



Principal Component Analysis (PCA) (aka Hauptkomponentenanalyse)

Unsupervised Learning

- Supervised Learning: Erkläre Zielgrösse durch erklärende Variablen
Ergebnis kann validiert werden (Fehlerrate, Kreuzvalidierung)
- Unsupervised Learning: Finde “interessante” Strukturen in Daten (z.B. Gruppen); es gibt keine Zielgrösse
Ergebnis kann nicht validiert werden; subjektiv

Beispiel 1: Visualisieren

- Wie visualisiert man hochdimensionale (>3) Datensätze ?

```
> head(USArrests)
      Murder  Assault UrbanPop  Rape
Alabama   13.2     236       58  21.2
Alaska    10.0     263       48  44.5
Arizona    8.1     294       80  31.0
Arkansas   8.8     190       50  19.5
California 9.0     276       91  40.6
Colorado   7.9     204       78  38.7
```

-
-
-



Beispiel 2: Komprimieren

- Wie komprimiert man viele Variablen in wenige Variablen, die die Daten gut beschreiben?

	Gen 1	Gen 2	...	Gen 6829	Gen 6830
Person 1	1.3	4.3		3.1	9.2
Person 2	8.2	5.5		3.2	5.8
...					



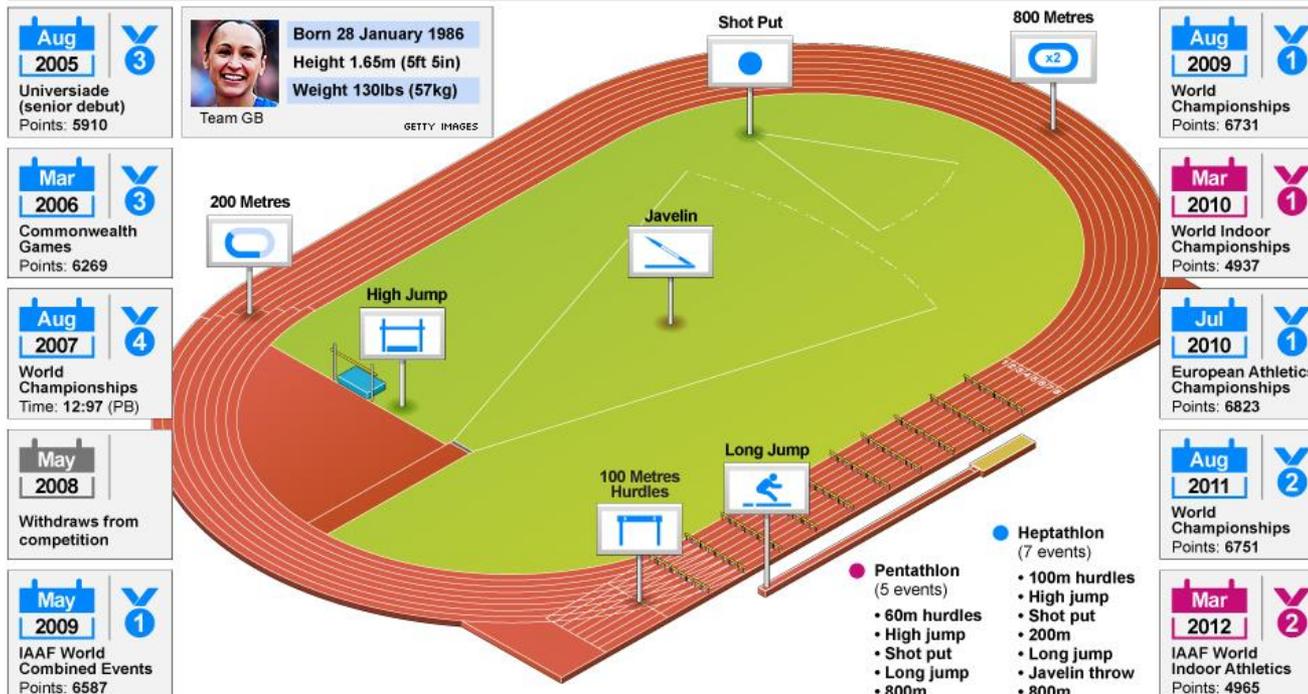
	Z1	Z2	...	Z9	Z10
Person 1	1.3	4.3		3.1	9.2
Person 2	8.2	5.5		3.2	5.8
...					



Beispiel 3: Unterscheiden

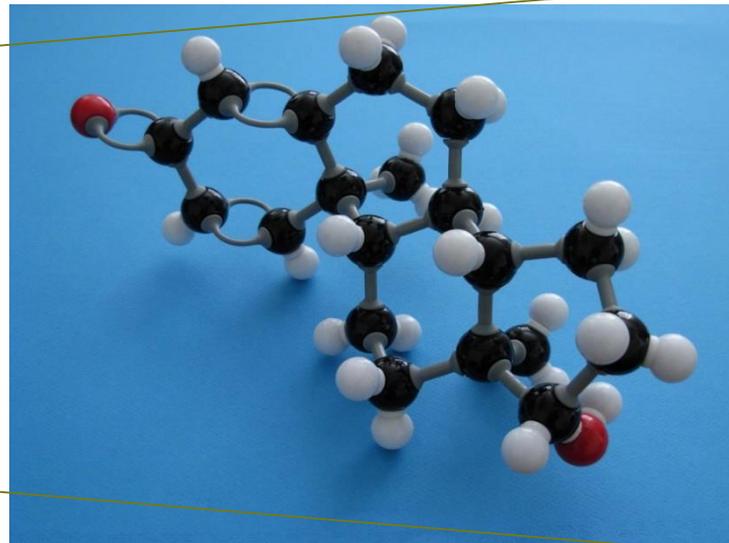
Wie erstellt man einen eindimensionalen Index, der Subjekte möglichst gut unterscheidet?

