LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Machine Learning

PCA

Bernd Bischl, Christoph Molnar, Daniel Schalk, Fabian Scheipl

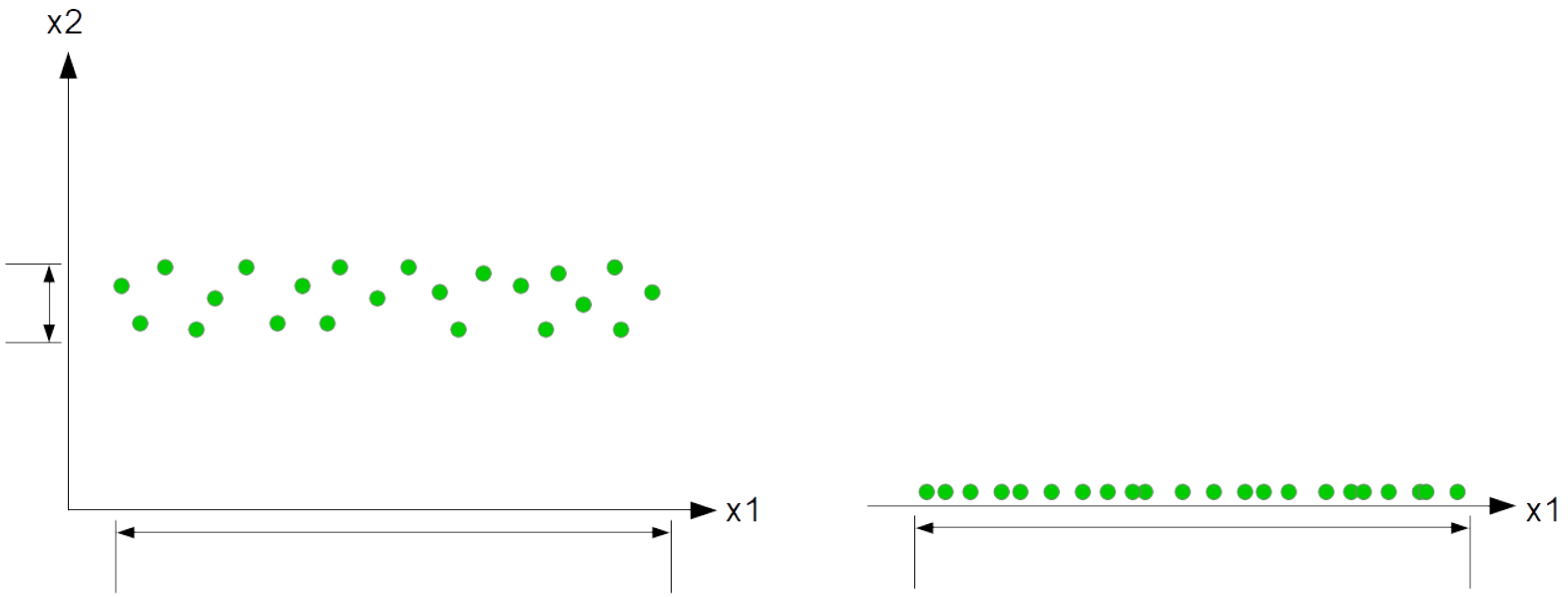
Department of Statistics – LMU Munich

Principal Component Analysis

PCA INTUITION

Motivational example I:

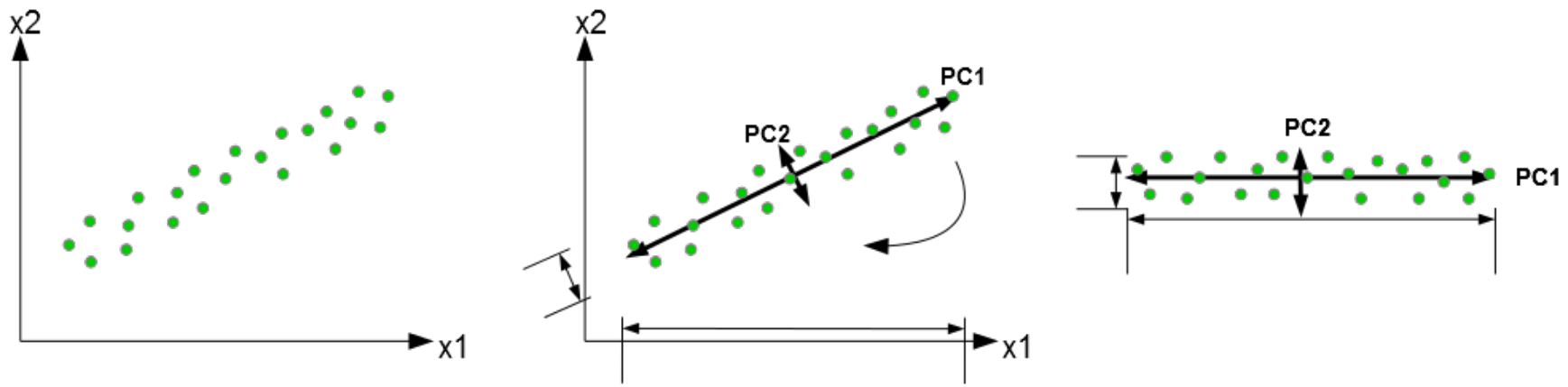
- Variable x_1 explains most of the variation.
- Variable x_2 has a lower variance than x_1.
- If we disregard x_2 and project the points into the 1-dimensional space of x_1, we do not lose much information w.r.t. variability.



PCA INTUITION

Motivational example II:

- x_1 and x_2 are correlated and have similar variances.
- Find a new orthogonal axes (e.g. PC1 and PC2), where PC1 explains most of the variation.
- Rotate the points and consider PC1 and PC2 as new coordinate system (situation as in the previous example).
- We can now project points onto PC1 and disregard PC2 (hopefully without losing much information).

