Data Anaysis and Unsupervised Learning

Introduction

Julien Chiquet

UMR MIA Paris-Saclay, AgroParisTech, INRAE

May 25, 2023

https://jchiquet.github.io/

Introduction

Exploratory analysis of (modern) data sets

Assume a table with n individuals described by p features/variables

Questions

Look for **patterns** or **structures** to summarize the data by

- Finding groups of “similar” individuals
- Finding variables important for these data
- Performing visualization

Challenges

- Size data may be large (“big data”: large n large p)
- Dimension data may be high dimensional (more variables than individual or $n \ll p$)
- Redundancy many variables may carry the same information
- Unsupervised we don’t necessary know what we are looking after

An example in genetics: 'snp'

Genetics variant in European population

Description: *medium/large data, high-dimensional*}

500, 000 Genetics variants (SNP – Single Nucleotide Polymorphism) for 3000 individuals (1 meter × 166 meter (height × width)

- SNP : 90 % of human genetic variations
- coded as 0, 1 or 2 (10, 1 or 2 allel different against the population reference)

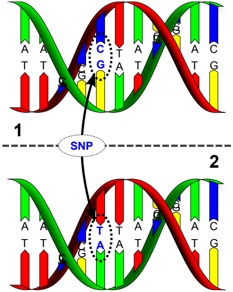


Figure 1: SNP (wikipedia)

Summarize 500,000 variables with 2 features

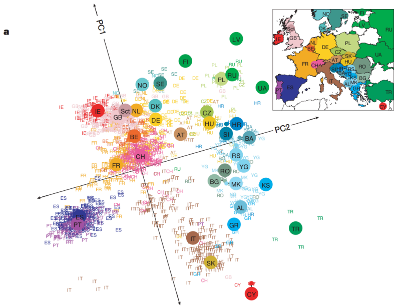


Figure 2: Dimension reduction + labels {source: Nature "Gene Mirror Geography Within Europe", 2008}

In the original messy $3,000 \times 500,000$ table, we may find - an extremely strong structure between individuals (**"clustering"**) - a very simple subspace where it is obvious (**"dimension reduction"**)

Theoretical argument: dimensionality Curse

Theorem (Folks theorem)

If $\mathbf{x}_1, \dots, \mathbf{x}_n$ *in the hypercube of dimension* p *such that their coordinates are i.i.d then*

$$p^{-1/2}\left(\max \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2 - \min \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2\right) = 0 + O\left(\sqrt{\frac{\log n}{p}}\right)$$

$$\frac{\max \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2}{\min \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2} = 1 + O\left(\sqrt{\frac{\log n}{p}}\right).$$

$\rightsquigarrow$ When p is large, all the points are almost equidistant\

Hopefully, the data **are not really leaving in p** dimension (think of the SNP example)

Dimension reduction: general goals

Main objective:

find a **low-dimensional representation** that captures the “essence” of (high-dimensional) data

Application in Machine Learning

Preprocessing, Regularization

- Compression, denoising, anomaly detection
- Reduce overfitting in supervised learning

Application in Statistics/Data analysis}

Better understanding of the data

- descriptive/exploratory methods
- visualization (difficult to plot and interpret $> 3d$!)

Dimension reduction: problem setup

Settings

- Training data : $\mathscr{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathbb{R}^p$, (i.i.d.)
- Space $\mathbb{R}^p$ of possibly high dimension ($n \ll p$)

Dimension Reduction Map

Construct a map Φ from the space $\mathbb{R}^p$ into a space $\mathbb{R}^q$ of smaller dimension:

$$\begin{aligned} \Phi : \quad & \mathbb{R}^p \to \mathbb{R}^q, q \ll p \\ & \mathbf{x} \mapsto \Phi(\mathbf{x}) \end{aligned}$$

How should we design/construct Φ?

