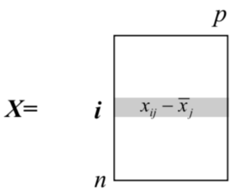
# Multivariate Analysis

Javier Ferrando Monsonis

# PCA

## PCA Analysis in $R^p$




We consider $\mathbf{X}$ as the centered dataset matrix of n observations and p features.

$$N = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}$$

$\mathbf{N}$ is a diagonal matrix containing weights (importance) for each of the observations in the data.
$\mathbf{u}_1 \in R^p$ is considered a unitary vector defining a direction in $R^p$.
$\Psi_{1i}$ represents the projection of the observation $i$ on $\mathbf{u}_1$. When projecting all the individuals on $\mathbf{u}_1$, we get

$$\Psi_1 = \mathbf{X} \cdot \mathbf{u}_1$$

The goal is to obtain orthogonal vectors $\mathbf{u}$ in the directions which maximizes the variance (or inertia $I_n$) of their $\Psi$, maximizing the sum of the individual's projections on $\mathbf{u}$. So, in the case of the First Principal Component the objective function we will try to maximize is

$$\max_{\mathbf{u}_1} I_{total} = \max_{\mathbf{u}_1} \sum_{i=1}^{n} w_i \Psi_{1i}^2 = \Psi_1^\intercal \mathbf{N} \Psi_1 = \mathbf{u}_1^\intercal \mathbf{X}^\intercal \mathbf{N} \mathbf{X} \mathbf{u}_1$$

Subject to $\mathbf{u}_1 \mathbf{u}_1^\intercal = \|\mathbf{u}_1\|_2^2 = 1$

Method of Lagrange multipliers $\rightarrow \mathcal{L}(x, y, \lambda) = f(x, y) - \lambda g(x, y)$,

$$\ell = \mathbf{u}_1^\intercal \mathbf{X}^\intercal \mathbf{N} \mathbf{X} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1^\intercal \mathbf{u}_1 - 1)$$

Setting $\frac{\partial \ell}{\partial u} = 0$

$$2\mathbf{X}^\intercal \mathbf{N} \mathbf{X} \mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0$$

$$\mathbf{X}^\intercal \mathbf{N} \mathbf{X} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

Since we are using a centered matrix, $\mathbf{X}^\intercal \mathbf{N} \mathbf{X} = Cov(\mathbf{X})$, $\mathbf{u}_1$ represents an eigenvector of $Cov(\mathbf{X})$ and $\lambda_1$ its associated eigenvalue. Taking the largest $\lambda_1$ will give the eigenvector with maximum variance (First Principal Component).
$\Psi_\alpha \in R^n$ where each component represent the projection of each individual $i$ on the Principal Component $u_\alpha$.
Since $u_1$ is a unitary vector we deduce from previous formulas that $\Psi_1^\intercal \mathbf{N} \Psi_1 = \lambda_1 = var(\Psi_1)$

$$I_{total} = \sum_{j=1}^{p} \sum_{i=1}^{n} w_i (x_{ij} - \overline{x}_j)^2 = \sum_{j=1}^{p} var(x_j) = \sum_{\alpha=1}^{p} \lambda_\alpha$$

Projected inertia on the first axis

$$I_1 = \sum_{i=1}^{n} \frac{1}{n} \Psi_{1i}^2 = \lambda_1$$

When working with standardized $\mathbf{X}$ matrix, $\mathbf{X}^\intercal \mathbf{N} \mathbf{X} = Cor(\mathbf{X})$

## PCA Analysis in $R^n$

$\mathbf{v}_1 \in R^n$ is considered a unitary vector defining a direction in $R^n$.
$\varphi_{1j}$ denotes the projections of variable j onto $\mathbf{v}_1$, $\mathbf{X}^\intercal \mathbf{N}^{1/2} \mathbf{v}_1$, when using a standardized matrix, $\varphi_1 = cor(\mathbf{X}, \Psi_1)$. The function to maximize is

$$\max_{\mathbf{v}_1} I_{total} = \max_{\mathbf{v}_1} \sum_{j=1}^{p} \varphi_{1j}^2 = \varphi_1^\intercal \varphi_1 = \mathbf{v}_1^\intercal \mathbf{N}^{1/2} \mathbf{X} \mathbf{X}^\intercal \mathbf{N}^{1/2} \mathbf{v}_1$$

