Optimized symbolic principal components for interval-valued variables
SDA-2017
Ljubljana, Slovenia

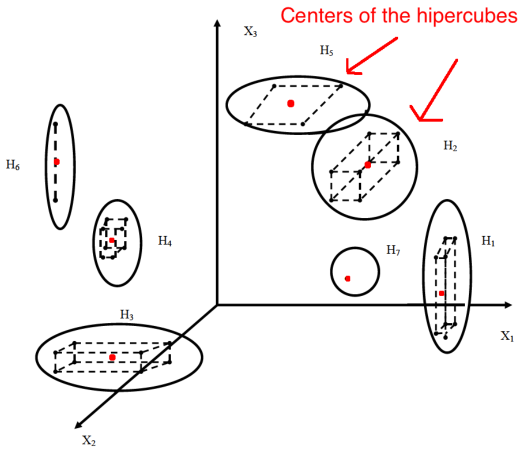
Jorge Arce [2]   Oldemar Rodríguez [1]

[1]University of Costa Rica, San José, Costa Rica;

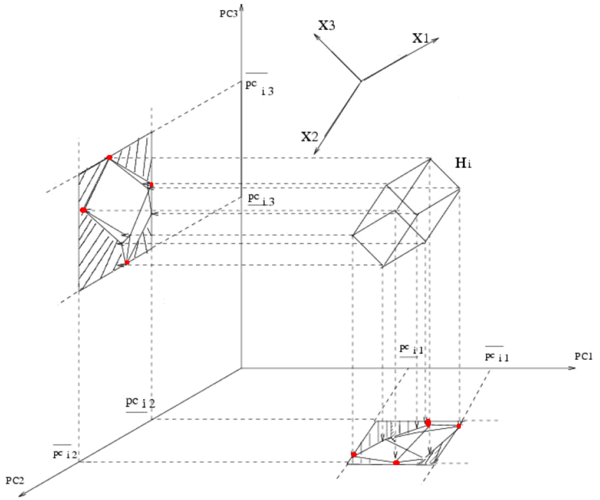
[2]National Bank of Costa Rica, San José, Costa Rica;

June 12, 2017

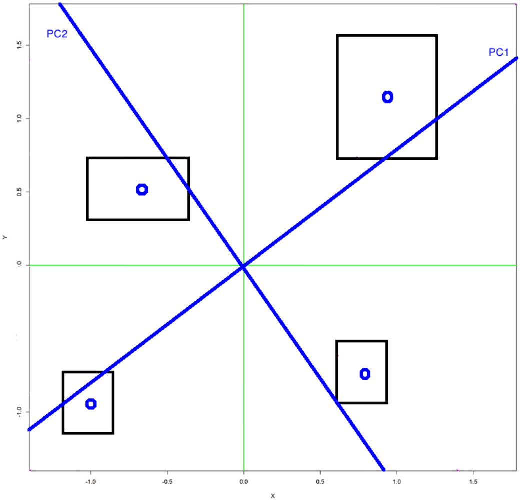
The centers principal components analysis is conducted by doing a classical analysis on the centered point observations X^c.



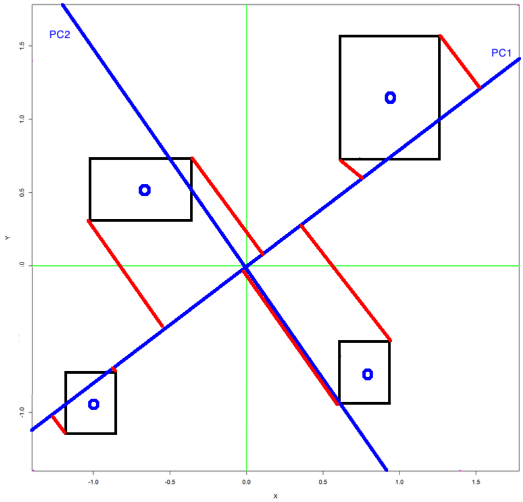
Then all the vertices of the hypercube are projected on principal components of the centers