Criterion

- Geometrical approach
- Reconstruction error
- Relationship preservation

Form of the map Φ

- **Linear** or non-linear ?
- tradeoff between **interpretability** and versatility ?
- tradeoff between high or **low** computational resource

Principal Component Analysis

Objectives

Individual/Observations

- similarity between observations with respect to all the variables
- Find pattern ($\sim$ partition) between individuals

Variables

- linear relationships between variables
- visualization of the correlation circle
- find synthetic variables

Link between the two

- characterization of the groups of individuals with variables
- specific observations to understand links between variables

Outline

1. Background: high-school algebra
2. Geometric approach to PCA
3. Principal axes and variance maximization
4. Representation and interpretation
5. Additional tools and Complements

Euclidean spaces

(Euclidean) distance between 2 vectors

$$\mathrm{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|.$$

Remark that when $\mathbf{x}$ and $\mathbf{y}$ are orthogonal and non zero, distances between $\mathbf{x}$ and $\mathbf{y}$ and $\mathbf{x}$ and $(-\mathbf{y})$ are the same. Then,

$$(\mathbf{x} - \mathbf{y})^\top(\mathbf{x} - \mathbf{y}) = (\mathbf{x} + \mathbf{y})^\top(\mathbf{x} + \mathbf{y}) \Leftrightarrow \mathbf{x}^\top\mathbf{y} = 0,$$

which motivates the following definition of orthorgonality:

Orthogonality

Two vectors $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$ are orthogonal iff $\mathbf{x}^\top\mathbf{y} = 0$.

Orthogonal Projection

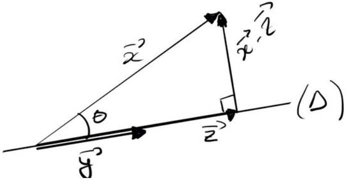
Geometric definition of the dot product

Orthogonal projection

It is the vector $\mathbf{z}$ such that

1. $\mathbf{z} = \lambda \mathbf{y}$
2. $\mathbf{y}$ is orthogonal to $\mathbf{x} - \mathbf{z}$

We find $\lambda = \mathbf{x}^\top \mathbf{y} / \|\mathbf{y}\|^2$



Thanks to Pythagoras theorem,

$$\cos(\theta) = \frac{\|\mathbf{z}\|}{\|\mathbf{x}\|} = \lambda \frac{\|\mathbf{y}\|}{\|\mathbf{x}\|}$$

and then we end with the following geometric definition of the dot product

Dot product: geometric definition

$$\mathbf{x}^\top \mathbf{y} = \cos(\theta) \|\mathbf{x}\| \, \|\mathbf{y}\|$$

Outline

The data matrix

The data set is a $n \times p$ matrix $\mathbf{X} = (x_{ij})$ with values in $\mathbb{R}$:

- each row $\mathbf{x}_i$ represents an individual/observation
- each col $\mathbf{x}^j$ represents a variable/attribute

$$
\mathbf{X} = \begin{array}{c} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_n \end{array}
\begin{array}{c}
\begin{array}{cccccc} \mathbf{x}^1 & \mathbf{x}^2 & \dots & \mathbf{x}^j & \dots & \mathbf{x}^p \end{array} \\
\begin{pmatrix}
x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\
x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np}
\end{pmatrix}
\end{array}
$$

Cloud of observation in $\mathbb{R}^p$

Individuals can be represented in the variable space $\mathbb{R}^p$ as a point cloud

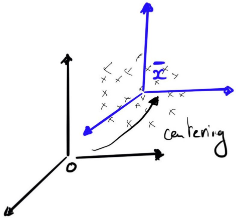


Figure 3: Example in $\mathbb{R}^3$

Center of Inertia

(or barycentrum, or empirical mean)

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{pmatrix} \sum_{i=1}^{n} x_{i1}/n \\ \sum_{i=1}^{n} x_{i2}/n \\ \vdots \\ \sum_{i=1}^{n} x_{ip}/n \end{pmatrix}$$

We center the cloud $\mathbf{X}$ around $\mathbf{x}$ denote this by $\mathbf{X}^c$

$$\mathbf{X}^c = \begin{pmatrix} x_{11}-\bar{x}_1 & \dots & x_{1j}-\bar{x}_j & \dots & x_{1p}-\bar{x}_p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1}-\bar{x}_1 & \dots & x_{ij}-\bar{x}_j & \dots & x_{ip}-\bar{x}_p \end{pmatrix}$$

Inertia and Variance

Total Inertia:

distance of the individuals to the center of the cloud

$$I_T = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)^2 = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\bar{\mathbf{x}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{dist}^2(\mathbf{x}_i,\bar{\mathbf{x}})$$

