Contents

Part I

Part: Unsupervised learning

1 Unsupervised learning

1.1 Introduction

- In the unsupervised learning setting, we we have only a set of features $X_1, X_2, \ldots, X_p$ measured on n observations.
- The goal is to discover (unknown and unanticipated) structures (*patterns*) in the data on $X_1, X_2, \ldots, X_p$.
- (We do not have a priviledged response variable Y to explain or predict.)

1.2 General objectives

We look for

- *relationships* between variables, units, and between both perspectives
- *reducing* data dimensions (in particular the space in which units live, $\Re^p$)
- *grouping* data (find subgroups among variables or units)
- *detecting anomalous* data

2 Principal component analysis

2.1 Objectives

Principal component analysis is a statistical methodology for the **dimensionality reduction**.

In particular, reduction **on the space of units**, $\Re^p$.

This also to

- produce a synthesis of the relationships between variables
- allow for a graphical representation on the plan or on $\Re^3$.

Dimensionality reduction

When the variables are numerous, it is difficult to grasp the existing structures in the data.

A question then arises: whether it is possible to represent observations, rather than in the original space $\Re^p$, in a low-dimension space ($\Re$, $\Re^2$, $\Re^3, \ldots$), minimizing the loss of information resulting from this reduction.

Demographic data for European countries (1999) (from the Statistical Yearbook 2002)

EU countries: first 15 composing **Europe of 15** (from 1995)

	Paesi	NAT	MOR	CR.N	MIG	CR.T	M.IN	VEC	N.FI	E.P	NU	E.M	V.M	V.F
1	Austria (1995)	9.7	9.7	−0.0	2.4	2.4	4.3	92.0	1.32	28.1	4.9	27.0	75.1	81.0
2	Belgio (1958)	11.2	10.3	0.9	1.6	2.5	5.3	95.0	1.61	28.5	4.3	26.1	74.8	81.1
3	Danimarca (1973)	12.4	11.1	1.3	1.8	3.1	4.2	80.6	1.73	29.6	6.7	29.7	74.2	79.0
4	Finlandia (1995)	11.2	9.5	1.6	0.7	2.3	3.6	81.4	1.74	29.6	4.7	27.7	73.8	81.0
5	Francia (1958)	12.6	9.2	3.4	0.8	4.3	4.8	84.3	1.77	29.3	4.8	27.6	74.9	82.4
6	Germania (1958)	9.4	10.3	−0.9	2.5	1.5	4.7	100.7	1.36	28.6	5.2	26.9	74.5	80.5
7	Grecia (1981)	9.7	9.8	−0.1	2.4	2.3	5.9	113.6	1.30	28.7	5.9	26.5	75.4	80.4
8	Irlanda (1973)	14.2	8.4	5.8	5.3	11.1	5.5	51.3	1.88	30.5	4.9	28.2	73.9	79.0
9	Italia (1958)	9.3	9.9	−0.6	1.8	1.2	5.2	124.5	1.22	30.0	4.8	27.6	76.0	82.1
10	Lussemburgo (1958)	12.9	8.8	4.2	10.9	15.0	2.9	75.4	1.73	29.4	4.8	27.4	74.7	81.2
11	NL (1958)	12.7	8.9	3.8	2.8	6.6	5.2	73.1	1.65	30.3	5.7	27.7	75.3	80.5
12	Portogallo (1986)	11.6	10.8	0.8	1.0	1.8	5.6	91.6	1.49	28.6	6.9	25.8	72.0	79.1
13	UK (1973)	11.8	10.6	1.1	2.8	3.9	5.8	81.9	1.68	28.4	5.1	27.0	75.0	79.8
14	Spagna (1986)	9.6	9.4	0.2	1.0	1.2	4.9	110.6	1.20	30.6	5.2	27.5	75.0	82.0
15	Svezia (1995)	10.0	10.7	−0.7	1.5	0.8	3.4	93.5	1.50	29.8	4.0	29.8	77.1	81.9
16	UE	10.4	9.9	0.5	1.9	2.4	5.0	97.2	1.48	29.2	5.1	27.3	74.9	81.1