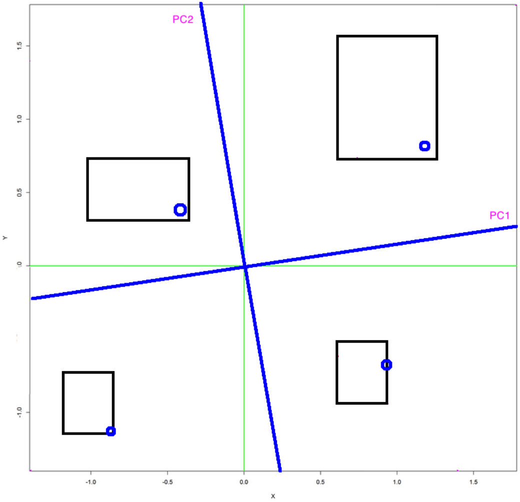
With the centers of the hypercubes principal components are computed



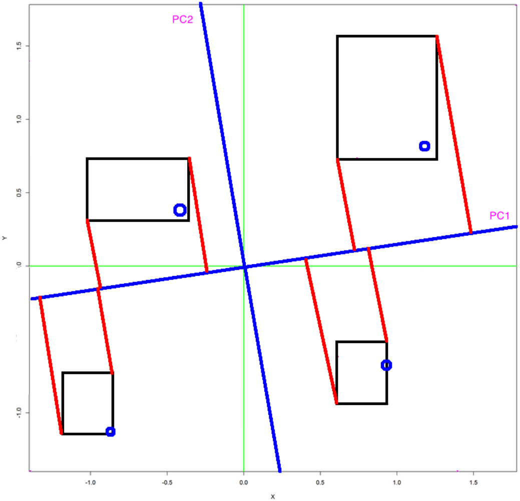
Vertices of the hypercubes are projected on principal components of the Centers



If we use differents points inside the hypercube the principal components will change



Vertices of the hypercubes are projected on principal components of the New Points



The question is:

- What are the best points inside the hypercube?
  - In order to maximize inertia or
  - In order to minimize squared distance between vertices and its projections.

From: Symbolic principal components for interval-valued observations Lynn Billard, Ahlame Douzal-Chouakria and E. Diday we know:

Then, the νth centers principal component can be written as

$$PC\nu^c = \sum_{j=1}^{p} (x_j^0 - \bar{X}_j^c) u_{\nu j}. \qquad (4.5)$$

In particular, let $\tilde{\boldsymbol{x}}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})$ be any point contained in the hypercube H_i described by $\boldsymbol{\xi}_i$. Thus, we can calculate the νth centers principal component for this $\tilde{\boldsymbol{x}}_i$ from (4.5) as

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

From: Symbolic principal components for interval-valued observations Lynn Billard, Ahlame Douzal-Chouakria and E. Diday we know:

Therefore, we can define the νth centers principal component as

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

where

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

and

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

From: Symbolic principal components for interval-valued observations Lynn Billard, Ahlame Douzal-Chouakria and E. Diday we know:

It can be shown that these reduce to

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

and

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

where $J_c^- = \{j | u_{\nu j} < 0\}$ and $J_c^+ = \{j | u_{\nu j} > 0\}$.

Optimized symbolic principal components for interval-valued variables

Definition ($Z \in X$)

Let be X an intervals symbolic matrix:

$$X = \begin{bmatrix} [a_{11}, b_{11}] & [a_{12}, b_{12}] & [a_{13}, b_{13}] & \dots & [a_{1p}, b_{1p}] \\ [a_{21}, b_{21}] & [a_{22}, b_{22}] & [a_{23}, b_{23}] & \dots & [a_{2p}, b_{2p}] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ [a_{n1}, b_{n1}] & [a_{n2}, b_{n2}] & [a_{n3}, b_{n3}] & \dots & [a_{np}, b_{np}] \end{bmatrix},$$

where $a_{ij} \leq b_{ij}$ *and let be* $Z = (z_{ij})$ *with* $z_{ij} \in \mathbb{R}$ *. We say that* $Z \in X$ *if* $z_{ij} \in \left[a_{ij}, b_{ij}\right]$ *for all* $i = 1, 2, \dots, n$ *and* $j = 1, 2, \dots, p$*.*

Optimized symbolic principal components for interval-valued variables

Let be X an interval data matrix size $n \times p$ and let be $Z \in X$, we do a classical PCA on Z then vth principal components of Z for the observation ξ_u with $v = 1, \ldots, n$, $u = 1, \ldots, p$,

$$y^Z_{uv} = \sum_{j=1}^{p} (z_{ju} - \bar{Z}_{(j)}) w^Z_{v_j}, \tag{1}$$

where $\bar{Z}_{(j)}$ is the mean of the variable $Z_{(j)}$ and $w^Z_v = (w^Z_{v_1}, \ldots, w^Z_{v_p})$ is the vth eigenvector of the variance-covariance matrix of Z.

