Principal Component Analysis

From a course by Anne M. Ruiz

Camille Mondon

January 5, 2026

The principal component analysis (PCA) is a method of exploratory multivariate data analysis which belongs to the family of **factorial methods** (Abdi and Williams 2010).

1 Nature of the data

The data are presented in a table $\mathbf{X} = (X_i^j)_{1\leq i\leq n}^{1\leq j\leq p}$ where rows $(X_1|\cdots|X_n)^\top$ are considered as individuals and columns $(X^1|\cdots|X^p)$ as quantitative variables. The size of the table is $n \times p$ where n is the number of individuals and p is the number of quantitative variables.

2 Objectives of PCA

PCA has two principal objectives:

- *Summarize* the table $\mathbf{X}$ in a **small number** d of new variables **uncorrelated** between them and which will **keep the maximum of the information** contained in the p initial variables.

  Intuitively, the new variables are obtained by "mixing together" the initial variables which are well correlated between each others.

  The number d of these new variables is *smaller* if the correlations between the p initial variables are *large*.

  As a by-product, PCA leads to a visualization of the correlations between the initial variables.

- *Interpret* the table $\mathbf{X}$ using the new variables and graphics such as scatterplots.

PCA allows in particular to **detect outliers** or to **find groups of individuals** having the same behavior with respect to the considered variables.

PCA is **unsupervised**: it ignores any response variable and only looks at the covariances or correlations among predictors.

3 Principles of PCA

3.1 What 'information' means

In order to achieve the aim of *"keeping the maximum of information contained in a set of data"*, we need a mathematically defined concept representing that information.

In this course, we consider (see Note 1 for a justification) that the information is based on the variability of the data and **is measured using the variance**:

- The 'information' brought by a quantitative variable is its variance.
- The 'information' brought by a random vector composed of p variables is the sum of the variances its variables.

In order to avoid confusions with other notions of information, from now on, we will talk about **inertia** instead of information.

Definition 3.1 (Inertia)**.** The inertia of a data table $\mathbf{X} = (X^1|\cdots|X^p) \in \mathbb{R}^{n\times p}$ is:

$$I_{\mathbf{X}} = \sum_{j=1}^{p} \mathrm{Var}(X^j).$$

The inertia is often called **total variance**.

Exercise 3.1. Is it the variance of the sum of the variables of **X**? What link with the sample covariance matrix of **X**?

Note 1: Inertia and dimension reduction

Our objective of *summarizing* the table **X** with less variables can be thought of as a problem of **dimension reduction**. Let us restrict, for ease of intepretation, to the **affine** dimension reduction problem, whose objective is to find, for a reduced dimension $d \leq p$:

- an affine **compression** transformation $\varphi : \mathbb{R}^p \to \mathbb{R}^d$
- an affine **decompression** transformation $\psi : \mathbb{R}^d \to \mathbb{R}^p$

that minimize the expected squared reconstruction error:

$$\min_{\varphi,\psi} \mathbb{E}\|X - \psi \circ \varphi(X)\|^2$$

where $X \in \mathbb{R}^p$ denotes a random variable following the empirical measure of the sample $(X_1, \ldots, X_n)$ (introduced to avoid writing averages over the sample).
It can be shown (Sangnier 2025, sec. 3.1.1) that this problem boils down to maximizing the inertia (or total variance) of the affine transformed data:

$$\max_{V\in\mathbb{R}^{p\times d}, V^\top V = I_d} \mathrm{tr}(V^\top \mathrm{Cov}[X]\, V) = \max_{V\in\mathbb{R}^{p\times d}, V^\top V = I_d} I_{\mathbf{X}V}$$

3.2 Standardized variables

Whenever the variables in a data table **X** are **not in the same physical unit** (e.g. euros, kilograms, meters, dimensionless), a unit conversion could be considered that would change the inertia of **X**.

That is why in that case we always **scale** the table **X** by standardizing each of its variables

$$x^j = \frac{X^j - \overline{X^j}}{\sigma_{X^j}},$$

so that $\overline{x^j} = 0$ and $\mathrm{Var}(x^j) = 1$. Therefore, standardized variables have the same inertia (equal to 1). The standardized table is denoted by **x**.

