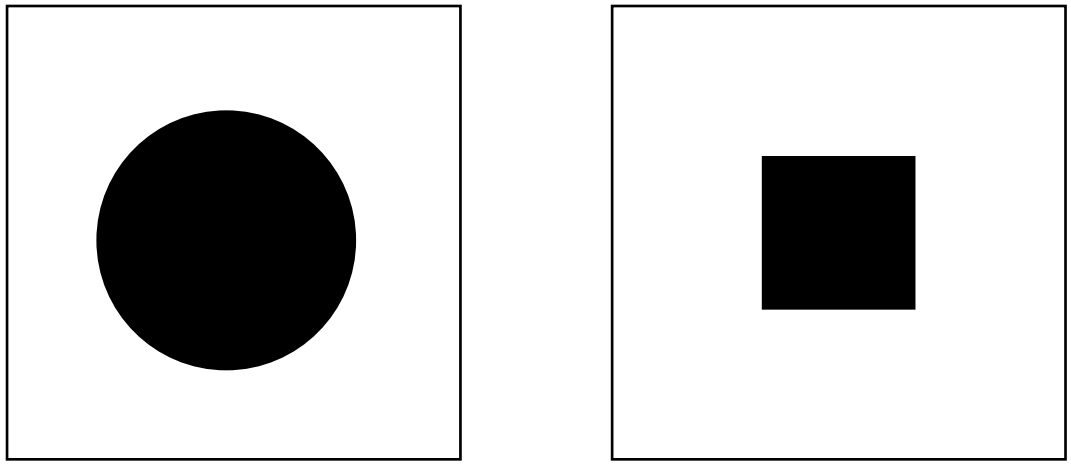
Kapitel 6

Optimale Merkmalsselektion

Einfaches Erkennungsbeispiel mit zwei Objekten (Kreis und Quadrat)



Es wird angenommen, dass die Bilder mit 512×512 Bildpunkte abgetastet werden.

Niemand käme auf die Idee hier 512×512 Bildpunkte als Merkmale der Objekte zu verwenden!! (1 Merkmal würde reichen: Fläche)

Frage: gibt es allgemeine Prinzipien zur Merkmalsauswahl ?

Merkmalsselektion mit linearen Transformationen

Die Komplexität eines Klassifikatorentwurfs wächst mit der Dimension N des Merkmalsraums. Gleichzeitig wächst der Bedarf an Stichproben!
Ziel einer Merkmalsselektion ist die Auswahl eines geeigneten Unterraumes. Die selektierten Merkmale müssen eine hohe Relevanz für die Charakterisierung der Klassen besitzen, aber zugleich eine hohe Diskriminierfähigkeit zwischen den Klassen garantieren. Sie müssen demnach innerhalb einer Klasse wenig variieren (Intraklassenabstand), aber gleichzeitig große Abstände zwischen den Klassen (Interklassenabstand) garantieren.

Es ist i.allg. wenig sinnvoll die Pixel eines Bildes direkt als Merkmale zu verwenden ($N=512^2=2^{18}=0{,}25$ Mio. Pixel). I.allg. gibt es eine hohe Redundanz in den Bildern durch starke Korrelation zwischen den Pixeln.

Es ist zudem wenig hilfreich einen Merkmalsraum durch Hinzunahme neuer Merkmale weiter zu vergrößern, wenn die neuen Merkmale stark korreliert sind mit den bereits vorhandenen.

Idee: Transformation der Originalbilder in einen neuen Merkmalsraum mit Hilfe linearer oder Euklidischer Transformationen (Verschieben und Drehung des Koordinatensystems (unitäre Transformation)). Dabei Reduktion auf wenige Merkmale und gleichzeitige Informationsverdichtung.

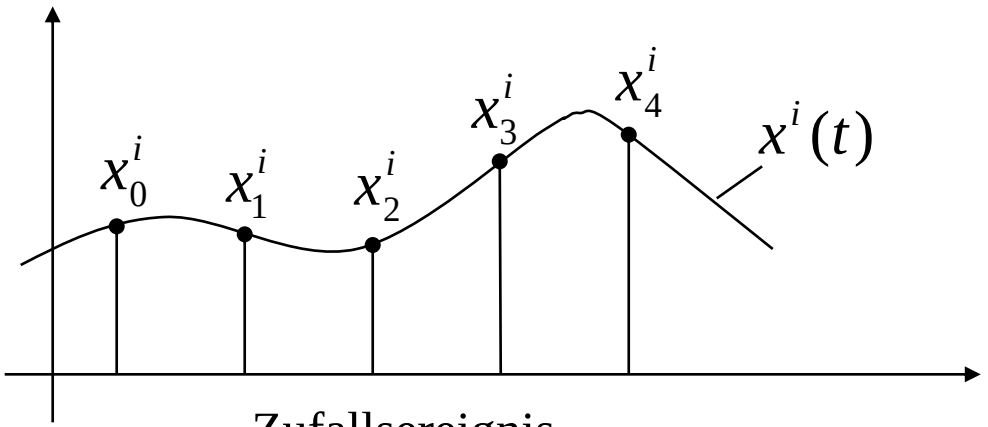
Charakterisierung von Zufallsereignissen in Vektorräumen

Ein *Zufallsereignis* $\mathbf{x}^i$ ist ein Element des Vektorraumes $\mathbb{X}$. Wobei im diskreten Fall das Elementarereignis aus einer geordneten Menge von Zahlenwerten

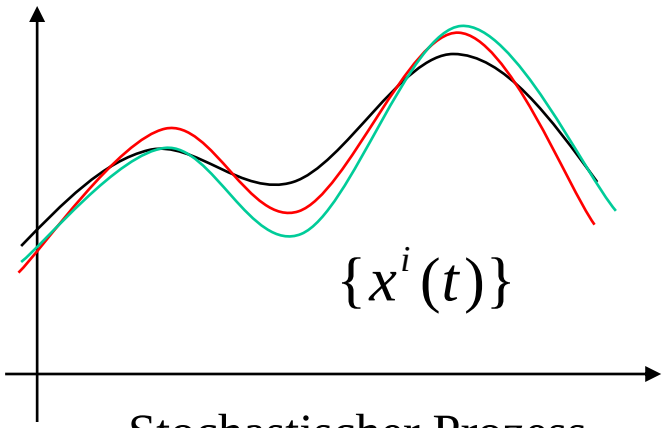
$\mathbf{x}^i := \{x_0^i, x_1^i, x_2^i, ..., x_{N-1}^i\}$

oder auch im kontinuierlichen Fall aus einer Zeit- oder Ortsfunktionen $x^i(t)$ besteht.

Ein *stochastischer Prozess* $\mathbf{x}$ besteht aus einer Menge von Ereignissen $\mathbf{x} := \{\mathbf{x}^j\}$.



Zufallsereignis



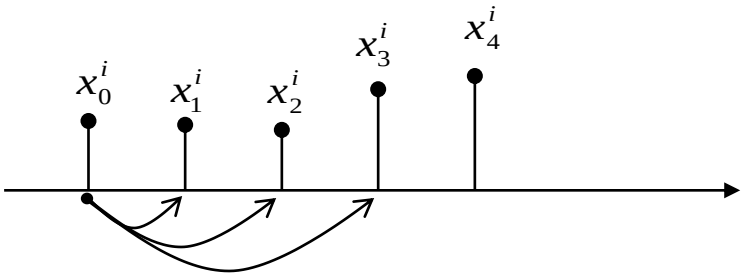
Stochastischer Prozess

Statistische Kenngrößen eines Prozesses

Erwartungswert: $\boldsymbol{\mu}_{\mathbf{x}} = \bar{\mathbf{x}} = E\{\mathbf{x}\} \approx \frac{1}{N}\sum_{i=0}^{N-1}\mathbf{x}^i$

Autokorrelation: $\mathbf{K} = \mathbf{R}_{\mathbf{xx}} = E\{\underbrace{\mathbf{x}\mathbf{x}^T}_{\text{dyad. Produkt!}}\} = \{E(x_i x_j)\} \approx \frac{1}{N}\sum_{i=0}^{N-1}\begin{bmatrix} x_0^i \cdot x_0^i & x_0^i \cdot x_1^i & x_0^i \cdot x_2^i \\ x_1^i \cdot x_0^i & x_1^i \cdot x_1^i & x_1^i \cdot x_2^i \\ x_2^i \cdot x_0^i & x_2^i \cdot x_1^i & x_2^i \cdot x_2^i \end{bmatrix}$

Die Elemente der Korrelationsmatrix beschreiben die Korrelation zwischen den einzelnen Vektorelementen $\{x_0,x_1\}$, $\{x_0,x_2\}$ ... in zeitlicher/örtlicher Richtung mit wachsendem Abstand zwischen den Elementen:



Autokovarianz: $\mathbf{C}_{\mathbf{xx}} = E\{(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\} = \mathbf{R}_{\mathbf{xx}} - \bar{\mathbf{x}}\cdot\bar{\mathbf{x}}^T$

Kreuzkorrelation: $\mathbf{R}_{\mathbf{xy}} = E\{\mathbf{x}\mathbf{y}^T\}$

Kreuzkovarianz: $\mathbf{C}_{\mathbf{xy}} = E\{(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{y}})^T\} = \mathbf{R}_{\mathbf{xy}} - \bar{\mathbf{x}}\cdot\bar{\mathbf{y}}^T$

Bei *linearen* Transformationen bleiben Gaußverteilungen erhalten

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{C_{xx}})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^T \mathbf{C_{xx}}^{-1} (\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})}$$

aus $\mathbf{y} = \mathbf{A}\mathbf{x}$ folgt für $p(\mathbf{y})$ wieder eine Normalverteilung mit:

$$\boldsymbol{\mu}_\mathbf{y} = \mathbf{A}\boldsymbol{\mu}_\mathbf{x}$$

und:

$$\mathbf{C_{yy}} = E\{(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})^T\} = \mathbf{A}E\{(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T\}\mathbf{A}^T$$

$$\Rightarrow \mathbf{C_{yy}} = \mathbf{A}\mathbf{C_{xx}}\mathbf{A}^T$$

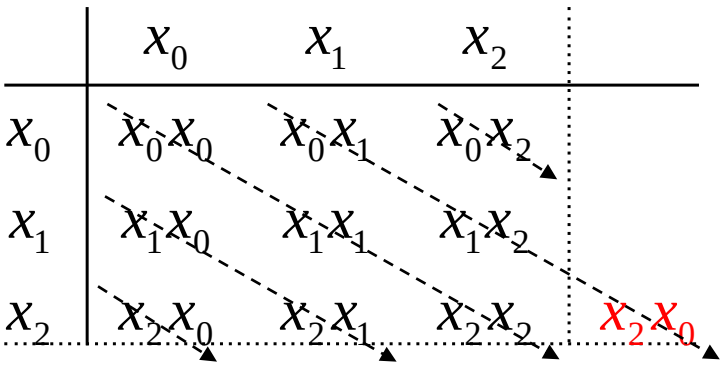
Berechnung der AKF aus der Autokorrelationsmatrix

Die Werte der lineraren (zyklischen) AKF ergeben sich aus den Diagonalsummen der (periodisch fortgesetzten) Autokorrelationsmatrix:

lineare AKF:

$$\begin{array}{lll} x_0 & x_1 & x_2 \\ x_0 & x_1 & x_2 \\ \hline x_0^2 + x_1^2 + x_2^2 \end{array} \qquad \begin{array}{llll} x_0 & x_1 & x_2 & \\ & x_0 & x_1 & x_2 \\ \hline & x_0 x_1 + x_1 x_2 & & \end{array} \qquad \begin{array}{llll} x_0 & x_1 & x_2 & \\ & & x_0 & x_1 \quad x_2 \\ \hline & & x_0 x_2 & \end{array}$$

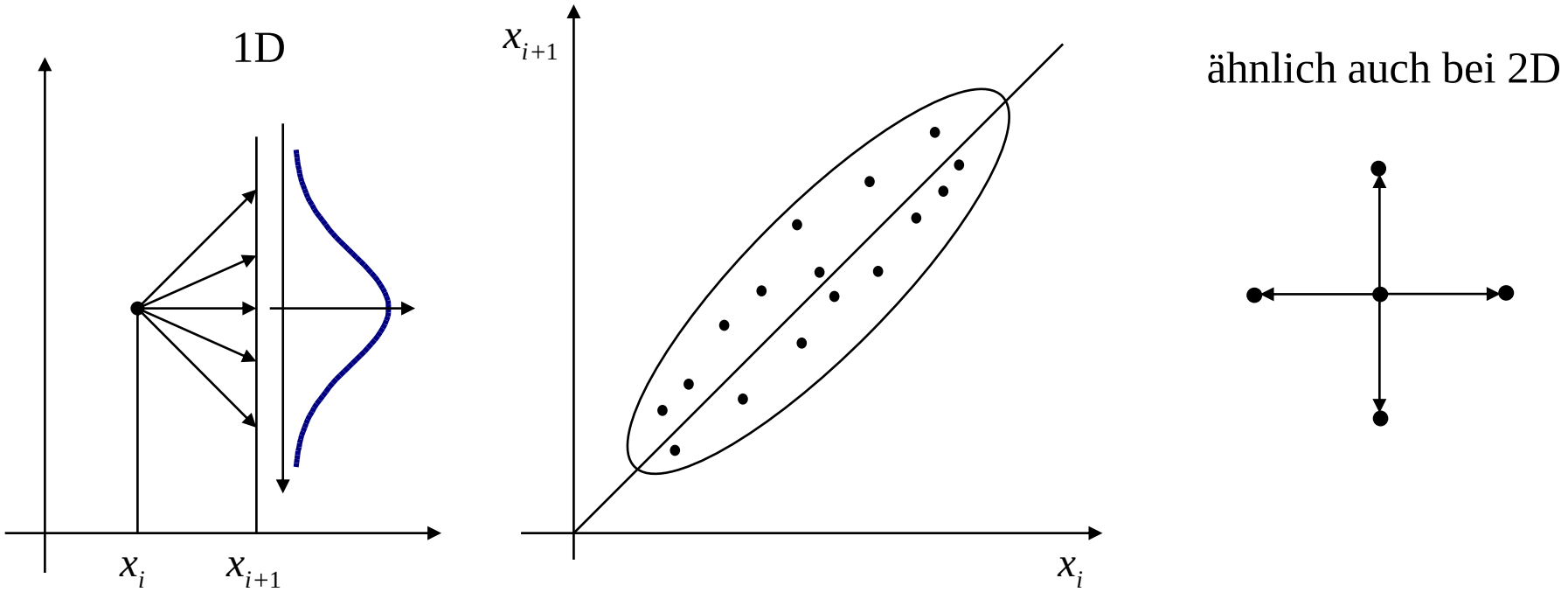
Autokorrelationsmatrix:



Dekorrelation benachbarter Signal- oder Pixelwerte im Vektorraum

Gegeben ein Signal- oder Pixelwert; wie groß ist die W., daß die benachbarte Signalamplitude ähnliche Werte annimmt?

I.a. hohe Korrelation in die Nachbarschaft! (1. Winkelhalbierende im Vektorraum)

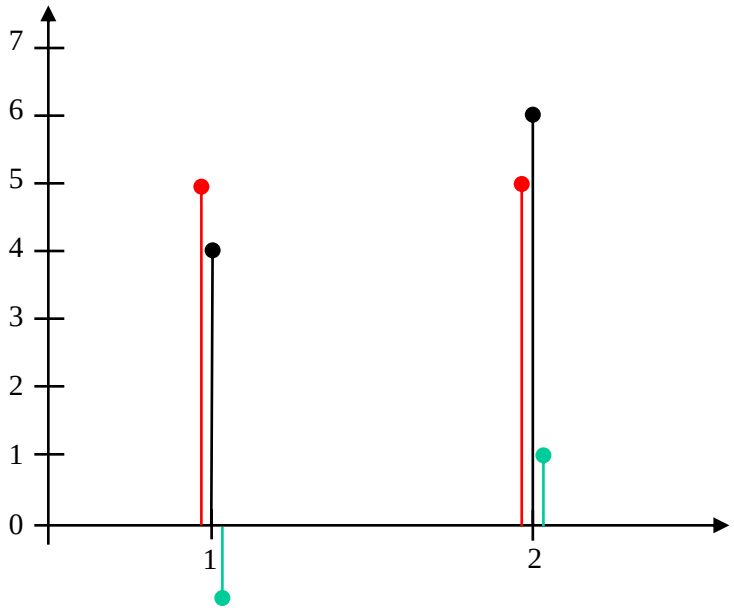


Einfaches Beispiel für den Übergang in einen neuen Merkmalsraum mit Hilfe einer orthogonalen Transformation

Gegeben sei ein Signal mit zwei Abtastwerten

$$\mathbf{x} = \begin{bmatrix} 4 \\ 6 \end{bmatrix} = 4\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 6\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$ Originalraum

$$= 5\begin{bmatrix} 1 \\ 1 \end{bmatrix} + 1\begin{bmatrix} -1 \\ 1 \end{bmatrix} = 5\sqrt{2} \cdot \tfrac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \sqrt{2} \cdot \tfrac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$ Merkmalsraum



Gleichanteil

Wechselanteil

Selektiert man nur die erste Komponente (Unterraum) im Originalraum, so erhält man eine Approximationsgüte von (*ein Abtastwert weglassen fällt auf !!*):

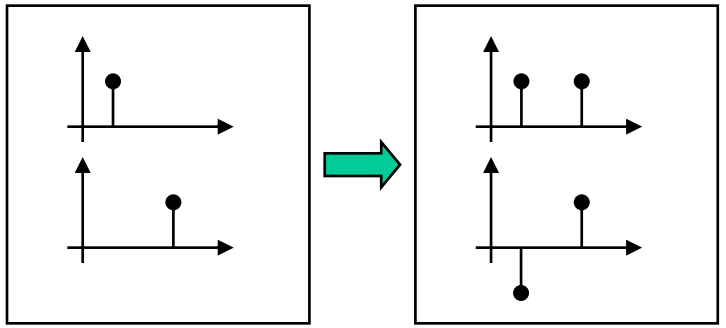
$$\frac{\|4 \quad 0\|}{\|4 \quad 6\|} = \frac{4}{7{,}21} = 55\%$$

Im neuen Merkmalsraum hingegen:

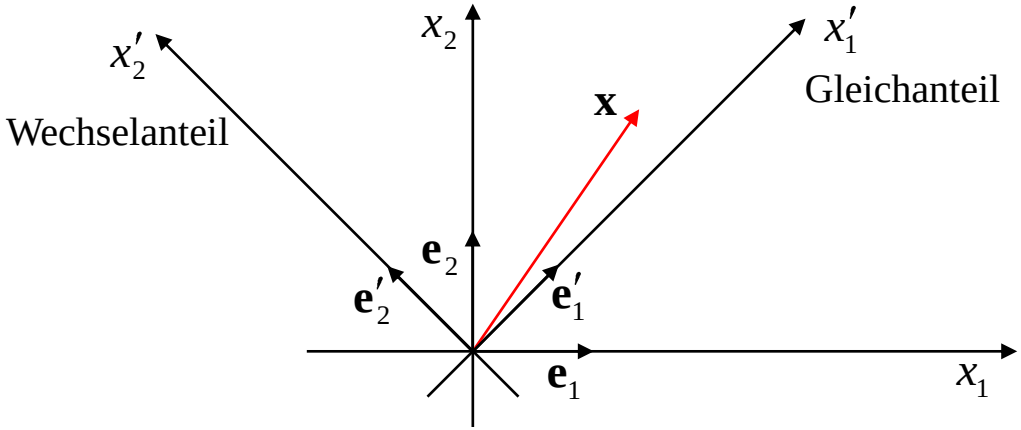
$$\frac{\|5\sqrt{2} \quad 0\|}{\|5\sqrt{2} \quad \sqrt{2}\|} = \frac{7{,}07}{7{,}21} = 98\%$$

Außerdem sind die neuen Werte unkorreliert!

Darstellung im Vektorraum durch Drehung des Koordinatensystems mit orthogonaler Transformation



Originalraum Merkmalsraum



$$\boxed{\mathbf{x}' = \mathbf{A}^T\mathbf{x}}$$

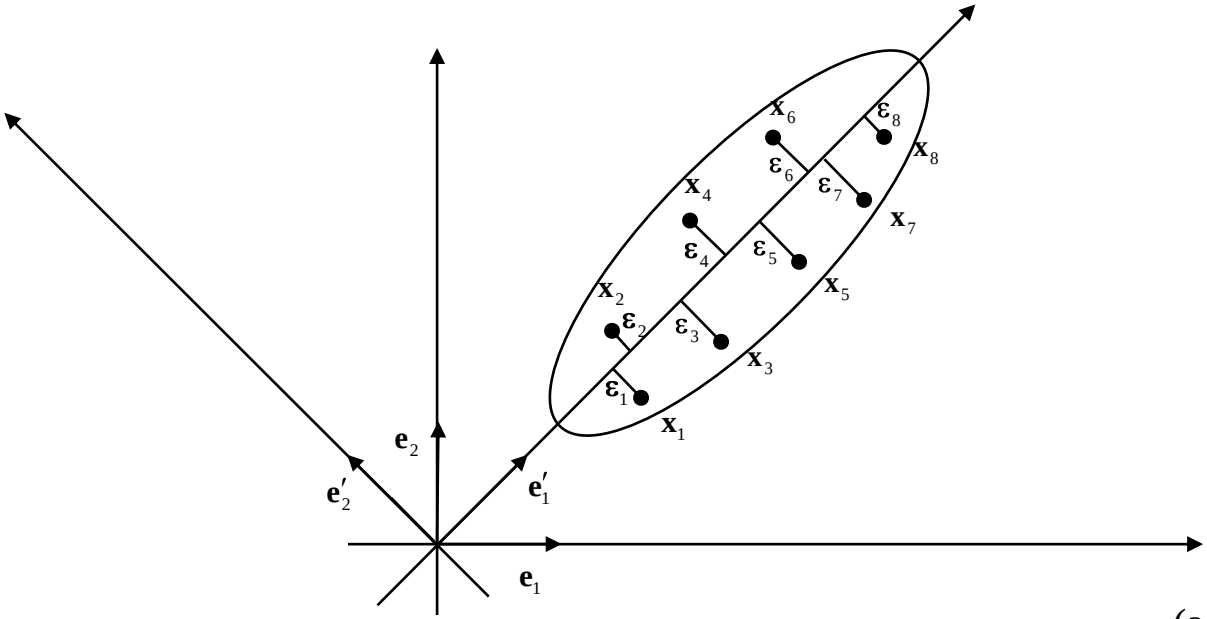
mit: $\mathbf{A}^T = \frac{1}{\sqrt{2}}\begin{bmatrix} +1 & +1 \\ -1 & +1 \end{bmatrix}$ und: $\mathbf{A}^{T\ -1} = \mathbf{A}^{T\ T} = \mathbf{A} = \frac{1}{\sqrt{2}}\begin{bmatrix} +1 & -1 \\ +1 & +1 \end{bmatrix}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 = x_1\begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2\begin{bmatrix} 0 \\ 1 \end{bmatrix} = x_1'\mathbf{e}_1' + x_2'\mathbf{e}_2' = x_1'\tfrac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + x_2'\tfrac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \underbrace{\frac{(x_1 + x_2)}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{\text{Gleichanteil}} + \underbrace{\frac{(x_2 - x_1)}{2}\begin{bmatrix} -1 \\ 1 \end{bmatrix}}_{\text{Wechselanteil}}$$

mit: $\mathbf{e}_1' = \mathbf{A}\mathbf{e}_1$ und $\mathbf{e}_2' = \mathbf{A}\mathbf{e}_2$

Optimale Merkmalsselektion mit unitären Transformationen

(Karhunen-Loeve oder Hauptachsentransformation)



Aufgabe: finde neue Basisvektoren

$$\{\mathbf{e}_i\} \xrightarrow{\mathbf{A}^T} \{\mathbf{e}'_i\}$$

$$\boxed{\mathbf{x}' = \mathbf{A}^T\mathbf{x}}$$

(adjungiert: konjug. komplex und transponiert)

Unitäre Transformation, im reellen orthogonale Transf. => Drehung des Koordinatensystems

$$\langle \mathbf{Ax}, \mathbf{Ay} \rangle = \langle \mathbf{x}, \mathbf{A}^*\mathbf{Ay} \rangle \overset{!}{=} \langle \mathbf{x}, \mathbf{y} \rangle$$

$$\Rightarrow \quad \mathbf{A}^*\mathbf{A} = \mathbf{I} \quad \Rightarrow \quad \mathbf{A}^{-1} = \mathbf{A}^*$$

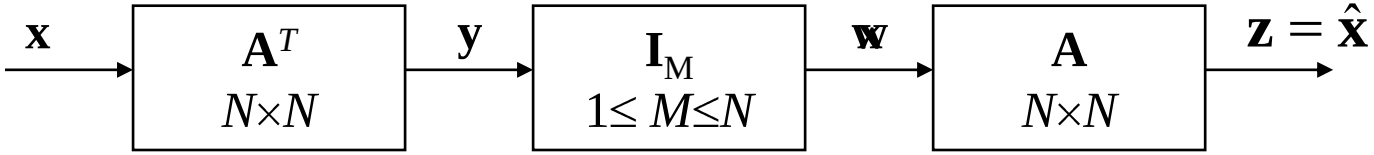
Ein einziges Vektorelement **x** kann, wenn das dazugehörige Basissystem frei gewählt werden kann (und auf Sende- und Empfangsseite bekannt ist) durch einen skalaren Wert charakterisiert werden, wenn der erste Basisvektor $\mathbf{e}'_1$ in Richtung **x** gewählt wird (Element kommt vor oder nicht):

$$\mathbf{x} = \alpha \frac{\mathbf{x}}{\|\mathbf{x}\|} + 0 \cdot \mathbf{e}'_2 + 0 \cdot \mathbf{e}'_3 + \ldots$$

Es geht jedoch i.allg. darum, für ein gesamtes *Ensemble von Vektoren* eine optimale Transformation in ein geeignetes Koordinatensystem zu finden, so dass im Mittel die Elemente des Ensembles mit möglichst wenig Koeffizienten charakterisiert werden können.

Wir beginnen mit der Bestimmung des *ersten* neuen Basisvektors $\mathbf{e}_1'$, welcher so gewählt wird, dass der *Approximationsfehler* für das Ensemble von n Vektoren *minimal* wird, oder die gesuchte Raumrichtung, die eine *maximale Information* des Ensembles repräsentiert.
Nach dem Projektionssatz erhält man den kleinsten Fehler bei orthogonaler Projektion auf den Unterraum, welcher durch $\mathbf{e}_1'$ aufgespannt wird; gesucht ist lediglich die richtige Raumrichtung. Ausgehend von einem Gütekriterium, wird eine optimale Lösung gesucht.

Vergleich mit einem optimalen Überragungskanal



$$\mathbf{w} = \mathbf{I}_M \mathbf{y} = y_1 \mathbf{e}'_1 + y_2 \mathbf{e}'_2 + \ldots + y_M \mathbf{e}'_M$$

die Bestapproximation muss für beliebige M gelten!

Ausgehend von einem quadratischen Gütekriterium ergibt sich:

$$J = \tfrac{1}{n}\{\|\boldsymbol{\varepsilon}_1\|^2 + \|\boldsymbol{\varepsilon}_2\|^2 + \ldots + \|\boldsymbol{\varepsilon}_n\|^2\}$$

$$= \tfrac{1}{n}\{\|\mathbf{x}_1 - \langle \mathbf{x}_1, \mathbf{e}'_1 \rangle \mathbf{e}'_1\|^2 + \|\mathbf{x}_2 - \langle \mathbf{x}_2, \mathbf{e}'_1 \rangle \mathbf{e}'_1\|^2 + \ldots\}$$

$$= E\{\|\mathbf{x} - \underbrace{\langle \mathbf{x}, \mathbf{e}'_1 \rangle \mathbf{e}'_1}_{\substack{\mathbf{Px} \\ \text{Projektion auf} \\ S(\mathbf{e}'_1)}}\|^2\}$$

für das Ensemble: $\mathbf{x} := \{\mathbf{x}_i\} \quad i = 1, 2, \ldots, n$

ZR: Bei einer orthogonalen *Projektion* auf einen Unterraum **Px** gilt gemäß dem Satz von Pythagoras:

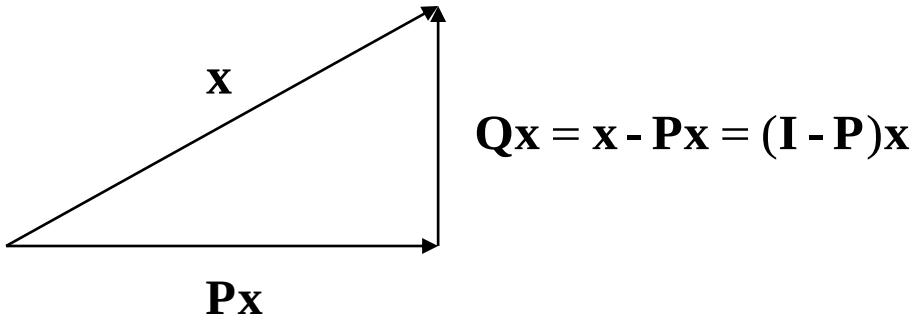
$$\|\mathbf{x}\|^2 = \|\mathbf{Px}\|^2 + \|(\underbrace{\mathbf{I}-\mathbf{P}}_{\mathbf{Q}})\,\mathbf{x}\|^2$$

$$\Rightarrow \boxed{\|\mathbf{x}-\mathbf{Px}\|^2 = \|\mathbf{x}\|^2 - \|\mathbf{Px}\|^2}$$

mit: $\langle \mathbf{x}, \mathbf{e}'_1 \rangle \mathbf{e}'_1 = \underbrace{(\mathbf{e}'_1 \mathbf{e}'^{T}_1)}_{\mathbf{P}}\,\mathbf{x}$

bei reellen Vektoren gilt:

$\langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{b}, \mathbf{a} \rangle$



und deshalb:

$$\langle\langle \mathbf{x},\mathbf{e}_1'\rangle\mathbf{e}_1',\langle \mathbf{x},\mathbf{e}_1'\rangle\mathbf{e}_1'\rangle = \langle \mathbf{x},\mathbf{e}_1'\rangle^2 \underbrace{\langle \mathbf{e}_1',\mathbf{e}_1'\rangle}_{=1}$$

$$J = E\{\|\mathbf{x}\|^2 - \overbrace{\|\langle \mathbf{x},\mathbf{e}_1'\rangle\mathbf{e}_1'\|^2}\}$$

$$= E\{\|\mathbf{x}\|^2 - \langle \mathbf{x},\mathbf{e}_1'\rangle^2\}$$

(Maximierung der Quadrate der FK)

$$= E\{\|\mathbf{x}\|^2 - \langle \mathbf{e}_1',\mathbf{x}\rangle\langle \mathbf{x},\mathbf{e}_1'\rangle\}$$

Nützliche Formeln:

$$\langle \mathbf{a},\mathbf{b}\rangle\langle \mathbf{c},\mathbf{d}\rangle = \mathbf{a}^T(\overbrace{\mathbf{b}\mathbf{c}^T}^{\text{dyad. Produkt}})\mathbf{d}$$

$$\mathbf{a}\langle \mathbf{b},\mathbf{c}\rangle = (\mathbf{a}\mathbf{b}^T)\mathbf{c}$$

$$\langle \mathbf{a},\mathbf{b}\rangle = \text{Spur}(\mathbf{a}\mathbf{b}^T)$$

Das Innenprodukt kann über die Spur des Aussenproduktes berechnet werden!

$$\langle \mathbf{A},\mathbf{B}\rangle = \text{Spur}(\mathbf{A}\ \underbrace{\mathbf{B}^*}_{\mathbf{B}\text{ adjungiert}})$$

bei Bildmatrizen

und deshalb:

$$J = E\{\|\mathbf{x}\|^2 - \mathbf{e}_1'^T(\mathbf{x}\mathbf{x}^T)\mathbf{e}_1'\}$$

$$= \underbrace{E\{\|\mathbf{x}\|^2\}}_{\sigma_x^2} - \mathbf{e}_1'^T \underbrace{E\{\mathbf{x}\mathbf{x}^T\}}_{\mathbf{R_{xx}}}\mathbf{e}_1'$$

σ_x^2: Varianz von x

$\mathbf{R_{xx}}$: Auto-korrelations-matrix

$$\Rightarrow \boxed{J = \text{Spur}(\mathbf{R_{xx}}) - \mathbf{e}_1'^T\mathbf{R_{xx}}\mathbf{e}_1' \overset{!}{=} \min_{\mathbf{e}_1'}}$$

Als Nebenbedingung geht ein, dass es sich bei dem neuen Basisvektor um einen Einheitsvektor handelt:

$$\|\mathbf{e}_1'\|^2 = \langle \mathbf{e}_1', \mathbf{e}_1' \rangle = 1$$

Der erste Term in J ist konstant und somit wird J minimiert, falls der folgende Ausdruck maximiert wird:

$$J' = \mathbf{e}_1'^T \mathbf{R}_{\mathbf{xx}} \mathbf{e}_1' \stackrel{!}{=} \max_{\mathbf{e}_1'}$$

Einbindung der Nebenbedingung in die Maximierung von J' durch einen Lagrange-Ansatz:

$$J'' = \mathbf{e}_1'^T \mathbf{R}_{\mathbf{xx}} \mathbf{e}_1' + \lambda(1 - \langle \mathbf{e}_1', \mathbf{e}_1' \rangle)$$

mit Hilfe von: $\frac{\partial \langle \mathbf{y}, \mathbf{y} \rangle}{\partial \mathbf{y}} = 2\mathbf{y}$ und: $\frac{\partial (\mathbf{y}^T \mathbf{R} \mathbf{y})}{\partial \mathbf{y}} = 2\mathbf{R}\mathbf{y}$ (falls $\mathbf{R}$ symm.)

ergibt sich aus der notwendigen Bedingung für ein Extremum:

$$\frac{\partial J''}{\partial \mathbf{e}_1'} = 2(\mathbf{R}_{\mathbf{xx}} \mathbf{e}_1' - \lambda \mathbf{e}_1') \stackrel{!}{=} \mathbf{0}$$

und daraus die Eigenwertgleichung:

$$\boxed{\mathbf{R}_{\mathbf{xx}} \mathbf{e}_1' = \lambda \mathbf{e}_1'}$$

Eingesetzt in J' ergibt: $\boxed{J' = \lambda \mathbf{e}_1'^T \mathbf{e}_1' = \lambda}$

Dieser Ausdruck wird maximal, wenn man unter allen Eigenwerten den maximalen $\lambda_1 = \lambda_{max}$ und den dazugehörigen Eigenvektor aussucht.

Man spaltet nun den eindimensionalen Unterraum entlang $\mathbf{e}_1'$ ab, fährt fort in dem verbleibenden Unterraum mit der Suche nach dem zweiten Basisvektor $\mathbf{e}_2'$ => λ_2 und dem zweitgrößten Eigenwert, usw.

Approximationsfehler

Der Approximationsfehler mit M Komponenten ($1 \leq M \leq N$) ergibt sich somit zu:

$$J_M = E\{\|\mathbf{x} - \mathbf{z}\|^2\} = E\{\|\mathbf{x} - \underbrace{\mathbf{A}\mathbf{I}_M\mathbf{A}^T}_{\substack{\text{orthog.}\\ \text{Projektion } \mathbf{P}}}\mathbf{x}\|\} \quad \text{mit:} \quad \mathbf{A} = \mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_N$$

auf den Unterraum, welcher durch die ersten M Eigenvektoren aufgespannt wird

$$J_M = E\{\|\mathbf{x} - \mathbf{P}\mathbf{x}\|^2\} = E\{\|\mathbf{x}\|^2 - \|\mathbf{P}\mathbf{x}\|^2\} = E\{\|\mathbf{x}\|^2 - \langle \mathbf{P}\mathbf{x}, \mathbf{P}\mathbf{x}\rangle\}$$

es gilt:

$\mathbf{A}^T\mathbf{A} = \mathbf{I}$ ($\mathbf{A}^T$ ist orthogonal)

$\mathbf{A}^T\mathbf{R}_{\mathbf{xx}}\mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ mit: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \lambda_N$

die Projektionsmatrix $\mathbf{P}$ ist idempotent und symmetrisch

und daraus folgt:

$$\langle \mathbf{P}\mathbf{x}, \mathbf{P}\mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{P}^T\mathbf{P}\mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{P}^2\mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{P}\mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{A}\mathbf{I}_M\mathbf{A}^T\mathbf{x}\rangle$$

$$= \langle \mathbf{I}_M\mathbf{A}^T\mathbf{x}, \mathbf{A}^T\mathbf{x}\rangle = \text{Spur}(\mathbf{I}_M\mathbf{A}^T(\mathbf{x}\mathbf{x}^T)\mathbf{A})$$

Approximationsfehler

und eingesetzt in das Gütemaß ergibt:

$$
\begin{aligned}
J_M &= E\{\text{Spur}(\mathbf{x}\mathbf{x}^T)\} - E\{\text{Spur}(\mathbf{I}_M \mathbf{A}^T(\mathbf{x}\mathbf{x}^T)\mathbf{A})\} \\
&= \text{Spur}(\mathbf{R}_{\mathbf{xx}} - \mathbf{I}_M \underbrace{\mathbf{A}^T\mathbf{R}_{\mathbf{xx}}\mathbf{A}}_{\text{diag}(\lambda_1,\lambda_2,\ldots,\lambda_N)}) \\
&= \text{Spur}(\mathbf{R}_{\mathbf{xx}}) - \sum_{i=1}^{M} \lambda_i = \sum_{i=M+1}^{N} \lambda_i
\end{aligned}
$$

d.h., der Approximationsfehler entspricht der Summe der nicht berücksichtigten Eigenwerte.

Die Karhunen-Loéve-Transformation (KLT)

Die Karhunen-Loéve-Transformation ist somit definiert als:

$$\boxed{\mathbf{y} = \mathbf{A}^T \mathbf{x}} \qquad \text{KLT}$$

$$\boxed{\mathbf{x} = \mathbf{A}\mathbf{y}} \qquad \text{KLT}^{-1}$$

und es gilt:

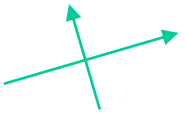
$$\mathbf{R}_{\mathbf{yy}} = E\{\mathbf{y}\mathbf{y}^T\} = E\{\mathbf{A}^T(\mathbf{x}\mathbf{x}^T)\mathbf{A}\} = \mathbf{A}^T\mathbf{R}_{\mathbf{xx}}\mathbf{A} = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$$

mit: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \lambda_N$

Die KLT kann wie hier auf der Grundlage der *Autokorrelationsmatrix*, oder aber auch aufbauend auf die *Autokovarianzmatrix* berechnet werden (der Erwartungswert wird zuvor abgezogen):

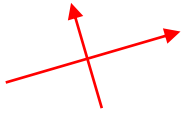
$$\boxed{\mathbf{y} = \mathbf{A}^T(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})}$$

$$\mathbf{y} = \mathbf{A}_1^T \mathbf{x}$$

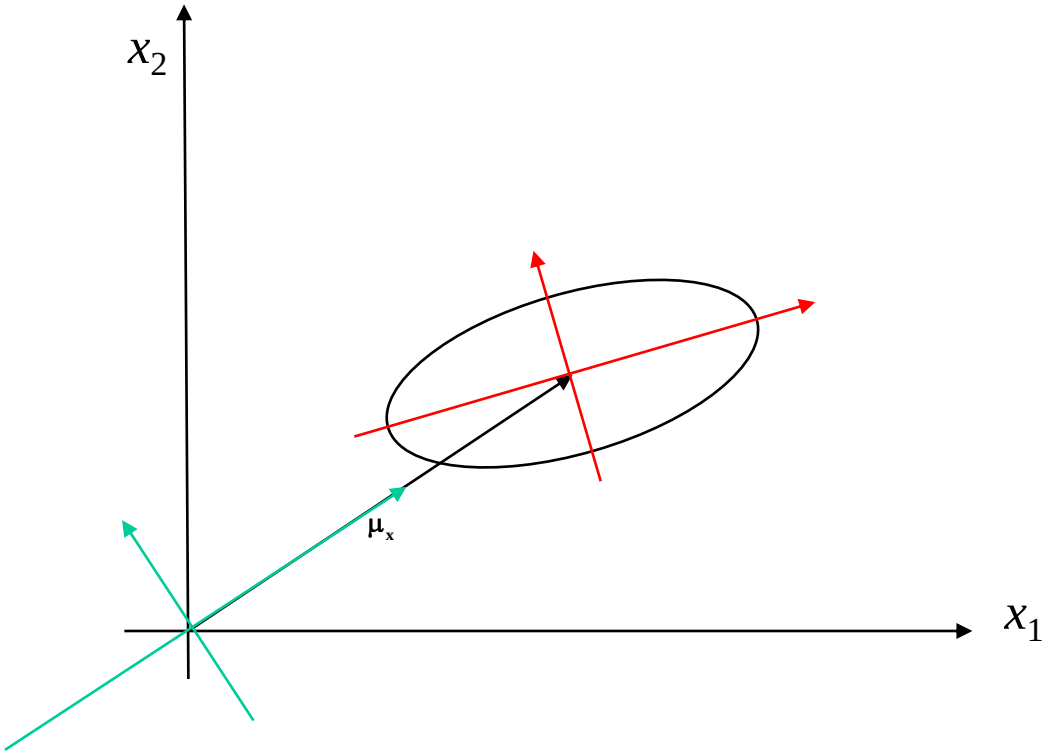


Die beste *lineare* Transformation

$$\mathbf{y} = \mathbf{A}_2^T (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$$



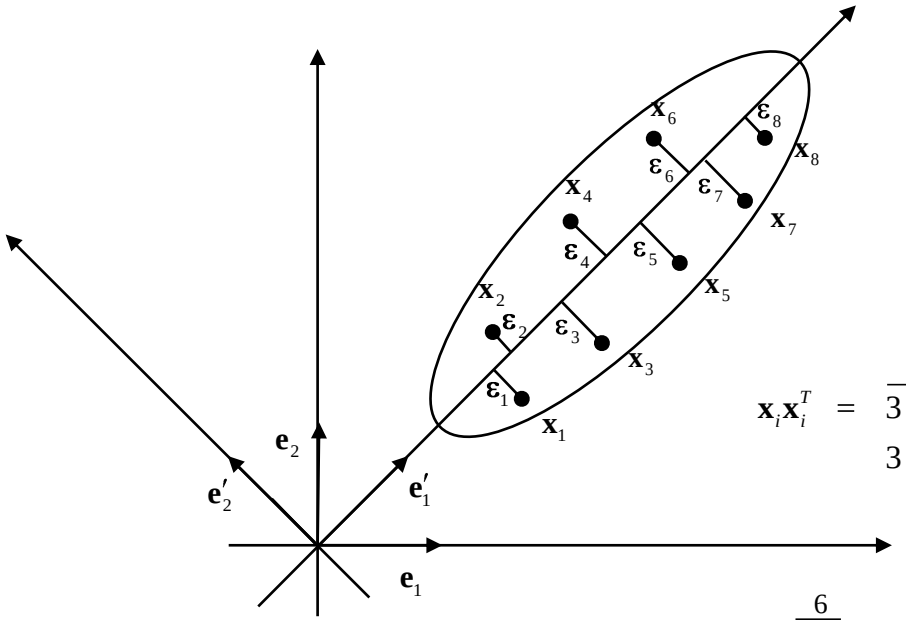
Die beste affine (Euklidische) Transformation



Beispiel zur KLT

KLT: $\mathbf{y} = \mathbf{A}^T\mathbf{x}$

KLT^{-1}: $\mathbf{x} = \mathbf{A}\mathbf{y}$



$$\mathbf{X} = \mathbf{x}_1 \mid \mathbf{x}_2 \mid \cdots \mid \mathbf{x}_6 = \left[\begin{array}{c|c|c|c|c|c} 3 & 5 & 7 & 8 & 10 & 12 \\ 3 & 7 & 5 & 10 & 8 & 12 \end{array}\right]$$

$\mathbf{x}_i\mathbf{x}_i^T$ =

	3	3		5	7		7	5		8	10		10	8		12	12
3	9	9	5	25	35	7	49	35	8	64	80	10	100	80	12	144	144
3	9	9	7	35	49	5	35	25	10	80	100	8	80	64	12	144	144

$$\mathbf{R}_{xx} = E\ \mathbf{x}_i\mathbf{x}_i^T \approx \frac{1}{6}\sum_{i=1}^{6} \mathbf{x}_i\mathbf{x}_i^T = \begin{bmatrix} 65{,}16 & 63{,}83 \\ 63{,}83 & 65{,}16 \end{bmatrix}$$

$$\lambda_1(\mathbf{R}_{xx}) = 129 \quad \lambda_2(\mathbf{R}_{xx}) = 1{,}33$$

$$\mathbf{e}'_1 = 1/\sqrt{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{e}'_2 = 1/\sqrt{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

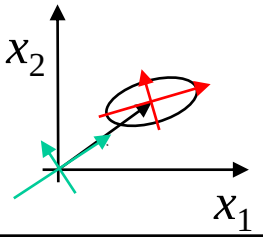
$$\mathbf{R}_{yy} = \mathbf{A}^T\mathbf{R}_{xx}\mathbf{A} = \begin{bmatrix} 129 & 0 \\ 0 & 1{,}33 \end{bmatrix} \quad \text{mit:} \quad \mathbf{A} = \mathbf{e}'_1, \mathbf{e}'_2 = 1/\sqrt{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

In diesem Fall führt die KLT von **x** auf die gleichen Ergebnisse wie die KLT von $(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})$!

Projektion in Richtung $\mathbf{e}'_1$: $\mathbf{P} = \mathbf{e}'_1\mathbf{e}'^T_1 = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ $\quad \mathbf{Q} = \mathbf{I} - \mathbf{P} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$

Cosinus-Transformation: $\mathbf{C} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ Identisch mit KLT!!

Aber hier nicht:



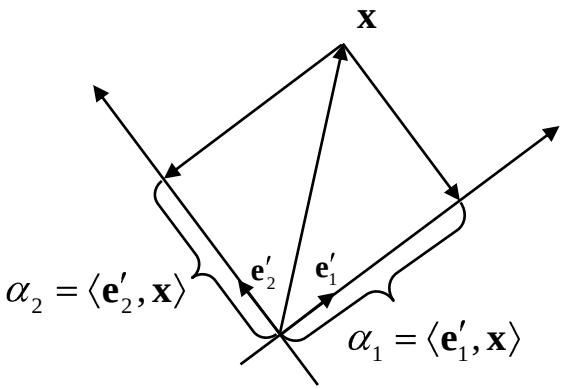
Zur Interpretation der KLT

Fourierreihe, Entwicklung nach den $\mathbf{e}'_i$

$$\mathbf{z} = \overbrace{\mathbf{A}\,\mathbf{I}_M \quad \underbrace{\mathbf{A}^T\mathbf{x}}}^{\sum \alpha_i \mathbf{e}'_i}$$

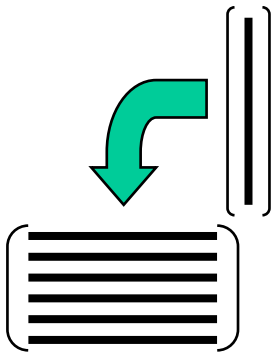
Berechnung der Fourierkoeffizienten α_i
Projektion auf den durch die $\mathbf{e}'_i$ aufgespannten Raum

Projektion auf den Unterraum



Projektion auf den Unterraum:

$$\mathbf{y} = \mathbf{I}_M \mathbf{A}^T \mathbf{x} = \begin{bmatrix} \langle \mathbf{e}'_1, \mathbf{x} \rangle \\ \langle \mathbf{e}'_2, \mathbf{x} \rangle \\ \vdots \\ \langle \mathbf{e}'_M, \mathbf{x} \rangle \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{bmatrix}$$



Fourierreihe (Entwicklung nach den Spaltenvektoren von **A**):

$$\mathbf{z} = \mathbf{A}\mathbf{w} = \sum_{i=1}^{M} \underbrace{\alpha_i}_{\text{FK}} \cdot \underbrace{\mathbf{e}'_i}_{\substack{\text{Spaltenvektoren} \\ \text{von } \mathbf{A}}}$$

Minimierung des Fehlers ist gleichbedeutend mit der Maximierung der Energie (Länge^2) im transformierten Bereich oder der Maximierung der Quadratsumme der Fourierkoeffizienten.

KLT für Bilder (2D)

Obiger Ansatz lässt sich direkt übertragen auf einen Vektor, welcher aus *gestapelten Zeilenvektoren* einer *Bildmatrix* der Dimension $N \times N$ besteht (da es ja nur um den Gesamtsummenfehler geht!). Damit ist jedoch ein Eigenwertproblem für symmetrische Matrizen der Dimension $N^2 \times N^2$ zu lösen. Ein Eigenwertproblem für eine Matrix $N \times N$ benötigt $O(N^3)$ Berechnungsschritte, also hier: $\boldsymbol{O(N^6)}$

Lässt sich hingegen ein Ensemble von Bildern $\mathbf{X} := \{\mathbf{X}_i\}$ der Dimension $N \times N$ durch das dyadische Produkt zweier eindimensionaler Ensembles der Dimension $N \times 1$

$$\mathbf{x}^1 := \{\mathbf{x}_i^1\} \qquad \mathbf{x}^2 := \{\mathbf{x}_i^2\}$$

modellieren gemäß:

$$\mathbf{X} := \mathbf{x}^1 \mathbf{x}^{2^T}$$

D.h. $\mathbf{X}$ ist separierbar!

KLT für Bilder (2D)

so lässt sich für jedes eindimensionale Ensemble eine KLT berechnen und man erhält:

$$\mathbf{Y} = \mathbf{A}^{1^T}(\mathbf{x}^1\mathbf{x}^{2^T})\mathbf{A}^2 = \mathbf{A}^{1^T}\mathbf{X}\mathbf{A}^2$$

2D-KLT bei separierbaren Bildern

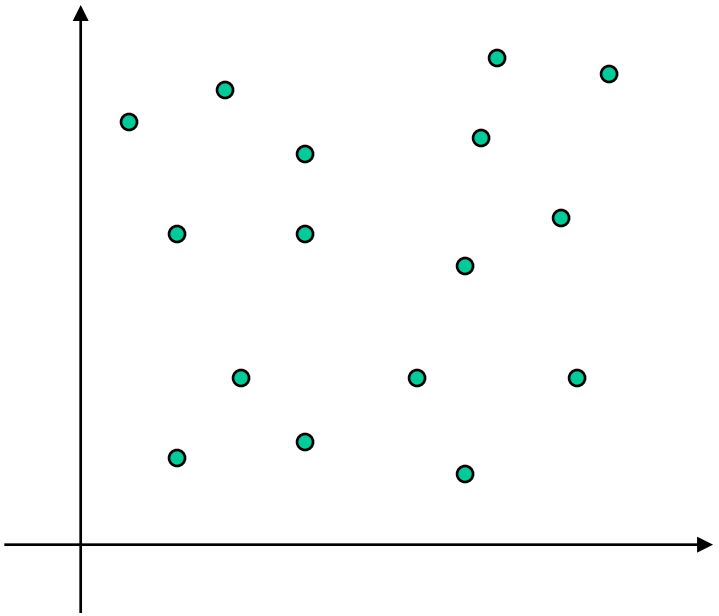
Damit sind nur 2 Eigenwertprobleme der Dimension $N \times N$ zu berechnen. Dies gibt einen Rechenzeitgewinn von: $O(N^6) / O(N^3) = O(N^3)$

Die Transformation mit separierbarem Kern reduziert sich ebenfalls im Aufwand, nämlich von $O(N^4)$ auf $O(2N^3)=O(N^3)$ (siehe DBV-I).

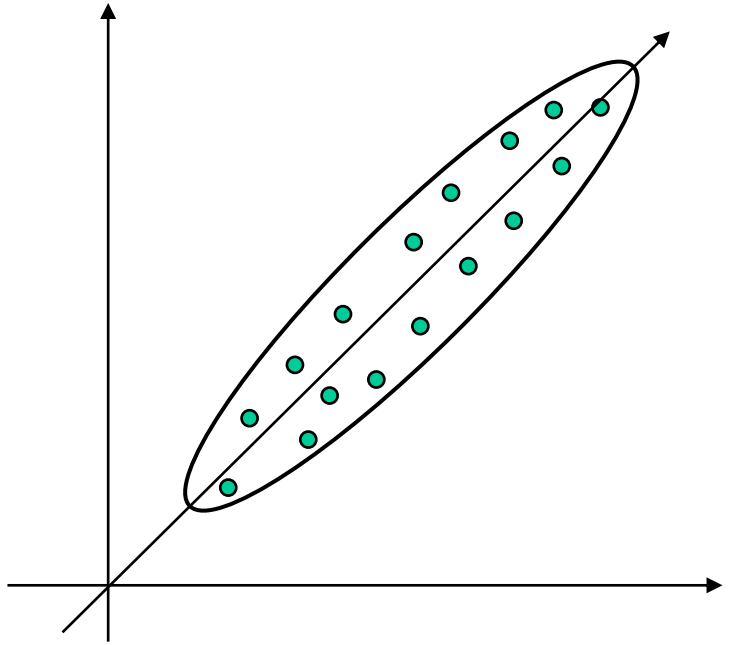
Eigenschaften der KLT

- Vorteile:
  - Die KLT ist optimal (bzgl. des quadr. Fehlers) im Hinblick auf bestmögliche Darstellung in Unterräumen mit orthogonaler Basis. Falls Vektorelemente stark korreliert sind, ergibt sich eine hohe Informationsverdichtung in wenigen Elementen der KLT. Die KLT profitiert von starken Korrelationen in den Vektorelementen.
  - Da $\mathbf{R}_{yy}$ eine Diagonalmatrix ist, sind die Werte in y unkorreliert!
- Nachteile:
  - Die KLT ist *datenabhängig* und muss für jeden Datensatz individuell berechnet werden.
  - Außerdem existiert für die KLT *kein schneller* Algorithmus.

Datenreduktion in Abhängigkeit vom Korrelationsgrad

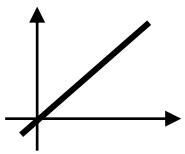


Daten unkorreliert (weißer Prozess)
KLT hat keine Bedeutung

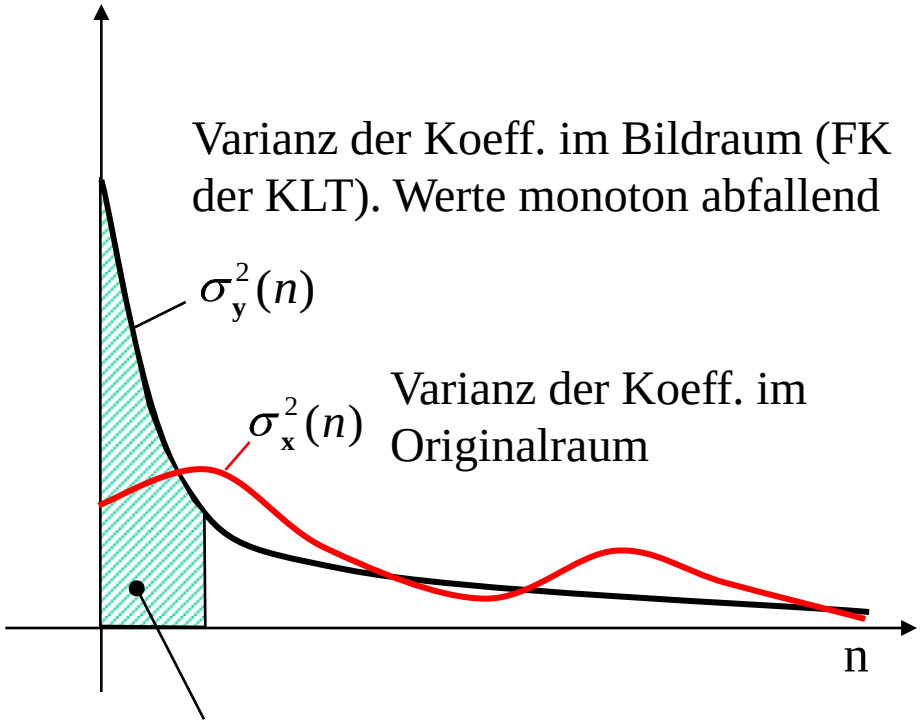


Daten stark korreliert. KLT bringt hohen Gewinn.

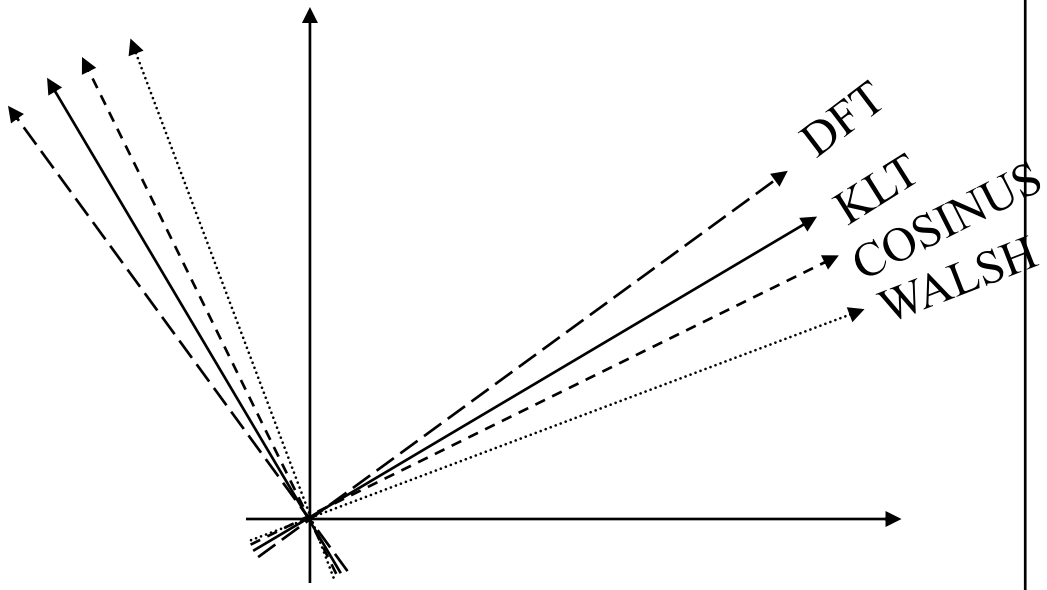
Extremfall: Bilder mit konstantem Grauwert
(hier genügt ein Vektor zur Darstellung)



Eigenschaften der KLT



Jede Teilsumme ist unabhängig von *M* maximal!



Verhalten unterschiedlicher unitärer Transformationen

Weitere Eigenschaften der KLT

Die KLT sorgt dafür, dass die Varianzen der transformierten Merkmale (Hauptdiagonalelemente der Kovarianzmatrix) maximal ungleichgewichtig sind (minimale Entropie):

$$\mathbf{y} = \mathbf{A}^T(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$$

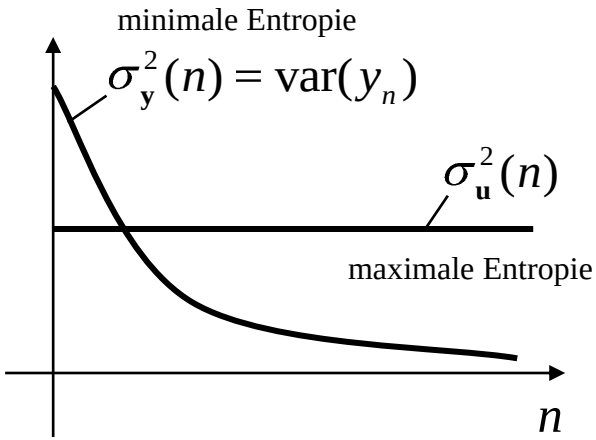
Eine Maximierung der Entropie, oder eine konstante Varianz aller Merkmale erreicht man durch eine ***Whitening-Transformation***:

$$\mathbf{u} = \Lambda^{-1/2}\mathbf{A}^T(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) \qquad \Lambda^{-1/2} = \mathrm{diag}(\lambda_1^{-1/2}, \lambda_2^{-1/2}, \ldots, \lambda_N^{-1/2})$$

Alle Merkmale haben die gleiche Varianz var(u_i)=1 (sphärisch invariante Verhältnisse). Die Energie ist gleichmäßig auf alle Merkmale verteilt.

Durch Multiplikation mit einer Diagonalmatrix bleiben die Variablen unkorreliert!

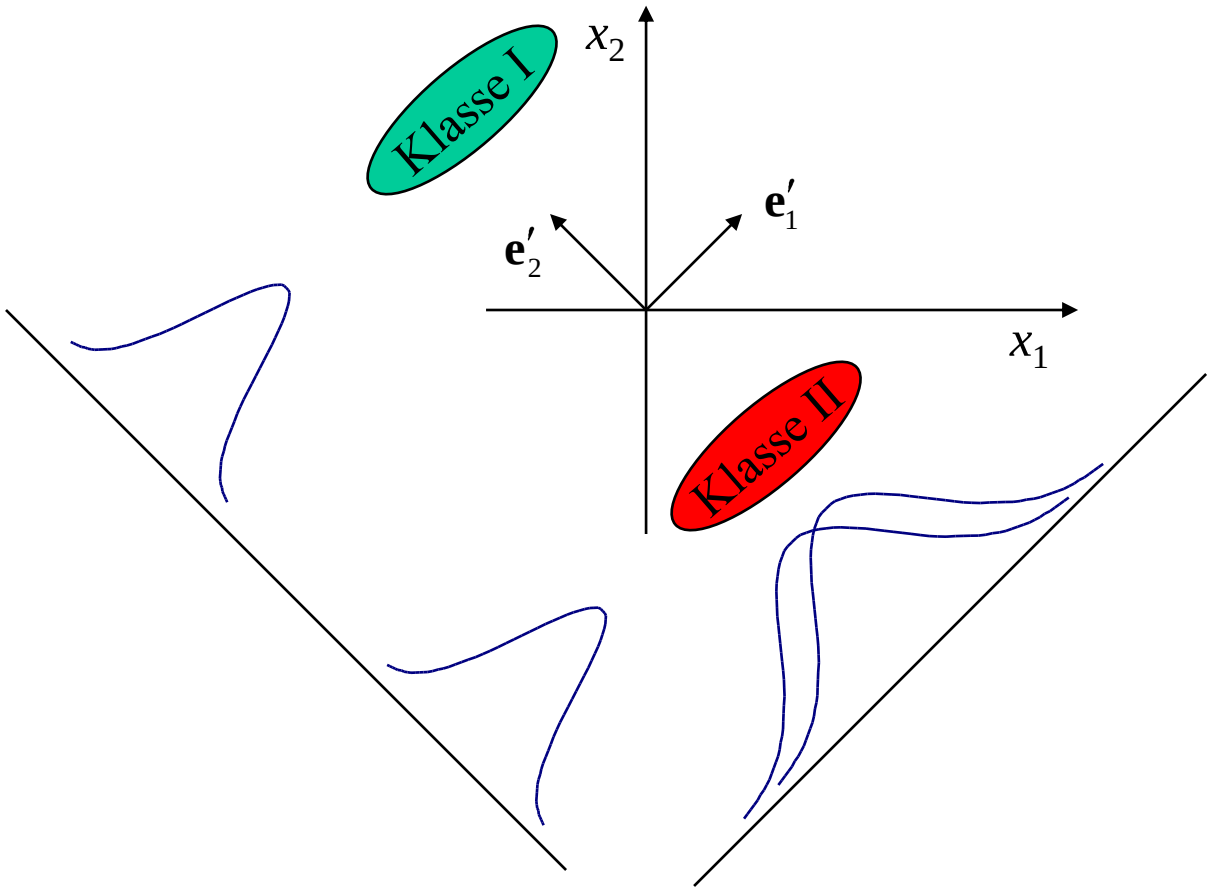
Whitening wird z.Bsp. benötigt, um größtmögliche Robustheit beim Wegfall einer Komponente zu bekommen (z.Bsp. In der Übertragungstechnik).



Weitere Eigenschaften der KLT

Die Optimalität der KLT bezüglich des minimalen Fehlerquadrats führt zu einer bestmöglichen Informationsverdichtung *eines* Ensembles und erlaubt uns eine Selektion der *M* dominanten Merkmalen von *N* Beobachtungswerten. Antwort auf die Frage: wie lassen sich alle Daten bestmöglich *repräsentieren*?

Dies führt jedoch nicht immer zwingend auf eine bestmögliche Klassenseparation, wenn mehrere Klassen zu unterscheiden sind. Eine diesbezügliche Optimierung führt zur sogenannten ***Diskriminanzanalyse***. Antwort auf die Frage: wie lassen sich die Daten bestmöglich *diskriminieren*?



In diesem Beispiel überlappen die Merkmale des ersten Eigenvektors, während das Merkmal des zweiten Eigenvektors die Klassen trennt!

Annahme: Varianz entlang $\mathbf{e}_1$´ grösser als Varianz entlang $\mathbf{e}_2$´.

Kovarianzmatrix wird vom gesamten Ensemble gerechnet, da es ja nur eine Art der Merkmalsauswahl geben kann!