Análisis de componentes principales para observaciones tipo intervalo

Cervantes, J. [1]

[1]Universidad de Costa Rica

Presentado por:
Cervantes

Octubre, 2024

Presentation Outline

Table 1 - Faces Dataset (Distances AD,. . . ,GH as in Figure 6, see text)

Subject	$X_1 = \text{AD}$	$X_2 = \text{BC}$	$X_3 = \text{AH}$	$X_4 = \text{DH}$	$X_5 = \text{EH}$	$X_6 = \text{GH}$
FRA1	[155.00, 157.00]	[58.00, 61.01]	[100.45, 103.28]	[105.00, 107.30]	[61.40, 65.73]	[64.20, 67.80]
FRA2	[154.00, 160.01]	[57.00, 64.00]	[101.98, 105.55]	[104.35, 107.30]	[60.88, 63.03]	[62.94, 66.47]
FRA3	[154.01, 161.00]	[57.00, 63.00]	[99.36, 105.65]	[101.04, 109.04]	[60.95, 65.60]	[60.42, 66.40]
HUS1	[168.86, 172.84]	[58.55, 63.39]	[102.83, 106.53]	[122.38, 124.52]	[56.73, 61.07]	[60.44, 64.54]
HUS2	[169.85, 175.03]	[60.21, 64.38]	[102.94, 108.71]	[120.24, 124.52]	[56.73, 62.37]	[60.44, 66.84]
HUS3	[168.76, 175.15]	[61.40, 63.51]	[104.35, 107.45]	[120.93, 125.18]	[57.20, 61.72]	[58.14, 67.08]
INC1	[155.26, 160.45]	[53.15, 60.21]	[95.88, 98.49]	[91.68, 94.37]	[62.48, 66.22]	[58.90, 63.13]
INC2	[156.26, 161.31]	[51.09, 60.07]	[95.77, 99.36]	[91.21, 96.83]	[54.92, 64.20]	[54.41, 61.55]
INC3	[154.47, 160.31]	[55.08, 59.03]	[93.54, 98.98]	[90.43, 96.43]	[59.03, 65.86]	[55.97, 65.80]
ISA1	[164.00, 168.00]	[55.01, 60.03]	[120.28, 123.04]	[117.52, 121.02]	[54.38, 57.45]	[50.80, 53.25]
ISA2	[163.00, 170.00]	[54.04, 59.00]	[118.80, 123.04]	[116.67, 120.24]	[55.47, 58.67]	[52.43, 55.23]
ISA3	[164.01, 169.01]	[55.00, 59.01]	[117.38, 123.11]	[116.67, 122.43]	[52.80, 58.31]	[52.20, 55.47]
JPL1	[167.11, 171.19]	[61.03, 65.01]	[118.23, 121.82]	[108.30, 111.20]	[63.89, 67.88]	[57.28, 60.83]
JPL2	[169.14, 173.18]	[60.07, 65.07]	[118.85, 120.88]	[108.98, 113.17]	[62.63, 69.07]	[57.38, 61.62]
JPL3	[169.03, 170.11]	[59.01, 65.01]	[115.88, 121.38]	[110.34, 112.49]	[61.72, 68.25]	[59.46, 62.94]
KHA1	[149.34, 155.54]	[54.15, 59.14]	[111.95, 115.75]	[105.36, 111.07]	[54.20, 58.14]	[48.27, 50.61]
KHA2	[149.34, 155.32]	[52.04, 58.22]	[111.20, 113.22]	[105.36, 111.07]	[53.71, 58.14]	[49.41, 52.80]
KHA3	[150.33, 157.26]	[52.09, 60.21]	[109.04, 112.70]	[104.74, 111.07]	[55.47, 60.03]	[49.20, 53.41]
LOT1	[152.64, 157.62]	[51.35, 56.22]	[116.73, 119.67]	[114.62, 117.41]	[55.44, 59.55]	[53.01, 56.60]
LOT2	[154.64, 157.62]	[52.24, 56.32]	[117.52, 119.67]	[114.28, 117.41]	[57.63, 60.61]	[54.41, 57.98]
LOT3	[154.83, 157.81]	[50.36, 55.23]	[117.59, 119.75]	[114.04, 116.83]	[56.64, 61.07]	[55.23, 57.80]
PHI1	[163.08, 167.07]	[66.03, 68.07]	[115.26, 119.60]	[116.10, 121.02]	[60.96, 65.30]	[57.01, 59.82]
PHI2	[164.00, 168.03]	[65.03, 68.12]	[114.55, 119.60]	[115.26, 120.97]	[60.96, 67.27]	[55.32, 61.52]
PHI3	[161.01, 167.00]	[64.07, 69.01]	[116.67, 118.79]	[114.59, 118.83]	[61.52, 68.68]	[56.57, 60.11]
ROM1	[167.15, 171.24]	[64.07, 68.07]	[123.75, 126.59]	[122.92, 126.37]	[51.22, 54.64]	[49.65, 53.71]
ROM2	[168.15, 172.14]	[63.13, 68.07]	[122.33, 127.29]	[124.08, 127.14]	[50.22, 57.14]	[49.93, 56.94]
ROM3	[167.11, 171.19]	[63.13, 68.03]	[121.62, 126.57]	[122.58, 127.78]	[49.41, 57.28]	[50.99, 60.46]

(a) Conjunto de datos rostros

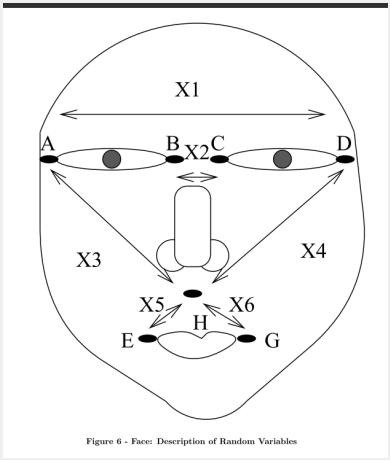


Figure 6 - Face: Description of Random Variables

(b) Medidas tomdas a los rostros

Figura 1: Ejemplo de datos sobre rostros

Esquema

- Definir un conjunto de datos sobre los que se va trabajar.
- Definir pesos.
- Matriz de varianza-covarianza.
- Determinación de autovalores y autovectores.
- Interpretación de los resultados.
- Implementación.

