Principal Component Analysis and Singular Value Decomposition

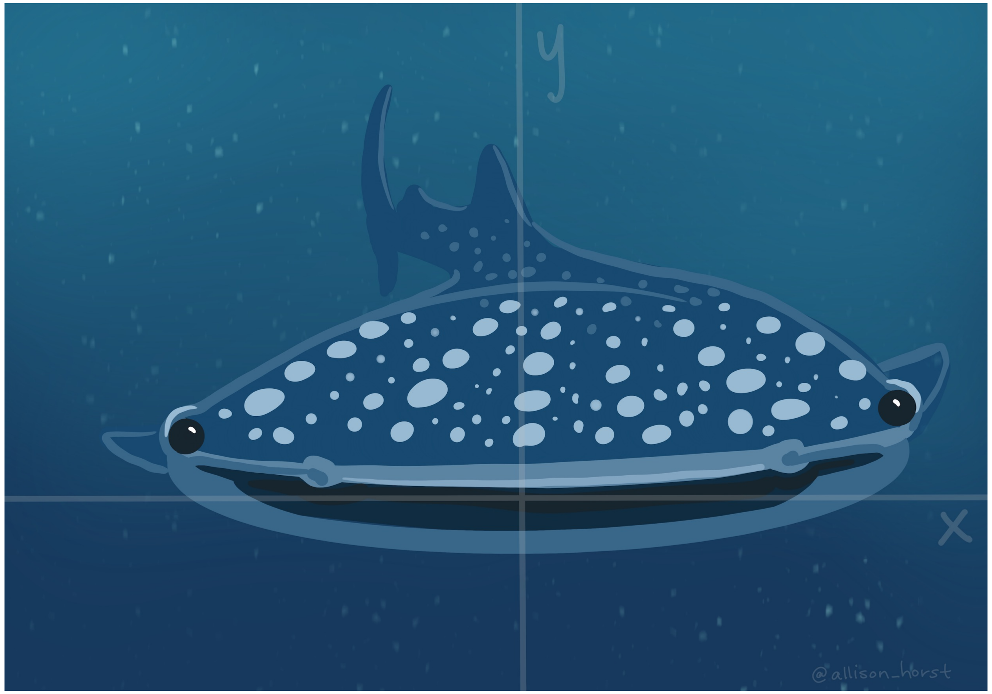
Vincent Guillemot

Mental images

Potato Chips Analysis

Cut the yummiest French fries

Whale versus krill: this is you (credit: Allison Horst)



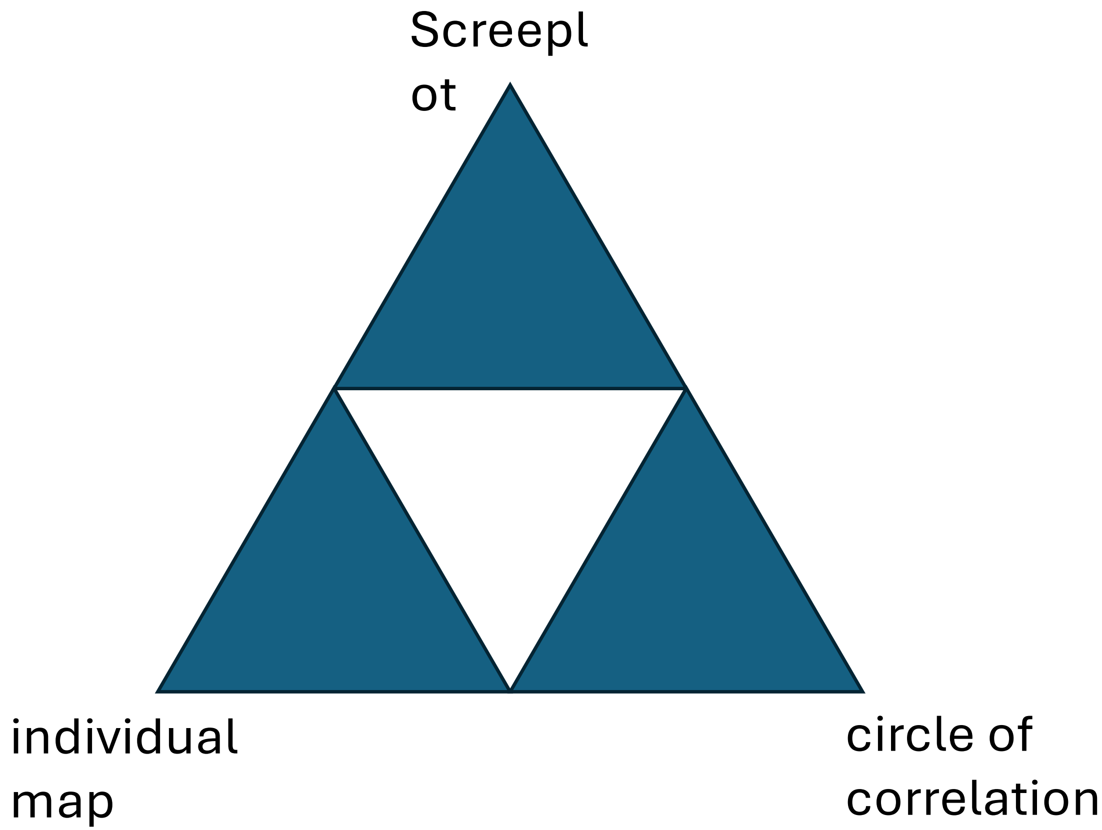
Eat the most krill (put on your 3D glasses)

Whale versus krill: this is your data (credit: Allison Horst)

Eat the most krill (put on your 3D glasses)

[Artwork by @allison_hors](#)

