Combinatorial Inference in Geometric Data Analysis

Brigitte Le Roux

Brigitte.LeRoux@mi.parisdescartes.fr www.mi.parisdescartes/∼lerb/

MAP5, CNRS UMR 8145 — Université Paris Descartes
CEVIPOF, CNRS UMR 7048 — SciencesPo. Paris

Naples, CARME 2015

For further developments, see the book entitled:
Combinatorial Inference in Geometric Data Analysis
B. Le Roux, S. Bienaise, J-L. Durand, London, Chapman & Hall/CRC, 2019.
`http://helios.mi.parisdescartes.fr/~lerb/livres/CIGDA/CIGDA_en.html`

Outline

Combinatorial Inference: General Introduction

Geometric Typicality test

Combinatorial Typicality Test

Homogeneity Test

General Introduction

Statistical inference methods for Geometric Data Analysis (GDA)

Combinatorial framework & permutation tests.

1. Descriptive phase:
   inspecting data cloud of individuals.
2. Inference phase:
   tests of typicality of a subcloud
   tests of homogeneity of subclouds

Principles of Combinatorial Tests

1. According to the inductive question (typicality or homogeneity), construct the *permutation space*;
2. Choose a *test statistic*;
3. Calculate the combinatorial *p–value*, i.e. calculate the proportion of elements of the permutation space more or as extreme than the observed one;
4. Determine (if possible) a *compatibility zone*.

Geometric typicality test

Comparison of the mean point of a cloud (group of observations) to a reference point.

Question

Is the reference point the "true mean point"?

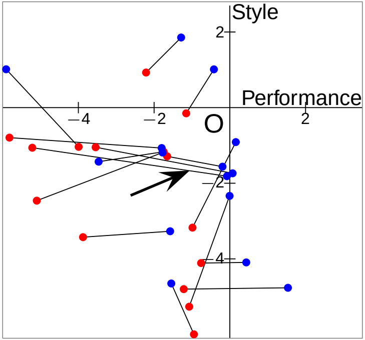
(combinatorial hypothesis under test).

Parkinson example*

15 patients observed twice:
before and after drug intake

2 variables pertaining to gait:
performance and *style*

2 matched groups of observations:
2 clouds of 15 points
(before and after)



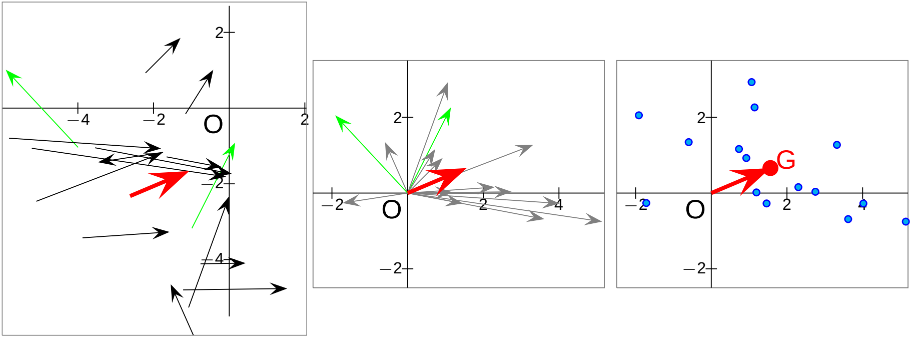
Question:

Does the drug have an effect?

[*] FERRANDEZ A-M. & BLIN O. (1991). A comparison between the effect of intentional modulations and the action of L–Dopa on gait in Parkinson's disease, *Behav. Brain Research*, 45, 177-183

Constructing the Pertinent Cloud

1) 15 **vector–effects** with the **mean vector–effect**;
2) Shift vectors by translation so that their initial points coincide with origine;
3) Take the endpoints of translated vectors to represent effects → the *cloud of 15 point–effects* (pertinent cloud).