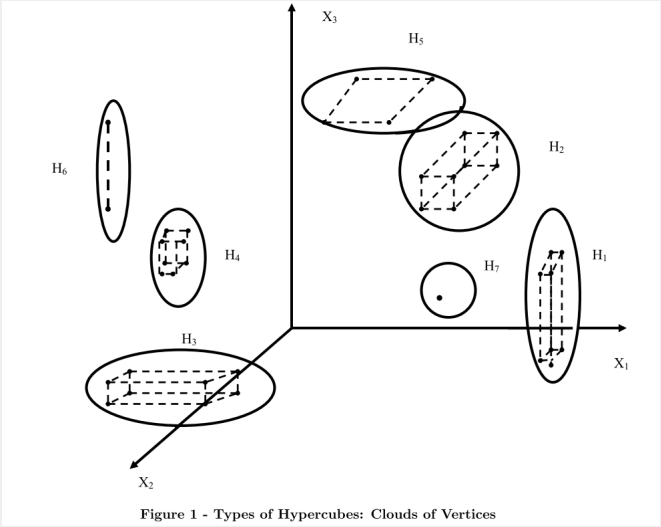


Figura 2: Representación gráfica del problema

Se supone que los datos cuenta con m observaciones y $\boldsymbol{\xi_i} = (\xi_{i1}, \ldots, \xi_{ip})$ donde

$$\xi_{ij} = [a_{ij}, b_{ij}], \quad i = 1, \ldots, m, \quad j = 1, \ldots, p$$

Se dice que un intervalo es trivial si $a_{ij} = b_{ij}$. Si q_i son los intervalos no triviales en $\boldsymbol{\xi_i}$ entonces el número de vértices de la observación es

$$n_i = 2^{q_i}$$

La cantidad de vértices para el conjuntos de datos es

$$n = \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} 2^{q_i}$$

Se construye la matriz $\mathbf{X}_{\xi_i}$

$$\begin{pmatrix} x^i_{11} & \cdots & x^i_{1p} \\ \vdots & & \vdots \\ x^i_{k1} & \cdots & x^i_{kp} \\ \vdots & & \vdots \\ x_{n^i_{i1}} & \cdots & x_{n^i_{ip}} \end{pmatrix}$$

Donde $\boldsymbol{x}^i_k = (x^i_{k1}, \ldots, x^i_{kp})$ es punto del vértice $k = 1, \ldots, n_i$ asociado al hipercubo H_i y representa la observación $\boldsymbol{\xi_i}$, $i = 1, ..., m$.

La matriz que representa completamente a el conjunto de datos es

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{\xi_1} \\ \vdots \\ \mathbf{X}_{\xi_m} \end{pmatrix} = \left(\begin{pmatrix} x_{11}^1 & \cdots & x_{1p}^1 \\ \vdots & & \vdots \\ x_{k1}^1 & \cdots & x_{kp}^1 \\ \vdots & & \vdots \\ x_{n_{i1}^i} & \cdots & x_{n_{ip}^i} \end{pmatrix} \right. \\ \vdots \\ \left. \begin{pmatrix} x_{11}^m & \cdots & x_{1p}^m \\ \vdots & & \vdots \\ x_{k1}^m & \cdots & x_{kp}^m \\ \vdots & & \vdots \\ x_{n_{i1}^m} & \cdots & x_{n_{ip}^m} \end{pmatrix} \right)$$

Una codificación alternativa puede ser aquela en la que se reemplazan los valores tipo intervalo $\boldsymbol{\xi_i}$ por el centro $\boldsymbol{x}_i^c = (x_{i1}^c, ..., x_{ip}^c)$ donde

$$x_{ij}^c = \frac{a_{ij} + b_{ij}}{2}$$

Entonces

$$\boldsymbol{X}^c = \begin{pmatrix} x_{11}^c & \cdots & x_{1p}^c \\ \vdots & & \vdots \\ x_{m1}^c & \cdots & x_{mp}^c \end{pmatrix}$$

Primero se establece que el peso de $\boldsymbol{\xi}_i$ es w_i, cada uno de los n_i vértices de $\boldsymbol{\xi}_i$ puede tener un peso w_i^k para $k = 1, ..., n_i$ y $i = 1, ..., m$ se requiere

$$w_i = \sum_{k=1}^{n_i} w_k^i, \quad \sum_{i=1}^{m} w_i = 1$$

Frecuentemente se define

$$w_i = \frac{1}{m}, \quad i = 1, ..., m$$

Sin embargo, esto omite la variación interna que existe en la observación. Entonces una alternativa puede ser

$$w_i = \frac{V_i}{\sum_{i=1}^{m} V_i}, \quad V_i = \prod_{a_{ij} \neq b_{ij}} (b_{ij} - a_{aij})$$

También se puede definir un tercer esquema en donde los pesos son inversamente proporcionales

$$w_i = \frac{1 - \frac{V_i}{\sum_{i=1}^{m} V_i}}{\sum_{i=1}^{m} 1 - \frac{V_i}{\sum_{i=1}^{m} V_i}}$$

Este tipo de forma de establecer los pesos es más apropiado si las observaciones son medidas de impresición.

Pesos para los vértices

Se puede asumir que los pesos para los n_i vertices son todos iguales de tal forma que

$$w_k^i = \frac{w_i}{n_i}, \quad k = 1, ..., n_i, \quad i = 1, ..., m$$

También, se pueden definir los pesos de acuerdo a un punto de referencia, puede ser el punto medio. De tal forma que si $a_{ij} < x_{ij}^0 < b_{ij}$. Con w_{ij}^a y w_{ij}^b el pesos en los puntos a_{ij} y b_{ij}

$$w_{ij}^a + w_{ij}^b = 1, \qquad w_{ij}^a a_{ij} + w_{ij}^b b_{ij} = x_{ij}^0$$

Entonces el peso w_i^k para el vértice k de $\boldsymbol{\xi}_i$ puede ser dado por

$$w_k^i = w_i \left[\prod_{j=1}^{q_i} w(x_{kj}^i) \right]$$

donde el peso asociado con la j-esima componente del vértice k es

$$w(x_{kj}^i) = w_{ij}^t, \quad \text{cuando} \quad x_{kj} = t_{ij}, \quad t = a, b$$

Ejemplo:

Para $\xi_i = ([a_{i1}, b_{i1}], [a_{i2}, b_{i2}])$. Para los 4 vértices del hipercubo H_i se cumple

$$w_1^i = w_i w_{i1}^a w_{i2}^a \quad w_2^i = w_i w_{i1}^a w_{i2}^b \quad w_3^i = w_i w_{i1}^b w_{i2}^a \quad w_4^i = w_i w_{i1}^b w_{i2}^b$$

Entonce lara matriz de pesos para $\boldsymbol{D}$ asociada con $\boldsymbol{X}$ es la matriz $n \times n$

$$\boldsymbol{D} = diag(w_1^1, ..., w_{n_1}^1, ..., w_{n_1}^m, ..., w_{n_m}^m)$$

Se define la matriz de varianza-covarianza relacionada con los vértices como $\boldsymbol{V}^*(v^*_{j_1 j_2})$, $j_1, j_2 = 1, ..., p$ como

$$\boldsymbol{V}^* = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X}$$

Se pueden obtener los promedios empíricos como

$$\bar{X}^*_j = \sum_{i=1}^{m} \sum_{k=1}^{n_i} w^i_k x^i_{kj} = \sum_{i=1}^{m} \alpha^a_{ij} a_{ij} + \alpha^b_{ij} b_{ij}$$

