

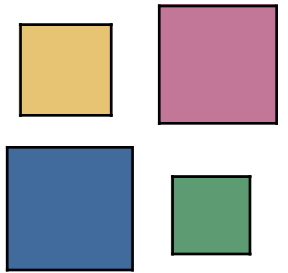
Multivariate Verteilungen

- Notation: Daten

Daten (n Beobachtungen, p Variablen) haben folgende Matrix Darstellung:

$$\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_p]$$
$$= \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}_{n \times p}$$

X_{ij} ist das Element in der i -ten Zeile und j -ten Spalte, d. h. i . Fall und j . Variable.



Wiederholung: Statistiken

Das Stichproben *Mittel* ist definiert als

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}, \quad j = 1, \dots, p.$$

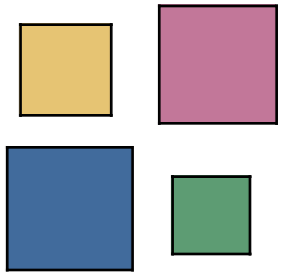
Die Stichproben *Varianz* ist definiert als

$$s_j^2 = s_{jj} = \frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2, \quad j = 1, \dots, p.$$

Die Stichproben *Covarianz* ist definiert als

$$s_{jk} = \frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k), \quad j, k = 1, \dots, p; j \neq k.$$

Die Stichproben *Korrelation* ist definiert als $r_{jk} = \frac{s_{jk}}{s_j s_k}$. Die Korrelation zwischen zwei Vektoren ist gleich dem Cosinus des Winkels zwischen den beiden zentrierten Vektoren, $\mathbf{X}_j - \bar{X}_j$, $\mathbf{X}_k - \bar{X}_k$.

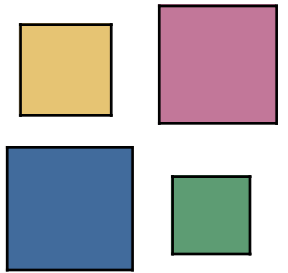


... in Matrix Form:

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

$$\mathbf{S}_n = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \cdots & \mathbf{S}_{1p} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \cdots & \mathbf{S}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{p1} & \mathbf{S}_{p2} & \cdots & \mathbf{S}_{pp} \end{bmatrix}$$

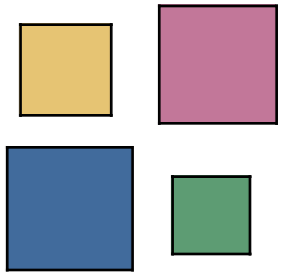
$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$



Geometrische Interpretation

- Die Projektion der Spalten X_i der Datenmatrix \mathbf{X} auf den Einsvektor $\mathbf{1} := (1, \dots, 1)_p$ ist der Vektor $\bar{x}_i \mathbf{1}$. Dieser Vektor hat Länge $\sqrt{n}|\bar{x}_i|$.
- Die Information aus \mathbf{S}_n erhält man aus den Devianz Vektoren $\mathbf{d}_i := x_i - \bar{x}_i \mathbf{1} = [x_{1i} - \bar{x}_i, \dots, x_{ni} - \bar{x}_i]'$. Das Quadrat der Länge von \mathbf{d}_i ist ns_{ii} , und das innere Produkt von \mathbf{d}_i und \mathbf{d}_j ist gleich ns_{ij}
- Die Stichproben Korrelation r_{ij} ist der Cosinus des Winkels ϑ_{jk} zwischen \mathbf{d}_j und \mathbf{d}_k , da

$$\mathbf{d}'_j \mathbf{d}_k = \sum_{i=1}^n (x_{ji} - \bar{x}_i)(x_{ki} - \bar{x}_i) = \|\mathbf{d}_j\| \cdot \|\mathbf{d}_k\| \cos(\vartheta_{jk})$$



Geometrische Interpretation: Beispiel

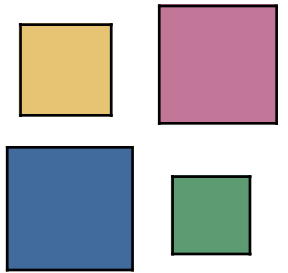
- Gegeben seien die Daten $\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$