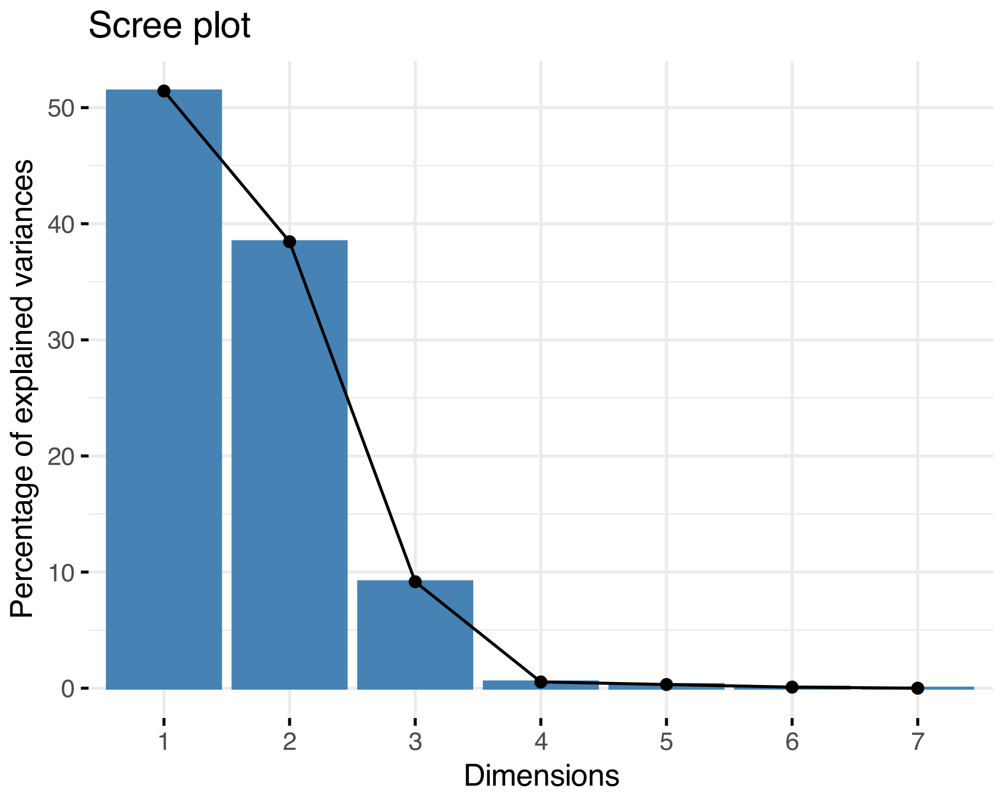
The tri-force of PCA



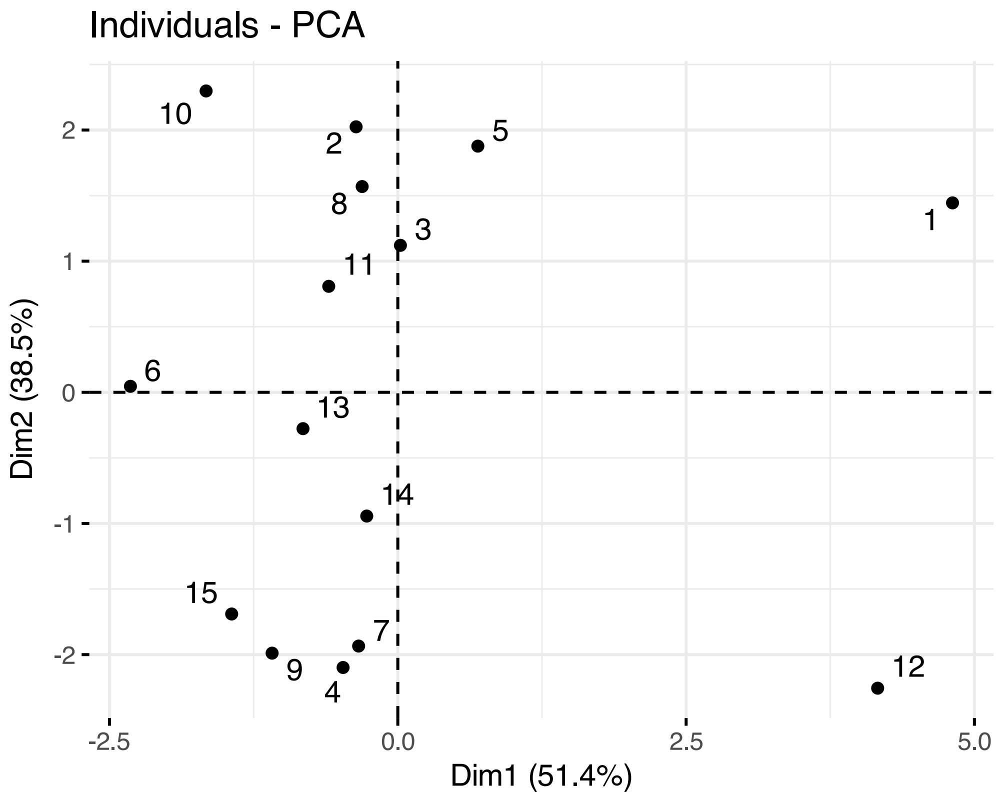
Example data

```r
n <- 15
dat.ex <- tibble(
  X1 = rnorm(n),
  X2 = rnorm(n),
  X3 = -X1,
  X4 = 2 * X2  + 0.25 * rnorm(n),
  X5 = X1 + X2 + 0.25 * rnorm(n),
  X6 = X1 - X2 + 0.25 * rnorm(n),
  X7 = rnorm(n)
)
```

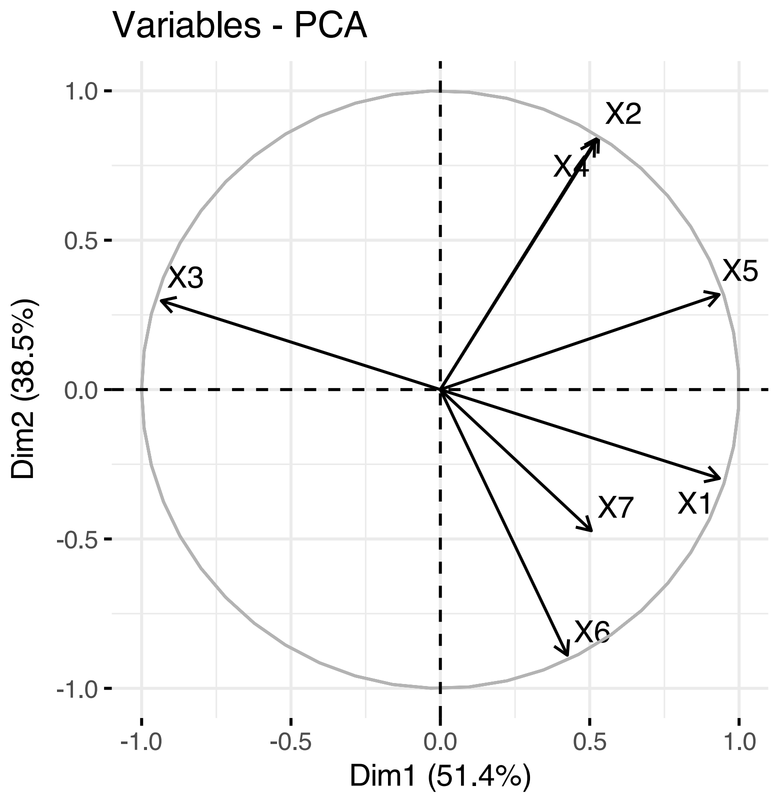
Example screeplot



Example individual map



Example circle of correlation



Vocabulary

French versus English

“Aaaaah, mais acépé en fait c’est la PCA !”

(Anonymous student, after 6 hours of teaching PCA in French)

English	French
PCA = principal component analysis	ACP = analyse en composantes principales
SVD = singular value decomposition	SVD = décomposition en valeurs singulières
EVD = eigenvalue decomposition	décomposition en éléments propres
ICA = independent component analysis	ICA = analyse en composantes indépendantes
MDS = multidimensional scaling	MDS = multidimensional scaling

R vocabulary

Base methods:

- `eigen` for eigenvalue decomposition, `svd` for singular value decomposition,
- `prcomp` and `princomp` for PCA,
- `biplot`

Nice packages:

- `FactoMineR`: `PCA`, `MFA`, `CA`, `MCA` and associates. In earlier versions, the graphs were “crude”…
- `factoextra`: “helper” package to make beautiful plots, and much more!
- `ade4`: more than “one block” type of analyses. Made by ecologists so ⇒ PCOA, coinertia analysis, STATIS, etc.
- `ExPosition`: made for psychometricians (they like PLS)

And a few nice books and papers

MOOC analyse de données de François Husson :
https://husson.github.io/MOOC_AnaDo/index.html

(also in English)

PCA paper(s) by Hervé Abdi:

https://personal.utdallas.edu/~herve/abdi-awPCA2010.pdf

(more?)

A little bit of Math

Notations

(non-universal) Conventions: matrices and vectors are **bold**

- n = number of observations, p = number of variables (only quantitative)
- i for an individual observation, and j for a single variable
- $\mathbf{X}$ = data matrix, with n rows and p columns, sometimes already centered, and scaled, to make our life easy
- $\mathbf{X}_j$ = variable j, and jth column of $\mathbf{X}$
- $\mathbf{w}$ a set of weights

A little detour: matrix multiplication

Take a pen and paper, and do this multiplication:

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 & 4 \\ -1 & 0 & 0 & 1 \end{bmatrix}$$

[Cool video: 5 ways to see matrix multiplication]

PCA

"Find a linear combination of the columns of the data that would capture the most information."

In mathematical words, find

$$\mathbf{Xw} = w_1\mathbf{X}_1 + \cdots + w_p\mathbf{X}_p$$

that maximizes… wait a minute! What are the dimensions?

- $\mathbf{X}$: n rows and p columns,
- $\mathbf{w}$: p rows and 1 columns,
- $\mathbf{Xw}$: n rows and 1 column.

The mathematical translation of the intuitions

Most popular intuition of PCA: how does it translate?

"PCA creates a linear combination of variables that maximizes variance."

$$\arg\max_{\|\mathbf{w}\|_2^2=1} \text{var}(\mathbf{X}\mathbf{w})$$

- Why $\| \mathbf{w} \|_2 = 1$?
- Dirty trick: $\text{var}(\mathbf{X}\mathbf{w}) = \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}$

Least “well-known” intuition of PCA: how does it translate?

