**
Latency	(Total)	.271	**.677**
	Integers	.065	.022
	Fraction	.028	.028
Exercises	Geometry	.023	.023
	Logic	**.097**	.044
	Measuremt	**.084**	.022
Exercises	(Total)	.296	.139
Total		1.	1.

Table 5.3: *EPGY*. Contributions to axes of the 15 variables (in bold: contributions greater than $\frac{1}{15} = .067$), and of the three types of performance (in bold: contributions greater than 1/3).

Making use of the $S \times T$ structure, the cloud of 45 categories can be subdivided into subclouds. For instance, for each type of performance, we can construct and examine the corresponding subcloud. As an example, Figure 5.3 depicts the subcloud of the 12 numbers-of-exercises categories in plane 1-2; this figure shows for this type of performance the coherence between strands, except for geometry.

5.5.3 Interpretation of axes

Interpretation of axis 1 ($\lambda_1 = .3061$)

There are 20 categories whose contributions to axis 1 are greater than average ($1/Q = 1/45 = .022 = 2.2\%$); to which we will add the low error rate category for Logic; the 21 categories account for 81% of the variance of axis, on which we base the interpretation of axis 1. The opposition between high error rates (right of axis) and low error rates (left) accounts for 35% of the variance of axis 1 (out of the 43% accounted for by all error rate categories). The contributions of short latency categories for the five strands are greater than the average contribution. These five categories are located on the right of origin (cf. Figure 5.4), exhibiting the link between high error rates and short latencies. The opposition between low error rates and short latencies accounts for 28% of the variance of axis 1, and the one between small and large numbers of exercises for 24%. The opposition between the 7 categories on the left and the 14 ones on the right accounts for 67% of the variance of axis 1.

> *The first axis is the axis of error rates and numbers of exercises.* It opposes on one side low error rates and small numbers of exercises and on the other side high error rates and large numbers of exercises, the latter being associated with short latencies.

Interpretation of axis 2 ($\lambda_2 = .2184$)

Conducting the analysis in the same way, we look for the categories most contributing to the axis; we find 15 categories, that can be depicted again in plane 1-2, see Figure 5.5; the analysis leads to the conclusion:

> *The second axis is the axis of latencies.* It opposes short latencies and long latencies, the latter being associated with high error rates and large numbers of exercises.

5.5.4 Cloud of individuals

The cloud of individuals (533 students) is represented on Figure 5.6; it consists in 520 observed response patterns, to which we add the following four extreme response patterns: Pattern 11111 11111 11111 (point A) (low error rates, short latencies, small number of exercises); Pattern 11111 44444 11111 (point B) (low error rates, long latencies, small number of exercises); Pattern 22332 11111 22332 (point D) (high error rates, short latencies, large number of exercises); and Pattern 22332 44444 22332 (point C) (high error rates, long latencies, large number of exercises). None of the 533 individuals matches any one of these extreme patterns, which will be used as landmarks for the cloud of individuals.

The individuals are roughly scattered inside the quadrilateral ABCD, with a high density of points along the side AB and a low density along the opposed side. This shows there are many students who make few errors whatever their latencies. On the other hand, students with high error rates are less numerous and very dispersed.

In structured individuals × variables tables, the cloud of individuals enables one to go farther than the cloud of categories, for investigating compound factors. As an example, let us study, for the measurement strand, the crossing of error rates and latencies. There are $3 \times 4 = 12$ composite categories, and for each of them there is associated a subcloud of individuals, each one with its mean point. Figure 5.7 shows the 3×4 derived cloud of mean points. The profiles are approximately parallel. There are few individuals with high error rates (3) (dotted lines), whose mean points are close to side CD. As one goes down along the AB direction, latencies increase, while error rates remain about steady; as one goes down along the AD direction, error rates increase, while latencies remain about steady.

	Axis 1	Axis 2	Plane 1-2
Between (Age×Gender)	.0306	.0099	.0405
Age	.0301	.0067	.0368
Gender	.0000	.0012	.0012
Interaction	.0006	.0016	.0022
Within (Age×Gender)	.2683	.2055	.4738
Total variance ($n = 468$)	.2989	.2154	.5143

Table 5.4: *EPGY*. Double breakdown of variances for the crossing Age × Gender.

