

Principal Components Analysis (PCA)

SE:MED 4. Semester

Werner Backfrieder

Backfrieder-Hagenberg

Principal Components Analysis (PCA)

- Motivation
 - Fragestellung: Ist in den 64 Ableitungen eines MEG ein gemeinsames Signal enthalten?
 - Sind die Signale korreliert?
- Datenanalyse, finden gemeinsamer Merkmale (Features).

Backfrieder-Hagenberg

Statistische Beschreibung von Signalen

- Signal x_i kann als Zufallsvariable X beschrieben werden.
- Die Verteilung der Zufallsvariablen $\{X\}_N$ wird beschrieben durch:

- Mittelwert

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- Varianz=mittlere quadratische Abweichung vom Mittelwert:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Backfrieder-Hagenberg

Varianz- und Kovarianz

- Varianz beschreibt die Streuung der Messwerte um ihren Mittelwert.
- Werden zwei Zufallsvariablen beobachtet, z.B. zwei Signale, Größe und Gewicht einer Personengruppe ..., beschreibt die Kovarianz die Verteilung:

$$s_{ij} = \frac{1}{N} \sum_{k=1}^N (x_{ik} - \mu_i)(x_{jk} - \mu_j)$$

- Kovarianz beschreibt das Produkt der Abweichung um den Mittelwert der einzelnen Verteilungen.

Backfrieder-Hagenberg

Covarianz-Matrix

$$\Sigma = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{pmatrix}$$

- Covarianz-Matrix aus drei Signalen
 - quadratische Matrix
 - symmetrisch $s_{ij}=s_{ji}$
 - Varianzen der Signale entlang der Hauptdiagonalen

Backfrieder-Hagenberg

Korrelation

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}} \cdot \sqrt{s_{jj}}}$$

- Graphische Interpretation:
 - Inneres Produkt
 - Vektoren zentriert: $\vec{x}_j = x_j - \mu$
 - Normalisiert auf Länge 1: $|\vec{x}_n|=1$
 - r_{ij} ist Cosinus des Winkels zwischen den Vektoren

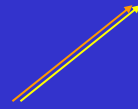
$$r_{ij} = \vec{x}_i \cdot \vec{x}_j = |\vec{x}_i| |\vec{x}_j| \cos(\alpha)$$

Backfrieder-Hagenberg

Korrelation: Beispiele

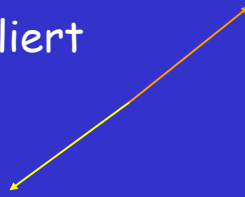
- positiv korreliert

$$r_{ij}=1$$



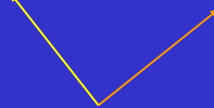
- Negativ korreliert

$$r_{ij}=-1$$



- unkorreliert

$$r_{ij}=0$$



Backfrieder-Hagenberg

PCA

- In den Signalen werden sogenannte **Hauptkomponenten (principal components, PCs)** gesucht
- PCs sind nicht korreliert, d.h. in statistischer Sicht sind die Variablen unabhängig von einander
- => Korrelationsmatrix ist diagonal
 - Mischelemente sind Null $r_{ij}=0$ wenn $i \neq j$

Backfrieder-Hagenberg

Eigenwert-Eigenvektor

- Für bestimmte Vektoren (Unterräume) einer Matrix A gilt:

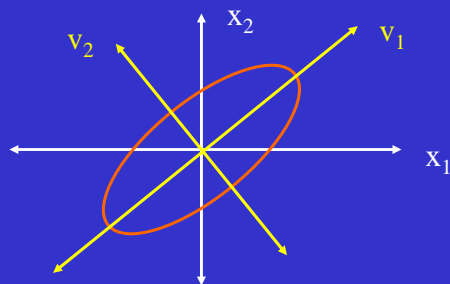
$$A \cdot v_i = \lambda_i v_i$$

- λ_i Eigenwert
- v_i Eigenvektor

- Bei einer $n \times n$ Matrix gibt es bis zu n verschiedene Eigenwerte und zugehörige Eigenvektoren
- Die Eigenvektoren stehen aufeinander normal: $\langle v_i | v_j \rangle = \delta_{ij}$