Proportional to the total variance

Let $\hat{\boldsymbol{\Sigma}}$ be the empirical variance-covariance matrix

$$I_T = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2 = \sum_{j=1}^{p}\frac{1}{n}\|\mathbf{x}^j-\bar{x}_j\|^2 = \sum_{j=1}^{p}\mathbb{V}(\mathbf{x}^j) = \mathrm{trace}\left(\hat{\boldsymbol{\Sigma}}\right)$$

⇝ Good representation has large inertia (much variability)

⇝ Large dispertion $\sim$ Large distances between points

Inertia with respect to an axis

The Inertia of the cloud wrt axe Δ is the sum of the distances between all points and their orthogonal projection on Δ.

$$I_\Delta = \frac{1}{n}\sum_{i=1}^{n} \mathrm{dist}^2(\mathbf{x}_i, \Delta)$$

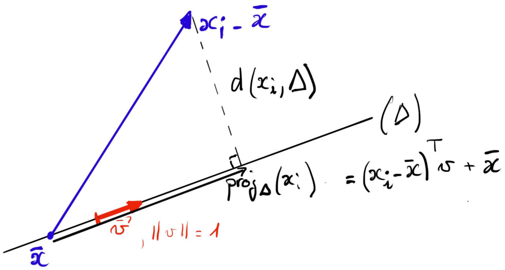
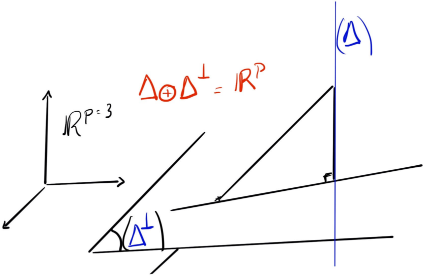


Figure 4: Projection of $\mathbf{x}_i$ onto a line Δ passing through $\bar{\mathbf{x}}$

Decomposition of total Inertia (1)

Let $\Delta^{\perp}$ be the orthogonal subspace of Δ in $\mathbb{R}^p$



Theorem (Huygens)

A consequence of the above (Pythagoras Theorem) is the decomposition of the following total inertia:

$$I_T = I_\Delta + I_{\Delta^\perp}$$

By projecting the cloud $\mathbf{X}$ *onto* Δ*, with loss the inertia measured by* $\Delta^{\perp}$

Decomposition of total Inertia (2)

Consider only subspaces with dimension 1 (that is, lines or axes). We can decompose $\mathbb{R}^p$ as the sum of p othogonal axis.

$$\mathbb{R}^p = \Delta_1 \oplus \Delta_2 \oplus \cdots \oplus \Delta_p$$

⇝ These axes form a new basis for representing the point cloud.

Theorem (Huygens)

$$I_T = I_{\Delta_1} + I_{\Delta_2} + \cdots + I_{\Delta_p}$$

Outline

Finding the best axis (1)

Definition of the problem

- The best axis Δ_1 is the "closest" to the point cloud
- Inertia of Δ_1 measures the distance between the data and Δ_1
- Δ_1 is defined by the director vector $\mathbf{u}_1$, such as $\|\mathbf{u}_1\| = 1$
- $\Delta_1^\perp$ is defined by the normal vector $\mathbf{u}_1$, such as $\|\mathbf{u}_1\| = 1$

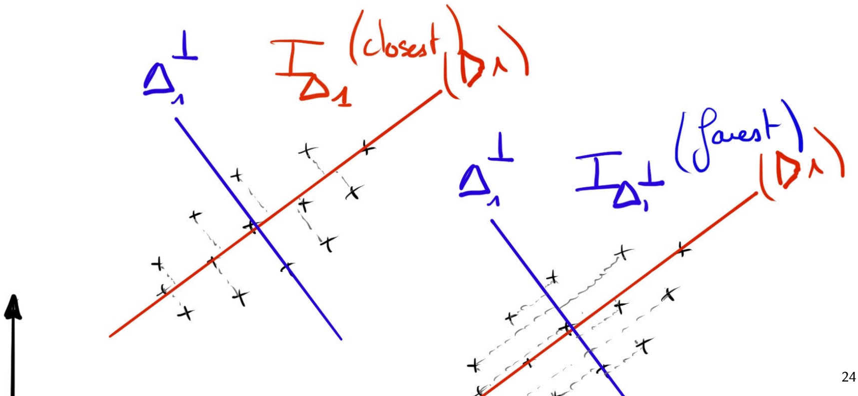
$\rightsquigarrow$ The best axis Δ_1 is the one with the minimal Inertia.