To illustrate the study of external factors in the cloud of individuals, we sketch the joint analysis of *age and gender*, allowing for missing data (57 and 46 respectively). Crossing age (in four classes) and gender (283 boys and 204 girls) generates $4 \times 2 = 8$ classes. A large dispersion is found within the 8 classes. In plane 1-2, the within variance is equal to .4738, and the between–variance only .0405: see Table 5.4. The variances of the two main effects and of the interaction between age and gender are also shown on this table. The crossing of age and gender is nearly orthogonal, therefore there is virtually no structural effect, that is, for each axis, the sum of the two main effect variances is close to the difference "between–variance of crossing minus interaction variance".

5.5.5 Euclidean classification

Distinguishing classes of gifted students was one major objective of the EPGY study. We have made a Euclidean classification, that is, an ascending hierarchical clustering with the inertia (Ward) criterion. The procedure first led to a 6–class partition, from which a partition into 5 classes (c1, c2, c3, c4, c5) was constructed and retained as a final partition. The between– and within–variances on the first two axes of this final partition are given in Table 5.5. A synopsis of this partition is presented on Table 5.6.

	Axis 1	Axis2
Between–Variance	.1964	.1277
Within–Variance	.1097	.0907

Table 5.5: *EPGY.* Between and within variances for the five–class partition.

	frequencies	Error rates	Latencies	Exercises
c1	111		short	small except in Geometry
c2	25	high		rather large
c3	142	high		
c4	178	low		rather small
c5	77		long	rather small

Table 5.6: *EPGY.* Synopsis of final five–class partition.

There are two compact classes of highly–performing students. One is class $c1$, close to point A, with short latencies and medium error rates; the other one is class c4, close to point B, with rather low error rates (especially in Geometry and in Logic) and medium to long latencies.

5.5.6 Conclusion of EPGY study

The geometric study shows the homogeneity of matters (strands), except geometry. It also shows how the differences among gifted students are articulated around two scales: that of error rates and numbers of exercises, and that of latencies. Within the geometric frame, differentiated clusters have been identified. The highly performing class c4, with low error rates and (comparatively) long latencies, is of special interest, in so far its profile is hardly reinforced by the current standards of educational testing!

5.6 Concluding comments

1. In Geometric Data Analysis, individuals×variables tables, beyond technical differences (PCA vs MCA), can be analyzed with a common strategy, with the joint study of the cloud of variables (or of categories) and the cloud of individuals.

2. Structured data arise whenever individuals or variables are described

by structuring factors. A cloud of individuals with structuring factors is no longer an undifferentiated set of points, it becomes a complex meaningful geometric object. The families of subclouds induced by structuring factors can be studied not only for their mean points but also for their dispersions. Concentration ellipses can be constructed, and various effects (between, within, interaction) investigated both geometrically and numerically.

Software Note

To carry out the analyses, we have used SPSS for the elementary descriptive statistics and the data codings. Then, programs in ADDAD format written by B. Le Roux, P. Bonnet and J. Chiche have been used for performing PCA and MCA. Finally, starting from principal coordinates calculated with ADDAD, the exploration of clouds and the double breakdowns of variance have been made with EyeLID. The EyeLID software, for the graphical investigation of multivariate data, was developed by Bernard, Baldy and Rouanet (1988), it combines two original features: a *Language for Interrogating Data* ("LID"), which designates relevant data sets in terms of structuring factors and constitutes a command language for derivations, and the *Visualization* ("Eye") of the clouds designated by EyeLID requests. For illustrations of the command language, see Bernard et al (1989) and Bonnet et al (1996). Concentration ellipses have been determined by the `ellipse` program by B. Le Roux and J. Chiche, which prepares a request file for the drawings done by the "freeware" WGNUPLOT.

An extensive (though limited in data size) version of ADDAD (Association pour le Développement et la Diffusion de l'Analyse des Données) and a DOS–version of EyeLID are available on the following ftp:

`ftp.math-info.univ-paris5.fr/pub/MathPsy/AGD`

ADDAD, EyeLID and `ellipse` programs can be downloaded from the Brigitte Le Roux's homepage:

`http://www.math-info.univ-paris5.fr/~lerb`

(under the "Logiciels" heading).

