Die Singulärwert-Zerlegung

Eine grundlegende Matrix-Faktorisierung

© Edgar M.E. Wermuth

Technische Hochschule Nürnberg
Fakultät für Angewandte Mathematik,
Physik u. Allgemeinwissenschaften

Stark erweiterte Fassung eines Vortrags vom 26.02.2010
(Erweiterungen 2014, 2022, 2023)

0. Vorbemerkung

Die Singulärwertzerlegung (singular value decomposition, **SVD**) einer beliebigen Matrix $A \in \mathbb{R}^{m\times n}$ ($\in \mathbb{C}^{m\times n}$) faktorisiert die Matrix auf so transparente Weise, dass *Umkehrabbildung* (sofern existent), *Lösungen von Gleichungssystemen*, *Rang*, *Kern* und *Bild* von A direkt ablesbar sind.

Im Falle der Nichtinvertierbarkeit ergeben sich unmittelbar *Näherungslösungen* und *Pseudo-Inverse*. Auch die bestmöglichen Näherungsmatrizen *niedrigeren Ranges* zu A lassen sich anhand der SVD sofort angeben.

Obwohl die SVD im wesentlichen schon von Beltrami (1873) und Jordan (1874) entdeckt und elegant hergeleitet wurde, erkannte man die zentrale Bedeutung dieser Matrix-Zerlegung erst relativ spät, ab Ende der 1930er Jahre. Erst da wurde sie Teil des Grundwissens der linearen Algebra.

Der Begriff „Singulärwert“ ist bei Matrizen erst seit ca. 1950 gebräuchlich. Robuste Berechnungsverfahren (Kahan/Golub/Reinsch) wurden in den 1960er Jahren entwickelt.

1. Ebene Geometrie der linearen Abbildungen

Eine lineare Abbildung

$$f : \vec{x} \mapsto A\vec{x} \quad (\vec{x} \in \mathbb{R}^n)$$

mit $A \in \mathbb{R}^{m\times n}$ ist einerseits dadurch gekennzeichnet, dass *gerade Linien auf gerade Linien* abgebildet werden. Bezeichnen wir das f-Bild von $\vec{x}$ kurz mit $\vec{x}'$, drückt sich das so aus:

$$t\,\vec{a} + (1-t)\,\vec{b} \quad (t \in \mathbb{R})$$

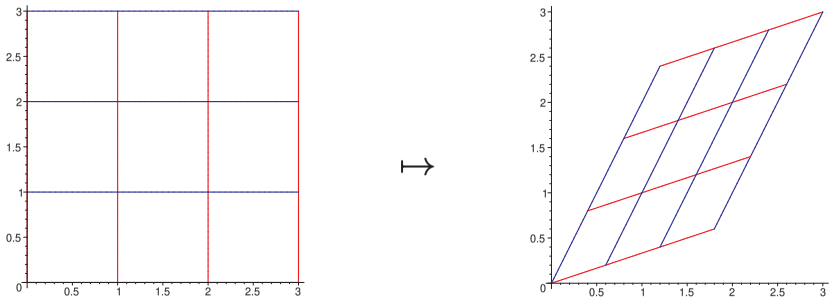
geht unter f über in

$$t\,\vec{a}' + (1-t)\,\vec{b}' \quad (t \in \mathbb{R}).$$

Im Falle $\vec{a}' = \vec{b}'$ *entartet* die Bildgerade zu einem Punkt.

Vollständiger wird das Bild, wenn wir die *zweidimensionale* Wirkung linearer Abbildungen beschreiben: Quadrate werden auf Parallelogramme abgebildet. Genauer:

Zueinander parallele Quadrate werden auf untereinander parallele und ähnliche Parallelogramme abgebildet.



Noch aufschlussreicher wird's, wenn wir die Bilder von *Kreisen* ins Spiel bringen.

Da die ebenen linearen Abbildungen – bis auf einen eventuellen einheitlichen Vergrößerungsfaktor – aus Quadraten sozusagen schief angeguckte Quadrate machen und damit auch die ganze Ebene einfach in ein schiefes Licht rücken, werden aus Kreisen schief angeguckte Kreise; und das sind bekanntlich Ellipsen!