“PCA creates a linear combination of variables that maximizes correlation.”

$$\underset{\mathbf{w}}{\operatorname{argmax}} \sum_{j=1}^{p} \operatorname{cor}\left(\mathbf{X}\mathbf{w}, \mathbf{X}_j\right)^2$$

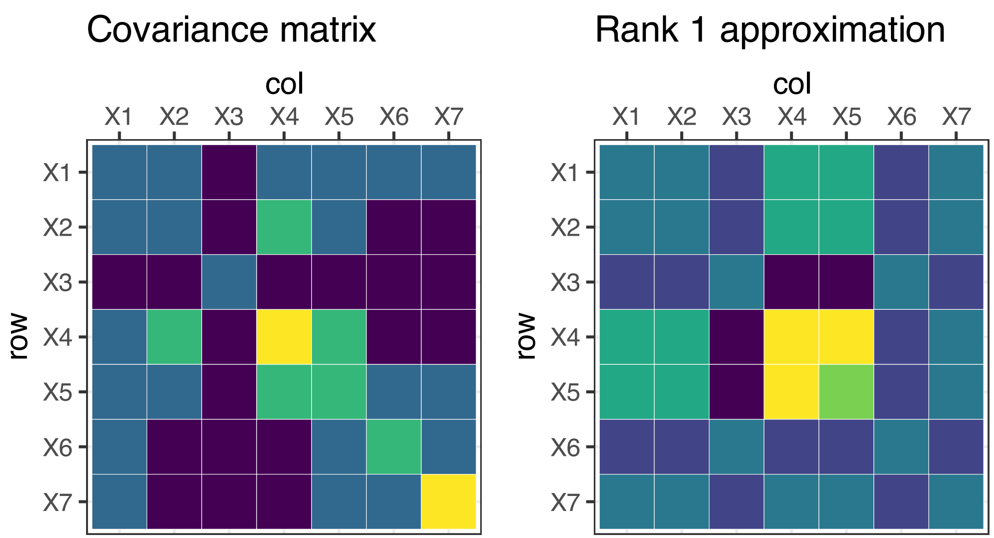
Second least "well-known" intuition of PCA: how does it translate?

"PCA creates the best lower rank approximation of the covariance matrix."

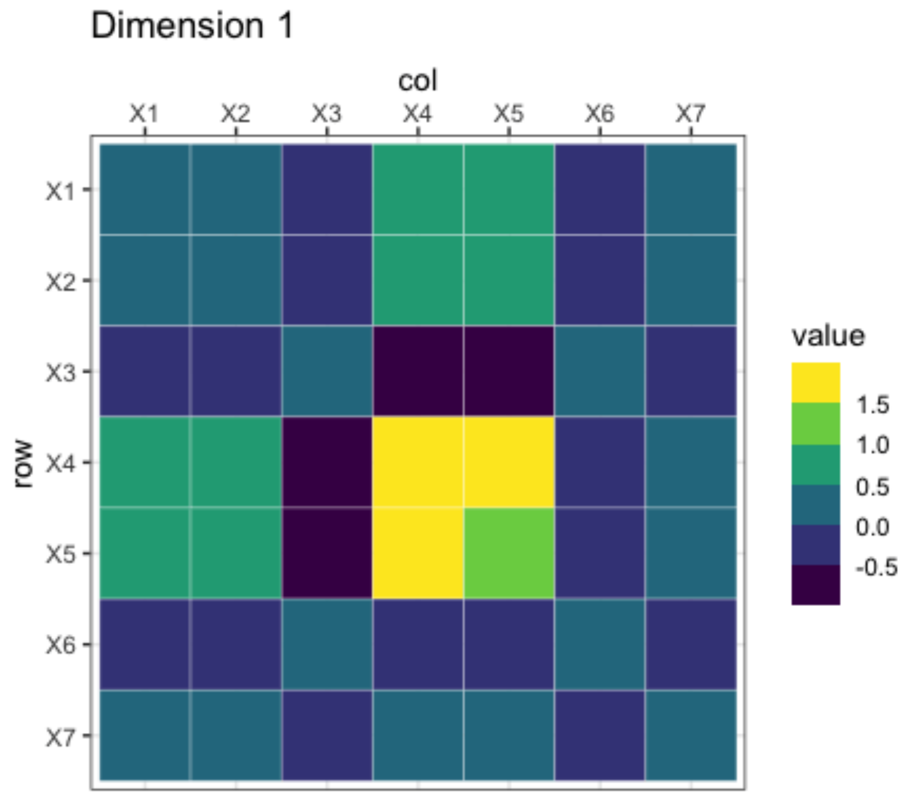
$$\arg\min_{\|w\|_2^2=1} \left\| \frac{1}{n}\mathbf{X}^\top\mathbf{X} - \lambda \mathbf{w}\mathbf{w}^\top \right\|_F^2$$

- $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$
- λ: the [blank] of the covariance matrix
- $\mathbf{w}$: the [blank] of the covariance matrix

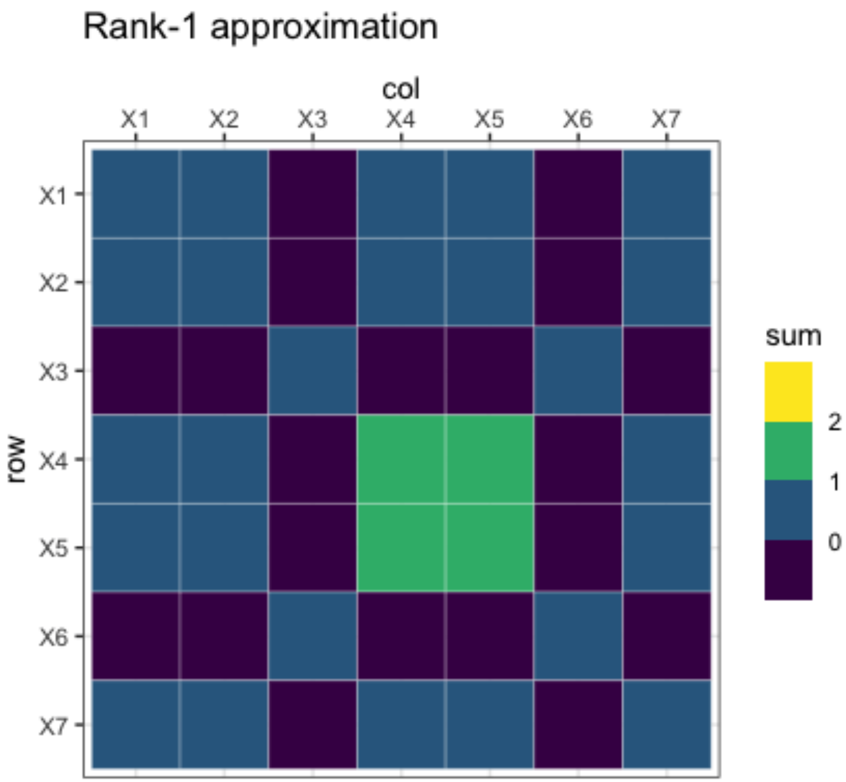
A little image



Rank-1 approximations



Increasing rank approximations



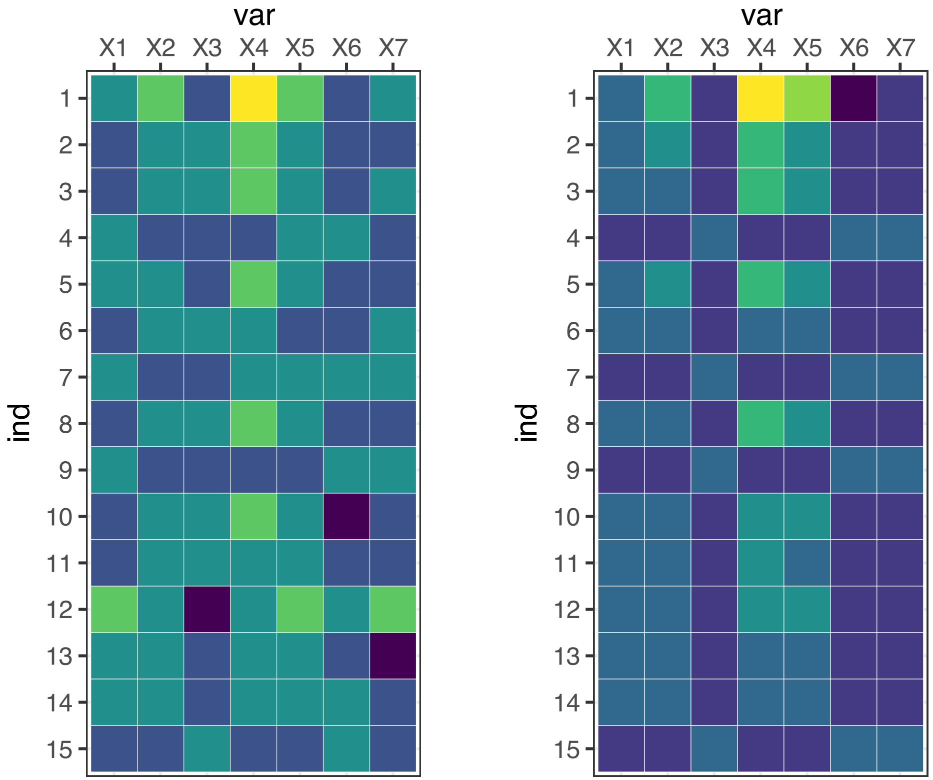
We can do the same kind of magic with the data itself