Jessica Ennis career history



Siebenkampf

PCA: “Gute” Projektion in wenige Dimensionen

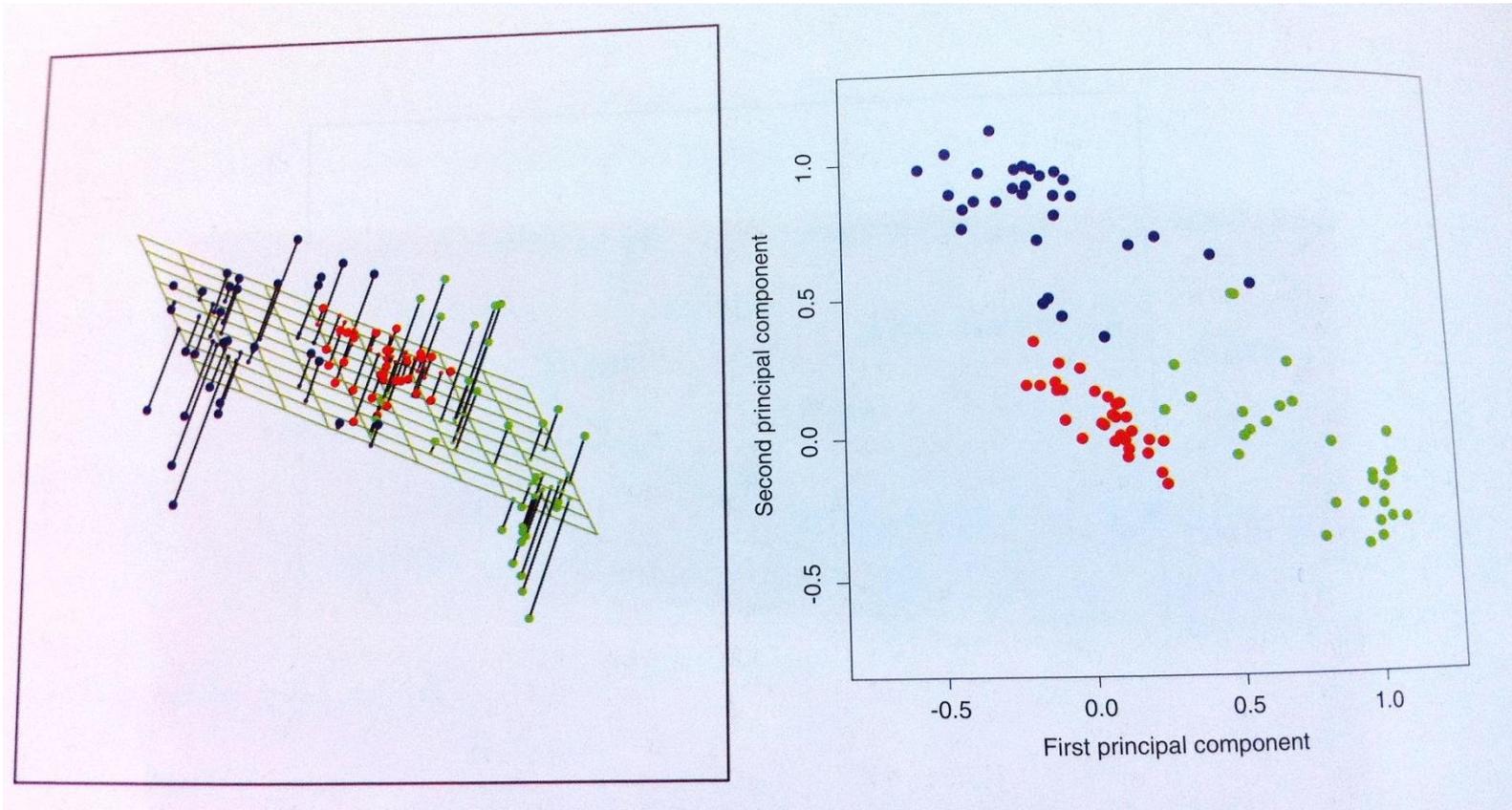


3D

2D

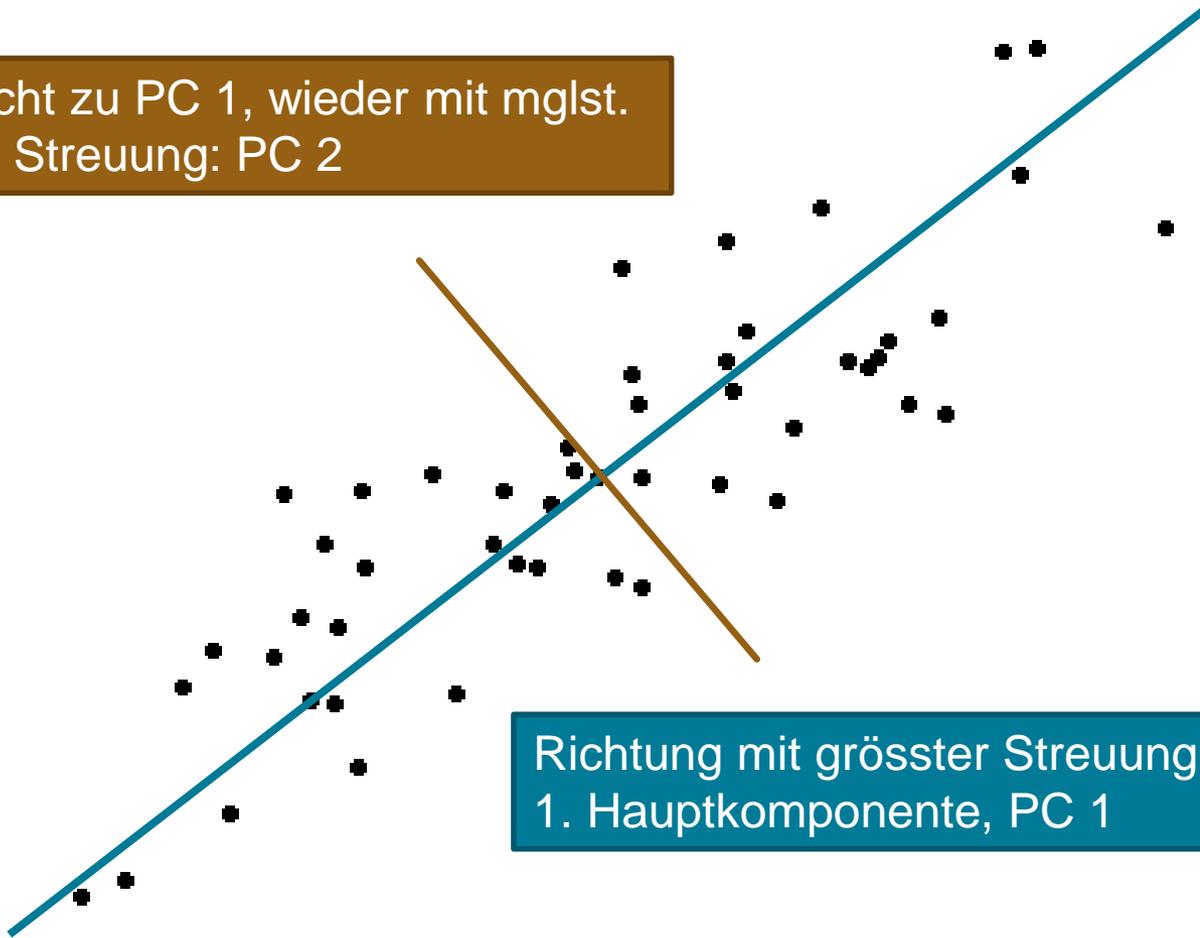
“Gut” = Möglichst viel Varianz

PCA: “Bester” Subraum (bzgl. Residuenquadratsumme)