- Mit $\bar{x} = (2, 3)$ folgt

$$\mathbf{d}_1 = \mathbf{x}_1 - \bar{x}_1 \mathbf{1} = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} \quad \mathbf{d}_2 = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

- Somit folgt: $\mathbf{d}'_1 \mathbf{d}_1 = [2 \quad -3 \quad 1] \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = 14 = 3s_{11}$

$$\mathbf{d}'_2 \mathbf{d}_2 = 8 = 3s_{22} \quad \mathbf{d}'_1 \mathbf{d}_2 = \mathbf{d}'_2 \mathbf{d}_1 = -2 = 3s_{12}$$



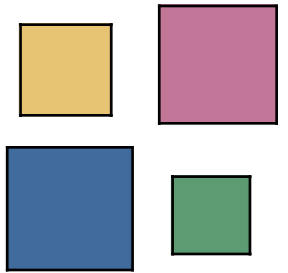
Eigenschaften von p -dim Statistiken

Ein Zufallsvektor ist ein Vektor dessen Elemente Zufallsvariablen (ZV) sind.

Seien $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$ und $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}$ Zufallsvektoren.

Dann gilt für $E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix}$, $E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$.

$E(\mathbf{A}\mathbf{X}) = \mathbf{A}E(\mathbf{X})$, wobei \mathbf{A} eine $p \times p$ Matrix von Konstanten ist.



...

$$\text{Für } \text{Var}(\mathbf{X}) = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_1, X_p) & \text{Cov}(X_2, X_p) & \dots & \text{Var}(X_p) \end{bmatrix}$$

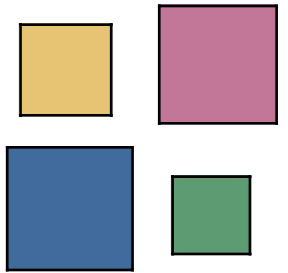
$$\text{Sei } E(\mathbf{X}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \text{ und } \text{Var}(\mathbf{X}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_{pp} \end{bmatrix}$$

Dann gilt

$$\boldsymbol{\Sigma} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})']$$

und für die Linearkombination $\mathbf{Z} = \mathbf{C}\mathbf{X}$

$$\boldsymbol{\mu}_{\mathbf{Z}} = E(\mathbf{Z}) = E(\mathbf{C}\mathbf{X}) = \mathbf{C}\boldsymbol{\mu}_{\mathbf{X}} \quad \text{und} \quad \boldsymbol{\Sigma}_{\mathbf{Z}} = \text{Cov}(\mathbf{Z}) = \text{Cov}(\mathbf{C}\mathbf{X}) = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}'$$



Projektionen

Eine lineare Kombination \mathbf{a} der Daten \mathbf{X} wird geschrieben als

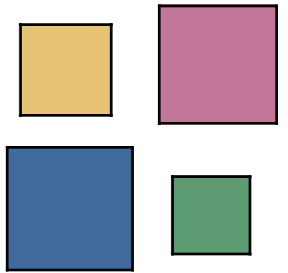
$$\mathbf{Xa} = \begin{bmatrix} a_1\mathbf{X}_{11} + a_2\mathbf{X}_{21} + \dots + a_p\mathbf{X}_{p1} & \dots \\ a_1\mathbf{X}_{n1} + a_2\mathbf{X}_{2n} + \dots + a_p\mathbf{X}_{pn} \end{bmatrix}_{n \times 1}$$

Eine 1-D Projektion von Daten auf einen Vektor $\alpha_{p \times 1}$ ist:

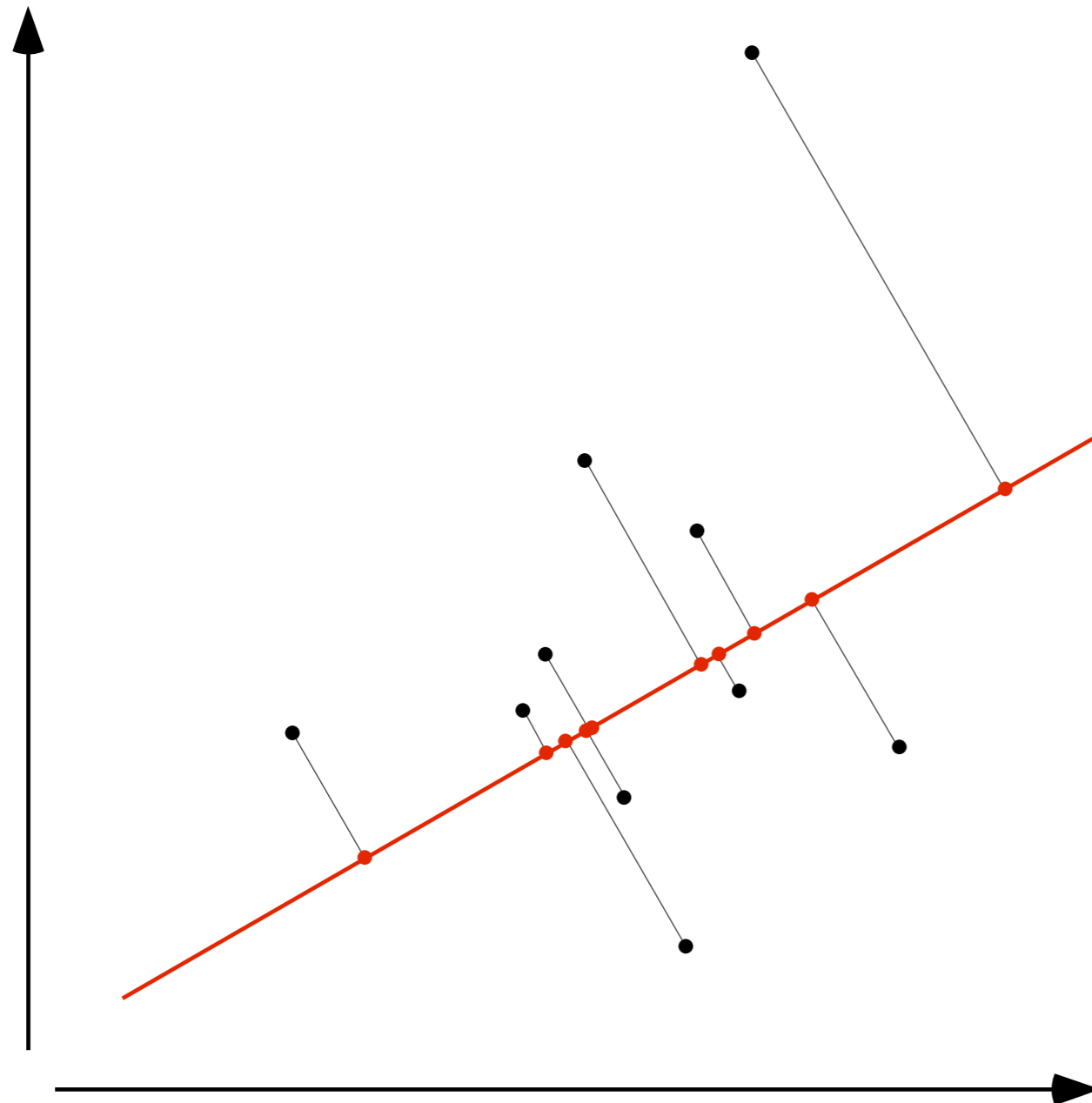
$$\mathbf{X}_1\alpha = \begin{bmatrix} \alpha_1\mathbf{X}_{11} + \alpha_2\mathbf{X}_{12} + \dots + \alpha_p\mathbf{X}_{1p} & \dots \\ \alpha_1\mathbf{X}_{n1} + \alpha_2\mathbf{X}_{2n} + \dots + \alpha_p\mathbf{X}_{np} \end{bmatrix}_{n \times 1}$$

wobei $\|\alpha\| = \sqrt{\alpha_1^2 + \dots + \alpha_p^2} = 1$ gilt.