Singular value decomposition can be used to approximate a rectangular matrix with a lower ranked matrix of the same dimension

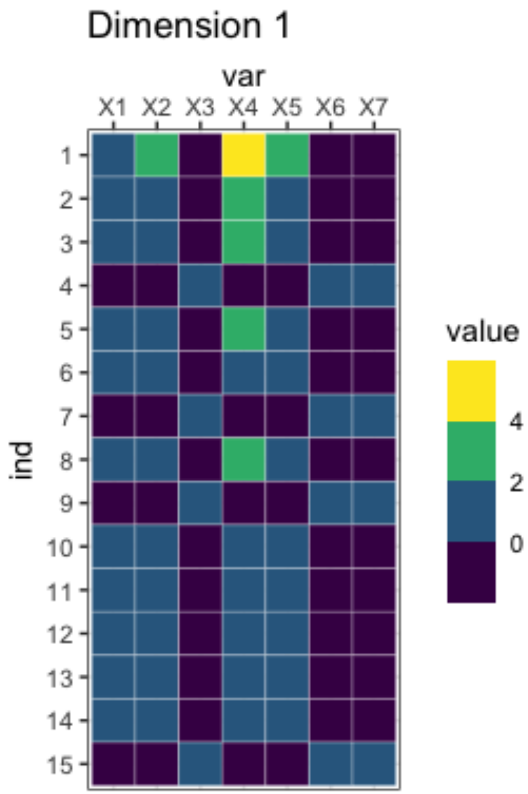
$$\arg\min_{\|\mathbf{u}\|_2^2=\|\mathbf{w}\|_2^2=1} \|\mathbf{X}-\delta\mathbf{u}\mathbf{w}^\top\|_F^2$$

- δ: singular value
- $\mathbf{u}$: left singular vector
- $\mathbf{w}$: right singular vector