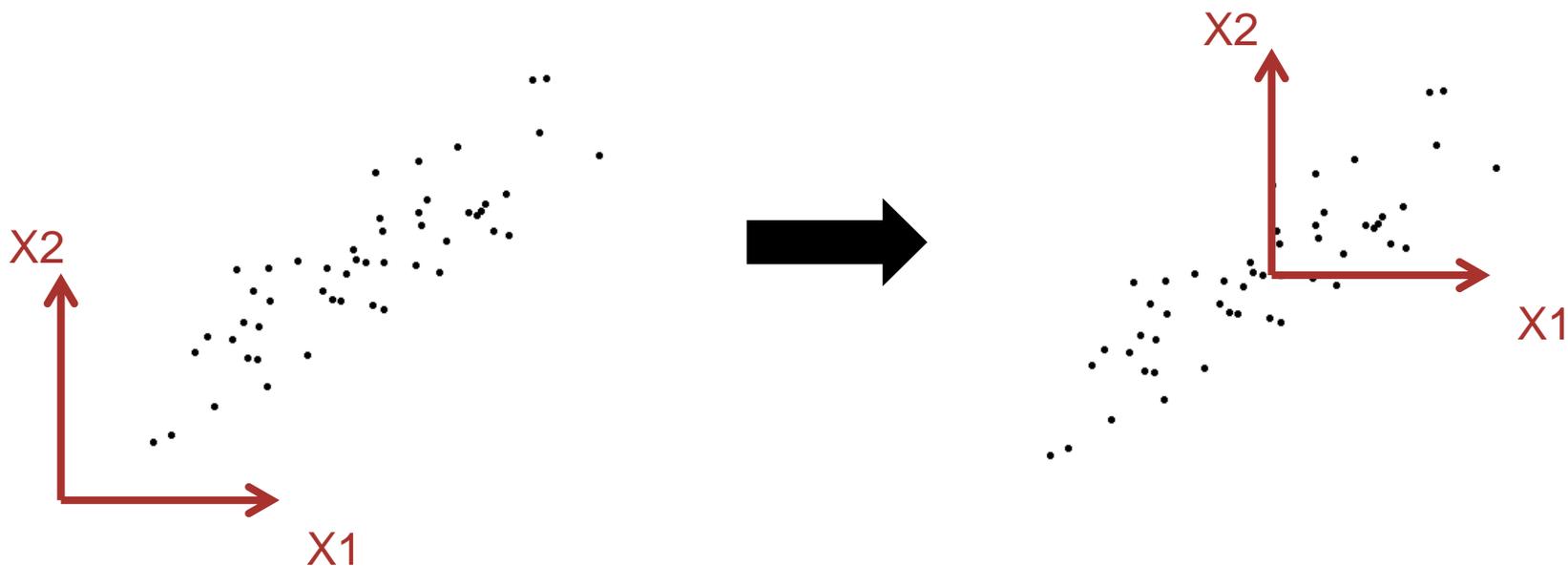
PCA: Intuition

Senkrecht zu PC 1, wieder mit mglst. grosser Streuung: PC 2

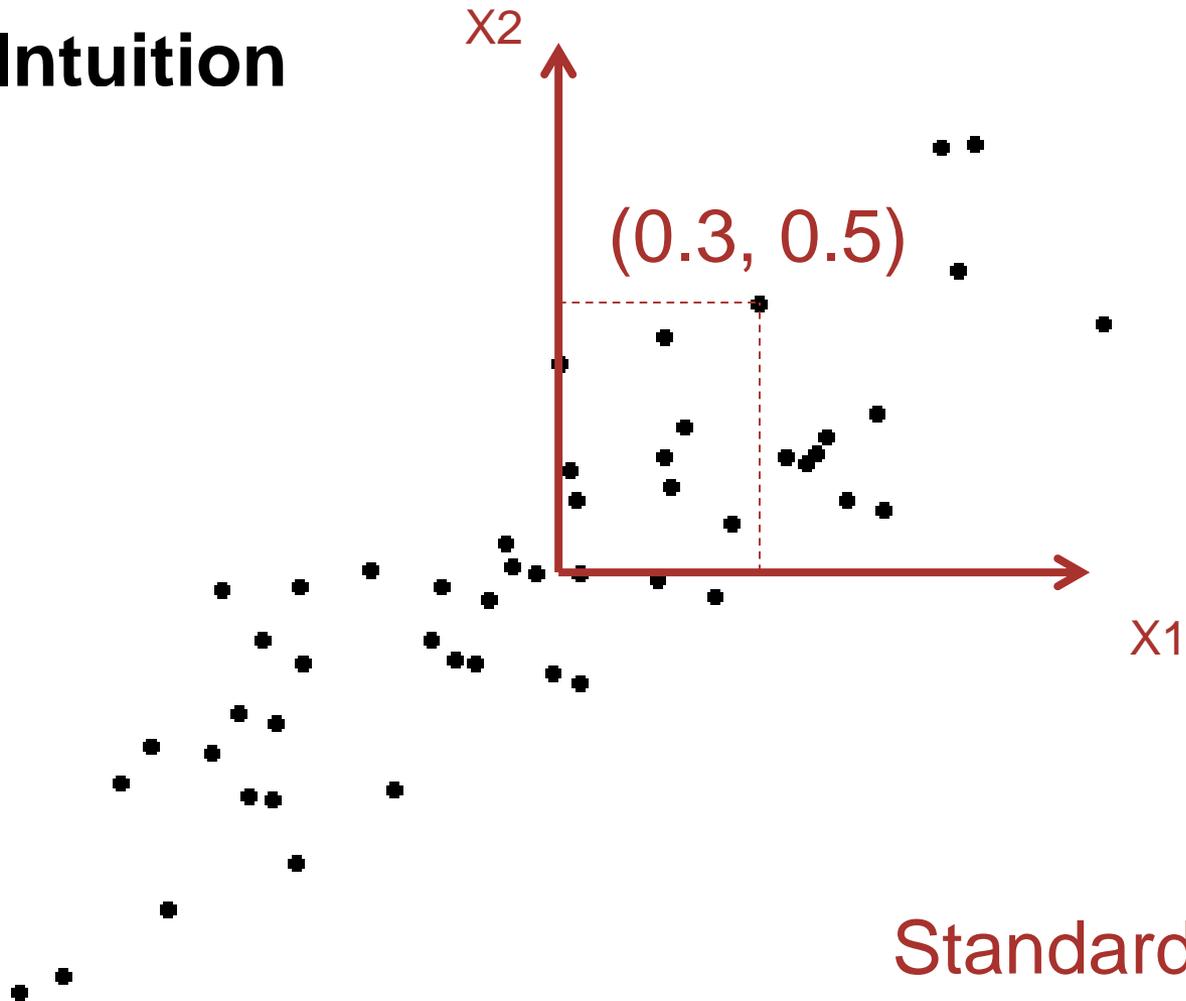


Richtung mit grösster Streuung:
1. Hauptkomponente, PC 1

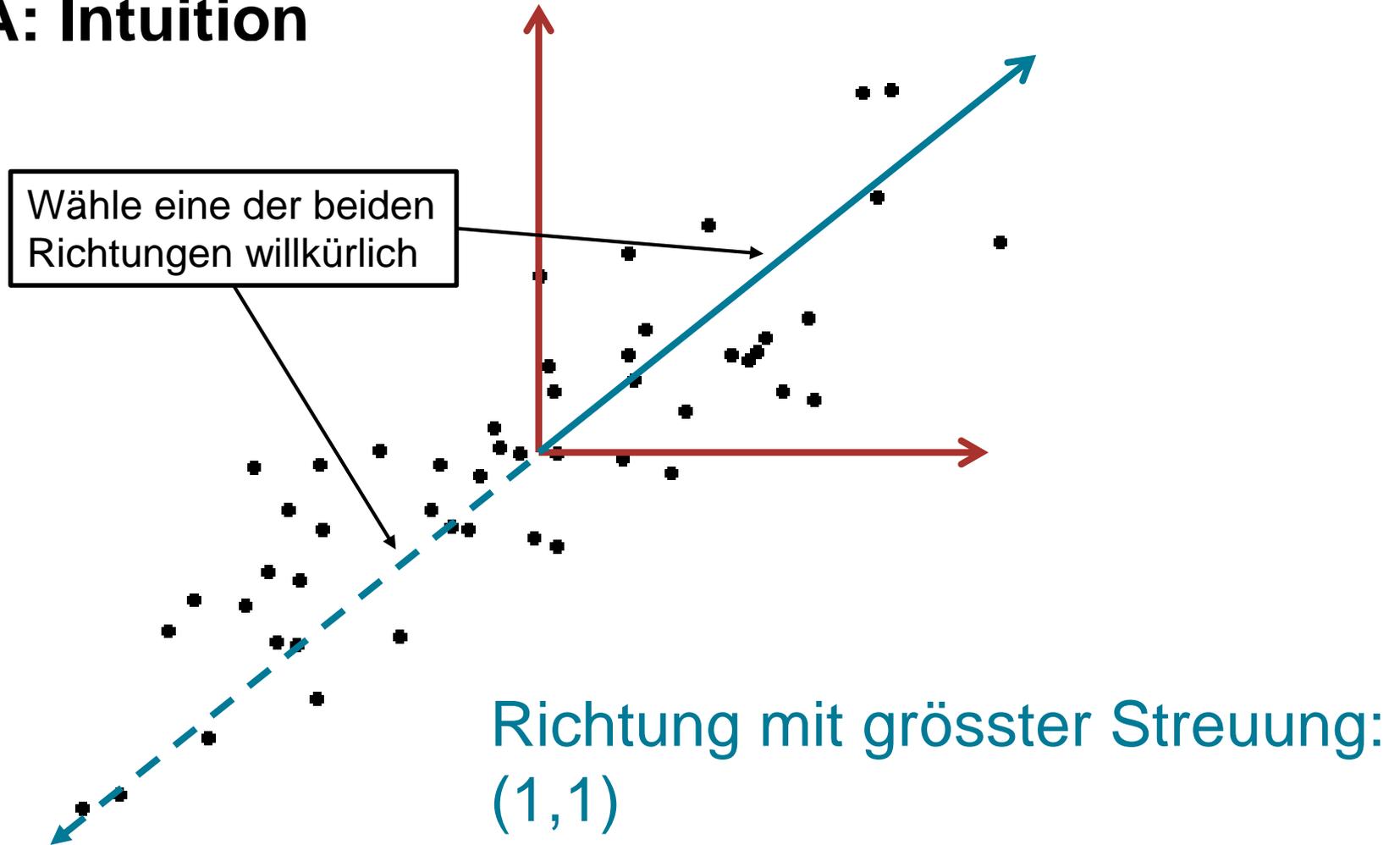
Konvention: Zentrieren



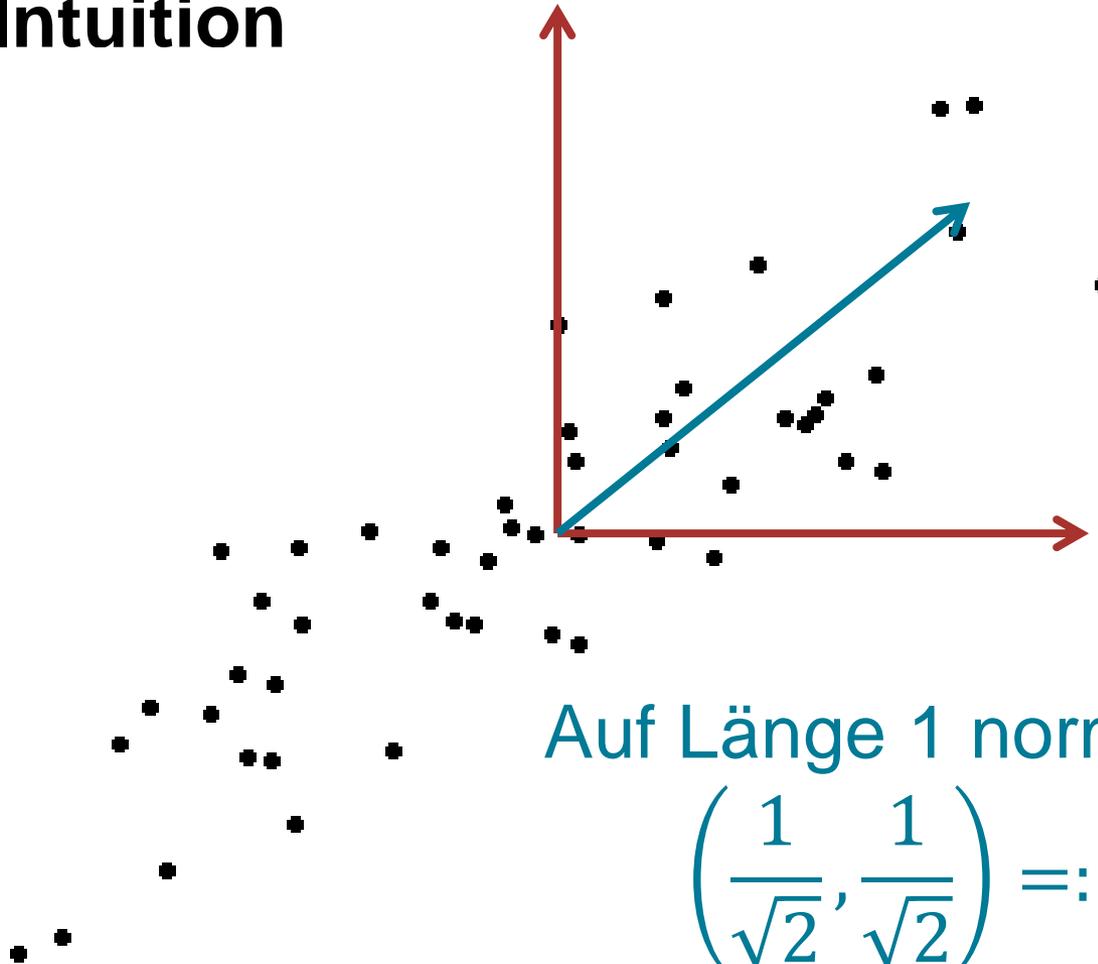
PCA: Intuition



PCA: Intuition



PCA: Intuition



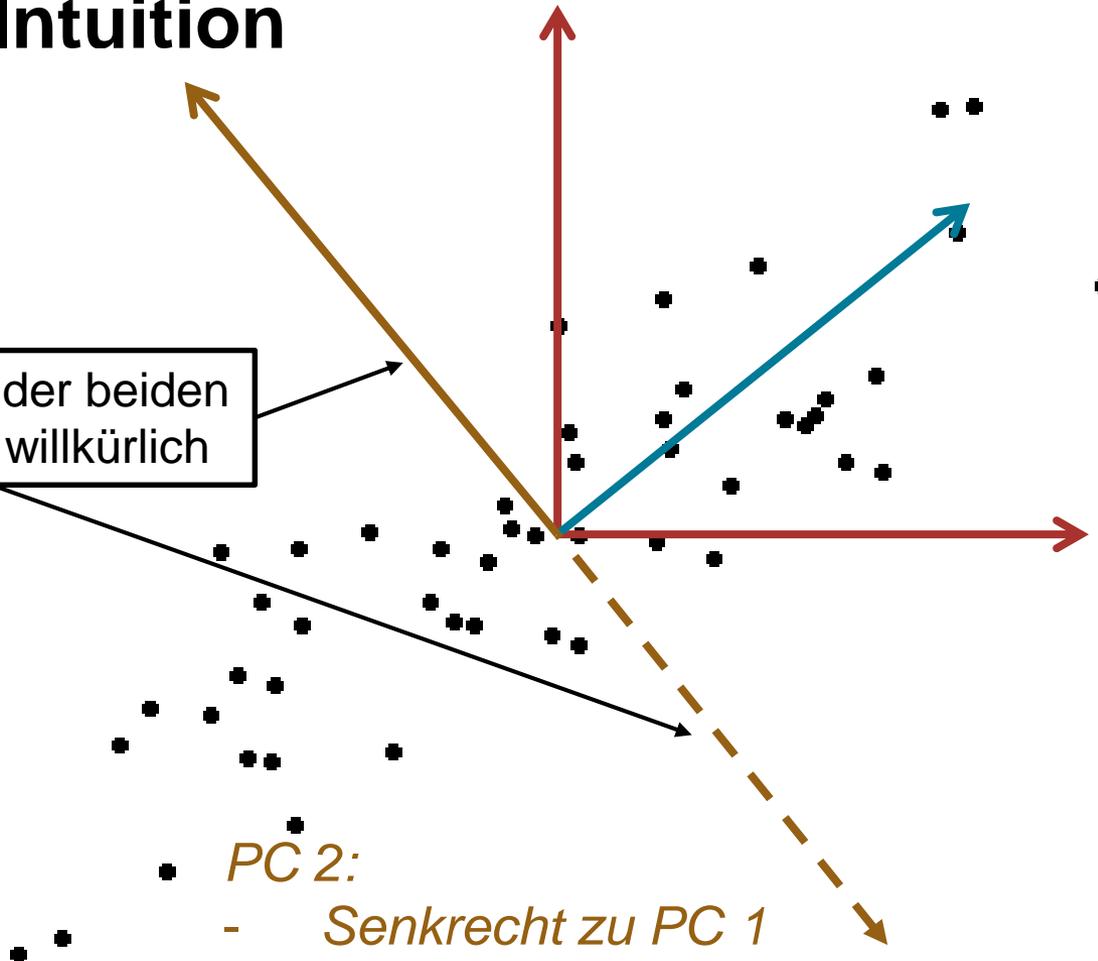
Auf Länge 1 normieren:

$$\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) =: (\phi_{11}, \phi_{21})$$