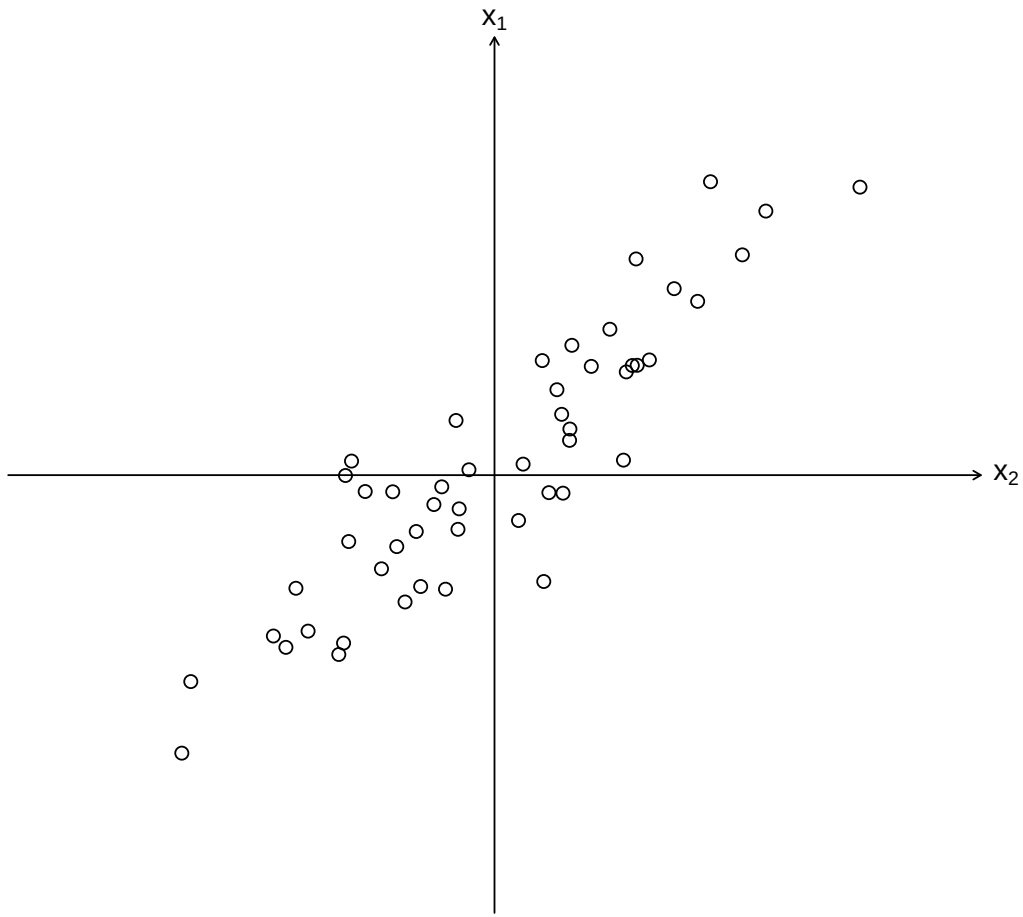


PCA INTUITION

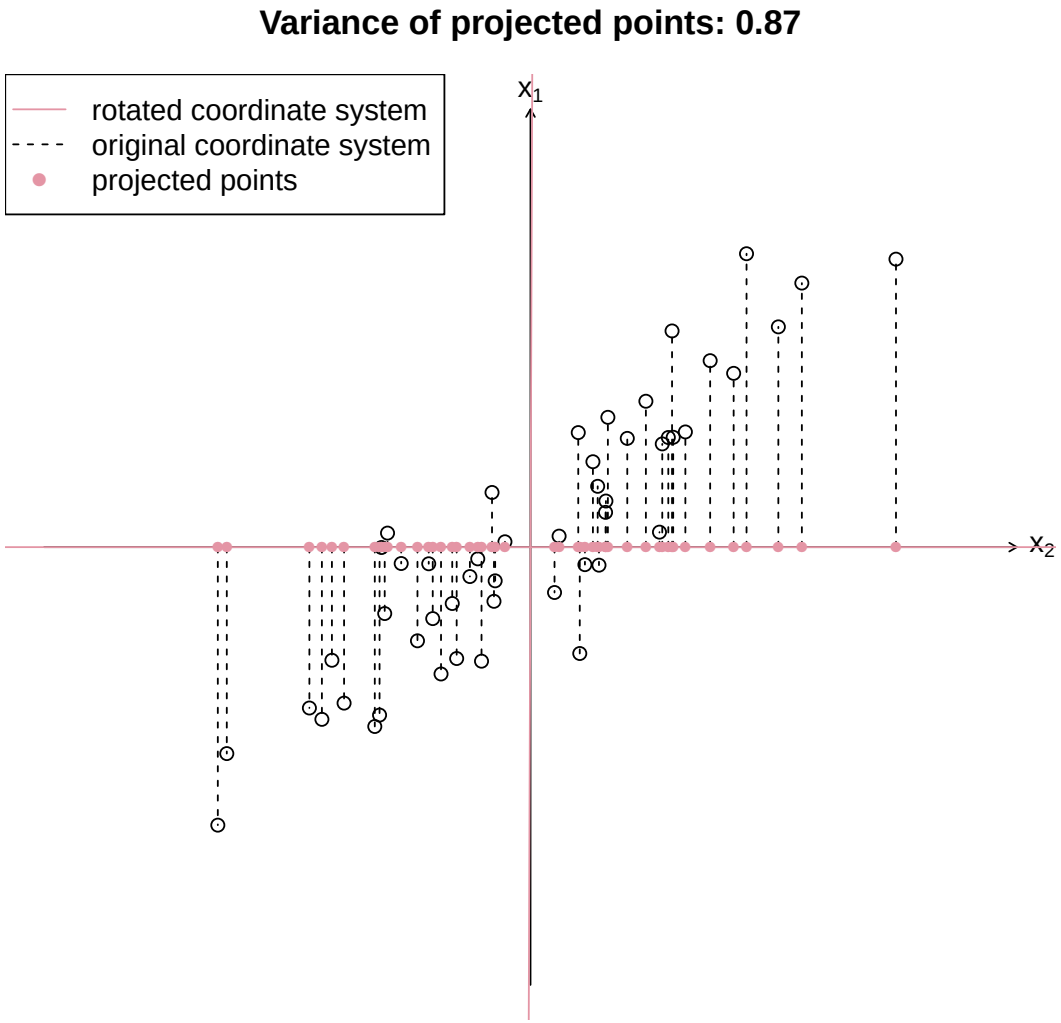
General procedure:

1. Rotate the original p-dimensional coordinate system until the first PC that explains most of the variation is found.
2. Fix the first PC and proceed with rotating the remaining $p - 1$ coordinates until the second PC (which is orthogonal to the first PC) is found that explains most of the *remaining* variation, etc.
3. We can reduce the dimensions by projecting the points onto the first, say $k < p$, PC.