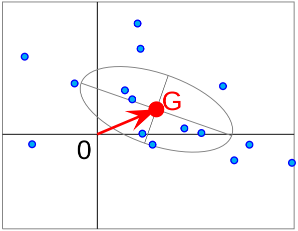


Pertinent Cloud

$\mathrm{D}^I = (\mathrm{D}^i)_{i=1,\ldots n}$: cloud of *n* points in a Euclidean space of dimension *L* equipped with an orthonormal basis.

Mean point G with coordinates $\mathbf{d}_\mathrm{O}$
Covariance matrix **V**

Parkinson Example



Pertinent cloud of 15 point–effects with:

- indicator ellipse
- mean vector– effect: $\overrightarrow{\mathrm{OG}}$ (in matrix notation: $\mathbf{d}_\mathrm{O}$)

Permutation Space

For the Parkinson example,

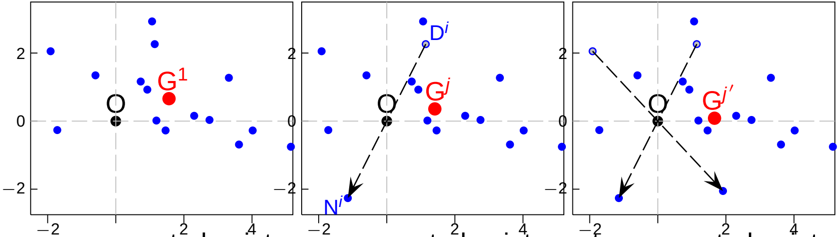
saying: "the drug has no effect"

means: "the values observed after drug intake could have been observed before."

Geometrically speaking:
one could observe a point as well as its symmetric with respect to reference point O.

Permutation space = set of 2^n clouds generated by changing each point D^i into its *symmetric* N^i with respect to O, then: $\overrightarrow{\mathrm{ON}^i} = -\overrightarrow{\mathrm{OD}^i}$.
Hence 2^n clouds D^{Ij} $(1 \leq j \leq 2^n)$
with their mean points $(\mathrm{G}^j)_{j=1,\ldots 2^n}$.

Parkinson Example: Three examples of clouds D^{Ij}

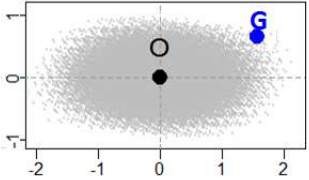


no permuted point
(basic cloud: $G^1{=}G$)

one permuted point

two permuted points

Permutation Space: cloud G^J of 2^n mean points $(G^j)_{j=1,\ldots 2^n}$



Properties:

For the cloud of mean points $(\mathrm{G})_{j=1,\ldots 2^n}$:

- The mean point is point O;
- The covariance matrix is $\mathbf{B}_{\mathrm{O}}/n$
  with $\mathbf{B}_{\mathrm{O}} = \mathbf{V} + \mathbf{d}_{\mathrm{O}}\mathbf{d}_{\mathrm{O}}^{\top}$ ($\mathbf{d}_{\mathrm{O}}$: column–vector of the components of the mean vector–effect $\overrightarrow{\mathrm{OG}}$).

($\mathbf{B}_0$ is the mean of squares and products matrix of basic cloud D^I with respect to point O)

Test Statistic and *p*–value

Given a point P, we take as *test statistic*, denoted T, the $\mathbf{B}_{\mathrm{O}}$–norm of vector $\overrightarrow{\mathrm{OP}}$, that is,

$$T : \mathrm{P} \mapsto |\overrightarrow{\mathrm{OP}}|_{B_{\mathrm{O}}} = (\mathbf{d}\mathbf{B}_0\mathbf{d}^{\top})^{1/2}$$

(**d** denotes the column vector of vector $\overrightarrow{\mathrm{OP}}$)

Calculate $T(\mathrm{G}^j)$, in brief $T(j)$, for $j = 1, \ldots 2^n$, hence the *distribution* of T.