Subject to $\mathbf{v}_1 \mathbf{v}_1^\intercal = \|\mathbf{v}_1\|_2^2 = 1$
Following the same optimization procedure that in $R^p$ we get

$$\mathbf{N}^{1/2} \mathbf{X} \mathbf{X}^\intercal \mathbf{N}^{1/2} \mathbf{v}_1 = \lambda_1 \mathbf{v}_1$$

Transition relationships between both fits

$$\mathbf{u}_\alpha = \lambda^{-1/2} \mathbf{X}^\intercal \mathbf{N}^{1/2} \mathbf{v}_\alpha$$

$$\mathbf{v}_\alpha = \lambda^{-1/2} \mathbf{N}^{1/2} \mathbf{X} \mathbf{u}_\alpha$$

## Singular Value Decomposition

Let's call $\mathbf{M} = \mathbf{N}^{1/2} \mathbf{X}$

$$\mathbf{M} \mathbf{u}_\alpha = \mathbf{v}_\alpha \sqrt{\lambda_\alpha}$$

$$\mathbf{M}^\intercal \mathbf{v}_\alpha = \mathbf{u}_\alpha \sqrt{\lambda_\alpha}$$

In matrix form

$$\mathbf{M} \mathbf{U} = \mathbf{V} \Lambda^{1/2} \rightarrow \mathbf{M} = \mathbf{V} \Lambda^{1/2} \mathbf{U}^\intercal$$

So, the singular values of $\mathbf{M}$ are the ones contained in the diagonal of $\Lambda^{1/2}$, been the eigenvalues of $\mathbf{M}\mathbf{M}^\intercal = \mathbf{N}^{1/2} \mathbf{X} \mathbf{X}^\intercal \mathbf{N}^{1/2}$ the square of the singular values obtained.

## Attributes from PCA RFactominer

Having pca$ind and pca$var as the objects returned by PCA function.

- coord

  Values of the projections of individuals and variables on the Principal Components

- cos2

  Contribution (importance) of a component to the squared distance of the observation to the origin (G) in the original cloud of points. Quality of the representations.

  $$cos^2(i, \alpha) = \frac{\Psi_{\alpha i}^2}{d_{i,G}^2}$$

  $$cos^2(j, \alpha) = \frac{\varphi_{\alpha j}^2}{s_j^2}$$

- contrib

  Contribution of an individual or variable to the variance explained by a component $\alpha$

  $$ctr(i, \alpha) = \frac{w_i \Psi_{\alpha i}^2}{\lambda_\alpha}$$

  $$ctr(j, \alpha) = \frac{\varphi_{\alpha j}^2}{\lambda_\alpha}$$

  Factominer $contrib multiplies by 100 these values, so the sum of contributions is 100.

- dist($ind)
- cor($var)

  Correlation between a component and a variable $\varphi_{\alpha j} = cor(x_j, \Psi_1)$ (standardized $\mathbf{X}$). How much information they share.

## Supplementary variables

### Categorical variables

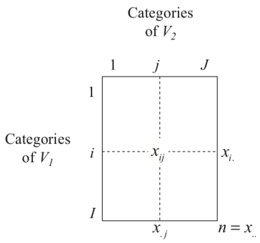
In $R^p$, its displayed the projection of the centroid of the individuals which share each of the categories onto the Principal Components.

### Continuous variables

In $R^n$, the correlation between the supplementary variable and the Principal Components are shown.

# Correspondence Analysis

## CA




$$x_{i\bullet} = \sum_{j=1}^{J} x_{ij} \quad x_{\bullet j} = \sum_{i=1}^{I} x_{ij} \quad n = x_{\bullet\bullet} = \sum_{i,j} x_{ij}$$

In CA it is also considered the probability tables associated with contingency tables as the general term $f_{ij} = x_{ij}/n$, the probability of carrying both the categories i (of V1) and those of j (V2)

$$f_{i\bullet} = \sum_{j=1}^{J} f_{ij} \quad f_{\bullet j} = \sum_{i=1}^{I} f_{ij} \quad f_{\bullet\bullet} = \sum_{i,j} f_{ij} = 1$$

### Independence Model and $\chi^2$ Test

$$\chi^2 = \sum_{i,j} \frac{(\text{Actual Sample Size} - \text{Theoretical Sample Size})^2}{\text{Theoretical Sample Size}},$$

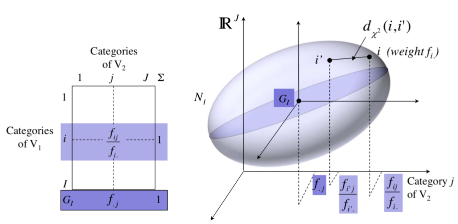
$$\chi^2 = \sum_{i,j} \frac{(nf_{ij} - nf_{i\bullet}f_{\bullet j})^2}{nf_{i\bullet}f_{\bullet j}} = n\sum_{i,j} \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}} = n\Phi^2,$$

If each category of $V_1$ where independent from every category of $V_2$

$$\forall i, j \quad \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$$

$f_{\bullet j}$ is the conditional probability $P(j|i) = \frac{P(j,i)}{P(i)}$. So, the probability of carrying category $j$ when carrying category $i$ does not depend on the category $i$ (in the independence model).

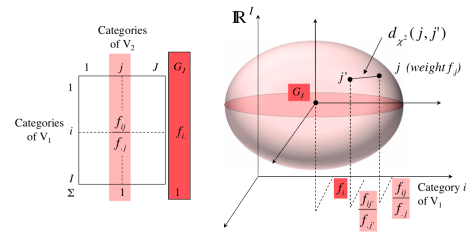
### The cloud of row profiles




Distance between two profiles: $d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$

Distance to the mean profile $G_I$: $d^2_{\chi^2}(i, G_I) = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2$

### The cloud of column profiles



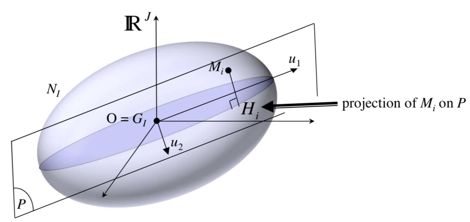

Distance between two profiles: $d^2_{\chi^2}(j,j') = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2$

Distance to the mean profile $G_J$: $d^2_{\chi^2}(j, G_J) = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.}\right)^2$

The further the data is from independence, the more the profiles spread from the origin

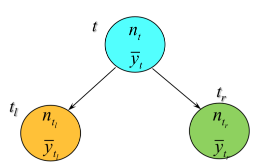
$$\begin{aligned} Inertia(N_I/G_I) &= \sum_{i=1}^{I} Inertia(i/G_I) = \sum_{i=1}^{I} f_{i.} d^2_{\chi^2}(i, G_I) \\ &= \sum_{i=1}^{I} f_{i.} \left(\sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2\right) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = \frac{\chi^2}{n} = \phi^2 \end{aligned}$$

$\phi^2$ measures the strength of the link

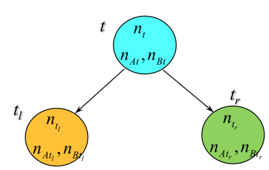



Find $\mathbf{P}$ that maximizes $\sum_{i=1}^{I} f_{i\bullet}(OH_i)^2$

# Decision trees




*Regression tree*




*Classification tree*

## Split criterion

### AID

AID split criterion is based on decomposition of variance

$$\sum_{i=1}^{n_t} (y_i - \overline{y}_t)^2 = \sum_{k=1}^{q} n_{t_k}(\overline{y}_{t_k} - \overline{y}_t) + \sum_{k=1}^{q}\sum_{i \in t_k}^{n_{t_k}} (y_i - \overline{y}_{t_k})^2$$

Where the first term of the equation refers to the variance between child nodes $t_k$ and parent node $t$ and the second term, the variance within child nodes.
$y_i$ denotes the response for every individual $i$ out of $n_t$ (number of individuals in node $t$), $q$ the number of children nodes (2 in a binary tree), $\overline{y}_t$ the mean response in node $t$.
We can now calculate the F statistic, $F = \frac{\text{between-nodes variability}}{\text{within-nodes variability}}$

$$F = \frac{\sum_{k=1}^{q} n_{t_k}(\overline{y}_{t_k} - \overline{y}_t)/q - 1}{\sum_{k=1}^{q}\sum_{i \in t_k}^{n_{t_k}} (y_i - \overline{y}_{t_k})^2/n - q}$$

The goal is to obtain the feature and its cutpoint that leads to the highest F value, increasing as much as possible the between-nodes variability.

### CHAID

CHAID split criterion is based on the Chi-square statistic comparing the frequency of each class and children node

$$\chi^2 = \sum_{k=1}^{m}\sum_{j=1}^{q} \frac{\left(n_{kt_j} - n_{k\cdot} \cdot \frac{n_{\cdot t_j}}{n_t}\right)^2}{n_k \cdot \frac{\cdot n_{t_j}}{n_t}}$$

The goal is to obtain the feature and its cutpoint that leads to the highest $\chi^2$ value.

## Impurity of a node

$p(j|t)$ probability of class $j$ in node $t$

### Categorical response

- Gini

$$i(t) = \sum_{i \neq j} p(j|t)p(i|t) = 1 - \sum_{j}^{q} p_j^2$$

- Information (Entropy)

$$i(t) = \sum_{j} p(j|t) log_2 p(j|t)$$

### Continuous response

- Variance

$$i(t) = \frac{\sum_{i \in t}(y_i - \overline{y}_t)}{n}$$

The objective is to maximize the decrement of impurity between the parent and its children. The decrement of impurity is defined as follows

$$\Delta i(t) = i(t) - \frac{n_{tl}}{n_t} i(t_l) - \frac{n_{tr}}{n_t} i(t_r)$$

## Cost of the tree

Cost of a node (classification tree)

$$r(t) = 1 - max_j p(j|t)$$

Cost of a node (regression tree)

$$r(t) = \frac{1}{n_t} \sum_{i \in t}^{n_t} (y_i - \overline{y}_t)^2$$

Cost of a classification tree

$$R(t) = \frac{\sum_{t \in T} p(t) r(t)}{r(root)} \cdot 100$$

Cost of a regression tree (guessing)

$$R(t) = \sum_{t \in T} \frac{1}{n_t} \sum_{i \in t}^{n_t} (y_i - \overline{y}_t)^2$$

The criterion to optimize is to minimize $R(t)$

### Penalization of complexity

Since the previous objective function would lead to large trees, we use $\alpha$ as a complexity parameter to control its size.

The new objective function becomes

$$Min(R(t) + \alpha |T|)$$

### Model selection

Training data: train trees with increasing values of $\alpha$. Each obtained tree will have $Min(R(t))$ within the set of trees with complexity $\alpha$ ($|T|$).
Validation data: calculate every tree $R(T)$ using validation data and get the optimum one.

## ROC and Concentration curves

**Confusion matrix**

| In Test data | Predicted class YES | Predicted class NO | |
|---|---|---|---|
| Real class YES | *TP* | *FN* | *P* |
| Real class NO | *FP* | *TN* | *N* |

$$Precision = \frac{1}{2}\left[\frac{TP}{TP + FP} + \frac{TN}{TN + FN}\right]$$

$$Accuracy = 1 - \frac{FN + FP}{n}$$

$$Recall = Sensitivity = \frac{TP}{P}$$

# Association Rules

## Support

$$Support(I_k) = \frac{|T|\{I_k\} \subseteq T}{|\tau|}$$

Probability of finding $I_k$ itemset in the set $T$ of transactions

## Confidence

$$Confidence(LHS \to RHS) = \frac{Support(LHS, RHS)}{Support(LHS)}$$

Probability of RHS, having occurred LHS

$$P(RHS|LHS) = \frac{P(RHS, LHS)}{P(LHS)}$$

## Lift

$$Lift(LHS \to RHS) = \frac{Support(LHS, RHS)}{Support(LHS) \cdot Support(RHS)}$$

Lift $\in [0, \infty]$, can be interpreted as how much better is a rule than a random prediction of the consequent (RHS).For Lift values $< 1$, should rely on Support(RHS) rather than following the rule.