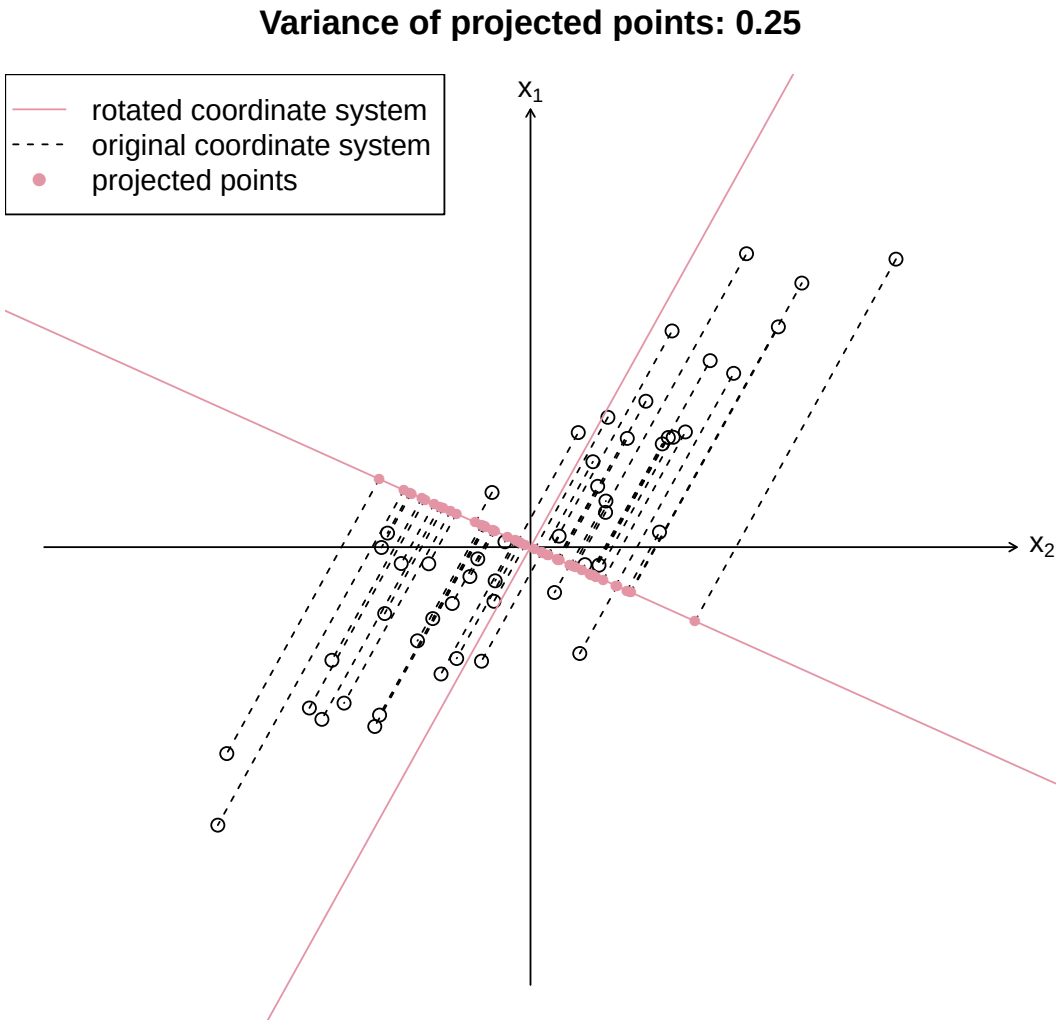
PCA INTUITION: FIND FIRST PC



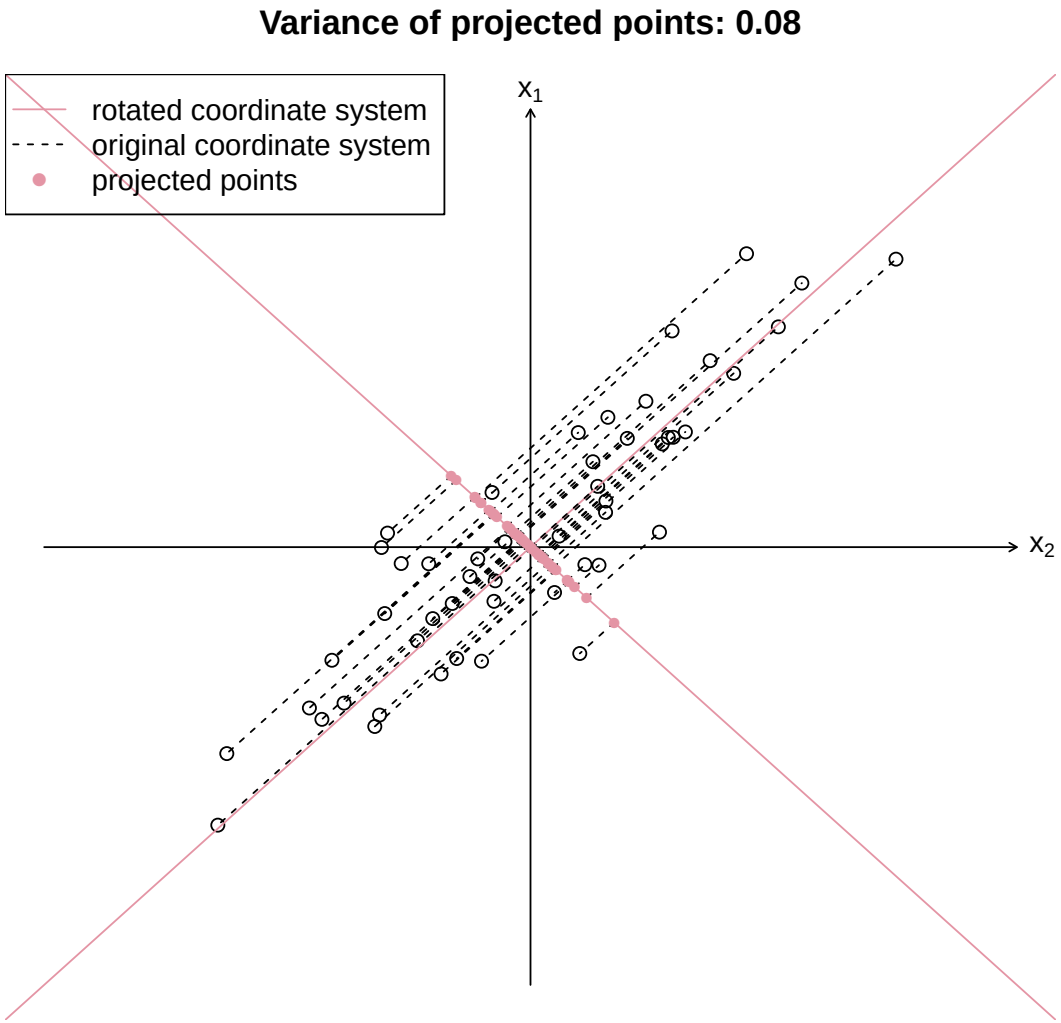
PCA INTUITION: FIND FIRST PC



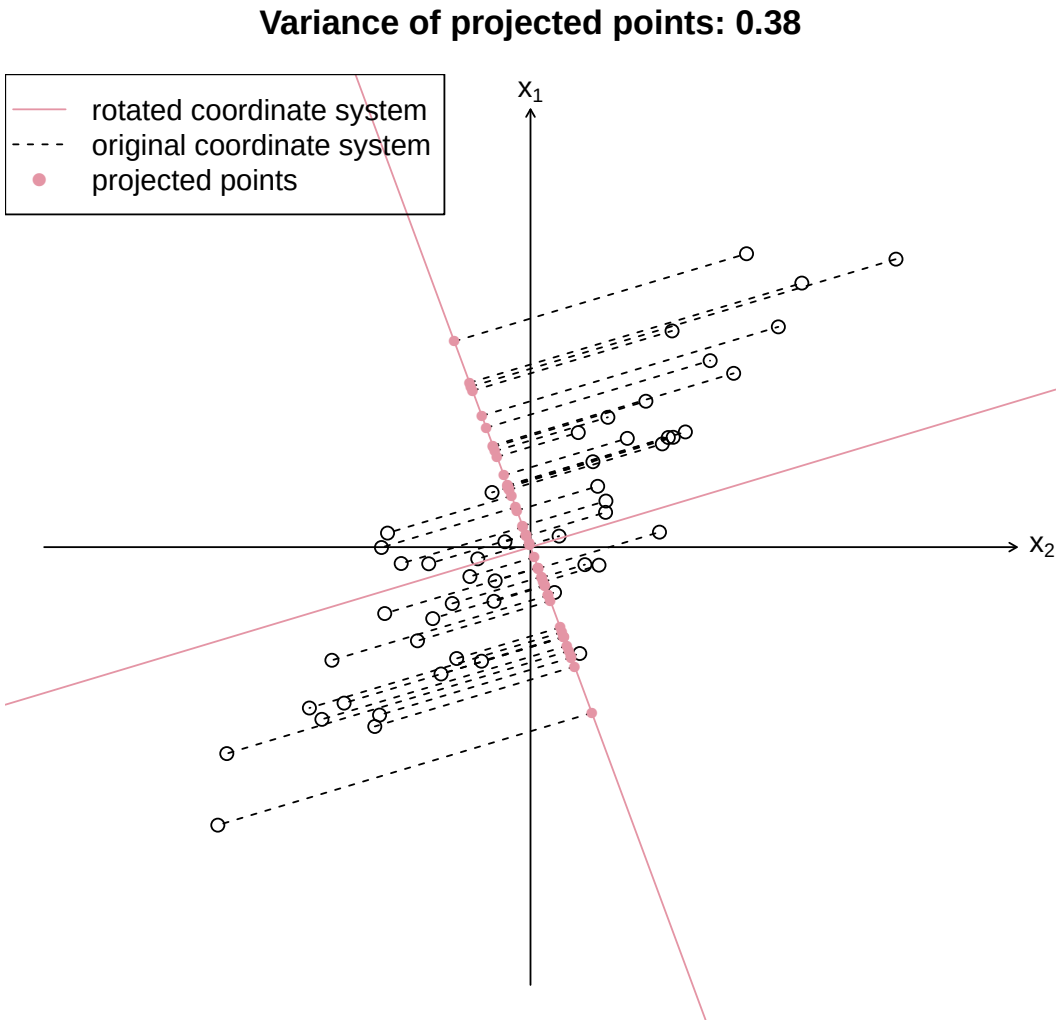
PCA INTUITION: FIND FIRST PC



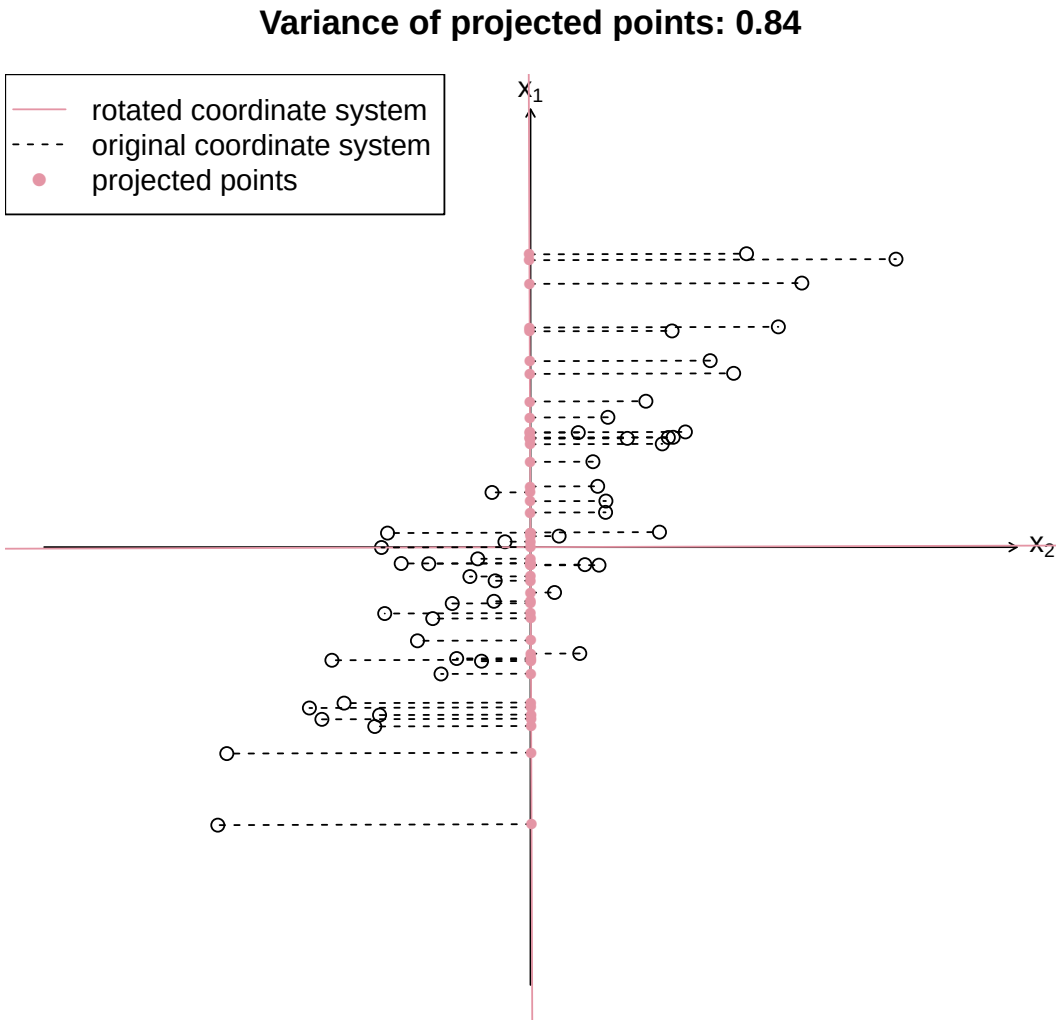
PCA INTUITION: FIND FIRST PC



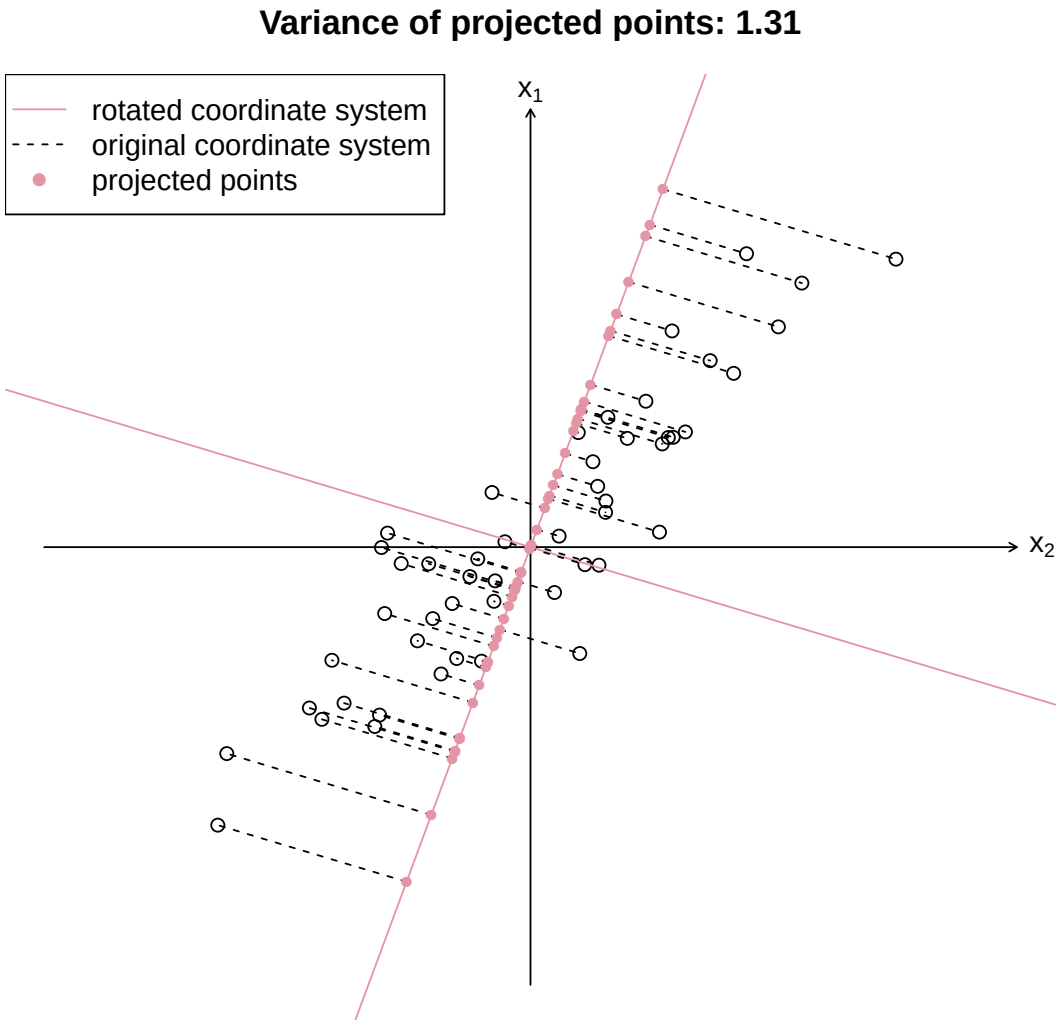
PCA INTUITION: FIND FIRST PC



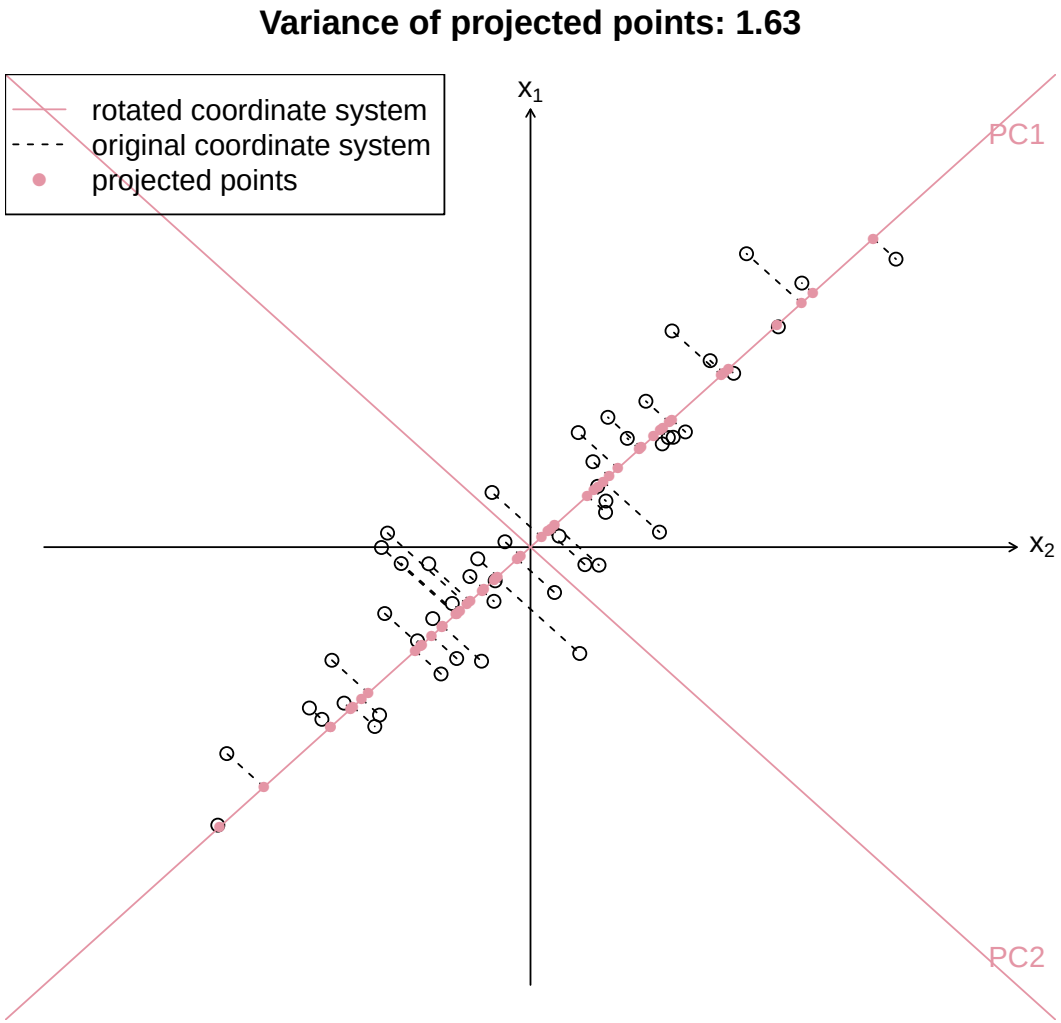
PCA INTUITION: FIND FIRST PC



PCA INTUITION: FIND FIRST PC

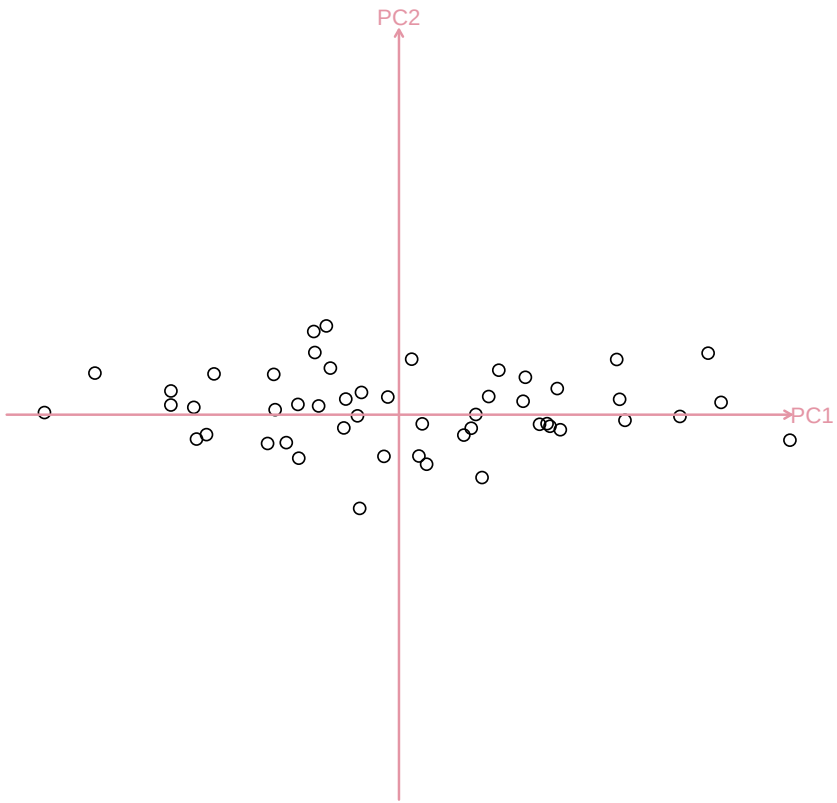


PCA INTUITION: FIND FIRST PC



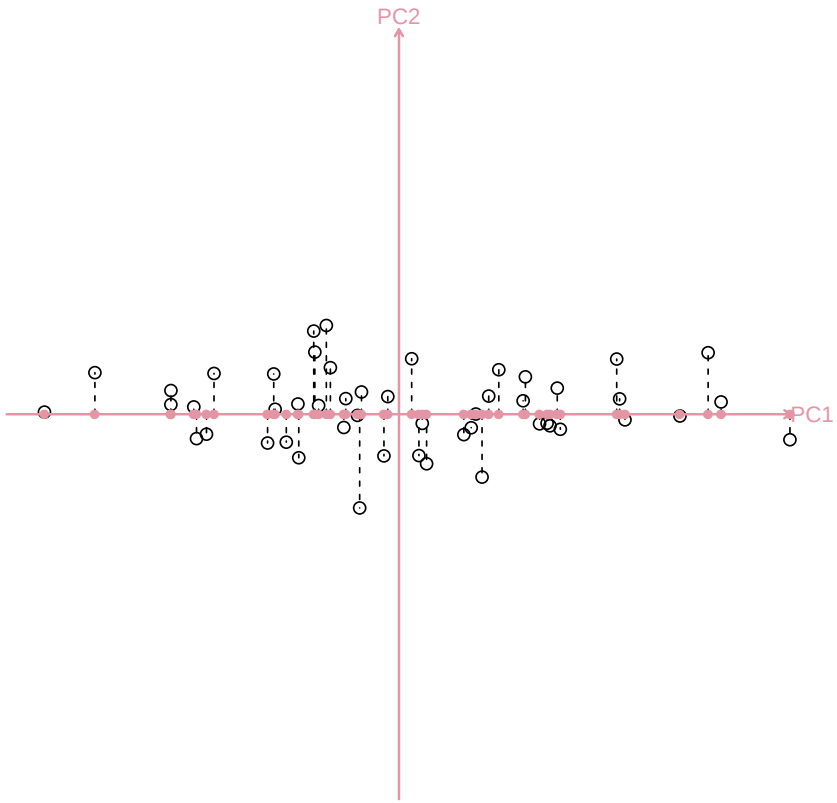
PCA INTUITION: REDUCE DIMENSIONALITY

Rotate the points and use PC1 and PC2 as new coordinate system.
Here, the PC1 axis explains most of the variance:



PCA INTUITION: REDUCE DIMENSIONALITY

Dimensionality can be reduced by projecting the points onto the PC1 (and by disregarding PC2). The hope is that we won't lose much information this way.



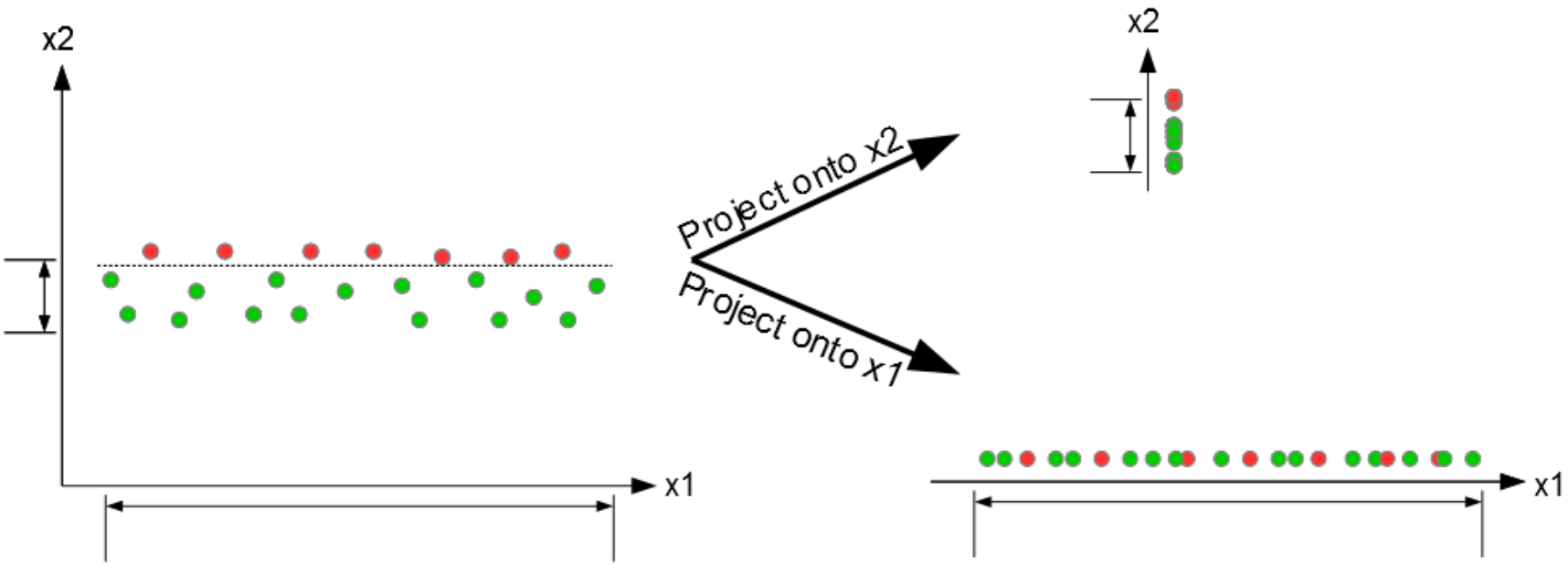
PCA INTUITION: SUMMARY

Idea: Transform an original set of correlated metric variables to a new set of uncorrelated (orthogonal) metric variables, called principal components (PC), that explain the variability in the data.

