# Estatistika metodoak distantzietan oinarrituriko ikuspegitik

*Itziar Irigoien*[1], *Concepción Arenas*[2], *Susana Ferreiro*[3], *Basilio Sierra*[1]

[1] Konputazio Zientziak eta Adimen Artifiziala.
Euskal Herriko Unibertsitatea (UPV/EHU)
[2] Departament d'Estadística. UB
[3] IK4-Tekniker

Jasota: 23/04/2013 Onartua: 17/07/2013

**Laburpena:** Lan honetan analisi anizkoitzaren baitan biltzen diren distantzietan oinarrituriko hainbat metodoren berrikusketa egin da. Oinarrizko kontzeptuak aurkeztu dira lehenik, ondoren metodoen funtsa laburki azaldu ahal izateko. Zehazki, erregresioa, diskriminazio-analisia, cluster-analisia, tipikotasuna eta sakonera aztertzen dituzten metodoak bildu ditugu. Metodologia honek berezitasun nabarmena du, edozein datu motaren gainean aplikagarria baita, datuek erakusten duten banaketa zein den ezagutu beharrik gabe. Azkenik, metodo hauen erabilgarritasuna erakusteko, funtzio-datu errealen gainean aplikatu eta lortutako emaitzak azaldu dira.

**Abstract:** A revision of some distance-based methods is presented within the Multivariate Analysis. First, basic concepts are introduced so that those as regression, discrimination, clustering, typicality and depth methods are briefly explained. The main characteristic of distance based methods is that they can be applied to any type of mixed data, where a known distribution of data is not necessary. Finally, these methods have been applied to a real functional data set and the obtained results are shown.

## 1. SARRERA

Edozein azterketaren helburu diren aleen inguruan bildutako informazio ordenatuari *datuak* esaten diogu eta estatistikako metodoak datu horietatik ezagutza ateratzea ahalbidetzen duten matematikako teknikak dira. Estatistikako metodo batzuek *alea* × *aldagaiak* ikuspegian oinarritzen dira eta ondorioz datu matrizetik abiatzen dira. Beste metodo batzuek, aldiz, aleen arteko berdintasunak/desberdintasunak direlakoak aztertzean datza;

*distantzietan oinarrituriko* metodoak dira. Metodo horiek distantzia matrizetik abiatzen dira, hau da, aleen arteko distantzia aztertzetik. Bi ikuspegi horiek elkarrekin osagarriak dira, eta oro har erraza da aleak×aldagaiak ikuspegitik distantzietan oinarritutakora aldatzea estatistika-distantzien bidez. Distantzietan oinarrituriko ikuspegitik abiatzeak baditu abantaila batzuk. Adibidez, datu matrizea aztertzen duten metodo batzuek aldagaiak jarraituak izatea eskatzen dute; aldiz, ez du muga hori distantzia matrizean oinarritzen den metodoak. Falta diren balioekin datu matrizean sortzen den zailtasuna ere distantzia matrize egokia aukeratuz gaindi daiteke faltako balioen estimaziorik egin gabe. Gainera, zenbait egoeratan, datuen gaineko banaketa jakin bat ontzat ematea behar duen datu matrizearekin lan egitea baino egokiagoa da distantzia matrizearekin aritzea. Izan ere, distantzietan oinarritzen diren metodoak erabiltzeko egokiak dira edozein datu mota izanik ere, datuek erakuts dezaketen banaketaren gainean inolako baldintzarik ezarri gabe.

Lan honetan azken urteetan garatutako distantzietan oinarrituriko hainbat metodo laburbildu dira. Bigarren atalean oinarrizko definizio batzuek eta hainbat propietate interesgarri bildu dira. Hirugarren atalean dago aukeraturiko metodoen aurkezpena eta interpretazioa. Laugarren atalean bildu ditugu datu errealen gainean metodo horiekin lorturiko emaitzak eta azkenik, bosgarren atalean, atera ditugun ondorioak.

## 2. OINARRIZKO DEFINIZIOAK

Har dezagun $\mathbf{Z}$ zorizko $p$-bektorea, $\lambda$ neurri egokiarekiko $f$ dentsitate-funtzioa duena eta R euskarria, eta izan bedi berau adierazten duen $n$ aledun $C$ multzoa. Izan bedi $\delta$ distantzia, $\mathbf{z}_i$, $i = 1, \ldots, n$ aleen artekoa. Distantzia estatistiko mota ugari dago eta egokia datu motaren arabera aukeratuko dugu. Izan bedi $\mathbf{X}$ $(n \times q)$ koordenatu nagusien matrizea [1, 2], zeinak $\Delta = (\delta_{ij})_{i,j=1,\ldots,n}$ distantzia matrizetik lortuak diren. Hau da, $\mathbf{X}$ren zutabeak $(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{B}(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)$ matrizearen bektore propioak dira, non $\mathbf{I}$ identitate matrizea den eta $\mathbf{B}$ matrizean $-\frac{1}{2}\delta_{ij}^2$ elementuak dauden. Bektore propio horiek dagozkien balio propioen arabera ordenatu dira handienetik txikienera eta bakoitzaren norma dagokion balio propioaren berdina aukeratu da. Onar daiteke $\mathbf{X}$ matrizeak $n$ aleak adierazten dituela $q$ dimentsiotan. Izan ere koordenatu horiek jatorrian zentratuak daude eta hain zuzen, $i$ eta $j$ aleen arteko Euklides distantzia $\delta_{ij}$ da. Beraz, balio propio negatiborik ez badago $\delta$ euklidearra dela esaten da eta $(\mathcal{R}, \delta)$ espazioa $(\mathbf{R}^q, \delta_{Euklides})$ espazioarekin uztar daiteke. Hau da, existitzen da $\psi : \mathcal{R} \rightarrow \mathbf{R}^q$ funtzioa zeinarekin $\delta_{ij}^2 = \|\psi(\mathbf{z}_i) - \psi(\mathbf{z}_j)^2\|$ erdiesten den. Hasierako $\mathbf{Z}$ datu matrizea koordenatu errealeko $\mathbf{X}$ matrizean bihurtu ondoren, analisi anizkoitzeko edozein teknika aplika diezaiokegu $\mathbf{X}$ri, datu kuantitatiboko ma-

trizeari. Ikuspegi hori, *distantzietan oinarriturikoa*, aldagai mistoekin erregresio linealaren ereduan [3] erabili zen lehen aldiz.

