Bonner Zentrum für Lehrerbildung (BZL)

Singulärwertzerlegung. Theorie und Anwendung

Bachelorarbeit im Fach: Mathematik

vorgelegt von Noah Pferdekamp
Matrikelnummer: 123456

Erstbetreuer: Prof. Dr. Philipp Hieronymi
Zweitgutachter: Prof. Dr. Thoralf Räsch
Mathematisches Institut

Wintersemester 2024/2025
Bonn, den 8. März 2025

Selbstständigkeitserklärung

Ich versichere hiermit, dass die Bachelorarbeit mit dem Titel „Singulärwertzerlegung. Theorie und Anwendung“ von mir selbst und ohne jede unerlaubte Hilfe selbstständig angefertigt wurde, dass sie noch an keiner anderen Hochschule zur Prüfung vorgelegen hat und dass sie weder ganz noch in Auszügen veröffentlicht worden ist. Die Stellen der Arbeit — einschließlich Tabellen, Karten, Abbildungen usw. —, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall kenntlich gemacht.

Bonn, den 8. März 2025
Signatur

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Notation

$\overline{z}$	Komplexe Konjugation
$\mathfrak{Re}$	Realteil
$\mathfrak{Im}$	Imaginärteil
$\mathcal{L}$	Lineare Hülle
$\|\|v\|\|$	Norm
$\langle v, w \rangle$	Skalarprodukt
$\mathbf{0}$	Nullmatrix
$\text{diag}(\lambda_1, \dots, \lambda_n)$	Diagonalmatrix
$\text{rg}(X)$	Rang
$\text{df}(X)$	Defekt
$\boldsymbol{I}$	Einheitsmatrix
$x \ll y$	viel kleiner
$\text{proj}_u(x)$	orthogonale Projektion von x auf u

Abkürzungen

SVD	Singular value decomposition (Singulärwertzerlegung)
PCA	Principal component analysis (Hauptkomponentenanalyse)
u.d.B.	unter den Bedingungen

1 Einleitung

2 Mathematische Theorie

In diesem Kapitel wird zunächst unter Verwendung vorher eingeführter Sätze und Definitionen die Existenz und eine fundamentale Eigenschaft der Singulärwertzerlegung formal bewiesen. Anschließend wird an einem konkreten Beispiel die Berechnung durchgeführt und die SVD visualisiert. Um das Kapitel abzuschließen, erfolgt eine Beschreibung der wichtigsten beiden Arten der Singulärwertzerlegung.

2.1. Beweisführung

Es wird davon ausgegangen, dass der Leser[1] mit den Grundlagen der linearen Algebra vertraut ist, insbesondere mit Matrizen und ihren Eigenschaften. Bekannte Definitionen werden nicht erneut aufgeführt, die einzige Ausnahme bildet Definition 2.1, da diese für jegliche Beweisführung und für das Verständnis in diesem Kapitel unerlässlich ist und deswegen eine Auffrischung sinnvoll erscheint.

Definition 2.1.
Sei $n \in \mathbb{N}$ und $A \in \mathbb{R}^{n \times n}$. Für

$$Av = \lambda v$$

heißen die Lösungen $\mathbb{R}^n \ni v \neq 0$ *Eigenvektoren* und die zugehörigen λ *Eigenwerte*.

Vorausgesetzte Sätze werden ohne weiteren Beweis verwendet, aber in wichtigen Fällen dennoch vor Verwendung kurz rekapituliert, wie in Wiederholung 2.2 verdeutlicht.

[1] Aus Gründen der besseren Lesbarkeit wird das generische Maskulinum verwendet, wobei alle Geschlechter mit eingeschlossen sind.

WIEDERHOLUNG 2.2 (Basisergänzungssatz).
Sei V ein beliebiger Vektorraum, $L \subseteq V$ linear unabhängig und $E \subseteq V$ ein Erzeugendensystem von V. Dann kann L durch Elemente aus E zu einer Basis von V ergänzt werden.

Um die Existenz der Singulärwertzerlegung für beliebige Matrizen zu beweisen, bedarf es der Hilfe eines anderen Satzes, des sogenannten Spektralsatzes. Dieser hat keine eindeutige Ausführung, sondern beschreibt vielmehr mehrere verwandte Aussagen der Mathematik, wobei sich in dieser Arbeit auf seine Folgerungen für symmetrische Matrizen beschränkt wird. Es ist ebenfalls wichtig zu betonen, dass der Spektralsatz zwar hier „nur“ für den Beweis der Singulärwertzerlegung verwendet wird, eine Bezeichnung als Hilfssatz jedoch irreführend wäre, da der Satz für sich genommen bereits eine bedeutende Aussage der linearen Algebra und Funktionalanalysis darstellt. Um den Spektralsatz für symmetrische Matrizen einführen und anschließend beweisen zu können, benötigen wir zunächst Lemma 2.3 und Wiederholung 2.4.

LEMMA 2.3.
Sei $z \in \mathbb{C}$ und $a, b \in \mathbb{R}$ mit $z = a + bi$. $z = \overline{z}$ gilt genau dann, wenn $z \in \mathbb{R}$.

Beweis. „⇒“ Durch $z = \overline{z}$ gilt

$$\begin{aligned} & a + bi = a - bi \\ \Leftrightarrow \quad & 2bi = 0. \end{aligned}$$

Da $i \neq 0$ muss $b = 0$, womit $\mathfrak{Im}(z) = 0$. Also ist $z = \mathfrak{Re}(z) \in \mathbb{R}$.
„⇐“ Folgt direkt aus der Definition. □

WIEDERHOLUNG 2.4 (Gram-Schmidtsches Orthonormalisierungsverfahren).
Sei V ein euklidischer Vektorraum und $\{u_1, \dots, u_n\}$ eine Menge von linear unabhängigen Vektoren in V. Dann kann eine Menge $\{v_1, \dots, v_n\}$ aus Vektoren in V konstruiert werden, sodass $\{v_1, \dots, v_n\}$ orthonormal ist und

$$\mathcal{L}\{v_1, \dots, v_n\} = \mathcal{L}\{u_1, \dots, u_n\}.$$

Satz 2.5 (Spektralsatz).
Sei $n \in \mathbb{N}$ und $A \in \mathbb{R}^{n\times n}$ quadratisch und symmetrisch. Dann gilt:

(i) A hat reelle Eigenwerte.

(ii) Es existiert eine orthogonale Matrix $R \in \mathbb{R}^{n\times n}$, sodass $R^{-1}AR = R^TAR = \Lambda \in \mathbb{R}^{n\times n}$ diagonal ist.

Beweis. Die Behauptungen werden nacheinander bewiesen (vgl. [Cra22]).

Zu (i). Sei $\lambda \in \mathbb{C}$ ein Eigenwert von A mit zugehörigem Eigenvektor $v \in \mathbb{C}^n$. Dann ist mit Definition 2.1

$$\begin{aligned} & Av = \lambda v \\ \Leftrightarrow \quad & A\overline{v} = \overline{\lambda}\overline{v}, \end{aligned} \tag{2.1}$$

da $A \in \mathbb{R}$ und somit $A = \overline{A}$ nach Lemma 2.3. Nun gilt zum einen

$$(Av)^T\overline{v} = \big(\lambda v\big)^T\overline{v} = \lambda v^T\overline{v} \tag{2.2}$$

und zum anderen

$$(Av)^T\overline{v} = v^TA^T\overline{v} \overset{A \text{ sym.}}{=} v^TA\overline{v} \overset{(2.1)}{=} v^T\overline{\lambda}\overline{v} = \overline{\lambda}v^T\overline{v}. \tag{2.3}$$

Mit (2.2)=(2.3) ergibt sich

$$\lambda v^T\overline{v} = \overline{\lambda}v^T\overline{v}.$$

Da $v \neq 0$ ist erhalten wir

$$\lambda = \overline{\lambda}.$$

Nach Lemma 2.3 ist dann $\lambda \in \mathbb{R}$, wodurch auch $v \in \mathbb{R}^n$ sein muss. □

Zu (ii). Induktion über $n \in \mathbb{N}$:
Induktionsanfang. Für $n = 1$ sind A und R Skalare. Setze $R = 1$. Damit ist R orthogonal, da $R^{-1} = R^T$ und $\mathbb{R} \ni A = R^{-1}AR$ trivialerweise diagonal.
Induktionshypothese. Die Behauptung (ii) gelte für festes, beliebiges $\mathbb{N} \ni n - 1$. Es soll gezeigt werden, dass sie dann auch für n gilt.
Induktionsschritt. Sei λ_1 ein beliebiger Eigenwert von A mit zugehörigem normalisiertem Eigenvektor v_1, also $\|v_1\| = 1$. Nach (i) gilt $\lambda_1 \in \mathbb{R}$ und $v_1 \in \mathbb{R}^n$. Mit dem Basisergänzungssatz (Wiederholung 2.2) kann v_1 durch Vektoren $u_2, \dots, u_n$ zu einer Basis von $\mathbb{R}^n$ ergänzt werden. Nun kann das Gram-Schmidt'sche Orthonormalisierungsverfahren (Wiederholung 2.4) angewendet

werden, wodurch eine orthonormale Basis $\{v_1, \dots, v_n\}$ von $\mathbb{R}^n$ konstruiert wird. Der Leser wird daran erinnert, dass orthonormale Vektoren normiert und orthogonal sind. Sei

$$P = \begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & | & | \end{bmatrix} \in \mathbb{R}^{n \times n}$$

mit $v_1, \dots, v_n$ als Spaltenvektoren und setze $\mathbb{R}^{n \times n} \ni B = P^{-1}AP = P^TAP$.