Remark

Remark 3.1. This scaling preserves the correlations among variables:

$$\text{cor}(x^k, x^j) = \text{cor}(X^k, X^j).$$

In this course, PCA is performed on standardized variables. It is a particular case of PCA sometimes called **standardized** PCA. It is also called **correlation** PCA (this will make sense later).

3.3 Principal components

3.3.1 Heuristics

We want the **first principal component** c^1 to be a new variable, that is a linear combination of the variables $x^1, x^2, \ldots, x^p$:

$$c^1 = \mathbf{x}v_1 = v_1^1 x^1 + v_1^2 x^2 + \ldots + v_1^p x^p$$

such that the inertia (i.e. the variance) of c^1 is maximal.

Remark

Remark 3.2. As we could always multiply v_1 by a constant greater than 1 to increase the inertia of c^1, we have to ensure that the coefficients of v_1 are normalized. A way of doing so is to make v_1 belong to the sphere in $\mathbb{R}^p$ (i.e. that it represents a direction in the space called the first **principal axis**):

$$\|v_1\|^2 := \sum_{k=1}^{p} (v_1^k)^2 = 1$$

It will make sense later why this makes computations easier than other constraints.

The **second principal component** $c^2 = \mathbf{x}v_2$ should be defined as a new variable *uncorrelated with* c^1, a linear combination of the variables $x^j, 1 \leq j \leq p$ (with $\|v_2\| = 1$) and of maximal variance.

The **third principal component** $c^3 = \mathbf{x}v_3$ should be *uncorrelated with* c^1 *and* c^2, a linear combination of the $x^j, 1 \leq j \leq p$ (with $\|v_3\| = 1$) and of maximal variance.

The p**-th principal component** $c^p = \mathbf{x}v_p$ should be *uncorrelated with* $c^1, c^2, \ldots, c^{p-1}$, a linear combination of the $x^j, 1 \leq j \leq p$ (with $\|v_p\| = 1$) and of maximal variance.

If we successfully define such p uncorrelated components, we group then in a table denoted $\mathbf{C} = (c^1|c^2|\cdots|c^p)$. In the end, we should have

$$\text{Var}(c^1) \geq \text{Var}(c^2) \geq \cdots \geq \text{Var}(c^p).$$

Indeed, each new principal component should maximize the variance among a increasingly restricted choice of variables (the ones that are uncorrelated with are the previous principals components), so the variance of the principal components decreases.

Note 2: Uncorrelatedness vs. orthogonality

Here we explain how the **iterative variance maximization** problem defining principal components relates to the **global inertia maximization** problem stated in Note 1.
Let $V = (v_1 \mid \cdots \mid v_p)$ be the matrix whose columns are the directions defining the principal components, so that

$$\mathbf{C} = \mathbf{x}V.$$

At first sight, it is **not obvious** that the matrix V should be orthogonal, since orthogonality was not explicitly

imposed in the definition of the principal components.
Indeed, the first principal direction is defined as

$$v_1 = \arg\max_{\|v\|=1} \mathrm{Var}(\mathbf{x}v) = \arg\max_{\|v\|=1} v^\top \mathbf{x}^\top \mathbf{x}\, v.$$

By first-order optimality on the unit sphere, the gradient of

$$f(v) = v^\top \mathbf{x}^\top \mathbf{x}\, v$$

at v_1 must be orthogonal to the tangent space of the sphere at v_1. This implies that

$$\mathbf{x}^\top \mathbf{x}\, v_1 = \lambda_1 v_1,$$

so v_1 is an eigenvector of $\mathbf{x}^\top \mathbf{x}$.
The second principal direction v_2 is obtained by maximizing the variance under the additional constraint that $\mathbf{x}v_2$ is uncorrelated with $\mathbf{x}v_1$. Using the relation

$$\mathrm{Cov}(\mathbf{x}v_1, \mathbf{x}v) = v_1^\top \mathbf{x}^\top \mathbf{x}\, v = \lambda_1 v_1^\top v,$$