1. *Principal Component (PC 1)*

PCA: Intuition

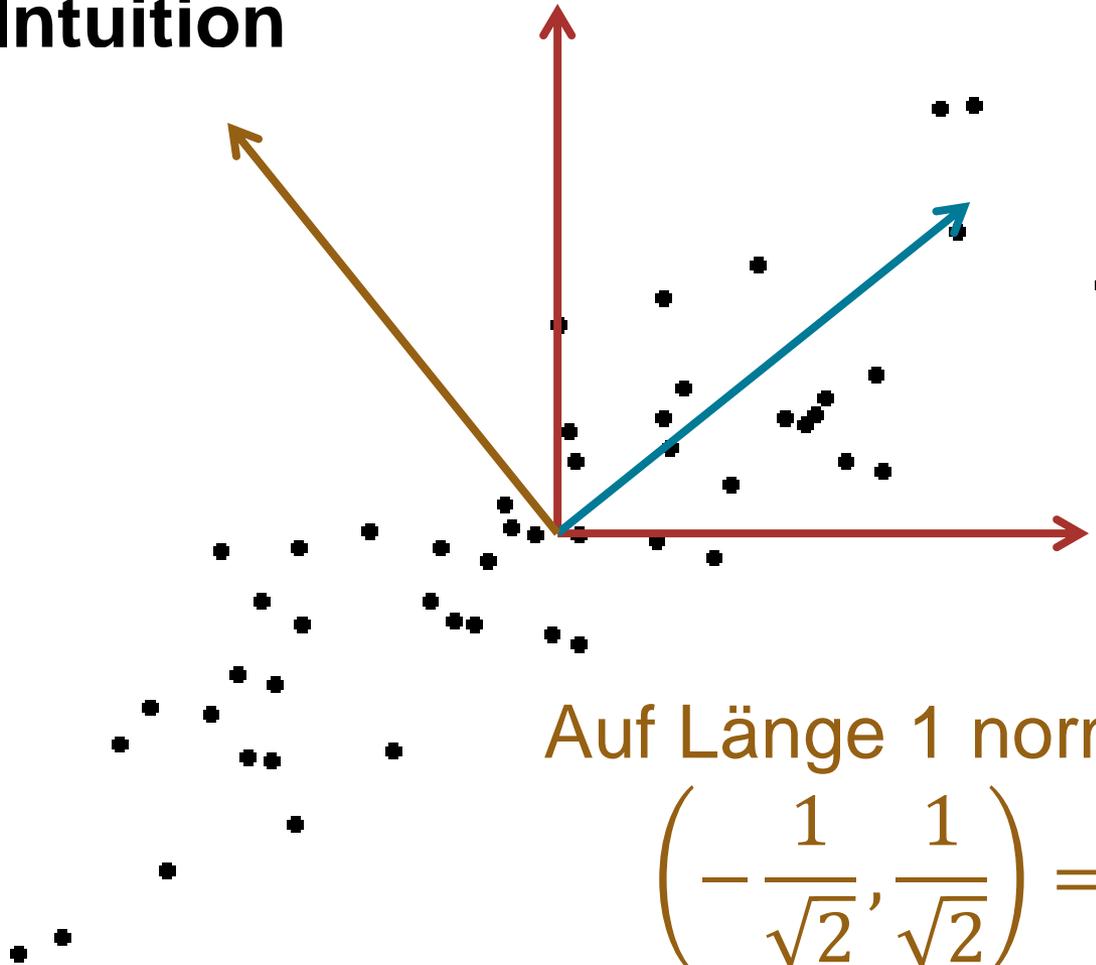
Wähle eine der beiden
Richtungen willkürlich



PC 2:

- *Senkrecht zu PC 1*
 - *Wieder Richtung mit grösster Streuung*
- In 2d gibt es nur noch zwei Möglichkeiten;
wähle eine willkürlich: $(-1, 1)$*

PCA: Intuition

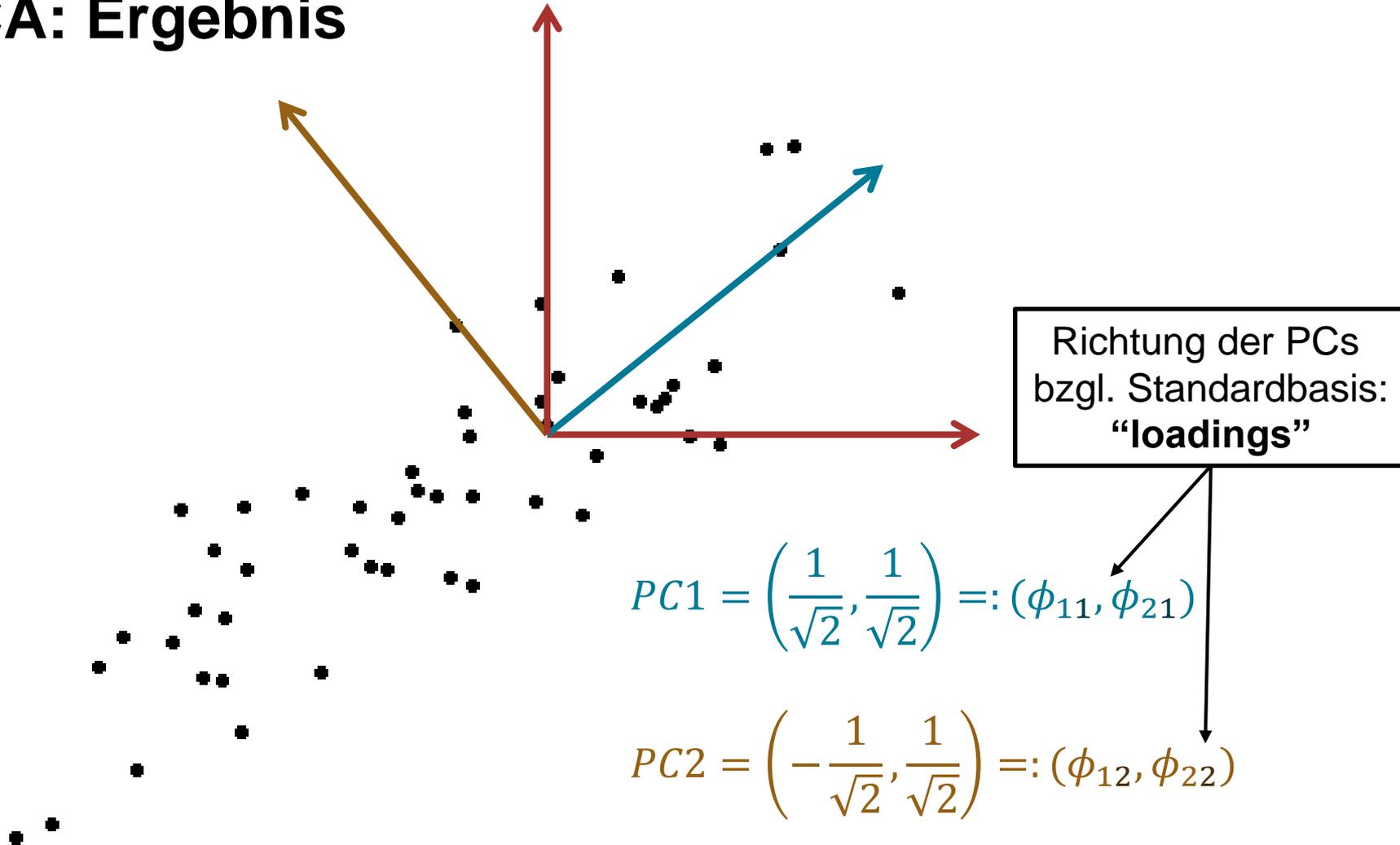


Auf Länge 1 normieren:

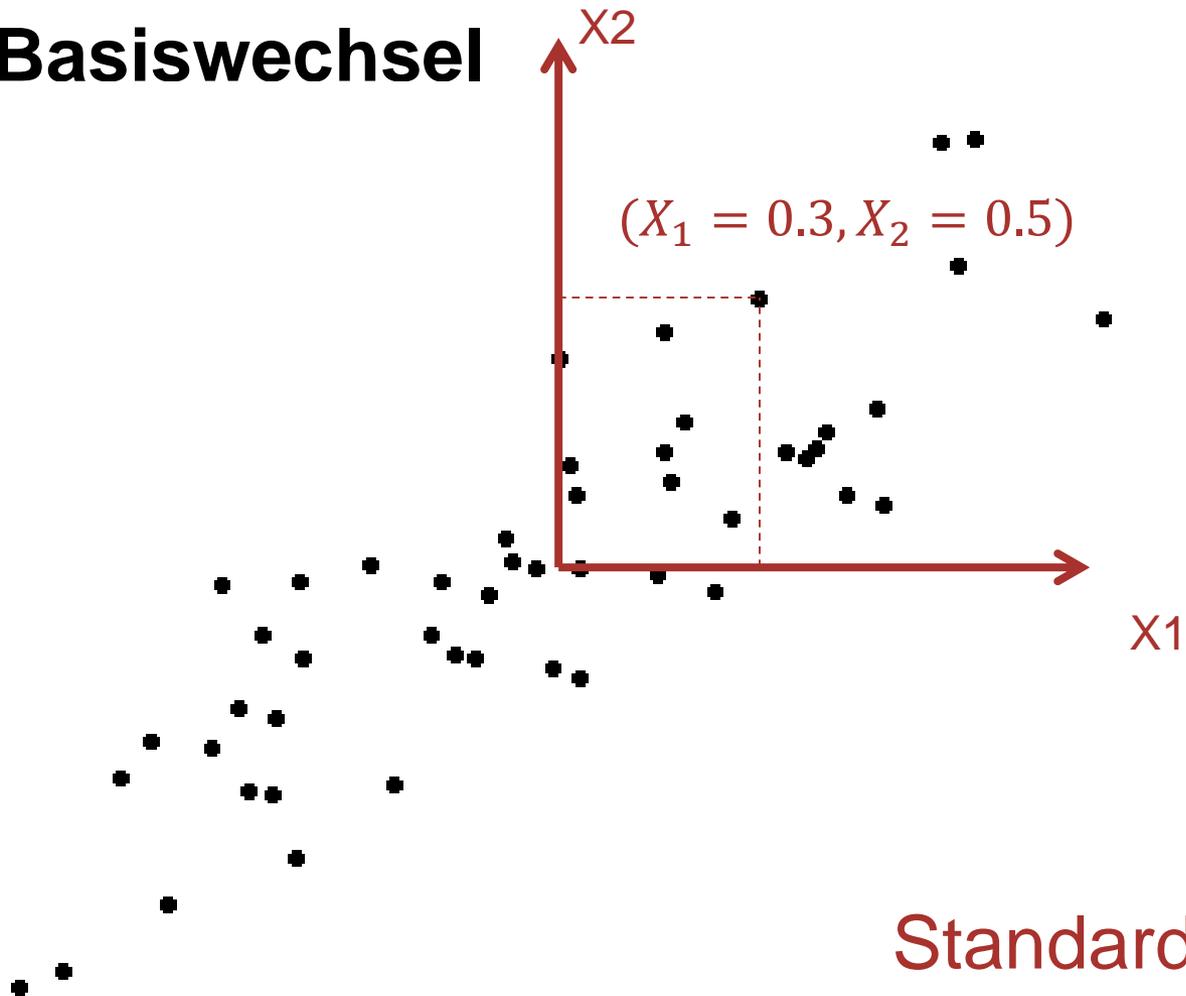
$$\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) =: (\phi_{12}, \phi_{22})$$