Eine 2-D Projektion kann erzeugt werden durch die Erweiterung von α auf $A_{p \times 2} = [\alpha_1 \ \alpha_2]$ mit orthonormalen Spalten, $\alpha_1' \alpha_2 = 0$. Analog gilt die Erweiterung auf d -D Projektionen.

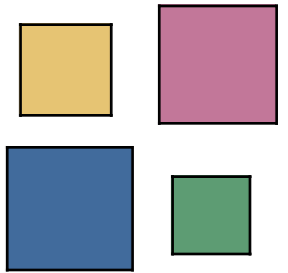


Geometrische Interpretation



2-dim. Daten X

1-dim. Projektion X_α



Distanzmaße

Sei $\mathbf{X} = (X_1 \ X_2 \ \dots \ X_p)'$ und $\mathbf{Y} = (Y_1 \ Y_2 \ \dots \ Y_p)'$ dann ist die Euclidische Distanz (“direkte Verbindung”) definiert als

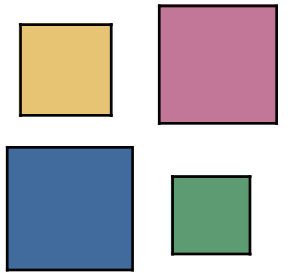
$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{(X_1 - Y_1)^2 + \dots + (X_p - Y_p)^2}$$

Als *statistische Distanz* (oder Mahalanobis Distanz) definiert man

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X} - \mathbf{Y})' \mathbf{S}^{-1} (\mathbf{X} - \mathbf{Y})}$$

Prinzipiell kann jedes Distanz Maß benutzt werden so lange es

- (1) $d(\mathbf{X}, \mathbf{Y}) = d(\mathbf{Y}, \mathbf{X})$,
 - (2) $d(\mathbf{X}, \mathbf{Y}) > 0$, if $\mathbf{X} \neq \mathbf{Y}$,
 - (3) $d(\mathbf{X}, \mathbf{Y}) = 0$, if $\mathbf{X} = \mathbf{Y}$,
 - (4) $d(\mathbf{X}, \mathbf{Y}) \leq d(\mathbf{X}, \mathbf{Z}) + d(\mathbf{Z}, \mathbf{Y})$, für jeden Zwischenpunkt \mathbf{Z} .
- erfüllt



Eigenwertzerlegung

Die Skalare, $\lambda_1, \dots, \lambda_p$, welche

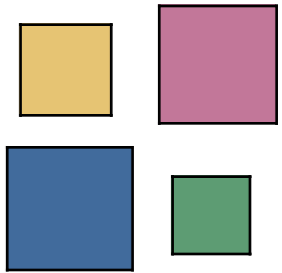
$$|\mathbf{A} - \lambda I_p| = 0$$

erfüllen, werden die Eigenwerte von \mathbf{A} genannt. Die von 0 verschiedenen Vektoren \mathbf{x} die

$$\mathbf{Ax} = \lambda \mathbf{x}$$

erfüllen, werden Eigenvektoren von \mathbf{A} genannt.

Jede quadratische, symmetrische Matrix (z.B. \mathbf{S}_n, \mathbf{R}) kann zerlegt werden in Eigenwerte und Eigenvektoren und dadurch beschrieben werden.



Geometrische Interpretation

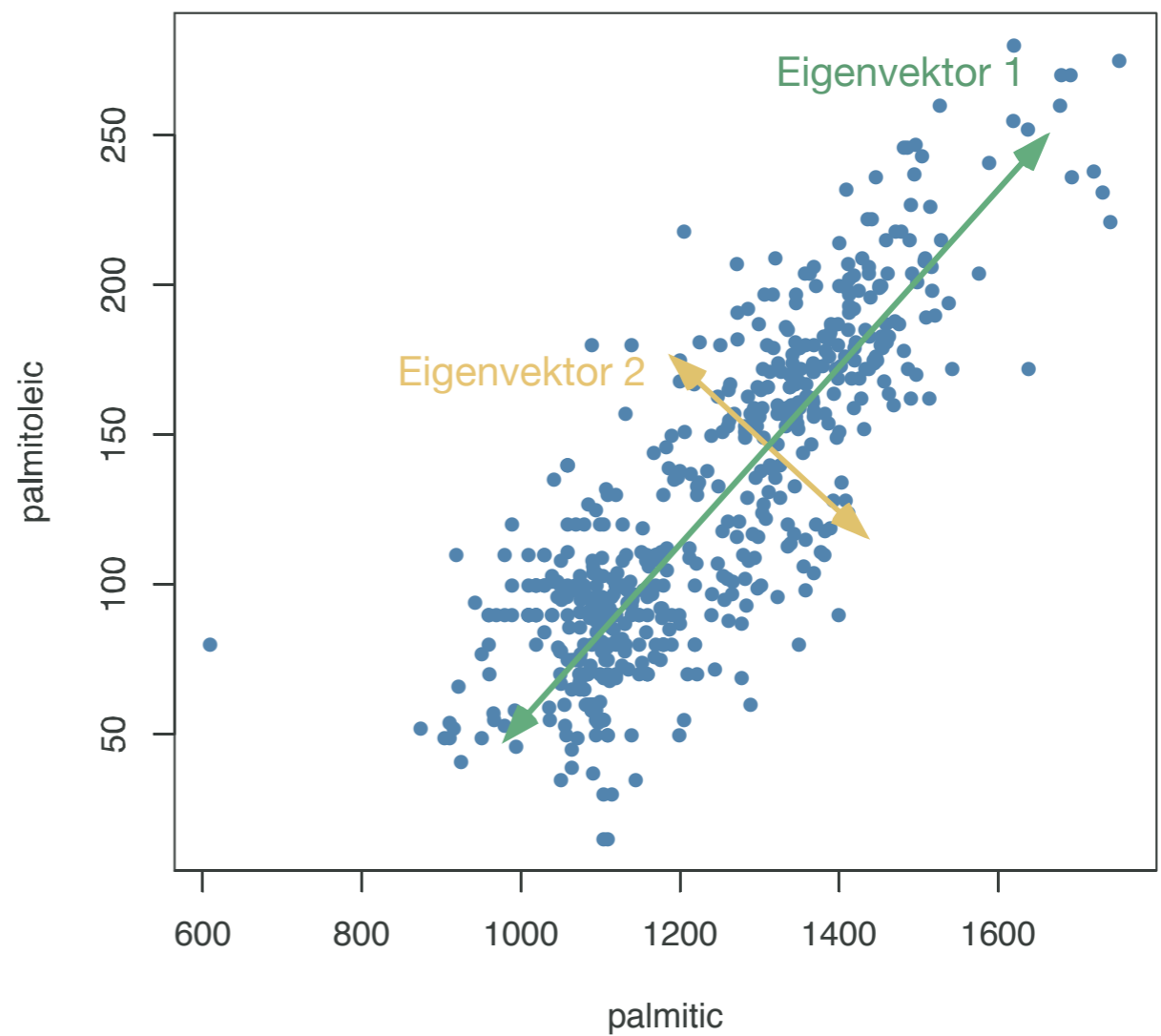
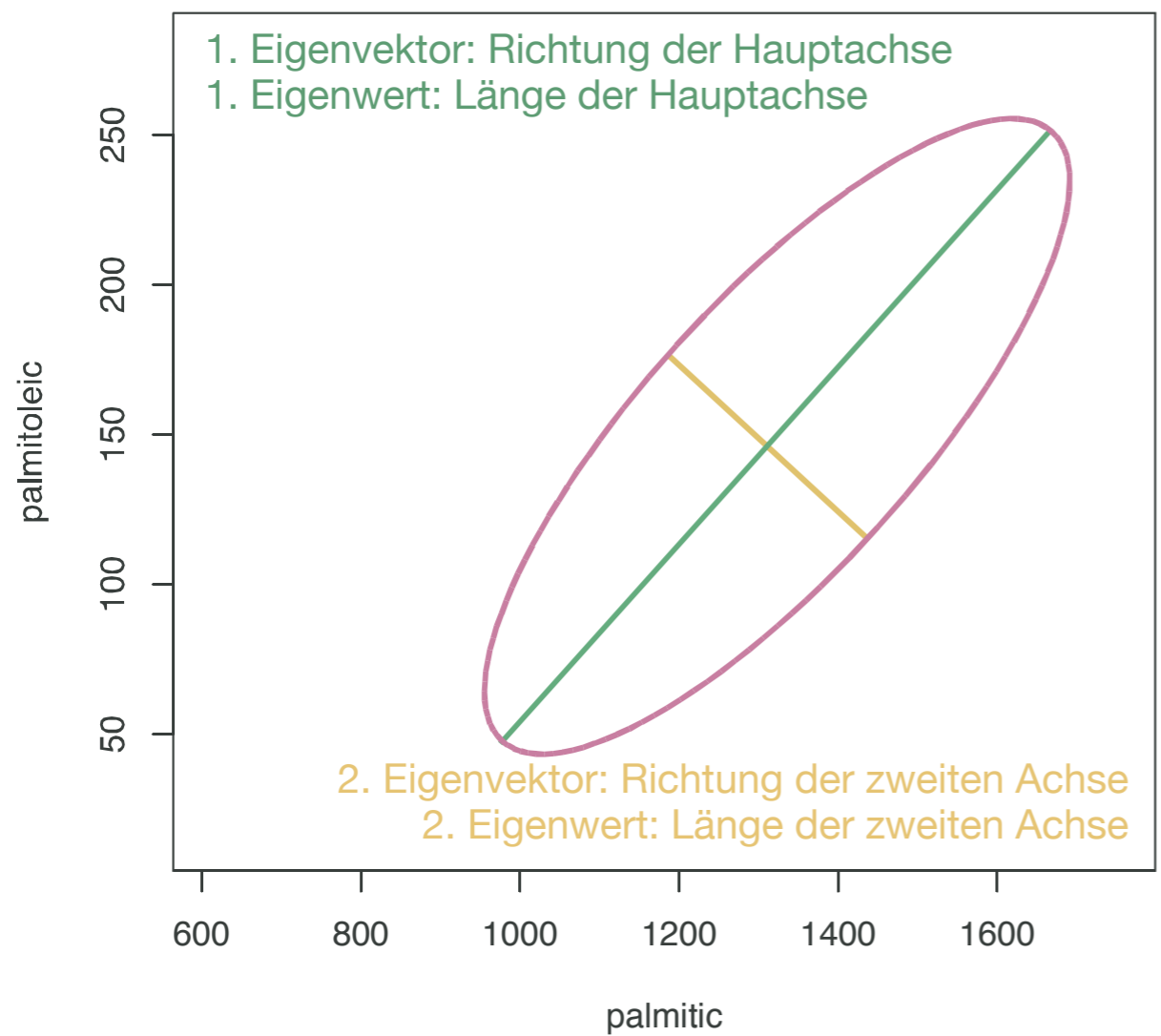
Eigenvektoren

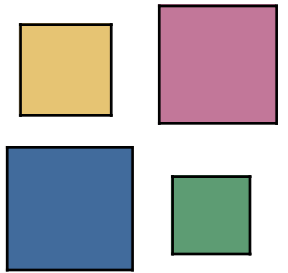
Varianz- Kovarianz Interpretation:

Ellipse repräsentiert

Varianz- Kovarianz Matrix

Daten Interpretation:





Multivariate Normalverteilung

- Univariater Fall

Klassische “Glockenkurve”

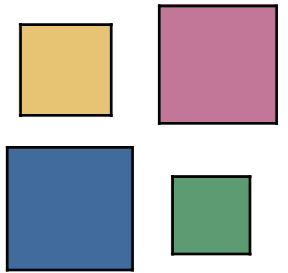
Dichte:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right\}$$

wo μ den Mittelwert und σ die Standardabweichung bezeichnen.

Die Wahrscheinlichkeit, dass Werte im Intervall einer Standardabweichung um den Mittelwert liegen, d.h. $P(\mu - \sigma < X < \mu + \sigma)$, ist ca. 0.68, für 2 Standardabweichungen, d.h. $P(\mu - 2\sigma < X < \mu + 2\sigma)$, ist der Wert 0.95.

Die Standard Normalverteilung hat Mittelwert 0 und Varianz 1.



Multivariate Normalverteilung

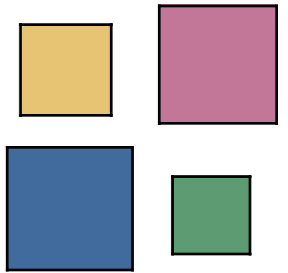
Die Normalverteilung kann leicht aus der univariaten Form in die multivariate Form abgeleitet werden. Der Mittelwert und die Varianz- Covarianz Parameter sind μ , bzw. Σ .

Die Dichtefunktion dieser p -dimensionalen Zufallsvariable hat die Form

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left\{-\frac{(\mathbf{X} - \mu)' \Sigma^{-1} (\mathbf{X} - \mu)}{2}\right\},$$

geschrieben als $N_p(\mu, \Sigma)$.

Die Standard Multivariate Normalverteilung hat demnach als Mittelwert Vektor den Nullvektor, und als Varianz- Covarianz Matrix die Einheitsmatrix.



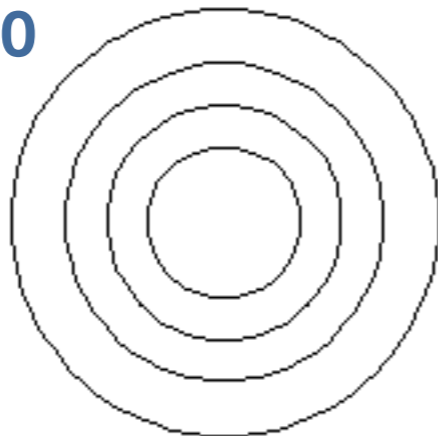
2-D Normalverteilung

$$f(x_1, x_2) = \frac{1}{2\pi \sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}} \exp\left\{-\frac{z}{2(1-\rho^2)}\right\}$$

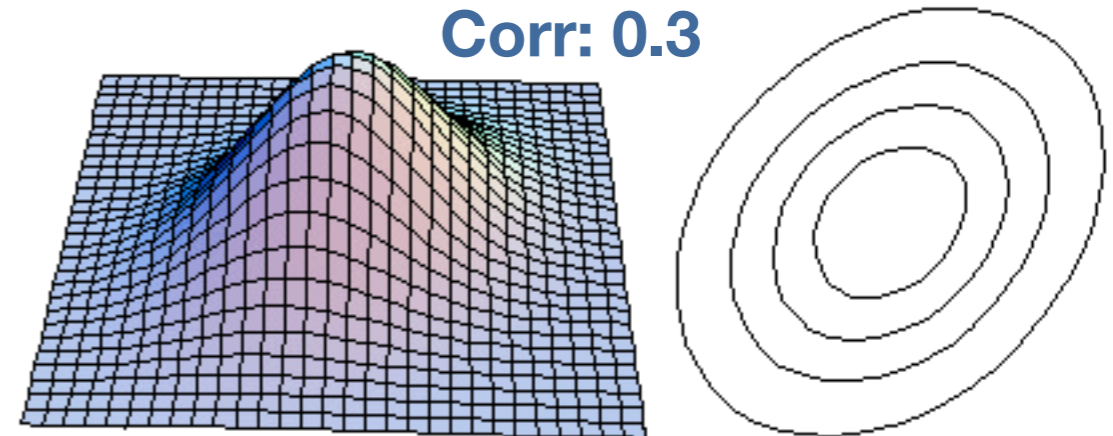
mit

$$z = \frac{(x_1 - \mu_1)^2}{\sigma_{11}} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}} + \frac{(x_2 - \mu_2)^2}{\sigma_{22}} \quad \text{und} \quad \rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$$

Corr: 0.0



Corr: 0.3



Corr: 0.6



Corr: 0.9