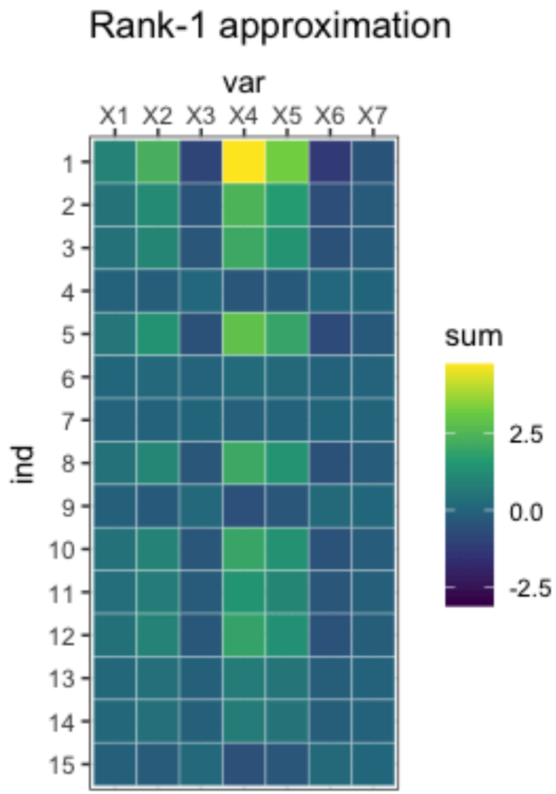
Rank 1 approximation



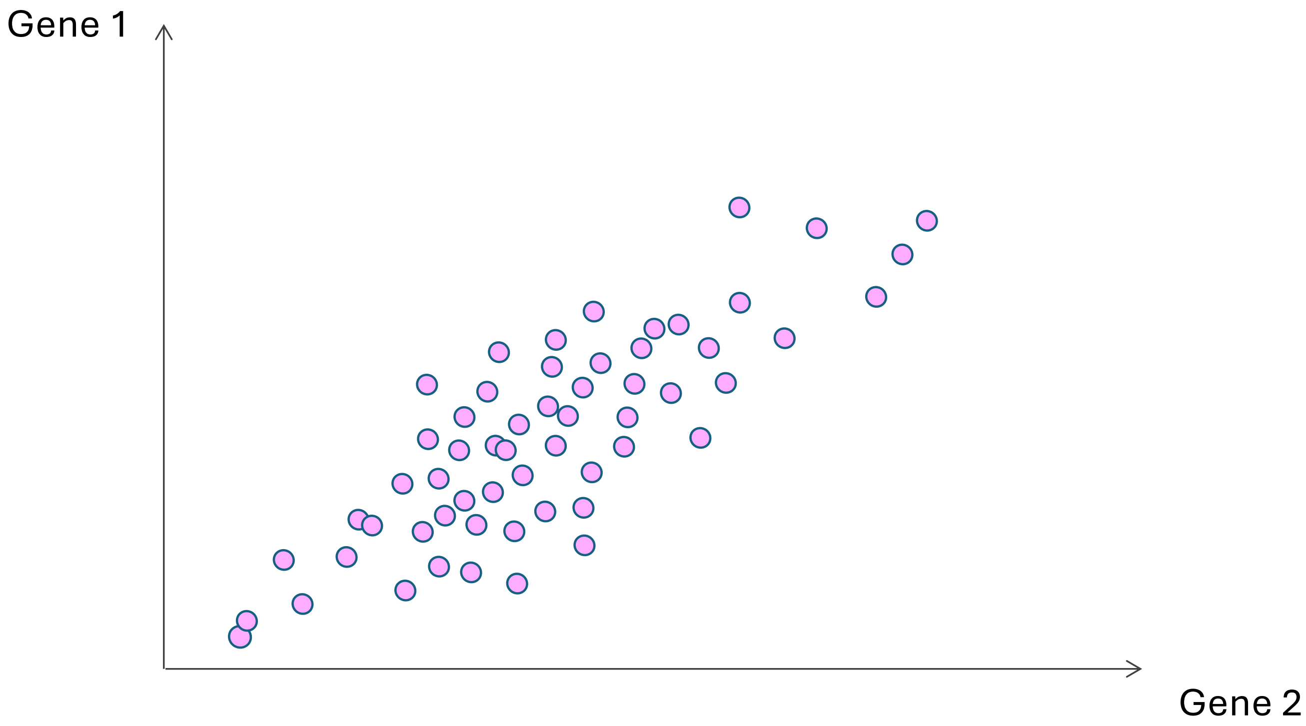
Rank-1 approximations



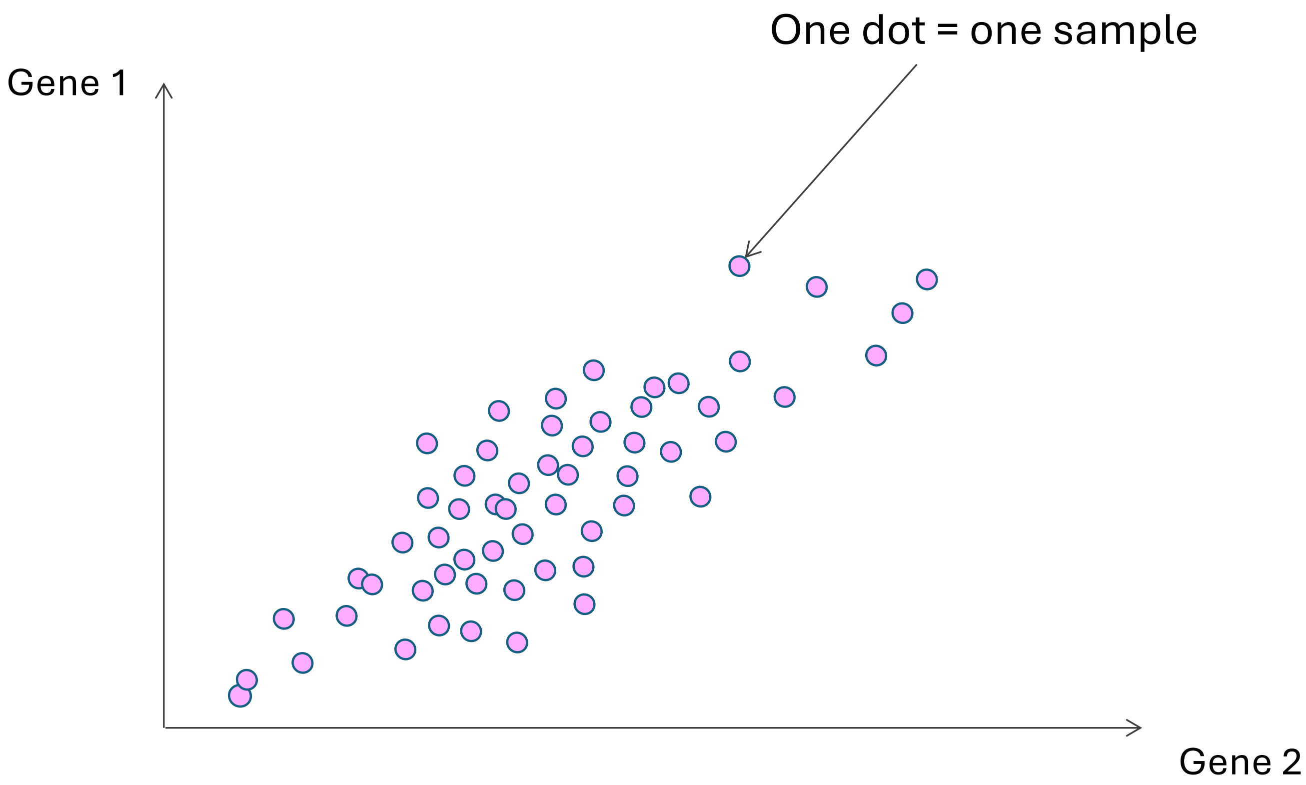
Increasing rank approximations



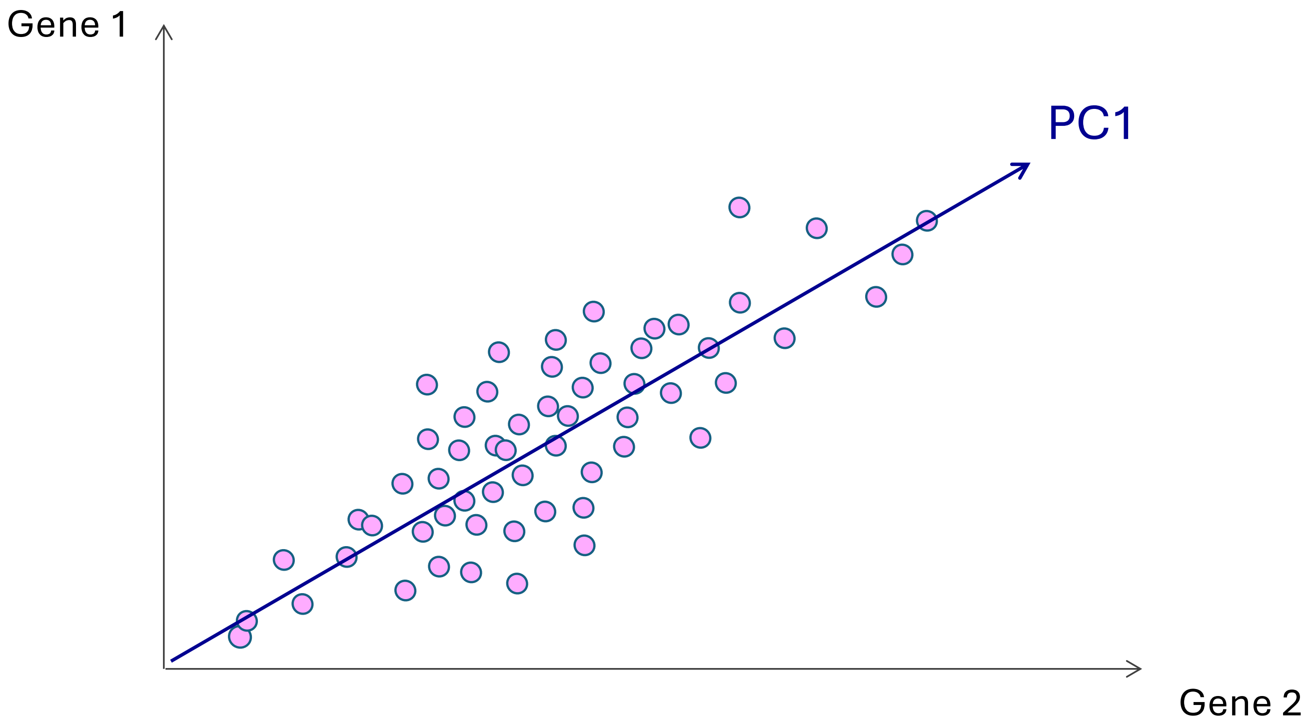
PCA: example in two dimensions



PCA: example in two dimensions

