Propuestas Robustas para el Análisis de Correlaciones Canónicas

Stella Maris Donato[1] Jorge G. Adrover[2]

[1]Universidad de Buenos Aires
[2]Universidad Nacional de Córdoba, CIEM y CONICET.

22 de setiembre de 2016

Esquema

- Revisión de Correlaciones Canónicas.

Esquema

- Revisión de Correlaciones Canónicas.
- Estimadores propuestos.

Esquema

- Revisión de Correlaciones Canónicas.
- Estimadores propuestos.
- Estudio de simulación.

Esquema

- Revisión de Correlaciones Canónicas.
- Estimadores propuestos.
- Estudio de simulación.
- Conclusiones.

Esquema

- Revisión de Correlaciones Canónicas.
- Estimadores propuestos.
- Estudio de simulación.
- Conclusiones.
- Referencias.

Revisión de Correlaciones Canónicas

Dados dos grupos de variables, el primero conformado por p variables y representado a través del vector aleatorio $\mathbf{x} = (x_1, x_2, ..., x_p)^t$, y el segundo formado por q variables y representado a través del vector aleatorio $\mathbf{y} = (y_1, y_2, ..., y_q)^t$. Es conveniente considerarlos conjuntamente, para lo cual se define el vector $\mathbf{z}$

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = (x_1, x_2, ..., x_p, y_1, y_2, ..., y_q)^t,$$

que tiene vector de medias

$$\boldsymbol{\mu} = E(\mathbf{z}) = \begin{pmatrix} E(\mathbf{x}) \\ E(\mathbf{y}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{pmatrix}$$

y matriz de covarianzas

$$\Sigma = E\left[(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^t\right] = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix}.$$

Revisión de Correlaciones Canónicas

Clásicamente, el primer par de vectores canónicos $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, ..., \alpha_{1p})^t$ y $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, ..., \beta_{1q})^t$ son los que resulven

$$(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1) = \underset{(\mathbf{a},\mathbf{b})\in\mathcal{A}_1}{\arg\max} Corr\left(\mathbf{a}^t\mathbf{x}, \mathbf{b}^t\mathbf{y}\right),$$

donde

$$\mathcal{A}_1 = \left\{ (\mathbf{a}, \mathbf{b}) \in \mathbb{R}^p \times \mathbb{R}^q : \begin{array}{l} Cov\left(\mathbf{a}^t\mathbf{x}, \mathbf{a}^t\mathbf{x}\right) = \mathbf{a}^t\Sigma_{\mathbf{xx}}\mathbf{a} = 1, \\ Cov\left(\mathbf{b}^t\mathbf{y}, \mathbf{b}^t\mathbf{y}\right) = \mathbf{b}^t\Sigma_{\mathbf{yy}}\mathbf{b} = 1 \end{array} \right\}.$$

Si el rango de $\Sigma_{\mathbf{xy}} > 1$, se buscan vectores canónicos de orden superior $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, ..., \alpha_{kp})^t$ y $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, ..., \beta_{kq})^t$.

Revisión de Correlaciones Canónicas

La definición es recursiva para $k = 2, 3, \ldots, min(p, q)$. Entonces, dado los primeros $k-1$ vectores canónicos $(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1), ..., (\boldsymbol{\alpha}_{k-1}, \boldsymbol{\beta}_{k-1})$, se define

$$(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = \underset{(\mathbf{a},\mathbf{b})\in\mathcal{A}_k}{\arg\max} Corr\left(\mathbf{a}^t\mathbf{x}, \mathbf{b}^t\mathbf{y}\right), \tag{1}$$

donde

$$\mathcal{A}_k = \left\{ (\mathbf{a}, \mathbf{b}) \in \mathbb{R}^p \times \mathbb{R}^q : \begin{array}{ll} Cov\left(\mathbf{a}^t\mathbf{x}, \mathbf{a}^t\mathbf{x}\right) = 1, & Cov\left(\mathbf{a}^t\mathbf{x}, \boldsymbol{\alpha}_j^t\mathbf{x}\right) = 0, \\ Cov\left(\mathbf{b}^t\mathbf{y}, \mathbf{b}^t\mathbf{y}\right) = 1, & Cov\left(\mathbf{b}^t\mathbf{y}, \boldsymbol{\beta}_j^t\mathbf{y}\right) = 0, \\ j = 1, 2, ..., k-1 & \end{array} \right\}$$

Revisión de Correlaciones Canónicas

Este planteo tiene una solución conocida

- Los vectores canónicos $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), k = 1, ..., \min(p, q)$ son los autovectores correspondientes a los autovalores $\lambda_1 \geq ... \geq \lambda_{\min(p,q)}$ de las matrices

$$\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}} \qquad y \qquad \Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}, \tag{2}$$

  respectivamente

Revisión de Correlaciones Canónicas

Este planteo tiene una solución conocida

- Los vectores canónicos $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), k = 1, ..., \min(p, q)$ son los autovectores correspondientes a los autovalores $\lambda_1 \geq ... \geq \lambda_{\min(p,q)}$ de las matrices

$$\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}} \qquad y \qquad \Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}, \tag{2}$$

  respectivamente

- Y las correlaciones canónicas están dadas por

$$\left[Corr\left(\boldsymbol{\alpha}_k^t\mathbf{x}, \boldsymbol{\beta}_k^t\mathbf{y}\right)\right]^2 = \lambda_k \tag{3}$$

Revisión de Correlaciones Canónicas

El análisis canónico se puede plantear en forma alternativa buscando las matrices $A^* \in \mathbb{R}^{r\times p}$ y $B^* \in \mathbb{R}^{r\times q}$ $(r \leq \min(p,q))$, y el vector $\mathbf{a}^* \in \mathbb{R}^r$ tales que:

$$(A^*, B^*, \mathbf{a}^*) = \underset{(A,B,\mathbf{a})\in\mathcal{C}}{\arg\min}\, E\left(\|A\mathbf{x} - B\mathbf{y} - \mathbf{a}\|^2\right) \tag{4}$$

donde

$$\mathcal{C} = \left\{(A, B, \mathbf{a}) : A \in \mathbb{R}^{r\times p}, B \in \mathbb{R}^{r\times q}, \mathbf{a} \in \mathbb{R}^r, A\Sigma_{\mathbf{xx}}A^t = I = B\Sigma_{\mathbf{yy}}B^t\right\}. \tag{5}$$

Cuya solución es la siguiente (Ver Seber (1984) y Brillinger (1975)):

- A^* y B^* son matrices que tienen en sus filas los autovectores correspondientes a los autovalores $\lambda_1 \geq ... \geq \lambda_r$ de las matrices en (2),

Revisión de Correlaciones Canónicas

El análisis canónico se puede plantear en forma alternativa buscando las matrices $A^* \in \mathbb{R}^{r\times p}$ y $B^* \in \mathbb{R}^{r\times q}$ $(r \leq \min(p,q))$, y el vector $\mathbf{a}^* \in \mathbb{R}^r$ tales que:

$$(A^*, B^*, \mathbf{a}^*) = \underset{(A,B,\mathbf{a})\in\mathcal{C}}{\arg\min}\, E\left(\|A\mathbf{x} - B\mathbf{y} - \mathbf{a}\|^2\right) \tag{4}$$

donde

$$\mathcal{C} = \left\{(A, B, \mathbf{a}) : A \in \mathbb{R}^{r\times p}, B \in \mathbb{R}^{r\times q}, \mathbf{a} \in \mathbb{R}^r, A\Sigma_{\mathbf{xx}}A^t = I = B\Sigma_{\mathbf{yy}}B^t\right\}. \tag{5}$$

Cuya solución es la siguiente (Ver Seber (1984) y Brillinger (1975)):

- A^* y B^* son matrices que tienen en sus filas los autovectores correspondientes a los autovalores $\lambda_1 \geq ... \geq \lambda_r$ de las matrices en (2),
- $\mathbf{a}^* = A^*\mu_{\mathbf{x}} - B^*\mu_{\mathbf{y}}$

Revisión de Correlaciones Canónicas

El análisis canónico se puede plantear en forma alternativa buscando las matrices $A^* \in \mathbb{R}^{r\times p}$ y $B^* \in \mathbb{R}^{r\times q}$ $(r \leq \min(p,q))$, y el vector $\mathbf{a}^* \in \mathbb{R}^r$ tales que:

$$(A^*, B^*, \mathbf{a}^*) = \underset{(A,B,\mathbf{a})\in\mathcal{C}}{\arg\min}\, E\left(\|A\mathbf{x} - B\mathbf{y} - \mathbf{a}\|^2\right) \tag{4}$$

donde

$$\mathcal{C} = \left\{(A, B, \mathbf{a}) : A \in \mathbb{R}^{r\times p}, B \in \mathbb{R}^{r\times q}, \mathbf{a} \in \mathbb{R}^r, A\Sigma_{\mathbf{xx}}A^t = I = B\Sigma_{\mathbf{yy}}B^t\right\}. \tag{5}$$

Cuya solución es la siguiente (Ver Seber (1984) y Brillinger (1975)):

- A^* y B^* son matrices que tienen en sus filas los autovectores correspondientes a los autovalores $\lambda_1 \geq ... \geq \lambda_r$ de las matrices en (2),
- $\mathbf{a}^* = A^*\boldsymbol{\mu}_{\mathbf{x}} - B^*\boldsymbol{\mu}_{\mathbf{y}}$
- Y las correlaciones canónicas cumplen la ecuación (3).