Then it is clear that $\beta(Z) = \{w^Z_1, \ldots, w^Z_p\}$ is an orthonormal basis of $\mathbb{R}^p$.

We define the centered and standarized matrix of vertices with respect to Z:

$$\tilde{X}^v(Z) = \begin{bmatrix} \begin{bmatrix} \frac{a_{11}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{a_{12}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{a_{1p}-\bar{Z}_{(p)}}{\sigma_{(p)}} \\ \cdots & \cdots & \ddots & \cdots \\ \frac{b_{11}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{b_{12}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{b_{1p}-\bar{Z}_{(p)}}{\sigma_{(p)}} \end{bmatrix} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ \begin{bmatrix} \frac{a_{i1}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{a_{i2}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{a_{ip}-\bar{Z}_{(p)}}{\sigma_{(p)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{b_{i1}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{b_{i2}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{b_{ip}-\bar{Z}_{(p)}}{\sigma_{(p)}} \end{bmatrix} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ \begin{bmatrix} \frac{a_{n1}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{a_{n2}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{a_{np}-\bar{Z}_{(p)}}{\sigma_{(p)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{b_{n1}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{b_{n2}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{b_{np}-\bar{Z}_{(p)}}{\sigma_{(p)}} \end{bmatrix} \end{bmatrix} \tag{2}$$

Optimized interval PCA

Let be ξ_i the ith observation of X with $i = 1, \ldots, n$, now the vertices of the hypercube will be projected as a supplementary elements in PCA of Z.

Definition

Let be:

$$\tilde{X}_i^v(Z) = \begin{bmatrix} \frac{a_{i1}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{a_{i2}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{a_{ip}-\bar{Z}_{(p)}}{\sigma_{(p)}} \\ \frac{a_{i1}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{a_{i2}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{a_{ip}-\bar{Z}_{(p)}}{\sigma_{(p)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{b_{i1}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{b_{i2}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{b_{ip}-\bar{Z}_{(p)}}{\sigma_{(p)}} \\ \frac{b_{i1}-\bar{Z}_{(1)}}{\sigma_{(1)}} & \frac{b_{i2}-\bar{Z}_{(2)}}{\sigma_{(2)}} & \cdots & \frac{b_{ip}-\bar{Z}_{(p)}}{\sigma_{(p)}} \end{bmatrix}. \tag{3}$$

where $\sigma_{(j)}$ is the standard deviation of $Z_{(j)}$. To simplify, let's denote each row of the matrix $\tilde{X}_i^v(Z)$ as follows $\tilde{x}_{i_{kj}}^v(Z)$, with $k = 1, \ldots, 2^{m_i}$, m_i the number of non-trivial intervals and $j = 1, \ldots, p$

Optimized interval PCA

Therefore, we can define the vth principal components as:

$$C^s(x^v_{i_j}) = \sum_{t=1}^{p} \tilde{x}^v_{i_{jt}}(Z) w_{s_t} \tag{4}$$

to $j = 1, \ldots, 2^{m_i}$, m_i in the number of non-trivial intervals.
where

$$\tilde{Y}^v_{iu} = \tilde{y}_{iu} = [\tilde{y}^{a_Z}_{iu}, \tilde{y}^{b_Z}_{iu}] \text{ con } u = 1, \ldots, p \tag{5}$$

where

$$\tilde{y}^{a_Z}_{iu} = \min_{j=1,\ldots,2^{m_i}} C^u(x^v_{i_j}) \tag{6}$$

$$\tilde{y}^{b_Z}_{iu} = \max_{j=1,\ldots,2^{m_i}} C^u(x^v_{i_j}) \tag{7}$$

Optimized interval PCA

It can be shown that these reduce to:

Theorem

$\tilde{Y}^{v_Z}_{iu}$ can be computed as:

$$\tilde{y}^{a_Z}_{ik} = \sum_{j \in J^-_Z} (b_{ij} - \overline{Z}_{(j)}) w^Z_{k_j} + \sum_{j \in J^+_Z} (a_{ij} - \overline{Z}_{(j)}) w^Z_{k_j}$$

$$\tilde{y}^{b_Z}_{ik} = \sum_{j \in J^-_Z} (a_{ij} - \overline{Z}_{(j)}) w^Z_{k_j} + \sum_{j \in J^+_Z} (b_{ij} - \overline{Z}_{(j)}) w^Z_{k_j}$$

where $J^-_Z = \{j | w^v_{kj} < 0\}$ y $J^+_Z = \{j | w^v_{kj} \geq 0\}, \overline{Z}_{(j)}$ is the mean j^{th} column.

Optimized interval PCA

So far we have found a way to do a PCA for each $Z \in X$. The idea is to look for the matrix $Z^\star$ that is optimal in some sense, for example:

- Minimize the squared distance from the vertices of the hypercube to the principal components of Z.
- Maximize the variance in the firsts components of Z

Minimize the squared distance from the vertices of the hypercube to the principal components of Z

Let X be an interval matrix size $n \times p$, $Z \in X$, and

$$\beta(Z) = \{w_1^Z, \ldots, w_s^Z\},$$

with $s \leq p$ where w_i^Z are the eigenvectors of Z variance-covariance matrix. Let be X^v the vertices matrix of X and

$$N = \sum_{i=1}^{n} 2^{m_i},$$

with m_i the number of non-trivial intervals for the observation ξ_i.

Optimized interval PCA

We want to minimized the function $\varphi(Z) : X \to \mathbb{R}^+$ defined as follows:

$$\varphi(Z) = \sum_{i=1}^{N} \| \tilde{X}_i^v(Z) - Pr_{\beta(Z)}(\tilde{X}_i^v(Z)) \|^2. \tag{8}$$

Since $Z \in X$ and X is the finite union of compact sets and the $\varphi(Z)$ is a continuous function then it has a maximum and a minimum.

Algorithm to compute $\varphi(Z)$

Algorithm 1 Calculation of φ

Require: X interval matrix $n \times p$, $Z \in X$, and
s number of principal components.

Ensure: $\varphi(Z)$

1: Apply a PCA on Z and compute:
2: $\beta = \{w_1, \ldots, w_s\}$, $s \leq p$ where w_i are the eigenvectors of Z.
3: Compute the vertices matrix X^v of X
4: Compute $\tilde{X}^v(Z)$
5: $\varphi(Z) = \sum_{i=1}^{N} \|\tilde{X}_i^v(Z) - Pr_{\beta(Z)}(\tilde{X}_i^v(Z))\|^2$.
6: return $\varphi(Z)$

Optimization Problem

$$
\begin{aligned}
&\text{Minimize} \quad \varphi(Z) = \sum_{i=1}^{N} \| \tilde{X}_i^v(Z) - Pr_{\beta(Z)}(\tilde{X}_i^v(Z)) \|^2 \\
&\text{Subject to} \quad \left\{
\begin{array}{c}
a_{11} \leq z_{11} \leq b_{11} \\
\vdots \\
a_{1j} \leq z_{1j} \leq b_{1j} \\
\vdots \\
a_{1p} \leq z_{1p} \leq b_{1p} \\
\vdots \\
a_{ij} \leq z_{ij} \leq b_{ij} \\
\vdots \\
a_{n1} \leq z_{n1} \leq b_{n1} \\
\vdots \\
a_{np} \leq z_{np} \leq b_{np}
\end{array}
\right.
\end{aligned}
\tag{9}
$$

Definition ($Z^\star$)

The matrix $Z \in X$ that solve the problem (9) is called the optimal matrix with respect to X^v and we will denote it by $Z^\star$.

Algorithm 2 Minimizing the squared distance PCA

Require: X $n \times p$ matrix, $Z \in X$, s number of principal components, `TOL` is a tolerance of variations and `N` the maximum number of iterations.

Ensure: $\tilde{Y}^{V_{Z^\star}}$

1: Let be $Z = X^c$ the centers matrix as an initial value.

2: Get $Z^\star$ using Broyden–Fletcher–Goldfarb–Shanno algorithm `BFGS(Z, function =` $\varphi(Z)$`,TOL,N)`

3: Get $\tilde{Y}^{V_{Z^\star}}$ applying the Theorem 1

4: **return** $\tilde{Y}^{V_{Z^\star}}$

Definition ($Z^\star$)

The matrix $Z \in X$ that solve the problem (9) is called the optimal matrix with respect to X^v and we will denote it by $Z^\star$.

Algorithm 3 Minimizing the squared distance PCA

Require: X $n \times p$ matrix, $Z \in X$, s number of principal components, `TOL` is a tolerance of variations and `N` the maximum number of iterations.

Ensure: $\tilde{Y}^{V_{Z^\star}}$

1: Let be $Z = X^c$ the centers matrix as an initial value.

2: Get $Z^\star$ using Broyden–Fletcher–Goldfarb–Shanno algorithm `BFGS(Z, function =` $\varphi(Z)$`,TOL,N)`

3: Get $\tilde{Y}^{V_{Z^\star}}$ applying the Theorem 1

4: **return** $\tilde{Y}^{V_{Z^\star}}$

Broyden–Fletcher–Goldfarb–Shanno algorithm

From an initial guess $\mathbf{x}_0$ and an approximate Hessian matrix B_0 the following steps are repeated as $\mathbf{x}_k$ converges to the solution:

1. Obtain a direction $\mathbf{p}_k$ by solving $B_k \mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
2. Perform a line search to find an acceptable stepsize α_k in the direction found in the first step.
3. Set $\mathbf{s}_k = \alpha_k \mathbf{p}_k$ and update $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$.
4. $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$.
5. $B_{k+1} = B_k + \dfrac{\mathbf{y}_k \mathbf{y}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}} \mathbf{s}_k} - \dfrac{B_k \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}} B_k}{\mathbf{s}_k^{\mathrm{T}} B_k \mathbf{s}_k}$.

$f(\mathbf{x})$ denotes the objective function to be minimized. Convergence can be checked by observing the norm of the gradient, $|\nabla f(\mathbf{x}_k)|$. Practically, B_0 can be initialized with $B_0 = I$, so that the first step will be equivalent to a gradient descent, but further steps are more and more refined by B_k, the approximation to the Hessian.

The first step of the algorithm is carried out using the inverse of the matrix B_k, which can be obtained efficiently by applying the Sherman–Morrison formula to the step 5 of the algorithm, giving

$$B_{k+1}^{-1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) B_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}.$$

This can be computed efficiently without temporary matrices, recognizing that B_k^{-1} is symmetric, and that $\mathbf{y}_k^{\mathrm{T}} B_k^{-1} \mathbf{y}_k$ and $\mathbf{s}_k^{\mathrm{T}} \mathbf{y}_k$ are scalar, using an expansion such as

$$B_{k+1}^{-1} = B_k^{-1} + \frac{(\mathbf{s}_k^{\mathrm{T}} \mathbf{y}_k + \mathbf{y}_k^{\mathrm{T}} B_k^{-1} \mathbf{y}_k)(\mathbf{s}_k \mathbf{s}_k^{\mathrm{T}})}{(\mathbf{s}_k^{\mathrm{T}} \mathbf{y}_k)^2} - \frac{B_k^{-1} \mathbf{y}_k \mathbf{s}_k^{\mathrm{T}} + \mathbf{s}_k \mathbf{y}_k^{\mathrm{T}} B_k^{-1}}{\mathbf{s}_k^{\mathrm{T}} \mathbf{y}_k}.$$

In statistical estimation problems (such as maximum likelihood or Bayesian inference), credible intervals or confidence intervals for the solution can be estimated from the inverse of the final Hessian matrix. However, these quantities are technically defined by the true Hessian matrix, and the BFGS approximation may not converge to the true Hessian matrix.

Maximize the variance on the firsts components of Z

Definition

Let be X a $n \times p$ interval matrix, $Z \in X$,

$$\beta(Z) = \{w_1^Z, \dots, w_s^Z\},$$

where $s \leq p$ and w_i^Z the eigenvectors of the variance-covariance matrix of Z and $\lambda(Z) = \lambda_1^Z, \dots, \lambda_s^Z$ the associated eigenvalues, we define the function

$$\Lambda(Z, s) : X \times \mathbb{N} \to \mathbb{R}^+$$

as follows:

$$\Lambda(Z, s) = \sum_{i=1}^{s} \lambda_i^Z. \tag{10}$$

Algorithm to compute $\Lambda(Z, s)$

Algorithm 4 Calculation of Λ

Require: X interval matrix $n \times p$, $Z \in X$, and s number of principal components.

Ensure: $\Lambda(Z, s)$

1: Apply a PCA on Z and compute:

2: $\lambda(Z) = \lambda_1^Z, \ldots, \lambda_s^Z$
the associated eigenvalues of the variance-covariance matrix of Z.

3: $\Lambda(Z, s) = \sum_{i=1}^{s} \lambda_i^Z$.

4: return $\Lambda(Z, s)$.

Optimization Problem

$$
\begin{aligned}
&\text{Maximize} \quad \Lambda(Z,s) = \sum_{i=1}^{s} \lambda_i^Z \\
&\text{Subject to} \quad \left\{
\begin{array}{c}
a_{11} \leq z_{11} \leq b_{11} \\
\vdots \\
a_{1j} \leq z_{1j} \leq b_{1j} \\
\vdots \\
a_{1p} \leq z_{1p} \leq b_{1p} \\
\vdots \\
a_{ij} \leq z_{ij} \leq b_{ij} \\
\vdots \\
a_{n1} \leq z_{n1} \leq b_{n1} \\
\vdots \\
a_{np} \leq z_{np} \leq b_{np}
\end{array}
\right.
\end{aligned}
\tag{11}
$$

Algorithm 5 Maximizing the variance PCA

Require: X $n \times p$ matrix, $Z \in X$, s number of principal components, `TOL` is a tolerance of variations and `N` the maximum number of iterations.

Ensure: $\tilde{Y}^{V_{Z^\star}}$

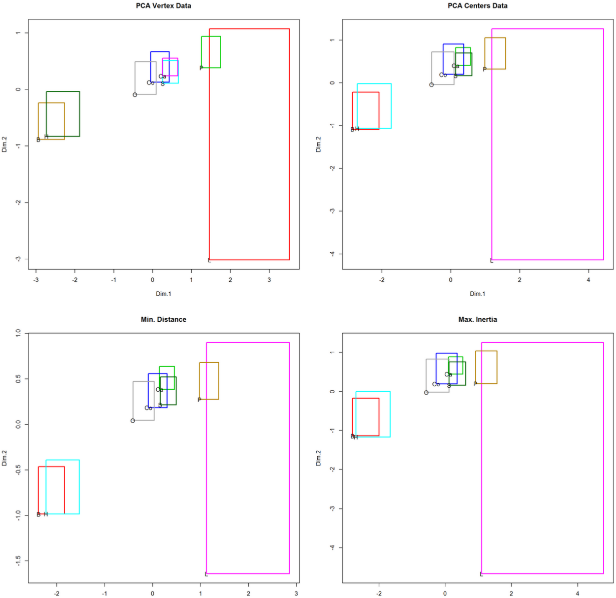
1: Let be $Z = X^c$ the centers matrix as an initial value.

2: Get $Z^\star$ using Broyden–Fletcher–Goldfarb–Shanno algorithm `BFGS(Z, function =` $\Lambda(Z, s)$`,TOL,N)`

3: Get $\tilde{Y}^{V_{Z^\star}}$ applying the Theorem 1

4: **return** $\tilde{Y}^{V_{Z^\star}}$

Oils Data Comparation Plot

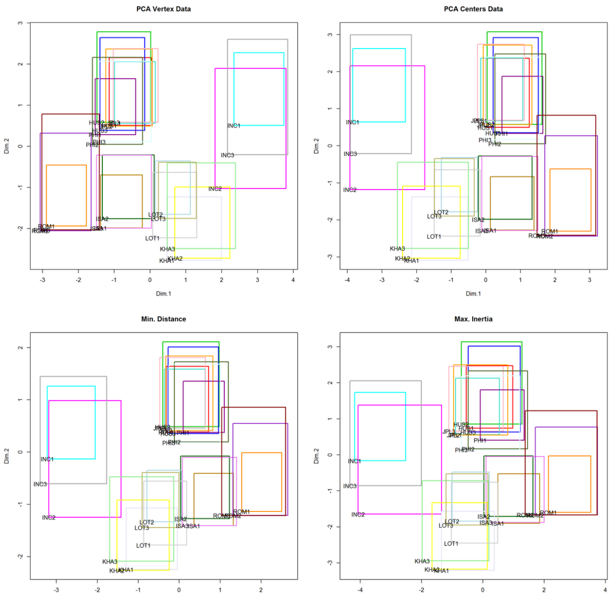


Oils Data Comparation

	ver.dist	cent.dist	min.dist	max.var
1	217.20491979452	226.345477905895	174.834606512487	230.524337279273

	inercia.vertex	inercia.centros	inercia.min.dist	inercia.max.var
1	67.7627771995287	74.6011009457592	73.9787243730559	77.948613220299
2	88.0249162228891	89.7748735221302	90.2834385024301	94.1681130876581
3	97.7724149902723	98.7261909065107	98.6319450630644	99.9999999999999

Faces Data Comparation Plot

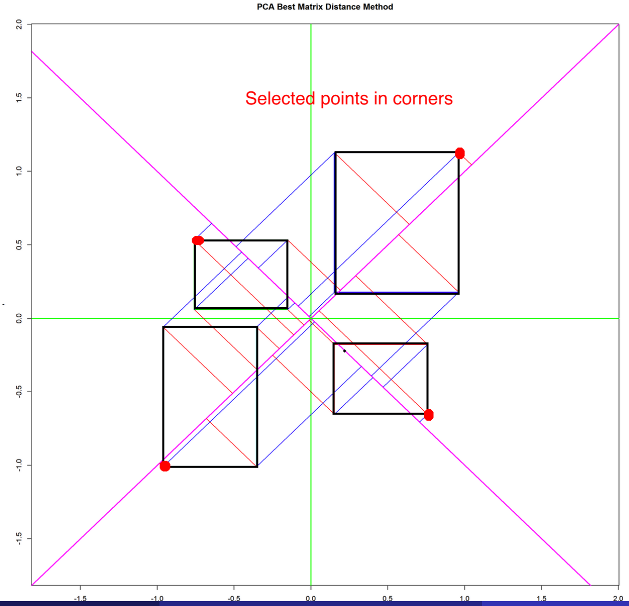


Faces Data Comparation

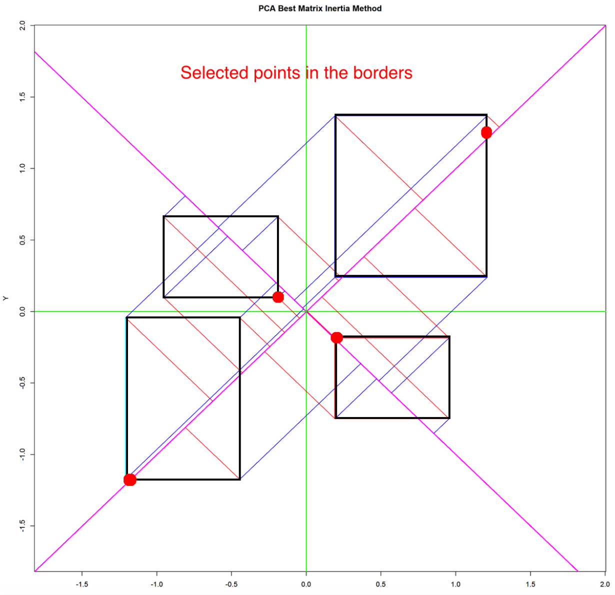
	ver.dist	cent.dist	min.dist	max.var
1	3614.34148419874	3942.95420912549	$\boxed{2820.78369329809}$	3867.53297496876

	inercia.vertex	inercia.centros	inercia.min.dist	inercia.max.var
1	40.0667097210037	43.088416163449	45.1627301022735	52.9691998846087
2	72.0071584552016	80.1717882349615	83.2077760605735	87.3778674340959
3	83.7124062384127	90.5023237657181	92.3406967688621	$\boxed{99.8318016506443}$
4	91.8005821950097	96.3366942724498	96.4459161680779	99.9306315461714
5	96.5842923422871	99.2474365249974	99.5858659478948	99.9873957795102

What points are selected?



What points are selected?



oldemar.rodriguez@ucr.ac.cr

Tomorrow I will show you how to do Optimized PCA in RSDA Package....

Thank You .....