- The objective is to investigate if only a few PC account for most of the variability in the original data.
- If the objective is fulfilled, we can use fewer PCs to reduce the dimensionality.
- The PCs remove collinearity of the input variables as they are orthogonal to each other.

PCA INTUITION: FINAL REMARKS

- PCA is used for dimensionality reduction by disregaring dimensions with lower variability.
- There is always an information loss, especially for other criteria.
- E.g., dimensionality reduciton can worsen the classification accuracy when the task is to classify two groups:



DERIVING THE FIRST PC MATHEMATICALLY

Aim: Find a new set of variables (PC scores) $\mathbf{pc}_1, \ldots, \mathbf{pc}_p$ based on the original data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ so that

- each PC score $\mathbf{pc}_1, \ldots, \mathbf{pc}_p$ is a linear combination of the original metric variables with coefficient weights (so-called **loading vectors**) $\mathbf{a}_1, \ldots, \mathbf{a}_p$, i.e.

$$\mathbf{pc}_j = a_{j1}\mathbf{x}_1 + a_{j2}\mathbf{x}_2 + \ldots + a_{jp}\mathbf{x}_p = \mathbf{X}\mathbf{a}_j.$$

- the set is mutually uncorrelated: $Cov(\mathbf{pc}_j, \mathbf{pc}_k) = 0, \ \forall j \neq k$.
- the variances of the PC scores decrease:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p, \quad \text{where } \lambda_k := Var(\mathbf{pc}_k).$$

DERIVING THE FIRST PC MATHEMATICALLY

We look for the loading vector $\mathbf{a}_1 = (a_{11}, a_{21}, \ldots, a_{p1})^\top$ that maximizes the variance of $\mathbf{pc}_1$:

$$\max_{\mathbf{a}_1} \; Var(\mathbf{pc}_1) = Var(\mathbf{X}\mathbf{a}_1) = \mathbf{a}_1^\top \Sigma \mathbf{a}_1$$

subject to the normalization constraint $\mathbf{a}_1^\top \mathbf{a}_1 = \sum_{k=1}^{p} a_{k1}^2 = 1$.

The constraint is required for identifiability reasons, otherwise we could maximize the variance by just increasing the values in $\mathbf{a}_1$.

Repeat this maximization step for the other PCs and additionally use the orthogonality constraint, i.e. for the second PC:

$$\mathbf{a}_2^\top \mathbf{a}_1 = 0.$$

EXKURS: HAUPTKOMPONENTENANALYSE (PCA)

