Data Mining für Technische Anwendungen – Hauptkomponentenanalyse

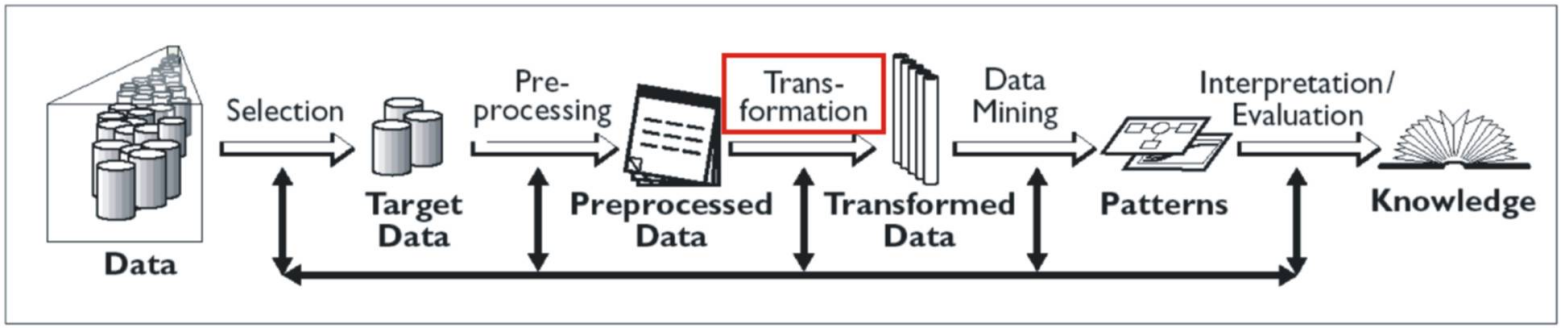
PD Dr.-Ing. habil. Sven Tomforde
Prof. Dr. Bernhard Sick

Universität Kassel
Fachbereich Elektrotechnik / Informatik
Fachgebiet „Intelligent Embedded Systems“

WS 2017/2018

Intelligent Embedded Systems

Worum geht es?



Datentransformation

Aus der Definition von KDD: Datenreduktion und **Datenprojektion mit dem Ziel der Verdichtung relevanter Informationen in einer geringeren Zahl von Variablen (Dimensionsreduktion)** und Identifikation relevanter Attribute (Merkmalsselektion)

Intelligent Embedded Systems

Agenda

- Motivation und Grundlagen
- Beispiel
- Abschließende Bemerkungen

Intelligent Embedded Systems

Motivation und Grundlagen

Intelligent Embedded Systems

Motivation – 1

- **gegeben:** ein Datensatz $\mathbf{X}$ mit Mustern $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$; die Muster sind D-dimensional, d. h., es gibt D Merkmale.
- **gesucht:** ein Datensatz $\mathbf{Y}$ mit Mustern $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M$; die Muster sind ebenfalls D-dimensional, d. h., es gibt D Merkmale, und $|\mathbf{X}| = |\mathbf{Y}|,\ N = M$;

  Der Informationsgehalt ist der gleiche wie im Datensatz $\mathbf{X}$, aber: der Informationsgehalt des Datensatzes $\mathbf{Y}$ ist gespeichert in den ersten, wenigen Merkmalen (Dimensionen).

  Man spricht hier auch von *Meta-Merkmalen*.
- **Hauptkomponentenanalyse:** eine Methode, einen solchen Datensatz zu finden.

Intelligent Embedded Systems

Motivation – 2

Was heißt *Informationsgehalt*?

- **Annahme:** hoher Informationsgehalt entspricht hoher Varianz!

Hauptziel der Hauptkomponentenanalyse:

- **Dimensionsreduktion:** es können weniger wichtige Dimensionen weggelassen werden, d. h., die Zahl D' der Meta-Merkmale im transformierten Datensatz ist $D' \ll D$.

Intelligent Embedded Systems

Motivation – 3

Nutzen der Hauptkomponentenanalyse.

- **Zeitersparnis**: durch Einsatz von DM-Algorithmen auf reduzierten Datensätzen.
- **Merkmalsselektion**: sehr einfach durch Wahl der wichtigsten Meta-Merkmale.
- **Verständnis**: besseres Erkennen von Strukturen in Daten z. B. durch Visualisierung des Datensatzes im Raum der zwei oder drei wichtigsten Meta-Merkmale.

Andere Namen für Hauptkomponentenanalyse:

- Principal Component Analysis (PCA), Hotelling Transformation, Karhunen-Loève-Transformation, ...

Intelligent Embedded Systems

Grundlagen – 1

Um einen Datensatz zu transformieren, wird zunächst das *arithmetische Mittel* jedes Merkmals i gebildet:

$$\mu_i := \frac{1}{N} \sum_{n=1}^{N} x_{in}$$

Anstelle der originalen Muster werden dann die mittelwertbereinigten Muster weiter verwendet, d. h.,

$$\forall_{n=1\ldots N} \forall_{i=1\ldots D} : x'_{in} = x_{in} - \mu_i.$$

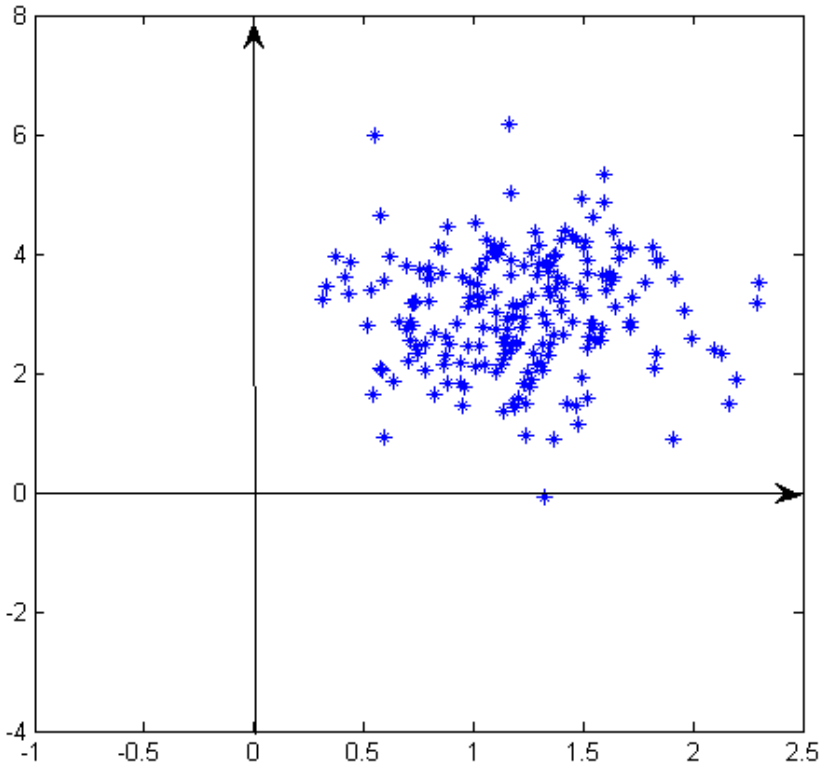
Oder auch mit

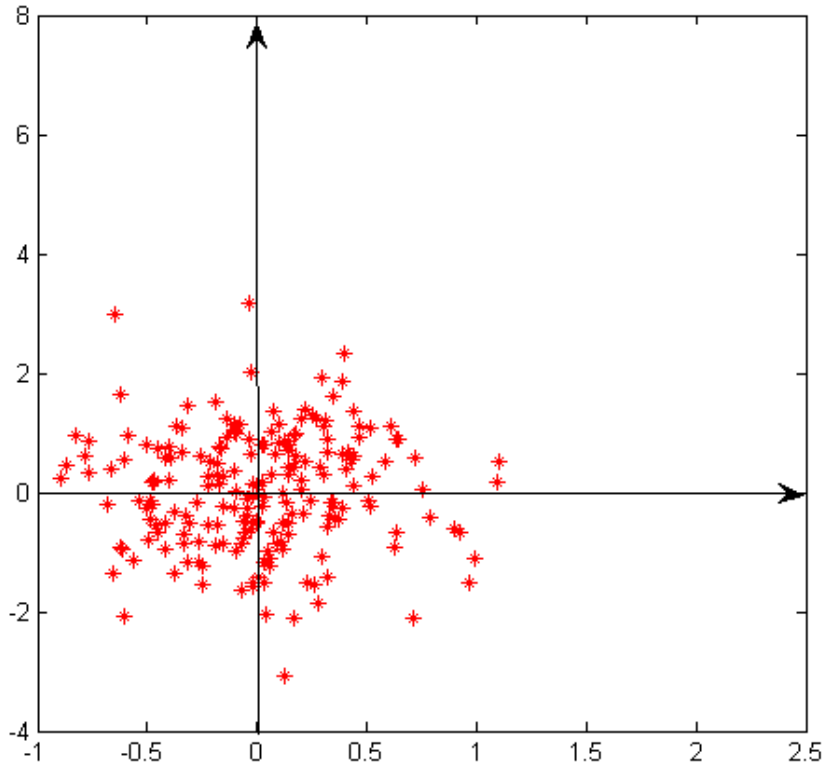
$$\boldsymbol{\mu} := \begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_D \end{pmatrix},$$

$\boldsymbol{\mu} - \mathbf{x}_n$ für $n = 1, 2, \cdots, N$.Dies entspricht geometrisch einer Translation (Verschiebung) der Daten.

Intelligent Embedded Systems

Beispiel zu Grundlagen – 1





Die blauen Punkte stellen den Originaldatensatz dar, mit den arithmetischen Mitteln $\mu_1 = 1,2$ und $\mu_2 = 3$. Die roten Datenpunkte stellen den Originaldatensatz nach Translation dar.

Intelligent Embedded Systems

Grundlagen – 2

Die *empirische Varianz* eines Merkmals i ist dann:

$$\sigma_i^2 := \frac{1}{N-1} \sum_{n=1}^{N} {x'}_{in}^2$$

Somit ist die *empirische Standardabweichung* des Merkmals i:

$$\sigma_i := \sqrt{\sigma_i^2}$$

Intelligent Embedded Systems

Grundlagen – 3

Benötigt wird auch die Kovarianz zweier Merkmale i und j:

$$s_{ij} := \frac{1}{N-1} \sum_{n=1}^{N} x'_{in} \cdot x'_{jn}$$

Eine Kovarianz s_{ii} (also eines Merkmals mit sich selbst) ist natürlich wieder die Varianz. Außerdem gilt $s_{ij} = s_{ji}$.

Intelligent Embedded Systems

Grundlagen – 4

Die Kovarianz wird immer paarweise, d. h., für zwei Merkmale berechnet.

Bei D-dimensionalen Daten gibt es $\frac{D!}{(D-2)!\cdot 2!}$ viele Kovarianzen.

Schreibt man die Kovarianzen in eine Matrix $\mathbf{C}$

$$\mathbf{C} := \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1D} \\ s_{21} & s_{22} & \dots & s_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ s_{D1} & s_{D2} & \dots & s_{DD} \end{pmatrix},$$

so ist diese Matrix symmetrisch. In der Diagonalen stehen die Varianzen der Merkmale.

Intelligent Embedded Systems

Grundlagen – 5

Ein *Eigenvektor* $\mathbf{v}$ einer solchen Matrix $\mathbf{C}$ ist ein D-dimensionaler Vektor, für den gilt:

$$\mathbf{C} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}.$$

Dabei heißt $\lambda \in \mathbb{R}$ *Eigenwert* zum Eigenvektor $\mathbf{v}$.

Es gilt:

- $\mathbf{C}$ hat D Eigenvektoren $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D$ mit den entsprechenden Eigenwerten $\lambda_1, \lambda_2, \ldots, \lambda_D$.
- Ohne Beschränkung der Allgemeinheit gelte für $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D$: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$.
- Die Eigenvektoren stehen senkrecht aufeinander, d. h., sie sind *orthogonal* zueinander.
- Vielfache eines Eigenvektors sind auch Eigenvektoren, wir verwenden diejenigen, die auf die Länge 1 normiert sind.

Intelligent Embedded Systems

Grundlagen – 6

Wichtigste Eigenschaft der Eigenvektoren:

- Der Eigenvektor mit dem höchsten Eigenwert gibt die Richtung an, in der der Datensatz die höchste Varianz aufweist.
- Der Eigenvektor mit dem zweithöchsten Eigenwert gibt eine dazu orthogonale Richtung an, in der der Datensatz die zweithöchste Varianz aufweist.
- usw.

Die Varianzen werden durch die jeweiligen Eigenwerte beschrieben!!!

Varianz $\rightarrow$ Informationsgehalt!

Intelligent Embedded Systems

Grundlagen – 7

Wie bekommt man Eigenwerte und Eigenvektoren?

Mathematische Bibliotheken für verschiedene Programmiersprachen bieten numerisch stabile Verfahren, die meist bereits längennormierte Eigenvektoren mit Eigenwerten liefern.

Intelligent Embedded Systems

Grundlagen – 8

Als nächstes wird eine bestimmte Zahl $D' \leq D$ von Eigenvektoren zur Transformation der Daten ausgewählt:

- Alle Eigenvektoren ($D' = D$) werden gewählt, wenn das Ziel der Hauptkomponentenanalyse z. B. eine Hauptachsentransformation zur Dekorrelation der Daten ist.

  In diesem Fall werden die mittelwertbereinigten Muster folgendermaßen transformiert:

$$\mathbf{y}_k = \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_D^{\mathrm{T}} \end{pmatrix} \mathbf{x}'_k.$$

  Dies entspricht einer Rotation der Daten.

Intelligent Embedded Systems

Grundlagen – 9

- Eine geringere Zahl von Eigenvektoren (meist $D' \ll D$) wird gewählt, wenn das Ziel der Hauptkomponentenanalyse eine Datenreduktion ist.

  In diesem Fall werden die mittelwertbereinigten Muster folgendermaßen transformiert:

$$\mathbf{y}_k = \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_{D'}^{\mathrm{T}} \end{pmatrix} \mathbf{x}'_k.$$

  Die transformierten Muster $\mathbf{y}_k$ haben also nur D' Dimensionen.

Intelligent Embedded Systems

Grundlagen – 10

Eine Rücktransformation der Daten ist möglich (z. B. üblich in der Bildverarbeitung, wo PCA u. a. zur Datenkompression eingesetzt wird), für $D' < D$ allerdings nur mit Informationsverlust.

Intelligent Embedded Systems

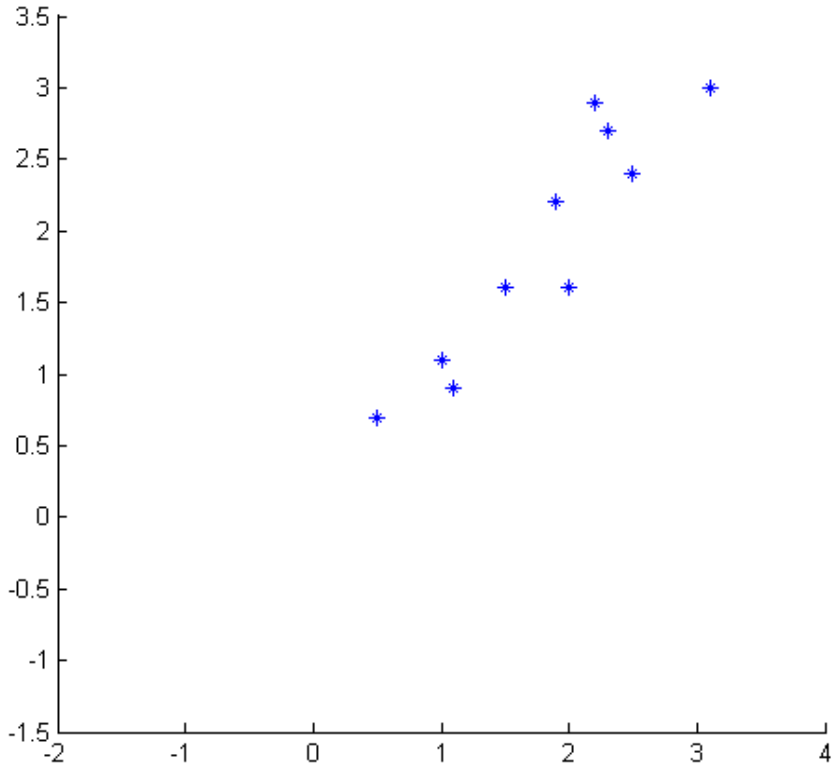
Beispiel

Intelligent Embedded Systems

Beispiel – 1

Zweidimensionaler Datensatz:

Muster	Merkmal 1	Merkmal 2
$\mathbf{x}_1$	2.5	2.4
$\mathbf{x}_2$	0.5	0.7
$\mathbf{x}_3$	2.2	2.9
$\mathbf{x}_4$	1.9	2.2
$\mathbf{x}_5$	3.1	3.0
$\mathbf{x}_6$	2.3	2.7
$\mathbf{x}_7$	2.0	1.6
$\mathbf{x}_8$	1.0	1.1
$\mathbf{x}_9$	1.5	1.6
$\mathbf{x}_{10}$	1.1	0.9



Intelligent Embedded Systems

Beispiel – 2

In jeder Dimension wird der Mittelwert von den Daten abgezogen.

Der Mittelwert der transformierten Daten ist dann 0.

$\mathbf{x}'_1$	0.69	0.49
$\mathbf{x}'_2$	-1.31	-1.21
$\mathbf{x}'_3$	0.39	0.99
$\mathbf{x}'_4$	0.09	0.29
$\mathbf{x}'_5$	1.29	1.09
$\mathbf{x}'_6$	0.49	0.79
$\mathbf{x}'_7$	0.19	-0.31
$\mathbf{x}'_8$	-0.81	-0.81
$\mathbf{x}'_9$	-0.31	-0.31
$\mathbf{x}'_{10}$	-0.71	-1.01

Intelligent Embedded Systems

Beispiel – 3

Anschließend wird die Kovarianzmatrix berechnet:

$$\mathbf{C} = \begin{pmatrix} 0.617 & 0.615 \\ 0.615 & 0.717 \end{pmatrix}$$

Da die Elemente abseits der Diagonalen positiv sind, besteht ein positiver Zusammenhang zwischen den beiden Merkmalen (vgl. Korrelationskoeffizient).

Intelligent Embedded Systems

Beispiel – 4

Die Eigenwerte und Eigenvektoren der Matrix $\mathbf{C}$ sind:

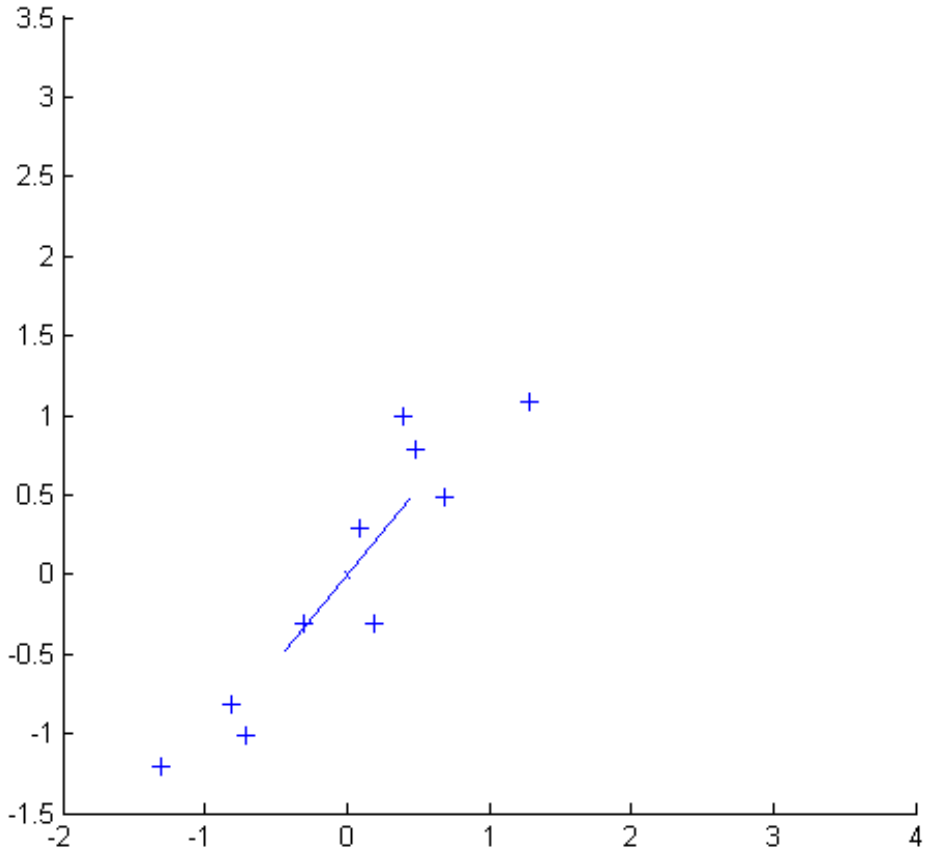
$$\mathbf{v}_1 = \begin{pmatrix} -0.678 \\ -0.735 \end{pmatrix} \text{ mit } \lambda_1 = 1.284$$

$$\mathbf{v}_2 = \begin{pmatrix} -0.735 \\ 0.678 \end{pmatrix} \text{ mit } \lambda_2 = 0.049$$

Die Eigenvektoren haben Länge Eins und stehen senkrecht aufeinander; $\mathbf{v}_1$ (höherer Eigenwert) beschreibt die erste Hauptkomponente, $\mathbf{v}_2$ die zweite.

Intelligent Embedded Systems

Beispiel – 5



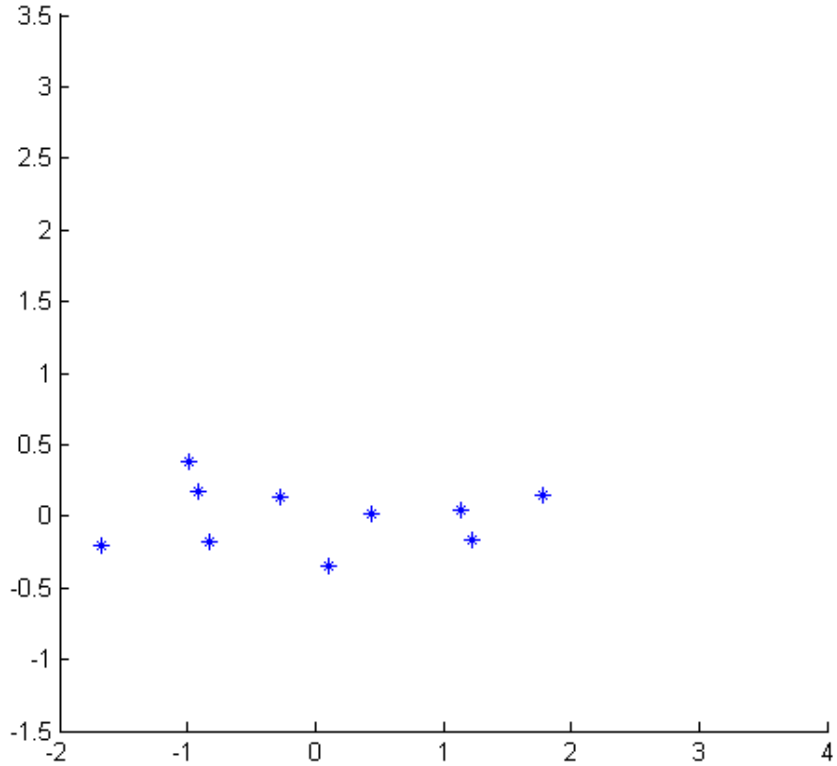
Vom Mittelwert ausgehend ist hier jeder Eigenvektor in beide Richtungen gezeichnet; Länge entspricht dem Eigenwert.

Intelligent Embedded Systems

Beispiel – 6

Transformation der Daten unter Verwendung beider Eigenvektoren:

$\mathbf{y}_1$	-0.828	-0.175
$\mathbf{y}_2$	1.778	0.143
$\mathbf{y}_3$	-0.992	0.384
$\mathbf{y}_4$	-0.274	0.130
$\mathbf{y}_5$	-1.676	-0.209
$\mathbf{y}_6$	-0.913	0.175
$\mathbf{y}_7$	0.099	-0.350
$\mathbf{y}_8$	1.145	0.046
$\mathbf{y}_9$	0.438	0.018
$\mathbf{y}_{10}$	1.224	-0.163



Intelligent Embedded Systems

Beispiel – 7

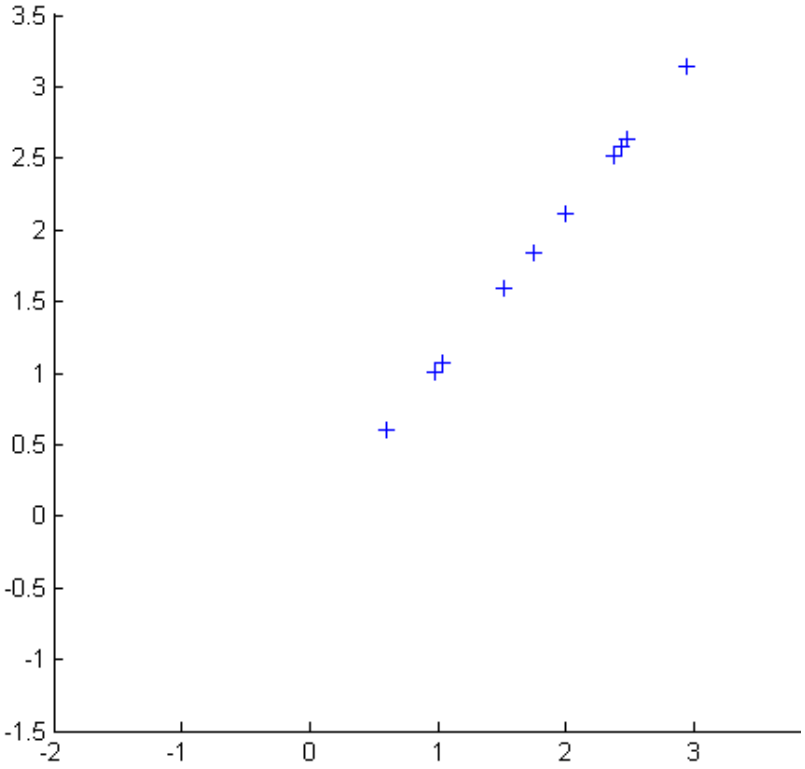
Transformation der Daten unter Verwendung des Eigenvektors mit dem höheren Eigenwert:

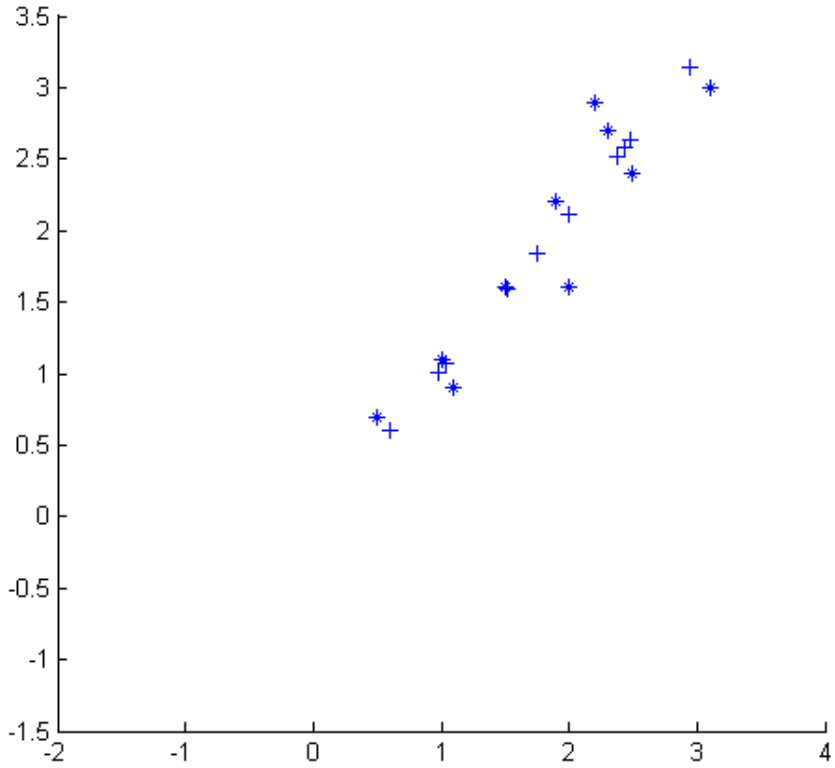
$\mathbf{y}_1$	-0.828
$\mathbf{y}_2$	1.778
$\mathbf{y}_3$	-0.992
$\mathbf{y}_4$	-0.274
$\mathbf{y}_5$	-1.676
$\mathbf{y}_6$	-0.913
$\mathbf{y}_7$	0.099
$\mathbf{y}_8$	1.145
$\mathbf{y}_9$	0.438
$\mathbf{y}_{10}$	1.224

... entspricht natürlich der ersten Spalte in der Tabelle der vorausgehenden Folie!!!

Intelligent Embedded Systems

Beispiel – 8





Rücktransformation dieser Daten zeigt den Informationsverlust!
(+: Rücktransformierte Daten, * Originaldatensatz)

(Entspricht Projektion der Daten auf die durch die erste Hauptkomponente beschriebene Achse.)

Intelligent Embedded Systems

Abschließende Bemerkungen

Intelligent Embedded Systems

Auswahl von Hauptkomponenten

Nach welchen Kriterien wird eine geeignete Zahl D' von Hauptkomponenten zur Datenreduktion bestimmt?

- Die Summe der Eigenwerte der wichtigsten D' Eigenvektoren sollte einen gewissen Anteil (z. B. mindestens 0.75) an der Summe aller D Eigenwerte ausmachen.
- Dimensionen werden weggelassen, wenn die Eigenwerte der entsprechenden Eigenvektoren geringer als der Durchschnitt aller Eigenwerte sind.
- Die Eigenwerte werden entsprechend der absteigenden Wichtigkeit der Eigenvektoren dargestellt. Wird diese Kurve an einer Stelle signifikant flacher, so werden die entsprechenden Dimensionen weggelassen (sog. Ellbogen- oder Kniekriterium).
- ...

Intelligent Embedded Systems

Veranschaulichung

Beispiele:

- *Applet*
  (`http://www.cs.mcgill.ca/~sqrt/dimr/dimreduction.html`)

Intelligent Embedded Systems

Ende

–

Noch Fragen zum Thema Hauptkomponentenanalyse?

Intelligent Embedded Systems