Backfrieder-Hagenberg

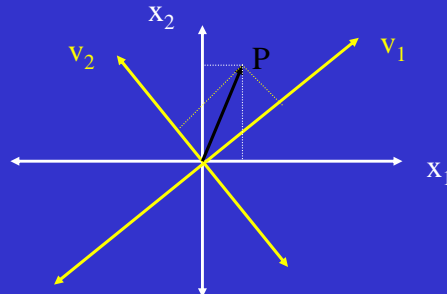
Hauptachsenform



- Ellipse komplizierte Form im Koordinatensystem x
- Einfache Form im System v
- Koordinaten v repräsentieren Eigenvektoren

Backfrieder-Hagenberg

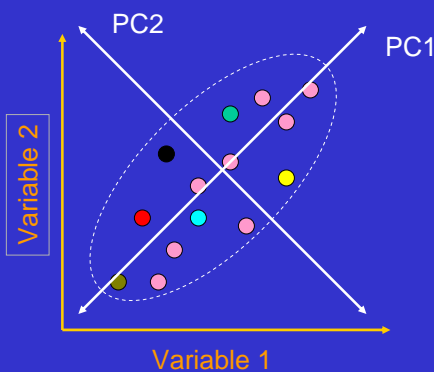
Koordinaten-Trafo



- Koordinaten des Punktes P sind das innere Produkt des Ortsvektors mit den Basisvektoren der Koordinatensysteme:
- $P_{x_1} = \langle x_1 | P \rangle$, $P_{x_2} = \langle x_2 | P \rangle \Leftrightarrow P_{v_1} = \langle v_1 | P \rangle$, $P_{v_2} = \langle v_2 | P \rangle$

Backfrieder-Hagenberg

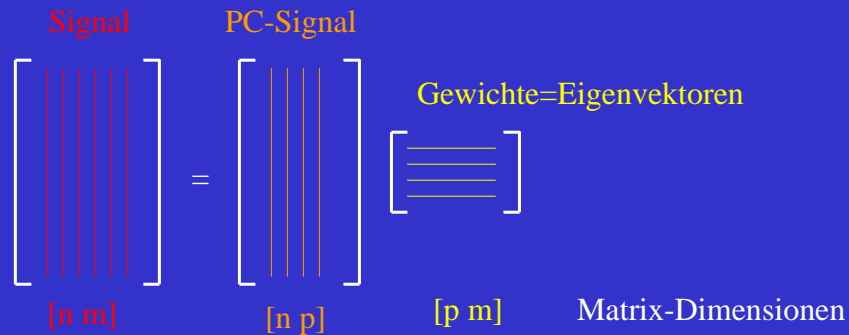
PCA: Prinzipien



- Koordinaten-Trafo
- Ursprung im Mittelwert
- PC1 in Richtung **größter Varianz**
- Höhere Dimensionen **normal** auf alle anderen
- **Unkorrelierte** Variablen

Backfrieder-Hagenberg

Datenmodell



- m Signale mit je n Samples
- $p < m$ Principal Components
- Signal ist gewichtete Summe der PCs

$$S_i = \sum_{k=1}^p v_{ki} PC_k$$

Backfriedler-Hagenberg

Anzahl der PCs

- p ist Anzahl der Komponenten
- **Spur** der Kovarianz-Matrix enthält die **Varianzen**
- Summe der Spur ist **Gesamtvarianz** des Systems
- So viele Eigenwerte auswählen, dass gilt:

$$\frac{\sum_{i=1}^p \lambda_i}{\sum_{i=1}^n \lambda_i} \geq 0.95$$

- Wahl des Signifikanzniveaus **willkürlich**

Backfriedler-Hagenberg

Berechnung der PC-Signale

- S Signalmatrix, Spalten enthalten Signale
- U Spalten enthalten p PC-Signale
- V enthält p Eigenvektoren multipliziert mit Wurzel der Eigenwerte
- $S=UV^T \Rightarrow SV=UV^TV \Rightarrow$
- $U=SV(V^TV)^{-1}$
- Least Squares Lösung für Signalmatrix

Backfrieder-Hagenberg

PC-Interpretation

- PC-Signale sind *nicht korreliert*
- Zerlegung des Signals in *unabhängige Variablen*
- 1. PC ist *mittleres Signal*
- 2. PC *orthogonale* Abweichung ...
- PCs $> p$ „sollen“ nur Rauschen enthalten
- Rekonstruktion der ursprünglichen Signale aus p -Hauptkomponenten
 - Entfernen des Rauschens \Rightarrow *PC Filter*

Backfrieder-Hagenberg