Donde se tiene que α^a_{ij} y α^b_{ij} son los pesos para la observación $\boldsymbol{\xi}_i$ cuando el valor x^i_{kj} es a_{ij} y b_{ij} respectivamente. Para $t = a, b$ entonces

$$\alpha^t_{ij} = \sum_{k=1}^{n_i} w^i_k = w^t_{ij} w_i, \quad \text{cuando} \quad \mathrm{x}^i_{kj} = t_{ij} \Rightarrow \alpha^a_{ij} + \alpha^b_{ij} = w_i$$

Entonces la varianza de v^*_{jj} de X_j puede ser reescrita como

$$v^*_{jj} = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w^i_k(x^i_{kj} - \bar{X}^*_j)^2 \Rightarrow v^*_{jj} = \sum_{i=1}^{m}[\alpha^a_{ij}(a_{ij} - \bar{X}^*_j)^2 + \alpha^b_{ij}(b_{ij} - \bar{X}^*_j)^2]$$

De forma análoga la covarianza $v^*_{j_1j_2}$ entre X_{j_1} y X_{j_2} puede ser reescrita como

$$v^*_{j_1j_2} = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w^i_k(x^i_{kj_1} - \bar{X}^*_{j_1})(x^i_{kj_w} - \bar{X}^*_{j_w})$$

Se puede demostrar que

$$v^*_{j_1j_2} = \sum_{i=1}^{m} w_i x^0_{ij_1} x^0_{ij_2}$$

Para matriz varianza-covarianza $\boldsymbol{X}^c$ la matriz de centros se define como

$$\boldsymbol{V}^c = (\boldsymbol{X}^c)^T \boldsymbol{D}^C \boldsymbol{X}\, c$$

con los elementos $(v^c_{j_1 j_2}), j_1, j_2 = 1, ..., p$. La varianza empírica sería v^c_{jj} de variable aleatoria X_j es

$$v^c_{jj} = \sum_{i=1}^{m} w_i (x^0_{ij} - \bar{X}^c_j)^2, \quad \text{con} \quad \bar{X}^c_j = \sum_{i=1}^{m} w_i x^0_{ij}$$

Si supone uniformidad en los intervalos y que todas las observaciones pesan igual entonces se tiene

$$v^c_{jj} = \sum_{i=1}^{m} w_i (w^a_{ij} a_{ij} + w^b_{ij} b_{ij})^2$$

De forma análoga la covarianza es

$$v^c_{j_1 j_2} = \sum_{i=1}^{m} w_i (x^0_{ij_1} - \bar{X}^c_{j_1})(x^0_{ij_2} - \bar{X}^c_{j_2}) = \sum_{i=1}^{m} w_i (w^a_{ij_1} a_{ij_1} + w^b_{ij_1} b_{ij_1})(w^a_{ij_2} a_{ij_2} + w^b_{ij_2} b_{ij_2})$$

Dado que se tienen la matriz $\mathbf{X} = (X_1, ..., X_p)$ y la matriz de varianza-covarianza $\boldsymbol{V}^*$ entonces se puede desarrollar un ACP clásico.
Se define $e_\nu = (e_{\nu 1}, ..., e_{\nu p})$, $\nu = 1, ..., p$ con $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ los autovectores y autovalores. Entonces

$$PC\nu = e_{\nu 1}X_1 + \cdots + e_{\nu p}X_p$$

Para observación $\boldsymbol{\xi}_i$ representada por los n_i vértices en $\boldsymbol{X}_{\xi_i}$ la ν-esima componen principal de los vértices se obtiene de

$$Y^*_{i\nu} = [y^a_{i\nu}, y^b_{i\nu}], \nu = 1, ..., s \geq p$$

Donde

$$y^a_{i\nu} = \min_{k \in L_i}\{y^i_{\nu k}\}, \qquad y^b_{i\nu} = \max_{k \in L_i}\{y^i_{\nu k}\}$$

donde $L_i = \{1, .., n_i\}$ es el conjunto de filas en $\boldsymbol{X}_{\xi_i}$ que describe los vértices del hipercubo $\boldsymbol{H}_i$ y $y^i_{\nu k}$ es el valor de la ν-esima componente principal para la fila k en L_i

Se puede demostrar que

$$y^a_{i\nu} = \sum_{j \in J^+}^{p} e_{\nu j}(a_{ij} - \bar{X}^*_j) + \sum_{j \in J^-}^{p} e_{\nu j}(b_{ij} - \bar{X}^*_j)$$

$$y^b_{i\nu} = \sum_{j \in J^-}^{p} e_{\nu j}(a_{ij} - \bar{X}^*_j) + \sum_{j \in J^+}^{p} e_{\nu j}(b_{ij} - \bar{X}^*_j)$$

Donde se tiene que $J^+ = \{j | e_{\nu j} > 0\}$ y $J^- = \{j | e_{\nu j} < 0\}$

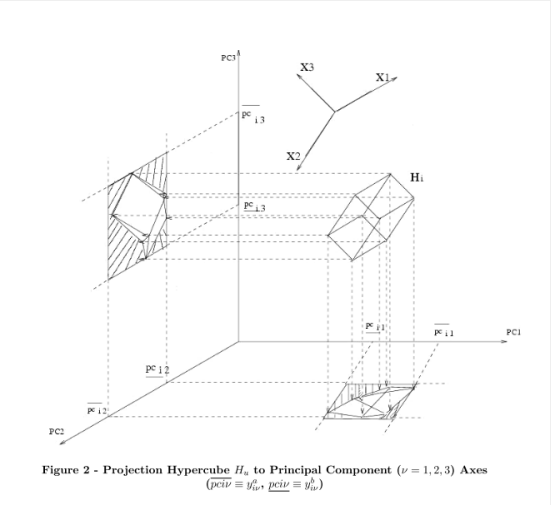


Figure 2 - Projection Hypercube H_u to Principal Component ($\nu = 1, 2, 3$) Axes ($\overline{pci\nu} \equiv y^a_{iu}$, $\underline{pci\nu} \equiv y^b_{iu}$)

Figura 3: Representación gráfica de la proyección en los planos

El resultado del min-max. Se puede obtener de la siguiente manera Si se toma cualquier punto $\tilde{\boldsymbol{x}}_i$ con $\tilde{x_{ij}} \in [a_{ij}, b_{ij}]$. Entonces la componente principal asociada con $\tilde{\boldsymbol{x}}_i$ es

$$\tilde{PC}\nu = \sum_{j=1}^{p} e_{ij}(\tilde{x}_{ij} - \bar{X}_j^*)$$

Se sigue que

$$\sum_{j=1}^{p} e_{ij}(\tilde{x}_{ij} - \bar{X}_j^*) \geq \sum_{j\in J^+}^{p} e_{\nu j}(a_{ij} - \bar{X}_j^*) + \sum_{j\in J^-}^{p} e_{\nu j}(b_{ij} - \bar{X}_j^*) = y_{i\nu}^a = \min_{k\in L_i}\{y_{\nu k}^i\}$$

y

$$\sum_{j=1}^{p} e_{ij}(\tilde{x}_{ij} - \bar{X}_j^*) \leq \sum_{j\in J^-}^{p} e_{\nu j}(a_{ij} - \bar{X}_j^*) + \sum_{j\in J^+}^{p} e_{\nu j}(b_{ij} - \bar{X}_j^*) = y_{i\nu}^b = \max_{k\in L_i}\{y_{\nu k}^i\}$$