	Paesi	NAT	MOR	CR.N	MIG	CR.T	M.IN	VEC	N.FI	E.P	NU	E.M	V.M	V.F
17	Albania*	0.2	5.1	−4.9	−3.8	−8.7	30.7	18.8	2.60	28.3	7.2	22.9	68.5	75.4
18	Bielorussia	9.2	14.1	−4.9	−10.9	−15.8	11.5	70.2	1.29	24.9	7.2	22.2	62.2	74.0
19	Bosnia*	13.1	7.6	5.6	−4.5	1.1	11.0	28.5	1.56	26.0	6.3	23.3	69.7	75.2
20	Bulgaria* (2007)	8.8	13.6	−4.8	0.0	−4.8	14.9	101.8	1.23	24.7	4.3	23.5	68.3	75.1
21	Cipro* (2004)	12.7	7.6	5.2	−0.8	4.4	6.0	48.6	1.84	28.6	12.8	25.6	75.3	80.4
22	Croazia* (2013)	10.4	11.5	−1.1	−1.0	−2.1	7.7	62.5	1.38	27.8	5.2	25.1	69.9	76.8
23	Estonia* (2004)	8.7	12.8	−4.2	−0.3	−4.5	9.6	80.6	1.24	26.6	3.9	24.5	65.5	76.3
24	Islanda*	14.8	6.9	7.9	4.0	11.9	2.4	49.6	1.99	28.7	5.6	29.8	77.8	81.5
25	Iugosl.	12.4	10.5	1.9	−0.0	1.8	13.8	62.6	1.77	26.8	5.3	24.2	70.0	74.9
26	Lettonia* (2004)	8.0	13.5	−5.6	−0.8	−6.3	11.6	82.2	1.16	26.8	3.9	24.2	64.8	75.4
27	Lituania* (2004)	9.8	10.8	−1.0	0.4	−0.6	8.6	67.5	1.35	26.5	4.8	23.1	67.0	77.2
28	Macedonia*	14.5	8.4	6.1	−1.0	5.1	14.7	41.4	1.76	26.2	7.0	23.3	70.4	74.5
29	Malta* (2004)	11.9	7.9	4.0	1.6	5.6	7.2	57.5	1.81	28.9	6.3	26.7	74.0	80.1
30	Moldova	10.1	11.4	−1.3	−1.4	−2.7	17.9	38.0	1.67	25.3	6.5	21.7	64.2	71.5
31	Norvegia	13.3	10.1	3.2	4.3	7.4	3.9	76.3	1.84	29.3	5.3	28.2	75.6	81.1
32	Polonia*(2004)	9.9	9.9	0.0	−0.4	−0.3	8.8	61.7	1.37	27.2	5.7	23.3	68.2	77.2
33	Rep. ceca* (2004)	8.7	10.7	−2.0	0.9	−1.1	4.6	83.1	1.13	26.9	5.2	24.1	71.4	78.2
34	Romania* (2007)	10.4	11.8	−1.4	−0.1	−1.5	18.5	71.2	1.30	25.6	6.2	23.2	67.1	74.1
35	Russia	8.3	14.7	−6.4	1.1	−5.3	16.9	68.4	1.17	25.7	6.2	22.1	59.9	72.4
36	S.Marino	11.3	7.5	3.8	11.3	15.1	3.3	106.3	1.30	32.2	8.7	28.4	80.4	82.6
37	Slovacchia* (2004)	10.4	9.7	0.7	0.3	1.0	8.2	57.5	1.33	26.4	5.1	23.2	69.0	77.2
38	Slovenia* (2004)	8.8	9.7	−0.9	5.6	4.7	4.2	86.0	1.21	28.0	3.9	26.3	71.8	79.3
39	Svizzera	11.0	8.7	2.2	3.5	5.7	4.6	87.6	1.48	29.7	5.7	27.7	76.8	82.5
40	Turchia*	21.1	6.3	14.8	0.6	15.5	37.9	17.0	1.10	26.6	6.2	22.6	66.5	71.2
41	Ungheria* (2004)	9.4	14.2	−4.8	0.0	−4.8	8.4	85.5	1.29	27.1	4.5	24.2	66.4	75.2

On May 1st 2004, the **Europe of 25** is formed.

Bulgaria and Romania were added in 2007.

On July 1th 2013, Croatia is joined.

Seven other countries are currently candidates: Albania, Bosnia-Herzegovina, Iceland, Kosovo, Macedonia, Montenegro, Serbia, Turkey.

In 2016, citizens of the United Kingdom after a referendum expressed their will to leave the European Union (51.9% instead of staying in the EU, 48.1%)

- NAT : Number of live births (per 1,000 population)
- MOR : Number of deaths (per 1,000 population)
- CR.N : Natural increase (per 1,000 population)
- MIG : Net migration estimate (per 1,000 population)
- CR.T : Total population increase (per 1,000 population)
- M.IN : Infant mortality rate
- VEC : Old-age index 65+/0-14 (%)
- N.FI : Mean number of children per woman
- E.P : Mean age at first birth (years)
- NU : Marriage index
- E.M : Mean age at first marriage (years)
- V.M, V.F : Life expectancy at birth (years) for males and females

2.2 Two ways of computing

2.2.1 Geometric via

Consider, for example, the case of only two variables:

- expected life for males, $x_1 = V.M$
- expected life for females, $x_2 = V.F$

collected in 40 European countries in 1999.