Metodologia horrek distantzia euklidearra behar du eta horrela ez bada, erraz lor daitezke $\delta_{ij}^{*}$ distantzia euklidearrak honela: $\delta_{ij}^{*} = (\delta_{ij}^{2} + h)^{1/2}\ \forall i, j$, non $h$ lehentxeago aipaturiko balio propio txikienaren balio absolutua den [4, 5].

Ondorengo definizioek, baimentzen dute besteak beste itxaropena, bariantza eta ale batetik populaziorako distantziaren kontzeptuak $(\mathcal{R}, \delta)$ espazio metrikora eramatea.

1. **Definizioa** [6]
Har dezagun $\mathbf{Z}$ zorizko $p$-bektorea, $\lambda$ neurri egokiarekiko $f$ dentsitate-funtzioa eta $\mathcal{R}$ euskarria dituena, eta izan bedi berau adierazten duen $n$ aleko $C$ multzoa. Izan bedi $\delta$ aleen arteko distantzia. *$C$ ren aldakortasun geometrikoa* $\delta$rekiko honela definitzen da:

$$V_{\delta}(C) = \frac{1}{2}\int_{\mathcal{R}\times\mathcal{R}} \delta^{2}(\mathbf{z}_i, \mathbf{z}_j) f(\mathbf{z}_i) f(\mathbf{z}_j) \lambda(d\mathbf{z}_i) \lambda(d\mathbf{z}_j).$$

2. **Definizioa** [7]
Har ditzagun $\mathbf{Z}_1$ eta $\mathbf{Z}_2$ zorizko $p$-bektoreak, hurrenez hurren $\lambda$ neurri egokiarekiko R euskarria eta $f_1$ eta $f_2$ dentsitate-funtzioak dituztenak. Izan bitez bektore horiek adierazten dituzten $n_1$ aleko $C_1$ eta $n_2$ aleko $C_2$ multzoak hurrenez hurren. Izan bedi $\delta$ aleen arteko distantzia. *$C_1$-en eta $C_2$-ren arteko distantzia (karratua)* honela definitzen da:

$$\Delta^{2}(C_1, C_2) = \int_{\mathcal{R}\times\mathcal{R}} \delta^{2}(\mathbf{z}_{(1)i}, \mathbf{z}_{(2)j}) f_1(\mathbf{z}_{(1)i}) f(\mathbf{z}_{(2)j}) \lambda(d\mathbf{z}_{(1)i}) \lambda(d\mathbf{z}_{(2)j}) - \\ - V_{\delta}(C_1) - V_{\delta}(C_2).$$

3. **Definizioa** [7]
Har dezagun $\mathbf{Z}$ zorizko $p$-bektorea, $\lambda$ neurri egokiarekiko $f$ dentsitate-funtzioa eta $\mathcal{R}$ euskarria dituena, eta izan bedi berau adierazten duen $n$ aleko $C$ multzoa. Izan bitez $\mathbf{z}_0 \in \mathbf{R}^p$ ale berria eta $\delta$ aleen arteko distantzia. $\mathbf{z}_0$ *alearen gertutasuna (karratua) $C$ multzora* $\delta$ distatziarekiko honela definitzen da:

$$\phi_{\delta}^{2}(\mathbf{z}_0, C) = \int_{\mathcal{R}} \delta^{2}(\mathbf{z}_0, \mathbf{z}) f(\mathbf{z}) \lambda(d\mathbf{z}) - V_{\delta}(C).$$

1. **PROPIETATEA** [7]
Aurreko definizioetan ezarritako balditza berdinetan, $E(\psi(\mathbf{Z}_r))$ eta $E(\|\psi(\mathbf{Z}_r)\|^2)$ ($r = 1, 2$) finituak izanik, ondorengo berdintzak betetzen dira:

$$V_\delta(C_r) = E(\|\psi(\mathbf{Z}_r) - E(\psi(\mathbf{Z}_r))\|^2), \quad (1)$$

$$\Delta^2(C_1, C_2) = \|E(\psi(\mathbf{Z}_1)) - E(\psi(\mathbf{Z}_2))\|^2, \quad (2)$$

$$\delta(\mathbf{z}_0, C_r) = \|\psi(\mathbf{z}_0) - E(\psi(\mathbf{Z}_r))\|^2, \quad r = 1, 2. \quad (3)$$

($\mathbf{R}^q$, $\delta_{Euklides}$) espazioan dugun batez bestekoaren kontzeptuaren parekoa defini dezakegu ($\mathcal{R}$, $\delta$) espazioan gertutasun funtzioarekin. $\delta$-batezbestekoa deituko diogu haserako espazioko $\mathbf{z}$ balioei baldin eta bere irudia $\psi$rekin $E(\psi(\mathbf{Z}))$ bada, hau da, $\psi(\mathbf{z}) = E(\psi(\mathbf{Z}))$. Sinplifikatuz, $\mathbf{z}$ balioak $E(\psi(\mathbf{Z}))$ renarekin identifikatuko ditugu.

Emaitza horiei esker, aurkeztu ditugun definizioek itxaropen, bariantza, populazioen arteko distantzia eta aletik populaziorako distantzien kontzeptuak berreskuratzen dituzte. Guzti hori aplikatu behar den egoeretan, $\delta$ distantzia datutzat hartzen da baina populazioen banaketa-funtzioak ezezagunak izaten dira. Ondorioz, kontzeptu horien estimatzaileak behar ditugu. Izan bitez $\mathbf{z}_{(1)1}, \ldots, \mathbf{z}_{(1)n_1}$ eta $\mathbf{z}_{(2)1}, \ldots, \mathbf{z}_{(2)n_2}$ laginak, hurrenez hurren $C_1$ eta $C_2$ multzoei dagozkienak. Modu naturalean ondorengo estimatzaileak ditugu:

*Aldakortasun geometriokoa*

$$\hat{V}_\delta(C_r) = \frac{1}{2n_r^2} \sum_{i,j \in C_r} \delta^2(\mathbf{z}_{(r)i}, \mathbf{z}_{(r)j}), \qquad r = 1, 2. \quad (4)$$

*$C_1$en eta $C_2$ren arteko distantzia (karratua)*

$$\hat{\Delta}^2(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{\substack{i \in C_1 \\ j \in C_2}} \delta^2(\mathbf{z}_{(1)i}, \mathbf{z}_{(2)j}) - \hat{V}_\delta(C_1) - \hat{V}_\delta(C_2). \quad (5)$$

*$\mathbf{z}_0$ alearen gertutasuna (karratua) $C_r$ra, $r = 1, 2$*

$$\hat{\phi}_\delta^2(\mathbf{z}_0, C) = \frac{1}{n_r} \sum_{i \in C_r} \delta^2(\mathbf{z}_0, \mathbf{z}_{(r)i}) - \hat{V}_\delta(C_r), \qquad r = 1, 2. \quad (6)$$

Estimatzaile horien kalkulurako $\delta_{ij}$ distantziak soilik ezagutu behar dira. Aipatutako (1), (2) eta (3) propietateen lagin bertsioak ondoregoak dira [7]:

$$\hat{V}_\delta(C) = \frac{1}{n_r}\sum_{i=1}^{n}\left\|\psi(\mathbf{z}_{(r)i})\right\|^2 - \left\|\overline{\psi(\mathbf{z}_{(r)})}\right\|^2 = \frac{1}{2}\sum_{i=1}^{n}\left\|\mathbf{x}_{(r)i})\right\|^2 - \left\|\overline{\mathbf{x}}_{(r)}\right\|^2 \quad (7)$$

$$\hat{\Delta}^2(C_1,C_2) = \left\|\overline{\psi(\mathbf{z}_{(1)})} - \overline{\psi(\mathbf{z}_{(2)})}\right\|^2 = \left\|\overline{\mathbf{x}}_{(1)} - \overline{\mathbf{x}}_{(2)}\right\|^2, \quad (8)$$

$$\hat{\phi}^2_\delta(\mathbf{z}_0,C) = \left\|\psi(\mathbf{z}_{(0)}) - \overline{\psi(\mathbf{z}_{(r)})}\right\|^2 = \left\|\mathbf{x}_{(0)} - \overline{\mathbf{x}}_{(r)}\right\|^2, \qquad r = 1, 2, \quad (9)$$

non $\psi : \mathcal{R} \to \mathbf{R}^q$ funtzioak $\mathbf{x}_{(r)i} = \psi(\mathbf{z}_{(r)i})$ koordenatu nagusiak ematen dituen, $\bar{\mathbf{x}}_{(r)}$ batezbestekoa $\psi(\mathbf{z}_{(r)i})$ balio guztiena den, hau da, $C_r$ ren zentroidea den ($r = 1, 2$) eta azkenik $\mathbf{x}_0 = \psi(\mathbf{z}_0)$ den.

2. **PROPIETATEA** [8]
$C$ ren $C_1, \ldots, C_k$ partiketa harturik, bariantza-analisiko identitate funtsezkoaren parera deskonposa daiteke aldakortasun geometrikoa,

$$n\hat{V}_\delta(C) = \sum_{r=1}^{k} n_r \hat{V}_\delta(C_r) + \frac{\mathbf{n}'\Delta\mathbf{n}}{n} \quad (10)$$

GUZTIZKO ald. = BARRUKO ald. + ARTEKO ald.

non $\mathbf{n} = (n_1, \ldots, n_k)'$den eta $\Delta_{rs} = \frac{1}{2}\hat{\Delta}^2(C_r, C_s)$

elementudun $\Delta$ ($k \times k$) matrize simetrikoa den, $r, s = 1, \ldots, k$. Hau da, $C_r$, $C_s$ bikote bakoitzaren artean kalkulaturiko (5) distantziaren (karratua) erdiek osatzen dute matrizea.

Propietate horrek, $C_1, \ldots, C_k$ partiketarako, $C$ multzoaren distantzietan oinarrituriko aldakortasunaren deskonposaketa eskaintzen du.

## 3. DISTANTZIETAN ONARRITURIKO METODOAK

Jarraian, labur azalduko dira distantzietan oinarrituriko hainbat metodo.

### 3.1. Erregresioa

Izan bedi $\mathbf{Y} = \gamma_0\mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}$ erregresio eredu lineala, non $\mathbf{Z}$ $n \times p$ datu matrizea ezaguna eta $p$ heinekoa den, $\mathbf{1} = (1, \ldots, 1)'$den, $\mathbf{Y}$ bektoreak zorizko aldagai kuantitatiboari dagokion $n$ neurketak dituen eta $\gamma_0$, $\boldsymbol{\gamma}$ estimatu beharreko parametroak diren. Gainera, ereduaren arabera $E(\mathbf{e}) = 0$ eta

$E(\mathbf{ee}') = \sigma^2\,\mathbf{I}$ dira, $\sigma^2$ ezezaguna delarik. Ereduak eskatzen du $\mathbf{Z}$ datu matrizean jasotako neurketak $p$ aldagai kuantitatiborenak izan daitezen, nahiz eta aldagai kualitatiboei dagozkien neurketak aldagai adierazleen bidez bertan txerta daitezkeen.

Honela idatz daiteke [3] eredu hori:

$$\mathbf{Y} = \beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{11}$$

non $\mathbf{X}$, $\beta_0$ eta $\boldsymbol{\beta}$ ondorengoak diren:

1. $\mathbf{X}$ datu matrizearen koordenatu nagusien matrizea da, hau da, $\mathbf{X} = \mathbf{HZV}$, $\mathbf{H} = (\mathbf{I} - \mathbf{11}'/n)$ zentratze-matrizearekin eta $\mathbf{V}$ bektore propioek osaturiko matrizearekin; azken horiek $(\mathbf{HZ})(\mathbf{HZ})' = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$ren espektro-deskonposaketari dagozkionak dira;
2. $\beta_0 = \gamma_0 + \bar{\mathbf{z}}'\boldsymbol{\gamma}$ da, $\mathbf{Z}$ren batezbestekoen bektorea da $\bar{\mathbf{z}}$;
3. $\boldsymbol{\beta} = \mathbf{V}'\boldsymbol{\gamma}$ da.

Demagun $\mathbf{Z}$ datu matrizean aldagai kuantitatibo, bitar eta kualitatiboei dagozkien neurketak bildu direla. Izan bitez $\boldsymbol{\Delta} = (\delta_{ij})$ distantzia matrizea eta $\mathbf{X}$ koordenatu nagusien $n \times q$ matrizea. Idatz dezagun $\mathbf{X} = (\mathbf{X}_{(k)}, \mathbf{W})$ eran non $\mathbf{X}_{(k)}$ matrizean $k < q$ zutabe aukeratu diren $\mathbf{X}$ matrizetik. Orduan, distantzietan oinarrituriko erregresio eredua honela definitzen da:

$$\mathbf{Y} = \beta_0\mathbf{1} + \mathbf{X}_{(k)}\boldsymbol{\beta}_{(k)} + \mathbf{e}. \tag{12}$$

Karratu txikienen estimazioak honakoak dira: $\hat{\beta}_0 = y^-$ eta $\hat{\boldsymbol{\beta}}_{(k)} = \boldsymbol{\Lambda}_k^{-1}\,\mathbf{X}_{(k)}^{-1}\mathbf{Y}$.

Datuak $q$ dimentsiotan badaude, hau da, $rg\{\mathbf{H}\,(-\frac{1}{2}\delta_{ij}^2)\mathbf{H}\} = q$ bada, $k = q$ ezarriz (11) ereduak (12) ereduarekin bat egiten du eta distantzietan oinarrituriko eredu osoa esaten zaio. Honela, erregresio eredu arrunta distantzietan oinarrituriko eredu gisa ikus daiteke, beti ere Euklides distantzia erabilita. Bestetik ordea, (12) ereduak edozein distantzia mota erabiltzea baimentzen du. Bereziki, polinomio ortogonalen gaineko erregresioaren baliokidea da [9] baldin eta erabilitako distantzia $\delta_{ij} = (z_i - z_j)^{1/2}$ bada $p = 1$ izanik. [3, 10] lanetan ikus daitezke emaitza osagarriak. Horretaz gain, aldagaien aukeraketarako $F$ testaren parekoa den distantzietan oinarritutako testa [11] proposatu zen. Berrikiago, erregresio funtzionalerako distantzietan oinarrituriko eredua [12] proposatu da.

### 3.2. **Diskriminazio-analisia**

Izan bitez $C_1$, …, $C_k$, $k$ populazio zeinak hurrenez hurren $\mathbf{Z}_1$, …, $\mathbf{Z}_k$ zorizko bektoreen bidez adieraziak diren; era berean, izan bedi $\mathbf{z}$ ale berria. Ale berri hori zein populaziotik datorren erabakitzeko diskriminazio-arau ugari

dago literaturan. Izan bedi $\delta$ distantzia egokia kasuko datu motarako. Orduan, distantzietan oinarrituriko diskriminazio-araua honela definitzen da [13, 6]:

Esleitu $\mathbf{z}$ alea $C_r$ populaziora, baldin eta $\phi_\delta(\mathbf{z}, C_r) = \min_{s=1,\ldots,k} \{\phi_\delta(\mathbf{z}, C_s)\}$ bada. (13)

Aipaturiko (3) emaitza kontuan hartuz, arau horrek oso interpretazio intuitiboa du: *Esleitu alea gertuen duen populaziora*. Gainera, diskriminazio-analisiko hainbat arau klasiko berreskuratzen ditu.

— Baldin eta $C_r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{I})$, $r = 1, 2$, badira eta $\delta$ Euklides distantzia bada, (13) arauak **diskriminatzaile euklidearrarekin** bat egiten du.
— Baldin eta $C_r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, $r = 1, 2$, badira, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ izanik, eta $\delta$ Mahalanobis distantzia bada, (13) arauak **Fisherren diskriminatzaile linealarekin** bat egiten du.
— Baldin eta $C_r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, $r = 1, 2$, badira, eta $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, **diskriminatzaile koadratikoarekin** bat egiten du (13) arauak baldin eta Mahalanobis distantzia, gehi konstate bat, erabilitzen badugu: $\delta^2 = (\mathbf{z}_1 - \mathbf{z}_2)' \boldsymbol{\Sigma}_r(\mathbf{z}_1 - \mathbf{z}_2) + log|\boldsymbol{\Sigma}_r|$, $\mathbf{z}_1 \neq \mathbf{z}_2$ ($r = 1, 2$). Erreparatu $\phi_\delta(\mathbf{z}, C_r)$ gertutasun funtzioen kalkuluan dauden bariantza-kobariantza matrizeak desberdinak direla.

Ezagunak diren beste arau batzuek ere berreskuratzen ditu distantzietan oinarrituriko diskriminazio-arauak [7]. Bestetik, arau horrek lagin bertsioa ere onartzen du (9)ren baitan eta ondorioz, populazioko banaketak ezezagunak direnean ere erabil daiteke.

$k > 2$ denean, eskuarki aldagai kanonikoekin ebazten da diskriminatzearen arazoa. Analisi kanonikoaren orokortze gisa uler daitekeen «Distantzietan oinarrituriko analisi kanonikoa» proposatzen dute [14] lanean. Datu ekologikoen gainean metodo hori aplikatu eta lortutako emaitzak eta, baita beste metodo batzuekin lortutakoen alderaketa ere, [15] lanean daude.

### 3.3. **Cluster-analisia**

Zuzenean distantzia matrizetik abiatzen den cluster-algoritmo ugari daude (ikus bitez [16] edo [17]), eta distantzia matrizetik abiatzen direnez, distantzietan oinarrituriko ak dira. Hala ere, algoritmo horiek ez dute aztertzen zein erlazio dagoen multzoaren aldakortasunarekin edo alea×aldagaia ikuspegiarekin. Bestetik, [8] lanean erakusten dute (10) aldakortasuna, bariantza-analisiko funtsezko identitatearen parekoa dena, baina ez dute lotzen cluster-analisiarekin. Bi alderdi horiek uztartu zituzten [18] lanean eta ondorengo emaitzak lortu zituzten, besteak beste:

— **GEVA-Ward**. Izan bitez $C$ multzoa eta $P = \{C_r : r = 1, \ldots, k\}$ bere partiketa $k$ clusterretan. $\sum_r n_r \hat{V}(C_r)$ espresioa $P$ partiketaren heterogeneota-

sun neurritzat har daiteke. Beraz, multzokatze algoritmo metakorren ikuspegitik egokia da $C_p$ eta $C_q$ clusterrak metatzea horiek badira aipaturiko espresioan handitze txikiena eragiten dutenak. Hau da, demagun algoritmoaren pauso batean $P = \{C_1, \ldots, C_p, \ldots, C_q, \ldots, C_k\}$ partiketa dugula $W(P) = \sum_{r=1}^{k} n_r \hat{V}(C_r)$ balioari lotuta. Hurrengo pausoan $Cp$ eta $Cq$ clusterrak metatuko ditugu eta $P' = \{C_1, \ldots, C_p \cup C_q, \ldots, C_k\}$ partiketa lortu, baldin eta dagokion $W(P')$ balioa minimoa bada. $C_p$ eta $C_q$ clusterren arteko distantzia honela definitzen da,

$$d(C_p, C_q) = (n_p + n_q)\hat{V}(C_p \cup C_q) - n_p\hat{V}(C_p) - n_q\hat{V}(C_q).$$

**Z** aldagaiak kuantitatiboak direnean eta $\delta$ Euklides distantzia, algoritmoak Warden sailkapen algoritmo klasikoarekin bat egiten du.

— **GEVA-Centroid**. Algoritmo metakor honetan $C_p$ eta $C_q$ clusterren arteko distantzia honela definitzen da,

$$d(C_p, C_q) = \hat{\Delta}^2 (C_p, C_q).$$

Aldagaiak kuantitatiboak direnean eta $\delta$ Euklides distantzia, algoritmoak zentroideen sailkapen algoritmo klasikoarekin bat egiten du.

Ondorioz, bai Warden algoritmoa eta bai zentroideena ere, egokiak dira, ez bakarrik aldagai kuantitatiboak dauden egoeretarako, baizik eta edozein egoeretarako, kasu bakoitzean $\delta$ distantzia egokia aukeratuz.

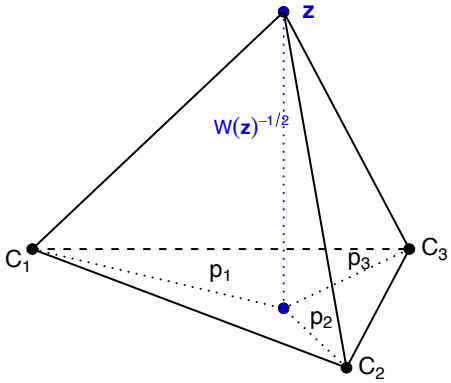
### 3.4. **Tipikotasuna**

$C_1, \ldots, C_k$ populazioak eta **z** ale berri bat emanik, alea aipaturiko zein populaziotik datorren erabakitzea da diskriminazio-analisiko helburu nagusia. Tipikotasunean, aldiz, alea aipaturiko populazioren batetik datorren ala ezagutzen ez den beste populazio batetik datorren erabakitzea da helburu nagusia. Helburu hori bete nahian, INCA estatistikoa honela definitzen da [19]:

$$W(\mathbf{z}) = \min_{\alpha_r}\{L(\mathbf{z})\}, \qquad \sum_{r=1}^{k}\alpha_r = 1 \tag{14}$$

$$\text{non} \quad L(\mathbf{z}) = \sum_{r=1}^{k}\alpha_r\phi^2(\mathbf{z}, C_r) - \sum_{1\le r\le s\le k}\alpha_r\alpha_s\Delta^2(C_r, C_s) \quad \text{den.}$$

Beraz, $W(\mathbf{z})$ estatistikoak populazioen barneko aldakortasuna ahalik eta txikiena izan dadin bilatzen du baina, baita, populazioen artekoa ahalik eta handiena izatea ere. Gainera, $W(\mathbf{z})$ estatistikoa geometrikoki interpreta daiteke, izan ere, **z** aletik $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ populazioen $\delta$-batezbestekoek osaturiko hiperplanora dagoen distantzia (karratua) da $W(\mathbf{z})$ (ikus 1. irudia $k = 3$ kasurako). Ondorioz, alea populazio horietarako ez dela tipikoa iradokitzen dute $W(\mathbf{z})$ ren balio handiek.




**1. irudia.** $W(\mathbf{z})$ estatistikoaren interpertazio geometrikoa $k = 3$ kasurako. Populazio bakoitzerako dagokion $\delta$-batezbestekoa erakutsi da.

Literaturan ezagunak diren beste estatistiko batzuekin bat egiten du INCAk. Hala nola, $C_r \sim N(\mu_r, \Sigma_r)$ denean eta $\Sigma_r = \Sigma$ $(r = 1, ..., k)$, $\delta$ Mahalanobis distantzia izanik, $W(\mathbf{z})$ estatistikoak [20] lanean proposaturikoarekin bat egiten du. Gainera, $k = 2$ kasuan, $W(\mathbf{z})$k [21] lanean proposaturikoa berreskuratzen du.

Behin $W(\mathbf{z})$ definituta, tipikotasun testa honela definitu eta gauzatzen da:

$$H_0 \quad : \quad \mathbf{z} \text{ alea } \sum_{r=1}^{k} \alpha_r E(\psi(\mathbf{Z}_r)) \text{ } \delta\text{-batezbestekoa duen populaziotik dator,}$$

$$H_1 \quad : \quad \mathbf{z} \text{ alea ezezaguna den populaziotik dator}$$

$W(\mathbf{z})$ adierazgarria bada, $\mathbf{z}$ alea ezezaguna den populaziotik datorrela onartuko dugu, bestela, ondorengo araua ezarriko da:

$$\mathbf{z} \text{ alea } C_r \text{ populazioan sailkatu, baldin eta } U_r(\mathbf{z}) = \min_{s=1,...,k} \{U_s(\mathbf{z})\} \text{ bada,} \quad (15)$$

non $U_s(\mathbf{z}) = \phi_s^2(\mathbf{z}) - W(\mathbf{z})$, $s = 1, ..., k$ diren.

Interpretazio geometrikoaren arabera, 1. Irudian erakusten den $p_r$ proiekzioa (karratua) da $U_r(\mathbf{z})$ eta (15) sailkapen arauak (13) arauarekin bat egiten du. Oro har, $W(\mathbf{z})$ren lagin-banaketa ez da ezaguna eta $N$ ale zoriz, itzulerarekin, $\bigcup_{r=1}^{k} C_r$ bilduratik aterata kalkulatzen da estatistikoaren $H_0$ peko banaketa.