Entonces para todo $\nu = 1, ..., p$

$$\tilde{PC}\nu \in [y^a_{i\nu}, y^b_{i\nu}]$$

Se puede obstener la correlación entre la componente principal $PC\nu$ y la variable aleatoria X_j como

$$C_{j\nu} = \text{Cor}(X_j, PC\nu) = e_{\nu k}\sqrt{\frac{\lambda_\nu}{\sigma_j^2}}$$

donde $e_{\nu j}$ es la componente el autovector $\boldsymbol{e}_\nu$ asociado con X_j y donde

$$\lambda_\nu = \text{Var}(PC_\nu)$$

es el ν-autovalor y σ_j^2 es la varianza de X_j

La alternativa es reemplazar $\boldsymbol{\xi}_i$ por los baricentros $\boldsymbol{x}_i^0 = (x_{i1}^0, ..., x_{ip}^0)$ donde

$$x_{ij}^0 = \sum_{k=1}^{n_i} \left(\frac{w_k^i}{w_i} x_{kj}^i \right) = x_{ij}^0 = w_{ij}^a a_{ij} + w_{ij}^b b_{ij}, \quad i = 1, ..., m, \quad j = 1, ..., p$$

Entonces se contruye la matriz

$$\begin{pmatrix} x_{11}^0 & \cdots & x_{1p}^0 \\ \vdots & & \vdots \\ x_{m1}^0 & \cdots & x_{mp}^0 \end{pmatrix}$$

La matriz de pesos es la matriz diagona $m \times m$

$$\boldsymbol{D}^c = \text{diag}(w_1, ..., w_m)$$

Con esto se realiza un análisis de componentes principales del tal forma que se obtiene $u_\nu = (u_{\nu 1}, ..., u_{\nu p})$ con $\nu = 1, ..., p$ siendo este ν-ésimo autovector y entonces la ν-ésima componente principal se puede escribir como

$$PC\nu^c = \sum_{j=1}^{p} (x_j^0 - \bar{X}_j^c) u_{\nu j}$$

En particular si se toma cualquier punto $\tilde{\boldsymbol{x}} = (\tilde{x}_{i1}, ..., \tilde{x}_{ip})$ cualquier punto en hipercubo H_i de $\boldsymbol{\xi}_i$ entonces se puede estimar la ν-ésima componente principal centrada como

$$PC\nu^c(\tilde{\boldsymbol{x}}_i) = \sum_{j=1}^{p} (\tilde{x}_{ij} - \bar{X}_j^c) u_{\nu j}$$

Entonces se puede definir la ν-esima componente principal a partir de los centros como

$$Z_{i\nu} = [z^a_{i\nu}, z^b_{i\nu}], \quad \nu = 1, ..., s \leq p$$

Donde

$$z^a_{i\nu} = \sum_{j=1}^{p} \min_{a_{ij} < \tilde{x}_{ij} < b_{ij}} \left\{ (\tilde{x}_{ij} - \bar{X}^c_j) u_{\nu j} \right\}$$

$$z^b_{i\nu} = \sum_{j=1}^{p} \max_{a_{ij} < \tilde{x}_{ij} < b_{ij}} \left\{ (\tilde{x}_{ij} - \bar{X}^c_j) u_{\nu j} \right\}$$

Se puede demostrar que

$$z_{i\nu}^{a} = \sum_{j \in J_c^-}^{p} (b_{ij} - \bar{X}_j) u_{\nu j} + \sum_{j \in J_c^+}^{p} (a_{ij} - \bar{X}_j) u_{\nu j}$$

$$z_{i\nu}^{b} = \sum_{j \in J_c^+}^{p} (b_{ij} - \bar{X}_j) u_{\nu j} + \sum_{j \in J_c^-}^{p} (a_{ij} - \bar{X}_j) u_{\nu j}$$

Donde $J_c^- = \{j | u_{\nu j} < 0\}$ y $J_c^+ = \{j | u_{\nu j} > 0\}$

Estima matriz de varianza-covarianza del método de los vértices es $O(m2^p)$ y el método de los centros es $O(m)$. El problema es que el método de los centros pierde información y el método de los vértices para p lo suficientemente grande la complejidad es alta.

Dicho lo anterior se quisiera conservar la mayor cantida de información como en el método de los centros, pero con la complejidad de los centros. Eso es lo que se tratará a continuación.

$$\bar{X}_j^c = \sum_{i=1}^{m} w_i x_{ij}^0 = \sum_{i=1}^{m} (\alpha_{ij}^a a_{ij} + \alpha_{ij}^b a b_{ij}), \text{ i.e } \bar{X}_j^c = \bar{X}_j^*$$

Sin embargo, las varianzas y covarianzas de tal forma que

$$\begin{aligned} & v_{jj}^* - v_{jj}^c \\ =& \sum_{i=1}^{m} [\alpha_{ij}^a (a_{ij} - \bar{X}_j^*)^2 + \alpha_{ij}^b (b_{ij} - \bar{X}_j^*)^2] - \sum_{i=1}^{m} w_i (w_{ij}^a a_{ij} + w_{ij}^b b_{ij})^2 \\ =& \sum_{i=1}^{m} w_i w_{ij}^a w_{ij}^b (b_{ij} - a_{ij})^2 = e_{jj} \end{aligned}$$

Si se compara la covarianza entre X_{j1} y X_{j2} con $j_1 \neq j_2$ para los métodos se tiene que

$$\begin{aligned} & v^c_{j_1 j_2} \\ =& \sum_{i=1}^{m} w_i (w^a_{ij_1} w^a_{ij_2} a_{ij_1} a_{ij_2} + w^a_{ij_1} w^b_{ij_2} a_{ij_1} b_{ij_2} \\ & + w^b_{ij_1} w^a_{ij_2} b_{ij_1} a_{ij_2} + w^b_{ij_1} w^b_{ij_2} b_{ij_1} b_{ij_2}) \\ =& v^*_{j_1 j_2} \end{aligned}$$

Entonces

$$\boldsymbol{V}^* = \boldsymbol{V}^c + \boldsymbol{E}$$

Donde $\boldsymbol{E}$ es una matriz $p \times p$ con elementos en la diagonal iguales a e_{jj}. Con esto se puede estimar $\boldsymbol{V}^*$ con complejidad $O(m)$.

La contribución del hipercubo H_i se puede cuantificar como

$$C^1_{i\nu} = Ctr(H_i, PC\nu) = w_i \sum_{k=1}^{n_i} \frac{w^i_k (y^i_{\nu k})^2}{[d(\boldsymbol{x}^i_k, \boldsymbol{G})]^2}$$

donde $y^i_{\nu k}$ es la ν-ésima componente principal para el vértice k, w^i_k es el peso del vértice y $d(\boldsymbol{x}^i_k, \boldsymbol{G})$ es la distancia euclidea entre el vértice $\boldsymbol{x}^i_k$ en la fila k de $X_{\boldsymbol{\xi}_i}$ y $\boldsymbol{G}$ definido como el centroide de todas las n rilas de $\boldsymbol{X}$.