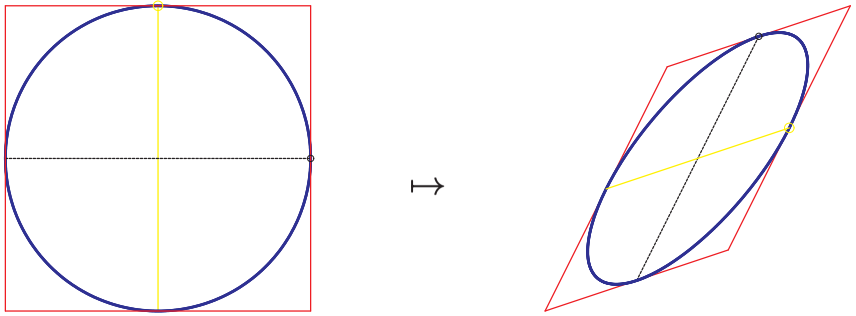
Durch eine lineare Abbildung werden *Kreise* auf *Ellipsen* abgebildet.

Sind $\vec{a}$ und $\vec{b}$ zwei zueinander senkrechte Radiusvektoren eines Kreises um $\vec{0}$, ist

$$\cos t \cdot \vec{a}' + \sin t \cdot \vec{b}' \quad (0 \le t \le 2\pi)$$

die zugehörige Bildellipse; im Ausnahmefall ist's nur eine Strecke oder bloß der Punkt $\vec{0}$.

Jedes einem Kreis umbeschriebene *Quadrat* wird auf ein seiner Bildellipse umbeschriebenes *Parallelogramm* abgebildet, und die beiden zueinander senkrechten Berührachsen des Kreises werden dabei auf Berührachsen der Ellipse bzgl. des Parallelogramms, zwei sog. konjugierte Durchmesser der Ellipse, abgebildet:



Übrigens haben alle der Ellipse umbeschriebenen Bild-Parallelogramme *denselben* Flächeninhalt: sie sind ja Bilder von *einem* Kreis umbeschriebenen Quadraten unter einunddersselben linearen Abbildung.

Wie sind nun diese Bild-Parallelogramme zu Berührquadraten des Urbildkreises von *beliebigen* der Ellipse umbeschriebenen Parallelogrammen möglichst einfach zu unterscheiden?

Antwort:

Konjugiert sind zwei Durchmesser einer Ellipse genau dann, wenn sie *parallel* zu den Seiten des zugehörigen Berührparallelogramms verlaufen.

Denn genau dann sind die Urbilder der Parallelogramme dem Kreis umbeschriebene *Quadrate* – und nicht nur Parallelogramme –, wegen Erhaltung der Parallelität unter linearen Abbildungen.

Unter den dem Kreis umbeschriebenen Parallelogrammen (Rauten) haben die Quadrate den kleinsten Flächeninhalt. Also haben auch die zu konjugierten Durchmessern einer Ellipse gehörenden *die kleinste Fläche unter allen Berührparallelogrammen der Ellipse*.

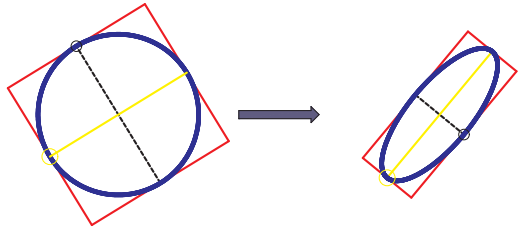
Insbesondere folgt aus dem Festgestellten:

Geometrische Quintessenz der SVD

Die beiden Hauptachsen der Ellipse sind Bilder zweier *zueinander senkrechter* Durchmesser des Urbildkreises.