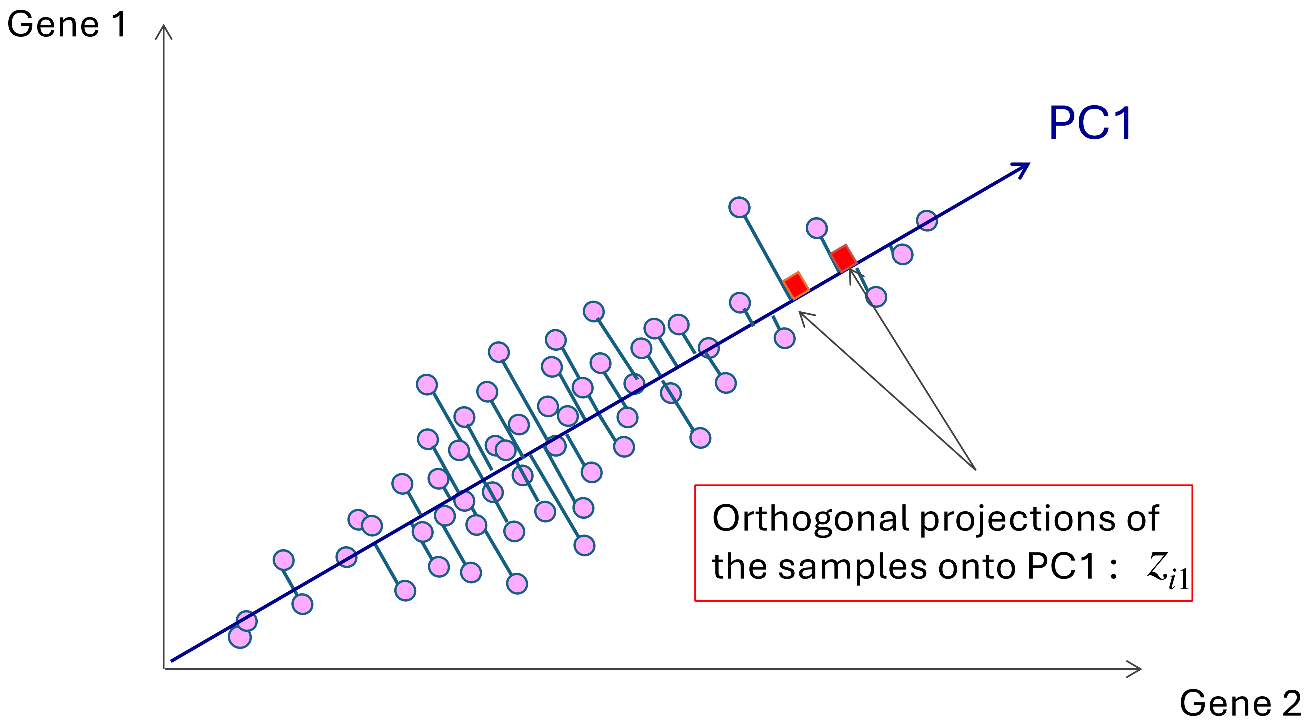


PCA: example in two dimensions

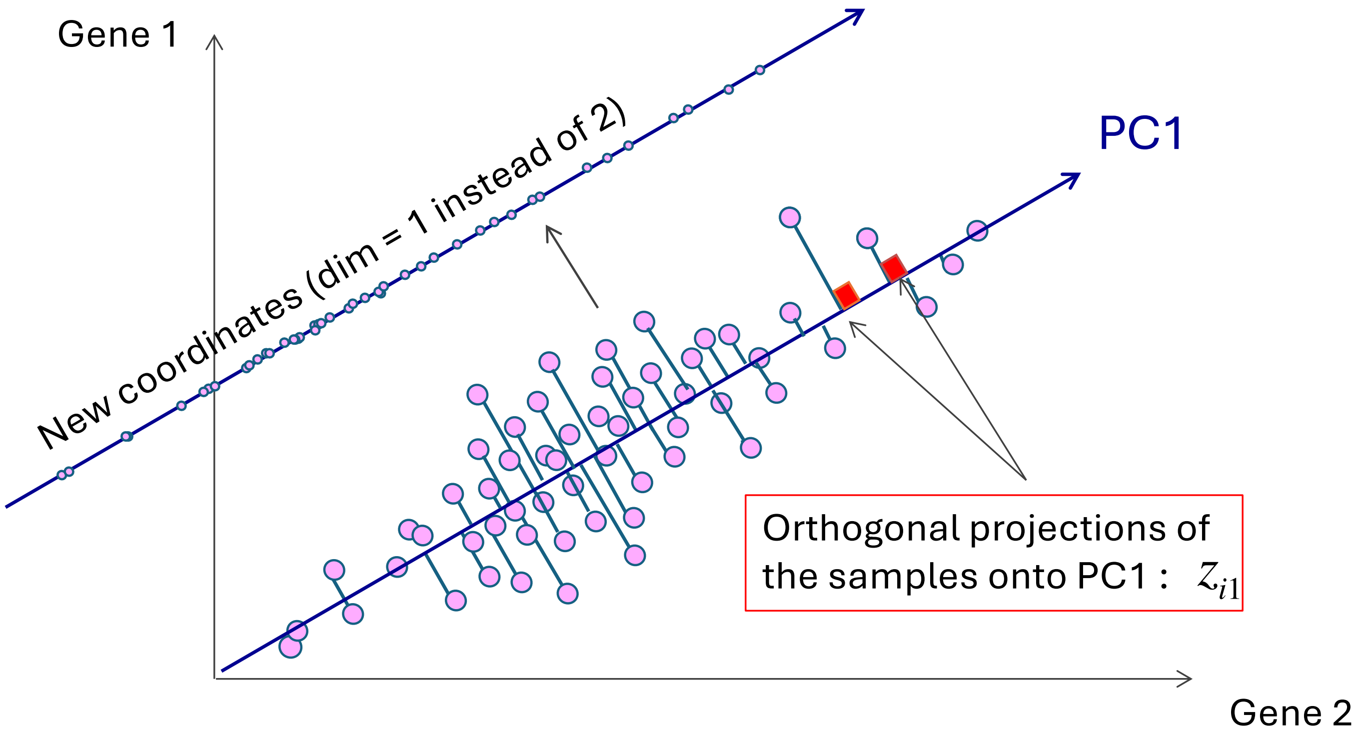


Find the direction(s)/space(s) with the property that the squared distance of the points to their orthogonal projection onto the space is minimized.
« The best line that fits the data»