Una segunda medida de contribución es

$$C_{i\nu}^2 = Ctr(H_i, PC\nu) = \frac{\sum_{k=1}^{n_i} w_k^i (y_{\nu k}^i)^2}{\sum_{k=1}^{n_i} w_k^i [d(\boldsymbol{x}_k^i, \boldsymbol{G})]^2}$$

La primera medida identifica el promedio del coseno al cuadrado entre los vértices y el eje de la ν-ésima componente principal. La segunda es el ratio entre la contribución de los vértices a la varianza explicada por λ_ν. La contribución absoluta a la varianza de λ_ν y la inercia total se cuantifica por

$$I_{i\nu} = Inertia(H_i, PC\nu) = \left[\sum_{k=1}^{n_i} w_k^i (y_{\nu k}^i)^2\right] / \lambda_\nu$$

$$I_i = Inertia(H_i) = \left[\sum_{k=1}^{n_i} w_k^i [d(\boldsymbol{x}_k^i, \boldsymbol{G})^2]\right] / I_T$$

Una ayuda visual se puede obterner al considerar solo aquellos vértices los cuales aportan un valor superior a un umbral α a la componente principal $PC\nu$

$$Y^*_{i\nu}(\alpha) = [y^a_{i\nu}(\alpha), y^b_{i\nu}(\alpha)]$$

Donde se tiene que

$$y^a_{i\nu}(\alpha) = \min_{k \in L_i}\{y^i_{\nu k} | Ctr(\boldsymbol{x}^i_k, PC\nu) \geq \alpha\} \qquad y^b_{i\nu}(\alpha) = \max_{k \in L_i}\{y^i_{\nu k} | Ctr(\boldsymbol{x}^i_k, PC\nu) \geq \alpha\}$$

$$Ctr(\boldsymbol{x}^i_k, PC\nu) = \frac{(y^i_{\nu k})^2}{[d(\boldsymbol{x}^i_k, \boldsymbol{G})]^2}$$

o

$$Ctr(\boldsymbol{x}^i_k, PC\nu_1, PC\nu_2) = Ctr(\boldsymbol{x}^i_k, PC\nu_1) + Ctr(\boldsymbol{x}^i_k, PC\nu_2)$$

Sea $Z = (z_{ij})_{i=1,2,...,m}$ con

$$z_{ij} = \frac{1}{\sqrt{m}} \frac{x_{ij}^c - \bar{X}_j^c}{\sigma_j^c}$$

De forma análoga se definen $\overline{z}_{ij}$ y $\underline{z}_{ij}$. Se sabe que ZZ^t y Z^tZ tienen los mismos q autovalores estrictamente positivos $\lambda_1, ..., \lambda_q$. Si $u_1, ..., u_q$ son los autovectores de Z^tZ y $v_1, ..., v_q$ son los autovectores de ZZ^t se puede demostrar que

$$u_l = \frac{Z^t v_l}{\sqrt{\lambda_l}}, \quad l = 1, .., q$$

$$v_l = \frac{Z u_l}{\sqrt{\lambda_l}}, \quad l = 1, .., q$$

$x_{ij}^c \leftarrow \frac{\underline{x}_{ij}+\overline{x}_{ij}}{2}$

$z_{ij} \leftarrow \frac{1}{\sqrt{m}} \frac{x_{ij}^c - \bar{X^c}_j}{\sigma_j^c}$

$H \leftarrow Z^t Z$

$\underline{z}_{ij} \leftarrow \frac{1}{\sqrt{m}} \frac{\underline{x}_{ij} - \bar{X^c}_j}{\sigma_j^c}$

$\overline{z}_{ij} \leftarrow \frac{1}{\sqrt{m}} \frac{\overline{x}_{ij} - \bar{X^c}_j}{\sigma_j^c}$

Se estiman los autovectores de H $v_1, ..., v_q$ y los autovalores $\lambda_1, ..., \lambda_q$

for $i = 1, ..., m$ **do**

for $j = 1, ..., q$ **do**

$$\underline{R}(X^i, Y^j) = \sum_{k=1, v_{kj}<0}^{m} \overline{z}_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \underline{z}_{ki} v_{kj}$$

$$\overline{R}(X^i, Y^j) = \sum_{k=1, v_{kj}>0}^{m} \overline{z}_{ki} v_{kj} + \sum_{k=1, v_{kj}<0}^{m} \underline{z}_{ki} v_{kj}$$

for $i = 1, ..., m$ **do**
    for $j = 1, ..., q$ **do**

$$u_{ij} = \frac{1}{\sqrt{\lambda_j}} \left(\sum_{k=1}^{m} z_{ik} v_{kj} \right)$$

for $i = 1, ..., m$ **do**
    for $j = 1, ..., q$ **do**

$$\underline{y}_{ij} = \sum_{k=1, u_{kj}<0}^{n} \overline{z}_{ik} u_{kj} + \sum_{k=1, u_{kj}>0}^{n} \underline{z}_{ik} u_{kj}$$

$$\overline{y}_{ij} = \sum_{k=1, u_{kj}<0}^{n} \underline{z}_{ik} u_{kj} + \sum_{k=1, u_{kj}>0}^{n} \overline{z}_{ik} u_{kj}$$

Ejemplo: Rostros

Table 2 - Vertices Principal Components, $\nu = 1, 2, 3$: Faces

Subject	$PC1$	$PC2$	$PC3$
FRA1	[-2.66, -1.61]	[0.27, 1.57]	[-0.29, 1.00]
FRA2	[-2.49, -1.03]	[-0.11, 1.61]	[-0.25, 1.01]
FRA3	[-2.99, -0.81]	[-0.40, 1.88]	[-0.88, 1.20]
HUS1	[-0.24, 1.10]	[0.39, 2.05]	[0.64, 2.13]
HUS2	[-0.40, 1.41]	[0.56, 2.65]	[0.29, 2.32]
HUS3	[-0.24, 1.42]	[0.43, 2.52]	[0.27, 2.17]
INC1	[-3.77, -2.29]	[-0.67, 1.23]	[-0.80, 0.69]
INC2	[-3.66, -1.35]	[-2.05, 0.92]	[-0.88, 1.83]
INC3	[-4.02, -1.86]	[-1.20, 1.41]	[-1.01, 1.50]
ISA1	[0.80, 2.00]	[-1.83, -0.46]	[-0.58, 0.58]
ISA2	[0.37, 1.86]	[-1.71, -0.08]	[-0.64, 0.73]
ISA3	[0.41, 2.11]	[-1.84, -0.12]	[-0.58, 1.20]
JPL1	[-0.36, 0.92]	[0.54, 2.03]	[-1.81, -0.43]
JPL2	[-0.34, 1.17]	[0.48, 2.37]	[-1.85, -0.07]
JPL3	[-0.52, 0.93]	[0.50, 2.28]	[-1.56, 0.25]
KHA1	[-1.18, 0.39]	[-3.07, -1.46]	[-1.19, 0.26]
KHA2	[-1.46, 0.15]	[-3.17, -1.32]	[-0.93, 0.61]
KHA3	[-1.71, 0.25]	[-2.95, -0.72]	[-1.25, 0.57]
LOT1	[-0.74, 0.61]	[-2.51, -0.87]	[-0.81, 0.61]
LOT2	[-0.69, 0.40]	[-1.94, -0.62]	[-0.80, 0.33]
LOT3	[-0.82, 0.34]	[-2.12, -0.70]	[-0.77, 0.52]
PHI1	[0.22, 1.51]	[0.56, 1.84]	[-1.40, -0.08]
PHI2	[-0.09, 1.66]	[0.33, 2.29]	[-1.81, 0.22]
PHI3	[-0.25, 1.38]	[0.25, 2.25]	[-2.01, -0.12]
ROM1	[2.19, 3.45]	[-1.20, 0.29]	[-0.51, 0.81]
ROM2	[1.85, 3.63]	[-1.30, 0.97]	[-0.83, 1.36]
ROM3	[1.48, 3.57]	[-1.33, 1.31]	[-0.79, 1.79]