Acknowledgments

This chapter is part of an ongoing joint project on “the comparison of the French and Norwegian social spaces”, headed by B. Le Roux and O. Korsnes and funded by the Aurora-program (CNRS France and NC Norway). I thank those who have encouraged and helped me to write this chapter, especially my much regretted friend Werner Ackermann.

Figure 5.1: *Basketball.* Basic cloud, plane 1-2: $8 \times 9 = 72$ (player, expert) points, joined to the eight player mean points.

Figure 5.2: *Basketball.* Additive cloud, plane 1-2: $8 \times 9 = 72$ (player, expert) points, showing parallelism property.

Figure 5.3: *EPGY.* Space of categories, plane 1-2, subcloud of the 12 numbers-of-exercises categories. Integers I1 I2, fractions F1 F2, geometry G1 G2 G3, logic L1 L2 L3, measurement M1 M2. Sizes of markers reflect frequencies.

Figure 5.4: *EPGY.* Cloud of categories, plane 1-2. Interpretation of axis 1: 21 categories most contributing to axis. Sizes of markers reflect frequencies.

Figure 5.5: *EPGY.* Cloud of categories, plane 1-2. Interpretation of axis 2: 15 categories most contributing to axis. Sizes of markers reflect frequencies.

Figure 5.6: *EPGY.* Cloud of individuals with extreme (unobserved) patterns A, B, C, D.

Figure 5.7: *EPGY.* Space of individuals, plane 1-2. For the measurement strand: mean points of the 12 composite categories error×latencies. Example: Point 14 is the mean point of individuals with low error rates (1) and long latencies (4) for measurement.

Figure 5.8: *EPGY.* Final five–class partition (c).

About the Author

Henry Rouanet is a guest researcher at the Centre de Recherche Informatique de Paris 5 (CRIP5), Université René Descartes, Paris. His main interests in statistics are analysis of variance and Bayesian inference. His main fields of application are psychology and social sciences. He has coauthored several books about statistics and Geometric Data Analysis.

Address: UFR Math-Info, Université René Descartes, 45 rue des Saints-Pères, F-75270 Paris Cedex 06, France.

E-mail: `rouanet@math-info.univ-paris5.fr`

References

ANDERSON T.W. (1958). *An Introduction to Multivariate Statistical Analysis.* New York: Wiley.

BENALI H., ESCOFIER B. (1987) Stabilité de l'analyse des correspondances multiples en cas données manquantes et de modlités à faible effectif, *Revue de statistique appliquée*, 35, 41-51.

BENZÉCRI J-P. & Coll. (1973). *L'Analyse des Données. Vol. 1: Taxinomie. Vol. 2: Analyse des Correspondances.* Paris: Dunod.

BENZÉCRI J-P. (1982). *Histoire et Préhistoire de l'Analyse des Données.* Paris: Dunod.

BENZÉCRI J-P. (1992). *Correspondence Analysis Handbook*, New York: Dekker.

BENZÉCRI J-P. (2003). Qu'est-ce que l'Analyse des Données?[Text read by B. Le Roux at the Conference on correspondence analysis , Barcelona]

BERNARD J-M., BALDY R. & ROUANET H. (1988). The Language for Interrogating Data (LID). In *Data Analysis and Informatics* (Ed. Diday E.), 461-468.

BERNARD J-M., LE ROUX B., ROUANET H. & SCHILTZ M-A. (1989). L'analyse des données multidimensionnelles par le langage d'interrogation de données (LID): au delà de l'analyse des correspondances, *Bulletin de Méthodologie Sociologique*, 23, 3-46.

BONNET P., LE ROUX B. & LEMAINE G. (1996). Analyse géométrique des données: une enquête sur le racisme, *Mathématiques et Sciences Humaines*, 136, 5-24.

BOURDIEU P. (1979). *La Distinction: Critique Sociale du Jugement.* Paris: Editions de Minuit (English translation: *Distinction* (1984). Boston (MA): Harvard University Press).

BOURDIEU P. (1999). Une révolution conservatrice dans l'édition, *Actes de la Recherche en Sciences Sociales*, Vol. 126-127, 3-28.

BOURDIEU P. (2001). *Langage et pouvoir symbolique.* Paris: Fayard.

BOURDIEU P. & SAINT–MARTIN M. (1978). Le Patronat, *Actes de la Recherche en Sciences Sociales*, Vol. 20-21, 3-82.