Azkenik, aipatu INCA estatistikoa datu multzoan dagoen cluster kopura estimatzeko [19] ere erabil daitekeela.

### 3.5. **Sakonera**

Analisi anizkoitzean, *sakonera* terminoak aleen zentraltasunari egiten dio erreferentzia. Ale sakonena, medianaren kontzeptuaren orokortzetzat jotzen da. Era asko daude **Z** zorizko *p*-bektoreak adierazten duen *C* populazioarekiko **z** aleak duen sakonera neurtzeko. Bereziki, distantzietan oinarrituriko sakonera [22] honela definitzen da:

$$I(\mathbf{z},C)=\left[1+\frac{\phi^2(\mathbf{z},C)}{V_\delta(C)}\right]^{-1} \tag{16}$$

Funtzio horrek [0, 1] tartean hartzen ditu balioak eta *C* populazioarekiko zentraltasun neurri bat esleitzen du. (16) funtzioaren balio altuak ale sakonekin uztartzen dira eta horretarako, batetik kontuan hartzen du **z** aleak *C* populazioarekin duen distantzia baina bestetik, baita populazioa beraren aldakortasuna ere. Sakoneraren definizio horrek bat egiten du ezagunak diren beste definizio batzuekin.

— *C* populazioa $\mathbf{Z} = Z$ aldagai kuantitatibo bakarrak adierazten duenean eta $\delta$ Euklides distantzia denean, (16) funtsean **z-score** delakoa da: $I(z,C)=\left[1+\frac{(z-\mu)^2}{\sigma^2}\right]^{-1}$, non $\mu = E(Z)$ eta $\sigma^2 = VAR(Z)$ diren.

— $C \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denean eta $\delta$ Mahalanobis distantzia, (16) **Mahalanobis sakoneraren** [23] parekoa da: $I(\mathbf{z},C)=\left[1+\frac{(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})}{p}\right]^{-1}$.

Sakonera erlatiboa delako kontzeptua [22] ere definitu da. $C_r$, $r = 1, \ldots, k$ klaseak harturik, $C_r$ klaseari dagokion alearen sakonera erlatiboa, klase horrekiko dagokion sakoneraren eta gainontzeko klaseekiko duen sakonera handienaren arteko diferentzia da. Ondorioz, sakonera erlatibo negatiboa duen alearen kokapena, berea ez den klase batekiko zentralagoa da berearekiko duena baino.

## 4. **ADIBIDE ERREALA: INDUSTRI OLIOAK**

Laburbildu diren tekniken erabilera erakusteko, teknika horiek industri olioen inguruko datu multzo batean aplikatu ditugu. Industrian olioak lubrikaziorako erabiltzen dira; kasu honetan, motorren lubrikaziorako erabiltzen direnak ditugu, hain zuzen. Olioaren egoerak ona izan behar du baldin eta motorrean kalterik ez bada eragingo. Olioaren egoera bere basikotasuna zenbakiaren (BZ) arabera neurtzen da; izan ere, horrek erakusten du zein den olioaren gordekin alkalinoa. Motorrean sortzen diren produktu azidoek olioaren degradazioa dakarte ordea. Oro har, olioaren hasierako BZ % 50 gutxitzen bada olioaren degradazioa gertu dagoela onartu eta ondorioz

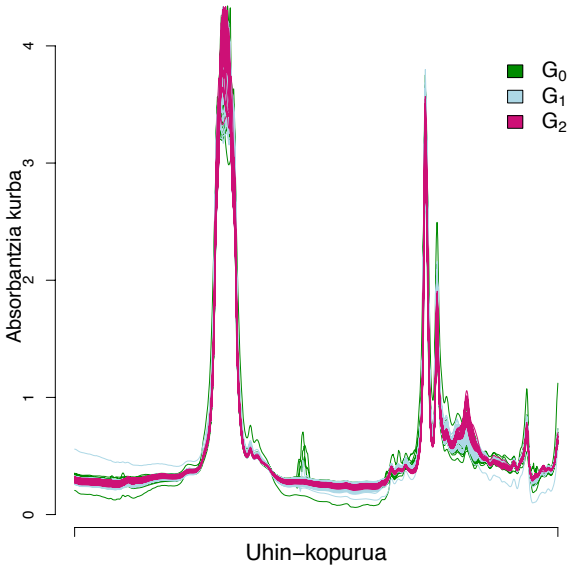
olioa aldatzea gomendatzen da. Hasierako BZ % 75 edo gehiago gutxitzen bada, olio aldaketa guztiz premiazkoa dela onartzen da.

**1. taula.** Olioen sailkapena, hasierako BZ gutxitzearen proportzioaren arabera.

| | Hasierako BZren gutxitzea | | |
|---|---|---|---|
| | 0-0.25 | 0.25-0.50 | 0.50-0.75 |
| $n$ | $G_0$ | $G_1$ | $G_2$ |
| 244 | 107 | 96 | 41 |

Datu multzo honetan 244 olio bildu dira (IK4-Tekniker), hiru multzotan sailkatuta bere BZren arabera (ikus 1. taula). Oraindik ere, $G_0$ eta $G_1$ multzoetako olioak dira egoera onean daudenak, eta komeniko litzateke aldiz $G_2$ multzokoak aldatzea. Ez dago egoera txarrean dagoen oliorik, hau da, ez dago hasierako BZ % 75 edo gehiago gutxitu zaionik.

Azterketa egiteko espektrograma jaso da olio bakoitzerako. Absorbantzian hartutako espektrograma, $4000cm^{-1}$-$500cm^{-1}$ tarteko 1751 neurketez osatuta (ikus 2. irudia).




**2. irudia.** Olioen absorbantzia-espektrogramak. Kurben koloreak olioaren BZ gutxitzearen araberakoak dira, $G_0$ (berdea), $G_1$ (urdina), $G_2$ (gorria).

Olio bakoitzerako espektrograma, izatez kurba bat, jaso denez, lan honetan olioak funtzio-datu gisa [24] aztertuko ditugu. Bi olio $i$ eta $j$ eta dagozkioen espetro-kurbak, $\chi_i(t)$ eta $\chi_j(t)$ harturik, $t \in (1, 1751)$, bi olioen arteko distantzia [25] honela neurtuko dugu:

$$\delta_{ij}^2 = \int (\chi_i'(t) - \chi_j'(t))^2 dt, \quad i, j = 1, \ldots, 244.$$

Azterketa R softwarearekin [26] egin da, bereziki `ICGE` [27] eta `npfda` [25] paketeekin.

**Diskriminzaio-analisia:** Jacknife zuzenketa kontuan hartuta, 2. taulak erakusten du zein den errakuntza matrizea (13) sailkapen araua erabiliz. Guztizko sailkatze okerra % 23.77 izan da. Multzoz multzo, ordea, % 20.56, % 31.25 eta % 14.63 izan dira hurrenez hurren. Ikus bedi gehien nahasten diren multzoak $G_0$ eta $G_1$ direla.

**2. taula.** Errakuntza matrizea (13) sailkapen arauarekin, Jacknife zuzenketa kontuan hartuta.

| | Jatorrizkoa BZren arabera | | |
|---|---|---|---|
| Esleitua | $G_0$ | $G_1$ | $G_2$ |
| $G_0$ | 85 | 25 | 0 |
| $G_1$ | 22 | 66 | 6 |
| $G_2$ | 0 | 5 | 35 |

**Cluster-analisia:** Aztertu da espektroaren arabera olioak nola banatzen diren eta baita lortutako clusterrak jatorriko $G_r$, $r = 0, 1, 2$ multzoekin nola uztartzen diren alderatu ere. Horretarako GEVA-Ward algoritmo metakorra aplikatu da. Hiru clusterreko partiketari dagozkien clusterrak eta hasierako BZ gutxitzearen araberako multzoen arteko erlazioa 3. taulan dago. Ikus dezakegunez, $C_1$ clusterreko aleen % 71.43 $G_0$ multzokoak dira eta gainontzekoak $G_1$ ekoak; $C_2$ clusterrean % 70.37 $G_1$ multzoko aleak dira, gainontzekoak $G_0$(% 20.99) eta $G_2$(% 4.94) clusterretan banaturik; azkenik, $C_3$ clusterra, batez ere, % 91.89, $G_2$ multzoko olioek osatzen dute, $G_1$eko gutxi batzuekin (% 8.1).

**Tipikotasuna:** Atal honetan tipikotasunaren arazoa azaldu da $k = 1$ egoerarako. $G_0$ multzoa ezaguntzat jo da eta tipikotasun-testa (3.4 atala) aplikatu zaie, bai $G_0$ multzoko olioei eta bai $G_1$ eta $G_2$ multzoko olioei ere,

**3. taula.** Olioen BZren gutxitzearen araberako banaketa eta GEVA-Wardekin 3 clusterreko partiketaren arteko erlazioa.

| | BZren arabera | | |
|---|---|---|---|
| Clusterrak | $G_0$ | $G_1$ | $G_2$ |
| $C_1$ | 90 | 36 | 0 |
| $C_2$ | 17 | 57 | 7 |
| $C_3$ | 0 | 3 | 34 |

$\alpha = 0.10$ finkatuta. Emaitzak hobeki ulertzeko, gogora dezagun $G_0$ multzoko olioak tipiko gisa sailkatu beharrekoak direla eta $G_1$ eta $G_2$ multzokoak berriz eztipiko gisa sailkatu beharrekoak. Espero zen bezala, $G_1$ multzoko olioak ez dira ongi bereizten $G_0$ multzokoetatik, % 14.58 soilik sailkatu da eztipiko gisa. Aldiz, $G_2$ multzoko olioen artean, olioen artean aldaketa beharko luketeenak, eztipiko bezala sailkatu da % 90.24 (4. taula).

**4. taula.** Tipikotasun-testaren emaitzak, $k = 1$ eta $\alpha = 0.1$rako, eta $G_0$ multzoa ezaguntzat jota.

| | Testatutakoak | | |
|---|---|---|---|
| | $G_0$ | $G_1$ | $G_2$ |
| Tipikoa $G_0$ rekiko | 96 | 82 | 4 |
| Ez tipikoa $G_0$ rekiko | 11 | 14 | 37 |
| | 107 | 96 | 41 |

**Sakoneren azteketa:** Olio bakoitzaren sakonera kalkulatu da dagokion multzoarekiko, $G_r$, $r = 0$, 1 edo 2, (16) adierazpenarekin. Multzo bakoitzeko kurba sakonenak konparatuz ondo ikus daiteke batetik $G_0$ eta $G_1$ multzokoak oso antzekoak direla eta bestetik, olioen arteko desberdintasunak gehien bat $1300cm^{-1} - 1100cm^{-1}$ heinean daudela. Espektroaren hein horretan amina eta fosfatoak pilatzen dira bereziki, gehigarriekin loturiko osagaiak, hain zuzen. Diferentzia horiek topatzea zentzuzko da, izan ere, gehigarriak baitira olioaren degradazio prozesuan sortzen diren azidoak neutralizatzen dituztenak. Multzo bakoitzeko sakonera handieneko eta txikieneko balioak kalkulatu dira. Sakonera balio handienak antzekoak dira hiru multzoetan (hurrenez hurren 0.987, 0.964 eta 0.989) baina balio txikienetan $G_2$ multzokoa da

handiena (0.083, 0.136 eta 0.191 hurrenez hurren). Sakonera erlatibo negatiboa duten olioen maiztasunak 21, 30 eta 6 dira $G_0$, $G_1$ eta $G_2$ multzoetan hurrenez hurren. Gainera, $G_0$ multzoko 21 olio horiek denak sakonagoak dira $G_1$ multzoarekiko; $G_1$ multzoko 30 olio horien artean 25 sakonagoak dira $G_0$ multzoarekiko eta 5 $G_2$ multzoarekiko; $G_2$ multzoko 6 olioak sakonagoak dira $G_1$ multzoarekiko. Emaitza horiek 5. Taulan laburbildu dira, sakonera erlatibo negatiboa duten olioak sakonen diren multzoan sailkatuz. Beraz, berriz ere ikusten da $G_0$ eta $G_1$ multzoak elkarren artean gehiago nahasten direla eta $G_2$ multzoa dela hobekien banantzen dena.