Estimadores propuestos - SM

- Los estimadores propuestos en este trabajo están basados en la idea de minimizar la distancia entre las variables canónicas y este problema de optimización tiene relación con Componentes Principales (PCA).

Estimadores propuestos - SM

- Los estimadores propuestos en este trabajo están basados en la idea de minimizar la distancia entre las variables canónicas y este problema de optimización tiene relación con Componentes Principales (PCA).
- Sea $A_o = A\Sigma_{\mathbf{xx}}^{1/2}$, $B_o = B\Sigma_{\mathbf{yy}}^{1/2}$, $D = (A_o - B_o) \in \mathbb{R}^{r \times m}$, $m = p + q$, y el vector aleatorio $\mathbf{z} = (\mathbf{x}^t \Sigma_{\mathbf{xx}}^{-1/2}, \mathbf{y}^t \Sigma_{\mathbf{yy}}^{-1/2})^t$.

Estimadores propuestos - SM

- Los estimadores propuestos en este trabajo están basados en la idea de minimizar la distancia entre las variables canónicas y este problema de optimización tiene relación con Componentes Principales (PCA).
- Sea $A_o = A\Sigma_{\mathbf{xx}}^{1/2}, B_o = B\Sigma_{\mathbf{yy}}^{1/2}, D = (A_o - B_o) \in \mathbb{R}^{r\times m}, m = p + q$, y el vector aleatorio $\mathbf{z} = (\mathbf{x}^t\Sigma_{\mathbf{xx}}^{-1/2}, \mathbf{y}^t\Sigma_{\mathbf{yy}}^{-1/2})^t$.
- Reformulando la expresión para los vectores estandarizados $\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x}$ y $\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y}$, se obtiene

$$\min_{(A,B,\mathbf{a})\in\mathcal{C}} E\left(\|A\mathbf{x} - B\mathbf{y} - \mathbf{a}\|^2\right) = \min_{(D,\mathbf{a})\in\mathcal{B}_{r,m}} E\left(\|D\mathbf{z} - \mathbf{a}\|^2\right)$$

con

$$\mathcal{B}_{r,m} = \left\{(D, \mathbf{a}) : \mathbf{a} \in \mathbb{R}^r, D = (A_o - B_o) \in \mathbb{R}^{r\times m}, A_oA_o^t = I_r = B_oB_o^t\right\}$$

que es similar al problema de PCA que trata Maronna (2005).

Estimador propuesto - SM

Notar que

- Si se estandarizan los vectores $\mathbf{x}$ e $\mathbf{y}$,

$$\mathbf{z} = \begin{pmatrix} \Sigma_{\mathbf{xx}}^{-1/2} \left(\mathbf{x} - E\left(\mathbf{x}\right)\right) \\ \Sigma_{\mathbf{yy}}^{-1/2} \left(\mathbf{y} - E\left(\mathbf{y}\right)\right) \end{pmatrix},$$

la matriz de covarianzas resultante para $\mathbf{z}$ es

$$M = \begin{pmatrix} I & \Sigma_{\mathbf{xx}}^{-1/2} \Sigma_{\mathbf{xy}} \Sigma_{\mathbf{yy}}^{-1/2} \\ \Sigma_{\mathbf{yy}}^{-1/2} \Sigma_{\mathbf{yx}} \Sigma_{\mathbf{xx}}^{-1/2} & I \end{pmatrix}.$$

Estimador propuesto - SM

Notar que

- Si se estandarizan los vectores $\mathbf{x}$ e $\mathbf{y}$,

$$\mathbf{z} = \begin{pmatrix} \Sigma_{\mathbf{xx}}^{-1/2} \left(\mathbf{x} - E\left(\mathbf{x}\right)\right) \\ \Sigma_{\mathbf{yy}}^{-1/2} \left(\mathbf{y} - E\left(\mathbf{y}\right)\right) \end{pmatrix},$$

  la matriz de covarianzas resultante para $\mathbf{z}$ es

$$M = \begin{pmatrix} I & \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2} \\ \Sigma_{\mathbf{yy}}^{-1/2}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1/2} & I \end{pmatrix}.$$

- Si $M \begin{pmatrix} \mathbf{v}^t & \mathbf{w}^t \end{pmatrix}^t = \lambda \begin{pmatrix} \mathbf{v}^t & \mathbf{w}^t \end{pmatrix}^t, 0 < \lambda \neq 1$, entonces

$$\begin{aligned} \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2}\Sigma_{\mathbf{yx}} \left(\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{v}\right) &= (\lambda - 1)^2 \left(\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{v}\right) \\ \Sigma_{\mathbf{yy}}^{-1/2}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}} \left(\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{w}\right) &= (\lambda - 1)^2 \left(\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{w}\right), \end{aligned}$$

Estimador propuesto

- Luego, si $\begin{pmatrix} \mathbf{v}^t & \mathbf{w}^t \end{pmatrix}^t$ es un autovector de M, entonces $\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{v}$ y $\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{w}$ son vectores canónicos relacionados con los grupos de variables $\mathbf{x}$ e $\mathbf{y}$, respectivamente.

Estimador propuesto

- Luego, si $\begin{pmatrix} \mathbf{v}^t & \mathbf{w}^t \end{pmatrix}^t$ es un autovector de M, entonces $\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{v}$ y $\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{w}$ son vectores canónicos relacionados con los grupos de variables $\mathbf{x}$ e $\mathbf{y}$, respectivamente.
- Y, si λ_j es un autovalor de M asociado con el autovector $\begin{pmatrix} \mathbf{v}_j^t & \mathbf{w}_j^t \end{pmatrix}^t$, entonces $|\lambda_j - 1|$ es una correlación canónica relacionada con los grupos de variables $\mathbf{x}$ e $\mathbf{y}$.

Estimador propuesto

- Luego, si $\begin{pmatrix} \mathbf{v}^t & \mathbf{w}^t \end{pmatrix}^t$ es un autovector de M, entonces $\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{v}$ y $\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{w}$ son vectores canónicos relacionados con los grupos de variables **x** e **y**, respectivamente.
- Y, si λ_j es un autovalor de M asociado con el autovector $\begin{pmatrix} \mathbf{v}_j^t & \mathbf{w}_j^t \end{pmatrix}^t$, entonces $|\lambda_j - 1|$ es una correlación canónica relacionada con los grupos de variables **x** e **y**.

Luego

ESTA CONEXIÓN NOS PERMITE CALCULAR LOS VECTORES CANÓNICOS A TRAVÉS DE UN ALGORITMO PARA COMPONENTES PRINCIPALES.

Estimadores propuestos - SM

Para acotar el efecto de *"casos" atípicos*, se considera una M-escala robusta σ y se define los **SM-vectores canónicos estandarizados robustos** como

$$(A_{SMo}, B_{SMo}, \mathbf{a}_{SM}) = \arg \min_{(A_o, B_o, \mathbf{a}) \in \mathcal{B}_{r,m}} \sigma(A_o, B_o, \mathbf{a}) \tag{6}$$

con $\sigma = \sigma(A_o, B_o, \mathbf{a})$ definida por la ecuación

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{\| A_o \widetilde{\mathbf{x}}_i - B_o \widetilde{\mathbf{y}}_i - \mathbf{a} \|^2}{\sigma} \right) = \delta, \tag{7}$$

donde $\rho : [0, \infty) \longrightarrow [0, 1]$, es no decreciente y derivable, $\rho(0) = 0$, $\lim_{x \to \infty} \rho(x) = 1$ y $0 < \delta < 1$. $\widetilde{\mathbf{x}} = \left(\widehat{\Sigma}_{\mathbf{xx}}^{(R)} \right)^{-1/2} \mathbf{x}$, $\widetilde{\mathbf{y}} = \left(\widehat{\Sigma}_{\mathbf{yy}}^{(R)} \right)^{-1/2} \mathbf{y}$.
Y, los **SM-vectores canónicos robustos** se definen de la siguiente manera,

$$A_{SM} = A_{SMo} \left(\widehat{\Sigma}_{\mathbf{xx}}^{(R)} \right)^{-1/2} \quad , \quad B_{SM} = B_{SMo} \left(\widehat{\Sigma}_{\mathbf{yy}}^{(R)} \right)^{-1/2} .$$