Beispiel:

Lineare Abbildung durch die Matrix $\begin{pmatrix} 0.4 & 0.6 \\ 0.8 & 0.2 \end{pmatrix}$



Der Kreis $\cos t \cdot \vec{u}_1 + \sin t \cdot \vec{u}_2 \ \ (0 \le t \le 2\pi)$ mit

$$\vec{u}_1 = \begin{pmatrix} 0.8506508 \\ 0.5257311 \end{pmatrix}, \qquad \vec{u}_2 = \begin{pmatrix} -0.5257311 \\ 0.8506508 \end{pmatrix}$$

wird abgebildet auf die Ellipse

$$\cos t \cdot \vec{u}'_1 + \sin t \cdot \vec{u}'_2 \ \ (0 \le t \le 2\pi)$$

mit $\sigma_1 = 1.0233346,\ \sigma_2 = 0.39087903$ und

$$\vec{u}'_1 = \sigma_1 \begin{pmatrix} 0.6407474 \\ 0.7677517 \end{pmatrix},\ \vec{u}'_2 = \sigma_2 \begin{pmatrix} 0.7677517 \\ -0.6407474 \end{pmatrix}.$$

Selbst in diesem einfachen Fall erfordert eine exakte Darstellung der Vektoren *in geschlossener Form* schon etwas unübersichtliche Wurzelausdrücke. (Siehe nächste Seite.)

Natürlich kann man das lineare Bild des Einheitskreises im $\mathbb{R}^2$ auch ganz elementar *analytisch* diskutieren: Mit $f(t) := |\vec{x}(t)|^2$, $\vec{x}(t) := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$, $f(t) = (a\cos t + b\sin t)^2 + (c\cos t + d\sin t)^2$, gilt $f'(t) = (b^2+d^2-a^2-c^2)\sin 2t + 2(ab+cd)\cos 2t$. Also $f'(t) = 0$ $\Leftrightarrow \tan 2t = \frac{2(ab+cd)}{a^2+c^2-b^2-d^2} \vee \cot 2t = \frac{a^2+c^2-b^2-d^2}{2(ab+cd)} \vee a^2+c^2 = b^2+d^2, ab+cd = 0$. Da tan und cot π-periodisch, gibt's 4 Nullstellen, außer im dritten Fall (Bild ein *Kreis*). Ferner: $f'(t_1) = 0 \ \Rightarrow \ \vec{x}(t_1 + \pi/2) \perp \vec{x}(t_1)$.

Algebraische Bestimmung von $\sigma_1,\ \sigma_2,\ \vec{u}_1,\ \vec{u}_2$ für eine invertierbare Matrix $A \in \mathbb{R}^{2\times 2}$:

Mit $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $A^{-1} = \frac{1}{\Delta}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$, $\Delta = ad - bc = \det A \neq 0$ gilt $\begin{pmatrix} u \\ v \end{pmatrix} = A\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax+by \\ cx+dy \end{pmatrix}$, $\begin{pmatrix} x \\ y \end{pmatrix} = A^{-1}\begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{\Delta}\begin{pmatrix} du-bv \\ -cu+av \end{pmatrix}$, also $x^2+y^2 = \frac{1}{\Delta^2}\Big((c^2+d^2)u^2 - 2(bd+ac)uv + (a^2+b^2)v^2\Big)$ und $x^2+y^2 = 1 \Leftrightarrow (c^2+d^2)u^2 - 2(bd+ac)uv + (a^2+b^2)v^2 = \Delta^2$, wobei $a^2+b^2 > 0,\ c^2+d^2 > 0$ wegen $\Delta \neq 0$.

Die Matrix $\begin{pmatrix} c^2+d^2 & -ac-bd \\ -ac-bd & a^2+b^2 \end{pmatrix}$ der quadratischen Form in u und v ist positiv definit, da $(a^2+b^2)(c^2+d^2) - (ac+bd)^2 = \boldsymbol{\Delta^2}$. Also ist das Bild des Einheitskreises $x^2+y^2 = 1$ eine Ellipse um den Nullpunkt. Mit $\boldsymbol{q} := a^2+b^2+c^2+d^2,\ \boldsymbol{q_-} := a^2+b^2-c^2-d^2$ und $\boldsymbol{q_\delta} := \sqrt{((a-d)^2+(b+c)^2)((a+d)^2+(b-c)^2)}$ gilt