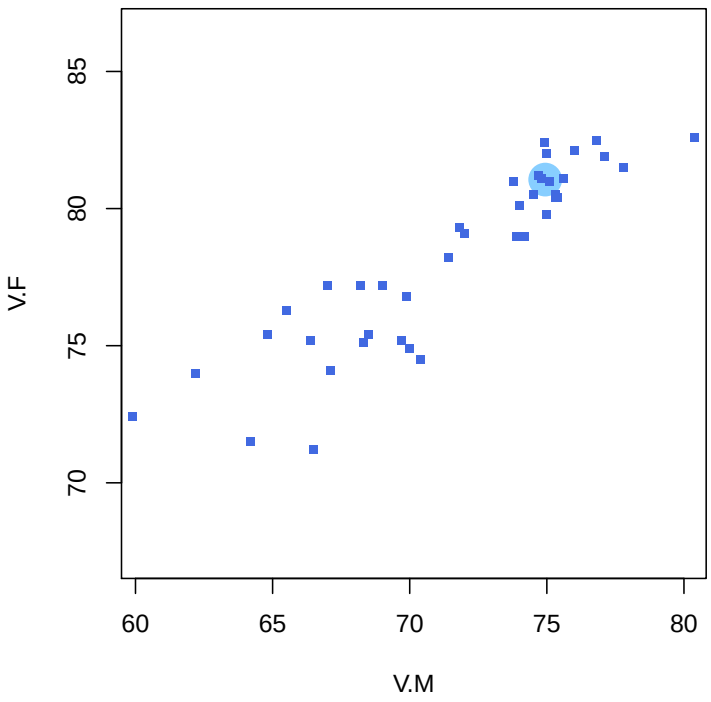

Figure 1: Original variables $X = (x_1, x_2)$

Variables x_1 and x_2 are clearly **non-orthogonal** (considering the variables centered in the mean vector, or, in statistical terms, they are not incorrelated, in fact the cofficient of linear correlation is 0.91).

- Suppose to **rotate** the coordinate axes so that the first axis is **in the direction of greater variability**.
- Also, rotate the axes so that the variables are **orthogonal in the new system (v_1,v_2)**.
- In the general case of p variables, this reasoning extends for a generic p number of variables.

Note that the origin of the Cartesian axis system has been translated into the centroid or mean vector $\bar{\mathbf{x}}$ of X.

In the following, we will consider the matrix of data centered in $\bar{\mathbf{x}}$, $\tilde{X} = X - \mathbf{1_n}\bar{\mathbf{x}}^{\mathbf{T}} = (\tilde{\mathbf{x}}_{\mathbf{1}}, \ldots, \tilde{\mathbf{x}}_{\mathbf{p}})$.