Estimador propuesto - SM

Consistencia Fisher

Si tomamos funcionales de dispersión $S_{\mathbf{x}}$ y $S_{\mathbf{y}}$ que son consistentes Fisher para $\Sigma_{\mathbf{xx}}$ y $\Sigma_{\mathbf{yy}}$ respectivamente, Luego, si se considera $\widetilde{\mathbf{x}}=S_{\mathbf{x}}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})$ e $\widetilde{\mathbf{y}}=S_{\mathbf{y}}^{-1/2}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})$ y buscamos las soluciones σ de

$$E\left[\rho\left(\frac{\|A\widetilde{\mathbf{x}}-B\widetilde{\mathbf{y}}-\mathbf{a}\|^2}{\sigma}\right)\right]=\delta.$$

Si tomamos la descomposición espectral de M dada por $M=\sum_{i=1}^{p+q}\gamma_i\mathbf{t}_i\mathbf{t}_i^t$, con $\gamma_1>\ldots>\gamma_r>\gamma_{r+1}\geq\ldots\geq\gamma_{p+q-r}>\gamma_{p+q-r+1}>\ldots>\gamma_{p+q}$, $\mathbf{t}_i'\mathbf{t}_j=\delta_{ij}, 1\leq i,j\leq p+q$, con

$$\delta_{ij}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}$$

Estimador propuesto - SM

Consistencia Fisher

Como $\mathbf{t}_i = (\mathbf{v}_i^t, \mathbf{w}_i^t)^t, \mathbf{v}_i \in \mathbb{R}^p, \mathbf{w}_i \in \mathbb{R}^q, i = 1, ..., p+q$, .si llamamos

$$A_o = \left(\frac{\mathbf{v}_{p+q-r+1}}{\|\mathbf{v}_{p+q-r+1}\|}, ..., \frac{\mathbf{v}_{p+q}}{\|\mathbf{v}_{p+q}\|} \right)^t \in \mathbb{R}^{r \times p},$$

$$B_o = \left(\frac{\mathbf{w}_{p+q-r+1}}{\|\mathbf{w}_{p+q-r+1}\|}, ..., \frac{\mathbf{w}_{p+q}}{\|\mathbf{w}_{p+q}\|} \right)^t \in \mathbb{R}^{r \times q},$$

$$\mathbf{a}_o = A_o \Sigma_{\mathbf{xx}}^{-1/2} \boldsymbol{\mu}_{\mathbf{x}} - B_o \Sigma_{\mathbf{yy}}^{-1/2} \boldsymbol{\mu}_{\mathbf{y}}.$$

En el siguiente teorema se establece la consistencia Fisher del estimador SM para vectores canónicos para familias elípticas con función densidad

$$f(\mathbf{z}, \boldsymbol{\mu}_0, \Sigma_0) = \frac{1}{\sqrt{|\Sigma_0|}} f_0((\mathbf{z} - \boldsymbol{\mu}_0)^t \Sigma_0^{-1} (\mathbf{z} - \boldsymbol{\mu}_0)),$$

donde $f_0 : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ es una función no creciente.

Estimador propuesto - SM

Consistencia Fisher

Y si además se verifica que:

- f_0 y ρ cumplen que existe un punto d en un intervalo no degenerado I de la intersección de los dominios tal que

$$\rho(u) < \rho(d) < \rho(v) \qquad \text{y} \qquad f_0(u) > f_0(d) > f_0(v)$$

para todo u y v en I, con $u < d < v$.

- **Teorema 1.** *Sea* $\mathbf{Z}$ *un vector aleatorio con distribución elíptica. Luego, los estimadores SM son estimadores funcionales consistentes Fisher, esto es*

$$(A_o, B_o, \mathbf{a}_o) = \arg \min_{\mathbf{a} \in \mathbb{R}^r, AA^t = I_r = BB^t} \sigma(A, B, \mathbf{a}).$$

- **Corolario 1**: *Sea* $\mathbf{Z}$ *un vector aleatorio con distribución elíptica. Si* ρ *es diferenciable, con* $\rho' = \psi$*, luego, para alguna constante* $c > 0$*,*

$$E_F\left[\psi\left(\frac{\left\|A_o\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x} - B_o\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y} - \mathbf{a}_o\right\|^2}{\sigma(A_o, B_o, \mathbf{a}_o)}\right)(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^t\right] = c\Sigma.$$

Estimador propuesto - SM

Consistencia

Los Teoremas que se enuncian a continuación establecen la consistencia de los estimadores SM de vectores canónicos para familias elípticas.

- **Teorema 2**. *Sea* $\mathbf{Z}$ *un vector aleatorio con densidad elíptica. Si* $m - r + 1 \leq n(1 - \delta)$, *el problema de encontrar* $(A, B, \mathbf{a}) \in B_{r,m}$ *que minimicen* $\sigma(A, B, a)$ *sujeto a (7) tiene al menos una solución con probabilidad 1.*
- **Teorema 3.** *Sea* $\mathbf{Z}$ *un vector aleatorio con densidad elíptica. Sea* $\left(A^{(n)}, B^{(n)}, \mathbf{a}^{(n)}\right)$ *una solución del problema planteado en el Teorema 2, entonces el estimador SM definido en (6) es un funcional consistente, esto es,*

$$\lim_{n\to\infty} A^{(n)} = A_o \quad , \quad \lim_{n\to\infty} B^{(n)} = B_o \quad y \quad \lim_{n\to\infty} \mathbf{a}^{(n)} = \mathbf{a}_o,$$

casi seguramente.

Estimador propuesto - SM

Función de Influencia

Sea F una familia de distribuciones elípticas, si ρ es dos veces diferenciable y, $S_{\mathbf{x}}$ y $S_{\mathbf{y}}$ son funcionales consistentes Fisher para $\Sigma_{\mathbf{xx}}$ y $\Sigma_{\mathbf{yy}}$ respectivamente.

Se muestra a continuación las gráficas de la norma de la función de influencia para el primer vector canónico $\mathbf{v}_1$, para el caso en el que F sea $N_{p+q}(\mathbf{0}, \Sigma)$ con

$$\Sigma = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix},$$

tal que $\Sigma_{\mathbf{xx}} = I_2, \Sigma_{\mathbf{yy}} = I_2$ y $\Sigma_{\mathbf{xy}} = \begin{pmatrix} 0,9 & 0 \\ 0 & 0,5 \end{pmatrix}$.