Finding the best axis (2)

Stating the optimization problem

Since $\Delta_1 \oplus \Delta_1^\perp = \mathbb{R}^p$ and $I_T = I_{\Delta_1} + I_{\Delta_1^\perp}$, then

$$\underset{\mathbf{u}\in\mathbb{R}^p\,:\,\|\mathbf{u}\|=1}{\text{minimize}}\ I_{\Delta_1} \Leftrightarrow \underset{\mathbf{u}\in\mathbb{R}^p\,:\,\|\mathbf{u}\|=1}{\text{maximize}}\ I_{\Delta_1^\perp}$$



Finding the best axis (3)

Stating the problem (algebraically)

Find $\mathbf{u}_1; \|\mathbf{u}_1\| = 1$ that maximizes

$$
\begin{aligned}
I_{\Delta_1^\perp} &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{dist}(\mathbf{x}_i, \Delta_1^\perp)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}_1^\top(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{u}_1 \\
&= \mathbf{u}_1^\top \left(\sum_{i=1}^{n} \frac{1}{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\right) \mathbf{u}_1 \\
&= \mathbf{u}_1^\top \hat{\mathbf{\Sigma}} \mathbf{u}_1
\end{aligned}
$$

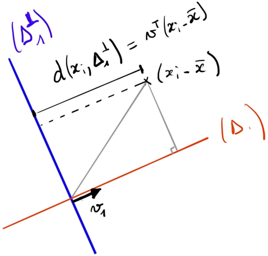


Figure 5: Geometrical insight

Finding the best axis (4)

We solve a simple constraint maximization problem with the method of Lagrange multipliers:

$$\underset{\mathbf{u}_1 : \|\mathbf{u}_1\|=1}{\text{maximize}}\, \mathbf{u}_1^\top \hat{\boldsymbol{\Sigma}} \mathbf{u}_1 \Leftrightarrow \underset{\mathbf{u}_1 \in \mathbb{R}^p, \lambda_1 > 0}{\text{maximize}}\, \mathbf{u}_1^\top \hat{\boldsymbol{\Sigma}} \mathbf{u}_1 - \lambda_1 (\|\mathbf{u}_1\|^2 - 1)$$

By straightforward (vector) differentiation, an using that $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

$$\begin{cases} 2\hat{\boldsymbol{\Sigma}}\mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0 \\ \mathbf{u}_1^\top \mathbf{u}_1 - 1 = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\boldsymbol{\Sigma}}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \\ \mathbf{u}_1^\top \hat{\boldsymbol{\Sigma}} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1^\top \mathbf{u}_1 = \lambda_1 = I_{\Delta_1}^{\perp} \end{cases}$$

- $\mathbf{u}_1$ is the first (normalized) eigen vector of $\hat{\boldsymbol{\Sigma}}$
- λ_1 is the first eigen value of $\hat{\boldsymbol{\Sigma}}$

$⊠$ **Δ_1 is defined by the first eigen vector of $\hat{\boldsymbol{\Sigma}}$**

$⊠$ **Variance "carried" by Δ_1 is equal to the largest eigen value of $\hat{\boldsymbol{\Sigma}}$**

Finding the following axes

Second best axis

Find Δ_2 with dimension 1, director vector $\mathbf{u}_2$ orthogonal to Δ_1 solving

$$\underset{\mathbf{u}_2 \in \mathbb{R}^p}{\text{maximize}}\, I_{\Delta_2^\perp} = \mathbf{u}_2^\top \hat{\boldsymbol{\Sigma}} \mathbf{u}_2, \quad \text{with } \|\mathbf{u}_2\| = 1, \mathbf{u}_1^\top \mathbf{u}_2 = 0.$$

$\rightsquigarrow$ $\mathbf{u}_2$ is the second eigen vector of $\hat{\boldsymbol{\Sigma}}$ with eigen value λ_2

And so on!