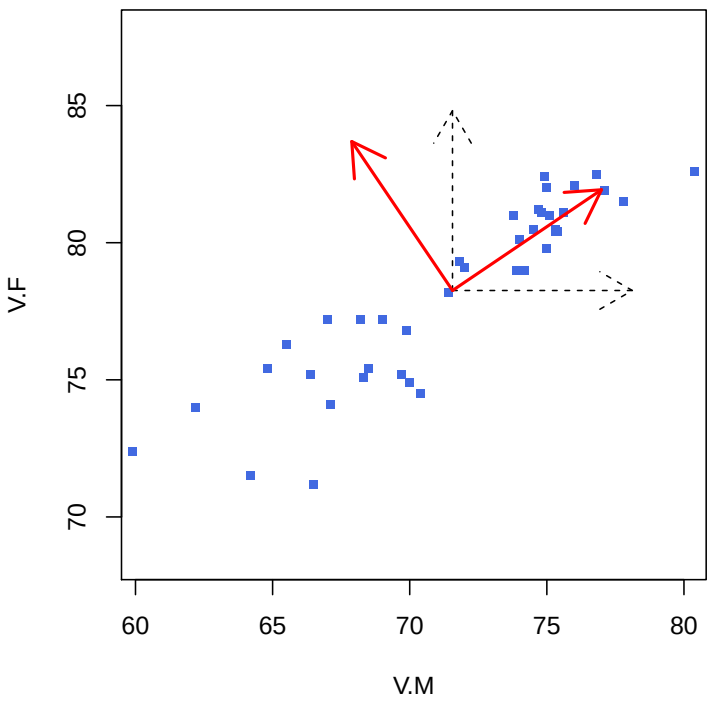

Figure 2: New reference axes $V = (v_1, v_2)$

Geometric interpretation: case $p = 2$

Values (or the sc **scores**) of new variables, $\mathbf{z_1}$ e $\mathbf{z_2}$ (sc **principal components**) are, $\mathbf{x_1}$ *and* $\mathbf{x_2}$ *expressed wrt the new axis system* V, or, in more formal terms, the linear combinations of original variables, $\tilde{\mathbf{x}}_\mathbf{1}$ e $\tilde{\mathbf{x}}_\mathbf{2}$, through the rotation matrix $V = (\mathbf{v_1}, \mathbf{v_2})$

$$Z = \tilde{X}V$$

where $Z = (\mathbf{z_1}, \mathbf{z_2})$ and $\tilde{X} = (\tilde{\mathbf{x}}_\mathbf{1}, \tilde{\mathbf{x}}_\mathbf{2})$, i.e. $\mathbf{z_1}$ and $\mathbf{z_2}$ are

$$\begin{aligned} \mathbf{z_1} &= \tilde{X}\mathbf{v_1} & \mathbf{z_2} &= \tilde{X}\mathbf{v_2} \\ \mathbf{z_1} &= \tilde{\mathbf{x}}_\mathbf{1} \cdot \mathbf{v_{11}} + \tilde{\mathbf{x}}_\mathbf{2} \cdot \mathbf{v_{21}} & \mathbf{z_2} &= \tilde{\mathbf{x}}_\mathbf{1} \cdot \mathbf{v_{12}} + \tilde{\mathbf{x}}_\mathbf{2} \cdot \mathbf{v_{22}} \end{aligned}$$

Computation of principal components

How do we calculate the new directions (or the rotation matrix V)? We have to obtain a rotation matrix $p \times p$, V, such that:

$$Z = \tilde{X}V \qquad et \qquad Z^T Z = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$$

i.e. columns of Z–the new variables–have to be orthogonal, and moreover

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0\,.$$

Since, replacing the expression above, we obtain that

$$Z^T Z = V^T \tilde{X}^T \tilde{X} V = \Lambda$$

diagonal matrix, $\lambda_1, \ldots, \lambda_p$ are the **eigenvalues** of $\tilde{X}^T \tilde{X}$ and the columns of V are the **eigenvectors** of $\tilde{X}^T \tilde{X}$.

2.2.2 Algebric via

1. We have to obtain the *linear combination* Z_1 of $X = (X_1, \ldots, X_p)^T$ *having the largest variance*, i.e.

$$\mathbf{v_1} \; : \; \underbrace{\max}_{\mathbf{v_1} \in \mathbf{R^p}} \; \mathrm{Var}\{\mathbf{v_1^T X}\} \quad \text{u.c.} \quad \mathbf{v_1^T v_1 = 1}$$

2. hence, to obtain the linear combination Z_2 of X, having the largest variance below that of Z_1 and *uncorrelated* with Z_1, i.e.

$$\mathbf{v_2} \; : \; \underbrace{\max}_{\mathbf{v_2} \in \Re^\mathbf{p}} \; \mathrm{Var}\{\mathbf{v_2^T X}\} \quad \text{s.v.} \quad \mathbf{v_1^T v_2 = 0} \text{ et } \mathbf{v_2^T v_2 = 1}$$

$\ldots p$. then, continue in obtaining coefficient vectors $\mathbf{v_j}$ that produce a linear compound with the largest variance and uncorrelated with the already defined linear compounds.

Computation

We have that

$$\mathrm{Var}(Z_1) = \mathrm{Var}(\mathbf{v_1^T X}) = \mathbf{v_1^T \Sigma v_1}$$

where Σ is the covariance matrix of X being estimated by the sample covariance matrix S

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^\mathbf{T}.$$

Notice that in algebric via we talked about variables and their theorical distributions (we utilized symbols X, Z, μ, Σ, P, etc.). But, what we really have is the sample data (or matrix $X = \{\mathbf{x_j}\}$ from which we derive estimates $\bar{\mathbf{x}}$, S, R, etc.).

Thus, from now on, when we talk about principal components we mean **sample principal components**, defined as linear combinations of observations $\mathbf{x_j}$ having the largest sample variance and mutually uncorrelated.

When helpful, we resort to geometric interpretation and, whenever possible, use a graphical representation.