Estimador propuesto - SM

Función de Influencia

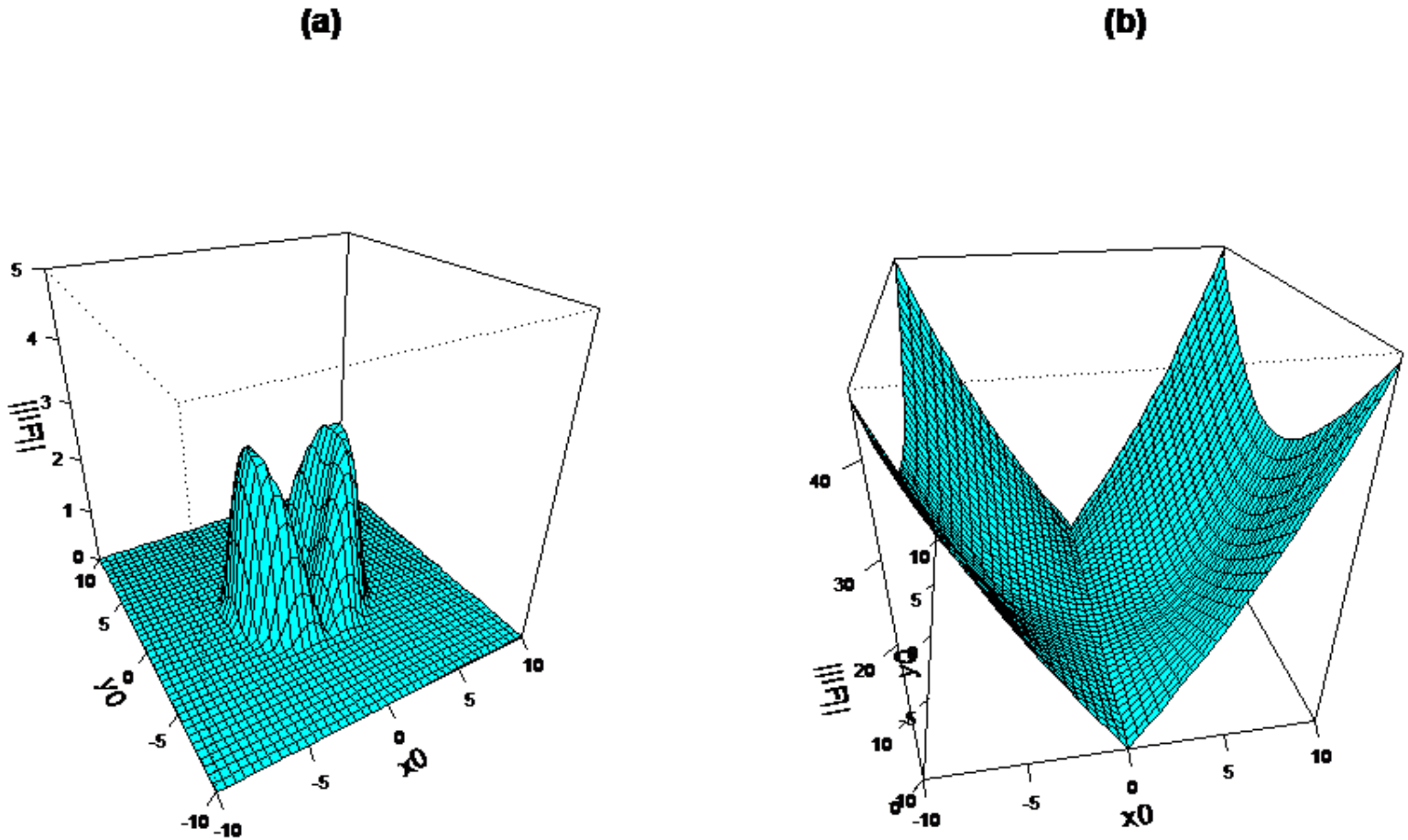


Figure: Norma de la Función de Influencia para $(x_0, y_0, 0, 0)$ del primer vector canónico de (a) Estimador SM y (b) Estimador Clásico.

Estimador propuesto - SM

Función de Influencia

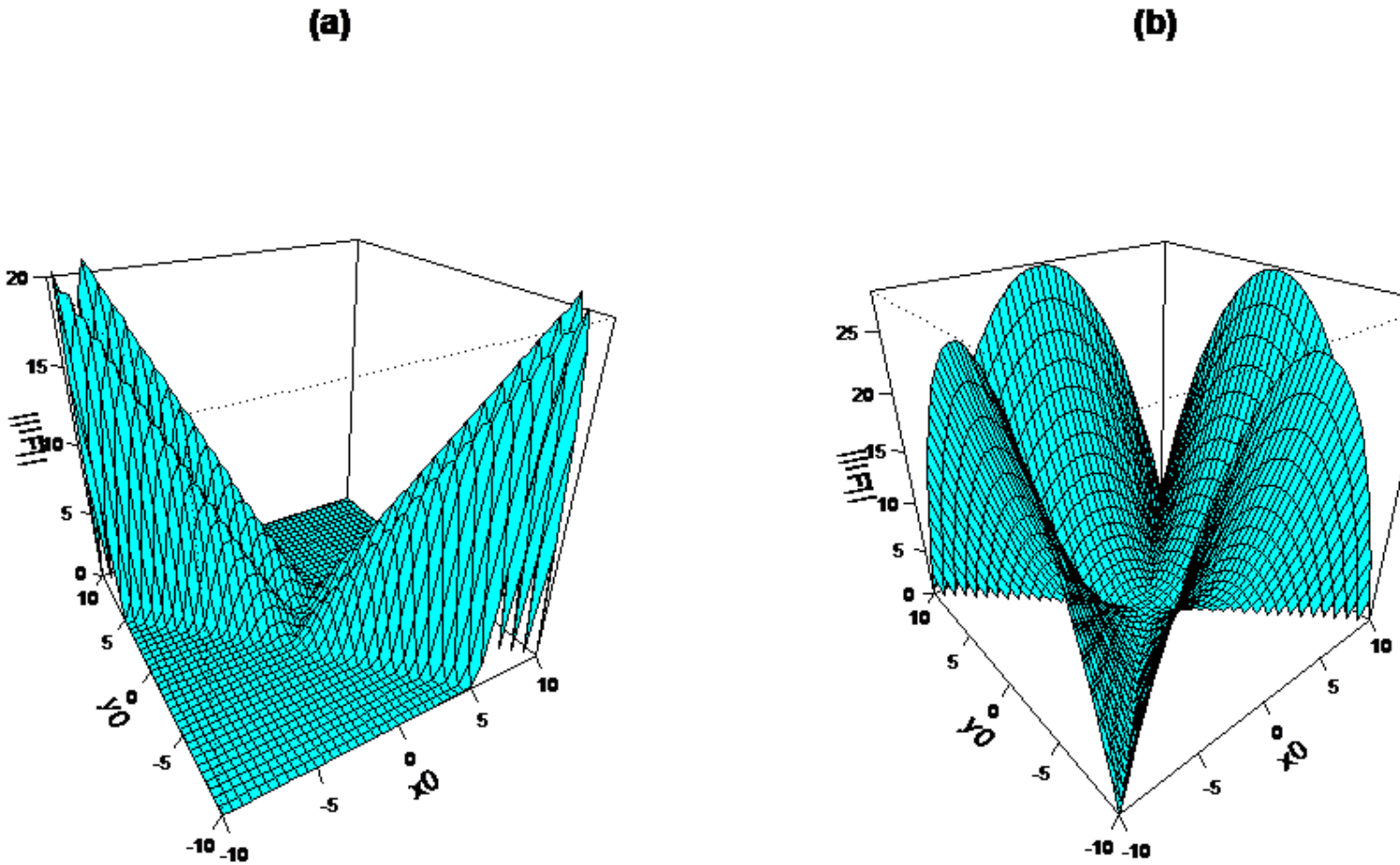


Figure: Norma de la Función de Influencia para $(x_0, 0, y_0, 0)$ del primer vector canónico de (a) Estimador SM y (b) Estimador Clásico.

Estimador propuesto - SMI

- Definimos Modelo de contaminaciones independientes cuando la contaminación puede afectar a las coordenadas de un vector en forma independiente (ver Alqallaf et al (2009) y Maronna and Yohai (2008)).

Estimador propuesto - SMI

- Definimos Modelo de contaminaciones independientes cuando la contaminación puede afectar a las coordenadas de un vector en forma independiente (ver Alqallaf et al (2009) y Maronna and Yohai (2008)).
- Para el caso de PCA, dada la muestra $\mathbf{z}_1, ..., \mathbf{z}_n$ en $\mathbb{R}^m$ (centrada), para encontrar vectores y componentes principales se debe encontrar $\mathbf{v}_1, ..., \mathbf{v}_q$ que forman una base del subespacio V de dimensión $q < m$, de forma de minimizar la siguiente expresión,

$$\begin{aligned}\sum_{i=1}^{n} \|\mathbf{z}_i - \widehat{\mathbf{z}}_i\|^2 &= \sum_{i=1}^{n} \left\| \mathbf{z}_i - \sum_{j=1}^{q} \left(\mathbf{z}_i' \mathbf{v}_j\right) \mathbf{v}_j \right) \right\|^2 \\ &= \sum_{i=1}^{n} \|\mathbf{z}_i - B\mathbf{a}_i\|^2 .\end{aligned}$$

Estimador propuesto - SMI

Para el caso de CCA, sean $\mathbf{z}_i = (\mathbf{x}_i', \mathbf{y}_i')', i = 1, ..., n$, con covarianzas muestrales $\widehat{\Sigma}_{\mathbf{xx}}$ y $\widehat{\Sigma}_{\mathbf{yy}}$ respectivamente.

- Podemos escribir $\widehat{\Sigma}_{\mathbf{xx}}^{-1/2}\mathbf{x}_i = \sum_{k=1}^{p} \frac{(\mathbf{x}_i'\mathbf{t}_k)}{\sqrt{\lambda_k}}\mathbf{t}_k$, $\lambda_1 > ... > \lambda_p$ y $\widehat{\Sigma}_{\mathbf{yy}}^{-1/2}\mathbf{y}_i = \sum_{k=1}^{q} \frac{(\mathbf{y}_i'\mathbf{v}_k)}{\sqrt{\gamma_k}}\mathbf{v}_k$, $\gamma_1 > ... > \gamma_q$.

Estimador propuesto - SMI

Para el caso de CCA, sean $\mathbf{z}_i = (\mathbf{x}_i', \mathbf{y}_i')'$, $i = 1, ..., n$, con covarianzas muestrales $\widehat{\Sigma}_{\mathbf{xx}}$ y $\widehat{\Sigma}_{\mathbf{yy}}$ respectivamente.