2. *Principal Component (PC 2)*

PCA: Ergebnis

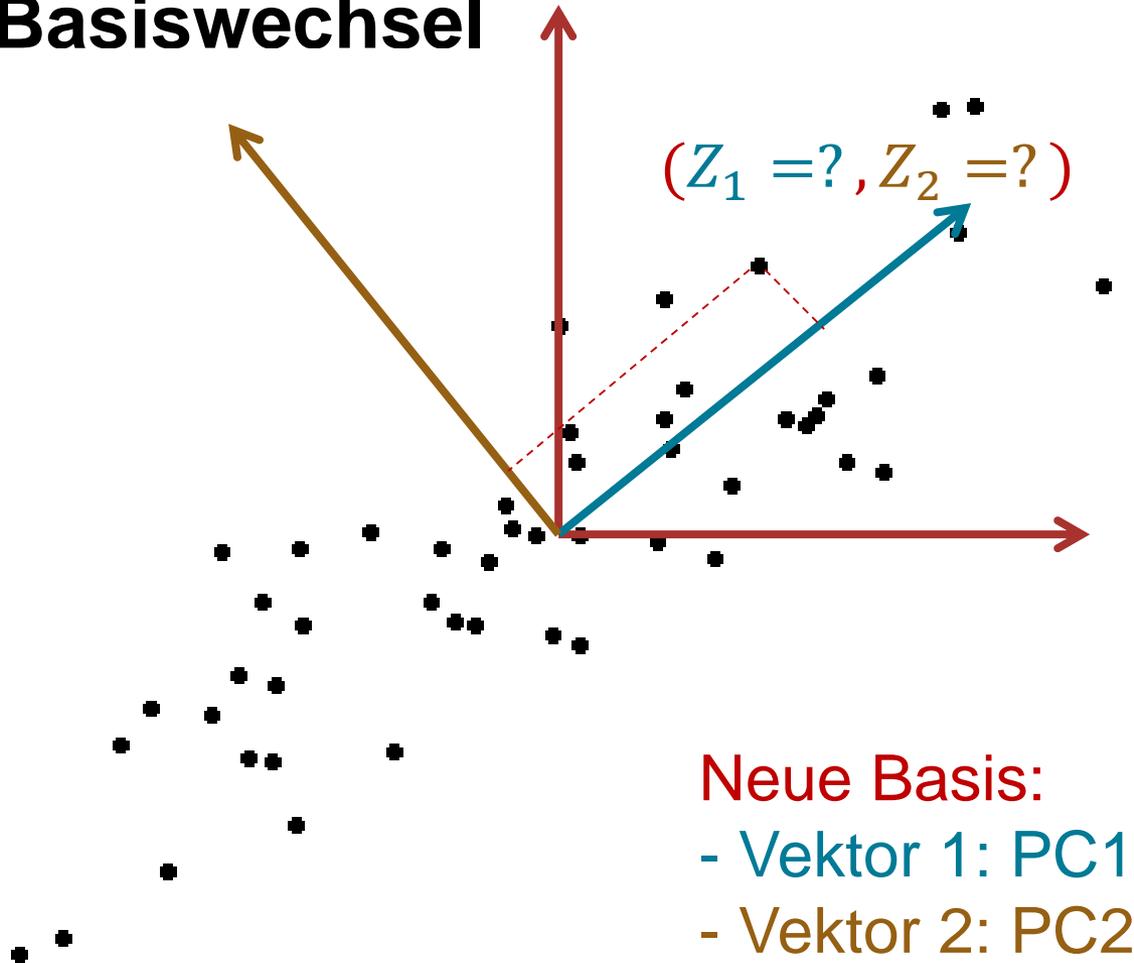


PCA: Basiswechsel

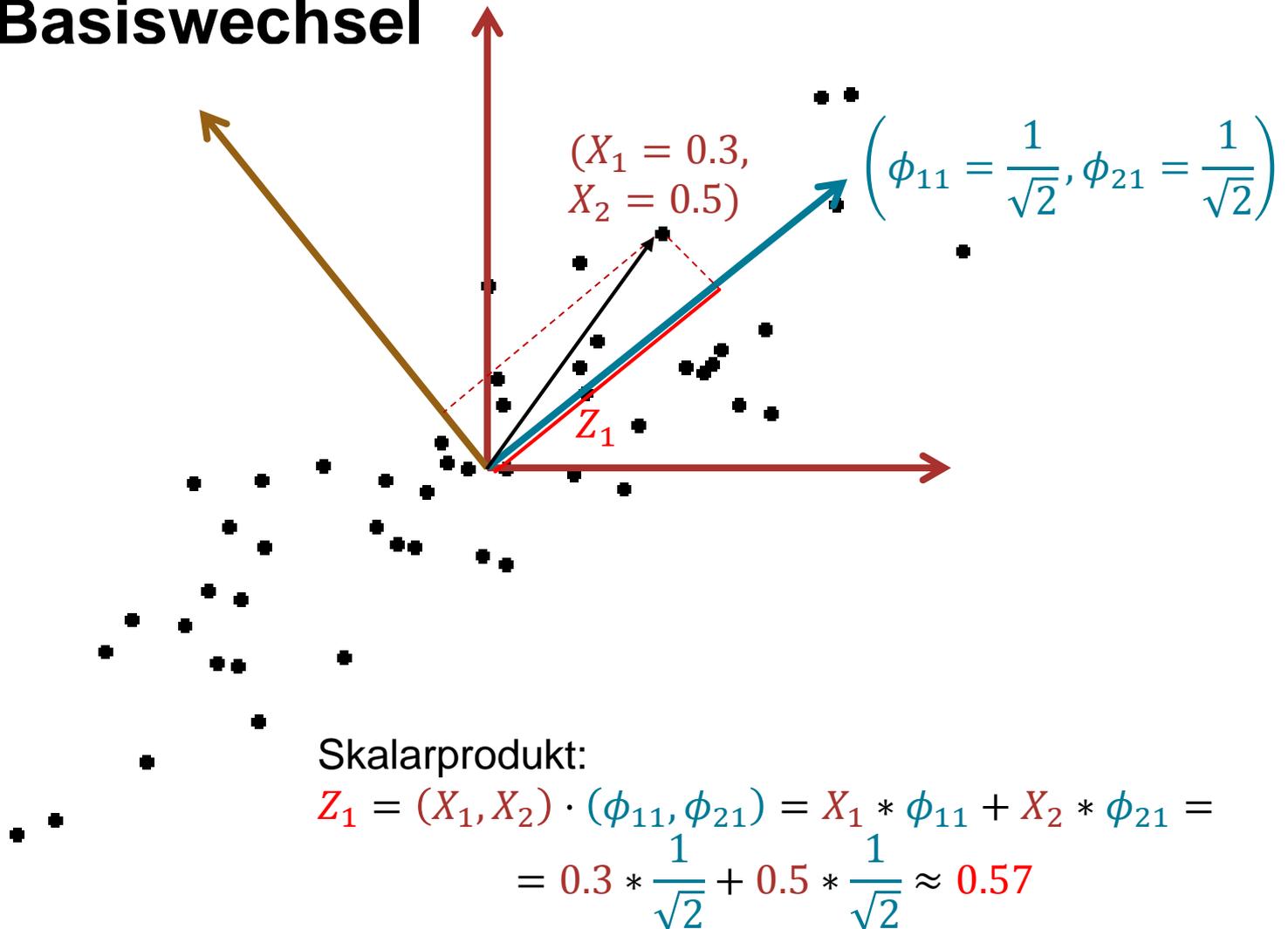


Standard basis

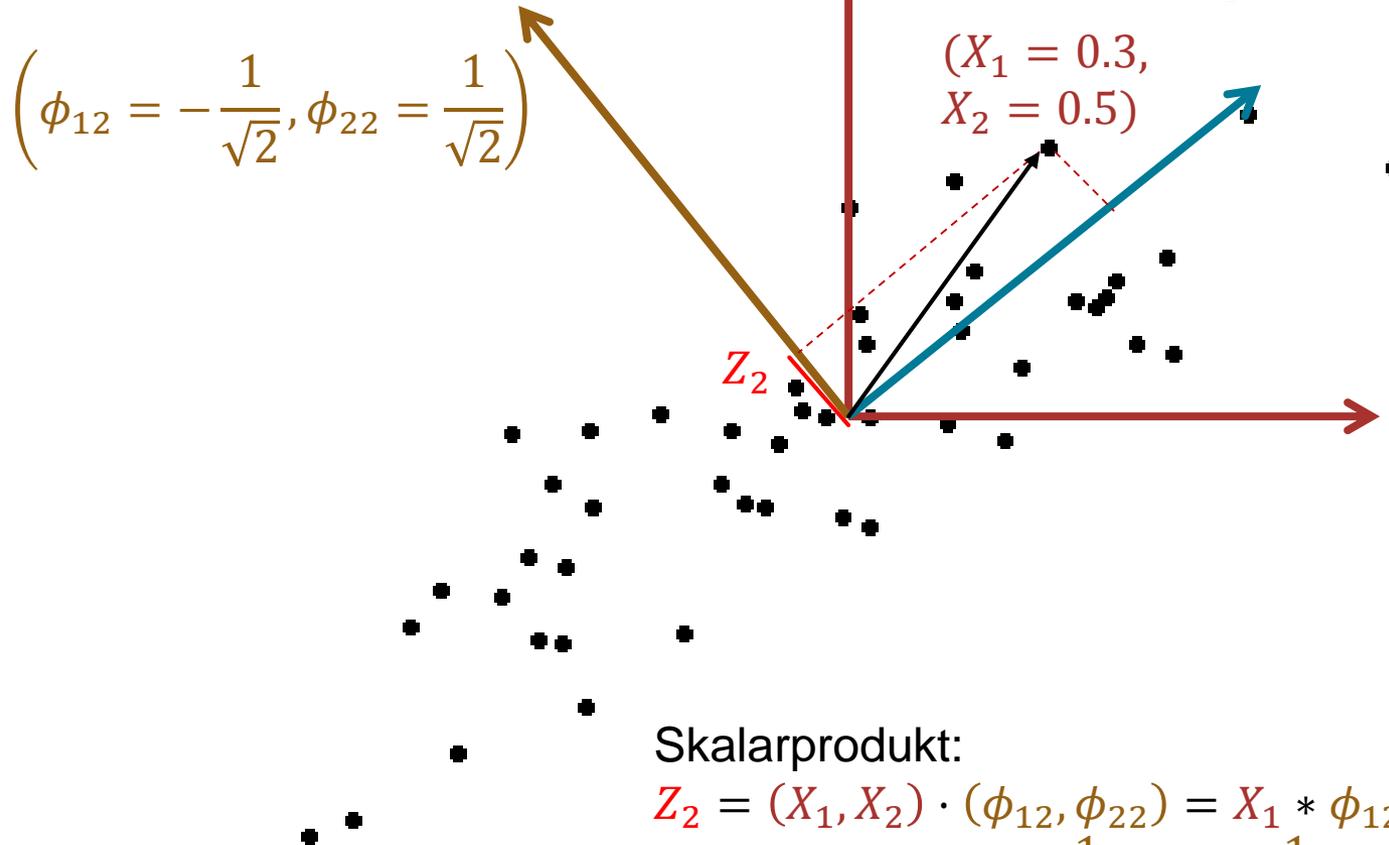
PCA: Basiswechsel



PCA: Basiswechsel



PCA: Basiswechsel

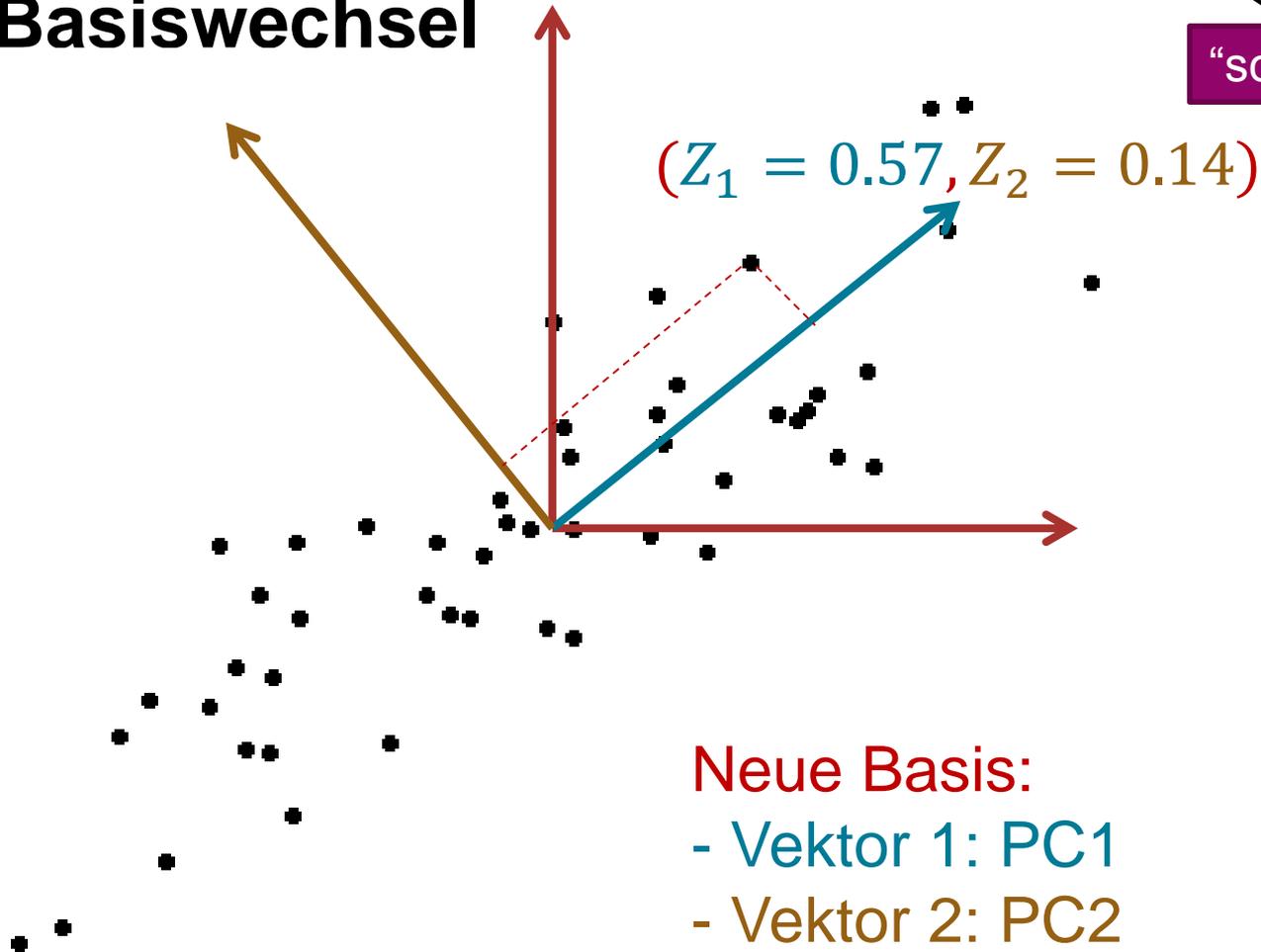


Skalarprodukt:

$$\begin{aligned}
 Z_2 &= (X_1, X_2) \cdot (\phi_{12}, \phi_{22}) = X_1 * \phi_{12} + X_2 * \phi_{22} = \\
 &= 0.3 * \frac{-1}{\sqrt{2}} + 0.5 * \frac{1}{\sqrt{2}} \approx \mathbf{0.14}
 \end{aligned}$$

	Koord. 1	Koord. 2
Std. Basis	$X_1 = 0.3$	$X_2 = 0.5$
PC Basis	$Z_1 = 0.57$	$Z_2 = 0.14$

PCA: Basiswechsel





Andere Koordinaten ?

Ihr Kollege hat unabhängig von Ihnen auch die PCA durchgeführt und berechnet nun die PC Koordinaten des Punktes (0.3, 0.5). Er bekommt bei Z_1 ein anderes Vorzeichen als Sie. Ist das ein Hinweis auf einen Fehler in seiner Analyse?

1. Ja
2. Nein

	Koord. 1	Koord. 2
Std. Basis	$X_1 = 0.3$	$X_2 = 0.5$
PC Basis 1	$Z_1 = 0.57$	$Z_2 = 0.14$
PC Basis 2	$Z_1 = -0.57$	$Z_2 = 0.14$

PCA: Basiswechsel mit Linearer Algebra

	Koord. 1	Koord. 2
Std. Basis	$X_1 = 0.3$	$X_2 = 0.5$
PC Basis	$Z_1 = 0.57$	$Z_2 = 0.14$

- Standard Basis und PC Basis sind je eine Orthonormal Basis (Achsen senkrecht, Länge 1)
- Basiswechsel: **Rotation**smatrix Φ
- Spalten der Rotationsmatrix sind *loadings*:

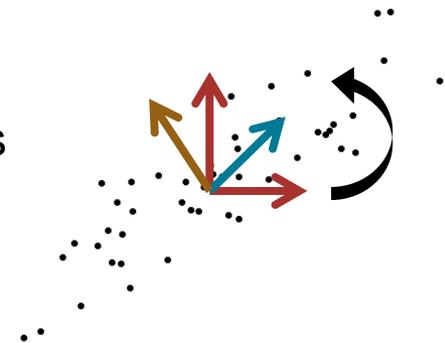
$$\Phi = \begin{pmatrix} \text{PC1} & \text{PC2} \\ 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{matrix} X1 \\ X2 \end{matrix}$$

- Basiswechsel mit Rotationsmatrix ist einfach:
 Φ : Von PC Basis nach Standardbasis
 Φ^{-1} : Von Standardbasis nach PC Basis

Bzgl. Std.basis

$$\Phi^{-1} = \begin{pmatrix} X1 & X2 \\ 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{matrix} \text{PC1} \\ \text{PC2} \end{matrix} ; Z = \Phi^{-1} * X = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} * \begin{pmatrix} 0.3 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.57 \\ 0.14 \end{pmatrix}$$

Bzgl. PC Basis
"scores"



Wie findet man 1.PC - Mathematik

- Zentriere Daten
- Angenommen, 1. PC ist in Richtung

$$\Phi_1 = (\phi_{11}, \phi_{21}, \dots, \phi_{p1})$$
- Betrachte Datenpunkt i :
Koordinaten bzgl. Standardbasis $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$
- Neue erste Koordinate von Datenpunkt x_i :

$$z_{i1} = \Phi_1 * x_i = \phi_{11} * x_{i1} + \phi_{21} * x_{i2} + \dots + \phi_{pi} * x_{ip}$$
- Koordinaten bzgl. PC-Basis $z_i = (z_{i1}, z_{i2}, \dots, z_{ip})$
- Kriterium für 1. PC (vgl. Gleichung (10.3) in ISLR):

$$\max_{\phi_{11}, \dots, \phi_{p1}} \left(\widehat{Var}(z_{i1}) \right) \text{ sodass Länge von } \Phi_1 = 1$$



Wie findet man 1. PC - Numerik

- Singulärwertzerlegung der Kovarianzmatrix (= Singular Value Decomposition, SVD)
- Schlechtere Alternative: Eigenwertzerlegung der Kovarianzmatrix

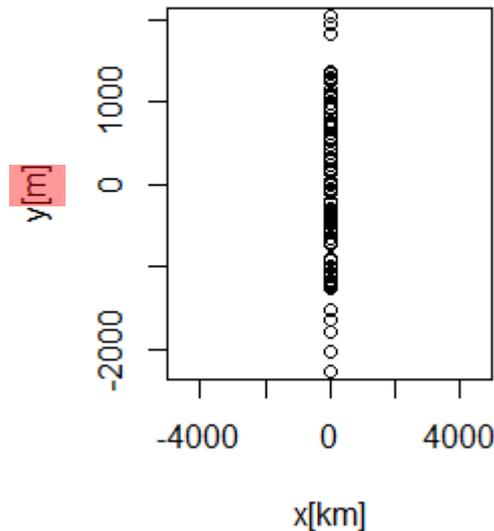
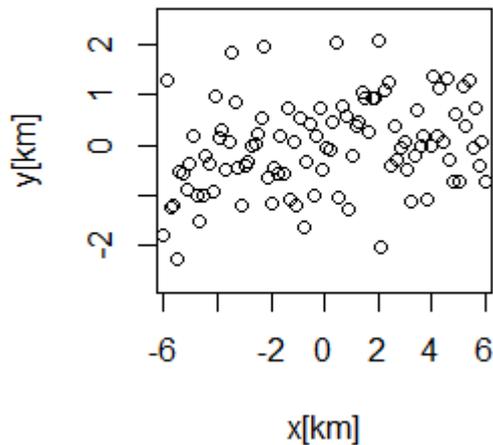
empfohlen

In R:

- Funktion *prcomp* verwendet **Singulärwertzerlegung** der Kovarianzmatrix
- Funktion *princomp* verwendet Eigenwertzerlegung der Kovarianzmatrix

To scale or not to scale ...