Definition

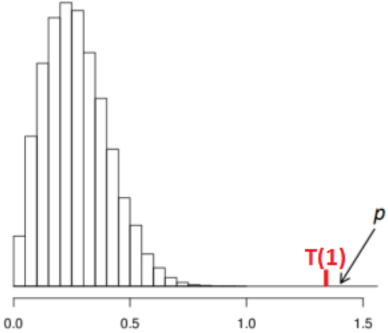
The combinatorial *p–value* is the *proportion* of points G^j whose distance $T(j)$ to point O is greater than or equal to the distance $T(1)$ from G to O.

Parkinson Example

Distribution of test statistic T.

Observed value:
$T(1) = 0.0809$.

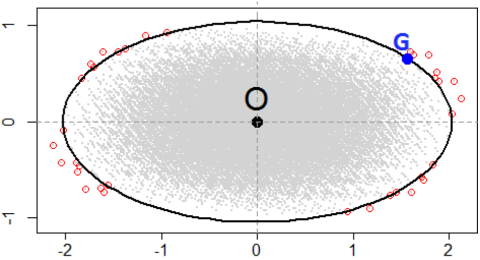
There are 36 points G^j whose distance to O is greater than or equal to the distance from G to O, hence:



$$p = 36/2^{15} = 0.0011 \text{ (highly significant result)}$$

Geometric interpretation of the *p*–value

Cloud G^J of the mean points of the $2^{15} = 32\,768$ possible clouds and ellipse of points at the same $\mathbf{B}_0$–distance to O as G.



p–value = proportion of points of cloud G^J outside (or on) the ellipse:

36 points are outside (or on) the ellipse, hence $p = \frac{36}{2^{15}}$.

Conclusion: asserting existence of effect

$p = 0.0011$: highly significant result (S**):

The data are compatible with the hypothesis that *the "true point" is point O*.

The data are in favor of a non–null effect of the drug.

Compatibility zone

By taking as reference point every point of the geometric space, we can define the $(1-\alpha)$ compatibility zone.

Definition

A point P is said *compatible* with point G at level $\alpha \iff$ the *p*–value associated with the reference point P is greater than α.

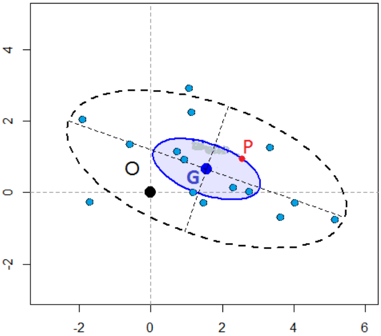
Definition

The *compatibility zone* is the set of points P that are compatible with point G at level α.

Point P is called *compatibility limit–point* if the corresponding *p*–value is equal to α.

Theorem

If point P is a compatibility limit–point at level α, any point of the inertia ellipsoid of the basic cloud going through P is also a compatibility limit–point at level α.



Combinatorial Typicality Test: Introduction

Comparison of a subcloud to a reference cloud.

Parkinson Example

Comparing patients after drug intake with healthy subjects.

Question

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

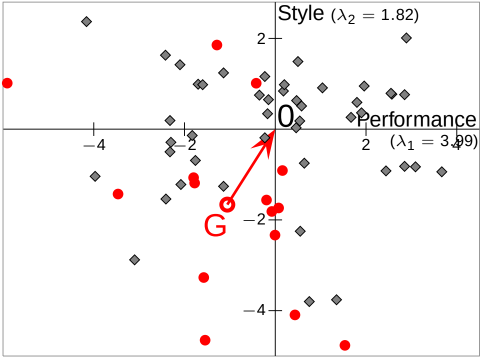
Principles of the Combinatorial Typicality Test

N: number of individuals in the reference population;
n: number of individuals in the group.

Parkinson Example.

Reference population: healthy subjects
→ cloud of 45 points with mean point O

Group of observations: 15 patients after drug intake
→ cloud of 15 points with mean point G



Parkinson Example: Descriptive appraisal

Effect of interest: geometric vector $\overrightarrow{\mathrm{OG}}$

with components $\mathbf{d} = \begin{pmatrix} -1.06 \\ -1.66 \end{pmatrix}$

magnitude of effect: **V**–norm of $\overrightarrow{\mathrm{OG}}$

$$|\overrightarrow{\mathrm{OG}}|_{\mathrm{V}} = (\mathbf{d}^{\top}\mathbf{V}^{-1}\mathbf{d})^{-1/2} = 1.34.$$

Descriptive conclusion: the effect is of large magnitude.