$$\boldsymbol{q^2 - 4\Delta^2} \;=\; \boldsymbol{q_-^2 + 4(ac+bd)^2} \;=\; \boldsymbol{q_\delta^2},$$

und die *Eigenwerte* $\lambda_{1,2}$ der Form-Matrix, d.h. die Wurzeln von $\lambda^2 - q\lambda + \Delta^2 = 0$, ergeben $\boxed{\boldsymbol{\sigma_1 = \frac{|\Delta|}{\sqrt{\lambda_2}} = \sqrt{\lambda_1},\ \sigma_2 = \frac{|\Delta|}{\sqrt{\lambda_1}} = \sqrt{\lambda_2}}}$

$\lambda_{1,2} = \frac{1}{2}\big(q \pm \sqrt{q^2-4\Delta^2}\big) = \frac{1}{2}\big(q \pm \sqrt{q_-^2 + 4(ac+bd)^2}\big) = \frac{1}{2}(q \pm q_\delta)$.

Klar: $q_\delta = 0 \;\Leftrightarrow\; (a = d \wedge b = -c) \vee (a = -d \wedge b = c)$

(Anschaulich: $A\vec{e}_1 \perp A\vec{e}_2 \;\wedge\; |A\vec{e}_1| = |A\vec{e}_2|$.)

Nur in diesem Fall sind die Eigenwerte gleich, $\lambda_1 = \lambda_2 = a^2+b^2$; die Bildellipse ist ein *Kreis* mit Radius $\sqrt{a^2+b^2} = \sigma_1 = \sigma_2$; die Einheitsvektoren $\vec{u}_1 \perp \vec{u}_2$ sind beliebig wählbar, da aus $\vec{x}_1 \perp \vec{x}_2$ stets $A\vec{x}_1 \perp A\vec{x}_2$ folgt ($\frac{1}{\sqrt{a^2+b^2}}A$ ist *orthogonal*).

Nun der Fall $q_\delta > 0$: $\vec{v} := (\frac{1}{2}(q_\delta + q_-), ac+bd)^T$, im Sonderfall $ac+bd = 0 \;\wedge\; q_- < 0$ stattdessen $\vec{v} := (0, 2q_\delta)^T$, ist unnormierter Eigenvektor der Form-Matrix zu $\lambda_2 = \frac{1}{2}(q - q_\delta)$; es ergibt sich $\boxed{\boldsymbol{\vec{u}_1 = A^{-1}\frac{\sigma_1}{|\vec{v}|}\vec{v}}}$ mit $|\vec{v}|^2 = \frac{1}{2}q_\delta(q_\delta + q_-)$ bzw. $|\vec{v}|^2 = 4q_\delta^2$, und $\boldsymbol{\vec{u}_2}$ folgt per Orthogonalität.

Die Tatsache, dass Kreise auf Ellipsen abgebildet werden, führte durch Diskussion umbeschriebener Quadrate ganz anschaulich zur Existenz orthogonaler „Hauptachsen“.

Hier nun eine vom Bisherigen unabhängige ganz einfache *exakte* Begründung.

Der *wesentliche* analytische Aspekt:

$$|\vec{u}+t\,\vec{v}| - |\vec{u}| = \frac{|\vec{u}+t\,\vec{v}|^2 - |\vec{u}|^2}{|\vec{u}+t\,\vec{v}| + |\vec{u}|} \approx \begin{cases} t\,\frac{(\vec{v},\vec{u})}{|\vec{u}|}, & \vec{v} \not\perp \vec{u}, \\ t^2\,\frac{|\vec{v}|^2}{2|\vec{u}|}, & \vec{v} \perp \vec{u} \end{cases}$$

für $\vec{u} \neq \vec{0}$ und kleine $|t|$ ($|\vec{x}| := \|\vec{x}\|_2$). Daher:

$$|\boldsymbol{\vec{u}_1}| = 1,\ |\boldsymbol{A\vec{u}_1}| = \max_{|\boldsymbol{\vec{u}}|=1} |\boldsymbol{A\vec{u}}|,\ \boldsymbol{\vec{u}_2 \perp \vec{u}_1}$$