Das Ziel ist, die Induktionshypothese auf eine symmetrische Untermatrix $C \in \mathbb{R}^{(n-1)\times(n-1)}$ von B anzuwenden. Dafür wird zunächst die Symmetrie von B gezeigt:

$$B^T = \left(P^TAP\right)^T = (AP)^TP = P^TA^TP \overset{A \text{ sym.}}{=} P^TAP = B.$$

Betrachte jetzt die erste Spalte von B. Die erste Spalte einer beliebigen Matrix erhält man durch Multiplikation mit dem kanonischen Einheitsvektor $\mathbf{e}_1$:

$$\begin{aligned}
B\mathbf{e}_1 &= P^TAP\mathbf{e}_1 \\
&= P^TAv_1 && (v_1 \text{ ist die erste Spalte von P}) \\
&= P^T\lambda_1 v_1 && (\lambda_1 \text{ ist der Eigenwert zu } v_1) \\
&= P^Tv_1\lambda_1 \\
&= \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ & \vdots & \\ — & v_n & — \end{bmatrix} v_1\lambda_1 \\
&= \begin{bmatrix} \langle v_1, v_1\rangle \\ \langle v_2, v_1\rangle \\ \cdots \\ \langle v_n, v_1\rangle \end{bmatrix} \lambda_1 \\
&= \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \lambda_1. && (\|v_1\| = 1 \text{ und bel. } v_i, v_j \in \{v_1, \dots, v_n\} \text{ orthogonal})
\end{aligned}$$

Mit der Darstellung als Blockmatrix und durch Symmetrie von B gilt somit $B = \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & C \end{bmatrix}$ mit $C \in \mathbb{R}^{(n-1)\times(n-1)}$ symmetrisch.

Nach Induktionshypothese gibt es ein orthogonales $Q \in \mathbb{R}^{(n-1)\times(n-1)}$ mit $Q^TCQ = D$ diagonal. Damit gilt

$$\begin{aligned} P^TAP &= B \\ &= \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & C \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & QDQ^T \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q \end{bmatrix} \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q^T \end{bmatrix}. \end{aligned}$$

Also ist

$$\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q^T \end{bmatrix} P^TAP \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q \end{bmatrix} = \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix}.$$

Definiere

$$R = P \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q \end{bmatrix}.$$

Es gilt

$$R^T = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q^T \end{bmatrix} P^T = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q^{-1} \end{bmatrix} P^{-1} = R^{-1}.$$

Dementsprechend ist R orthogonal und $R^{-1}AR = R^TAR = \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix} = \Lambda$ diagonal, da D diagonal ist. □

Korollar 2.6.
Seien die Voraussetzungen aus Satz 2.5 gegeben.
Dann gilt, dass die Diagonalwerte von $\Lambda \in \mathbb{R}^{n\times n}$ die Eigenwerte der Matrix A und die Spalten von R die zugehörigen normierten Eigenvektoren von A sind.

Beweis. Nach dem Spektralsatz gibt es ein orthogonales $R = [r_1 \dots r_n]$ mit $r_1, \dots, r_n \in \mathbb{R}^n$ und $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, sodass

$$R^{-1}AR = \Lambda.$$

Dann ist

$$AR = R\Lambda,$$

oder spaltenweise

$$Ar_i = \lambda_i r_i, \quad \text{für } i = 1, \dots, n.$$

Da R orthogonal ist, sind nach Definition die Spaltenvektoren von R, also $r_1, \dots, r_n$, orthonormal. Für beliebiges r_i gilt somit $\|r_i\| = 1$, wodurch $r_i \neq \mathbf{0}$ sein muss. Mit Definition 2.1 sind also $r_1, \dots, r_n$ die (normierten) Eigenvektoren von A mit zugehörigen Eigenwerten $\lambda_1, \dots, \lambda_n$.

Hinweis. Mithilfe von Zeilen- und Spaltenvertauschungen innerhalb von R und Λ kann Λ so geordnet werden, dass $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Diese Sortierung wird für den Rest der Arbeit angenommen. □

BEMERKUNG 2.7.
Durch Korollar 2.6 lässt sich direkt die nützliche Aussage treffen, dass bei einer symmetrischen Matrix Eigenvektoren zu verschiedenen Eigenwerten orthogonal zueinander sind.

Nun kann mithilfe der vorangegangenen Sätze die Existenz der Singulärwertzerlegung für beliebige reelle Matrizen bewiesen werden. Der Beweis orientiert sich dabei an [Che20].

SATZ 2.8 (Singulärwertzerlegung).
Seien $m, n \in \mathbb{N}$ und $X \in \mathbb{R}^{m \times n}$.
Dann existieren orthogonale Matrizen $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$ und eine Matrix $\Sigma \in \mathbb{R}^{m \times n}$, sodass

$$X = U \Sigma V^T.$$

Beweis. Sei $C = X^T X \in \mathbb{R}^{n \times n}$ und $r = \operatorname{rg}(X) \leq \min(m, n)$. Dann ist C symmetrisch und positiv semidefinit, da zum einen

$$C^T = \left(X^T X\right)^T = X^T X = C$$

und zum anderen für beliebiges $\mathbb{R}^n \ni w \neq 0$ gilt:

$$w^T C w = w^T (X^T X) w = \left(Xw\right)^T \left(Xw\right) = \langle Xw, Xw \rangle = \|Xw\| \geq 0.$$

Damit sind alle Eigenwerte von C positiv oder gleich null. Nach Korollar 2.6 gibt es dann ein orthogonales

$$\mathbf{V} = \begin{bmatrix} v_1 \dots v_n \end{bmatrix} \in \mathbb{R}^{n \times n}$$

und diagonales $\mathbb{R}^{n\times n} \ni \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ mit $\lambda_1 \geq \dots \geq \lambda_r > 0 = \lambda_{r+1} = \dots = \lambda_n$, sodass $C = V\Lambda V^T$. Definiere $\sigma_i := \sqrt{\lambda_i}$ für $i = 1, \dots, r$ und

$$\Sigma_{ij} = \begin{cases} \sigma_i, & i = j. \\ 0, & i \neq j. \end{cases}$$
$$\Leftrightarrow \quad \mathbf{\Sigma} = \begin{bmatrix} \operatorname{diag}(\sigma_i, \dots, \sigma_r) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

als Blockmatrix. Definiere außerdem

$$u_i = \frac{1}{\sigma_i} X v_i \in \mathbb{R}^m, \quad \text{für } 1 \leq i \leq r.$$

Dann sind $u_1, \dots, u_r$ orthornomal:

$$\begin{aligned} u_i^T u_j &= \left(\frac{1}{\sigma_i} X v_i\right)^T \left(\frac{1}{\sigma_i} X v_j\right) \\ &= \frac{1}{\sigma_i \sigma_j} v_i^T \underbrace{X^T X}_{=C} v_j \\ &= \frac{1}{\sigma_i \sigma_j} v_i^T (\lambda_j v_j) && (\lambda_j \text{ ist Eigenwert zu } v_j) \\ &= \frac{\sigma_j}{\sigma_i} v_i^T v_j && (\lambda_j = \sigma_j^2) \\ &= \begin{cases} 1, & i = j. \\ 0, & i \neq j. \end{cases} && (v_i, v_j \text{ orthonormal}) \end{aligned}$$

Wie bereits im Beweis des Spektralsatzes können $u_1, \dots, u_r$ mithilfe des Basisergänzungssatzes (Wiederholung 2.2) und des Gram-Schmidt-Verfahrens (Wiederholung 2.4) durch Vektoren $u_{r+1} \dots, u_m \in \mathbb{R}^m$ zu einer orthonormalen Basis von $\mathbb{R}^m$ ergänzt werden. Damit ist

$$\mathbf{U} = \begin{bmatrix} u_1 \dots u_r u_{r+1} \dots u_m \end{bmatrix} \in \mathbb{R}^{m\times m}$$

orthogonal. Es bleibt zu zeigen, dass $XV = U\Sigma$ ist, also:

$$X\begin{bmatrix} v_1 \dots v_r v_{r+1} \dots v_n \end{bmatrix} = \begin{bmatrix} u_1 \dots u_r u_{r+1} \dots u_m \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \mathbf{0} \\ & & \ddots & & \\ & & & \sigma_r & \\ & \mathbf{0} & & & \mathbf{0} \end{bmatrix}.$$

Für $1 \leq i \leq r$ gilt $Xv_i = u_i\sigma_i$ nach Konstruktion.
Für $i > r$ soll gezeigt werden, dass $Xv_i = 0u_i = \mathbf{0}$ ist. Betrachte dafür

$$X^TXv_i = Cv_i \overset{(*)}{=} 0v_i = \mathbf{0}$$

(∗) *Der zugehörige Eigenwert zum Eigenvektor v_i ist 0 für $i > r$.*

Damit muss wie erwünscht $Xv_i = \mathbf{0}$ gelten, oder $X^T = \mathbf{0}$, wodurch ebenfalls $Xv_i = \mathbf{0}$ folgt. Dementsprechend ist $X = U\Sigma V^T$ und die Aussage ist bewiesen. □

BEMERKUNG 2.9.

- Die Diagonalwerte von Σ heißen **Singulärwerte** von X und werden meist absteigend sortiert.
- Die Spalten von U heißen **linke Singulärvektoren** von X.
- Die Spalten von V heißen **rechte Singulärvektoren** von X.

Nachdem der zentrale Beweis dieses Kapitels geführt und die Existenz der Singulärwertzerlegung für beliebige reelle Matrizen gezeigt wurde, wird nun auf eine wichtige Eigenschaft der SVD eingegangen.

Dafür werden zunächst die vier fundamentalen Unterräume zu einer Matrix nach [Str03, S. 185] definiert:

DEFINITION 2.10.
Sei $A \in \mathbb{R}^{m\times n}$. Definiere folgende Unterräume zu A:

- *Spaltenraum:* $\text{Bild}(A) = \{b \in \mathbb{R}^m \mid \exists x \in \mathbb{R}^n, Ax = b\}$.
- *Zeilenraum:* $\text{Bild}(A^T) = \{z \in \mathbb{R}^n \mid \exists y \in \mathbb{R}^m, A^Ty = z\}$.
- *Kern/Nullraum:* $\text{Kern}(A) = \{x \in \mathbb{R}^n \mid Ax = \mathbf{0}\}$.
- *Linkskern:* $\text{Kern}(A^T) = \{y \in \mathbb{R}^m \mid A^Ty = \mathbf{0}\}$.

Die vier Unterräume geben umfangreichen Aufschluss über die Wirkung einer Matrix auf verschiedene Vektoren und stehen dabei in Verbindung mit zahlreichen Themen der linearen Algebra, wie beispielsweise dem Lösen von Gleichungssystemen.

Die Art der Beziehung zwischen der Singulärwertzerlegung und den vier fundamentalen Unterräumen wird in Korollar 2.12 zusammengefasst. Der

Beweis wird nach [Joh21, S. 214 f.] geführt, vor der Beweisführung wird allerdings Wiederholung 2.11 benötigt.

Wiederholung 2.11 (Dimensionssatz).
Sei $A \in \mathbb{R}^{m\times n}$. Dann gilt:

$$\mathrm{df}(A) + \mathrm{rg}(A) = n.$$

Korollar 2.12.
Seien $m, n \in \mathbb{N}$, $X \in \mathbb{R}^{m\times n}$ und $r = rg(X)$.
Dann existieren orthogonale Matrizen $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ und eine Diagonalmatrix $\Sigma \in \mathbb{R}^{m\times n}$, sodass

$$X = U\Sigma V^T$$

und es gilt:

- Die ersten r Spalten von U sind eine Basis des Spaltenraums von X.
- Die letzten $m - r$ Spalten von U sind eine Basis des Linkskerns von X.
- Die ersten r Spalten von V sind eine Basis des Zeilenraums von X.
- Die letzten $n - r$ Spalten von V sind eine Basis des Kerns von X.

Beweis. Mit der Singulärwertzerlegung (Satz 2.8) erhalten wir orthogonales

$$U = \begin{bmatrix} u_1 \dots u_r u_{r+1} \dots u_m \end{bmatrix} \in \mathbb{R}^{m\times m},$$
$$V = \begin{bmatrix} v_1 \dots v_r v_{r+1} \dots v_n \end{bmatrix} \in \mathbb{R}^{n\times n}$$

und diagonales

$$\Sigma = \begin{bmatrix} \mathrm{diag}(\sigma_i, \dots, \sigma_r) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m\times n},$$

sodass $X = U\Sigma V^T$. Sei nun $i \in \{1, \dots, n\}$ und betrachte

$$Xv_i = U\Sigma V^T v_i \overset{(**)}{=} U\Sigma \mathbf{e}_i = U\sigma_i \mathbf{e}_i = \sigma_i U\mathbf{e}_i = \sigma_i u_i.$$

(**) *Für $i, j \in \{1, \dots, n\}$ sind v_i, v_j orthonormal.*

Fall 1: $1 \leq i \leq r$.
Damit ist $\sigma_i > 0$ und

$$X\frac{v_i}{\sigma_i} = u_i.$$

Nach Definition 2.10 sind dann $u_1, \dots, u_r \in \text{Bild}(X)$. Nun ist $\dim(\text{Bild}(X)) = \text{rg}(X) = r$ und da $\mathcal{B}_S = \{u_1, \dots, u_r\}$ genau r orthonormale Vektoren enthält, bildet $\mathcal{B}_S$ eine Basis vom $\text{Bild}(X)$, also vom Spaltenraum.

Fall 2: $i \geq r + 1$.
Damit ist $\sigma_i = 0$ und

$$Xv_i = \mathbf{0}.$$

Nach Definition 2.10 sind dann $v_{r+1}, \dots, v_n \in \text{Kern}(X)$. Durch Wiederholung 2.11 wissen wir, dass $\dim(\text{Kern}(X)) = \text{df}(X) = n - r$. Mit $\mathcal{B}_K = \{v_{r+1}, \dots, v_n\}$ haben wir $n - r$ orthonormale Vektoren gegeben, also bildet $\mathcal{B}_K$ eine Basis vom $\text{Kern}(X)$.

Die Beweise für die Basen des Linkskerns und des Zeilenraums werden analog gezeigt, indem

$$X^T u_i = V\Sigma U^T u_i$$

betrachtet wird. □

Damit ist die Beweisführung dieser Arbeit abgeschlossen und die Berechnung der Singulärwertzerlegung kann an einem Beispiel veranschaulicht und visualisiert werden.

2.2. BEISPIEL UND VISUALISIERUNG

BEISPIEL 2.13. Sei

$$A = \begin{bmatrix} 1 & -1 & 3 \\ 3 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 3}.$$

Wir wollen nun die SVD dieser Matrix finden. Dafür muss der Beweis der Singulärwertzerlegung (Satz 2.8) mithilfe unserer konkreten Werte schrittweise nachvollzogen werden. Zuerst wird also

$$A^T A = \begin{bmatrix} 10 & 2 & 6 \\ 2 & 2 & -2 \\ 6 & -2 & 10 \end{bmatrix} \in \mathbb{R}^{3\times 3}$$

bestimmt. Davon sollen die Eigenwerte mit zugehörigen normierten Eigenvektoren berechnet werden. Für die Eigenwerte muss zunächst das charakteristi-

sche Polynom

$$\det(A - \lambda \boldsymbol{I}) = 0$$

gesetzt und die Lösungen λ_i für $i = 1, 2, 3$ gefunden werden. Auf die genaue Berechnung wird an dieser Stelle verzichtet, das Ergebnis lautet:

$$\lambda_1 = 16, \quad \lambda_2 = 6, \quad \lambda_3 = 0.$$

Mit $\sigma_j = \sqrt{\lambda_j}$ für $j = 1, 2$ (da $\operatorname{rg}(A) = 2$) erhalten wir

$$\Sigma = \begin{bmatrix} 4 & 0 & 0 \\ 0 & \sqrt{6} & 0 \end{bmatrix} \in \mathbb{R}^{2\times 3}.$$

Durch die Lösungen $\mathbb{R}^3 \ni v_i \neq 0$ von

$$(A - \lambda \boldsymbol{I})v = \boldsymbol{0}$$

ergeben sich die normierten Eigenvektoren

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \quad v_2 = \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}, \quad v_3 = \begin{bmatrix} -\frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}$$

und damit

$$V^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} \in \mathbb{R}^{3\times 3}.$$

Es muss also nur noch $U \in \mathbb{R}^{2\times 2}$ bestimmt werden, welches spaltenweise durch

$$u_j = \frac{1}{\sigma_j} X v_j$$

ausgedrückt wird. Wir erhalten also

$$u_1 = \frac{1}{4} \begin{bmatrix} 1 & -1 & 3 \\ 3 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

und

$$u_2 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & -1 & 3 \\ 3 & 1 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Dadurch ist

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

und die Berechnung ist abgeschlossen mit

$$A = \begin{bmatrix} 1 & -1 & 3 \\ 3 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & \sqrt{6} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} = U\Sigma V^T.$$

Damit genauer verstanden wird, was genau durch die Singulärwertzerlegung geschieht, betrachten wir die Wirkung der Matrix A aus Beispiel 2.13 auf einen Vektor $v \in \mathbb{R}^3$, indem wir v und Av grafisch darstellen. Da dies an einem einzelnen Vektor schwer visualisiert werden kann, multiplizieren wir A mit allen Punkten auf der Einheitssphäre, also allen

$$v \in \left\{ \begin{bmatrix} \cos(x)\sin(y) \\ \sin(x)\sin(y) \\ \cos(y) \end{bmatrix} \in \mathbb{R}^3 \mid x \in [0, 2\pi],\ y \in [0, \pi] \right\}.$$

Das Ergebnis ist in Abbildung 2.1 dargestellt. Um nachzuvollziehen, wie dieses Ergebnis zustande gekommen ist, verwenden wir die Singulärwertzerlegung

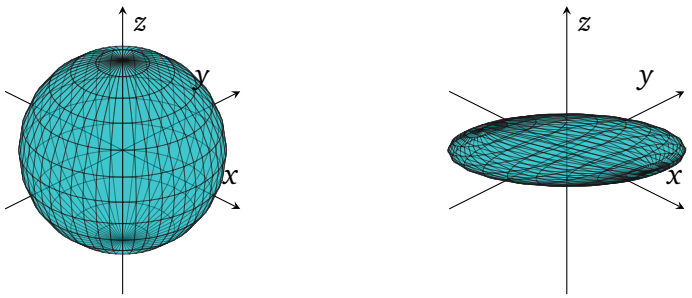


Abb. 2.1. Wirkung von A auf die Einheitssphäre

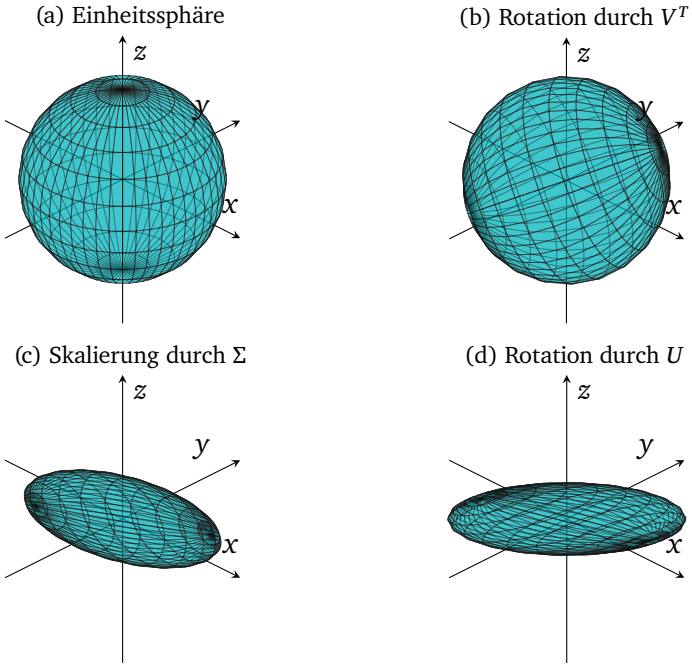


Abb. 2.2. Visualisierung der Singulärwertzerlegung

und veranschaulichen die Zwischenschritte von $Av = U\Sigma V^T v$ anhand von einzelner Plots (siehe Abbildung 2.2).

Grundlage für diese Visualisierung ist das Wissen, dass im euklidischen Raum orthogonale Matrizen Drehungen und Diagonalmatrizen Skalierungen (entlang der Hauptachsen) darstellen. An dieser Stelle ist auch wichtig zu erwähnen, dass in Abbildung 2.2c und Abbildung 2.2d der zu beobachtende Wertebereich vergrößert wurde, also eine größere Streckung durch Σ erfolgt ist, als auf dem Plot zu sehen ist. Außerdem sei angemerkt, dass ab Abbildung 2.2c die Darstellung der z-Achse überflüssig und eigentlich nicht zu empfehlen ist, da sich dort durch die Dimensionsreduktion mittels Σ im zweidimensionalen Raum bewegt wird. Um eine Vergleichbarkeit der Plots zu verbessern, wird die Darstellung dennoch beibehalten.

Zusammenfassend zerlegt also die Singulärwertzerlegung eine Matrix in grundlegende geometrische Transformationen: Drehung, Skalierung und gegebenenfalls Dimensionsreduktion oder -erhöhung.

Mit dieser Erkenntnis wird sich dem letzten Abschnitt des theoretischen Teils zugewandt, in dem die wichtigsten beiden Arten der SVD definiert werden.

2.3. ARTEN DER SINGULÄRWERTZERLEGUNG

Die Singulärwertzerlegung, die im vorherigen Teil der Arbeit beschrieben wurde, ist die klassische und vollständige Zerlegung. In den tatsächlichen Anwendungsgebieten, welche im nächsten Kapitel ausgeführt werden, finden häufig Variationen Verwendung.

DEFINITION 2.14.
Seien $m, n \in \mathbb{N}$, $A \in \mathbb{R}^{m\times n}$, $\operatorname{rg}(A) = r$ und $A = U\Sigma V^T$ die vollständige SVD von A mit $U \in R^{m\times m}$, $\Sigma \in \mathbb{R}^{m\times n}$ und $V \in \mathbb{R}^{n\times n}$.
Definiere die *reduzierte SVD* von A:

$$A = U_r \Sigma_r V_r^T$$

mit

$$\begin{aligned} U_r &= [u_1 \dots u_r] \in \mathbb{R}^{m\times r}, \\ \Sigma_r &= \operatorname{diag}(\sigma_i, \dots, \sigma_r) \in \mathbb{R}^{r\times r}, \\ V_r &= [v_1 \dots v_r] \in \mathbb{R}^{n\times r}. \end{aligned}$$

Die in Definition 2.14 beschriebene Art der Singulärwertzerlegung besitzt den Vorteil, dass eine exakte Berechnung mit deutlich weniger Speicherbedarf möglich ist, insbesondere bei großen Matrizen $X \in \mathbb{R}^{m\times n}$ mit $\operatorname{rg}(X) \ll \min(m, n)$.

Nachteilig ist, dass die Basen für den Kern und Linkskern „verloren gehen“ (siehe Definition 2.10), dies spielt für die meisten Anwendungen allerdings eine untergeordnete Rolle.

Bevor zur nächsten Variation übergegangen werden kann, betrachten wir eine weitere Darstellungsform von Definition 2.14.

BEMERKUNG 2.15.
Sei

$$A = [u_1 \dots u_r] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_r^T \end{bmatrix}$$

wie in Definition 2.14.
Dann ist $A_{11} = u_{1,1}\sigma_1 v_{1,1}^T + u_{2,1}\sigma_2 v_{2,1}^T + \cdots + u_{r,1}\sigma_r v_{r,1}^T$.
Für die andere Komponenten von A kann die Summe analog gebildet

werden. Wir erhalten also

$$\begin{aligned} A_{ij} &= \sum_{k=1}^{r} \sigma_k u_{k,i} v_{k,j}^T \\ \Leftrightarrow \quad A &= \sum_{i=1}^{r} \sigma_i u_i v_i^T. \end{aligned}$$

Damit kann A als *Summe von Rang-1-Matrizen* dargestellt werden, da für $i \in \{1, \dots, r\}$ die Zeilen von $(\sigma_i u_i v_i^T) \in \mathbb{R}^{m \times n}$ ein Vielfaches von v_i^T und die Spalten ein Vielfaches von u_i sind.
Beachte, dass $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ und $\|u_i\| = 1 = \|v_i\|$, also gilt komponentenweise: $\sigma_1 u_1 v_1^T \geq \sigma_2 u_2 v_2^T \geq \cdots \geq \sigma_r u_r v_r^T$.

Nun wird die womöglich interessanteste Art der Singulärwertzerlegung für diverse Anwendungsfälle definiert.

Definition 2.16.
Seien $m, n \in \mathbb{N}$, $A \in \mathbb{R}^{m \times n}$, $\operatorname{rg}(A) = r$ und $A = U_r \Sigma_r V_r^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ die reduzierte SVD von A mit $U_r \in R^{m \times r}$, $\Sigma_r \in \mathbb{R}^{r \times r}$ und $V_r \in \mathbb{R}^{n \times r}$.
Für $k < r$ sei die *trunkierte SVD* definiert mit:

$$A \approx \sum_{i=1}^{k} \sigma_i u_i v_i^T = A_k.$$

Zur Definition 2.16 sei angemerkt, dass sich A_k für größere Werte von k zunehmend A annähert, aber da komponentenweise $\sigma_1 u_1 v_1^T \geq \cdots \geq \sigma_r u_r v_r^T$ gilt (siehe Bemerkung 2.15), kann bereits mit $k \ll r$ eine gute Approximation erzielt werden. Außerdem sollte betont werden, dass $\operatorname{rg}(A_k) = k < r$ ist, aufgrund dessen spricht man auch von einer *Niedrigrang-Approximation*.

Mit den beiden, in diesem Abschnitt eingeführten, Variationen der Singulärwertzerlegung sind alle wesentlichen Grundlagen für die Vertiefung verschiedener Anwendungsmöglichkeiten der SVD gelegt. Damit wird der theoretische Teil dieser Arbeit beendet und zur ersten praktischen Anwendung übergegangen.

3 Hauptkomponentenanalyse

Die Hauptkomponentenanalyse (engl. *Principal Component Analysis*, PCA) ist ein Verfahren zur Dimensionsreduktion von Daten. Genauer: Es handelt sich um eine Methode, um komplexe Daten auf ihr Wesentliches zu reduzieren, was eine Weiterverarbeitung und Visualisierung erleichtert.

In diesem Kapitel wird zunächst die Intuition hinter der PCA erläutert, bevor der mathematische Hintergrund und insbesondere die Verbindung zur Singulärwertzerlegung beschrieben wird. Abschließend betrachten wir ein konkretes Anwendungsbeispiel und veranschaulichen dieses mithilfe der Programmiersprache `Python`.

3.1. Intuition der PCA

Angenommen, beim Familienessen käme die Frage auf, welche der mitgebrachten Weine sich am ähnlichsten sind. Um diese Frage zu beantworten, überlegt sich die Familie verschiedene Merkmale und ordnet jedem Wein für jedes Merkmal Zahlen zwischen -3 und 3 zu. Dadurch können die Weine als Punkte im Raum bezüglich der verschiedenen Werte dargestellt und anschließend analysiert werden, welche Weine sich gruppieren, also sich ähneln.

In Abbildung A.1, zu finden im Anhang A, wird dies für verschiedene $n :=$ *Anzahl der Merkmale* verdeutlicht.

Das Problem wird schnell ersichtlich: Eine visuelle Interpretation ist zwar möglich, allerdings nur im niedrig-dimensionalen Raum, für eine größere Anzahl an Merkmalen (Dimensionen) besteht die Notwendigkeit, die Anzahl zu reduzieren. Diese Dimensionsreduktion stellt in vielen Fällen auch unabhängig von der visuellen Interpretation eine sinnvolle Maßnahme dar. In Bezug auf unser Beispiel bestehe die Möglichkeit, dass „Alkoholgehalt“ und „Schwere“ stark korrelieren und somit redundant für eine Rekonstruktion der Weine sind. Eine andere Möglichkeit zur Reduzierung der Merkmale ist dadurch gegeben, dass gewisse Merkmale wenig Informationen über die Charakteristiken der Weine

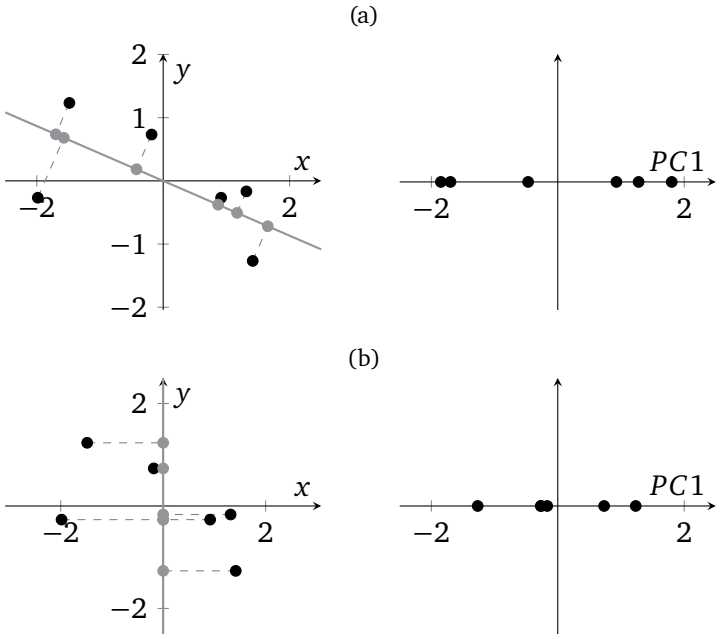


Abb. 3.1. Projektionen im zweidimensionalen Raum

enthalten, wie beispielsweise die Flaschenform oder die Fließgeschwindigkeit.

Die Hauptkomponentenanalyse konstruiert neue, unkorrelierte Richtungen, die sich als Linearkombinationen aus den bestehenden Merkmalen zusammensetzen. Dadurch könnten beispielsweise „Alkoholgehalt“ und „Schwere“ zu einer neuen *Hauptrichtung* zusammengefasst werden. Anschließend werden die Weine auf den durch diese Richtung definierten Unterraum projiziert. Die projizierten Koordinaten entlang dieser neuen Achse ergeben die erste *Hauptkomponente* (PC1). Der Unterschied zwischen Hauptrichtung und Hauptkomponente wird am Ende der Intuition präzisiert.

Die Projektion wird in Abbildung 3.1 veranschaulicht. Der Unterschied zwischen Abbildung 3.1a und Abbildung 3.1b zeigt folgendes Problem auf: Wie kann die neue Richtung optimal gewählt werden, sodass unsere Daten trotz der Dimensionsreduktion so originalgetreu wie möglich rekonstruiert werden können? Dafür gibt es zwei alternative, aber äquivalente, Formulierungen:

1. Die Richtung wird so konstruiert, dass ein Minimum an Informationen verloren wird.

2. Die Richtung wird so konstruiert, dass eine maximale Varianz erhalten bleibt.

Eine Äquivalenz dieser beiden Aussagen kann durch den Satz des Pythagoras hergeleitet werden, indem das rechtwinklige Dreieck zwischen einem Punkt, dessen Projektion und dem Ursprung betrachtet wird. Wird der Informationsverlust (Distanz zwischen der Projektion und dem Original) minimiert, maximiert sich die Varianz (Abstand zum Ursprung). Es sei angemerkt, dass dafür von einer Zentrierung der Daten ausgegangen wird, also von einem Mittelwert gleich null.

Mit diesem Hintergrund wird intuitiv ersichtlich, dass Abbildung 3.1a im Vergleich zu Abbildung 3.1b vorzuziehen ist, was sich im Ergebnis widerspiegelt, in dem die räumliche Verteilung im Wesentlichen erhalten bleibt.

In vielen mathematischen Texten erfolgt keine klare Differenzierung zwischen den Begriffen Hauptrichtung und Hauptkomponente. Stattdessen wird der Begriff Hauptkomponente häufig synonym für beide verwendet. In dieser Arbeit definieren wir die erste Hauptrichtung als den Einheitsvektor, der den eindimensionalen Unterraum aufspannt, auf den die Daten mit maximaler Varianz projiziert werden. Die erste Hauptkomponente bezeichnet die projizierten Koordinaten der Datenpunkte entlang dieser Hauptrichtung. In höheren Dimensionen wird die zweite Hauptrichtung so gewählt, dass sie orthogonal zur ersten Hauptrichtung liegt und die verbleibende Varianz maximiert. Dies kann beliebig fortgesetzt werden, die PCA besitzt allerdings die nützliche Eigenschaft, dass die Hauptrichtungen nach „Wichtigkeit“ sortiert sind, also die ersten Richtungen bereits den Großteil der Varianz erklären. Eine genauere Erläuterung dieser Aussage wird dabei in der folgenden mathematischen Herleitung gegeben.

3.2. Mathematische Herleitung

Bevor die vorangegangenen Überlegungen formalisiert werden können, rekapitulieren wir durch Wiederholung 3.1 die Formel der orthogonalen Projektion.

Wiederholung 3.1.
Sei $n \in \mathbb{N}$ und $u, x \in \mathbb{R}^n$ mit $\|u\| = 1$.
Dann ist der orthogonal projizierte Vektor $\text{proj}_u(x)$ von x auf u gegeben durch

$$\text{proj}_u(x) = \langle x, u \rangle u = (x^T u)u.$$

Das Ziel der Herleitung in diesem Abschnitt wird durch Anwendung 3.2 zusammengefasst.

ANWENDUNG 3.2 (PCA).
Seien $n, d \in \mathbb{N}$ und

$$X = [x_1 \dots x_d] \in \mathbb{R}^{n \times d}$$

eine standardisierte[a] Datenmatrix, die d Merkmale über n Objekte hinweg speichert und

$$\mathbb{R}^d \ni x^{(i)} := X_{i,:} \quad \text{für } i \in \{1, \dots, n\},$$

die i-te (transponierte) Zeile von X, also ein Datenpunkt.
Bei der Projektion der Daten auf $\mathbb{R}^k$ für $\mathbb{N} \ni k < d$ mit dem Ziel, dass die maximale Varianz der $x^{(i)}$ erhalten bleiben soll, wird die Basis von $\mathbb{R}^k$ durch die ersten k Hauptrichtungen $u_j \in \mathbb{R}^d$ für $j \in \{1, \dots, k\}$ gegeben. Diese entsprechen den Eigenvektoren der Matrix $\frac{1}{n} X^T X \in \mathbb{R}^{d \times d}$, sortiert in absteigender Reihenfolge der zugehörigen Eigenwerte.

[a] Siehe Schritt 1 im Beweis

Beweis. Der Beweis orientiert sich zum Großteil an [NM23, S. 166-169] und [Hsu16, S. 32 f.].

Schritt 1. Vorbereitung der Daten:
Wir standardisieren zunächst X, indem

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j} \quad \text{für alle } j \in \{1, \dots, d\},$$

wobei

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_j^{(i)}, \quad \sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_j^{(i)} - \mu_j)^2$$

jeweils die Mittelwerte, bzw. die Varianzen der einzelnen Merkmale, also der Spalten sind. Durch die Subtraktion des Mittelwerts werden die Daten um den Ursprung zentriert. Die Division durch die Standardabweichung verhindert Ungenauigkeiten, die aufgrund unterschiedlicher Skalen der Merkmale entstehen könnten. Falls Merkmal A beispielsweise das Bruttoinlandsprodukt und Merkmal B die Geburtenrate verschiedener Länder darstellt, wird dadurch eine Vergleichbarkeit gewährleistet.

Schritt 2. Herleitung für $k = 1$:
Es sei daran erinnert, dass die erste Hauptrichtung der Richtung entspricht,

die die Varianz der Projektion maximiert. Dafür wird zunächst der Mittelwert μ_{proj} der projizierten Vektoren berechnet:

$$\mu_{\text{proj}} = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)^T}u\right)u = \left(\left(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}\right)^T u\right)u = \mathbf{0},$$

da durch die Standardisierung der Spaltenmittelwert für jede Spalte von X null beträgt, wodurch

$$\sum_{i=1}^{n} x^{(i)} = \mathbf{0}.$$

Die Entfernung (Abweichung) vom Mittelwert, also hier vom Ursprung, für einen beliebigen Vektor $x^{(i)}$ beträgt

$$\|\text{proj}_u\left(x^{(i)}\right)\| = \|\left(x^{(i)^T}u\right)u\| = |\left(x^{(i)^T}u\right)|\,\|u\| = |x^{(i)^T}u|.$$

Damit ist die Varianz der projizierten Punkte gegeben durch

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)^T}u\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}x^{(i)^T}ux^{(i)^T}u\\ &= \frac{1}{n}\sum_{i=1}^{n}u^T x^{(i)}x^{(i)^T}u \qquad \text{(Skalarprodukt kommutativ)}\\ &= u^T\left(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}x^{(i)^T}\right)u\end{aligned}$$

Definiere

$$\Sigma := \frac{1}{n}\sum_{i=1}^{n}x^{(i)}x^{(i)^T} = \frac{1}{n}X^T X \in \mathbb{R}^{d\times d}.$$

Bei standardisierten Daten ist diese Matrix als *Kovarianzmatrix* bekannt, in unserem Fall ist sie die Kovarianzmatrix der verschiedenen Merkmale. Damit haben wir das Ziel der Herleitung auf folgendes Optimierungsproblem reduziert:

$$\begin{aligned}\max \quad & u^T\Sigma u,\\ \text{u.d.B.} \quad & \|u\| = 1.\end{aligned}$$

Dieses Optimierungsproblem wird in der Literatur meist mithilfe von Lagrange-Multiplikatoren gelöst. In dieser Arbeit werden wir einen anderen Ansatz verfolgen und mit dem, im vorherigen Kapitel bewiesenen, Spektralsatz (Satz 2.5)

vorgehen. Beachte dafür zunächst, dass Σ symmetrisch ist:

$$\Sigma^T = \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)}x^{(i)^T}\right)^T = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)}x^{(i)^T}\right)^T = \Sigma.$$

Damit sind die Voraussetzungen für den Spektralsatz erfüllt, womit

$$\Sigma = R\Lambda R^T$$

für orthogonales $R = [r_1 \dots r_d] \in \mathbb{R}^{d\times d}$ und diagonales $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ mit $\lambda_1 \geq \cdots \geq \lambda_d$. Für $w := R^T u$ gilt dann

$$u^T\Sigma u = u^T R\Lambda R^T u = w^T\Lambda w = w^T \begin{bmatrix} \lambda_1 w_1 \\ \vdots \\ \lambda_d w_d \end{bmatrix} = \sum_{i=1}^{d} \lambda_i w_i^2.$$

Nach Bedingung ist $\|u\| = 1$, wodurch:

$$\|w\| = \|R^T u\| = \sqrt{\langle R^T u, R^T u\rangle} = \sqrt{u^T R R^T u} = \|u\| = 1.$$

Da

$$\sum_{i=1}^{d} \lambda_i w_i^2 = \lambda_1 w_1 + \lambda_2 w_2 + \cdots + \lambda_d w_d$$

und $\lambda_1 \geq \cdots \geq \lambda_d$ wird der Ausdruck maximiert für $w = \mathbf{e}_1$. Es folgt

$$u = Rw = R\mathbf{e}_1 = r_1,$$

womit u nach Korollar 2.6 gleich dem zugehörigen Eigenvektor zum größten Eigenwert von Σ ist.

Schritt 3. Herleitung für beliebiges k mit vollständiger Induktion über k:
Induktionsanfang. Gegeben durch Schritt 2.
Induktionshypothese. Angenommen die ersten $k-1$ Eigenvektoren $r_1, \dots, r_{k-1}$ von Σ maximieren die Varianz der Projektion.
Induktionsschritt. Wir wollen zeigen, dass

$$\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\left(x^{(i)^T}u_j\right)^2$$

maximiert wird für $u_k = r_k$. Betrachte dafür

$$\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\left(x^{(i)^T}u_j\right)^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{k-1}\sum_{i=1}^{n}\left(x^{(i)^T}u_j\right)^2}_{A} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)^T}u_k\right)^2}_{B}.$$

Nach Induktionshypothese ist A maximal für $u_1 = r_1, u_2 = r_2, \dots, u_{k-1} = r_{k-1}$. Um B zu maximieren, wird analog zu Schritt 2 vorgegangen. Es wird zunächst das Optimierungsproblem

$$\begin{aligned} \max \quad & u_k^T \Sigma u_k, \\ \text{u.d.B.} \quad & \|u_k\| = 1, \\ & \langle u_k, u_i \rangle = 0 \quad \text{für alle } i \in \{1, \dots, k-1\}, \end{aligned}$$

aufgestellt mit zusätzlicher Orthogonalitätsbedingung. Anschließend kann, mithilfe der Eigenschaft, dass bei symmetrischen Matrizen Eigenvektoren zu verschiedenen Eigenwerten orthogonal sind (Bemerkung 2.7), das Optimierungsproblem durch $u_k = r_k$ gelöst werden, womit der Beweis abgeschlossen ist.

Hinweis. Die i-te Hauptkomponente ist dann durch Xu_i gegeben, wodurch die Projektionen der Datenpunkte auf u_i im durch u_i aufgespannten Unterraum zusammengefasst werden. □

Korollar 3.3.
Seien die Voraussetzungen aus Anwendung 3.2 gegeben.
Die Varianz der Datenpunkte entlang einer beliebigen Hauptrichtung $u_j \in \mathbb{R}^d$ für $j \in \{1, \dots, k\}$ entspricht genau dem zugehörigen Eigenwert λ_j zu u_j.

Beweis. Wie bereits im vorherigen Beweis gezeigt, ist die Varianz der Datenpunkte entlang von u_j gegeben durch

$$u_j^T \Sigma u_j.$$

Mit dem Wissen, dass u_j ein Eigenvektor von Σ mit zugehörigem Eigenwert λ_j ist, kann dies umgeformt werden zu

$$u_j^T(\lambda_j u_j) = \lambda_j u_j^T u_j = \lambda_j,$$

da $\|u_j\| = 1$. □

BEMERKUNG 3.4.
Daraus ergibt sich die bereits in der Intuition erwähnte Eigenschaft der Hauptkomponentenanalyse: Ein Großteil der Varianz kann bereits durch die ersten Hauptkomponenten erklärt werden, da die Eigenwerte der Kovarianzmatrix Σ in absteigender Reihenfolge die jeweilige Varianz angeben.

3.3. VERBINDUNG ZUR SVD

Das Ziel dieses Abschnitts ist, die Hauptrichtungen, -komponenten und die Varianzen mithilfe der Singulärwertzerlegung der Datenmatrix auszudrücken, statt mit der Spektralzerlegung der Kovarianzmatrix. Diese Art der Berechnung wird in der Praxis häufig verwendet, da sie einen entscheidenden Vorteil besitzt, auf den am Ende des Abschnitts genauer eingegangen wird.

Sei also erneut $X \in \mathbb{R}^{n\times d}$ eine standardisierte Datenmatrix mit zugehöriger Kovarianzmatrix

$$\Sigma = \frac{1}{n} X^T X = V \Lambda V^T \in \mathbb{R}^{d\times d} \tag{3.1}$$

nach dem Spektralsatz für orthogonales $V \in \mathbb{R}^{d\times d}$, welches die Eigenvektoren von Σ als Spaltenvektoren enthält und diagonales $\Lambda \in \mathbb{R}^{d\times d}$ mit den zugehörigen Eigenwerten auf der Diagonalen.

Sei außerdem

$$X = USV^T$$

die Singulärwertzerlegung von X mit „diagonalem“ $S \in \mathbb{R}^{n\times d}$ und orthogonalem $U \in \mathbb{R}^{n\times n}$. Da die Eigenvektoren von $\Sigma = \frac{1}{n} X^T X$ und $X^T X$ identisch sind (folgt direkt aus Definition 2.1; nur die Eigenwerte unterscheiden sich um einen Faktor n), ist V die gleiche Matrix wie in (3.1). Folglich sind die Hauptrichtungen durch die Spalten von V gegeben.

Um die Varianzen zu ermitteln, betrachten wir

$$\frac{1}{n} X^T X = \frac{1}{n} V S^T U^T U S V^T = V \frac{S^T S}{n} V^T = V \Lambda V^T.$$

Für $\operatorname{rg}(X) = r$ und $i \in \{1, \dots, r\}$ kann damit ein beliebiger Eigenwert λ_i von Σ, also die Varianz zur i-ten Hauptrichtung, durch den Singulärwert σ_i von X ausgedrückt werden mit

$$\lambda_i = \frac{\sigma_i^2}{n},$$

wobei $\lambda_i = 0$ für $i > r$.

Die Hauptkomponente zu einer beliebigen Hauptrichtung v wird, wie bereits in der Herleitung angemerkt, durch Xv beschrieben. Es lassen sich also alle Hauptkomponenten zusammenfassen durch

$$XV = USV^TV = US.$$

Damit lässt dich die Hauptkomponentenanalyse vollständig durch die Singulärwertzerlegung ausdrücken:

- Die Hauptrichtungen durch die Spaltenvektoren von V.
- Die Hauptkomponenten durch US.
- Die Varianzen durch $\frac{\sigma_i^2}{n}$ mit $n :=$ Anzahl der Objekte.

Der Vorteil, die PCA mit der SVD von X zu berechnen, statt mit der Spektralzerlegung von Σ, zeichnet sich durch die numerische Stabilität der Berechnung aus. Um dies zu zeigen, wird ein Exkurs in die algorithmische Mathematik vorgenommen, wobei die für uns wichtigen Konzepte in Wiederholung 3.5 rekapituliert werden.

Wiederholung 3.5.

- Die *Kondition* eines Problems gibt an, wie sich kleine Eingabefehler auf den Ausgabefehler auswirken. Bei einem gut konditionierten Problem führen kleine Eingabefehler zu kleinen Ausgabefehlern.
- Ein Algorithmus ist dann *numerisch stabil*, falls der nicht vermeidbare Fehler, bedingt durch die Kondition, nicht weiter verstärkt wird.
- Für eine Matrix $A \in \mathbb{R}^{m\times n}$ mit $\operatorname{rg}(A) = r$ ist die *Konditionszahl* κ gegeben durch
$$\kappa(A) = \frac{\sigma_1}{\sigma_r},$$
wobei σ_1 den größten und σ_r den kleinsten Singulärwert von A darstellt.
- Desto kleiner die Konditionszahl ist, desto besser konditioniert ist das Problem.

Wir zeigen nun, dass Σ schlechter konditioniert ist als X und somit anfälliger für Ungenauigkeiten bei numerischen Berechnungen, wie dem Bestimmen

von Eigenwerten. Nach Wiederholung 3.5 kann die Konditionszahl mithilfe der Singulärwerte von Σ bestimmt werden, die als die Wurzeln der positiven Eigenwerte von $\Sigma^T\Sigma = \Sigma^2$ gegeben sind.

Sei v ein beliebiger Eigenvektor von Σ zum Eigenwert λ, dann gilt

$$\Sigma^2 v = \Sigma\Sigma v = \Sigma\lambda v = \Sigma v\lambda = \lambda^2 v.$$

Daraus folgt, dass die Eigenvektoren von Σ und Σ^2 identisch sind, während die Eigenwerte quadriert werden. Die Singulärwerte von Σ entsprechen damit den Eigenwerten von Σ und

$$\kappa(\Sigma) = \frac{\sigma_1^2/n}{\sigma_r^2/n} = \left(\frac{\sigma_1}{\sigma_r}\right)^2 = \kappa(X)^2.$$

Die Konditionszahl von Σ ist somit das Quadrat der Konditionszahl von X, was insbesondere bei schlecht konditionierten Matrizen zu erheblichen Ungenauigkeiten führen kann.

Diese Folgerung veranschaulichen wir an einem Beispiel. Sei

$$X = \begin{bmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{bmatrix},$$

die sogenannte *Läuchli-Matrix*. Die quadrierten Singulärwerte von X sind bekannt und gegeben durch

$$\sigma_1^2 = 3 + \varepsilon^2, \quad \sigma_2^2 = \varepsilon^2, \quad \sigma_3^2 = \varepsilon^2.$$

Um diese Werte zu berechnen, verwenden wir zum einen die Singulärwertzerlegung von X und zum anderen die direkte Berechnung der Eigenwerte von X^TX. Mithilfe von `Python` erreichen wir dies durch:

```
# Quadrate der Singulärwerte von X
singular_values = np.linalg.svd(X, compute_uv=False)
singular_values_squared = singular_values**2
# Eigenwerte von X^T X
eigvals = np.linalg.eigvalsh(X.T @ X)[::-1]
```

Die Funktionen `np.linalg.eigvalsh` und `np.linalg.svd` berechnen dabei jeweils die Eigenwerte einer reell symmetrischen Matrix, bzw. die Singulärwertzerlegung einer Matrix, wobei in diesem Beispiel durch `compute_uv=False`

ausschließlich die Singulärwerte berechnet werden. Der vollständige Code ist zu finden unter Code B.1 in Anhang B.

Für $\varepsilon = 10^{-3},\ 10^{10}$ und 10^{-20} erhalten wir damit folgenden Output:

```
Epsilon: 0.001
Eigenwerte vs. Quadrate der Singulärwerte:
3.000e+00        3.000e+00
1.000e-06        1.000e-06
1.000e-06        1.000e-06
-
Epsilon: 1e-10
Eigenwerte vs. Quadrate der Singulärwerte:
3.000e+00        3.000e+00
-1.770e-17       1.000e-20
-5.849e-16       1.000e-20
-
Epsilon: 1e-20
Eigenwerte vs. Quadrate der Singulärwerte:
3.000e+00        3.000e+00
-1.770e-17       1.000e-40
-5.849e-16       1.000e-40
-
```

Es lässt sich beobachten, dass für $\varepsilon = 10^{-3}$ beide Berechnungsmethoden übereinstimmen und exakt arbeiten. Für $\varepsilon = 10^{-10}$ und $\varepsilon = 10^{-20}$ treten jedoch Ungenauigkeiten in der direkten Eigenwertberechnung von $X^T X$ auf, während die SVD weiterhin zu stabilen Ergebnissen führt.