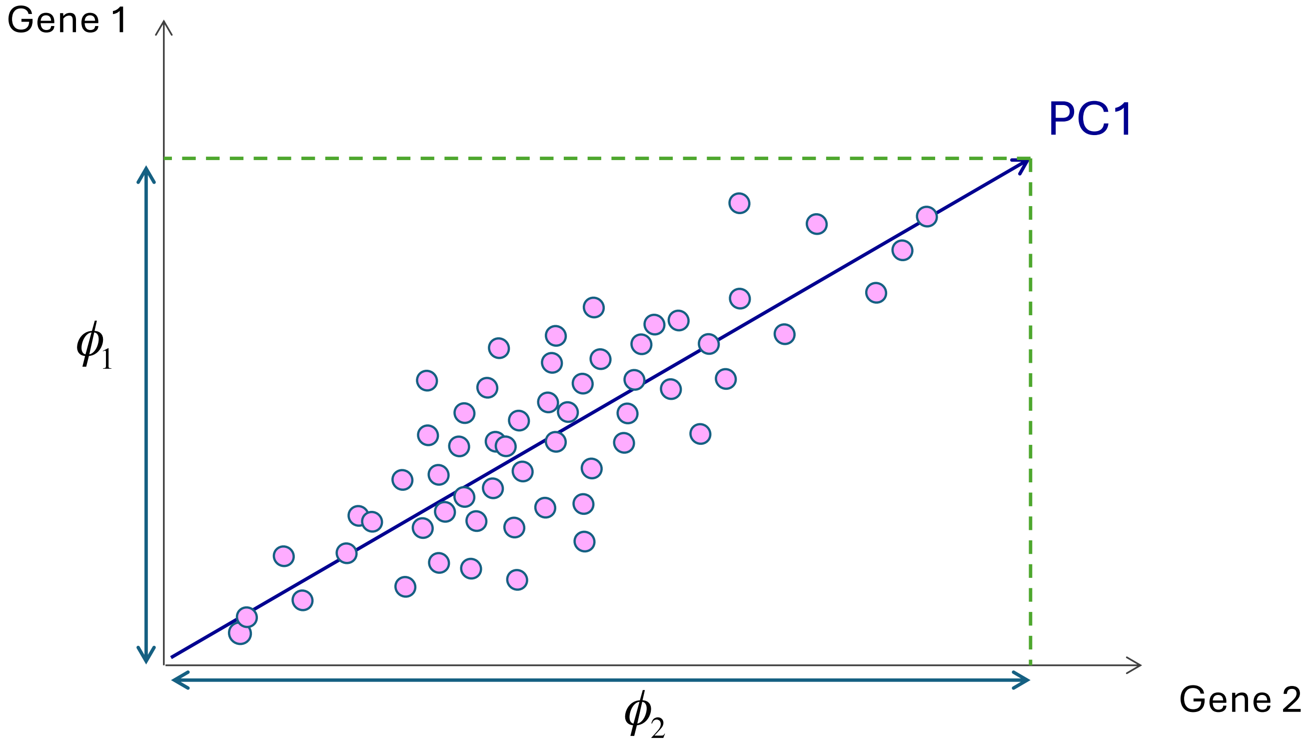
PCA: example in two dimensions



PCA: example in two dimensions

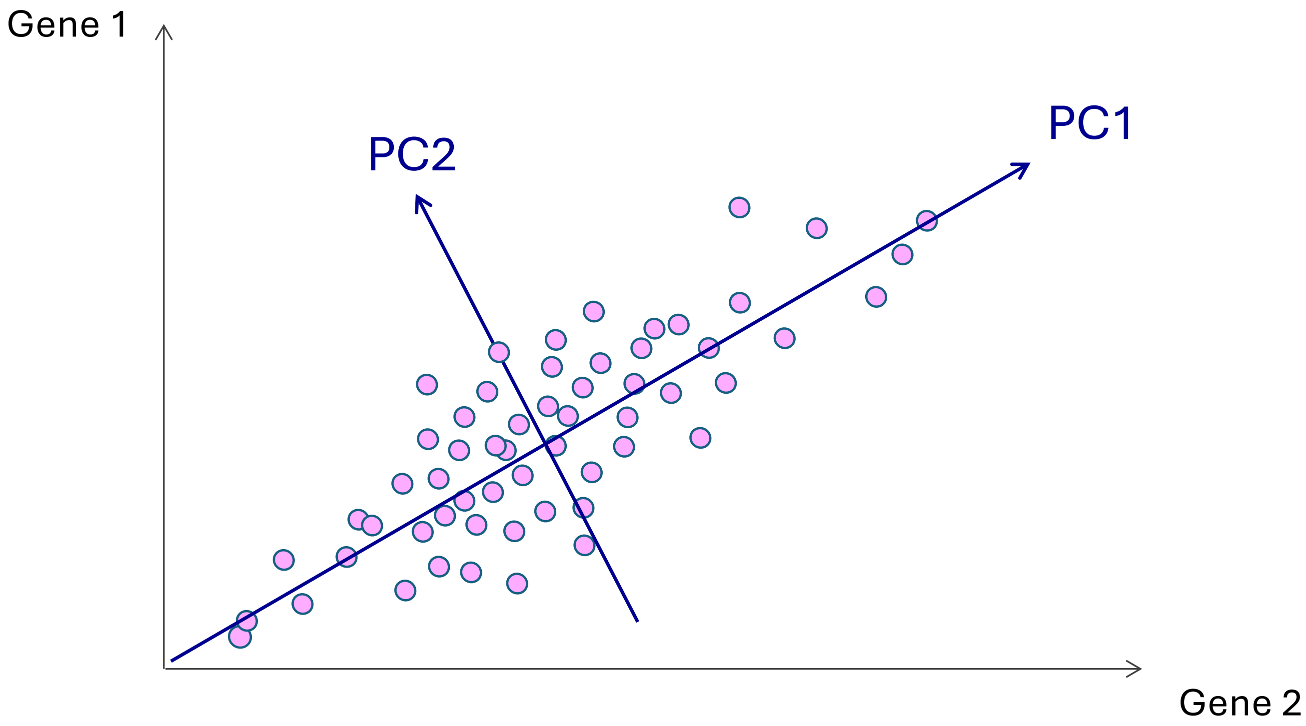


PCA: example in two dimensions



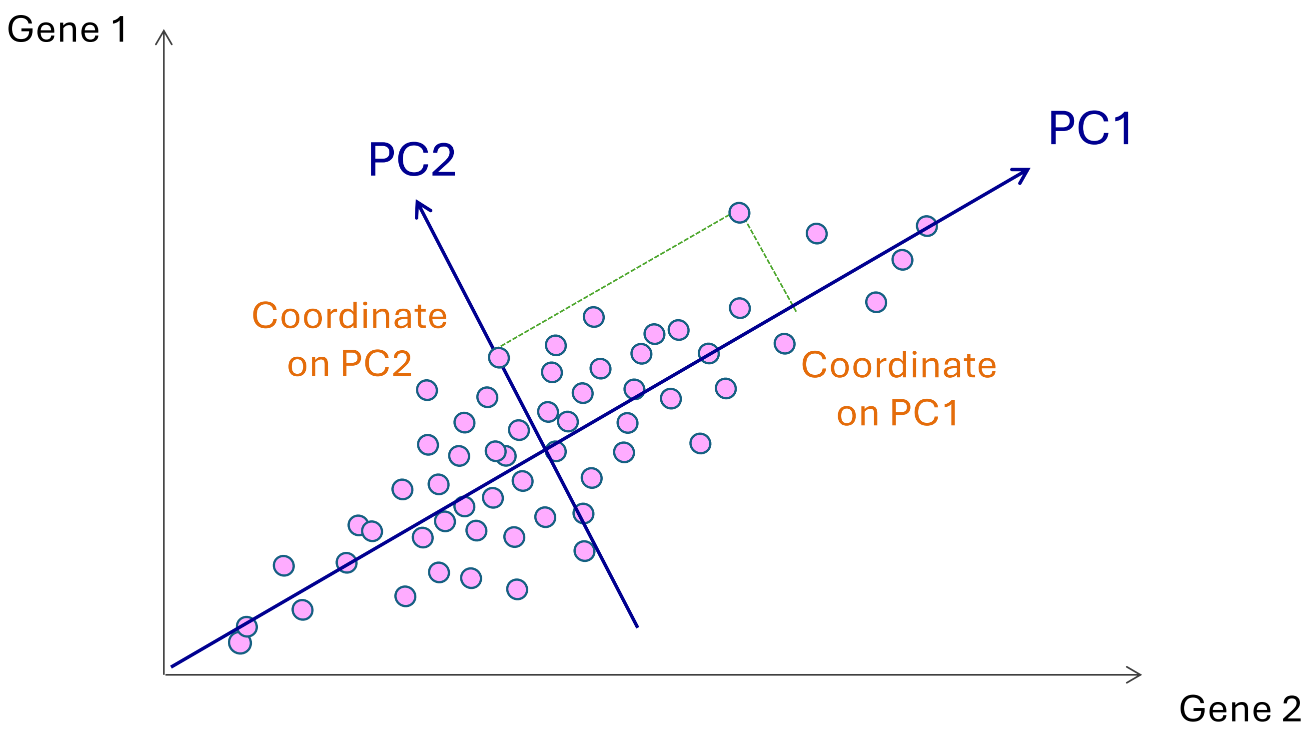
(PC1) $Z_1 = \phi_1 Gene_1 + \phi_2 Gene_2$

PCA: example in two dimensions



The second principal component is the linear combination of the variables that has maximal variance among all linear combinations that are orthogonal (uncorrelated) with PC1.

PCA: example in two dimensions



How to do it?

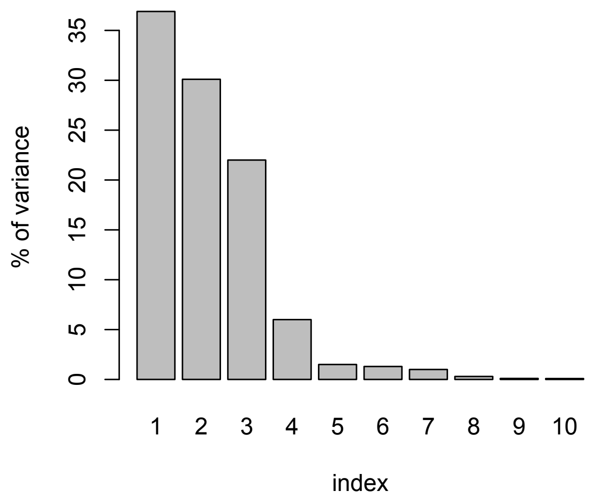
The old way ...

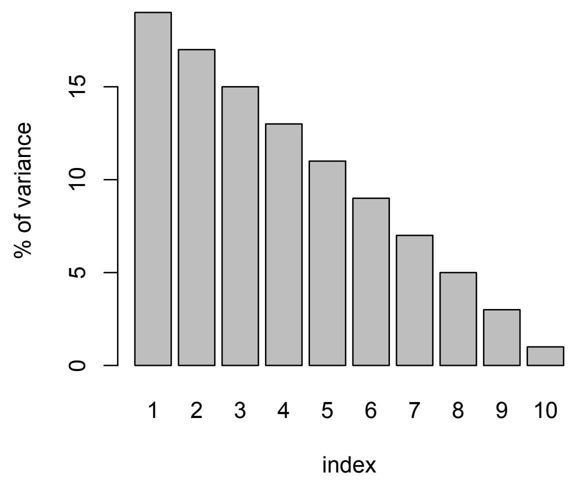
How many PCs?

- There are a couple ways of deciding how many principal components to use but **no correct answer**.
- Consider what is your reason of using PCA in the first place:
  - *How many do you need?* Sometimes PCA is used in a classification framework.
  - *How much variance do you want?* We could decide that we need a certain fraction of the total variance, and choose the number of PCs allowing to achieve it.
  - *How many eigenvalues are big?* i.e. *how many PCs capture large variance?*
- All these methods are arbitrary.

How many PCs?
Examples of methods

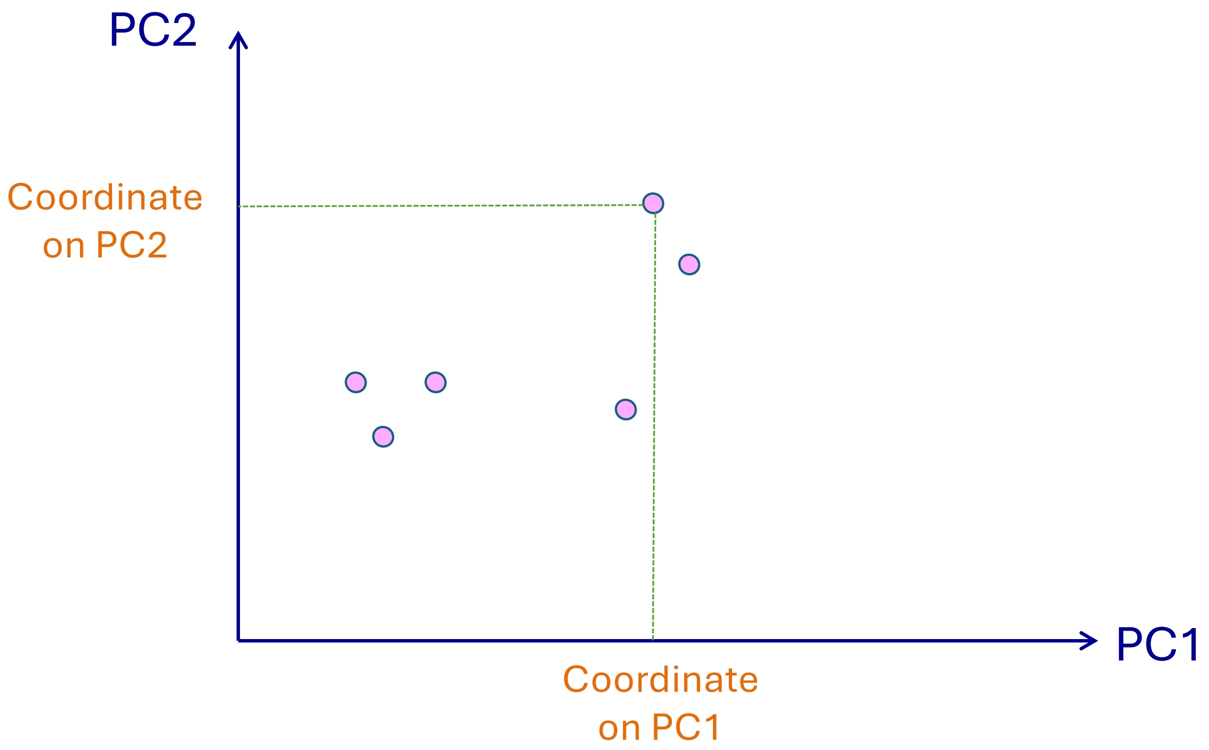
- Always useful to have a look at the plot of the fraction of total variance explained by each component.
- *Scree* graph (Cattell, 1966) : Look at the plot of the percentage of variance l_k against k and decide which value of k the defines an *'elbow'* in the graph (subjective). In practice, rarely easy to choose.