Inductive question

Is the hypothesis *that the "true mean point" of patients after drug intake coincides with the mean point of healthy subjects* compatible with the data or not?

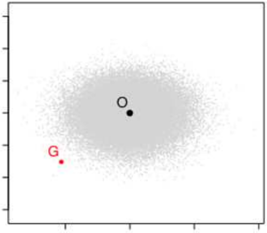
Construction of the test

- One considers the set of $\binom{N}{n}$ subsets of size n from the set of N individuals;
- A subcloud with mean point G^j is associated wich each subset
  $\rightarrow$ the cloud of $\binom{N}{n}$ mean points: $\mathrm{G}^J = (\mathrm{G}^j)_{j=1,\ldots\binom{N}{n}}$

Properties: Its mean point is point O;
its covariance matrix is $\mathbf{W} = \frac{N-n}{N-1}\,\frac{\mathbf{V}}{n}$.

- Test statistic. $T : P \mapsto |\overrightarrow{\mathrm{OP}}|_V$
- p = proportion of points G^j such that $T(j) \geq T(1)$

Parkinson Example



$$p = 0/\binom{45}{15}$$

Conclusion

The hypothesis *that the "true mean point" of patients after treatment coincides with the mean point of healthy subjects* is incompatible with the data.

Homogeneity problems

I: set of N individuals → cloud of N points
mean point O
covariance matrix **V**

Partition of I into C groups of sizes $(n_c)_{c \in C}$ → C subclouds ($c \in C$)
mean points: G^c weighted by n_c
covariance matrices: $\mathbf{W}_c$

Recall that: $\mathbf{V} = \mathbf{B} + \mathbf{W}$

between covariance matrix: **B**
within covariance matrix: $\mathbf{W} = \sum \frac{n_c}{N} \mathbf{W}_c$

Question

Are the groups heterogeneous?

Dog Example*

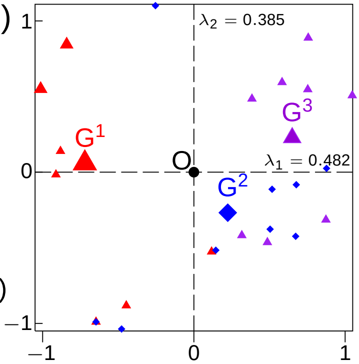
- ▶ *27 individuals* (breeds of dog)
- ▶*6 categorized variables*:

(size, weight, velocity, intelligence, affection, aggressiveness):

3+3+3+3+2+2=16 categories

- ▶*Structuring factor:* the function of the dog:

(companion ($c1$) (10), hunting ($c2$) (9) and working ($c3$) (8)).



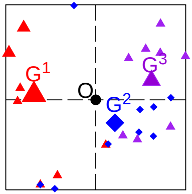
Dog breeds in the first principal plane of the MCA.

*Example from Tenenhaus, *Statistique: Méthodes pour décrire expliquer et prévoir*, Dunod, 2007

- *Global comparison*:

  comparison of C groups.

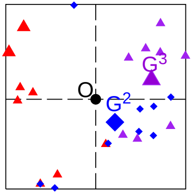
  $C = \{c1, c2, c3\}$



- *Partial comparison*:

  comparison of C' groups ($C' \subset C$).

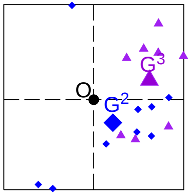
  $C' = \{c2, c3\}$



- *Specific comparison*:

  comparison of C' groups after restriction to individuals belonging to groups $c \in C'$ ($C' \subset C$).

  $C' = \{c2, c3\}$



Test Principle

- A set of $J = \dfrac{N!}{\prod\limits_{c \in C} n_c!}$ *possible clouds* of the same structure is generated by exchanging points between the C subclouds;
- the possible subcloud of rank j has *C mean points* $(\mathrm{G}^{cj})_{c \in C}$ with weigths $(n_c)_{c \in C}$,
- *Test statistic*

$$T^2 : (\mathrm{G}^{cj})_{c \in C} \longmapsto T^2(j) = \frac{1}{2} \sum \sum \frac{n_c n_{c'}}{N^2} \left[\mathrm{G}^{cj}\mathrm{G}^{c'j}\right]_V^2$$

$\left[\mathrm{G}^{cj}\mathrm{G}^{c'j}\right]_V^2 = \mathbf{d}_{(cc')j}^\top \mathbf{V}^{-1} \mathbf{d}_{(cc')j}$ denotes the squared $\mathbf{V}$–norm of vector $\overrightarrow{\mathrm{G}^{cj}\mathrm{G}^{c'j}}$ with components $\mathbf{d}_{(cc')j}$;

- *p–value* = proportion of possible clouds such that: $T^2(j) \geq T^2(1)$ $\quad$ ($j = 1$ corresponds to the initial cloud).

Partial comparison $C' \subset C$

Comparison of C' groups with $N' = \sum_{c \in C'} n_c$.

c_r: group of size $N - N'$ constituted by the union of groups $c \notin C'$.

$\longrightarrow$ C' subclouds of sizes $(n_c)_{c \in C'}$
1 subcloud with $N - N'$ points.

Partial comparison $C' \subset C$

- Permutation Space
  set of the $J = \frac{N!}{\prod\limits_{c \in C'} n_c!(N-N')}$ possible clouds generated by exchanging points between the $C'+1$ subclouds. Hence the *possible cloud* of rank j has $C'+1$ mean points: $(\mathrm{G}^{cj}, n_c)_{c \in C'}$ and $(\mathrm{G}^{rj}, n_r = N - N')$.
- Test statistic
  $T^2 : (\mathrm{G}^{cj})_{c \in C'} \longmapsto T^2(j) = \frac{1}{2} \sum\limits_{c \in C'} \sum\limits_{c' \in C'} \frac{n_c n_{c'}}{N'^2} \left[\mathrm{G}^{cj}\mathrm{G}^{c'j}\right]^2_V$
- The *p*–value is the proportion of possible clouds such that: $T^2(j) \geq T^2(1)$ ($j = 1$ corresponds to the initial cloud, i.e. the possible cloud without exchange).

Particular case of 2 groups

Group c_1 with weight n_1 and group c_2 with weight n_2

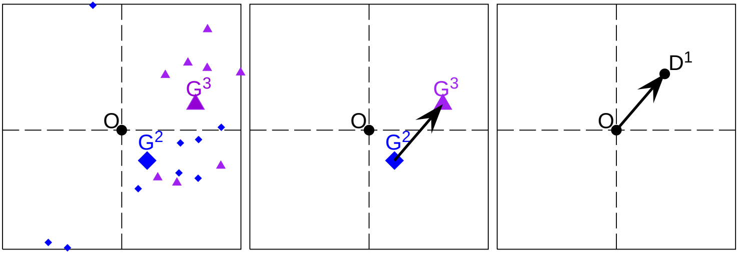
$$T^2 : (\mathrm{G}^{c_1 j}, \mathrm{G}^{c_2 j}) \longmapsto T^2(j) = \frac{n_1\, n_2}{(n_1+n_2)^2}[\mathrm{G}^{c_1 j}\mathrm{G}^{c_2 j}]^2_V$$

Test statistic T^2 is equivalent to statistic E defined by

$$E : (\mathrm{G}^{c_1 j}, \mathrm{G}^{c_2 j}) \longmapsto E(j) = [\mathrm{G}^{c_1 j}\mathrm{G}^{c_2 j}]_V$$

Case of 2 groups: Cloud of deviation–points

Given j, consider the deviation between the two mean points $\mathrm{G}^{c_1 j}$ and $\mathrm{G}^{c_2 j}$, the deviation–vector $\overrightarrow{\mathrm{G}^{c_1 j}\mathrm{G}^{c_2 j}}$ is translated so that the point $\mathrm{G}^{c_1 j}$ coincides with O, hence the terminal point D^j: $\overrightarrow{\mathrm{OD}^j} = \overrightarrow{\mathrm{G}^{c_1 j}\mathrm{G}^{c_2 j}}$.

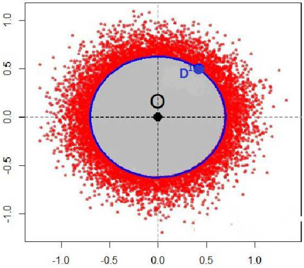


Properties of cloud D^J and *p*–value

- The mean point is point O.
- The covariance matrix is equal to $\frac{N}{N-1}\frac{\mathbf{V}}{\widetilde{n}}$, with $\widetilde{n} = 1/(\frac{1}{n_1} + \frac{1}{n_2})$.

Dog example: comparison $(c2, c3)$

p–value = proportion of points outside (or on) the ellipse going through point D^1.
$p = 0.126$ (Monte–Carlo method): non–significant.



Compatibility zone

Denote $\mathrm{M}^{I_{c1}}$ and $\mathrm{M}^{I_{c2}}$ the two subclouds associated with c_1 and c_2 with mean points G^1 and G^2.

Given a vector $\overrightarrow{u}$, consider

1. the subcloud $\mathrm{M}_u^{I_{c1}}$ translation by vector $n_2\overrightarrow{u}$ of cloud $\mathrm{M}^{I_{c1}}$ with mean point $\mathrm{G}_u^1 = \mathrm{G}^1 + n_2\overrightarrow{u}$.
2. the subcloud $\mathrm{M}_u^{I_{c2}}$ translation by vector $-n_1\overrightarrow{u}$ of cloud $\mathrm{M}^{I_{c2}}$ with mean point $\mathrm{G}_u^2 = \mathrm{G}^1 + n_2\overrightarrow{u}$.
3. the sucloud M^{I_r} of $N' = N - (n_1 + n2)$ initial points.

Calculate the p–value.

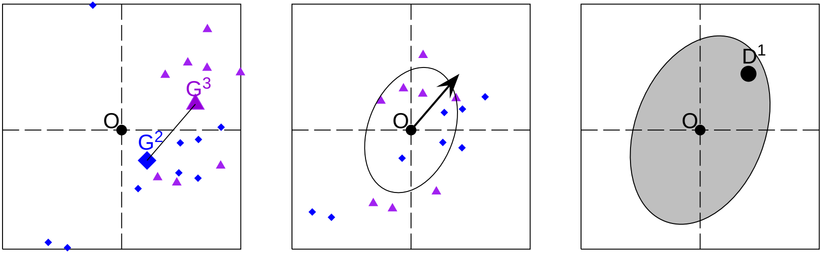
Compatibility zone

Definition.
The set of points P such that $\overrightarrow{\mathrm{OP}} = \overrightarrow{\mathrm{G}^1\mathrm{G}^2} - (n_1 + n_2)\overrightarrow{u}$ and with a *p*–value greater than α defines the compatibility zone at level $(1 - \alpha)$.

For a *specific comparison* ($C = \{c1, c2\}$), it can be shown that the compatibility zone depends on the within covariance $\mathbf{W} = \frac{n_1\mathbf{V}_1 + n_2\mathbf{V}_2}{n_1 + n_2}$.

Dog example: specific comparison $(c2, c3)$

Cloud with 2 groups



within–cloud
with indicator ellipse

compatibility–zone

Summary

- The framework is basically combinatorial, not random.
- Combinatorial procedures depend on data size, therefore they are not descriptive in character.
  They are performed on all possible data sets, not just on the data themselves, therefore, they are inductive in character.
- Combinatorial inference is seen to be in harmony with the inductive philosophy of GDA: description first, inference later