(a) Componentes principales, método de vértices sin

Table 4 - Vertices Principal Components, $\nu = 1, 2$, $\alpha = 0.2$: Faces

	Principal Component		# Vertices Retained	
Subject	$PC1$	$PC2$	$\nu = 1$	$\nu = 2$
FRA1	[-2.66, -1.61]	[1.12, 1.57]	64	12
FRA2	[-2.49, -1.03]	[0.94, 1.61]	64	18
FRA3	[-2.99, -0.81]	[0.67, 1.87]	64	17
HUS1	[0.87, 1.10]	[0.81, 2.05]	3	49
HUS2	[0.86, 1.41]	[0.97, 2.65]	6	56
HUS3	[0.68, 1.42]	[0.88, 2.52]	11	50
INC1	[-3.77, -2.29]	0.28	64	0
INC2	[-3.66, -1.35]	[-2.05, -1.64]	64	8
INC3	[-4.02, -1.85]	0.11	64	0
ISA1	[0.80, 2.00]	[-1.83, -0.70]	64	51
ISA2	[0.67, 1.86]	[-1.71, -0.51]	52	38
ISA3	[0.66, 2.11]	[-1.84, -0.46]	60	41
JPL1	[0.92, 0.92]	[0.60, 2.03]	1	60
JPL2	[0.64, 1.17]	[0.79, 2.37]	7	57
JPL3	[0.81, 0.93]	[0.59, 2.28]	3	60
KHA1	-0.39	[-3.07, -1.46]	0	64
KHA2	[-1.46, -1.09]	[-3.17, -1.32]	4	64
KHA3	[-1.71, -0.83]	[-2.95, -0.72]	12	64
LOT1	-0.07	[-2.61, -0.87]	0	64
LOT2	-0.14	[-1.94, -0.62]	0	64
LOT3	-0.24	[-2.12, -0.70]	0	64
PHI1	[0.63, 1.51]	[0.63, 1.84]	36	59
PHI2	[0.62, 1.66]	[0.66, 2.29]	26	51
PHI3	[0.62, 1.38]	[0.62, 2.25]	18	54
ROM1	[2.19, 3.45]	-0.46	64	0
ROM2	[1.85, 3.63]	-0.17	64	0
ROM3	[1.48, 3.57]	[1.28, 1.31]	64	2

(b) Componentes principales, método de vértices con

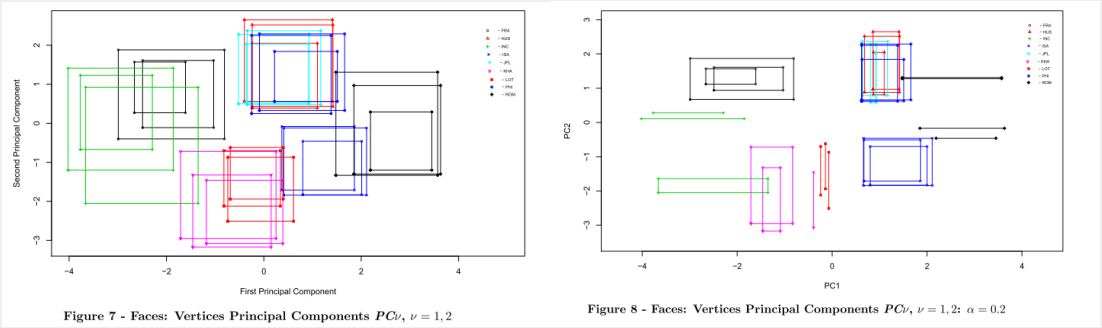


Figure 7 - Faces: Vertices Principal Components $PC\nu$, $\nu = 1, 2$

Figure 8 - Faces: Vertices Principal Components $PC\nu$, $\nu = 1, 2$: $\alpha = 0.2$

(a) Sin restricción sobre contribución

(b) Con restricción sobre contribución

Figura 5: Plano principal ACP, método de los vértices

Table 5 - Vertices PC Inertia: Faces

$PC\nu$	Eigenvalue λ_ν	% Inertia	Cumulative Inertia
$PC1$	2.560	42.7	42.7
$PC2$	1.798	30.0	72.7
$PC3$	0.642	10.7	83.4
$PC4$	0.476	7.9	91.3
$PC5$	0.335	5.6	96.9
$PC6$	0.188	3.1	100

(a) Inercia de los autovalores

X_j	$PC1$	$PC2$	$PC3$
AD	0.6444	0.5889	0.1717
BC	0.4903	0.6663	-0.1403
AH	0.8374	-0.1968	-0.3707
DH	0.8913	0.0885	0.1649
EH	-0.4749	0.6248	-0.5607
GH	-0.4283	0.7554	0.3377

(b) Correlación

Figura 6: Medidas importantes ACP, método de los vértices

Table 2 - Vertices Principal Components, $\nu = 1, 2, 3$: Faces

Subject	$PC1$	$PC2$	$PC3$
FRA1	[-2.66, -1.61]	[0.27, 1.57]	[-0.29, 1.00]
FRA2	[-2.49, -1.03]	[-0.11, 1.61]	[-0.25, 1.01]
FRA3	[-2.99, -0.81]	[-0.40, 1.88]	[-0.88, 1.20]
HUS1	[-0.24, 1.10]	[0.39, 2.05]	[0.64, 2.13]
HUS2	[-0.40, 1.41]	[0.56, 2.65]	[0.29, 2.32]
HUS3	[-0.24, 1.42]	[0.43, 2.52]	[0.27, 2.17]
INC1	[-3.77, -2.29]	[-0.67, 1.23]	[-0.80, 0.69]
INC2	[-3.66, -1.35]	[-2.05, 0.92]	[-0.88, 1.83]
INC3	[-4.02, -1.86]	[-1.20, 1.41]	[-1.01, 1.50]
ISA1	[0.80, 2.00]	[-1.83, -0.46]	[-0.58, 0.58]
ISA2	[0.37, 1.86]	[-1.71, -0.08]	[-0.64, 0.73]
ISA3	[0.41, 2.11]	[-1.84, -0.12]	[-0.58, 1.20]
JPL1	[-0.36, 0.92]	[0.54, 2.03]	[-1.81, -0.43]
JPL2	[-0.34, 1.17]	[0.48, 2.37]	[-1.85, -0.07]
JPL3	[-0.52, 0.93]	[0.50, 2.28]	[-1.56, 0.25]
KHA1	[-1.18, 0.39]	[-3.07, -1.46]	[-1.19, 0.26]
KHA2	[-1.46, 0.15]	[-3.17, -1.32]	[-0.93, 0.61]
KHA3	[-1.71, 0.25]	[-2.95, -0.72]	[-1.25, 0.57]
LOT1	[-0.74, 0.61]	[-2.51, -0.87]	[-0.81, 0.61]
LOT2	[-0.69, 0.40]	[-1.94, -0.62]	[-0.80, 0.33]
LOT3	[-0.82, 0.34]	[-2.12, -0.70]	[-0.77, 0.52]
PHI1	[0.22, 1.51]	[0.56, 1.84]	[-1.40, -0.08]
PHI2	[-0.09, 1.66]	[0.33, 2.29]	[-1.81, 0.22]
PHI3	[-0.25, 1.38]	[0.25, 2.25]	[-2.01, -0.12]
ROM1	[2.19, 3.45]	[-1.20, 0.29]	[-0.51, 0.81]
ROM2	[1.85, 3.63]	[-1.30, 0.97]	[-0.83, 1.36]
ROM3	[1.48, 3.57]	[-1.33, 1.31]	[-0.79, 1.79]

(a) ACP, método de los vértices

Table 9 - Centers Principal Components, $\nu = 1, 2, 3$: Faces

	$PC1$	$PC2$	$PC3$
FRA1	[-2.969, -1.747]	[0.236, 1.722]	[-0.376, 1.033]
FRA2	[-2.729, -1.111]	[-0.197, 1.789]	[-0.372, 1.067]
FRA3	[-3.286, -0.856]	[-0.532, 2.077]	[-1.032, 1.301]
HUS1	[-0.374, 1.162]	[0.376, 2.284]	[0.772, 2.419]
HUS2	[-0.583, 1.482]	[0.592, 2.975]	[0.355, 2.592]
HUS3	[-0.403, 1.506]	[0.461, 2.822]	[0.384, 2.418]
INC1	[-4.065, -2.381]	[-0.902, 1.289]	[-0.905, 0.744]
INC2	[-3.909, -1.213]	[-2.509, 0.933]	[-0.954, 2.022]
INC3	[-4.332, -1.837]	[-1.495, 1.469]	[-1.115, 1.615]
ISA1	[0.881, 2.239]	[-2.063, -0.477]	[-0.603, 0.708]
ISA2	[0.388, 2.038]	[-1.933, -0.073]	[-0.688, 0.868]
ISA3	[0.444, 2.355]	[-2.088, -0.112]	[-0.618, 1.388]
JPL1	[-0.567, 0.888]	[0.667, 2.379]	[-2.068, -0.536]
JPL2	[-0.593, 1.167]	[0.575, 2.764]	[-2.101, -0.145]
JPL3	[-0.782, 0.916]	[0.593, 2.645]	[-1.805, 0.206]
KHA1	[-1.097, 0.641]	[-3.507, -1.653]	[-1.280, 0.368]
KHA2	[-1.429, 0.386]	[-3.645, -1.505]	[-0.993, 0.743]
KHA3	[-1.730, 0.468]	[-3.393, -0.817]	[-1.346, 0.714]
LOT1	[-0.794, 0.742]	[-2.864, -0.977]	[-0.874, 0.707]
LOT2	[-0.773, 0.472]	[-2.205, -0.687]	[-0.879, 0.376]
LOT3	[-0.928, 0.408]	[-2.435, -0.778]	[-0.848, 0.586]
PHI1	[0.114, 1.574]	[0.722, 2.178]	[-1.582, -0.030]
PHI2	[-0.270, 1.740]	[0.454, 2.689]	[-2.017, 0.220]
PHI3	[-0.450, 1.440]	[0.356, 2.671]	[-2.258, -0.168]
ROM1	[2.407, 3.838]	[-1.270, 0.436]	[-0.549, 0.902]
ROM2	[1.961, 4.041]	[-1.394, 1.200]	[-0.899, 1.491]
ROM3	[1.529, 3.978]	[-1.436, 1.585]	[-0.874, 1.918]