this uncorrelatedness condition is equivalent to $v_2 \perp v_1$.
Repeating the same gradient argument on the unit sphere intersected with the orthogonal complement of $\mathrm{span}(v_1)$ shows that

$$\mathbf{x}^\top \mathbf{x}\, v_2 = \lambda_2 v_2,$$

with $v_2 \perp v_1$.
By induction, the iterative construction yields an orthonormal family $(v_1, \ldots, v_p)$ of eigenvectors of $\mathbf{x}^\top \mathbf{x}$. As a consequence,

$$V^\top V = I_p.$$

which is the orthogonality condition imposed on the matrix V from Note 1 (when $d = p$). This reasoning is similar to that of Jolliffe (2002) but without relying on the Lagrangian method.

3.3.2 Definition

Now we should prove using the covariance matrix of $\mathbf{x}$ (i.e. the correlation matrix of $\mathbf{X}$) that principal components such as in Section 3.3.1 can be properly defined.

The correlation matrix

$$R = \frac{1}{n-1} \mathbf{x}^\top \mathbf{x}$$

is symmetric and non-negative, because if

$$u \in \mathbb{R}^p, u^\top R u = \mathrm{Var}(\mathbf{x}u) \geq 0$$

Applying the **spectral theorem** to R, we obtain an orthonormal basis $(v_1, \ldots, v_p)$ of $\mathbb{R}^p$ of **eigenvectors** of R and a decreasing sequence $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ of **eigenvalues** of R such that

$$R\, v_j = \lambda_j\, v_j.$$

Definition 3.2 (Principal components & scores)**.** If $1 \leq j \leq p$, we define the j-th principal component of $\mathbf{x}$ as:

$$c^j = \mathbf{x} v_j$$

If $1 \leq i \leq n$, then c_i^j is called the **score** of individual i on the j-th principal component.

Definition 3.3 (Principal axes & loadings)**.** If $1 \leq j \leq p$, v_j is called the j**-th principal axis**. If $1 \leq k \leq p$, then v_j^k is called the **loading** of variable x^k on principal component c^j.

Note 3: Connection with Courant–Fisher min-max theorem

The iterative variance-maximization procedure for defining principal components can be seen as a consequence of the **Courant–Fisher min-max theorem** applied to the symmetric matrix $R = \frac{1}{n-1}\mathbf{x}^\top\mathbf{x}$. It implies that the j-th largest eigenvalue λ_j satisfies

$$\lambda_j = \max_{\substack{v \in \mathbb{R}^p \setminus \{0\} \\ v \perp v_1, \ldots, v_{j-1}}} \frac{v^\top R v}{v^\top v},$$

where $v_1, \ldots, v_{j-1}$ are the eigenvectors associated with the $j-1$ largest eigenvalues.
Thus, the j-th principal axis is the **direction of maximal variance orthogonal to the previous principal axes**.

3.3.3 Properties

Proposition 3.1.

- *The principal components are centered:* $\overline{c^j} = 0$ *for* $1 \leq j \leq p$.
- *The eigenvalues of* R *are the variance explained by the principal components:* $\mathrm{Var}(c^j) = \lambda_j$ *for* $1 \leq j \leq p$.
- *The principal components are uncorrelated:* $r(c^j, c^k) = 0$ *for* $1 \leq j \neq k \leq p$.
- *Principal Component Analysis preserves the inertia:*

$$I_{\mathbf{C}} = \sum_{j=1}^{p} \lambda_j = I_{\mathbf{x}} = p.$$

Whereas each column of x brings the same inertia (equal to 1), the columns of table C bring an inertia which decreases with their index.

The first objective of PCA is then reached: table x can be summarized by a table containing less columns **if the last principal components have a little inertia**.

Remark

Remark 3.3. PCA could be introduced at the population level instead of the sample level, but this is outside of the scope of this course.

Note 4: Geometric approach to PCA

PCA can be introduced by other criteria than a statistical one (dimension reduction by minimizing the expected reconstruction error, or variance maximization). The **geometric approach** is notably often presented (only in the **sample** framework, not in the population framework, since it requires working with the data table).
The strength of this geometric approach is that it works without any statistical assumption on the data set, so interpretation is possible **even in high-dimensional contexts**.
The idea is that instead of relying on the correlation matrix $\frac{1}{n-1}\mathbf{x}^\top\mathbf{x}$, we can multiply the data table $\mathbf{x}$ the other way around, to obtain the Gram matrix:

$$\mathbf{x}\mathbf{x}^\top = (x_i^\top x_\ell)_{1 \leq i \leq n}^{1 \leq \ell \leq n}$$

which encodes the geometric structure of the data set $\mathbf{x}$ (distances and angles).

We can show that the Gram matrix possesses the same nonzero eigenvalues than the correlation matrix, which is the core result of **singular value decomposition (SVD)**. This decomposition allows to work directly with the data matrix, and is the mathematical foundation behind the **implementation** of PCA in practical algorithms. It unifies the geometric and statistical approaches (recovering by multiplying the data table by its transpose one way or the other):

$$\mathbf{x} = U\Sigma V^\top$$

where

- $r \leq \min(n, p)$ is the rank of $\mathbf{x}$;
- $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{p\times r}$ have orthonormal columns:

$$U^\top U = I_r \text{ and } V^\top V = I_r.$$

- $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ is diagonal.
- The singular values σ_j are the square roots of the eigenvalues of R.

Then $\mathbf{C} = U\Sigma = \mathbf{x}V$ are the principal components and V is the loading matrix.
The **Eckart-Young-Mirsky low-rank approximation theorem** states that for any $1 \leq d \leq r$, the truncated SVD matrix $\mathbf{X}_{(d)} = U_{(d)}\Sigma_{(d)}V_{(d)}^\top$ obtained by keeping only the d first columns of U, Σ and V is optimal in the following sense:

$$\|\mathbf{X} - \mathbf{X}_{(d)}\|_F = \min_{\text{rank}(M)\leq d} \|\mathbf{X} - M\|_F$$

where

$$\|M\|_F^2 = \text{tr}(M^\top M) = \sum_{i=1}^{n}\sum_{j=1}^{p} M_i^{j\,2} = \sum_{i=1}^{n} \|M_i\|^2 = \sum_{j=1}^{p} \|M^j\|^2.$$

In particular, this theorem justifies why the **biplot** ($d = 2$) is based on the best rank-2 approximation of the data set.

4 Component selection criteria

4.1 Percentage of explained variance criterion

When performing PCA, one must choose a number d of principal components that is sufficient to summarize the total inertia of the data while losing as little inertia as possible. The **total inertia** of standardized variables is $I_\mathbf{x} = p$, and the inertia of the j-th component is

$$I_{c^j} = \text{Var}(c^j) = \lambda_j.$$

The proportion of inertia explained by the first principal component c^1 is

$$i_1 = \frac{\lambda_1}{p},$$

by the first two components c^1, c^2 is

$$i_2 = \frac{\lambda_1 + \lambda_2}{p},$$

and in general, by the first d components $c^1, \ldots, c^d$ is

$$i_d = \frac{\sum_{j=1}^{d} \lambda_j}{p}.$$

Finally, the proportion of inertia explained by all p principal components is

$$i_p = \frac{\sum_{j=1}^{p} \lambda_j}{p} = \frac{p}{p} = 100\%.$$

Criterion: Choose d so that the proportion of explained variance is sufficiently large, typically at least 80 %.

4.2 Kaiser criterion

The initial variables have a variance equal to 1 (standardized).

Criterion: Select the principal components whose variance is **greater than 1**, because they have more inertia than the initial variables. This means than:

$$d = |\{j, \lambda_j > 1\}|$$

4.3 Scree test criterion

The differences between eigenvalues are examined :

$$\lambda_1 - \lambda_2, \lambda_2 - \lambda_3, \ldots$$

In general, these differences decrease.

Criterion : Retain the d principal components so that the difference $\lambda_d - \lambda_{d+1}$ is large and the differences $\lambda_j - \lambda_{j+1}$ for $d+1 \leq j \leq p-1$ are small (**elbow method**).

5 Interpreting PCA

5.1 Interpreting the components

5.1.1 Presentation of the problem

Let us assume that, after using one of the previous criteria, d (small) principal components (or d dimensions or d factors) have been selected.

One of the difficulties of PCA (and of factorial analysis in general) is the interpretation of the principal components.

PCA leads to a reduction of the variables number(from p to d) but if the signification of the initial variables is known, it is not the case for the principal components.

5.1.2 Interpretation of the loadings

For all $1 \leq j \leq d$:

$$c^j = \mathbf{x} v_j = \sum_{k=1}^{p} v_j^k x^k$$

The composition of the each principal component c^j is known and the important x^k variables are associated to large loadings v_j^k (since they have the same variance).

But this method is rarely used (the size of the coefficients is assessable with difficulty).

5.1.3 Interpretation of the correlations between the components and the initial variables

Proposition 5.1 (Correlations between variables and components)**.** *For any* $1 \leq j,k \leq p$*:*

$$\text{cor}(c^j, x^k) = \sqrt{\lambda_j} v_j^k$$

Moreover for a fixed variable x^k*, the squared correlations with all the principal components sum up to 1:*

$$\sum_{j=1}^{p} \text{cor}(c^j, x^k)^2 = 1$$

This means that we could represent the initial variables $x^1, \ldots, x^p$ as points on the unit hypersphere of the p-dimensional Euclidean space of random variables generated by the principal components. The coordinates of the initial variables are their correlations with the principal components.

But in general, it is preferable to represent the correlations by considering the selected principal components 2 by 2 and to interpret them graphically (this is possible because we selected a small number d of components so there is only $\frac{d(d-1)}{2}$ possible correlation plots.)

In that case, the points are not located on the circle but rather in the disk centered at the origin and of rayon 1. The bounding circle (called the **circle of correlations**) is often drawn because it is helpful for the interpretation.

In a given principal plane, the variables that are too close to the origin (i.e. far from the circle) are not interesting because because they are weakly correlated with the selected principal components and therefore are not useful for their interpretation.

But if we exclude these variables, we can interpret each principal component according to strong correlations (positive and negative).

Remark

Remark 5.1. The **quality of representation** of a variable x^k by component c^j can be measured by $\text{cor}(x^k, c^j)^2$. In a principal plane, the quality of representation of a variable will be its squared distance to the origin (or norm).

Note 5: Space of variables

Note

Definition 5.1 (The L^2 Hilbert space of random variables)**.** Let $(\Omega, \mathscr{B}, \mathbb{P})$ be a probability space. The space

$$L^2(\Omega, \mathbb{P}) = \{X : \Omega \to \mathbb{R} \mid \mathbb{E}[X^2] < \infty\}$$

(where we say that $X = Y$ when $P(X = Y) = 1$) is a Hilbert space equipped with the inner product

$$\langle X, Y \rangle = \mathbb{E}[XY].$$

In this framework:

- the **expectation** $\mathbb{E}[X]$ is the inner product with the constant variable 1,
- the **variance** is the squared norm of the centered variable,

$$\text{Var}(X) = \|X - \mathbb{E}[X]\|^2,$$

- the **standard deviation** is the norm of the centered variable,

$$\sigma_X = \|X - \mathbb{E}[X]\|,$$

- the **covariance** is the inner product of centered variables,

$$\mathrm{Cov}(X,Y) = \langle X - \mathbb{E}[X],\, Y - \mathbb{E}[Y] \rangle,$$

- the **correlation** is the cosine of the angle between centered variables,

$$\mathrm{cor}(X,Y) = \frac{\langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle}{\|X - \mathbb{E}[X]\| \, \|Y - \mathbb{E}[Y]\|}.$$

  Thus, **orthogonality** transcribes **uncorrelatedness**.

Centering a variable corresponds to its **orthogonal projection** onto the closed subspace

$$\{X \in L^2(\Omega, \mathbb{P}) | \mathbb{E}[X] = 0\},$$

which is the kernel of the expectation operator.
Scaling a variable to unit variance corresponds to normalizing it to have unit norm in this space, so that it belongs to the **unit sphere**.
That is where our standardized variables $x^1, \ldots, x^p$ are located (since we work on sample PCA, we assume that x^k follows the empirical measure on the k-th column of $\mathbf{x}$). More precisely they generate a **finite-dimensional subspace** (of dimension less than p)

$$\mathscr{X} = \mathrm{span}(x^1, \ldots, x^p) = \{\mathbf{x}v, v \in \mathbb{R}^p\},$$

when considering all the possible linear combinations of the initial variables.
On this space, the inner product is induced by the empirical covariance:

$$\langle x^k, x^\ell \rangle = \mathrm{Cov}(x^k, x^\ell).$$

Thus, the correlation matrix is simply the **Gram matrix** of the variables in this Euclidean space.
PCA consists in constructing an **orthogonal basis** (not orthonormal because the PCs have not a unit standard deviation)

$$(c^1, \ldots, c^p)$$

of $\mathscr{X}$ that diagonalizes the covariance operator and orders directions by decreasing variance (inertia).
The principal components are new orthogonal vectors in the space of variables, and the loading vectors $(v_1, \ldots, v_p)$ describe the change of basis from the original variables to this orthonormal basis.
Because the initial variables have unit norm and the principal components are orthogonal (they are not scaled), the **coordinates of the initial variables in the basis of scaled principal components are the correlations with the non-scaled principal components**. This proves that the sum of squared correlations for a given variable is just its variance, which is 1.
In parallel, PCA induces an orthonormal basis in the **space of individuals** through the dual geometry of the data matrix. This duality explains why PCA simultaneously organizes variables (via correlations) and individuals (via distances).

5.2 Interpreting the individuals

5.2.1 Graph of the individuals

After selecting d principal components, each individual i can be represented by its **scores**

$$(c_i^1, \ldots, c_i^d),$$

which are the coordinates of the projection of x_i onto the subspace spanned by the principal axes $v_1, \ldots, v_d$.

In order to interpret the data set, individuals are represented in the space generated by the selected principal components, usually taken two by two. When two components (c^j, c^ℓ) are considered, the corresponding representation is called a **principal plane**.

These graphs are interpreted as standard scatter plots, taking into account the interpretation of the principal components. However, as for variables, not all individuals are equally well represented in a given plane, and individuals that are **poorly represented** should not be over-interpreted.

Remark (Projecting shrinkens the distances)

Remark 5.2 (Projecting shrinkens the distances). PCA can be understood as an orthogonal projection of the data cloud in $\mathbb{R}^p$ onto a subspace of dimension d, chosen so as to preserve as much inertia as possible. Since any projection shortens distances, the interpretation of the graphs of individuals is reliable only if the projection preserves a large proportion of their distance to the origin.

5.2.2 Measure of the quality of representation of the individuals

To quantify how well an individual is represented in a reduced space, we compare its distance to the origin before and after projection.

Initially, an individual $x_i = (x_i^1, \dots, x_i^p)$ has squared distance to the origin

$$\|x_i\|^2 = \sum_{j=1}^{p} (x_i^j)^2 = \sum_{j=1}^{p} (c_i^j)^2,$$

where the second equality follows from the fact that the principal axes form an orthonormal basis of $\mathbb{R}^p$.

After projection onto the subspace spanned by the first d principal axes, the squared distance becomes

$$\|P_d x_i\|^2 = \sum_{j=1}^{d} (c_i^j)^2.$$

Definition 5.2. The **quality of representation** of individual i in the d-dimensional principal subspace is therefore measured by the ratio

$$\frac{\sum_{j=1}^{d} (c_i^j)^2}{\sum_{j=1}^{p} (c_i^j)^2},$$

which represents the proportion of the individual's inertia preserved by the projection.
In particular, the quality of representation of individual i on a single principal component c^j is given by

$$\frac{(c_i^j)^2}{\sum_{l=1}^{p} (c_i^l)^2} = \cos^2 \theta_i^j,$$

where θ_i^j is the angle between the vector x_i and the principal axis v_j.
This quantity is called the **squared cosine** and measures how strongly the individual is aligned with the principal axis associated with component c^j.

Remark

Remark 5.3. The squared cosines sum up to 1 if we keep all the principal components:

$$\sum_{j=1}^{p} \cos^2 \theta_i^j = 1.$$

So in practice, individuals for which

$$\sum_{j=1}^{d} \cos^2 \theta_i^j$$

is small are poorly represented in the principal planes and should be interpreted with caution.

5.2.3 Contributions of individuals for a component

In the previous section, we focused on one individual that we explained from the selected principal components.

Now we consider a given principal component c^j and we want to understand how each individual contributes to c^j.

Definition 5.3. We define the **contribution** of an individual x_i to a principal component c^j as its squared score normalized by the variance of the component:

$$\text{ctr}_i^j = \frac{(c_i^j)^2}{\lambda_j}.$$

Remember that the squared scores sum up to the variance of the component:

$$\sum_{i=1}^{n} (c_i^j)^2 = \lambda_j$$

so the sum of the contributions of all individuals is 1:

$$\sum_{i=1}^{n} \text{ctr}_i^j = 1.$$

5.3 The biplot

Recall that the score matrix is

$$\mathbf{C} = \mathbf{x}V,$$

so since V is orthogonal:

$$\mathbf{x} = \mathbf{C}V^\top,$$

which is sometimes known as the **reconstruction formula**, because it recovers the original data table from the scores and loadings.

The **biplot** is based on a approximated reconstruction enabling visual representation in lower dimension, using only the d selected principal components.

Let $\mathbf{C}_{(d)} = (c^1, \dots, c^d) \in \mathbb{R}^{n \times d}$ be the scores of the n individuals on the selected d principal components, and $V_{(d)} = (v_1, \dots, v_d) \in \mathbb{R}^{p \times d}$ be the corresponding loading vectors (principal axes). then

$$\mathbf{x} \approx \mathbf{C}_{(d)} V_{(d)}^\top.$$

In order to extract a maximum of properties of the biplot representation, we define **rescaled score matrix** representing the **individuals**:

$$G = \mathbf{C}_{(d)} \Lambda_{(d)}^{-1/2} \in \mathbb{R}^{n \times d},$$

and the **rescaled loading matrix** representing the **variables**:

$$H = V_{(d)} \Lambda_{(d)}^{1/2} \in \mathbb{R}^{p \times d},$$

where $\Lambda_{(d)} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ contains the variances of the first d principal components.

The **rescaled reconstruction formula** is now:

$$\mathbf{x} \approx GH^\top.$$

Then the biplot is constructed by representing:

- **Individuals** as points given by the rows $g_1, \ldots, g_n$ of G,
- **Variables** as vectors given by the rows $h_1, \ldots, h_p$ of H.

This graphical representation has the following properties (that hold exactly if $d = p$ or as approximation is $d < p$):

1. The **cosine of the angle between any two vectors representing variables** equals the **correlation coefficient** between the corresponding variables:
$$\frac{h_k^\top h_\ell}{\|h_k\|\|h_\ell\|} = \text{cor}(x^k, x^\ell),$$
because $HH^\top = (n-1)\mathbf{x}^\top\mathbf{x} = (n-1)R$.
2. The **cosine of the angle between a variable vector h_k and a principal axis e_j** equals the correlation between that variable and the principal component:
$$\frac{h_k^j}{\|h_k\|} = \text{cor}(x^k, c^j) = \sqrt{\lambda_j}\, v_j^k,$$
because $\|h_k\| = 1$.
3. The **inner product between the individual g_i and variable h_j** recovers the value of the observation x_i^j:
$$g_i^\top h_j = x_i^j$$
4. The **Euclidean distance between the individuals g_i and $g_{i'}$** in G is proportional to the Mahalanobis distance between them in the original variable space.

These properties justify why the biplot is a faithful low-dimensional representation: it simultaneously encodes correlations between variables, contributions of variables to components, and the positions of individuals in the principal component space.

6 Additional concepts and extensions

6.1 Size factor

When all the initial variables are positively correlated between them, the first principal component defines a size factor.

A symmetric matrix (in our case the correlation matrix) having all its terms positive possesses a first eigenvector whose coordinates have the same sign.

If they are chosen positive, the first principal component is then positively correlated with all the variables.

6.2 Rotation methods

One of the difficulties of PCA is the **interpretation** of the axes.

But when there are numerous variables with average correlations, the interpretation is difficult.

The role of rotation methods is to return the more clear-cut correlations, making the axes revolve which involves an easier reading.

The **VARIMAX rotation** makes the axes turn preserving their orthogonality but the first factor is not any more the axis of larger variance.

The VARIMAX rotation aims of maximizing the variance of the correlations in each column of the table of the correlations between the principal components and the initial variables.

6.3 Kernel PCA

The geometric formulation of PCA based on the **Gram matrix**

$$\mathbf{x}\mathbf{x}^\top = (x_i^\top x_\ell)_{1\le i,\ell\le n}$$

suggests a crucial observation: **PCA only depends on inner products between observations**.

This remark opens the door to a nonlinear extension of PCA.

6.3.1 Inner products and feature maps

Let $\mathscr{H}$ be a (possibly infinite-dimensional) Hilbert space and let

$$\Phi : \mathbb{R}^p \longrightarrow \mathscr{H}$$

be a nonlinear **feature map**. Instead of working with the original data points $x_i \in \mathbb{R}^p$, we consider their images

$$\Phi(x_i) \in \mathscr{H}.$$

Applying PCA in $\mathscr{H}$ would require computing the Gram matrix

$$G_{i\ell} = \langle \Phi(x_i), \Phi(x_\ell) \rangle_{\mathscr{H}}.$$

However, $\mathscr{H}$ may be very high-dimensional or even infinite-dimensional, making this approach infeasible directly.

6.3.2 The kernel trick

A **kernel function** is a symmetric function

$$k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$$

such that there exists a feature map Φ with

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathscr{H}}.$$

This allows us to compute inner products in $\mathscr{H}$ **without explicitly computing** Φ. Typical examples include:

- Linear kernel: $k(x, y) = x^\top y$
- Polynomial kernel: $k(x, y) = (x^\top y + c)^d$
- Gaussian (RBF) kernel:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

6.3.3 Kernel PCA

Kernel PCA consists in performing PCA on the transformed data $\Phi(x_1), \ldots, \Phi(x_n)$ using only the kernel matrix

$$K = (k(x_i, x_\ell))_{1 \leq i, \ell \leq n}.$$

After centering the kernel matrix, one computes its eigenvalues and eigenvectors. The principal components are then expressed as nonlinear functions of the original data.

Kernel PCA thus:

- generalizes PCA to **nonlinear structures**,
- preserves the **geometric interpretation** of PCA,
- relies entirely on linear algebra in the sample space.

7 Conclusion

PCA is an **unsupervised** statistical method which can be applied:

- to a $n \times p$ data table $\mathbf{x}$
- for p quantitative variables
- $p > 3$
- when some variables are strongly correlated.

to obtain an interpretable compressed representation of the data set.

> Remark
>
> *Remark* 7.1. If $R = I_p$, then PCA is not useful.

Abdi, Hervé, and Lynne J. Williams. 2010. "Principal Component Analysis." *WIREs Computational Statistics* 2 (4): 433–59. https://doi.org/10.1002/wics.101.

Jolliffe, I. T. 2002. *Principal Component Analysis*. Springer Series in Statistics. New York: Springer-Verlag. https://doi.org/10.1007/b98835.

Sangnier, Maxime. 2025. "Introduction to Machine Learning." Sorbonne Université - MS2A. https://perso.lpsm.paris/~msangnier/mlM2.html.