doi.org/10.26398/IJAS.0029-018

COMBINATORIAL TYPICALITY TEST IN GEOMETRIC DATA ANALYSIS

Solène Bienaise,
Coheris, *92150 Suresnes - France*

Brigitte Le Roux[1]
MAP5, UMR 8145, Université Paris Descartes (Sorbonne Paris Cité), France
Cevipof, UMR 7048, Sciences Po, Paris, France

Abstract. In the present article, we present a method of statistical inference for Geometric Data Analysis (GDA) that is not based on random modeling but on a combinatorial framework, that highlights the role of permutation tests. The method is applicable to any Individuals×Variables table, with structuring factors on indi-viduals, and numerical variables possibly produced by a GDA method. We develop procedures dealing with the typicality of a subcloud with respect to an overall cloud of individuals, which is the generalization of the test–values to the multidimen-sional case in a combinatorial framework. We outline the geometric interpretation of the observed p–value and study a compatibility zone (confidence zone). We pro-pose exact and approximate solutions. The method is applied to data from medical research on Parkinson's disease.

Keywords: *Geometric data analysis; Combinatorial inference; Permutation tests; Case study.*

1. INTRODUCTION

In the present article, we outline statistical inference procedures for Geometric Data Analysis (GDA) that are based on a combinatorial framework, and that highlight the role of permutation tests. The methods relate mainly to studying a *Euclidean cloud*, that is, a family of statistical observations conceptualized as points in a multidimensional space.

Permutation tests were initiated by Fisher (1935) and Pitman (1937), then further developed by Pesarin (2001), Edgington (2007), Good (2012) and others. Because permutation tests are computationally intensive, it took the advent of powerful computers to make them practical and, thus, it is only recently that they are really used. Permutation tests generally consist of three types: exact, resampling, and approximate tests. In an *exact test*, a suitable test statistic is computed on the observed data; the data

[1] Corresponding author: Brigitte LeRoux, email: Brigitte.LeRoux@mi.parisdescartes.fr

are "permuted" in all possible rearrangements of the objects considering the structure of the data, and the test statistic is computed for each rearrangement. The exact p–value is the proportion of rearrangements whose test statistic values are as extreme as or more extreme than the observed one. When the number of permutations is too large, we use a *Monte Carlo method* by selecting a random subset of all the possible rearrangements of the data and the p–value is calculated from the subset. Another alternative, for large data sets, consists in an approximate test by replacing the discrete distribution of the test statistic by a *classical distribution*.

A lot of statistical inference work —generally using assumptions like homoscedasticity, normality, random sampling, etc.— has been done not only in multivariate statistics in general, but also in Correspondence Analysis (CA)[2]. To name a few references: Lebart (1976), Gilula and Haberman (1986), Saporta and Hatabian (1986), Daudin et al. (1988), Gifi (1990), Le Roux and Rouanet (2004, 2010), etc., not to speak of the work done in traditional Multivariate Analysis and directly applicable to GDA, such as Anderson (1963) or Rao (1964).

The paper is organized as follows. Firstly, we introduce the typicality problem. Secondly, we deal with the exact test of typicality and the approximate one. Thirdly, we give some results for the unidimensional case. Finally we apply the method to a research case study, namely an experimental research on Parkinson's disease.

2. THE TYPICALITY PROBLEM

We will now present typicality situations and characterize the typicality problem.

Consider the following situations.

- **Committee**. Among the members of a club, a committee is appointed. Can the *committee* be declared to be *atypical of the club* with respect to the average age, the sex ratio, etc.?

- **Gifted children**. In a follow–up study on five gifted children, a psychologist found that for a certain task the mean grade of the group is 20, whereas, for a reference population of children of the same age, the mean is known to be 15 and the standard deviation 6.

[2] Benzécri has constantly insisted on the inductive logic embodied in CA; and in Benzécri & al (1973, Vol 2), there is an entire chapter (pp. 210-230) on inference in CA.

Is the psychologist entitled to claim that *the group of gifted children is, on average, superior to the reference group of children?*

- **Parkinsonian patients.** In an experiment on the gait of Parkinsonian patients, there are two groups of subjects: a reference group of healthy subjects and an experimental group of Parkinsonian patients observed after drug intake. For every subject, variables pertaining to gait were recorded. The question is:

  Is it possible to assimilate patients after drug intake to healthy subjects?

  This example will be discussed in detail in Section 5.

The preceding situations exemplify what we call *typicality situations*. In such a situation, there is a given *group of observations* and also a known *reference population*[3]. Some statistic is considered such as the mean of a variable of interest, the distance between mean points, etc. Intuitively, the problem can be formulated as follows:

"*Can the group of observations be assimilated to the reference population? Is it typical of it?*", or more specifically "*How can a typicality level be assessed for the group of observations with respect to the population, according to some statistic of interest?*"

In typicality situations, it is tempting to do a significance test. Yet often, the conventional statistical framework is not valid, since no randomness is assumed in the data generating the process. Even for a random sample, typicality may be raised as an issue that is perfectly distinct from randomness. In order to offer a solution to the typicality problem, the basic idea is to compare the group of the observations to the samples of the reference population, where samples are simply defined as *subsets* of the reference population.

In this paper, starting with a data cloud (descriptive phase), we proceed to the inference phase, dealing with the problem of *typicality* of clouds[4]. The methods are applicable to any cloud constructed from an Individuals×Variables table with structuring factors on individuals (Le Roux, 2014b); the variables can be numerical, or provided by any GDA method (Principal

[3] In the context of finite sampling, population will always refer to a finite set of statistical individuals.

[4] Most of the results of this paper are drawn from the PhD dissertation by S. Bienaise (2013, Chapter 3); see also Le Roux et al. (2018, Chapter 3)

Component Analysis, Correspondence Analysis, Multiple Correspondence Analysis, etc.).

3. TEST OF TYPICALITY

Let us consider a set I of N individuals constituting the reference population and a set C of n $(1 < n < N)$ individuals constituting a group of observations (C can be a subset of I or not). The set I is represented by a cloud of N points in a multidimensional Euclidean space, called *reference cloud* and denoted M^I, whose mean point is denoted O. Similarly, the set C is represented by a cloud of n points, denoted M^C, whose mean point is C.

Actually, the study will be done in the affine support of the cloud or in a subspace of the affine support of the cloud. For instance, in GDA we will often study the cloud in the subspace of the first principal axes.

Without loss of generality, we will suppose that the affine support is referred to an orthonormal Cartesian frame with origin–point O (the mean point of the reference cloud). In this frame, the covariance structure of the reference cloud is simply defined by its covariance matrix, which we denote $\mathbf{V}$, and which is invertible.

3.1. PRINCIPLE OF THE EXACT TEST

To answer the question, in a combinatorial framework, we construct the typicality test as follows.

1. *Sample set.* A sample is defined as an n–element subset of the reference population ("samples" in a purely set–theoretic sense). The set of all $\binom{N}{n}$ n–element subsets defines a *sample set.* Let J be the set indexing the samples and I_j the subset of the n individuals of sample j.

2. *Sample cloud.* A subcloud $(\mathrm{M}^i)_{i\in I_j}$ is associated with each sample j; its mean point is denoted H^j, with $\mathrm{H}^j = \sum_{i\in I_j} \mathrm{M}^i/n$. The cloud of the mean points of the $\binom{N}{n}$ subclouds is called the *sample cloud* and denoted $\mathrm{H}^J = (\mathrm{H}^j)_{j\in J}$ (see Figure 3, for the Parkinson Study).

3. *Test statistic.* Then we choose a *test statistic* that is an index of magnitude of the deviation between the points of the sample cloud and the mean point of the reference cloud, namely the origin–point

O. This index —as is commonly done in multivariate analysis— is the Mahalanobis norm[5] of deviations with respect to the covariance structure of the reference cloud; it will be denoted D^2.

Let $\mathbf{y}_j$ be the column–vector of the coordinates of point H^j, the squared Mahalanobis distance between the point H^j of the sample cloud and the mean point O of the reference cloud is equal to $\mathbf{y}_j^\top \mathbf{V}^{-1} \mathbf{y}_j$ and termed $D^2(j)$.

We denote D^2_{obs} the value of this statistic for the group of observations, that is, the Mahalanobis distance between the mean point C of the cloud M^C and the mean point O of the reference cloud M^I: one has $D^2_{\mathrm{obs}} = \mathbf{y}_{\mathrm{C}}^\top \mathbf{V}^{-1} \mathbf{y}_{\mathrm{C}}$ where $\mathbf{y}_{\mathrm{C}}$ denotes the column–vector of coordinates of point C.

4. *The p–value of the test.* The proportion of samples $j \in J$ such that $D^2(j) \geq D^2_{\mathrm{obs}}$ defines the p–value of the test.

5. *Conclusion.* We state the conclusion of the test:

   If the p–value is less than or equal to a conventional level α, the deviation from point C to point O will be said to be *statistically significant* (in a combinatorial sense) at level α.

   If the p–value is greater than α, the deviation from point C to point O will be said to be *not significant* at level α; points C and O will be said to be *compatible* at level α.

The p–value can be taken as defining the *level of typicality* of the group of observations with respect to the reference population, for the *Mean Point*: the smaller the value of p, the lower the typicality.

Properties of the sample cloud. The two following properties[6] are the generalization to multidimensional data of the ones of the classical theory of sampling in a univariate finite population.

Property 3.1. *The mean point of the sample cloud is the mean point of the reference cloud.*

---

[5] In multivariate analysis, it is usual to consider the statistic $\mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}$, where $\mathbf{S} = \frac{N}{N-1}\mathbf{V}$ is the covariance matrix corrected by the number of degrees of freedom.

[6] Proofs of these sampling properties for a cloud are given in Bienaise (2013, pp. 52-53).

Property 3.2. *The covariance matrix of the sample cloud, denoted* $\mathbf{W}$*, is proportional to that of the reference cloud:*

$$\mathbf{W} = \frac{1}{n} \times \frac{N-n}{N-1}\mathbf{V}$$

Property 3.3. *The squared Mahalanobis distance between every point* H^j *of the sample cloud and point* O *is less than or equal to* $\frac{N-n}{n}$*.*

$$\forall j \in J,\ D^2(j) \leq \tfrac{N-n}{n}$$

Proof. Given $j \in J$, let $\mathcal{D}$ be the line going through points O and H^j. If $D^2_{\mathcal{D}}(i)$ denotes the squared Mahalanobis distance between point O and the projection of point M^i on line $\mathcal{D}$, one has $\sum_{i\in I} D^2_{\mathcal{D}}(i)/N = 1$, since the "Mahalanobis variance" of cloud M^I is equal to 1 along all the directions of the affine support of the cloud. On line $\mathcal{D}$, the between–variance of the partition of cloud M^I into two subclouds $(\mathrm{M}^i)_{i\in I_j}$ (whose mean point is H^j) and $(\mathrm{M}^i)_{i\notin I_j}$ is equal to $\frac{n}{N-n}D^2(j)$. The within–between decomposition of variance leads us to $\sum_{i\in I} D^2(i)/N = \frac{n}{N-n}D^2(j) +$ within–variance, hence $\frac{n}{N-n}D^2(j) \leq 1$.

Geometric interpretation of the p**–value.** Recall that (see Cramér, 1946, p. 300) the principal κ–hyperellipsoid of a cloud is defined as the set of points such that the Mahalanobis distance to its mean point is equal to κ. Hence, the p–value of the typicality test can be interpreted as the proportion of points of the sample cloud H^J located outside or on the hyperellipsoid of the reference cloud going through point C (see Figure 3).

3.2. COMPATIBILITY REGION

The typicality test can be applied to every sample of the reference population viewed as a particular group of observations. For any specified α, the test will separate out those samples that are atypical at the α–level. The *fundamental typicality property* states that the proportion of these samples is at most α — "at most" rather than "equal to", owing to the discreteness of the sampling distribution of D^2.

Now our concern is to find a region of compatibility around point C. For this purpose, given any point P in the space, we consider the cloud P^I, which is the image of cloud M^I by the translation of vector $u = \mathrm{P} - \mathrm{O}$ (deviation from point O to point P). Cloud P^I has the same covariance structure as the cloud M^I, then the Mahalanobis distances attached to

clouds M^I and P^I are the same. Taking now cloud P^I as a reference cloud, we construct the sample cloud $\mathrm{H}_{\mathrm{P}}^J$, which is the translation of cloud H^J by vector $u = \mathrm{P} - \mathrm{O}$. The mean point of this cloud is P, its covariance matrix is $\mathbf{W}$ and $\forall j \in J$, $D_{\mathrm{P}}^2(j) = D^2(j)$ (since the translation implies that $\mathrm{H}_{\mathrm{P}}^j - \mathrm{P} = \mathrm{H}^j - \mathrm{O}$).

Considering every cloud P^I as a reference cloud, we say that point P is *compatible with point* C *at level* α if the corresponding p–value is greater than α. Hence the following definition.

Definition 3.1 (Compatibility region)**.** *The* $1-\alpha$ *compatibility region is the set of points* P *for which the proportion of samples* $j \in J$ *such that* $D_{\mathrm{P}}^2(j) \geq D_{\mathrm{P}}^2(\mathrm{C})$ *is greater than* α*.*

Remark. From the property $D_{\mathrm{P}}^2(j) = D^2(j)$, one deduces that the proportions of points $\mathrm{H}_{\mathrm{P}}^j$ whose squared distance $D_{\mathrm{P}}^2(j)$ (distance to point P) is greater than $D_{\mathrm{P}}^2(\mathrm{C})$ is equal to the proportion of points H^j whose squared distance $D^2(j)$ (distance to point O) is greater than $D_{\mathrm{P}}^2(\mathrm{C})$, with $D_{\mathrm{P}}^2(\mathrm{C}) = (\mathbf{y}_{\mathrm{C}} - \mathbf{y}_{\mathrm{P}})^\top \mathbf{V}^{-1}(\mathbf{y}_{\mathrm{C}} - \mathbf{y}_{\mathrm{P}})$, $\mathbf{y}_{\mathrm{P}}$ being the column–vector of coordinates of point P.

Lemma 3.1. *Given two points* P *and* Q *at the same Mahalanobis distance* κ *from point* C*, the proportion of samples* $j \in J$ *such that* $D_{\mathrm{P}}^2(j) \geq D_{\mathrm{P}}^2(\mathrm{C})$ *is equal to the proportion of samples* $j \in J$ *such that* $D_{\mathrm{Q}}^2(j) \geq D_{\mathrm{Q}}^2(\mathrm{C})$*.*

Proof. The translation of the κ–hyperellipsoid of the reference cloud M^I such that its center is point C writes $(\mathbf{y} - \mathbf{y}_{\mathrm{C}})^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{y}_{\mathrm{C}}) = \kappa^2$. Points P and Q belonging to this κ–hyperellipsoid verify $D_{\mathrm{P}}^2(\mathrm{C}) = D_{\mathrm{Q}}^2(\mathrm{C}) = \kappa^2$. The proportion of points $\mathrm{H}_{\mathrm{Q}}^j$ such that $D_{\mathrm{Q}}^2(j) \geq \kappa^2$ is equal to that of points $\mathrm{H}_{\mathrm{P}}^j$ such that $D_{\mathrm{P}}^2(j) \geq \kappa^2$, since, by construction, $\forall j \in J$, $D_{\mathrm{P}}^2(j) = D^2(j) = D_{\mathrm{Q}}^2(j)$.

From this lemma, we deduce the following property.

Property 3.4. *Let* κ_α^2 *be the maximum value of* $(D^2(j))_{j \in J}$ *for which the proportion of* $j \in J$ *verifying* $D^2(j) \geq \kappa_\alpha^2$ *is greater than* α*. The* $1-\alpha$ *compatibility region for point* C *is defined as the set of points* P *whose Mahalanobis distance to point* C *is less than or equal to* κ_α*.*

Geometrically, the $1-\alpha$ compatibility region for point C is the set of points inside or on the κ_α–hyperellipsoid of cloud M^I.

3.3. APPROXIMATE TEST

Let r denote the dimension of cloud M^I, that is, of its affine support. When N and n are both large, the sample cloud H^J (with $\binom{N}{n}$ points) can be fitted by an r–dimensional normal distribution whose center is the origin–point O and whose covariance matrix is $\mathbf{W} = \frac{1}{n} \times \frac{N-n}{N-1} \times \mathbf{V}$. Therefore the distribution of $\mathbf{y}^\top \mathbf{W}^{-1} \mathbf{y}$ is a χ^2 distribution with r degrees of freedom, denoted χ^2_r and the distribution of the test statistic $n \times \frac{N-1}{N-n} \times D^2$ can be approximated by χ^2_r.

- The combinatorial p–value can be approximated by

$$\widetilde{p} = p(\chi^2_r \geq n \times \tfrac{N-1}{N-n} \times D^2_{\mathrm{obs}})$$

- Let us denote $\chi^2_r[\alpha]$ the critical value of the chi-square distribution with r d.f., at level α, that is, $p(\chi^2_r \geq \chi^2_r[\alpha]) = \alpha$. The $1-\alpha$ approximate compatibility region is the set of points P for which the Mahalanobis distance to point G is less than $\widetilde{\kappa}_\alpha$, with $\widetilde{\kappa}^2_\alpha = \frac{1}{n} \times \frac{N-n}{N-1} \times \chi^2_r[\alpha]$.

  Remark. Nowadays it is possible to perform *exact combinatorial tests* using the enumeration of all samples as soon as their number is not too large (say $< 1,000,000$). Most often, the cardinal of the sample space is too large to enumerate all its points. According to many authors, we can inspect this sample space by using a *Monte Carlo method*, that is, by resampling algorithm from the sample space[7]. In this case we also use a hat, as in $\widehat{p}$, to indicate an estimate when referring to a Monte Carlo estimation. In any case, the *approximate test* provides an order of magnitude of the p–value and of the compatibility region.

4. PARTICULAR CASE: UNIDIMENSIONAL CLOUD

The preceding test can be applied as is to a unidimensional cloud but it leads to a two–sided test. By choosing another test statistic, namely the difference of means, we will be able to conclude by taking into account the sign of the difference (Le Roux, 1998).

The pertinent data are:

1. a reference population I of size N with which is associated a numerical variable $x^I = (x^i)_{i \in I}$ with mean $\overline{x}$ and variance v;

[7] Because of increasing computing power, by 2010, p–values based on exact enumeration sometimes exceeded 10,000,000 samples. A number of 1,000,000 resampling is not only recommended but common (Johnston et al., 2007).

2. a group of observations of size n whose *observed mean* is denoted m_{obs}.

Exact p–value. The steps of the test are the following ones:

1. Construct the sample space, that is, the set J of all $\binom{N}{n}$ n–elements subset of I: $(I_j)_{j\in J}$;

2. Choose a test statistic, for instance the mean, and associate with each sample j the value of the statistic, here the mean $m^j = \sum_{i\in I_j} x^i/n$.

3. If $m_{obs} > \overline{x}$, consider the samples whose means are greater than or equal to m_{obs}; the proportion of these samples defines the *observed upper level* (one–sided p–value) of the test.

This proportion will be taken as defining the level of extremality (or level of typicality) for the mean, with respect to the reference population (on the positive side). The smaller the value of p, the lower the typicality. For any conventional level α, if $p \leq \alpha/2$ the group of observations is declared to be atypical (on the positive side) of the reference population with respect to the mean at level $\alpha/2$.

When $m_{obs} < \overline{x}$, the *observed lower level* will be considered similarly and the typicality level defined accordingly.

Compatibility interval. As for a multidimensional cloud, given $a \in \mathbb{R}$, we consider the "shifted cloud" associated with the variable $y^I = (x^i + a)_{i\in I}$ whose mean is $\overline{y} = \overline{x} + a$ and whose variance is v.

Given α, the mean of y^I is said to be compatible with m_{obs} at level α if it is not atypical. The $1-\alpha$ compatibility interval is defined as the set of means $\overline{y}$ compatible with m. It is easily shown that it is equal to

$$[m_{obs} - (\overline{m}_\alpha - \overline{x})\ ;\ m_{obs} - (\underline{m}_\alpha - \overline{x})]$$

where $\overline{m}_\alpha$ (resp. $\underline{m}_\alpha$) is the greater (resp lower) value m^j for which the proportion of samples j verifying $m^j \geq \overline{m}_\alpha$ (resp. $m^j \leq \underline{m}_\alpha$) is greater than $\alpha/2$.

Approximate p–value. After the classical theory of sampling in a finite population, the mean of the statistic *Mean* is equal to $\overline{x}$ and its variance to $\frac{N-n}{N-1} \times \frac{v}{n}$ (see Cramér, 1946, p. 523). Consider now the scaled deviation

between means. We denote this statistic by Z and its observed value by $z_{obs} = \frac{m_{obs} - \overline{x}}{\sqrt{\frac{N-n}{N-1} \times \frac{v}{n}}}$. Both test statistics (i.e., *Mean* and *Scaled deviation*) lead to the same p–values: they produce equivalent tests.

When n and $N-n$ are both large, the *Mean* is approximately normal so that the variable Z is approximately normal $\mathcal{N}(0,1)$. Then, for $z_{obs} > 0$, the approximate p–value is equal to the proportion of the normal distribution greater than z_{obs} (observed upper level).
Remark: z_{obs} is also called *test–value* (see Lebart et al., 2006). The observed value D^2_{obs} is its multivariate generalization.

Approximate compatibility interval. If Z is distributed $\mathcal{N}(0,1)$ and if $z[\alpha]$ is such that $p(|Z| > z[\alpha]) = \alpha$, the $1-\alpha$ *approximate compatibility interval* is:

$$\left[m - z[\alpha] \sqrt{\tfrac{N-n}{N-1} \times \tfrac{v}{n}} \; ; \; m + z[\alpha] \sqrt{\tfrac{N-n}{N-1} \times \tfrac{v}{n}} \right]$$

Particular case of Multiple Correspondence Analysis. Let us take as a "reference population" the projected cloud of the N (active) individuals onto principal axis ℓ (with mean 0 and variance λ_ℓ), and as a group of observations a set of individuals of size n_k associated with a category k. The mean of coordinates of individuals of the group on axis ℓ is denoted $\overline{y}^k_\ell$ and the coordinate of the category–point k is denoted y^k_ℓ, with $y^k_\ell = \overline{y}^k_\ell / \sqrt{\lambda_\ell}$ (see Le Roux and Rouanet, 2010, p. 45).

To study the typicality of class k (for the mean), we can take as a test statistic the scaled deviation Z whose observed value is (see Le Roux, 1998):

$$z_{obs} = (\overline{y}^k_\ell / \sqrt{\lambda_\ell}) \sqrt{n_k \times \tfrac{N-1}{N-n_k}} = \sqrt{N-1}\, y^k_\ell \sqrt{\tfrac{f_k}{1-f_k}} = \sqrt{N-1} \cos\theta_{k\ell}$$

where $\cos\theta_{k\ell}$ is equal to the coordinate of k on axis ℓ, namely y^k_ℓ, divided by the distance between the category–point and the origin, namely $\sqrt{(f_k/(1-f_k))}$; the square of $\cos\theta_{k\ell}$ is the quality of representation of category k on axis ℓ.

5. THE PARKINSON'S STUDY

To present our strategy we will use a study about Parkinsonian patients' gait. The design of the experiment was as follows (Ferrandez and Blin,

1991): there were two groups of subjects, namely a reference group of 45 healthy subjects, observed once, and an experimental group of 15 Parkinsonian patients observed twice, before and after drug intake (L-Dopa). Six numerical variables pertaining to gait were recorded, namely Velocity, Length of Stride, Swing Duration, Stride Duration, Stance Duration, and Double Support Duration.

An extensive geometric descriptive analysis was done by Le Roux (2014a, pp.347-361). In the present paper, the descriptive analysis will be summarized, and we will concentrate on inferential analyses for comparing patients after drug intake to healthy subjects.

5.1. CLOUDS CONSTRUCTED BY PRINCIPAL COMPONENT ANALYSIS

A standard Principal Component Analysis (PCA)[8] is performed on healthy subjects' data, putting the patients' data as supplementary elements. The PCA reveals that the cloud is nearly two–dimensional reflecting dependencies in the definition of variables as well as in the data themselves. The first two axes account for 97% of the variance of the cloud. In addition, the quality of representation of healthy subjects in plane 1-2 is quite good: it exceeds .84 for 38 subjects, and goes below .50 for only three of them (who are near the mean point) and that of patients is very high (all are above 0.88). Therefore, the projections of clouds on the first principal plane will make up the basic data set. The two–dimensional representation of the cloud of the 45 healthy subjects (gray points), together with that of the 15 patients after drug intake (black points), are shown in Figure 1. The axes are the principal axes of the cloud of the healthy subjects, and the origin–point is its mean point (denoted O). The variances of the first two principal axes (eigenvalues) are $\lambda_1 = 3.9927$ and $\lambda_2 = 1.8224$.

The first principal axis can be interpreted as a *performance axis*, that is, the performances increase from left (poor) to right (fair); and the second axis as a *style axis*, that is, for an equal performance, Length of Stride is longer above and shorter below.

All procedures presented in the sequel can be visualized in Figure 1, which will provide an intuitive guide throughout the section.

---

[8] Since the six variables are not on a common scale, a PCA of correlations is performed.

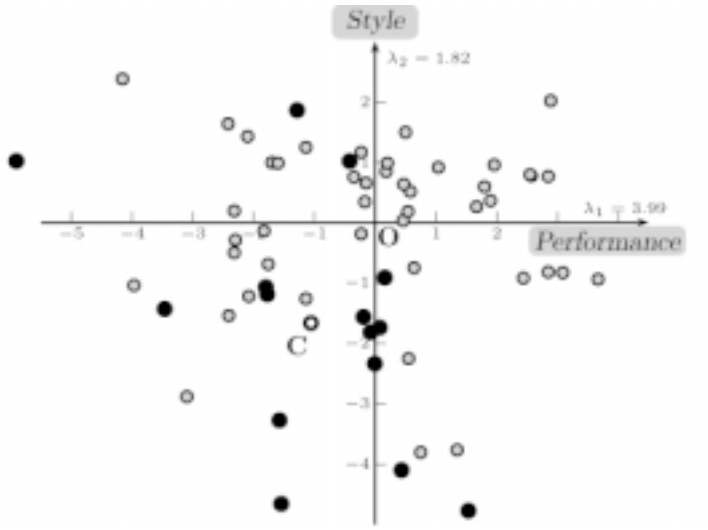

Figure 1: Plane 1-2. The cloud of 45 healthy subjects (gray circles) with its mean point O and the cloud of 15 patients after drug intake (black circles) with its mean point C.

Descriptive findings

The following findings emerge from the PCA.

1. On the whole, performances are poorer for patients, that is to say, most patient points lie on the left side of Figure 1.

2. The mean point of patients after drug intake (point C) still lies on the left side of Figure 1 and remains quite distant from the mean point of healthy subjects (origin point O).

5.2. TYPICALITY TEST FOR PLANE CLOUDS

In the sequel, we will compare patients after drug intake with healthy subjects (reference population) by using combinatorial typicality test to establish the existence of a difference between the two mean points O and C.

The question is:

Is it possible to assimilate patients after drug intake to healthy subjects?

The *effect of interest* is the deviation between the two mean points O and C, namely the geometric vector C−0 with coordinates $(-1.057; -1.663)$.

The covariance matrix of healthy subjects (reference cloud referred to its principal axes) is the diagonal matrix of eigenvalues. Hence, the magnitude of the observed effect is:

$$D^2_{obs} = \begin{pmatrix} -1.057 \\ -1.663 \end{pmatrix} \begin{pmatrix} \frac{1}{3.9927} & 0 \\ 0 & \frac{1}{1.82237} \end{pmatrix} \begin{pmatrix} -1.057 & -1.663 \end{pmatrix} = 1.798$$

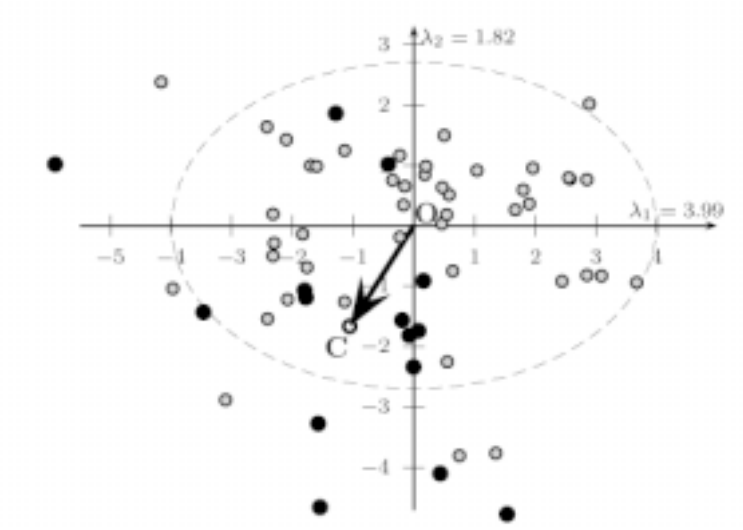

Figure 2: Cloud of the 45 healthy subjects (reference cloud) with its concentration ellipse (κ = 2) and cloud of the 15 patients after drug intake with its mean point C, in plane 1-2.

Descriptively, we conclude that *the deviation between patients after drug intake and healthy subjects is of* large magnitude.

We will now attempt to evaluate the atypicality of the group of Parkinsonians by performing the typicality test.

The number of possible samples is equal to the binomial coefficient $\binom{45}{15} = 3.44 \times 10^{11}$. This number is too large to construct the whole set of possible samples. So we use the Monte Carlo method in order to generate a subset of possible samples (see Figure 3).

Figure 4 shows the distribution of the statistic D^2 based on 500,000 resamples.

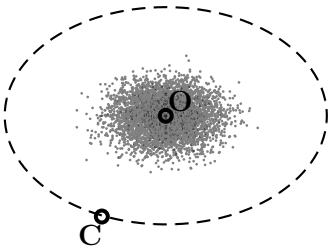

Figure 3: Sample cloud (5,000 points of the sample cloud) and principal ellipse of the reference cloud going through point C.

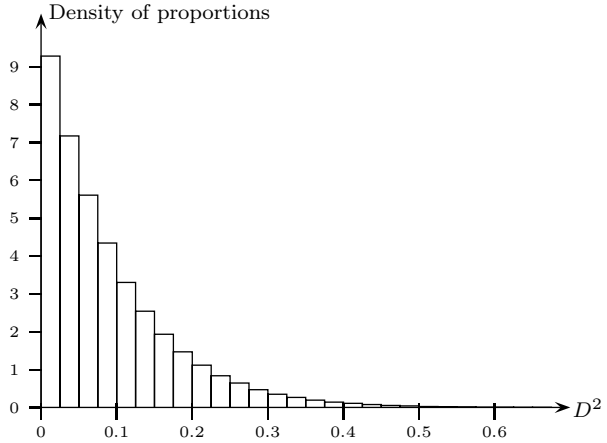

Figure 4: Distribution of test statistic D^2 (based on 500,000 samples among the $\binom{45}{15}$ possible samples); observed value D^2_{obs} = 1 .7 9 8 .

Among the 500,000 resamples, no mean point of possible sample clouds is found outside of or on the ellipse of the cloud of the healthy subjects going through point C, hence $\widehat{p} = 0/500,000$. We can conclude that the group of patients after drug intake is *atypical* of the reference population and say that:

The data are in favor of a difference between patients after drug intake and healthy subjects ($\widehat{p} < 0.001$).

Compatibility region

We will determine the 95% compatibility region, that is, the set of points that are compatible with point C. This region is depicted in Figure 5: it is delineated by an inertia ellipse of the cloud of the healthy subjects with

$\widehat{\kappa} = 0.514$ which is translated so that its center is point C. The mean point O of the healthy subjects is outside the ellipse: the group of patients after drug intake is atypical of the reference population at level .05.

Moreover the compatibility ellipse being located in the South-West quadrant of Figure 5, all points in the other three quadrants lead to significant results: the patients cannot be assimilated to healthy subjects, as they are less performant and make smaller steps.

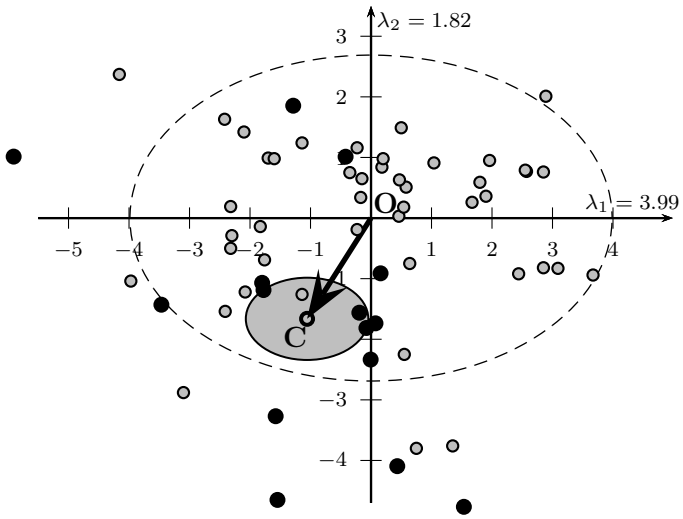

Figure 5: **Cloud of the 45 healthy subjects (reference cloud) with its concentration ellipse (dashed line; $\kappa = 2$) and cloud of the 15 patients after drug intake with its mean point** C **and the .95%–compatibility region ($\kappa = 0.514$).**

Approximate test

One has $D^2_{\text{obs}} = 1.798$, hence $\widetilde{p} = p(\chi^2_2 \geq 15 \times \frac{44}{30} \times 1.798) = 2.57 \times 10^{-9}$. The *approximate observed level* of the test is near 0, hence we have the same conclusion as for the exact test. One has $\chi^2_2[.05] = 5.991$ hence the *compatibility region* at level $\alpha = .05$ is defined by the κ–ellipse with $\widetilde{\kappa} = \sqrt{5.991 \times \frac{1}{15} \times \frac{45-15}{45-1}} = 0.522$. The approximate compatibility region is somewhat larger than the exact one ($0.522 > 0.514$).

5.3. TYPICALITY TEST FOR PERFORMANCE AXIS

The question is: *Are patients after drug intake atypical of healthy subjects as far as performance is concerned?*

Figure 6 shows the cloud of the healthy subjects (grey points) and that of the Parkinsonian patients (black points) on the performance axis.

The mean performance of the healthy subjects is 0 and the variance is the eigenvalue, that is, 3.993. The mean performance of the patients is -1.057, and the scaled deviation between means is $-1.057/\sqrt{3.993} = -0.53$, hence the *descriptive conclusion*:

The mean performance of patients after drug intake is inferior *to that of healthy subjects, and the* difference is large.

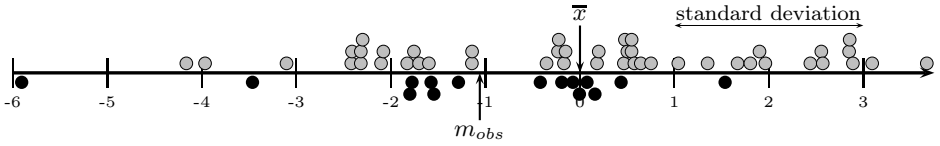

Figure 6: **Reference cloud of healthy subjects (grey points) and observed cloud of patients (black points) on the performance axis.**

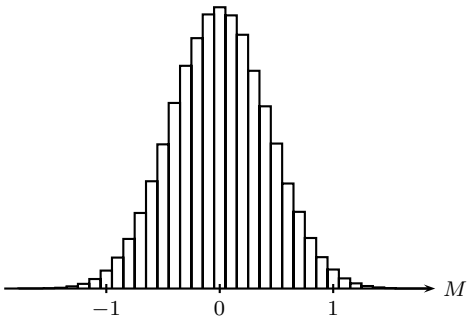

Figure 7: **Distribution of test statistic M based on 500,000 samples, observed value $m_{obs} = -1.057$.**

There are 3,045 resamples over 500,000 that have a mean that is superior or equal to the observed one, hence $\widehat{p} = .006$. The 95% compatibility interval is $[-1.89\,;\,-0.22]$.

The mean performance of patients after drug intake is significantly lower *than that of healthy subjects.*

6. CONCLUSION

The typicality test presented in this paper deals with *Euclidean clouds*. Our example shows the case of a cloud constructed by PCA, but all other methods can be used: examples using MCA can be found in in Bienaise (2013, Chapter 3) or in Le Roux et al. (2018, Chapter 7).

Here, we took a test statistic linked to means, but it is possible to choose *any statistic of interest* (e.g. variance of clouds). Permutation remains the method of choice to test novel or other statistics whose distributions are poorly known.

Combinatorial inference provides an efficient approach to testing when the data do not conform to the distributional assumptions of the statistical method one wants to use (e.g. normality). Furthermore, results of permutation are valid with observations that are not a random sample of some statistical population. Unfortunately, it is generally difficult for permutation tests to express *power functions* in closed form, useful for explicit calculations (see e.g. Pesarin, 2001, p. 63-66). In the combinatorial typicality test, the reference population is known, so we feel it would be preferable to consider *compatibility regions*.

The combinatorial inference can be used for addressing many questions as, for instance: comparing groups of individuals, studying the correlations between variables, interactions between factors, etc. In each case, according to the question under study, we have to choose a "permutation system" in order to construct the "permutation space" and the distribution of the statistic of interest.

Nowadays, the speed of computers makes it possible to perform any statistical test using the permutation method. The chief advantage is that one does not have to worry about distributional assumptions of classical test procedures; the disadvantage is the amount of computer time required to perform a large number of permutations, each one being followed by computation of the test statistic. This disadvantage vanishes as computer science evolves, especially through the parallelization of algorithms.

Software. All computations have been worked out with Coheris-Spad software[9] by including specific routines for the typicality tests that we wrote in R language. The SPAD project with the R scripts[10] is available from authors.

---

[9] The software is distributed by Coheris (www.coheris.com).

[10] R Core Team (www.R-project.org): a language and environment for statistical computing.

REFERENCES

Anderson, T. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics,* 34(1):122–148.

Bienaise, S. (2013). *Méthodes d'inférence combinatoire sur un nuage euclidien/Étude statistique de la cohorte EPIEG.* PhD thesis, Université Paris Dauphine, CEREMADE.

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

Daudin, J.-J., Duby, C. and Trecourt, P. (1988). Stability of principal component analysis studied by the bootstrap method. *Statistics: A Journal of Theoretical and Applied Statistics,* 19(2):241–258.

Edgington, E. (2007). *Randomization Tests*. London: Chapman & Hall/CRC, 4th edition.

Ferrandez, A.-M. and Blin, O. (1991). A comparison between the effect of intentional modulations and the action of L–Dopa on gait in Parkinson's disease. *Behavioural Brain Research,* 45:177–183.

Fisher, R. (1935). The fiducial argument in statistical inference. *Annals of Eugenics,* 6(4):391–398.

Gifi, A. (1990). *Nonlinear Multivariate Analysis.* Chicester: Wiley.

Gilula, Z. and Haberman, S. (1986). Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association,* 81(395):780–788.

Good, P. (2012). *A Practitioner's Guide to Resampling for Data Analysis, Data Mining, and Modeling.* London: Chapman & Hall/CRC.

Johnston, J. E., Berry, K. J. and Mielke, P. W. (2007). Permutation tests: precision in estimating probability values. *Perceptual and Motor Skills,* 105(3):915–920.

Le Roux, B. (1998). Inférence combinatoire en analyse géométrique des données. *Mathématiques et Sciences Humaines,* 144:5–14.

Le Roux, B. (2014a). *Analyse Géométrique des Données Multidimensionnelles.* Paris: Dunod.

Le Roux, B. (2014b). Structured data analysis. In Blasius, J. and Greenacre, M., editors, *Visualization and Verbalization of Data,* pages 185–203. London: Chapman & Hall.

Le Roux, B., Bienaise, S. and Durand, J.-L. (2018). *Combinatorial Inference in Geometric Data Analysis.* London: Chapman and Hall/CRC.

Le Roux, B. and Rouanet, H. (2004). *Geometric Data Analysis. From Correspondence Analysis to Structured Data Analysis.* Dordrecht: Kluwer.

Le Roux, B. and Rouanet, H. (2010). *Multiple Correspondence Analysis, 163.* QASS. CA, Thousand Oaks: SAGE Publications.

Lebart, L. (1976). The significance of eigenvalues issued from correspondence analysis. In *Proceedings in Computational Statistics, Physica Verlag, Vienna,* pages 38–45.

Lebart, L., Morineau, A. and Piron, M. (1995/2006). *Statistique Exploratoire Multidimensionnelle.* Paris: Dunod.

Pesarin, F. (2001). *Multivariate Permutation Tests: with Applications in Biostatistics.* Chichester: Wiley.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society,* 4(1):119–130.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya: The Indian Journal of Statistics, Series A,* pages 329–358.

Saporta, G. and Hatabian, G. (1986). Régions de confiance en analyse factorielle. *Data analysis and informatics,* pages 499–508.