Then,

$$\begin{aligned}
&\underbrace{\max}_{\mathbf{v_1} \in \Re^\mathbf{p}} \; \mathbf{v_1^T S v_1} \quad \text{s.v. } \mathbf{v_1^T v_1 = 1} \\
&\underbrace{\max}_{\mathbf{v_2} \in \Re^\mathbf{p}} \; \mathbf{v_2^T S v_2} \quad \text{s.v. } \mathbf{v_1^T v_2 = 0} \text{ et } \mathbf{v_2^T v_2 = 1} \\
&\vdots \\
&\underbrace{\max}_{\mathbf{v_p} \in \Re^\mathbf{p}} \; \mathbf{v_p^T S v_p} \quad \text{s.v. } \mathbf{v_j^T v_p = 0} \; \forall \mathbf{j < p} \text{ et } \mathbf{v_p^T v_p = 1}
\end{aligned}$$

By applying the method of *Lagrange multiplers*, we obtain that $\mathbf{v_j}$ corresponds to the eigenvector of S associated with the j-the largest eigrnvalue λ_j, or, V and $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ are such that:

$$S = V \Lambda V^T$$

2.3 Derived properties

- The jth *PC* is given by:
$$\mathbf{z_j} = \tilde{\mathbf{X}}\mathbf{v_j}$$
where $\tilde{X}$ is the data matrix centered in the mean vector ($\tilde{X} = X - \mathbf{1_n}\bar{\mathbf{x}}^\mathbf{T}$)
- *Score* z_{ij} of individual i on the jth PC is obtained as:
$$\tilde{\mathbf{x}}_\mathbf{i}^\mathbf{T}\mathbf{v_j}$$
where $\tilde{\mathbf{x}}_\mathbf{i} = \mathbf{x_i} - \bar{\mathbf{x}}$

Given the spectral (orthogonal) decomposition $S = V\Lambda V^T$,

- $\text{Var}(\mathbf{z_j}) = \text{Var}(\mathbf{Xv_j}) = \mathbf{v_j^T S v_j} \overset{\mathbf{Sv_j}=\lambda_\mathbf{j}\mathbf{v_j}}{=} \mathbf{v_j^T}\lambda_\mathbf{j}\mathbf{v_j} \overset{\mathbf{v_j^T v_j}=\mathbf{1}}{=} \lambda_\mathbf{j}$
- whence the sum of variances of the overall p PC or of eigenvalues is equal to the trace of the matrix being diagonalized:
$$\sum_{j=1}^{p} \lambda_j = \text{trace}(S)$$
with $\text{trace}(S) = \sum_{j=1}^{p} var(\mathbf{x_j}) = \mathbf{VAR_T}$.

VAR_T, referenced to as **total variance**, is an estimate of one of the possible measures of multidimensional variability.

- Hence, the *proportion of total variance* explained by the jth PC is equal to
$$\frac{\lambda_j}{VAR_T}$$
- and the rate of total variance *explained by the first* p' *PC* ($p' < p$) is equal to
$$\frac{\sum_{j=1}^{p'} \lambda_j}{VAR_T}$$

- The covariance of original variables X with the PC is given by:
$$cov(X, \mathbf{z_j}) = \mathbf{cov}(\mathbf{X}, \mathbf{Xv_j}) = \mathbf{Sv_j} = \lambda_\mathbf{j}\mathbf{v_j}$$
- whence
$$cov(\mathbf{x_r}, \mathbf{z_s}) = \lambda_\mathbf{s}\mathbf{v_{rs}}, \quad \mathbf{r, s} = \mathbf{1}, \ldots, \mathbf{p}$$
$$cor(\mathbf{x_r}, \mathbf{z_s}) = \frac{\mathbf{cov}(\mathbf{x_r}, \mathbf{z_s})}{\mathbf{s_r}\sqrt{\lambda_\mathbf{s}}}$$
$$\frac{\sqrt{\lambda}_s v_{rs}}{s_r}$$
If the variable $\mathbf{x_r}$ is standardized, it remains only $cor(\mathbf{x_r}, \mathbf{z_s}) = \sqrt{\lambda}_\mathbf{s}\mathbf{v_{rs}}$.

2.4 PC from matrix R

PCs can also be derived from the (sample) correlation matrix R of the original variables.

- When you do not want different variances of the original variables affect the PCA results.
  - if $\mathbf{x_1}, \ldots, \mathbf{x_p}$ are variables of different type ($\rightarrow$ (arbitrary) measurement units are different);
  - or if there is a significant difference between variances of $\mathbf{x_1}, \ldots, \mathbf{x_p}$ ($\rightarrow$ variables with larger variances tend to dominate the first PCs).
- In general, PCs derived from R are different from PCs derived from S.
- PCs derived from R are equivalent to PCs derived from standardized variables.