PCA is roughly a matrix factorisation problem

$$\hat{\boldsymbol{\Sigma}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top, \quad \mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2, & \dots & \mathbf{u}_p \end{pmatrix}, \quad \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$$

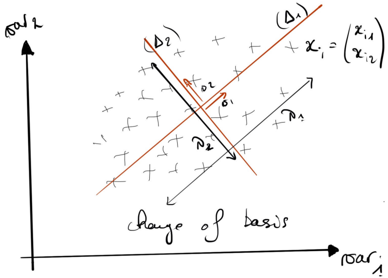
- $\mathbf{U}$ is an orthogonal matrix of normalized eigen vectors.
- $\boldsymbol{\Lambda}$ is diagonal matrix of ordered eigen values.

Interpretation in $\mathbb{R}^p$

$\mathbf{U}$ describes a new orthogonal basis and a rotation of data in this basis

$\rightsquigarrow$ PCA is an appropriate rotation on axes that maximizes the variance

$$\left\{\begin{array}{ccccc} \Delta_1 & \oplus & \dots & \oplus & \Delta_p \\ \mathbf{u}_1 & \perp & \dots & \perp & \mathbf{u}_p \\ \lambda_1 & > & \dots & > & \lambda_p \\ I_{\Delta_1^\perp} & > & \dots & > & I_{\Delta_p^\perp} \end{array}\right.$$



Outline

Contribution of each axis and quality of the representation}

Δ_k is carrying inertia/variance defined by its orthogonal, thus

$$I_T = I_{\Delta_1^\perp} + \cdots + I_{\Delta_p^\perp} = \lambda_1 + \cdots + \lambda_p$$

Relative contribution of axis k

$$\mathrm{contrib}(\Delta_k) = \frac{\lambda_k}{\sum_{k=1}^p \lambda_j} = \frac{\lambda_k}{\mathrm{trace}\left(\hat{\Sigma}\right)} \times 100$$

⇝ Percentage of explained inertia/variance explained

Global quality of the representation on the first k axes

$$\mathrm{contrib}(\Delta_1, \ldots, \Delta_k) = \frac{\lambda_1 + \cdots + \lambda_k}{\mathrm{trace}\left(\hat{\Sigma}\right)} \times 100$$

A few axes may explain a large proportion of the total variance.

⇝ This paves the way for dimension reduction

Contribution of each axis and quality of the representation}

Δ_k is carrying inertia/variance defined by its orthogonal, thus

$$I_T = I_{\Delta_1^\perp} + \cdots + I_{\Delta_p^\perp} = \lambda_1 + \cdots + \lambda_p$$

Relative contribution of axis k

$$\text{contrib}(\Delta_k) = \frac{\lambda_k}{\sum_{k=1}^p \lambda_j} = \frac{\lambda_k}{\text{trace}\left(\hat{\boldsymbol{\Sigma}}\right)} \times 100$$

⇝ Percentage of explained inertia/variance explained

Global quality of the representation on the first k axes

$$\text{contrib}(\Delta_1, \dots, \Delta_k) = \frac{\lambda_1 + \cdots + \lambda_k}{\text{trace}\left(\hat{\boldsymbol{\Sigma}}\right)} \times 100$$

A few axes may explain a large proportion of the total variance.

⇝ This paves the way for dimension reduction

Individuals: representation in the new basis

Projection

The projection of $\mathbf{x}_i$ onto axis Δ_k is $c_{ik}\mathbf{u}_k$, with

$$c_{ik} = \mathbf{u}_k^\top(\mathbf{x}_i - \bar{\mathbf{x}}),$$

the coordinate of i in the basis $\mathbf{u}_k$ (along axis Δ_k).

Coordinates

Coordinates of i in the new basis $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ is thus

$$\mathbf{c}_i = (\mathbf{U}^\top(\mathbf{x}_i - \bar{\mathbf{x}}))^\top = (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{U} = \mathbf{X}_i^c \mathbf{U}, \quad \mathbf{c}_i \in \mathbb{R}^p.$$

- $\mathbf{U}$ are often the called the **loadings**, or **weights**
- $\mathbf{c}_i$ are the **scores** or **coordinates** in the new space for the individuals

Warning: about distances after projection

Close projection doesn't mean close individuals!