(b) ACP, método de los centros

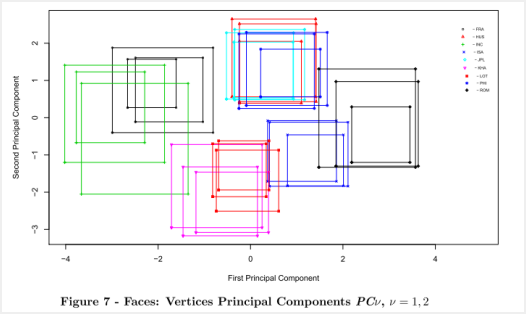


Figure 7 - Faces: Vertices Principal Components $PC\nu$, $\nu = 1, 2$

(a) ACP, método de los vértices

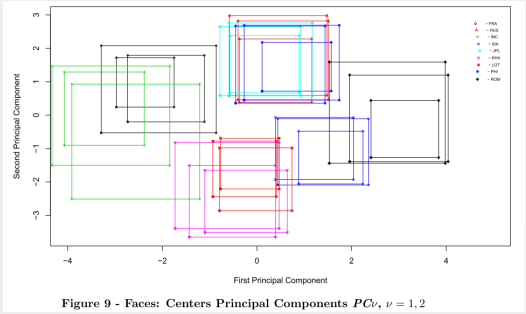


Figure 9 - Faces: Centers Principal Components $PC\nu$, $\nu = 1, 2$

(b) ACP, método de los centros

Figura 8: Comparativa método de los centros y método de los vértices

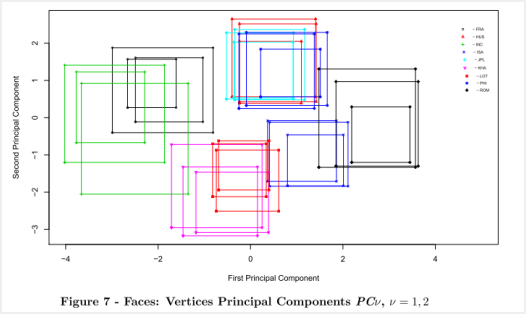


Figure 7 - Faces: Vertices Principal Components $PC\nu$, $\nu = 1, 2$

(a) ACP, método de los vértices

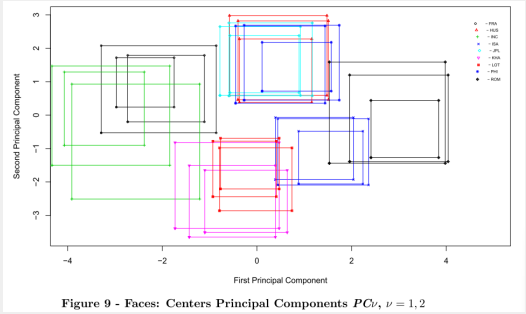


Figure 9 - Faces: Centers Principal Components $PC\nu$, $\nu = 1, 2$

(b) ACP, método de los centros

Figura 9: Comparativa método de los centros y método de los vértices

	GRA	FRE	IOD	SAP
Linsed (L)	[0.93, 0.935]	[−27, −18]	[170, 204]	[118, 196]
Perilla (P)	[0.93, 0.937]	[−5, −4]	[192, 208]	[188, 197]
Cotton (Co)	[0.916, 0.918]	[−6, −1]	[99, 113]	[189, 198]
Sesame (S)	[0.92, 0.926]	[−6, −4]	[104, 116]	[187, 193]
Camellia (Ca)	[0.916, 0.917]	[−25, −15]	[80, 82]	[189, 193]
Olive (O)	[0.914, 0.919]	[0, 6]	[79, 90]	[187, 196]
Beef (B)	[0.86, 0.87]	[30, 38]	[40, 48]	[190, 199]
Hog (H)	[0.858, 0.864]	[22, 32]	[53, 77]	[190, 202]

Table 1: Oils and Fats data table.

(a) Tabla de aceites y grasas Ichino

	PC1	PC2	PC3	PC4
L	[1.275, 4.733]	[−1.353, 4.428]	[−1.025, 1.289]	[−0.989, 0.989]
P	[1.059, 1.701]	[−1.128, −0.343]	[−1.508, −1.046]	[−0.134, 0.334]
Co	[−0.236, 0.399]	[−0.969, −0.213]	[−0.170, 0.368]	[−0.246, 0.204]
S	[0.154, 0.658]	[−0.745, −0.179]	[−0.027, 0.342]	[−0.369, 0.028]
Ca	[0.151, 0.613]	[−0.881, −0.437]	[0.807, 1.204]	[0.113, 0.538]
O	[−0.594, 0.100]	[−0.775, 0.043]	[0.019, 0.545]	[−0.645, −0.101]
B	[−3.046, −2.226]	[0.234, 1.162]	[−0.392, 0.152]	[−0.530, 0.193]
H	[−2.900, −1.841]	[0.020, 1.135]	[−0.729, 0.171]	[−0.105, 0.720]

Table 4: Principal components with Duality Center Method.

(b) ACP, método de los centros

Figura 10: Comparativa método de los centros y método de los vértices

	GRA	FRE	IOD	SAP
Linsed (L)	[0.93, 0.935]	[−27, −18]	[170, 204]	[118, 196]
Perilla (P)	[0.93, 0.937]	[−5, −4]	[192, 208]	[188, 197]
Cotton (Co)	[0.916, 0.918]	[−6, −1]	[99, 113]	[189, 198]
Sesame (S)	[0.92, 0.926]	[−6, −4]	[104, 116]	[187, 193]
Camellia (Ca)	[0.916, 0.917]	[−25, −15]	[80, 82]	[189, 193]
Olive (O)	[0.914, 0.919]	[0, 6]	[79, 90]	[187, 196]
Beef (B)	[0.86, 0.87]	[30, 38]	[40, 48]	[190, 199]
Hog (H)	[0.858, 0.864]	[22, 32]	[53, 77]	[190, 202]

Table 1: Oils and Fats data table.

(a) Tabla de aceites y grasas de Ichino

	PC1	PC2	PC3	PC4
L	[1.275, 4.733]	[−1.353, 4.428]	[−1.025, 1.289]	[−0.989, 0.989]
P	[1.059, 1.701]	[−1.128, −0.343]	[−1.508, −1.046]	[−0.134, 0.334]
Co	[−0.236, 0.399]	[−0.969, −0.213]	[−0.170, 0.368]	[−0.246, 0.204]
S	[0.154, 0.658]	[−0.745, −0.179]	[−0.027, 0.342]	[−0.369, 0.028]
Ca	[0.151, 0.613]	[−0.881, −0.437]	[0.807, 1.204]	[0.113, 0.538]
O	[−0.594, 0.100]	[−0.775, 0.043]	[0.019, 0.545]	[−0.645, −0.101]
B	[−3.046, −2.226]	[0.234, 1.162]	[−0.392, 0.152]	[−0.530, 0.193]
H	[−2.900, −1.841]	[0.020, 1.135]	[−0.729, 0.171]	[−0.105, 0.720]

Table 4: Principal components with Duality Center Method.

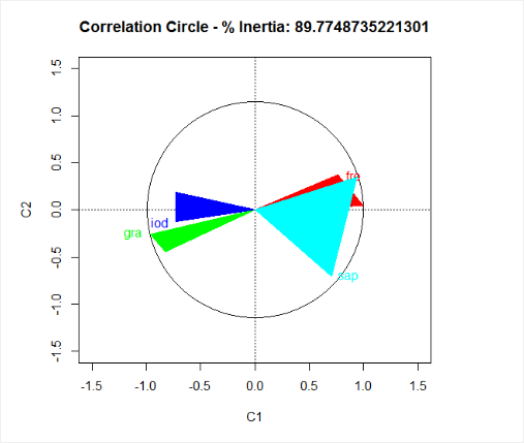
(b) Componente principales de tabla de aceites y grasas Ichino

Figura 11: Tabla de aceites y grasas

	PC1	PC2	PC3	PC4
GRA	[0.827, 1.000]	[−0.443, −0.265]	[−0.038, 0.087]	[−0.238, −0.084]
FRE	[−1.000, −0.760]	[0.044, 0.372]	[−0.428, −0.220]	[−0.288, 0.019]
IOD	[0.726, 1.000]	[−0.124, 0.191]	[−0.565, −0.401]	[−0.024, 0.161]
SAP	[−1.000, 0.190]	[−1.000, 0.371]	[−0.442, 0.163]	[−0.231, 0.325]

Table 2: Symbolic correlations between the variables and principal components with Duality Center Method.

(a) Correlaciones de ACP a tabla de aceites y grasas Ichino



(b) Circulo de correlaciones de ACP a tabla de aceites y grasas Ichino

Figura 12: Correlaciones ACP mediante método de los centros a tabla de aceites y grasas Ichino

Conclusiones

- El método de los centros es un análisis que se realiza entre los observaciones, mientras que el método de los vértices es un análisis entre las observaciones y también sobre si mismas.
- Se vio que el la complejidad del método de los centros y el método de los vértices es la misma.
- En el trabajo se presenta el concepto de la contribución de vértices cuestión que no se puede tratar con los análisis clásicos.
- Se presentó la dualidad en el caso del ACP simbólico con el método de los centros.
- Los autores consideran que dado el crecimiento de los data sets es importante desarrollar métodos para tratar con objetos como histogramas y multivalores.