CAZES P. (1982). Note sur les éléments supplémentaires en Analyse des Correspondances, *Les Cahiers de l'Analyse des Données*, 7, 9-23 & 133-154.

CHICHE J., LE ROUX B., PERRINEAU P. & ROUANET H. (2000). L'espace politique des électeurs français à la fin des années 1990, *Revue française de science politique*, 50, 463-487.

CRAMÉR H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

ESCOFIER B. (2003). *Analyse des correspondances: recherches au cœur de l'analyse des données.* Rennes: Presses universitaires de Rennes.

GREENACRE M. (1981). Practical correspondence analysis. In *Interpreting Multivariate Data* (Ed. V. Barnett), 119-146. Chichester: Wiley.

HJELLBREKKE J., LE ROUX B., KORSNES O., LEBARON F., ROSENLUND L., ROUANET H. (2005). The Norwegian field of power anno 2000, *European Societies* (to appear).

KENDALL M.G. & STUART A. (1973). *The Advanced Theory of Statistics*, Volume 2. London: Griffin.

LE ROUX B. (1999). Analyse spécifique d'un nuage euclidien: application à l'étude des questionnaires, *Mathématiques, Informatique et Sciences Humaines*, 146, 65-83.

LE ROUX B., CHICHE J. (2004). Specific multiple correspondence analysis, *Hellenike Tetradia Analysis Dedomenon*, Thessalonike, 4, 30-41.

LE ROUX B. & ROUANET H. (1984). L'analyse multidimensionnelle des données structurées, *Mathématiques et Sciences Humaines*, 85, 5-18.

LE ROUX B. & ROUANET H. (1998). Interpreting axes in Multiple Correspondence Analysis: Method of the contributions of points and deviations. In *Visualization of Categorical Data*, (Eds. Blasius J. & Greenacre M.), 197-220, San Diego: Academic Press.

LE ROUX B. & ROUANET H. (2004).*Geometric Data Analysis: from Correspondence Analysis to Structured Data*. Dordrecht: Kluwer.

LE ROUX B. & ROUANET H. (2004).*Individual differences in gifted students*, `http://epgy.stanford.edu/research/GeometricDataAnalysis.pdf`.

LEBARON F. (2000). *La croyance économique: les économistes entre science et politique*. Paris: Seuil.

LEBARON F. (2001). Economists and the economic order: The field of economists and the field of power in France, *European Societies*, 3 (1), 91-110.

LEBART L. (1975). L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples. *Consommation*, 2, 73-96.

LEBART L. & FÉNELON J-P. (1971). *Statistique et Informatique Appliquées.* Paris: Dunod.

LEBART L., MORINEAU A., TABARD N. (1977).*Techniques de la description statistique: méthodes et logiciels pour l'analyse des grands tableaux.* Paris: Dunod.

LEBART L., MORINEAU A., WARWICK K. (1984). *Multivariate Descriptive Statistical Analysis.* New-York: Wiley.

ROSENLUND L. (2000). Social structures and change; applying Pierre Bourdieu's approach and analytic framework. Working papers from Stavanger University College , 85/20000. *Dr. Philo thesis*, Stavanger, Norway.

ROUANET H., ACKERMANN W. & LE ROUX B. (2000). The geometric analysis of questionnaires: the Lesson of Bourdieu's La Distinction, *Bulletin de Méthodologie Sociologique*, 65, 5-18.

ROUANET H. & LE ROUX B. (1993). *Analyse des données multidimensionnelles.* Paris: Dunod.

ROUANET H., LEBARON F., LE HAY V., ACKERMANN W. & LE ROUX B. (2002). Régression et analyse géométrique des données: réflexions et suggestions, *Mathématiques & Sciences humaines*, 160, 13-45.

TOCK K. & SUPPES P. (2002). The High Dimensionality of Students' Individual Differences in Performance in EPGY's K6 Computer-Based Mathematics Curriculum, `http://epgy.stanford.edu/research`.

WOLFF M., ROUANET H. & GROSGEORGE B. (1998). Analyse d'une expertise professionnelle: l'évaluation des jeunes talents au basket–ball de haut niveau, *Le Travail Humain*, 61, 281-303.