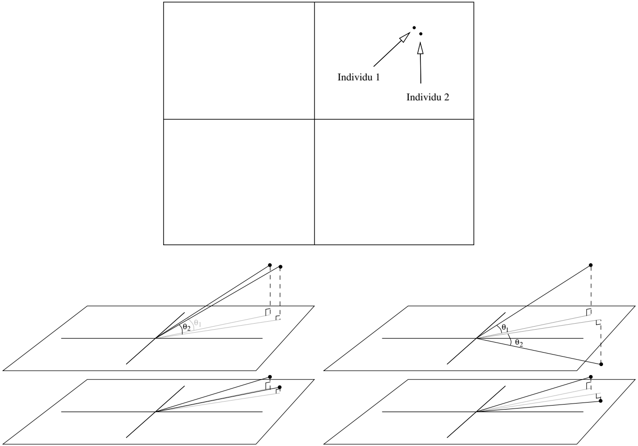


Figure 6: Same projections but different situations (source: E. Matzner)

⇝ Only work when individuals are well represented in the lower space

Individual: representation

Quality

- An individual i is well represented by Δ_k if it is close to this axis.
- In other word, vector $\mathbf{x}_i - \bar{\mathbf{x}}$ and $\mathbf{u}_k$ are close to collinear

Use the cosine of the angle between $\mathbf{x}_i - \bar{\mathbf{x}}$ and $\mathbf{u}_k$ to measure collinearity:

$$\cos^2(\theta_{ik}) = \frac{\left(\mathbf{u}_k^\top(\mathbf{x}_i - \bar{\mathbf{x}})\right)^2}{\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 \cancel{\|\mathbf{u}_k\|^2}}$$

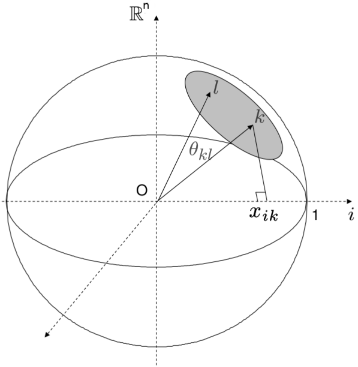
Contribution

- Inertia "explained" by Δ_k is inertia of $\Delta_k^\perp$
- $I_{\Delta_k^\perp} = n^{-1} \sum_{i=1}^n \mathrm{dist}^2(\Delta_k^\perp, \mathbf{x}_i)$

Contribution is the proportion of variance/inertia carried by individual i:

$$n^{-1}\mathrm{dist}^2(\Delta_k^\perp, \mathbf{x}_i) \quad \left(\mathbf{u}_k^\top(\mathbf{x}_i - \bar{\mathbf{x}})\right)^2$$

Cloud of variables



Direct equivalence between geometry and statistics (collinearity ≡ correlation)

$$\cos(\theta_{kl}) = \frac{\langle \mathbf{x}^k, \mathbf{x}^\ell \rangle}{\|\mathbf{x}^k\|\|\mathbf{x}^\ell\|} = \rho(\mathbf{x}^k, \mathbf{x}^\ell)$$

Principal Components

Dual representation

A symmetric reasoning can be made in $\mathbb{R}^n$ for the variables, like with the individuals in $\mathbb{R}^p$.
$\rightsquigarrow$ New axes are linear combinaison of the original variables, which can be seen as **new variables** in the new latent space

Principal component

It is the linear combinasion formed by the orginal variables with weights given by the loadings $\mathbf{u}_k = (u_{k1}, \dots, u_{kj}, \dots, u_{kp})$

$$\mathbf{f}_k = \sum_{j=1}^{p} u_{kj}(\mathbf{x}^j - \bar{x}_j) = \mathbf{X}^c \mathbf{u}_k, \quad \mathbf{f}_k \in \mathbb{R}^n$$

Sometimes called ”factors” in factor analysis, as latent (hidden) variables.