Messungen auf einer Landkarte (z.B. Bodenschätze)



Welche Einheiten ?

To scale or not to scale ...

Faustregeln:

- Daten immer **zentrieren**
- Falls alle Variablen in der **gleichen Einheit** sind: **Nicht skalieren**
- Falls Variablen in **unterschiedlichen Einheiten** sind: **Skalieren**



Beispiel 1: Visualisierung

```
> head(USArrests)
      Murder  Assault UrbanPop  Rape
Alabama   13.2    236      58  21.2
Alaska    10.0    263      48  44.5
Arizona    8.1    294      80  31.0
Arkansas   8.8    190      50  19.5
California 9.0    276      91  40.6
Colorado   7.9    204      78  38.7
      ■
      ■
      ■
```



Beispiel 1: Interpretation der PCs

```
> pr.out$rotation
          PC1      PC2
Murder    -0.5358995  0.4181809
Assault   -0.5831836  0.1879856
UrbanPop  -0.2781909 -0.8728062
Rape      -0.5434321 -0.1673186
```

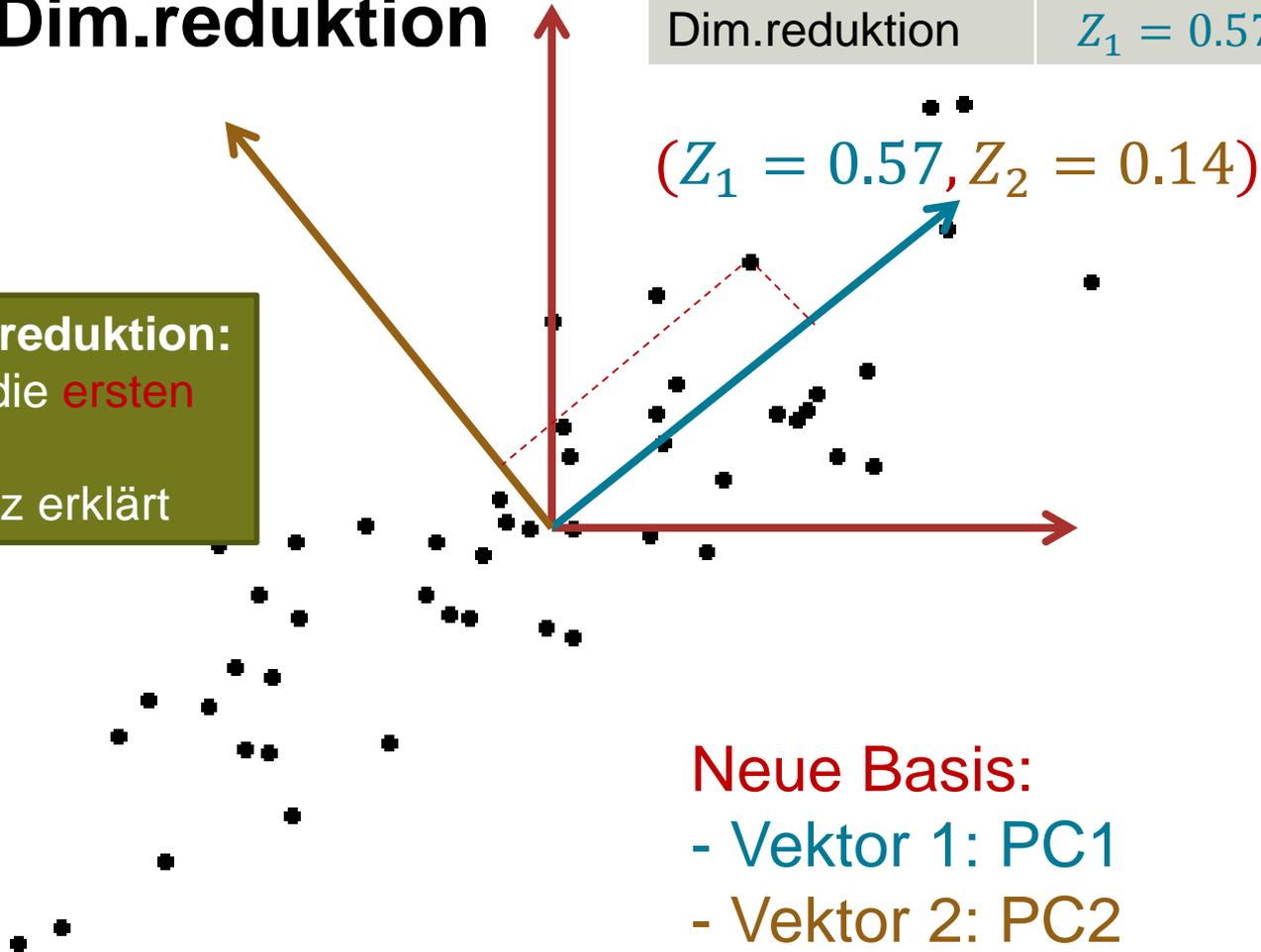
- PC 1 ist gross, wenn v.a. Murder, Assault und Rape klein sind
→ PC 1 spiegelt “Verbrechen” wieder
- PC 2 ist gross, wenn UrbanPop klein ist
→ PC 2 spiegelt “Verstädterung” wieder



	Koord. 1	Koord. 2
Std. Basis	$X_1 = 0.3$	$X_2 = 0.5$
PC Basis	$Z_1 = 0.57$	$Z_2 = 0.14$
Dim.reduktion	$Z_1 = 0.57$	-

PCA: Dim.reduktion

Dimensionsreduktion:
Behalte nur die **ersten paar** PC's
→ viel Varianz erklärt



Wie viele PCs?

- Es gibt so viele PCs wie es X-Variablen gibt

0 PCs
perfekt komprimiert
Varianz in Daten nicht erfasst



alle PCs
nicht komprimiert
Varianz in Daten perfekt erfasst

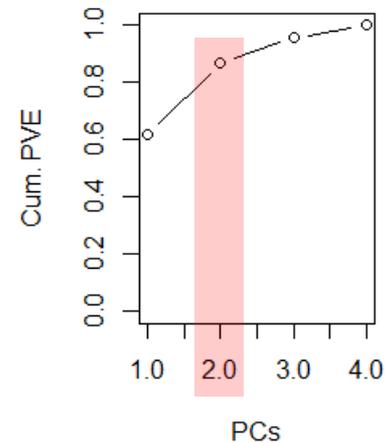
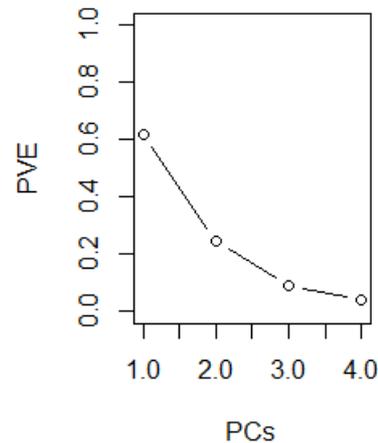
Kompromiss



Screeplot: Wie viele PCs bei USArrests?



- Ziel: Möglichst viel Varianz in den Daten erfassen
- Varianz entlang der PCs nimmt ab



Faustregel: 80% der Varianz erklären
→ 2 PCs in diesem Bsp

Beispiel 2: NCI60 Data

	Gen 1	Gen 2	...	Gen 6829	Gen 6830
Person 1	0.3	1.18	...	-0.34	-1.93
...					
Person 64	0.35	-0.27	...	-0.15	1.21



- 64 Krebszell-Linien; je 6830 Gene
- Wie fasst man die Anzahl Gene zusammen ?
- (Vgl. ISLR 10.6.1)



Beispiel 2: NCI60 Data

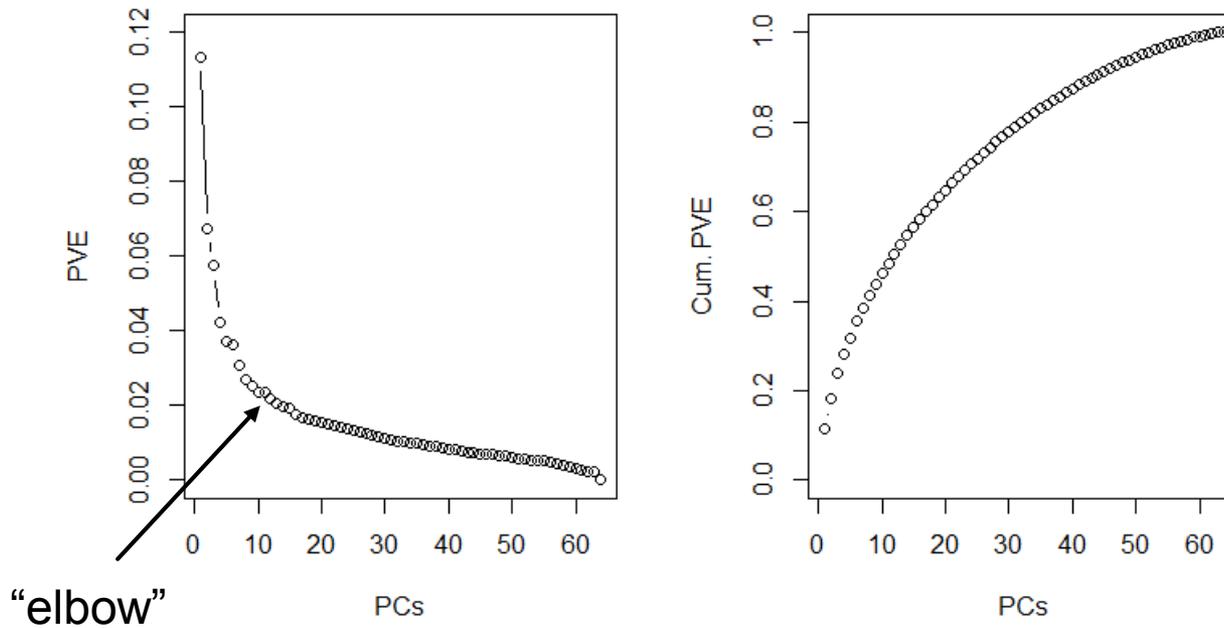
	Gen 1	Gen 2	...	Gen 6829	Gen 6830
Person 1	0.3	1.18	...	-0.34	-1.93
...					
Person 64	0.35	-0.27	...	-0.15	1.21



	Z1	Z2	...	Z9	Z10
Person 1	1.3	4.3		3.1	9.2
...					
Person 64	8.2	5.5		3.2	5.8

Wie viele PCs?

Beispiel 2: Scree-Plot

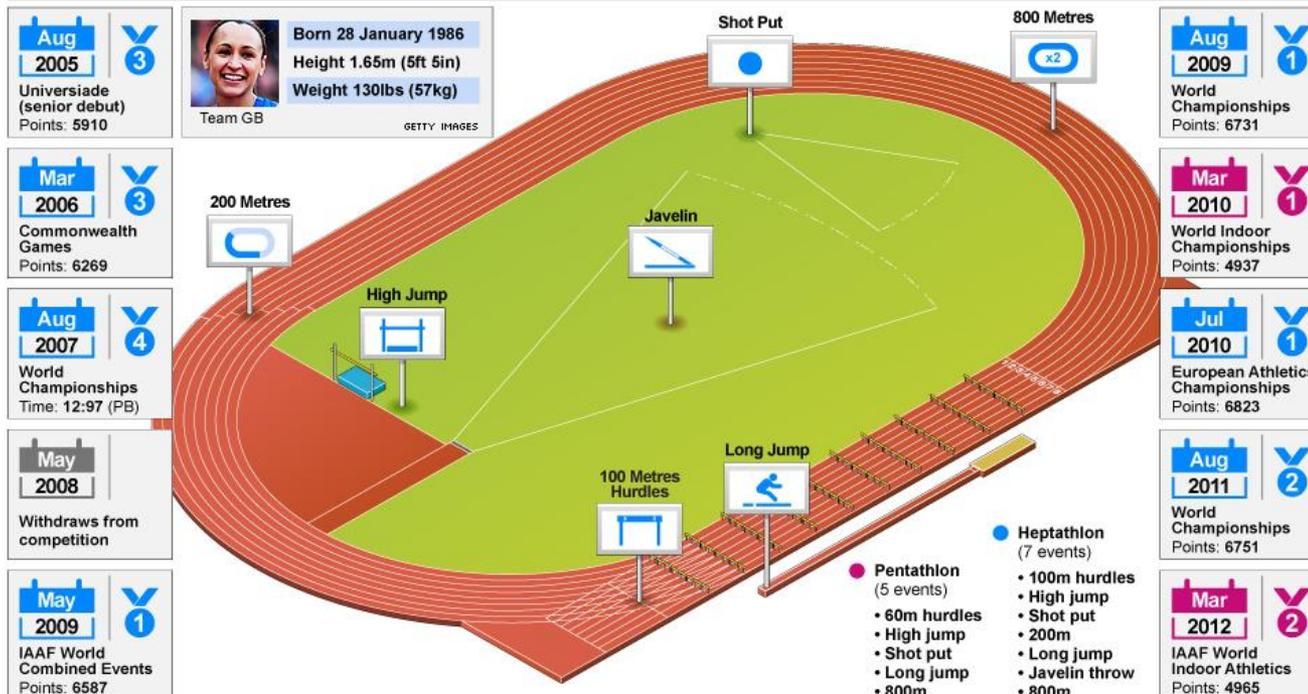


- Die ersten ca. 10 PCs haben grosses PVE; danach flacht Kurve ab
- Mit ca. 30 PCs hat man 80% der Varianz erklärt
- Zum Vergleich: die restlichen PCs erklären nur noch 20% der Varianz

Beispiel 3: Siebenkampf

Wie erstellt man einen eindimensionalen Index, der Subjekte möglichst gut unterscheidet?

Jessica Ennis career history



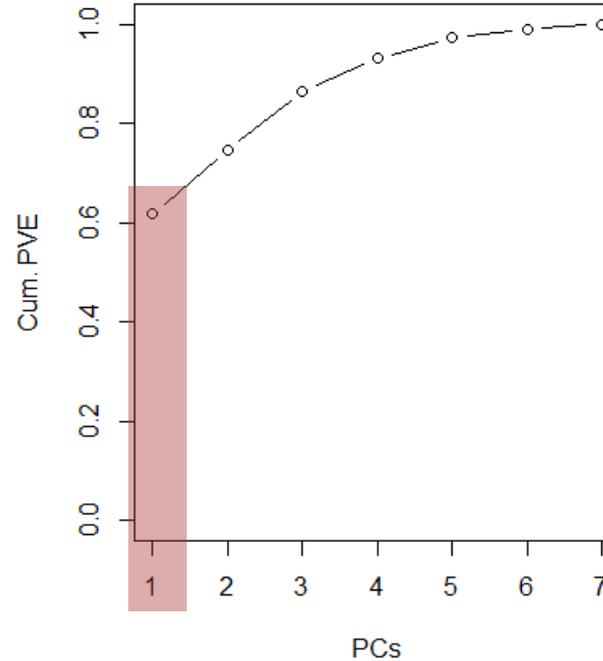
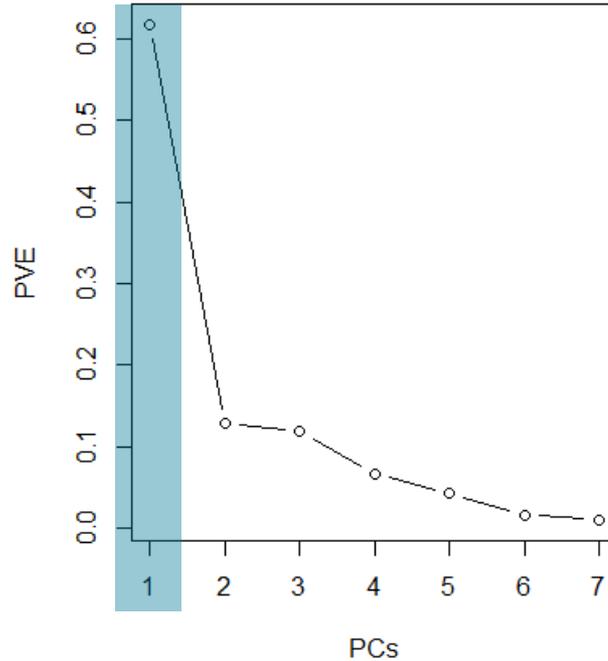
Siebenkampf

Bsp 3: Korrelationsmatrix

```
> cor(dat2)
      hurdles highjump      shot run200m longjump javelin run800m
hurdles 1.000000 0.5817409 0.7666860 0.8300371 0.8893472 0.3324779 0.5587794
highjump 0.5817409 1.0000000 0.4646854 0.3909024 0.6626910 0.3480793 0.1523350
shot      0.7666860 0.4646854 1.0000000 0.6694330 0.7840380 0.3430333 0.4082925
run200m   0.8300371 0.3909024 0.6694330 1.0000000 0.8106176 0.4707969 0.5731902
longjump  0.8893472 0.6626910 0.7840380 0.8106176 1.0000000 0.2870826 0.5233809
javelin   0.3324779 0.3480793 0.3430333 0.4707969 0.2870826 1.0000000 0.2559348
run800m   0.5587794 0.1523350 0.4082925 0.5731902 0.5233809 0.2559348 1.0000000
```

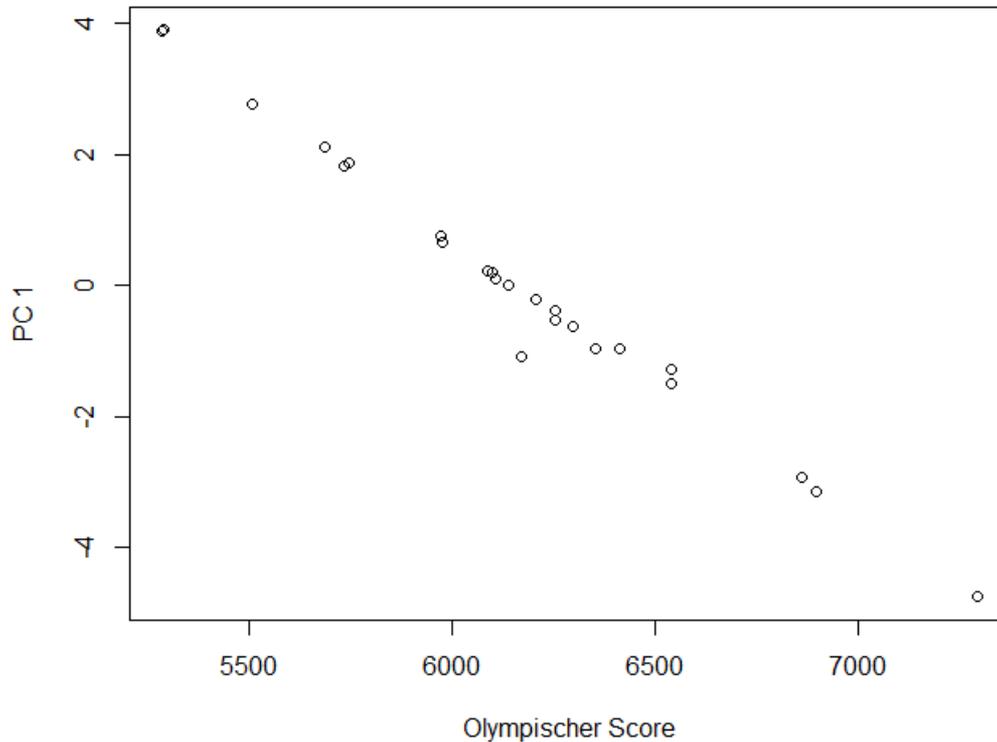
Bsp 3: Scree-Plot

PC 1 ist die “Richtung”, in der die Punkte am meisten streuen → ideal für Ranking



PC 1 erklärt schon über 60% der Varianz !

PC 1 vs. Olympischer Score



PC 1 gibt den olympischen Score (mit kleinen Ausnahmen) gut wieder.