2.5 Choosing the appropriate number of components

In general, three guidelines are followed

- components must be sufficient to explain a certain portion of (total) variance
- using the **screeplot**, stop at the component at which the curve shows an 'elbow' (since it identifies the last component that explains a considerable portion of the total variance)
- hold the sole components whose variance is higher than the average variance of the components.

In any case, there is no incontrovertible rule and the issue is debated until recently (see a comparison of 20 'stopping rules' in Peres-Neto, Jackson, and Somers (2005)).

2.6 Summary

Pros

- PCA is a technique for *linear* **dimensionality reduction**
  - summarizes information of a large number of variables in a limited number of components, linear combinations of original variables
- components **extract the maximum** *variance* from the original variables
- components are **orthogonal**

Contras

- components are **difficult to interpret**
  - original variables have non-trascurable loadings on more than one component.
  - in general, only the first component is a *block-component*, the others being of *difference-component* type (still more difficult to interpret)
  - Some solutions consist in obtaining *sub-optimal components* in place of the original PCs (rotation methods; Simple component analysis (SCA) (Rousson and Gasser 2004) implemented in R pkg `sca`).
- Outcome of PCA is **sensitive to outliers**
  - a solution in R pkg `pcaMethods` (2014) (Bioconductor repository), which, moreover, allows PCA on data with missing values

---

- (Linear) PCA is **not adequate**
  - to address any **nonlinear** relationships between original variables

    - There are various extensions of PCA. One is the sc *nonlinear PCA* (Linting M. and Kooij A.J 2007) wich essentially consists in a ordinary PCA where original variables are transformaed in order to maximize the variance explained by the k PCs. It is called *optimal scaling*.
- to deal with **non numerical** variables.
    - *Multiple correspondence analysis*—technique for dimensionality reduction for categorial variables, can be interpreted as a PCA of contingency matrix.

3 Biplot

3.1 Introduction

- Biplot can be seen as a multidimensional generalization of the common scatter plot, by overlapping the representation of unit profiles to the representation of the variables.
- **Bi** therefore means that it is able to provide a joint representation of the row and column units of the data X, i.e. a **2-way** representation.
- In practice, such representation is generally approximated in 2 dimensions in such a way that the resulting reference system gives an *optimal approximation* of the values contained in X.

3.2 PC Biplot

We will consider the **Principal Components Biplot** which is one derivation or, better, another way of reading the PCA.

This technique is based on the **singular value decomposition**.

3.2.1 A graphical sketch of the 2-way representation

3.2.2 Singular value decomposition

Given X(n×p), rank $r(\leq \min(n,p))$,

$$X = U\,\Gamma\,V^T = \sum_{k=1}^{r} \gamma_k u_k v_k^T$$

- U(n×r): matrix of normalized eigenvectors of XX^T (ortonormal: $U^TU=I$), sc **left singular vectors**
- V(r×r): matrix of normalized eigenvectors of X^TX (ortonormal: $V^TV=I$), sc **right singular vectors**
- Γ(r×r): matrix of $\sqrt{(}\lambda_j)$, λ_js **singular values** of XX^T (or XX^T)

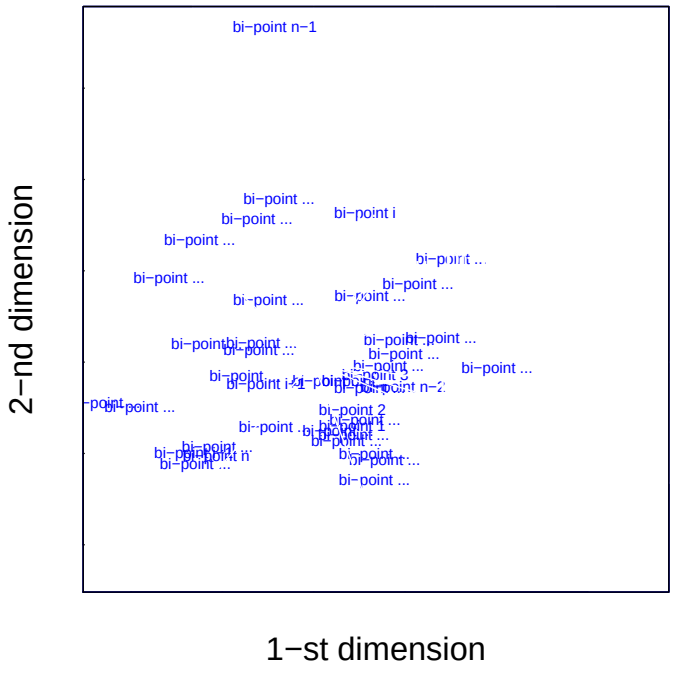

Figure 3: Biplot: units (points).