Gegeben: n Datenpunkte mit jeweils p Merkmalen $^{(*)}$

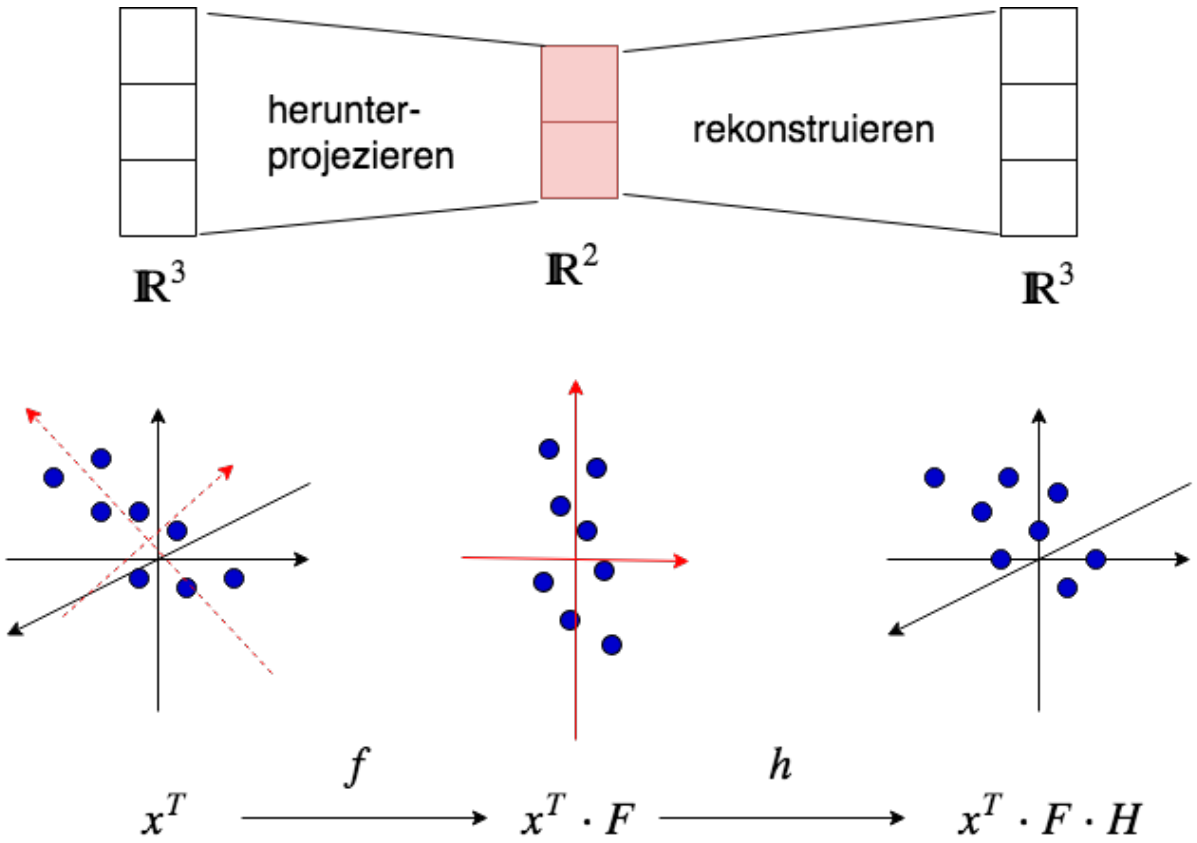
Ziel: Projektion der n Datenpunkte in einen k-dimensionalen Raum ($k < p$) mit möglichst wenig Informationsverlust

Idee:

- Finde eine lineare Abbildung $f : \mathbb{R}^p \to \mathbb{R}^k$, die jeden Datenpunkt $\boldsymbol{x}$ auf einen k-dimensionalen Punkt $\boldsymbol{z}$ abbildet.
- Verliere dabei so wenig wie möglich Information.
- Möglichst wenig Information geht verloren, wenn wir den Punkt $\boldsymbol{z}$ möglichst gut rekonstruieren können, d. h. wir können eine lineare Funktion $h : \mathbb{R}^k \to \mathbb{R}^p$ finden, sodass $\boldsymbol{x} \approx h(\boldsymbol{z})$.

$^{(*)}$ Wir nehmen an, die Datenpunkte sind um 0 zentriert.

EXKURS: HAUPTKOMPONENTENANALYSE (PCA)



Die linearen Abbildungen f, h beschreiben wir durch die Multiplikation mit Matrizen: $f : \boldsymbol{x}^T \mapsto \boldsymbol{x}^\top \boldsymbol{F} =: \boldsymbol{z}$ und $h : \boldsymbol{z}^\top \mapsto \boldsymbol{z}^\top \boldsymbol{H}$

EXKURS: HAUPTKOMPONENTENANALYSE (PCA)

Ziel: Minimiere Rekonstruktionsfehler zwischen Daten $\boldsymbol{X}$ und den projezierten und rekonstruierten Daten $\boldsymbol{XFH}$.

$$\min_{\boldsymbol{F} \in \mathbb{R}^{m \times k}, \boldsymbol{H} \in \mathbb{R}^{k \times n}} \|\mathbf{X} - \boldsymbol{XFH}\|_F^2.$$

Definieren wir $\boldsymbol{XF} =: \boldsymbol{W}$, so entspricht das gerade dem Problem der Matrixapproximation. Die Lösung ist somit gegeben durch

$$\begin{aligned} \boldsymbol{W} &= \boldsymbol{XF} = \mathbf{U}_k (\Sigma_k)^{1/2} \\ \boldsymbol{H} &= (\Sigma_k)^{1/2} \mathbf{V}_k^\top, \end{aligned}$$

wobei die Matrizen $\boldsymbol{U}_k \Sigma_k \boldsymbol{V}_k$ der trunkierten Singulärwertzerlegung von $\boldsymbol{X}$ entsprechen.

EXKURS: HAUPTKOMPONENTENANALYSE (PCA)

Die Spalten von $\mathbf{W}$ entsprechen den Hauptkomponenten (transformierte Variablen), die Zeilen von $\boldsymbol{H}$ sind die Hauptachsen.

Die Matrix $\boldsymbol{F}$, die die Funktion definiert, mithilfe derer die Punkte in den niedrigdimensionalen Raum projeziert werden, lässt sich herleiten

$$\begin{aligned} \boldsymbol{W} = \boldsymbol{X}\boldsymbol{F} &= \mathbf{U}_k(\Sigma_k)^{1/2} \\ \boldsymbol{U}_k\Sigma_k\boldsymbol{V}_k\boldsymbol{F} &\approx \mathbf{U}_k(\Sigma_k)^{1/2} \qquad | \boldsymbol{V}_k^{-1}\Sigma_k^{-1}\boldsymbol{U}_k^{-1} \cdot \\ \boldsymbol{F} &= \boldsymbol{V}_k^T(\Sigma_k)^{1/2} \end{aligned}$$

(was einer Drehung $\boldsymbol{V}_k^T$ und einer Skalierung $(\Sigma_k)^{1/2}$ entspricht).

EXAMPLE: THE OLYMPIC HEPTATHLON DATA

The `heptathlon` data set in the R package **HSAUR3** contains the competition results of 25 athletes in 7 disciplines for the Olympics held in Seoul in 1988.

Aim: Rank the athletes according to their overall performance in all 7 disciplines.

Idea: Use PCA to reduce the dimensionality (i.e., reduce the results of the 7 disciplines to one dimension) and compare the scores of the first PC with the official scores.

EXAMPLE: THE OLYMPIC HEPTATHLON DATA

Variables of the `heptathlon` data:

- `hurdles`: results 100m hurdles (in seconds).
- `highjump`: results high jump (in m).
- `shot`: results shot putt (in m).
- `run200m`: results 200m race (in seconds).
- `longjump`: results long jump (in m).
- `javelin`: results javelin (in m).
- `run800m`: results 800m race (in seconds).
- `score`: total score of the official scoring system.

EXAMPLE: THE OLYMPIC HEPTATHLON DATA

The variables `hurdles`, `run200m` and `run800m` are time measurements, i.e. low values are better. For all other variables high values are better.

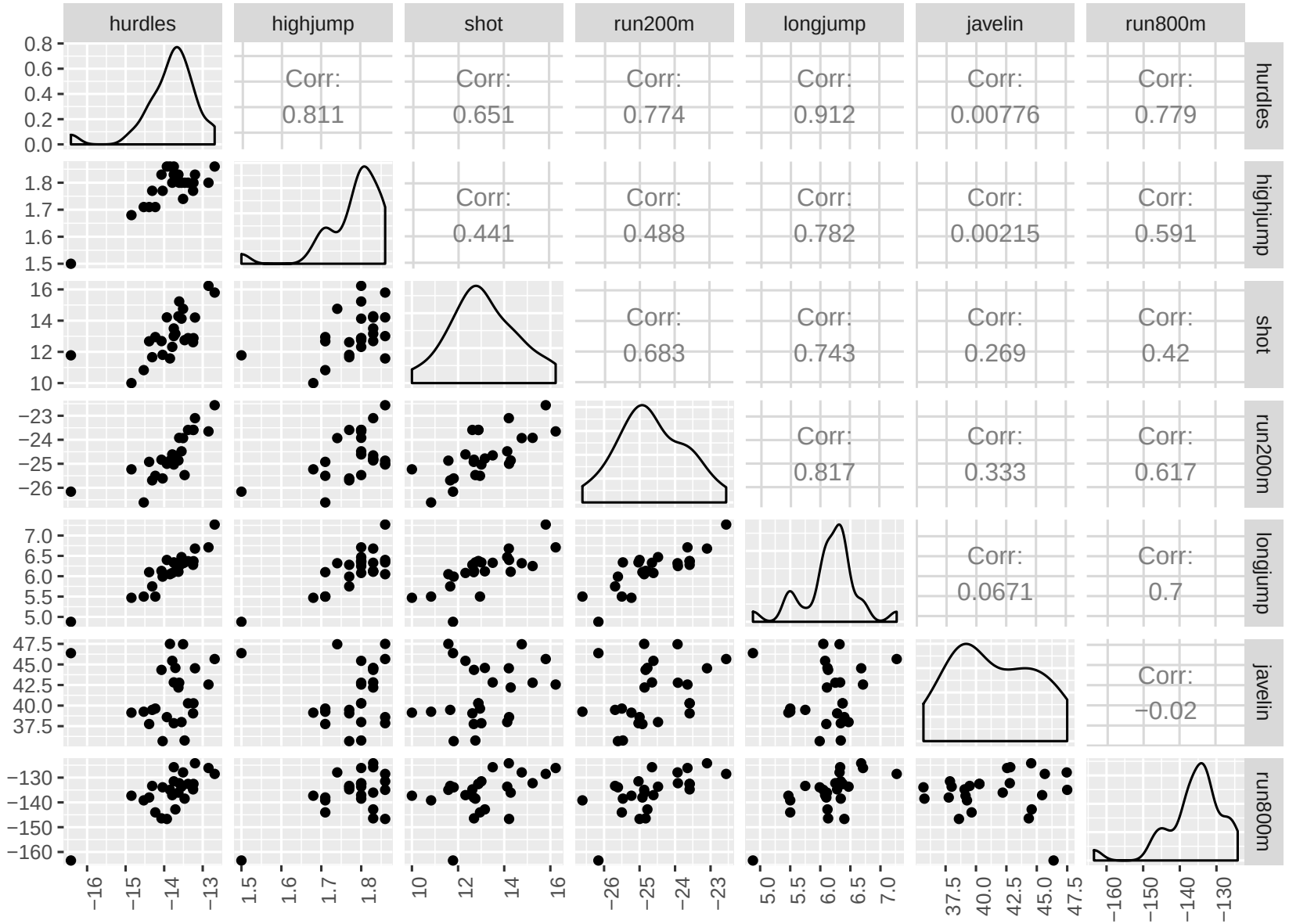
Results of the best and worst participant:

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersee (USA)	12.7	1.86	15.8	22.6	7.27	45.7	129	7291
Launa (PNG)	16.4	1.50	11.8	26.2	4.88	46.4	163	4566

We use negative time measurements so that higher values are better and therefore all variables have the same direction:

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersee (USA)	-12.7	1.86	15.8	-22.6	7.27	45.7	-129	7291
Launa (PNG)	-16.4	1.50	11.8	-26.2	4.88	46.4	-163	4566

SCATTER PLOT MATRIX

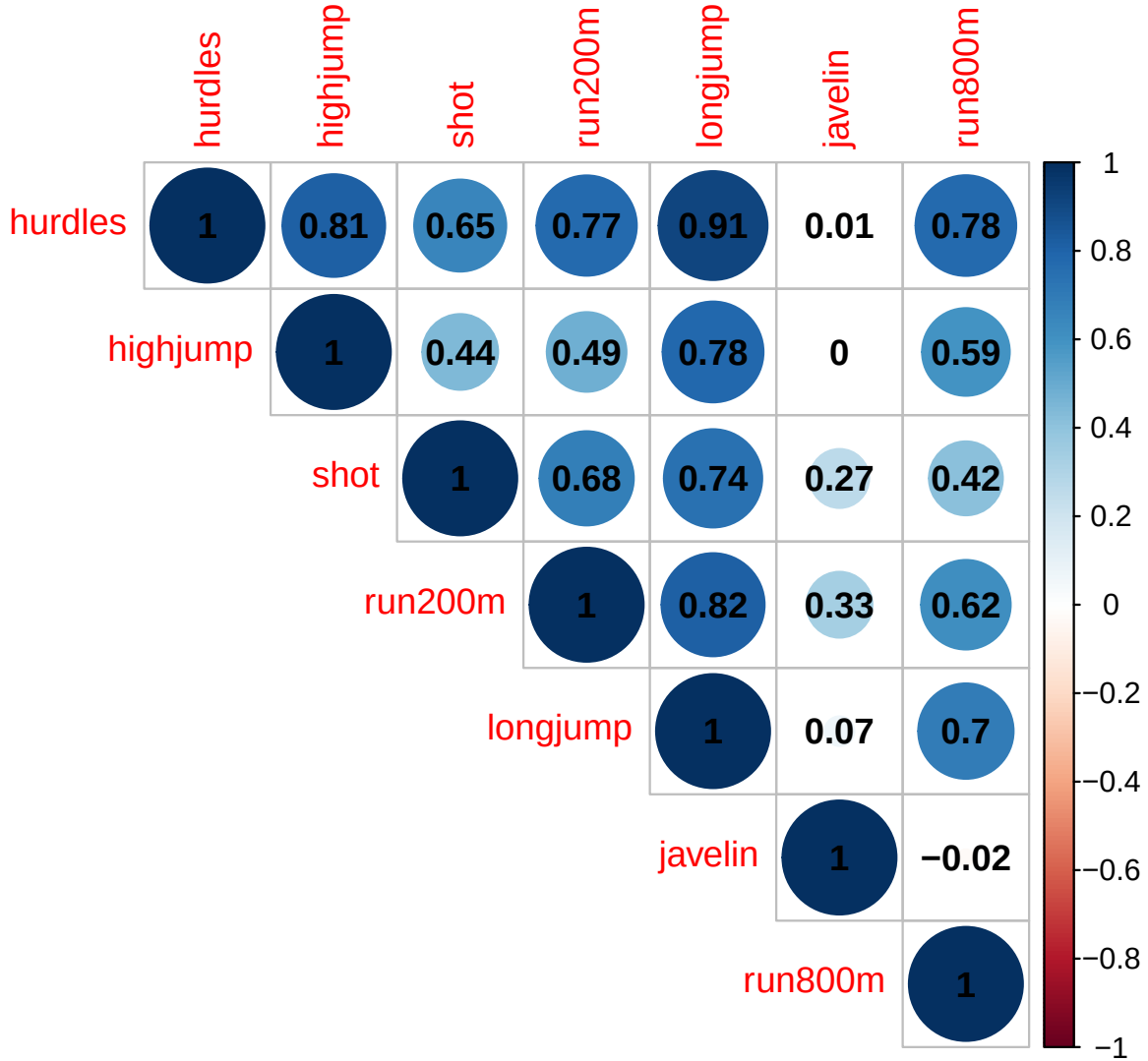


CORRELATION MATRIX

We can compute all pairwise correlations of the variables (without the `score` column):

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1.000	0.811	0.651	0.774	0.912	0.008	0.779
highjump	0.811	1.000	0.441	0.488	0.782	0.002	0.591
shot	0.651	0.441	1.000	0.683	0.743	0.269	0.420
run200m	0.774	0.488	0.683	1.000	0.817	0.333	0.617
longjump	0.912	0.782	0.743	0.817	1.000	0.067	0.700
javelin	0.008	0.002	0.269	0.333	0.067	1.000	-0.020
run800m	0.779	0.591	0.420	0.617	0.700	-0.020	1.000

CORRELOGRAM



PERFORMING THE PCA

Remember: The first PC is the linear combination

$$\mathbf{pc}_1 = a_{11}\mathbf{x}_1 + a_{12}\mathbf{x}_2 + \cdots + a_{1p}\mathbf{x}_p$$

and has the largest sample variance among all other PCs.

- If variables are on very different scales, PCA should be carried out on the correlation matrix (which is equivalent to the correlation matrix if normalized variables are used).
- As the variables of the heptathlon data are on different scales, we perform the PCA based on the correlation matrix.
- Alternatively, we could also perform the PCA based on the covariance matrix but on the normalized heptathlon data.