- Podemos escribir $\widehat{\Sigma}_{\mathbf{xx}}^{-1/2}\mathbf{x}_i = \sum_{k=1}^{p} \frac{(\mathbf{x}_i'\mathbf{t}_k)}{\sqrt{\lambda_k}}\mathbf{t}_k$, $\lambda_1 > ... > \lambda_p$ y $\widehat{\Sigma}_{\mathbf{yy}}^{-1/2}\mathbf{y}_i = \sum_{k=1}^{q} \frac{(\mathbf{y}_i'\mathbf{v}_k)}{\sqrt{\gamma_k}}\mathbf{v}_k$, $\gamma_1 > ... > \gamma_q$.
- Entonces cuando miramos la ecuación que queremos minimizar resulta

$$\sum_{i=1}^{n} \left\| C\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x}_i - B\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y}_i - \mathbf{a} \right\|^2$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{r} \left(C_j \sum_{k=1}^{p} \frac{(\mathbf{x}_i'\mathbf{t}_k)}{\sqrt{\lambda_k}}\mathbf{t}_k - B_j \sum_{k=1}^{q} \frac{(\mathbf{y}_i'\mathbf{v}_k)}{\sqrt{\gamma_k}}\mathbf{v}_k - a_j \right)^2$$

$$\approx \sum_{i=1}^{n}\sum_{j=1}^{r} \left((\mathbf{C}_j, \mathbf{B}_j) \begin{pmatrix} T & 0 \\ 0 & V \end{pmatrix} (\mathbf{a}_i', \mathbf{b}_i')' - a_j \right)^2 .$$

Estimador propuesto - SMI

- Para reducir la influencia de observaciones atípicas, proponemos usar un M-estimador de escala

$$(C^*, B^*, \boldsymbol{\mu}^*) = \arg \min_{\boldsymbol{\mu}, C, B} \sum_{j=1}^{r} \widehat{\sigma}_j^2 (\boldsymbol{\mu}, C, B),$$

Estimador propuesto - SMI

- Para reducir la influencia de observaciones atípicas, proponemos usar un M-estimador de escala

$$(C^*, B^*, \boldsymbol{\mu}^*) = \arg \min_{\boldsymbol{\mu}, C, B} \sum_{j=1}^{r} \widehat{\sigma}_j^2 (\boldsymbol{\mu}, C, B),$$

- donde $\widehat{\sigma}_j = \widehat{\sigma}_j (\boldsymbol{\mu}, C, B)$ satisface

$$g\left(\mu_j, \mathbf{C}_j, \mathbf{B}_j\right) = \frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{(\mathbf{C}_j, \mathbf{B}_j) \begin{pmatrix} T & 0 \\ 0 & V \end{pmatrix} (\mathbf{a}_i', \mathbf{b}_i')' - \mu_j}{\widehat{\sigma}_j} \right) = \delta,$$

con $\boldsymbol{\mu} \in \mathbb{R}^r$, $C \in \mathbb{R}^{r \times p}$, $B \in \mathbb{R}^{r \times q}$, para $r \leq \min(p, q)$.

Estudio de simulación

Medidas para evaluar el rendimiento de los estadísticos

Para realizar el estudio de simulación, se consideraron muestras aleatorias $\mathbf{z}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix} \backsim N_{p+q}(\mathbf{0}, \Sigma)$, donde Σ que tiene todos los elementos diagonales iguales a 1 y los elementos fuera de la diagonal son todos iguales a $h \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ (estructura de covarianzas utilizadas por Danilov et al (2012)).
Para evaluar el rendimiento de los estimadores, se consideró la siguiente medida propuesta por Branco et al (2005):
Error cuadrático medio para vectores canónicos.

$$MSE(\widehat{\theta}_k) = \frac{1}{m} \sum_{j=1}^{m} \cos^{-1} \left(\frac{\left| \theta_{k,j}^t \widehat{\theta}_{k,j} \right|}{\|\theta_{k,j}\| \left\| \widehat{\theta}_{k,j} \right\|} \right),$$

siendo m la cantidad de replicaciones, $\widehat{\theta}_k$ el estimador del $k-$ésimo vector canónico, $\theta_{k,j}$ el $k-$ésimo vector canónico

Estudio de simulación

Se consideraron dos estructuras de contaminación:

1. Contaminación por celdas (independientes): celdas de la matriz de datos son elegidas al azar con probabilidad ε y su contenido reemplazado por un elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.

Estudio de simulación

Se consideraron dos estructuras de contaminación:

1. Contaminación por celdas (independientes): celdas de la matriz de datos son elegidas al azar con probabilidad ε y su contenido reemplazado por un elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.
2. Contaminación por individuos (casos): El contenido de las primeras $n\varepsilon$ filas de la matriz de datos son reemplazadas con elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.

Estudio de simulación

Se consideraron dos estructuras de contaminación:

1. Contaminación por celdas (independientes): celdas de la matriz de datos son elegidas al azar con probabilidad ε y su contenido reemplazado por un elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.
2. Contaminación por individuos (casos): El contenido de las primeras $n\varepsilon$ filas de la matriz de datos son reemplazadas con elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.

- Tamaño de la muestra: 100.

Estudio de simulación

Se consideraron dos estructuras de contaminación:

1. Contaminación por celdas (independientes): celdas de la matriz de datos son elegidas al azar con probabilidad ε y su contenido reemplazado por un elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.
2. Contaminación por individuos (casos): El contenido de las primeras $n\varepsilon$ filas de la matriz de datos son reemplazadas con elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.

- Tamaño de la muestra: 100.
- Número de replicaciones: 150.

Estudio de simulación

Se consideraron dos estructuras de contaminación:

1. Contaminación por celdas (independientes): celdas de la matriz de datos son elegidas al azar con probabilidad ε y su contenido reemplazado por un elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.
2. Contaminación por individuos (casos): El contenido de las primeras $n\varepsilon$ filas de la matriz de datos son reemplazadas con elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.

- Tamaño de la muestra: 100.
- Número de replicaciones: 150.
- $p = q = 5$.

Estudio de simulación

Se consideraron dos estructuras de contaminación:

1. Contaminación por celdas (independientes): celdas de la matriz de datos son elegidas al azar con probabilidad ε y su contenido reemplazado por un elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.
2. Contaminación por individuos (casos): El contenido de las primeras $n\varepsilon$ filas de la matriz de datos son reemplazadas con elemento proveniente de $N_{p+q}\left(k\mathbf{v}_0, 0.5\Sigma\right), k \in \{1, ..., 12\}$ y $\mathbf{v}_0$ pertenece al subespacio de los autovectores de Σ asociados al menor autovalor.

- Tamaño de la muestra: 100.
- Número de replicaciones: 150.
- $p = q = 5$.
- $\varepsilon = 0.2$.

Estudio de simulación

Los estimadores que se consideran en el estudio de simulación son:

- **Estimador clásico.**

Estudio de simulación

Los estimadores que se consideran en el estudio de simulación son:

- **Estimador clásico.**
- **Estimador robusto de 2 pasos para la matriz de covarianzas (Agostinelli et al (2015)).**

Estudio de simulación

Los estimadores que se consideran en el estudio de simulación son:

- **Estimador clásico.**
- **Estimador robusto de 2 pasos para la matriz de covarianzas (Agostinelli et al (2015)).**
- **Estimador SM (Adrover and Donato (2015)).**Para computar la escala se usó la función (Maronna (2005) que hace convergente al algoritmo iterativo)

$$\rho(t) = \min\left\{1, 1-(1-t)^3\right\},$$

con $\delta = 0,5$.

Estudio de simulación

Los estimadores que se consideran en el estudio de simulación son:

- **Estimador clásico.**
- **Estimador robusto de 2 pasos para la matriz de covarianzas (Agostinelli et al (2015)).**
- **Estimador SM (Adrover and Donato (2015)).**Para computar la escala se usó la función (Maronna (2005) que hace convergente al algoritmo iterativo)

$$\rho(t) = \min\left\{1, 1-(1-t)^3\right\},$$

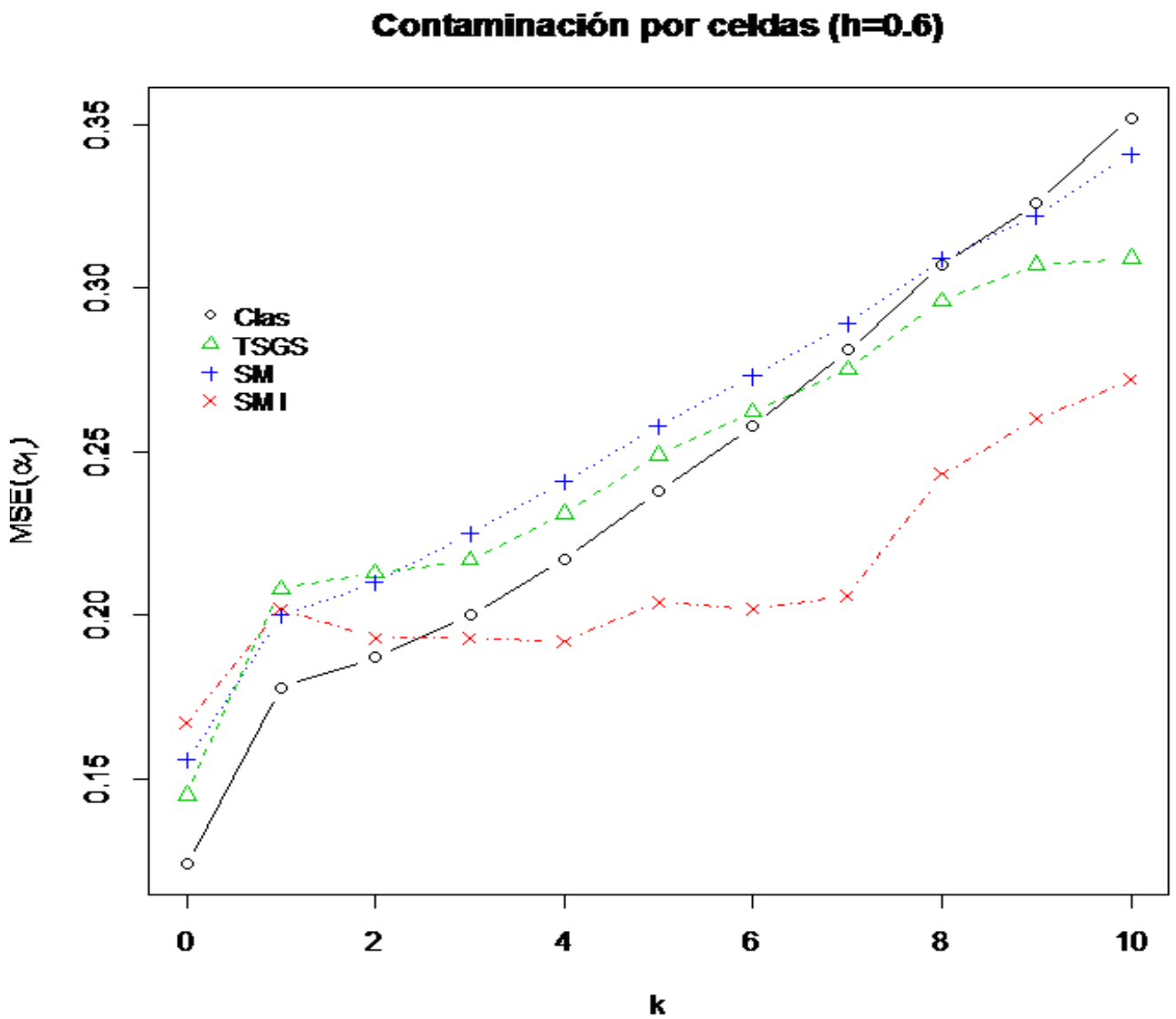
con $\delta = 0,5$.

- **Estimador SMI.** Para computar el algoritmo se utilizó la función bicuadrada de Tukey

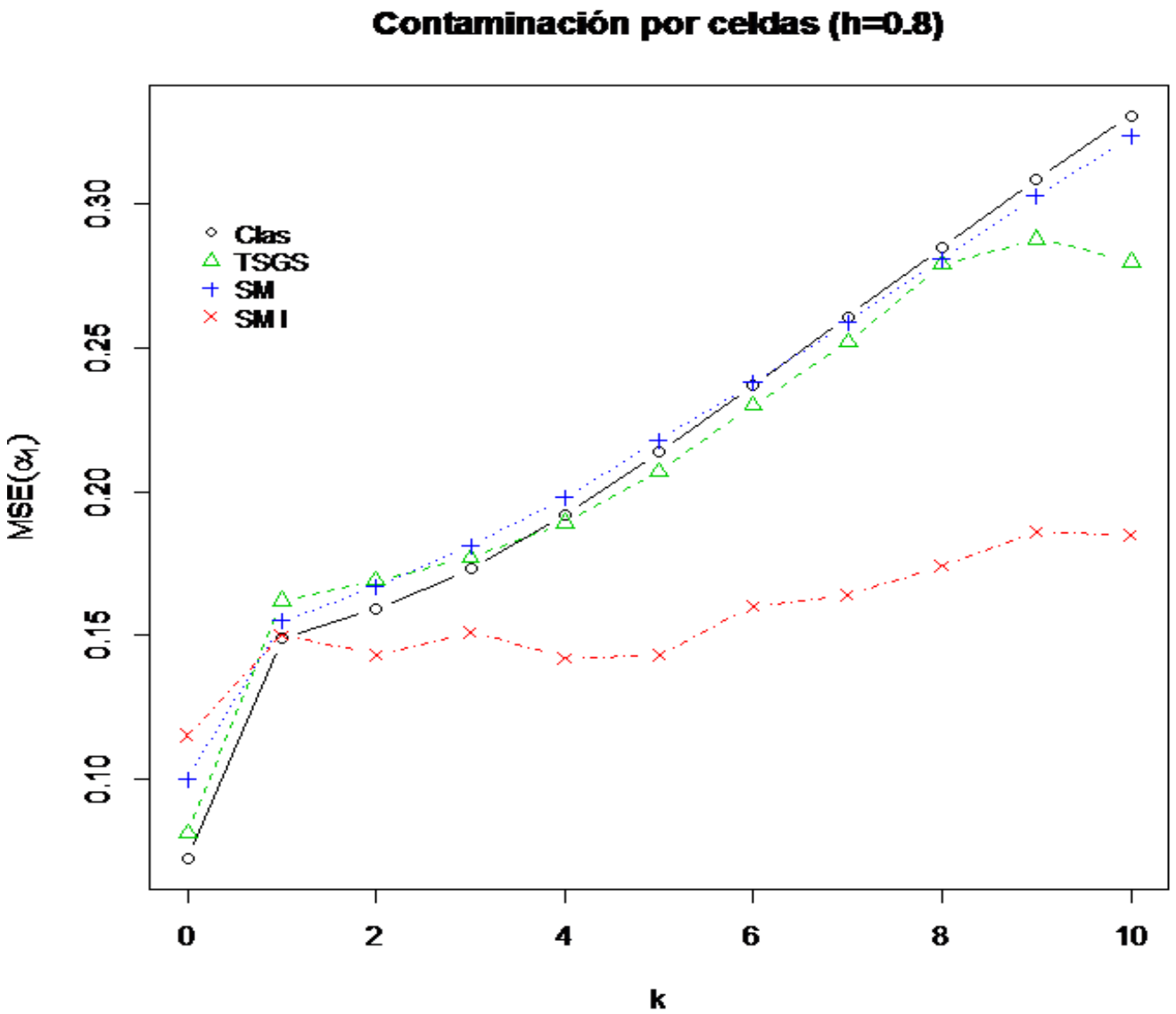
$$\rho(t) = \min\left\{1, 1-\left(1-\left(\frac{t}{c}\right)^2\right)^3\right\},$$

con $c = 1.54764$ y $\delta = 0,5$. (Boente and Salibian Barrera (2015)).

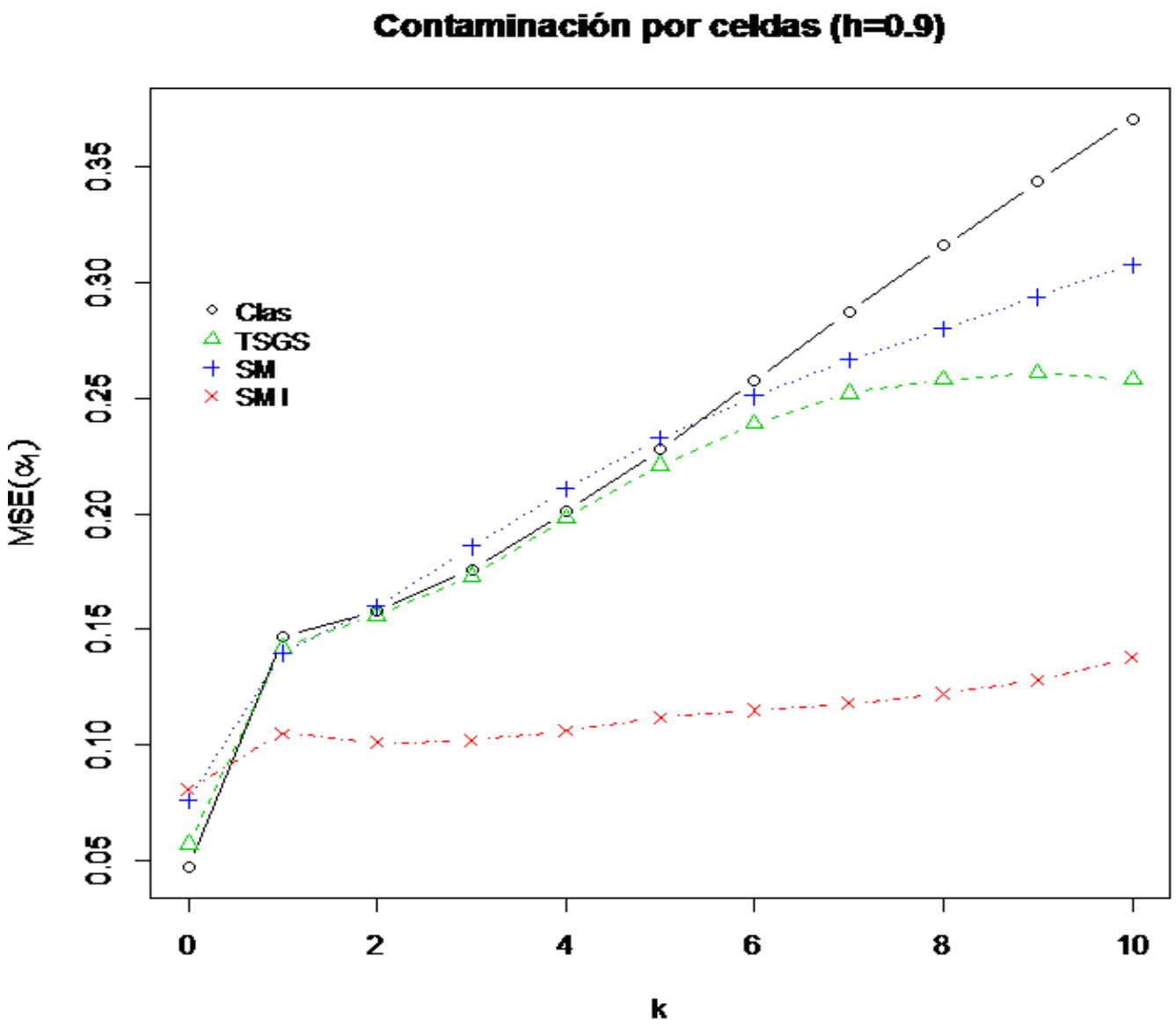
Estudio de simulación



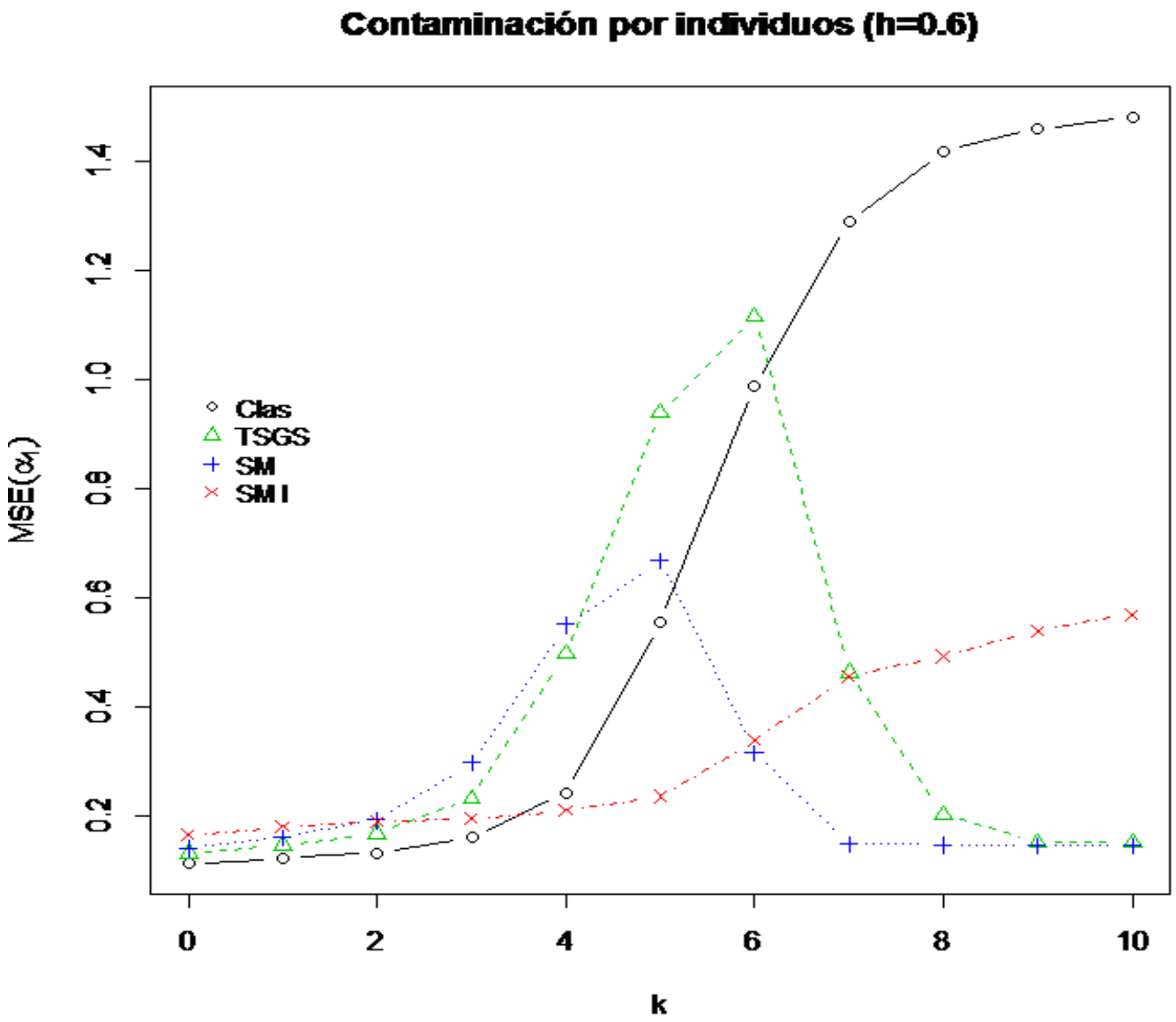
Estudio de simulación



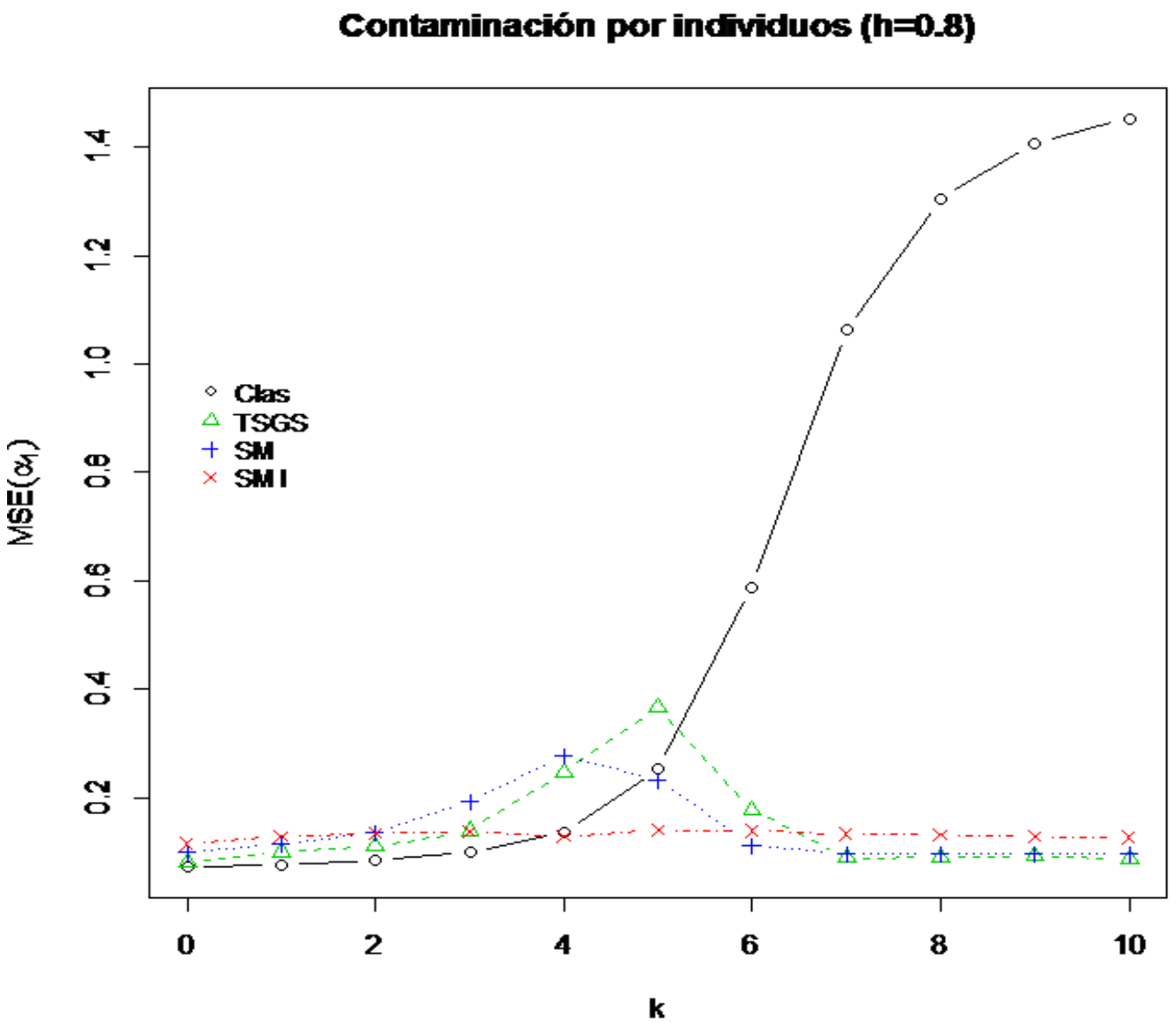
Estudio de simulación



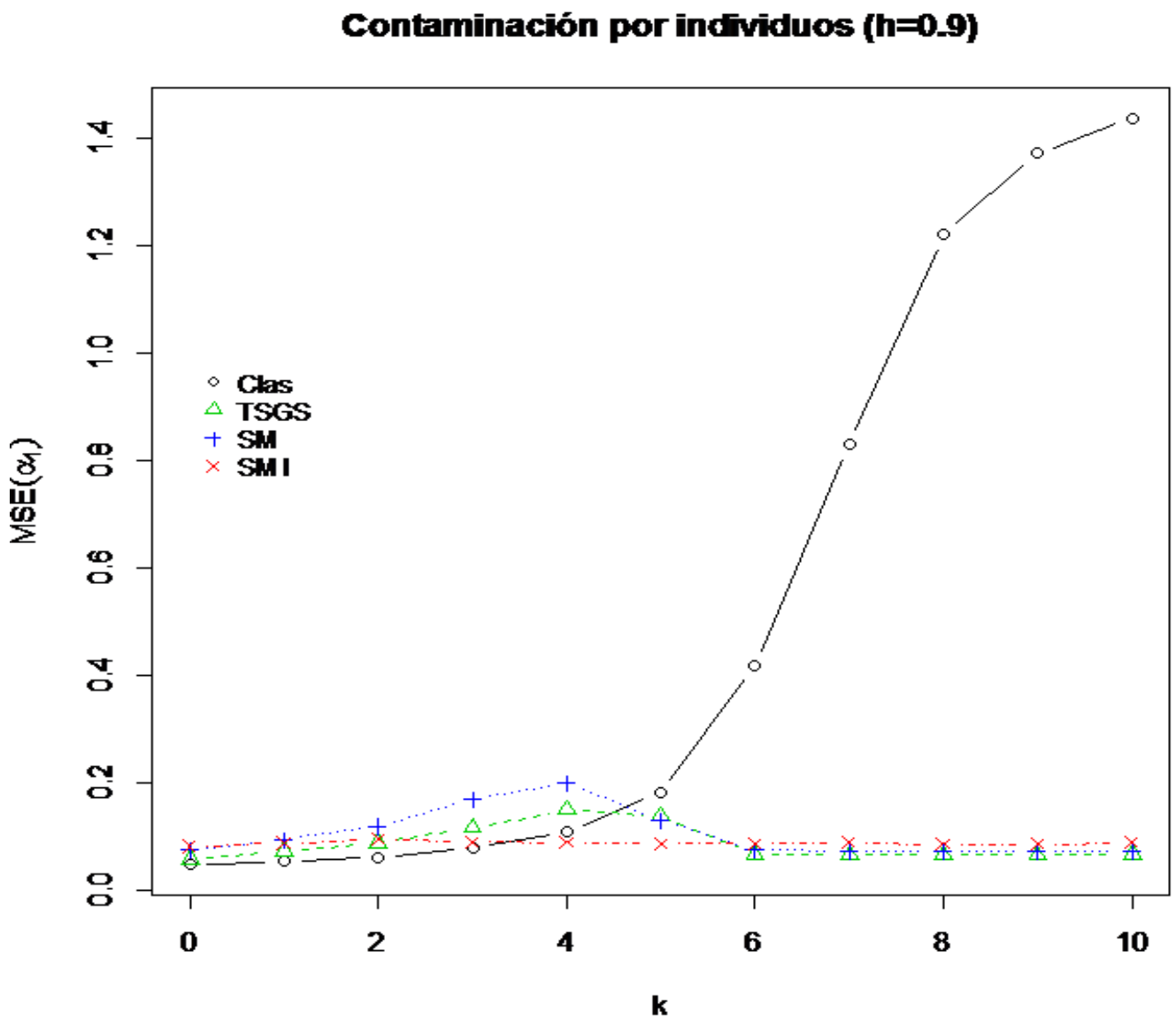
Estudio de simulación



Estudio de simulación



Estudio de simulación



Conclusiones

1. Se presentaron dos estimadores robustos para vectores canónicos, uno para el modelo de contaminaciones por celdas y otro para contaminaciones por individuo

Conclusiones

1. Se presentaron dos estimadores robustos para vectores canónicos, uno para el modelo de contaminaciones por celdas y otro para contaminaciones por individuo
2. Se exploró la relación CCA - PCA que permitió desarrollar los algoritmos iterativos.

Conclusiones

1. Se presentaron dos estimadores robustos para vectores canónicos, uno para el modelo de contaminaciones por celdas y otro para contaminaciones por individuo
2. Se exploró la relación CCA - PCA que permitió desarrollar los algoritmos iterativos.
3. Se comparó el rendimiento de los estimadores propuestos a través de un estudio de simulación.

Conclusiones

1. Se presentaron dos estimadores robustos para vectores canónicos, uno para el modelo de contaminaciones por celdas y otro para contaminaciones por individuo
2. Se exploró la relación CCA - PCA que permitió desarrollar los algoritmos iterativos.
3. Se comparó el rendimiento de los estimadores propuestos a través de un estudio de simulación.
4. El estimador SMI muestra el mejor comportamiento global para los dos modelos de contaminación analizados.

Conclusiones

1. Se presentaron dos estimadores robustos para vectores canónicos, uno para el modelo de contaminaciones por celdas y otro para contaminaciones por individuo
2. Se exploró la relación CCA - PCA que permitió desarrollar los algoritmos iterativos.
3. Se comparó el rendimiento de los estimadores propuestos a través de un estudio de simulación.
4. El estimador SMI muestra el mejor comportamiento global para los dos modelos de contaminación analizados.
5. Los estimadores SM y TSGS tienen mejor rendimiento en el modelo de contaminación por individuo.

Conclusiones

1. Se presentaron dos estimadores robustos para vectores canónicos, uno para el modelo de contaminaciones por celdas y otro para contaminaciones por individuo
2. Se exploró la relación CCA - PCA que permitió desarrollar los algoritmos iterativos.
3. Se comparó el rendimiento de los estimadores propuestos a través de un estudio de simulación.
4. El estimador SMI muestra el mejor comportamiento global para los dos modelos de contaminación analizados.
5. Los estimadores SM y TSGS tienen mejor rendimiento en el modelo de contaminación por individuo.
6. Muchas gracias!!!

Referencias bibliográficas

- Adrover, J. G., Donato, S. M. (2015) "A robust predictive approach for canonical correlation analysis". Journal of Multivariate Analysis. 133: 356-376.
- Agostinelli, C., Leung, A., Yohai, V. and Zamar, R. (2015) "Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination" Test. 24(3): 441-461.
- Alqallaf, F., Van Aelst, S., Yohai, V. J. and Zamar, R. H. (2009). "Propagation of outliers in multivariate data". The Annals of Statistics 37(1): 311-331.
- Boente, G. and Salibian Barrera, M. (2015). "S-estimators for functional principal component analysis". Journal of the American Statistical Association 110(511): 1100-1111.
- Brillinger, David. (1975). *Time Series: Data Analysis and Theory*. Estados Unidos de América. Holt, Rinehart and Winston.

Referencias bibliográficas

- Branco, J.A., Croux, C., Filzmoser, P., and Oliveira, M.R. (2005) "Robust Canonical Correlations: A Comparative Study". *Computational Statistics*, 20(2): 203-229.
- Danilov, M., Yohai, V. and Zamar, R. (2012) "Robust Estimation of Multivariate Location and Scatter in the presence of Missing Data". *Journal of the American Statistical Association*, 107(499): 1178-1186.
- Maronna, R.A. (2005) "Principal Components and Orthogonal Regression Based on Robust Scales". *Technometrics*, 47(3): 264-273.
- Maronna, R. A. and Yohai, V. J. (2008) "Robust Low-Rank Approximation of Data Matrices With Elementwise Contamination" *Technometrics*. 50(3): 295-304.
- Seber, G. A. F. (1984). *Multivariate Observations*. Estados Unidos de América. John Wiley & Sons, Inc.