**5. taula.** Olioen BZren gutxitzearen araberako banaketa eta multzo sakonenaren arabera lortutako partizioaren arteko erlazioa.

| | BZren arabera | | |
|---|---|---|---|
| Multzo sakonena | $G_0$ | $G_1$ | $G_2$ |
| $G_0$ | 86 | 25 | 0 |
| $G_1$ | 21 | 66 | 6 |
| $G_2$ | 0 | 5 | 35 |

## 5. ONDORIOAK

Distantzietan oinarrituriko metodoek duten abantaila nagusietako bat da beraien aldaberatasuna, hau da, oso egoera desberdinetan erabiltzeko egokiak dira. Honela, bai datu mistoetan eta bai datuen azpiko banaketarik ezagutzen ez denean ere egokiak dira. Aldagai guztiak jarraituak diren egoeretan, distantzia egokia erabiliz, metodo hauek metodo klasikoekin bat egiten dute. Aldiz, egoera jakinetan, berariazko distantzia erabiliz, metodo hauek eskaintzen dituzten emaitzak ezin dira metodo klasikoen bidez lortu. Horregatik, alea×aldagaia ikuspegiak eskaintzen dituen aukeren osagarri direnak eskaintzen dituzte. Industria olioen azterketaren adibideak hemen laburbildutako metodoen erabilgarritasuna erakutsi du batetik, eta bestetik agerian utzi du era berean, metodo hauek edozein arlotako ikertzaileentzat interesgarriak izan daitezkeela.

## ERREFERENTZIAK

[1] GOWER J. C. 1966. «Some distance properties of latent root and vector methods used in multivariate analysis». *Biometrika,* **53**, 325-338.

[2] KRZANOWSKI W. J. eta MARRIOTT F. H. C. 1994. *Multivariate analysis. Part 1: Distributions, Ordination and Inference*. Kendall's Library of Statistics, Edward Arnold.

[3] CUADRAS C. M. eta ARENAS C. 1990. «A distance based regression model for prediction with mixed data». *Communications in Statistics A. Theory and Methods,* **19**, 2261-2279.

[4] LINGOES J. C. 1971. «Some boundary conditions for a monotone analysis of symmetric matrices». *Psychometrika,* **36**, 195-203.

[5] GOWER J. C. eta LEGENDRE P. 1986. «Metric and euclidean properties of dissimilarity coefficients». *Journal of Classification,* **3**, 5-48.

[6] CUADRAS C. M. eta FORTIANA J. 1995. «A continuous metric scaling solution for a random variable». *Journal of Multivariate Analysis,* **32**, 1-14.

[7] CUADRAS C. M., FORTIANA J. eta OLIVA F. 1997. «The proximity of an individual to a population with applications in discriminant analysis». *Journal of Classification,* **14**, 117-136.

[8] GOWER J. C. eta KRZANOWSKI W. J. 1999. «Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance». *Journal of the Royal Statistical Society: Series C (Applied Statistics*), **48**, 505-519.

[9] CUADRAS C. M. eta FORTIANA J. 1993. *Multivariate Analisys, Future Directions,* vol. 2, chap. Continuous metric scaling and prediction, pp. 47-66. Elsevier Science Publishers.

[10] CUADRAS C. M., ARENAS C. eta FORTIANA J. 1996. «Some computational aspects of distance-based model for prediction». *Communications in Statistics. Simulation and Computation,* **25**, 593-609.

[11] BOJ E., GRANE A., FORTIANA J. eta CLARAMUNT M. 2007. «Selection of predictors in distance-based regression». *Communications in Statistics B - Simulation and Computation,* **36**, 87-98.

[12] BOJ E., DELICADO P. eta FORTIANA J. 2010. «Distance-based local linear regression for functional predictors». *Computational Statistics and Data Analysis,* **54**, 429-437.

[13] CUADRAS C. M. 1989. *Statistical Data Analysis and Inference,* chap. Distance analysis in discrimination and classfication using both continuous and categorical variables, pp. 459-473. Elsevier Science Publishers B.V.

[14] ANDERSON M. J. eta ROBINSON J. 2003. «Generalized discriminant analysis based on distances». *Australian and New Zealand Journal of Statistics,* **45**, 301-318.

[15] ANDERSON M. J. eta WILLIS T. J. 2003. «Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology». *Ecology,* **84**, 511-525.

[16] KAUFMAN L. eta ROUSSEEUW P. 1990. *Finding groups in data. An introduction to cluster analysis* . Wiley.

[17] GORDON A. D. 1999. *Classification*. Monograph on Statistics and Applied Probability, Chapman and Hall, 2 edn.

[18] IRIGOIEN I., ARENAS C., FERNÁNDEZ E. eta MESTRES F. 2010. «GEVA: geometric variability-based approaches for identifying patterns in data». *Computational Statistics,* **25**, 241-255.

[19] IRIGOIEN I. eta ARENAS C. 2008. «INCA: New statistic for estimating the number of clusters and identifying atypical units». *Statistics in Medicine,* **27**, 2948-2973.

[20] RAO C. R. 1962. «Use of discriminant and allied functions in multivariate analysis». *Sankhya-Serie A*, **24**, 149-154.

[21] CUADRAS C. M. eta FORTIANA J. 2000. *Statistics for the 21st Century,* chap. The importance of geometry in multivariate analysis and some applications, pp. 93-108. Marcel Dekker.

[22] IRIGOIEN I., MESTRES F. eta ARENAS C. 2012. «The depth problem: identifying the most representative units in a data group». *IEEE/ACM Transactions on Computational Biology and Bioinformatics* .

[23] LIU R. eta SINGH K. 1993. «A quality index based on data depth and multivariate rank test». *Journal of the American Statistical Association,* **88**, 252-260.

[24] RAMSAY J. eta SILVERMAN B. 1997. *Functional Data Analysis*. Springer Series in Statistics.

[25] FERRATY F. eta VIEU P. 2006. *Nonparametric Functional Data Analysis. Theory and Practice*. Springer Series in Statistics.

[26] R CORE TEAM 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

[27] IRIGOIEN I., SIERRA B. eta ARENAS C. 2012. «ICGE: an r package for detecting relevant clusters and atypical units in gene expression». *BMC Bioinformatics,* **13**.