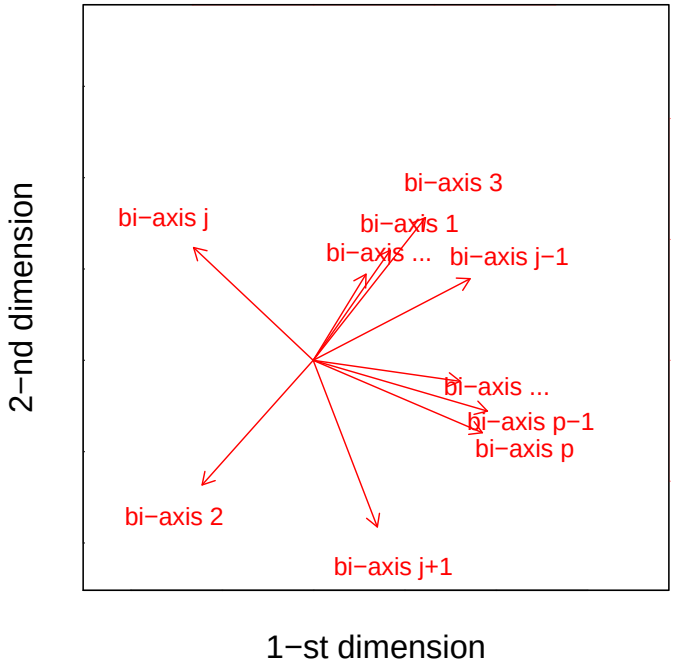

Figure 4: Biplot: variables (axes).

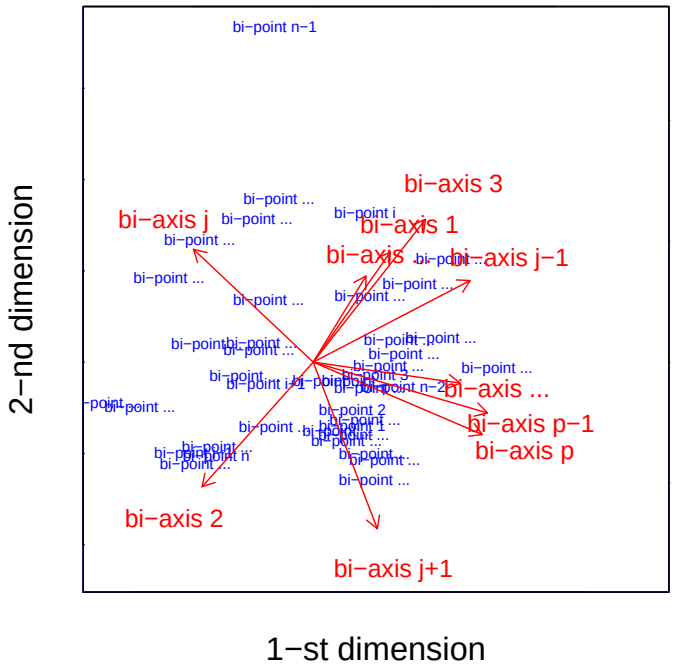

Figure 5: Biplot.

3.2.3 Least Squares approximation

The relationship $X = U\Gamma V$ allows to reconstruct the matrix X exactly by using all the (r associated with non-null eigenvalues) PCs.

However, if the first k PCs take into account a high share of the total variance, an approximate relationship can be written:

$$\underbrace{X}_{n\times p} \cong \underbrace{U}_{n\times k}\ \underbrace{\Gamma}_{k\times k}\ \underbrace{V'}_{k\times p}$$

This representation of X is the best (linear) approximation of rank k according to the **Least Squares (LS) criterion**.

In particular, the case $k = 2$ is of interest, in order to have one representation on the Cartesian plane.

Least squares property

The matrix Y, $n \times p$, of rank k, that best approximates X, by minimizing the sum of squared residuals, i.e.

$$\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - y_{ij})^2 = tr[(X - Y)(X - Y)']$$

is the s.v.d. (of order k) of X:

$$Y = \sum_{h=1}^{k} \gamma_h \mathbf{u}_h \mathbf{v}_h'$$

3.2.4 Infinite declinations of biplot

Consider approximation $Y = U_2 \, \Gamma_2 \, V_2'$, we can decompose Γ_2:

$$Y = U_2 \, \Gamma_2^{\alpha} \, \Gamma_2^{1-\alpha} \, V_2' \qquad (0 \leq \alpha \leq 1)$$

Set

$$G = \underbrace{U_2}_{n \times 2} \underbrace{\Gamma_2^{\alpha}}_{2 \times 2} \qquad H = \underbrace{\Gamma_2^{1-\alpha}}_{2 \times 2} \underbrace{V_2'}_{2 \times p}$$

then

$$X \cong GH \quad \text{with} \quad x_{ij} \cong \mathbf{g}_i' \, \mathbf{h}_j$$

i.e. x_{ij} is approximated by the scalar product of the point-unit $\mathbf{g}_i'$, $i = 1, \ldots, n$, and of the point-variable $\mathbf{h}_j'$, $j = 1, \ldots, p$.

Interpretation of biplot changes according to the value assigned to α in Y decomposition.

Cases of special interest are $\alpha = 0$ and $\alpha = 1$.

3.3 Different approaches to PCs

Derived variables

PCs can be obtained as a set of uncorrelated variables $Z_k = Xv_h$, $h = 1, \ldots, k < (<) p$, linear combinations of the original variables, which explain most of the variability of the original data set.

$Z_1 = Xv_1$ is the projection of data along the direction of largest variability, and thereby has the largest variance among all the normalized projections. v_1 is the eigenvector corresponding to the largest eigenvalue of the sample dispersion matrix of X. Z_2 and v_2 correspond to the second largest eigenvalue.

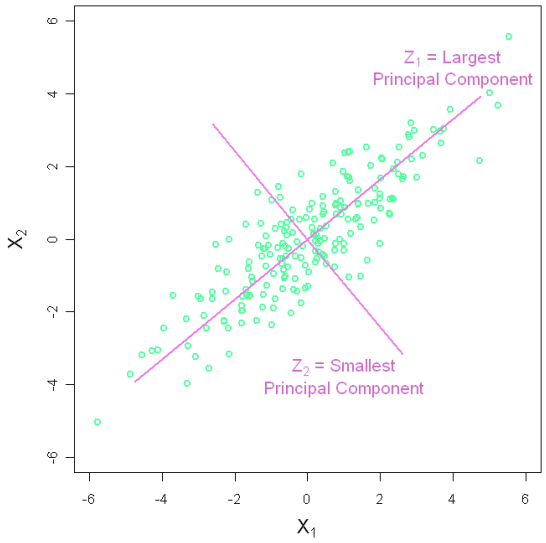

Figure 6: PCs as derived variables.

Approximation minimizing distance

PCs can be obtained approximating the (centered) data matrix X $n \times p$ via the best matrix $\widehat{X}$ of rank $k < (<)p$ according to the criterion of minimum distance. (This approach usually justifies the use of the s.v.d.)

The s.v.d. $\widehat{X}_k = U\Gamma_k V^T$ solves

$$\min_{rank(\widehat{X}_k = k)} ||X - \widehat{X}_k||$$

LS approximation

PCs can be obtained approximating the original data set of n points in $\Re^p$ via the best linear function of dimension $k < (<)p$ according to the least squares criterion.

Finding the linear space $f(\lambda) = \mu + V_k\lambda$ that best approximates the data according to the LS:

$$\min_{\mu, \{\lambda_i\}, V_q} \sum_{i=1}^{n} ||x_i - \mu - V_k\lambda_i||^2$$

Solution: $\mu = \bar{x}$, v_k, $\lambda_i = V_k^T(x_i - \bar{x})$

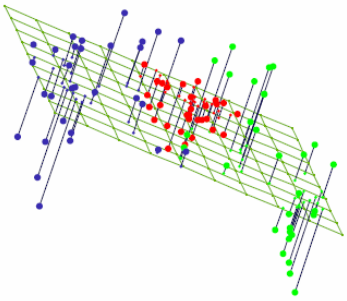

Figure 7: PCs as LS approximatiion.

References

Linting M., Groenen P.J.F., Meulman J.J., and van der Kooij A.J. 2007. "Nonlinear Principal Components Analysis: Introduction and Application." *Psychological Methods* 12 (3). Blackwell Publishing Ltd: 336–58. doi:10.1037%2F1082-989X.12.3.336.

Peres-Neto, Pedro R., Donald A. Jackson, and Keith M. Somers. 2005. "How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited." *Computational Statistics & Data Analysis* 49 (4): 974–97. doi:https://doi.org/10.1016/j.csda.2004.06.015.

Rousson, Valentin, and Theo Gasser. 2004. "Simple Component Analysis." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 53 (4). Blackwell Publishing Ltd: 539–55. doi:10.1111/j.1467-9876.2004.05359.x.