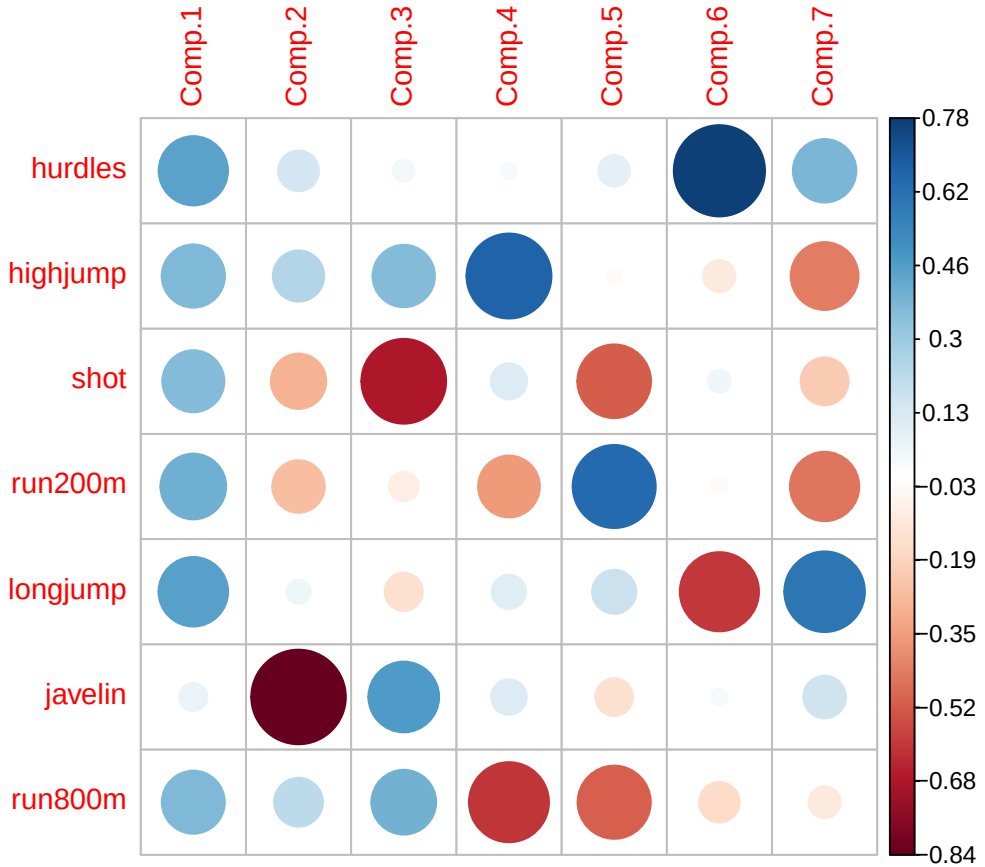
LOADINGS

The loadings $\mathbf{a}_1, \ldots, \mathbf{a}_p$ are given by

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
hurdles	0.45	0.16	0.05	0.03	0.09	0.78	0.38
highjump	0.38	0.25	0.37	0.68	-0.02	-0.10	-0.43
shot	0.36	-0.29	-0.68	0.12	-0.51	0.05	-0.22
run200m	0.41	-0.26	-0.08	-0.36	0.65	-0.02	-0.45
longjump	0.46	0.06	-0.14	0.11	0.18	-0.59	0.61
javelin	0.08	-0.84	0.47	0.12	-0.14	0.03	0.17
run800m	0.37	0.22	0.40	-0.60	-0.50	-0.16	-0.10

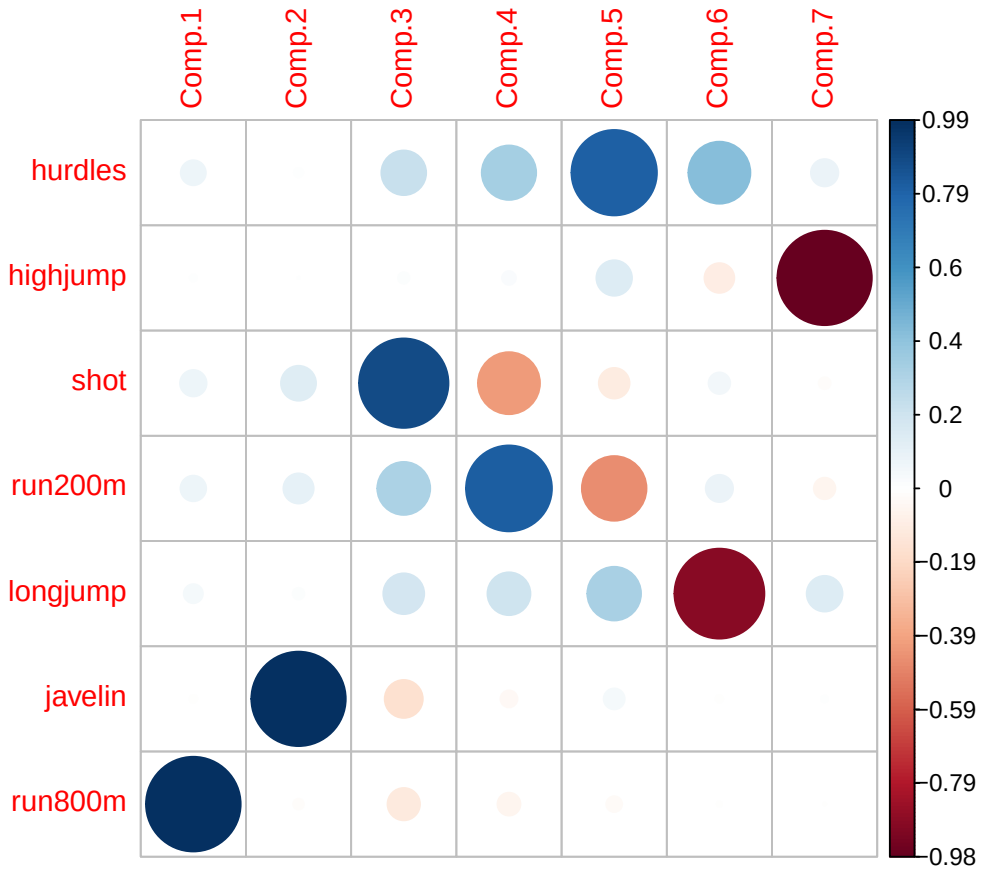
LOADINGS

Visualize the coefficient weights (loadings) of the linear combinations of the PC scores:



LOADINGS

If we perform a PCA on the covariance matrix (without normalizing the data), each component mainly loads on a single variable:



Reason: Variables have very different scales (e.g., time measurement of 200m and 800m run).

PROPORTION OF EXPLAINED VARIANCE

- The total variance of the p PC scores is equal the total variance of the original variables, i.e.,

$$\sum_{j=1}^{p} \lambda_j = s_1^2 + s_2^2 + \cdots + s_p^2,$$

  where λ_j is the variance of the jth PC and s_j^2 is the sample variance of variable $\mathbf{x}_j$.

- The proportion of explained variance of the j-th PC is

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

- The first k PCs account for a proportion

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

PROPORTION OF EXPLAINED VARIANCE

In the example above, the proportion of explained variance is given by

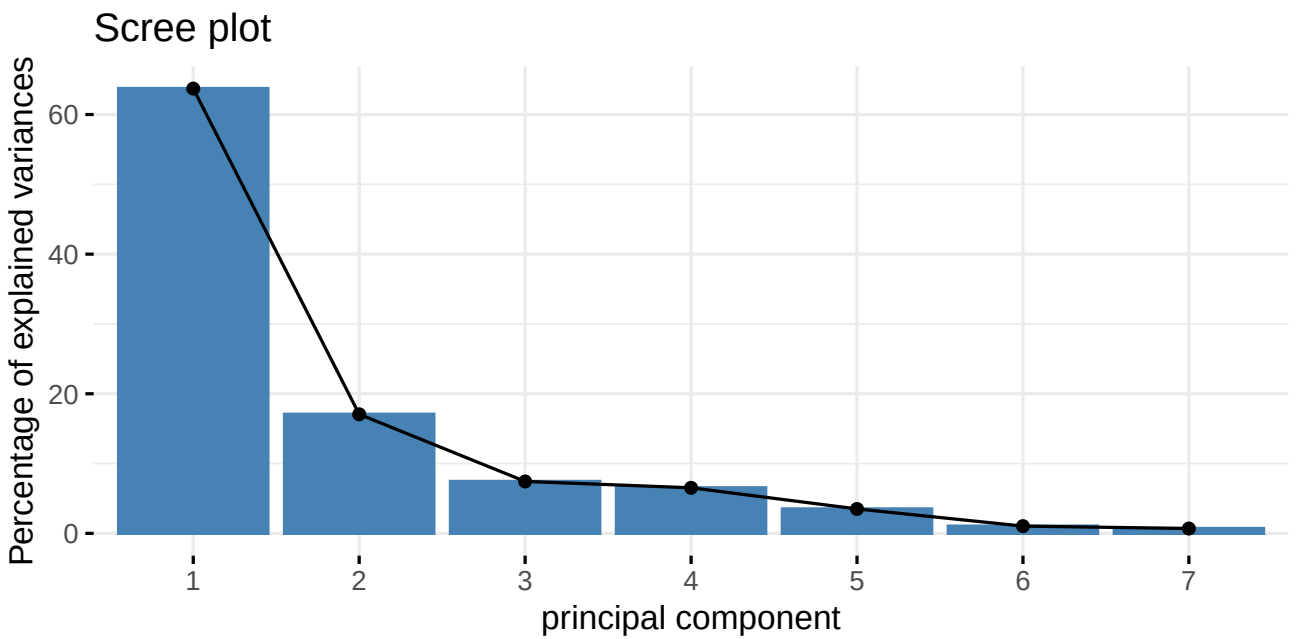
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	2.112	1.093	0.722	0.676	0.495	0.270	0.221
Proportion of Variance	0.637	0.171	0.074	0.065	0.035	0.010	0.007
Cumulative Proportion	0.637	0.808	0.882	0.948	0.983	0.993	1.000

Question: How do we choose the number of PCs?

CHOOSING THE NUMBER OF PCS

Two simple rules of thumb for choosing the number of PCs:

1. Retain the first k components, which explain a large proportion of the total variation, e.g., 70-80%.
2. Use a scree plot: Plot the component variances vs. the component number and look for an *elbow*. For components after the *elbow*, the variance decreases more slowly.



PC SCORES VS. OFFICIAL SCORES

The first PC explains $63,72\%$ of the variation, the loadings of the first PC are:

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
Comp.1	0.453	0.377	0.363	0.408	0.456	0.075	0.375

Dimensionality reduction:

- Project all 8 features onto the first PC.
- Compare the scores of the first PC with the official scores used to rank the athletes.

PC SCORES VS. OFFICIAL SCORES

The scores of the first PC $\mathbf{pc}_1$ have a similar ranking as the scores of the official scoring system, i.e., we can reduce the dimension to the first PC without losing much information: