

Singulärwertzerlegung

wr@isg.cs.uni-magdeburg.de

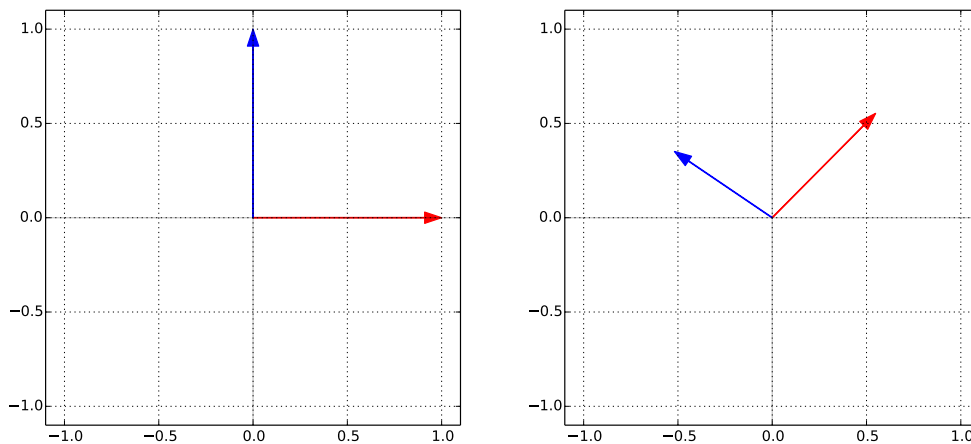
SoSe 2018

1 Singulärwertzerlegung einer 2D Transformationsmatrix

Betrachten wir die 2D Transformationsmatrix:

$$A = \begin{pmatrix} 0.55 & -0.52 \\ 0.55 & 0.35 \end{pmatrix}. \quad (1)$$

Für die Standard-Basis (links) erzeugt diese folgende Transformation (rechts):



Dies kann als eine verallgemeinerte Scherungstransformation des Einheitsquadrates, welches durch die Standard-Basis aufgespannt wird, verstanden werden. Nach der Anwendung von A erhalten wir also ein verkleinertes, verzerrtes und rotiertes Einheitsquadrat.

Betrachten wir nun die Singulärwertzerlegung der Matrix A :

$$A = USV^T = \begin{pmatrix} 0.87 & -0.50 \\ 0.50 & 0.87 \end{pmatrix} \begin{pmatrix} 0.80 & 0.0 \\ 0.0 & 0.60 \end{pmatrix} \begin{pmatrix} 0.94 & 0.34 \\ -0.34 & 0.94 \end{pmatrix}^T \quad (2)$$

Wir können A also als drei einzelne Transformationen auffassen. Für $y = Ax$ wird zunächst V^T auf x angewendet, dann S , und zuletzt U ,

$$y_V = V^T x \quad (3a)$$

$$y_S = S y_V \quad (3b)$$

$$y = U y_S. \quad (3c)$$

Wir werden nun den Effekt der einzelnen Transformationen im Detail betrachten.

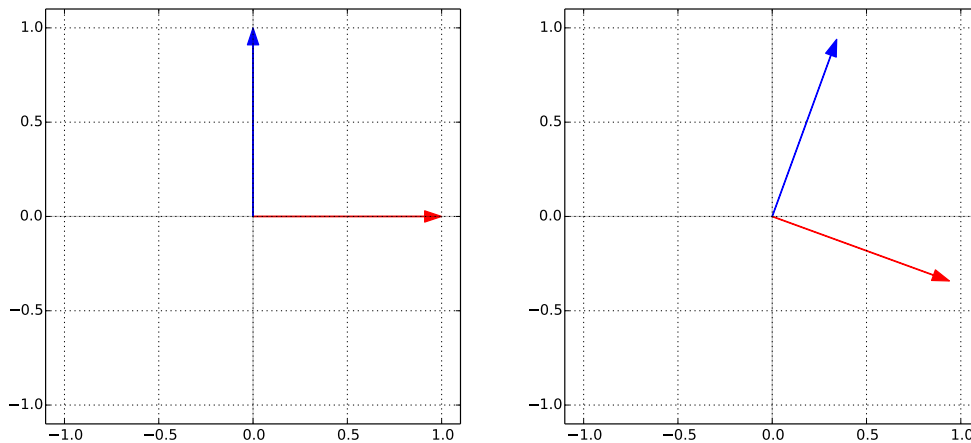
Die Matrix V , und damit auch V^T , ist bei "Konstruktion" der Singulärwertzerlegung eine orthogonale Matrix, d.h.

$$V^T V = \begin{pmatrix} 0.94 & -0.34 \\ 0.34 & 0.94 \end{pmatrix} \begin{pmatrix} 0.94 & 0.34 \\ -0.34 & 0.94 \end{pmatrix} = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} = \text{Id} \quad (4)$$

Wenn wir das Matrixprodukt näher betrachten, dann berechnen wir in Gleichung 4 vier Skalarprodukte, welche die Elemente der Identitätsmatrix formen:

$$\text{Id}_{ij} = \langle v_i, v_j \rangle = \begin{cases} 1 & \text{wenn } i = j \\ 0 & \text{ansonsten} \end{cases} \quad (5)$$

wobei v_i die i -te Spalte von V ist, also $v_1 = (0.94, -0.34)$ und $v_2 = (0.34, 0.94)$. Da $\langle v_1, v_2 \rangle = 0$ sind die Vektoren orthogonal und zwei orthogonale Vektoren formen im \mathbb{R}^2 notwendiger Weise eine orthogonale Basis für den Vektorraum. Auf der rechten Seite in der folgenden Abbildung sind die Spalten von V als Vektoren dargestellt:

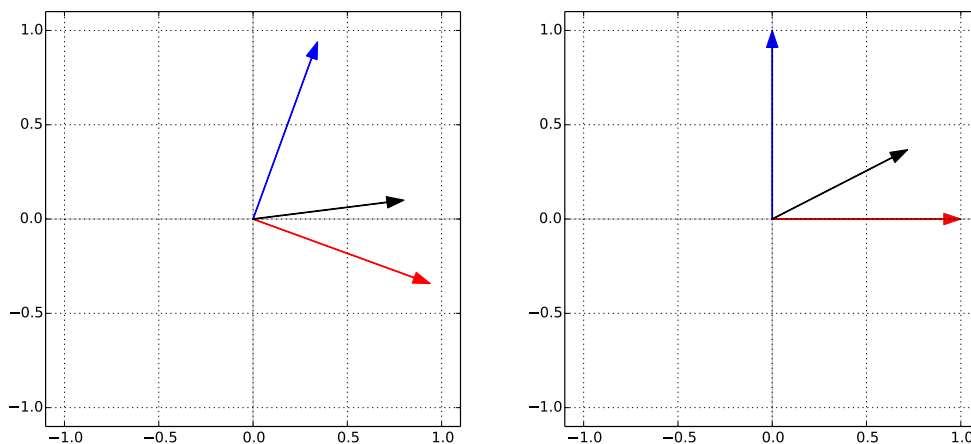


Wir sehen also, dass wir (v_1, v_2) als eine rotierte Version der Standardbasis (e_1, e_2) auffassen können. In der Tat haben wir für unser Beispiel (bei Konstruktion)

$$V^T = \begin{pmatrix} 0.94 & 0.34 \\ -0.34 & 0.94 \end{pmatrix} = \begin{pmatrix} \cos(-20.0^\circ) & -\sin(-20.0^\circ) \\ \sin(-20.0^\circ) & \cos(-20.0^\circ) \end{pmatrix} = R(-20.0^\circ) \quad (6)$$

d.h. V^T ist die Rotationsmatrix $R(-20.0^\circ)$ um -20.0° im Uhrzeigersinn. Die Elemente einer Rotationsmatrix $R(\theta)$ sind also nicht nur durch $\sin(\theta), \cos(\theta)$ gegeben, sondern sie sind auch die kartesischen Koordinaten der rotierten Basis in der Standardbasis (e_1, e_2) ausgedrückt.

Wir haben bereits in den vorhergehenden Tutorien besprochen, dass Vektoren geometrische Objekte sind, die unabhängig von ihrer Beschreibung in einem Koordinatensystem sind: was sich ändert, wenn man ein anderes Koordinatensystem verwendet sind die Koordinaten, zum Beispiel a_1, a_2 , welche den Vektor bezüglich der Basis beschreiben. Für einen gegebenen, fixen Vektor erhalten wir im rotierten Koordinatensystem also neue Koordinaten. Allerdings ist geometrisch auch einfach zu sehen, dass die Beziehung des fixen Vektors zum um den Winkel $-\theta$ rotierten Koordinatensystem die gleiche ist, wie die eines um θ gedrehten Winkels zu einem fixen Koordinatensystem:



Aber wenn die geometrische Beziehung die gleiche ist, dann müssen auch die Koordinaten in beiden Fällen übereinstimmen. Die Elemente des rotierten Vektors, oder des fixen Vektors bezüglich des inversen

rotierten Koordinatensystems, sind also durch

$$R(\theta)\vec{a} = \vec{a}_\theta \quad (7)$$

gegeben.

Wir können uns den Zusammenhang zwischen Rotation eines Vektors \vec{a} und den Koordinaten des rotierten Koordinatensystems (\vec{v}_1, \vec{v}_2) auch noch etwas mathematischer vor Augen führen. Bei Definition sind die Komponenten von \vec{a} bezüglich v_i durch das Skalarprodukt gegeben:

- Warum ist dies der Fall? Wenn wir einen beliebigen Vektor \vec{a} in die Basis (\vec{v}_1, \vec{v}_2) , welche von den Spalten von V aufgespannt wird, projizieren, dann müssen wir $\langle \vec{a}, \vec{v}_i \rangle$ berechnen.
- Rotation um θ oder Basiswechsel in ein Koordinatensystem, welches um $-\theta$ gedreht ist. Für die SVD wird oft die zweite Perspektive bevorzugt, insbesondere im Kontext von PCA.
- S ist eine Skalierungsmatrix, wie man einfach bei einer Anwendung auf die Einheitsvektoren sehen kann.
- Wir können uns V^T also als die Wahl eines Koordinatensystems vorstellen, in der der Effekt der ursprünglichen Matrix A eine einfache Skalierung ist.
- Und was ist mit U ? U ist wieder eine Rotation des skalierten Vektors. Natürlich können wir das auch wieder als einen Wechsel des Koordinatensystems vorstellen.
- Qualitative Diskussion des Falls wenn A nicht quadratisch ist: zunächst zeigen, wie eine Projektion auf die X -Achse als Matrix dargestellt werden kann. Dann Struktur der SVD für ein 2×3 Matrix zeigen:

$$C = USV^T = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix} \begin{pmatrix} s_{11} & 0.0 & 0.0 \\ 0.0 & 0.0 & s_{22} \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{pmatrix} \quad (8)$$

V stellt dann eine orthogonale Basis für die Eingangsdaten dar und U für die Ausgangsdaten.

- Alle Ideen funktionieren konzeptionell genauso in \mathbb{R}^n und Vektoren nicht mehr geometrische Objekte sondern beliebige Daten darstellen.
- Zusammenfassung: V und U stellen Koordinatensystem dar, welche

2 Eigenschaften der Singularwertzerlegung

Definition Sei $A \in \mathbb{R}^{m \times n}$ eine beliebige reelle Matrix. Dann gibt es orthogonale Matrizen $U \in \mathbb{R}^{m \times m}$ und $V \in \mathbb{R}^{n \times n}$, so dass

$$U^T A V = \Sigma = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m \times n} \quad (9)$$

mit $p = \min(m, n)$ und $\sigma_1 \geq \dots \geq \sigma_p \geq 0$.

Eigenschaften der SVD Sei $A \in \mathbb{R}^{m \times n}$ eine Matrix und $A = U \Sigma V^T$ eine Singularwertzerlegung von A mit $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_p)$, wobei $p = \min(m, n)$. Dann gelten die folgenden Aussagen.

- Bezeichnet u_i und v_i die Spalten von U und V , so gilt:

$$A v_i = \sigma_i u_i \quad A^T u_i = \sigma_i v_i \quad \text{für } i = 1, \dots, p$$

- Ist $\sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_p = 0$, so gilt:

$$\begin{aligned} \text{rang } A &= r \\ \ker A &= \text{span}(v_{r+1}, \dots, v_p) \\ \text{im } A &= \text{span}(u_1, \dots, u_r) \end{aligned}$$

- Die Quadrate $\sigma_1^2, \dots, \sigma_p^2$ der Singulärwerte sind die Eigenwerte von $A^T A$ und $A A^T$ zu den Eigenvektoren v_1, \dots, v_p bzw. u_1, \dots, u_p .

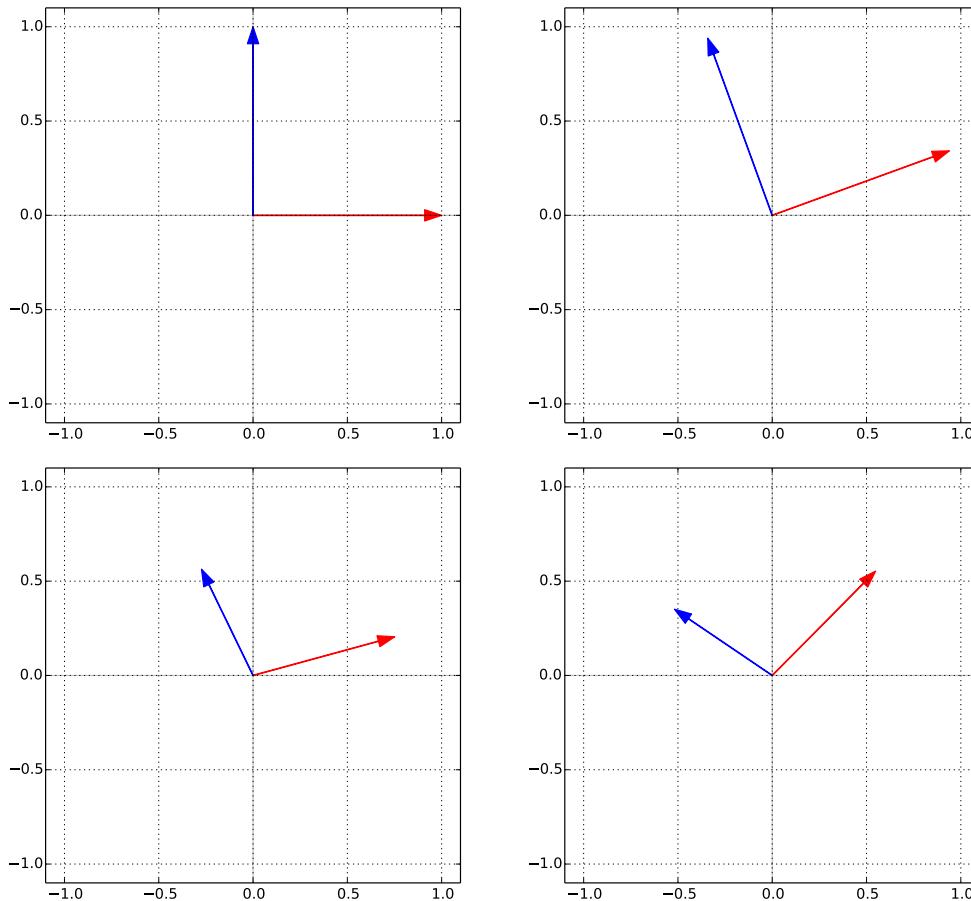


Abbildung 1: Zerlegung einer Transformation der Euklidischen Standardbasis (links oben) durch die SVD: Im ersten Schritt erfolgt eine Rotation durch V^T . Im zweiten Schritt führt S eine Skalierung der rotierten Vektoren durch. Zuletzt erfolgt eine erneute Rotation durch U .

Die Eigenschaften können wie folgt gezeigt werden:

- a) Einsetzen der Singularwertzerlegung für A ergibt:

$$Av_i = U\Sigma V^T v_i = U\Sigma e_i = U\sigma_i e_i = \sigma_i u_i$$

$$A^T u_i = (U\Sigma V^T)^T u_i = V\Sigma U^T u_i = V\Sigma e_i = V\sigma_i e_i = \sigma_i v_i$$

- b) Aus $Av_i = \sigma_i u_i$ folgt $Av_i \neq 0$ für $i = 1, \dots, r$ und $Av_i = 0$ für $i = r + 1, \dots, p$. Sei nun $x = \sum_{i=1}^n x_i v_i \in \mathbb{R}^{n \times n}$. Dann gilt:

$$Ax = A\left(\sum_{i=1}^n x_i v_i\right) = \sum_{i=1}^n x_i Av_i = \sum_{i=1}^n x_i \sigma_i u_i = \sum_{i=1}^r x_i \sigma_i u_i \in \text{span}(u_1, \dots, u_r)$$

- c) Einsetzen der Singularwertzerlegung für A ergibt die folgenden Eigenwertzerlegung für $A^T A$ und AA^T :

$$A^T A = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma U^T U\Sigma V^T = V\Sigma\Sigma V^T$$

$$AA^T = U\Sigma V^T (U\Sigma V^T)^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma U^T$$

3 Pseudoinverse

Sei $A \in \mathbb{R}^{m \times n}$ eine Matrix und $A = U\Sigma V^T$ eine Singularwertzerlegung von A mit $\text{rang } A = r$ und $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$. Dann ist die Pseudoinverse $A^+ \in \mathbb{R}^{n \times m}$ gegeben durch

$$A^+ = V\Sigma^+U^T \quad \text{mit} \quad \Sigma^+ = \text{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0, \dots, 0).$$

Das Programm `src/least_squares_pseudoinverse.py`, welches wir auf der Homepage zur Verfügung stellen, demonstriert die geometrischen Transformationen, welche bei der Lösung eines Ausgleichsproblems mit Hilfe der Pseudo-Inversen auftreten:

4 Ausgleichsrechnung und Singularwertzerlegung

Sei $A \in \mathbb{R}^{m \times n}$ eine Matrix und $b \in \mathbb{R}^m$ ein Vektor. Wir wollen die Lösung des Ausgleichsproblems $Ax = b$ unter Verwendung der SVD von A bestimmen.

Nehmen wir dazu zunächst an, dass Σ eine Diagonalmatrix mit absteigend sortierten Diagonalelementen $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$ ist. Dann haben wir:

$$\Sigma x - b = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \sigma_1 x_1 - b_1 \\ \vdots \\ \sigma_n x_n - b_n \end{pmatrix} \quad (10)$$

Bezeichnet r den Rang von Σ , so ist $\sigma_i = 0$ für $i = r + 1, \dots, n$. Wir erhalten also:

$$\|\Sigma x - b\|^2 = \sum_{i=1}^n (\sigma_i x_i - b_i)^2 \quad (11a)$$

$$= \sum_{i=1}^r (\sigma_i x_i - b_i)^2 + \sum_{i=r+1}^n b_i^2 \quad (11b)$$

Der erste Term ist genau dann minimal, wenn $x_i = \frac{b_i}{\sigma_i}$ für $i = 1, \dots, r$ ist. Der zweite Term in Gleichung (11b) ist konstant und spielt daher bei einer Minimierung keine Rolle. Die Komponenten x_i können also für $i = r + 1, \dots, n$ beliebig gewählt werden:

$$x = \left(\frac{b_1}{\sigma_1}, \dots, \frac{b_r}{\sigma_r}, \xi_{r+1}, \dots, \xi_n \right)^T \quad (12)$$

Mit Hilfe der Pseudoinversen

$$\Sigma^+ = \begin{pmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_r} \\ & & & 0 \end{pmatrix} \quad (13)$$

von Σ lässt sich x auch schreiben als:

$$x = \Sigma^+ b + \sum_{i=r+1}^n \xi_i e_i \quad (14)$$

Sei nun A beliebig und $A = U\Sigma V^T$ eine SVD von A . Da U eine orthogonale Matrix ist, gilt $\|U^T z\| = \|z\|$ für alle $z \in \mathbb{R}^m$. Daraus folgt:

$$\|Ax - b\|^2 = \|U^T(Ax - b)\|^2 = \|U^T Ax - U^T b\|^2 \quad (15a)$$

$$= \|U^T U \Sigma V^T x - U^T b\|^2 = \|\Sigma V^T x - U^T b\|^2 \quad (15b)$$

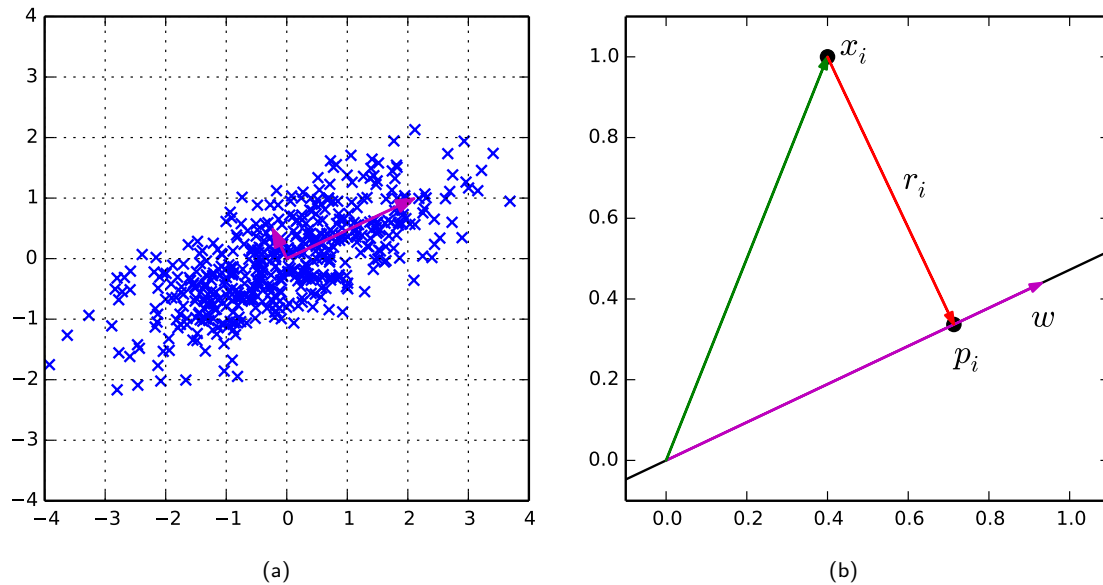


Abbildung 2: (a) Eine Menge von zweidimensionalen Datenpunkten. (b) Projektion p_i des Datenpunktes x_i auf auf den Einheitsvektor w . Der Fehler der Projektion ist r_i .

Sei nun $y = V^T x$. Aufgrund der Orthogonalität von V gilt dann $\|y\| = \|V^T x\| = \|x\|$. Da V außerdem vollen Rang hat (d.h. für alle y gibt es ein x mit $V^T x = y$), können wir also, anstatt über alle x zu minimieren, auch über alle y minimieren:

$$\arg \min_{x \in \mathbb{R}^n} \|Ax - b\|^2 = \arg \min_{x \in \mathbb{R}^n} \|\Sigma V^T x - U^T b\|^2 \Leftrightarrow \arg \min_{y \in \mathbb{R}^n} \|\Sigma y - U^T b\|^2 \quad (16)$$

Bei letzterem Minimierungsproblem handelt es sich um ein Ausgleichsproblem in Diagonalgestalt, dessen Lösung sich somit aus Gleichung (14) ergibt:

$$x = Vy = V \left(\Sigma^+ U^T b + \sum_{i=r+1}^n \xi_i e_i \right) = V \Sigma^+ U^T b + \sum_{i=r+1}^n \xi_i v_i \quad (17)$$

Hierbei bezeichnet v_i die i -te Spalte von V .

Ist der Rang r kleiner der Anzahl der Spalten von A so erhalten wir einen $n - r$ -dimensionalen Lösungsraum. Bei vielen technischen Anwendungen reicht es oft eine Lösung zu wählen, z.B. die Lösung mit minimaler Norm. Anhand von Gleichung (17) sieht man, dass x genau dann minimal ist, wenn der Summenterm verschwindet. Da die Spalten von V linear unabhängig sind, ist dies gleichbedeutend mit $\xi_i = 0$ für $i = r + 1, \dots, n$. Die Lösung mit kleinster Norm x^+ für das Ausgleichsproblem $Ax - b$ ist somit eindeutig bestimmt durch:

$$x = V \Sigma^+ U^T b \quad (18)$$

5 Hauptkomponentenanalyse für Punktwolken

Ziel der Hauptkomponentenanalyse (engl. principal component analysis (PCA)) ist es einen Datensatz, welcher in Form einer Menge von Punkten im \mathbb{R}^n gegeben ist, in einem neuen Koordinatensystem darzustellen. Die erste Koordinatenachse wird dabei so gewählt, dass der Fehler, welcher bei der Projektion der Datenpunkte auf die Koordinatenachse entsteht, minimal unter allen Koordinatenachsen ist. Analog werden die anderen Koordinatenachsen gewählt. Die zweite Koordinatenachse wird zum Beispiel so gewählt, dass der Fehler der Projektion auf den Unterraum, welcher durch die erste und zweite Koordinatenachse aufgespannt wird, minimal ist.

Wir wollen uns die Hauptkomponentenanalyse an einem einfachen Beispiel anschauen. Gegeben sei dazu eine Menge von zweidimensionalen Punkten $x_i \in \mathbb{R}^2$, $i = 1, \dots, m$ (Abbildung 2(a)). Der Einfachheit halber wollen wir annehmen, dass der Schwerpunkt der Daten im Ursprung liegt. Gesucht ist nun

eine neue Koordinatenachse, so dass beim projizieren der Datenpunkte ein minimaler Fehler entsteht. Eine Koordinatenachse lässt sich einfach mittels eines Einheitsvektor w (Abbildung 2) beschreiben. Die Projektion p_i des Datenpunktes x_i auf den durch w aufgespannten Unterraum (d.h. die durch w definierte Gerade) ist dann gegeben durch:

$$p_i = \langle x_i, w \rangle w \quad (19)$$

Der Fehler, welcher bei der Projektion entsteht, lässt sich leicht mittels des Satzes von Pythagoras bestimmen:

$$\|r_i\|^2 = \|x_i\|^2 - \|p_i\|^2 \quad (20)$$

Summieren wir nun $\|r_i\|^2$ für alle Datenpunkte auf, so erhalten wir ein Maß für den Fehler der Projektion der Datenpunkte auf die Koordinatenachse w . Indem wir über alle Koordinatenachsen minimieren, erhalten wir die Koordinatenachse mit minimalem Fehler:

$$\arg \min_{\|w\|=1} \sum_{i=1}^m \|r_i\|^2 = \arg \min_{\|w\|=1} \sum_{i=1}^m \left(\underbrace{\|x_i\|^2}_{=\text{const}} - \|p_i\|^2 \right) = \arg \max_{\|w\|=1} \sum_{i=1}^m \|p_i\|^2 \quad (21)$$

Dabei nutzten wir aus, dass $\|x_i\|^2$ bzgl. der Minimierung konstant ist und vernachlässigt werden kann. Außerdem nutzen wir aus, dass das Minimieren des Negativen eines nicht negativen Ausdrucks äquivalent zum Maximieren des Ausdrucks ist.

Wir wollen nun das Maximierungsproblem unter Verwendung der SVD lösen. Dazu müssen wir zunächst Gleichung (21) in Matrixnotation schreiben. Als erstes schreiben wir $\|p_i\|^2$ als Skalarprodukt

$$\|p_i\|^2 = \|\langle x_i, w \rangle w\|^2 = \langle x_i, w \rangle^2 \|w\|^2 = \langle x_i, w \rangle^2, \quad (22)$$

wobei wir ausnutzen, dass w ein Einheitsvektor ist. Die Datenpunkte fassen wir in einer Matrix

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix} \quad (23)$$

zusammen und zwar so, dass jede Zeile genau einem Datenpunkt entspricht. Der Vektor Xw besteht dann aus den Skalarprodukten $x_i^T w = \|p_i\|$ und wir erhalten:

$$\sum_{i=1}^m \|p_i\|^2 = \|Xw\|^2 \quad (24)$$

Gleichung (21) können wir also auch schreiben als:

$$\arg \max_{\|w\|=1} \sum_{i=1}^m \|p_i\|^2 = \arg \max_{\|w\|=1} \|Xw\|^2 \quad (25)$$

Sei nun $X = U\Sigma V^T$ eine SVD von X . Im Tutorium haben wir gezeigt, dass $\max_{\|w\|=1} \|Xw\|^2 = \sigma_1^2$. Das Maximum wird dabei von der ersten Spalte von V angenommen, denn es gilt:

$$\|Xv_1\|^2 = \|U\Sigma V^T v_1\|^2 = \|U\Sigma e_1\|^2 = \|\Sigma e_1\|^2 = \|\sigma_1 e_1\|^2 = \sigma_1^2 \quad (26)$$

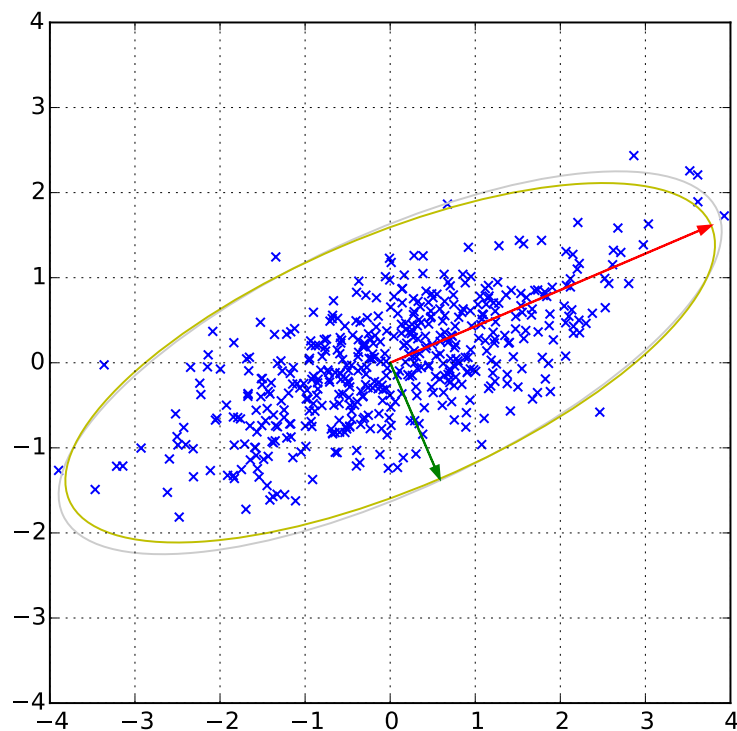
Zusammengefasst erhalten wir somit:

$$\arg \max_{\|w\|=1} \|Xw\|^2 = v_1 \quad (27)$$

In folgendem Beispiel werden die Hauptkomponenten einer künstlich erzeugten Punktwolke berechnet. Die Erzeugung der Punktwolke erfolgt in zwei Schritten. Zunächst wird eine Matrix mit zufälligen (normalverteilten) Koordinaten erzeugt (`np.random.randn`). Durch Multiplikation der ersten und zweiten Spalte mit $\sqrt{2.0}$ und $\sqrt{0.25}$ wird dann erreicht, dass die erste und zweite Koordinate der Punkte unterschiedliche Varianzen haben (nämlich 2.0 und 0.25). Das bedeutet also, dass die Punktwolke stärker entlang der ersten Koordinatenachse ausgedehnt ist. Durch Rotation um 25° erhalten wir somit einen Datensatz welcher entlang der 25° -Achse ausgedehnt ist.

Für die Bestimmung der Hauptkomponenten wird zunächst der Schwerpunkt der Punktwolke berechnet und dann abgezogen um den Schwerpunkt in den Koordinatenursprung zu verschieben. Nun wird eine SVD der Punktwolke bestimmt. Zu beachten ist hier, dass von `numpy.linalg.svd` nicht V sondern V^T zurückgegeben wird. Das bedeutet die i -te Spalte von V findet sich in $V[i]$. Die Singulärwerte stellen ein Maß für die Ausdehnung der Daten in Richtung der jeweiligen Hauptachse dar (genauer: $\sqrt{\sigma_i^2/(m-1)}$ ist die korrigierte Stichprobenvarianz). Zur besseren Visualisierung werden daher die Hauptkomponenten mit dem dreifachen der geschätzten Standardabweichungen multipliziert.

Im folgenden ist die Ausgabe eines Programmlaufes zu sehen. Die erste Hauptachse ist in rot und die zweite Hauptachse ist in grün eingezeichnet. Zur besseren Visualisierung ist außerdem die durch die skalierten Hauptkomponenten definierte Ellipse eingezeichnet.




```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
from math import sqrt, atan2

N = 500
std_x = sqrt(2.0)
std_y = sqrt(0.25)
theta = np.deg2rad(25.0)

# generate point cloud
A = np.random.randn(N, 2)
A[:,0] *= std_x
A[:,1] *= std_y
cn = np.cos(theta)
sn = np.sin(theta)
R = np.array([[cn,-sn],[sn,cn]])
A = R.dot(A.T).T

# subtract mean to move center to coordinate origin
A -= A.mean(0)
# plot point cloud
plt.scatter(A[:,0], A[:,1], c='r', marker='x')

# compute principal components
U,S,V = np.linalg.svd(A, full_matrices=0)

# project data points onto reduced space
P = V[0:1,:].dot(A.T)
# Reconstruct projected data points
Ap = (V[0:1,:].T.dot(P)).T
plt.scatter(Ap[:,0], Ap[:,1], marker='x', c='b')

# plotting

# compute variance (using the formula from statistics)
std1 = sqrt(S[0]**2 / (N-1))
std2 = sqrt(S[1]**2 / (N-1))
pc1 = V[0] * 3.0 * std1
pc2 = V[1] * 3.0 * std2

# show ellipse defined by principal components and estimated standard deviation
plt.gca().add_artist(matplotlib.patches.Ellipse(xy=(0,0), width=6*std1,
                                                height=6*std2,
                                                angle=np.rad2deg(atan2(pc1[1], pc1[0])),
                                                fill=False, color='y'))

# plot principal components
plt.arrow( 0.0, 0.0, pc1[0], pc1[1], color='r', width=0.005, length_includes_head=True)
plt.arrow( 0.0, 0.0, pc2[0], pc2[1], color='g', width=0.005, length_includes_head=True)

plt.xlim(-4, 4)
plt.ylim(-4, 4)
plt.grid(True)
plt.gca().set_aspect('equal')
plt.show()
```

A Formale Definition der Singularwertzerlegung

Satz 1. Sei $A \in \mathbb{R}^{m \times n}$ eine beliebige reelle Matrix. Dann gibt es orthogonale Matrizen $U \in O(m)$ und $V \in O(n)$, so dass

$$U^T A V = \Sigma = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m \times n} \quad (28)$$

mit $p = \min(m, n)$ und $\sigma_1 \geq \dots \geq \sigma_p \geq 0$.

Beweis. Induktion nach n . Für $n = 1$ ist nichts zu zeigen. Sei nun die Aussage für $n - 1$ bereits bewiesen. Es bezeichne $S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ die n -dimensionale Einheitsphäre. Die Menge S^{n-1} ist abgeschlossen und beschränkt in \mathbb{R}^n und daher kompakt. Die stetige Funktion $x \mapsto \|Ax\|$ nimmt somit auf S^{n-1} ihr Maximum an. Bezeichnet $\sigma_1 := \max_{\|x\|=1} \|Ax\|$, dann gibt es also Vektoren $u_1, v_1 \in S^{n-1}$, so dass

$$A v_1 = \sigma_1 u_1.$$

Der Vektor u_1 kann zu einer Orthonormalbasis u_1, \dots, u_m von \mathbb{R}^m und der Vektor v_1 kann zu einer Orthonormalbasis v_1, \dots, v_n von \mathbb{R}^n vervollständigt werden. Bezeichne $U_1 = (u_1, \dots, u_m)$ die Matrix mit Spalten u_i ($i = 1, \dots, m$) und $V_1 = (v_1, \dots, v_n)$ die Matrix mit Spalten v_i ($i = 1, \dots, n$). Dann sind $U_1 \in O(m)$ und $V_1 \in O(n)$ orthogonale Matrizen und wir erhalten

$$S = U_1^T A V_1 = \begin{pmatrix} \sigma_1 & w^T \\ 0 & B \end{pmatrix}$$

mit $w \in \mathbb{R}^{n-1}$ und $B \in \mathbb{R}^{(m-1) \times (n-1)}$. Sei $y = V_1(\sigma_1, w)^T = \sigma_1 v_1 + w_1 v_2 + \dots + w_{n-1} v_n$. Dann folgt:

$$\begin{aligned} \sigma_1^2 &= \max_{\|x\|=1} \|Ax\|^2 \geq \frac{\|Ay\|^2}{\|y\|^2} = \frac{\|U_1 S V_1^T y\|^2}{\|y\|^2} = \frac{\|S V_1^T y\|^2}{\|y\|^2} \\ &= \frac{(\sigma_1^2 + \|w\|^2)^2 + \|Bw\|^2}{\sigma_1^2 + \|w\|^2} \geq \frac{(\sigma_1^2 + \|w\|^2)^2}{\sigma_1^2 + \|w\|^2} = \sigma_1^2 + \|w\|^2 \end{aligned}$$

Und daher $w = 0$. Nach Induktionsvoraussetzung gibt es eine Zerlegung $\Sigma_B = U_B^T B V_B$. Daraus folgt:

$$S = U_1^T A V_1 = \begin{pmatrix} \sigma_1 & 0 \\ 0 & U_B \Sigma_B V_B^T \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & U_B \end{pmatrix}}_{:=U_2} \underbrace{\begin{pmatrix} \sigma_1 & 0 \\ 0 & \Sigma_B \end{pmatrix}}_{:=\Sigma} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & V_B^T \end{pmatrix}}_{:=V_2^T} = U_2 \Sigma V_2^T$$

Die Matrizen U_2 und V_2 sind orthogonal und Σ ist eine Diagonalmatrix. Setzen wir nun $U := U_1 U_2$ und $V := V_1 V_2$, so erhalten wir:

$$U^T A V = U_2^T U_1^T A V_1 V_2 = U_2^T S V_2 = \Sigma \quad \square$$

Die Zerlegung $A = U \Sigma V^T$ heißt *Singularwertzerlegung* von A . Die Diagonalelemente σ_i heißen *Singularwerte* von A . Die Spalten von U werden als *linke Singulärvektoren* und die von V *rechte Singulärvektoren* von A bezeichnet.

B Berechnung der Singularwertzerlegung

Wir wollen eine Singularwertzerlegung für die folgende 2×2 -Matrix bestimmen:

$$A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix} \quad (29)$$

Dazu berechnen wir zunächst die Eigenwerte und Eigenvektoren von

$$A^T A = \begin{pmatrix} 25 & 15 \\ 15 & 25 \end{pmatrix}. \quad (30)$$

Die Eigenwerte von $A^T A$ zu sind die Nullstellen des charakteristischen Polynoms. Für dieses ergibt sich:

$$\det(A^T A - I\lambda) = \det \begin{pmatrix} 25 - \lambda & 15 \\ 15 & 25 - \lambda \end{pmatrix} = \lambda^2 - 50\lambda + 400 \quad (31)$$

Und wir erhalten für die Eigenwerte:

$$\lambda^2 - 50\lambda + 400 = 0 \quad \Leftrightarrow \quad \lambda_1 = 40, \lambda_2 = 10 \quad (32)$$

Um einen Eigenvektor v_1 zum Eigenwert $\lambda_1 = 40$ zu bestimmen müssen wir das folgende homogene Gleichungssystem lösen:

$$(A^T A - I\lambda_1)v_1 = 0 \quad \Leftrightarrow \quad \begin{pmatrix} -15 & 15 \\ 15 & -15 \end{pmatrix} v_1 = 0 \quad (33)$$

Lösen des Gleichungssystems und anschließendes normieren des Lösungsvektors ergibt:

$$v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \quad (34)$$

Ein Eigenvektor v_2 zum Eigenwert λ_2 kann auf die gleiche Weise durch Lösen von $(A^T A - I\lambda_2)v_2 = 0$ bestimmt werden. Da ein solcher Eigenvektor orthogonal zu v_1 sein muss kann man diesen jedoch auch einfach mittels einer Rotation von v_1 um 90 Grad erhalten:

$$v_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \quad (35)$$

Entsprechend der bereits diskutierten Eigenschaften sind die Singulärwerte von A folglich gegeben durch:

$$\sigma_1 = \sqrt{\lambda_1} = \sqrt{40} \quad \sigma_2 = \sqrt{\lambda_2} = \sqrt{10} \quad (36)$$

Zu bestimmen bleiben die linken Singulärvektoren u_1 und u_2 . Diese erhält man aus den Eigenschaften der Singulärwertzerlegung:

$$u_1 = \frac{Av_1}{\sigma_1} = \frac{1}{\sqrt{40}} \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix} \quad (37a)$$

$$u_2 = \frac{Av_2}{\sigma_2} = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix} \quad (37b)$$

Eine Singulärwertzerlegung von A ist somit gegeben durch:

$$\underbrace{\begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix}}_{=A} = \underbrace{\begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix}}_{=U} \underbrace{\begin{pmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{pmatrix}}_{=\Sigma} \underbrace{\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^T}_{=V^T} \quad (38)$$

C Matrixnorm und Singulärwertzerlegung

Sei $A \in \mathbb{R}^{m \times n}$ und es bezeichne σ_1 den größten Singulärwert von A . Dann gilt:

$$\|A\| := \max_{\|x\|=1} \|Ax\| = \sigma_1 \quad (39)$$

Für positive reelle Zahlen ist Quadrieren streng monoton. D.h. es gilt $a < b \Leftrightarrow a^2 < b^2$ für alle $a, b \geq 0$. Äquivalent zu Gleichung (39) kann also das Maximum von $\|Ax\|^2$ gesucht werden.

Betrachten zunächst den Fall, dass A eine Diagonalmatrix mit absteigend sortierten Diagonalelementen $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$ ist. Dann gilt

$$\Sigma x = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \sigma_1 x_1 \\ \vdots \\ \sigma_n x_n \end{pmatrix} \quad (40)$$

und wir können abschätzen:

$$\|\Sigma x\|^2 = \sum_{i=1}^n (\sigma_i x_i)^2 \leq \sigma_1^2 \sum_{i=1}^n x_i^2 = \sigma_1^2 \|x\|^2 \quad (41)$$

Daraus erhalten wir $\max_{\|x\|=1} \|\Sigma x\|^2 \leq \sigma_1^2$. Da für den Vektor $x = (1, 0, \dots, 0)^T$ Gleichheit angenommen wird, folgt:

$$\max_{\|x\|=1} \|\Sigma x\|^2 = \sigma_1^2 \quad (42)$$

Sei nun A beliebig und $A = U\Sigma V^T$ eine SVD von A . Dann gilt:

$$\|Ax\|^2 = \langle Ax, Ax \rangle = (Ax)^T Ax = x^T A^T Ax \quad (43a)$$

$$= x^T (U\Sigma V^T)^T (U\Sigma V^T) x = x^T V \Sigma^T U^T U \Sigma V^T x \quad (43b)$$

$$= x^T V \Sigma^T \Sigma V^T x \quad (43c)$$

Sei nun $y = V^T x$. Da $\|y\| = \|V^T x\| = \|x\|$ und V vollen Rang hat, können wir anstatt über alle x zu maximieren auch über alle y maximieren:

$$\max_{\|x\|=1} \|Ax\|^2 = \max_{\|x\|=1} x^T V \Sigma^T \Sigma V^T x = \max_{\|y\|=1} y^T \Sigma^T \Sigma y = \max_{\|y\|=1} \|\Sigma y\|^2 = \sigma_1^2 \quad (44)$$