Variable representation in the new space

Connection with original variables

- essential for interpretation
- answer to the question: how to read the axes of the individual map
- use correlation to measure connection to original variable

$$\mathbb{V}(\mathbf{f}_k) = \frac{1}{n}\mathbb{V}(\mathbf{X}^c\mathbf{u}_k) = \mathbf{u}_k^\top \frac{1}{n}(\mathbf{X}^c)^\top \mathbf{X}^c\mathbf{u}_k = \mathbf{u}_k^\top \hat{\mathbf{\Sigma}}\mathbf{u}_k = \lambda_k$$

$$\mathrm{cov}(\mathbf{f}_k, (\mathbf{x}^j - \bar{x}_j)) = \mathbf{u}_k \top \mathbf{X}^{c\top}\mathbf{X}^c e_j = \mathbf{u}_k \lambda_k e_j = \lambda_k \mathbf{u}_{kj}$$

$$\mathrm{cor}(\mathbf{f}_k, (\mathbf{x}^j - \bar{x}_j)) = \sqrt{\frac{\lambda_k}{\mathbb{V}(\mathbf{x}^j)}}\mathbf{u}_{kj}$$

Warning: about angle after projection

Close projection doesn't mean close variable!

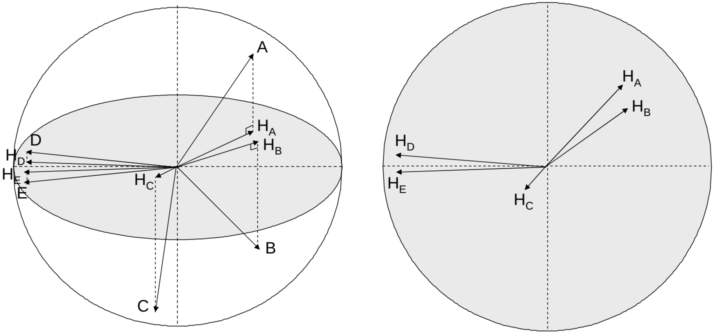


Figure 7: Same angle but different situations {(source: J. Josse)}

⇝ Only work when variables are well represented in the latent space

Variable representation

Quality

- An variable j is well represented by Δ_k if its projection is close to $\mathbf{f}_k$.
- High collinearity means high absolute correlation and high cosine.
- use cosine to the square of the angle between the original and new variables.

$\rightsquigarrow$ The projection of j must be close to the boundary of the correlation circle

Contribution

Similarly to individuals, we can measure the contribution of the original variables to the construction of the new ones.

Outline

Unifying view of variables and individuals

Principal components

The full matrix of principal component connects individual coordinates to latent factors:

$$\mathrm{PC} = \mathbf{X}^c\mathbf{U} = \begin{pmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \dots & \mathbf{f}_p \end{pmatrix} = \begin{pmatrix} \mathbf{c}_1^\top \\ \mathbf{c}_2^\top \\ \dots \\ \mathbf{c}_n^\top \end{pmatrix}$$

- new variables (latent factor) are seen column-wise
- new coordinates are seen row-wise

⇝ Everything can be interpreted on a single plot, called the biplot

Reconstruction formula

Recall that $\mathbf{F} = (\mathbf{f}_1, \dots, \mathbf{f}_p)$ is the matrix of Principal components. Then,

- $\mathbf{f}_k = \mathbf{X}^c \mathbf{u}_k$ for projection on axis k
- $\mathbf{F} = \mathbf{X}^c \mathbf{U}$ for all axis.

Using orthogonality of $\mathbf{U}$, we get back the original data as follows, without loss ($\mathbf{U}^T$ performs the inverse rotation of $\mathbf{U}$):

$$\mathbf{X}^c = \mathbf{F}\mathbf{U}^\top$$

We obtain an approximation $\tilde{\mathbf{X}}^c$ (compression) of the data $\mathbf{X}^c$ by considering a subset $\mathcal{S}$ of PC, typically $\mathcal{S} = 1, \dots, q$ with $q \ll p$.

$$\tilde{\mathbf{X}}^c = \mathbf{F}_{\mathcal{S}} \mathbf{U}_{\mathcal{S}}^\top = \mathbf{X}^c \mathbf{U}_{\mathcal{S}} \mathbf{U}_{\mathcal{S}}^\top$$

$\rightsquigarrow$ This is a rank-q approximation of $\mathbf{X}$ (information captured by the first q axes).