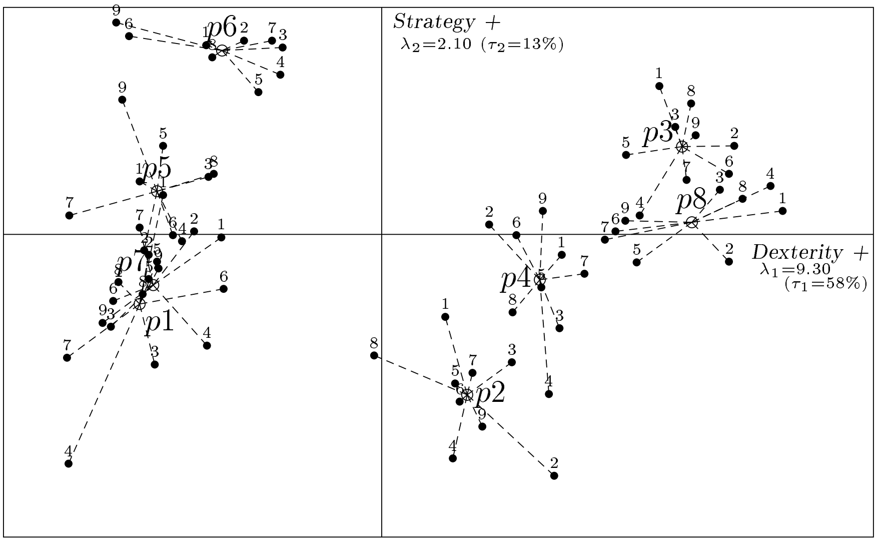

Figure 5.1: *Basketball.* Basic cloud, plane 1-2: 8 × 9 = 72 (player, expert) points, joined to the eight player mean points.

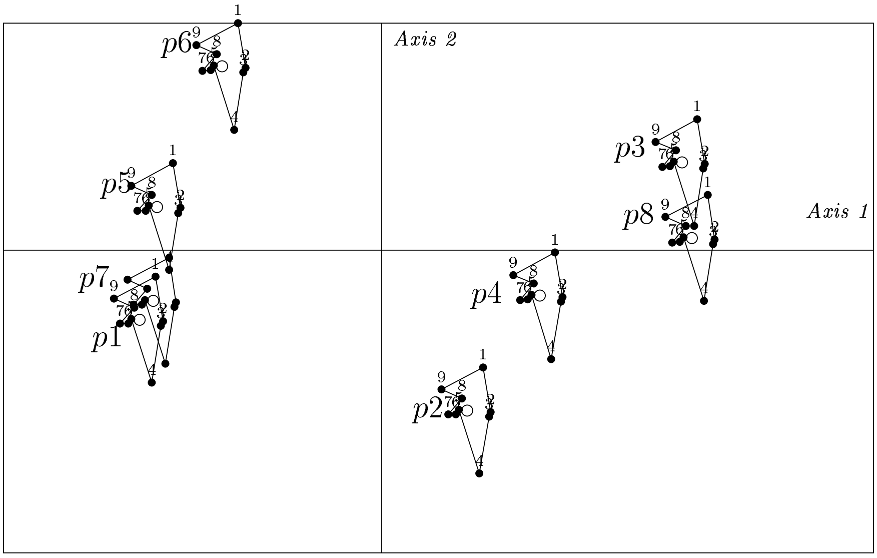

Figure 5.2: *Basketball.* Additive cloud, plane 1-2: 8 × 9 = 72 (player, expert) points, showing parallelism property.