Damit sind die Vorteile der Singulärwertzerlegung bei der Berechnung der Hauptkomponentenanalyse gezeigt und es wird zum nächsten Anwendungsbeispiel übergegangen.

4 Empfehlungssysteme

Empfehlungssysteme sind eine Anwendungsart der Singulärwertzerlegung, mit der bereits die meisten Menschen in Kontakt gekommen sind. Seien es Filmempfehlungen bei Netflix oder Produktempfehlungen bei Amazon, die Wahrscheinlichkeit ist groß, dass diese mithilfe einer Abwandlung der SVD generiert werden. In diesem Kapitel wird zunächst, ähnlich wie im vorherigen Kapitel, die Grundidee von Empfehlungssystem mithilfe eines intuitiven Ansatzes veranschaulicht. Anschließend wird diese Idee mathematisch formalisiert und vertieft. Um das Kapitel abzuschließen, wird mithilfe von `Python` ein eigenes (simples) Empfehlungssystem für Filme programmiert, basierend auf tatsächlichen Bewertungen aus einer Datenbank.

4.1. Intuition

Für die Intuition verweilen wir beim Beispiel der Filmempfehlungen. Die Ausgangslage für so ein Empfehlungssystem ist in Tabelle 4.1 veranschaulicht.

Gegeben sei eine Nutzer-Item-Matrix, in der jede Zeile einen Nutzer und jede Spalte einen Film repräsentiert, wobei die einzelnen Einträge die abgegebenen

Tab. 4.1. Nutzer-Item-Matrix

	Items			
Nutzer	Film A	Film B	Film C	Film D
Nutzer 1		4	2	0
Nutzer 2	0	2	3	5
Nutzer 3	1	2		
Nutzer 4		4	3	3
Nutzer 5	4	2	1	1
Nutzer 6	5			2

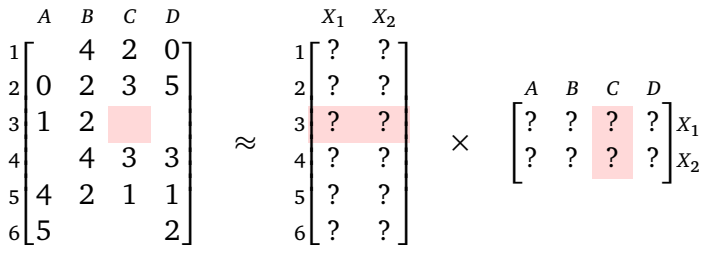


Abb. 4.1. Nutzer-Matrix und Item-Matrix

Bewertungen der Nutzer für den jeweiligen Film darstellen. Das Ziel des Empfehlungssystems besteht darin, die fehlenden Bewertungen so präzise wie möglich zu approximieren, um darauf basierend Empfehlungen generieren zu können. Hierfür wird die Annahme getroffen, dass die Bewertungen nicht unabhängig erfolgen, sondern einer bestimmten Struktur folgen. Es wird also angenommen, dass es zugrunde liegende Muster gibt, nach denen Nutzer mit ähnlichen Präferenzen auch tendenziell ähnliche Bewertungen vergeben. Ein Beispiel dafür wäre, dass Nutzer mit einer Vorliebe für Horrorfilme diese potenziell höher bewerten als andere Nutzer.

Diese Muster werden als *latente Merkmale* bezeichnet. Zur Approximation der fehlenden Bewertungen wird die Nutzer-Item-Matrix als Produkt zweier Matrizen dargestellt: einer Nutzer-Matrix, in der die Nutzer durch die latenten Merkmale beschrieben werden, und einer Item-Matrix mit der Beschreibung der Filme durch die Merkmale. Dieses Konzept wird in Abbildung 4.1 verdeutlicht mit den latenten Merkmalen X_1 und X_2.

Für eine Approximation des hervorgehobenen fehlenden Wertes muss dann nur das Skalarprodukt aus den markierten Vektoren gebildet werden.

Literaturverzeichnis

[Che20] G. Chen. *Lecture 5: Singular Value Decomposition (SVD)*. San José State University, 2020. URL: https://www.sjsu.edu/faculty/guangliang.chen/Math253S20/lec5svd.pdf (Stand: 20. 01. 2025).

[Cra22] T. Crawford. *Oxford Linear Algebra: Spectral Theorem Proof*. 2022. URL: https://tomrocksmaths.com/2022/11/18/oxford-linear-algebra-spectral-theorem-proof/ (Stand: 02. 01. 2025).

[Hsu16] D. Hsu. *Machine Learning Theory. COMS 4772. Topic 5: Principal Component Analysis.* Columbia University, 2016. URL: https://www.cs.columbia.edu/~djhsu/AML/lectures/notes-pca.pdf (Stand: 25. 02. 2025).

[Joh21] N. Johnston. *Advanced Linear and Matrix Algebra*. Cham: Springer, 2021.

[NM23] A. Ng und T. Ma. *Machine Learning. CS229 Lecture Notes.* Stanford University, 2023. URL: https://cs229.stanford.edu/main_notes.pdf (Stand: 23. 02. 2025).

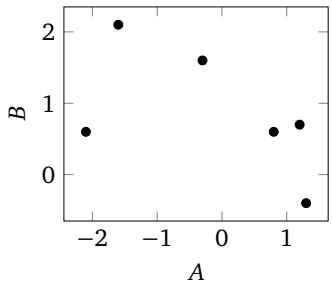
[Str03] G. Strang. *Lineare Algebra*. Berlin, Heidelberg: Springer, 2003.

Abbildungen

A

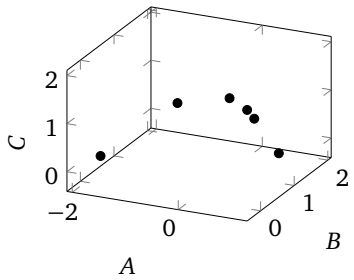
(a) $n = 2$

	Merkmale	
Weine	A	B
Wein 1	1,3	-0,4
Wein 2	-1,6	2,1
Wein 3	1,2	0,7
Wein 4	-2,1	0,6
Wein 5	0,8	0,6
Wein 6	-0,3	1,6



(b) $n = 3$

	Merkmale		
Weine	A	B	C
Wein 1	1,3	-0,4	1,9
Wein 2	-1,6	2,1	0,2
Wein 3	1,2	0,7	0,4
Wein 4	-2,1	0,6	-0,2
Wein 5	0,8	0,6	1,1
Wein 6	-0,3	1,6	0.8



(c) $n = 4$

	Merkmale			
Wein	A	B	C	D
Wein 1	1,3	-0,4	1,9	0.7
Wein 2	-1,6	2,1	0,2	0.9
Wein 3	1,2	0,7	0,4	1.3
Wein 4	-2,1	0,6	-0,2	0.5
Wein 5	0,8	0,6	1,1	-0.8
Wein 6	-0,3	1,6	0.8	-1.3

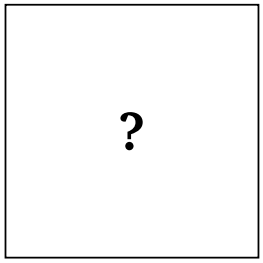


Abb. A.1. Darstellung von Daten in verschiedenen Dimensionen

Programmcode

B

```python
# lauchli.py

import numpy as np


# Funktion zur Erstellung der Läuchli-Matrix für gegebenes epsilon
def laeuchli_matrix(epsilon):
    L = np.zeros((4, 3))
    L[0, :] = 1
    np.fill_diagonal(L[1:, :], epsilon)
    return L


# Verschiedene Werte für epsilon
epsilons = [1e-3, 1e-10, 1e-20]

for epsilon in epsilons:
    L = laeuchli_matrix(epsilon)

    # Eigenwerte von L^T L
    eigvals = np.linalg.eigvalsh(L.T @ L)[::-1]

    # Quadrate der Singulärwerte von L
    singular_values = np.linalg.svd(L, compute_uv=False)
    singular_values_squared = singular_values**2

    print(f"Epsilon: {epsilon}")
    print("Eigenwerte vs. Quadrate der Singulärwerte:")
    for ev, sv in zip(eigvals, singular_values_squared):
        print(f"{ev:.3e}          {sv:.3e}")
    print("-")
```

Code B.1. Berechnungsunterschiede Läuchli-Matrix