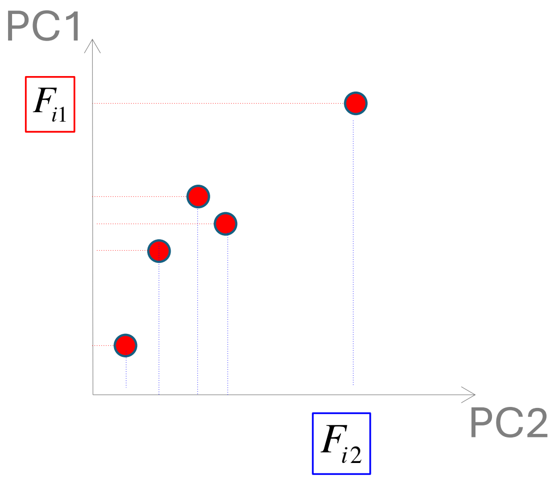
PCA plot of the individuals

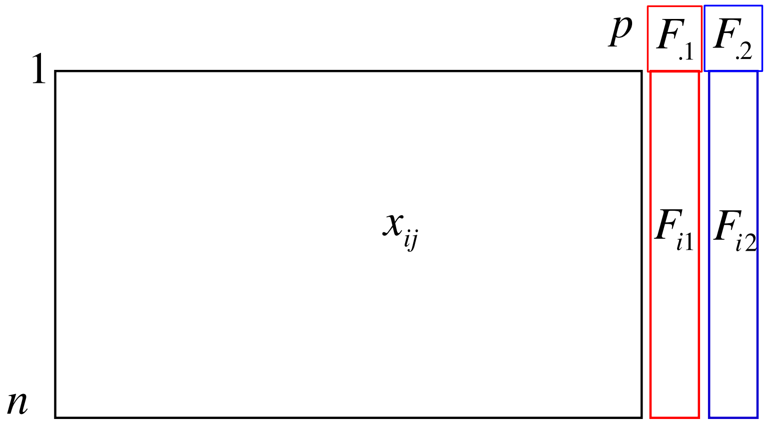
- Plot of the individual projections onto the space spanned by for example the first and the second principal component (PC).



Interpretation of the PCA plot of the individuals using the variables

Let's consider the coordinates of the individuals on the PCs as variables.

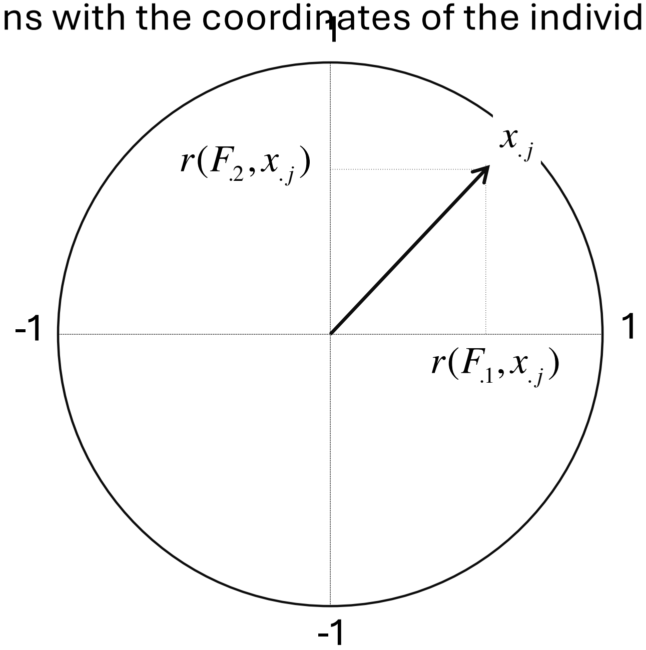




Interpretation of the PCA plot of the individuals using the variables

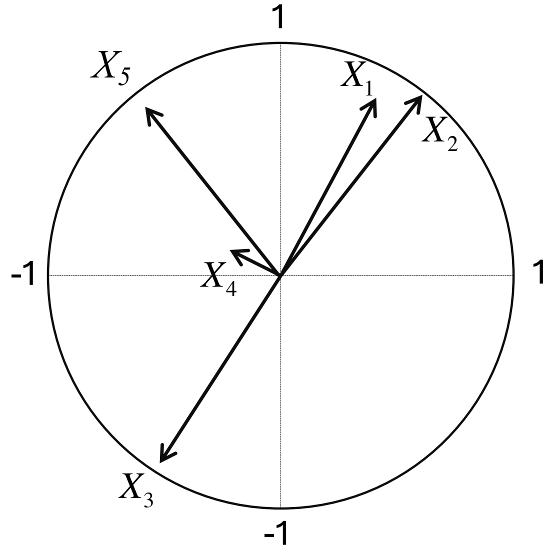
- We can study the correlation between the initial variables and the PCs.
- The correlation between a PC and a variable estimates the information they share.

- **Correlation circle:** The variables can be plotted as points in the component space using their correlations with the coordinates of the individuals on the PCs.



Interpretation of the PCA plot of the individuals using the variables

- An arrow close to the circle indicates a variable that is well characterized by the represented PCs (X_1, X_2, X_3, X_5).
- An arrow close to the center indicates a variable whose properties are not captured/described by the represented PCs (X_4).
- Two arrows close to the circle and to each other indicate a positive correlation between the two variables (X_1, X_2).
- Two arrows close to the circle and in opposite direction indicates a negative correlation between the two variables (X_1, X_3 or X_2, X_3).
- Two arrows close to the circle and at a right angle to one another are not correlated (X_1 with X_5 or X_2 with X_5).

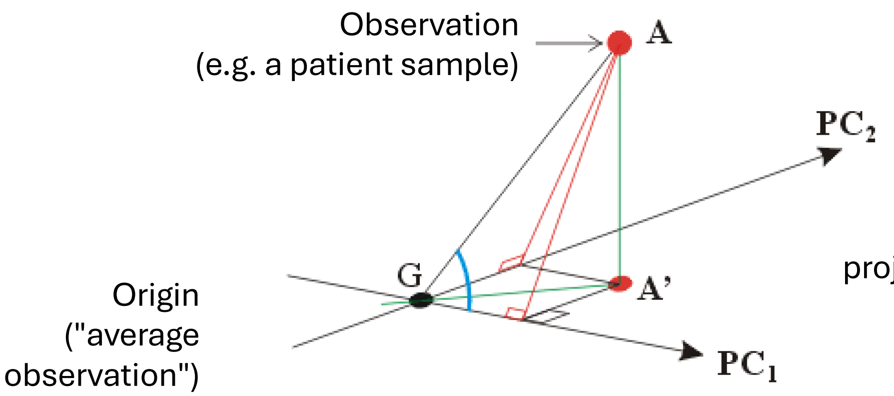


PCA in *R* with package *FactoMineR*

Squared cosine of the subjects with the components

- Is the distance between the origin and the projection of a point in the principal plane a good approximation of the true distance of the observation to the origin ?

- Rather than giving the distance to the principal plane, softwares often output the ratio of the projected distance to the true distance. This quantity corresponds to the square of the cosine of the angle from the right triangle made with the origin, the observation, and its projection on the component.



Cos2(blue angle) =
projected distance (PC1) / true distance

```
library(dplyr)
library(FactoMineR)
library(factoextra)
library(ggplot2)


n <- 15
dat.ex <- tibble(
 X1 = rnorm(n),
 X2 = rnorm(n),
 X3 = -X1,
 X4 = 2 * X2  + 0.25 * rnorm(n),
 X5 = X1 + X2 + 0.25 * rnorm(n),
 X6 = X1 - X2 + 0.25 * rnorm(n),
 X7 = rnorm(n)
)
```

```
res.pca.ex <- PCA(dat.ex, scale.unit = TRUE, graph =
FALSE)
fviz_screeplot(res.pca.ex)
fviz_pca_ind(res.pca.ex, repel = TRUE)
fviz_pca_var(res.pca.ex, repel = TRUE)
```