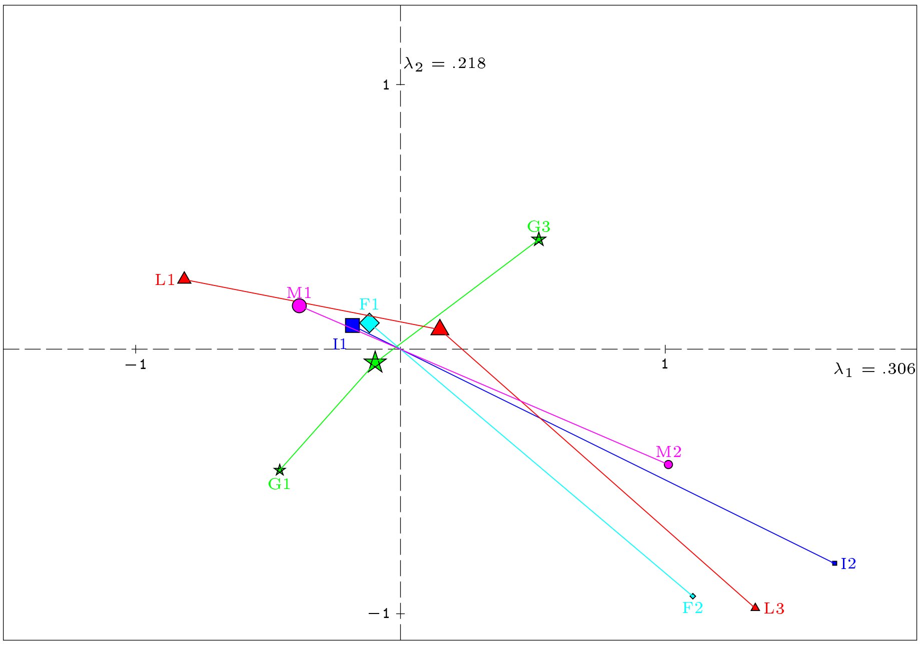

Figure 5.3: *EPGY*. Space of categories, plane 1-2, subcloud of the 12 numbers-of-exercises categories. Integers I1 I2, fractions F1 F2, geometry G1 G2 G3, logic L1 L2 L3, measurement M1 M2. Sizes of markers reflect frequencies.

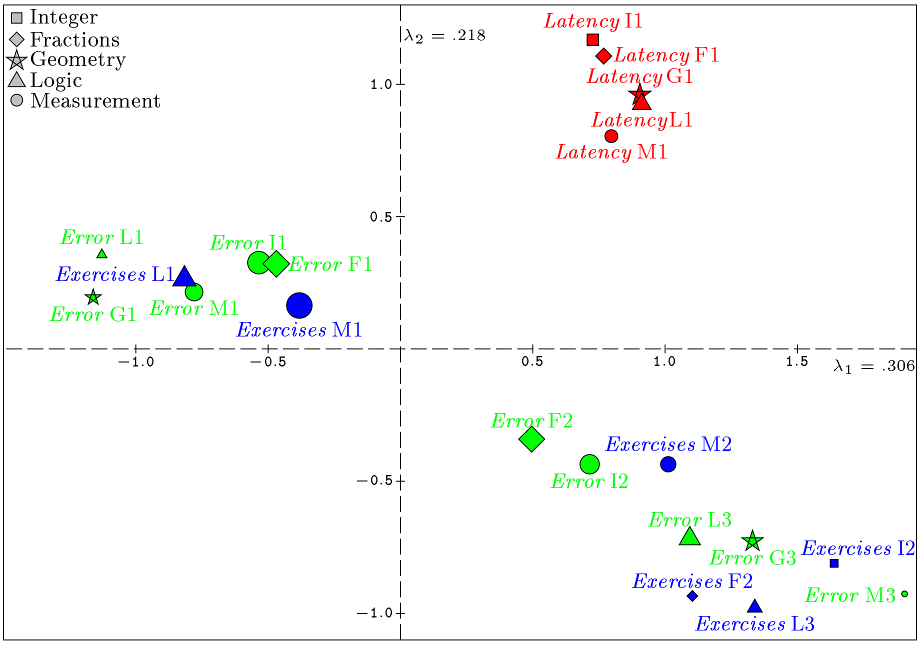

Figure 5.4: *EPGY*. Cloud of categories, plane 1-2. Interpretation of axis 1: 21 categories most contributing to axis. Sizes of markers reflect frequencies.

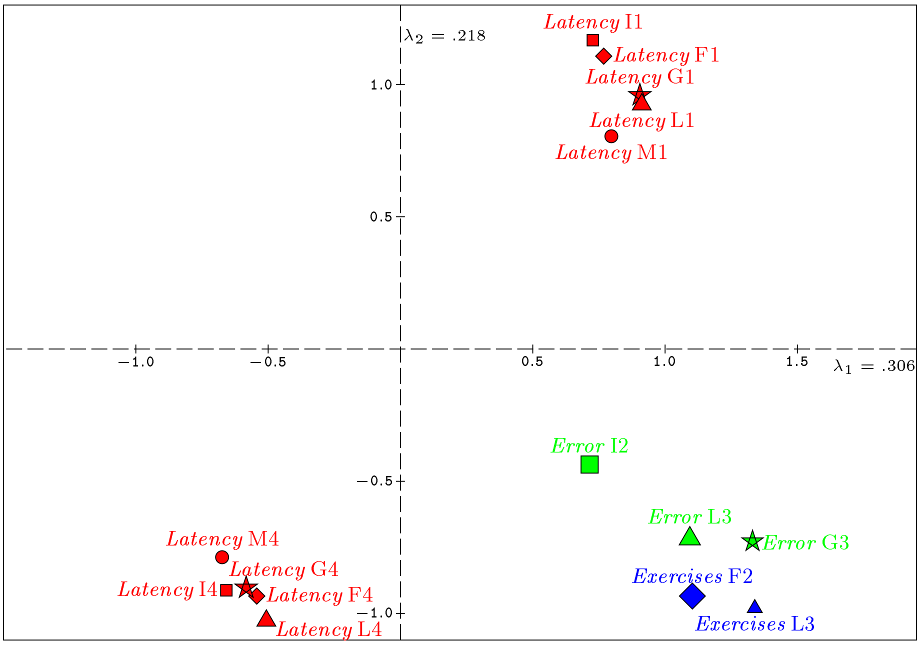

Figure 5.5: *EPGY*. Cloud of categories, plane 1-2. Interpretation of axis 2: 15 categories most contributing to axis. Sizes of markers reflect frequencies.

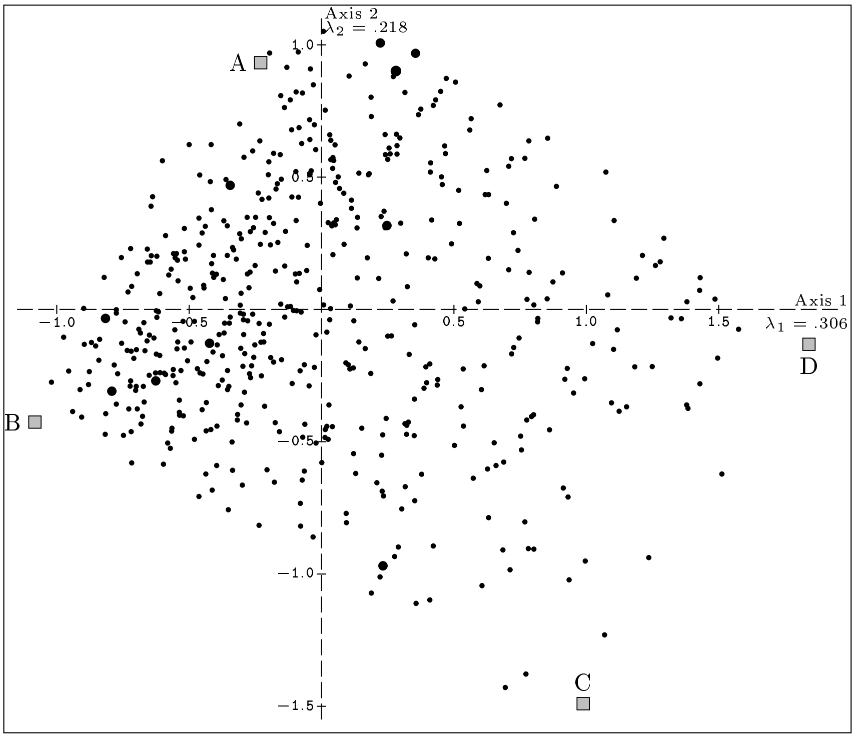

Figure 5.6: *EPGY.* Cloud of individuals with extreme (unobserved) patterns A, B, C, D.

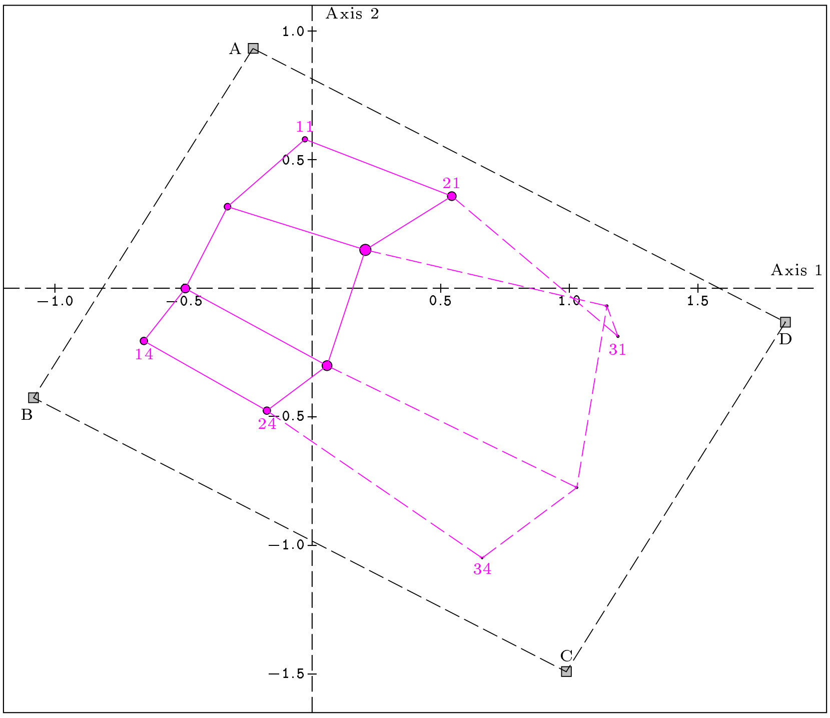

Figure 5.7: *EPGY*. Space of individuals, plane 1-2. For the measurement strand: mean points of the 12 composite categories error×latencies. Example: Point 14 is the mean point of individuals with low error rates (1) and long latencies (4) for measurement.

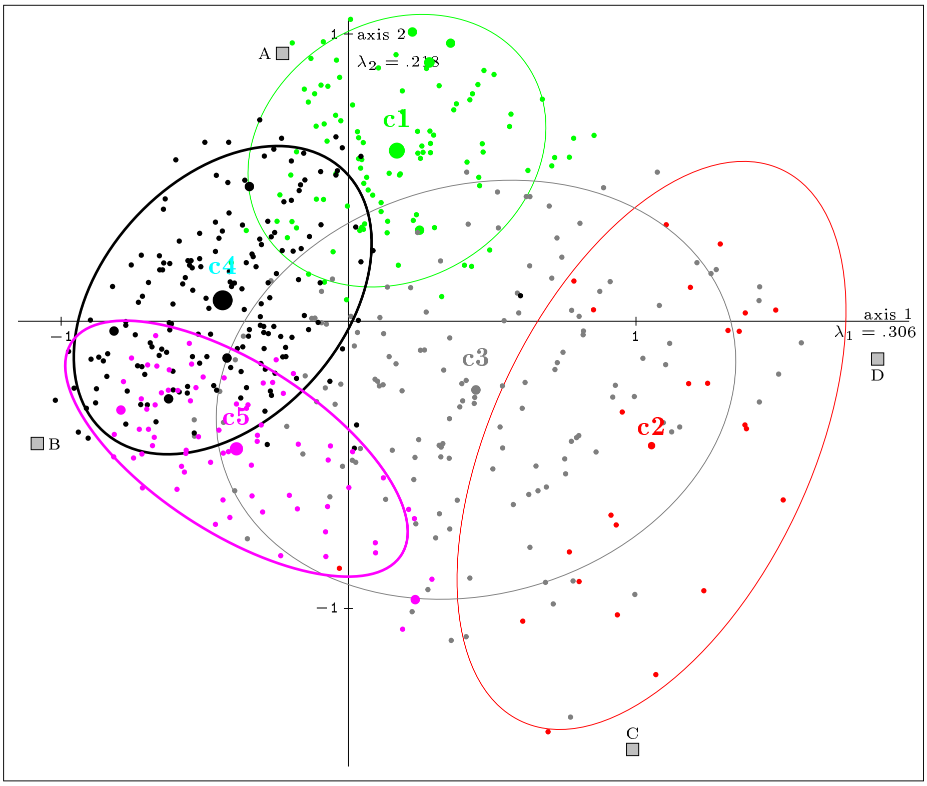

Figure 5.8: *EPGY*. Final five–class partition (c).