$$\Rightarrow \quad \boldsymbol{A\vec{u}_2 \perp A\vec{u}_1}.$$

Kurz: $\boldsymbol{A(\vec{u}_1^{\,\perp}) \subseteq (A\vec{u}_1)^\perp}$. Anderenfalls wäre $\frac{|A\vec{u}_1 + t\,A\vec{u}_2|}{|\vec{u}_1 + t\,\vec{u}_2|} \approx \frac{|A\vec{u}_1| + t\frac{(A\vec{u}_2, A\vec{u}_1)}{|A\vec{u}_1|}}{|\vec{u}_1| + t^2\frac{|\vec{u}_2|^2}{2|\vec{u}_1|}} > \frac{|A\vec{u}_1|}{|\vec{u}_1|}$ für kleine $|t|$ mit $t\,(A\vec{u}_2, A\vec{u}_1) > 0$. (Im *komplexen* Fall $t\,\mathrm{e}^{\mathrm{i}\varphi}$ statt t mit $\varphi = \arg(A\vec{u}_1, A\vec{u}_2)$.)

Gilt außerdem $|\vec{u}_2| = 1$, so durchläuft $\cos t \cdot \vec{u}_1 + \sin t \cdot \vec{u}_2$ den Einheitskreis und $\cos t \cdot A\vec{u}_1 + \sin t \cdot A\vec{u}_2$ eine Ellipse mit den Haupt(halb)achsen $A\vec{u}_1$ und $A\vec{u}_2$.

Nebenbei *folgt*: $\cos t \cdot \vec{a} + \sin t \cdot \vec{b}$ durchläuft auch im Falle $\vec{a} \not\perp \vec{b}$ eine Ellipse: $A = (\vec{a}, \vec{b})$,...

2. Singulärwerte und Singulärvektoren einer Matrix

Sei $A \in \mathbb{R}^{m\times n}$ gegeben.
Dann gibt es mindestens ein $\vec{u}_1 \in \mathbb{R}^n$ mit $|\vec{u}_1| = 1$ und

$$\boldsymbol{|A\vec{u}_1| = \max_{|\vec{u}|=1} |A\vec{u}|\,.}$$

($\vec{u}_1$ ist absolute Maximumsstelle der stetigen Funktion $\vec{u} \mapsto |A\vec{u}|$ auf der kompakten Menge $|\vec{u}| = 1$.)

Wir setzen

$$\boldsymbol{\sigma_1 := |A\vec{u}_1|,\ \vec{v}_1 := \frac{1}{\sigma_1} A\vec{u}_1, \quad \text{falls } \sigma_1 > 0.}$$

Insbesondere natürlich $\vec{v}_1 \in \mathbb{R}^m$.

Falls $\sigma_1 = 0$, ist A die $m \times n$-Nullmatrix, ein Trivialfall. Sei also $\sigma_1 > 0$ angenommen.

Wir betrachten nun – im Falle $n \geq 2$ – irgendein $\vec{u} \in \mathbb{R}^n$ mit

$$\boldsymbol{\vec{u} \perp \vec{u}_1, \qquad |\vec{u}| = 1.}$$

Bild des Kreises $\cos t \cdot \vec{u}_1 + \sin t \cdot \vec{u}$ $(0 \leq t \leq 2\pi)$ unter A ist die Ellipse

$$\cos t \cdot \sigma_1 \vec{v}_1 + \sin t \cdot A\vec{u} \quad (0 \leq t \leq 2\pi).$$

Nach Definition ist $\sigma_1 \vec{v}_1$ eine *große* Halbachse dieser Ellipse, und $A\vec{u}$, der zu $\sigma_1 \vec{v}_1$ *konjugierte* Halbmesser, dementsprechend eine *kleine* Halbachse dieser Ellipse: $A\vec{u} \perp A\vec{u}_1$. Auch im Entartungsfall $A\vec{u} = \vec{0}$ bleibt dies – auf triviale Weise – richtig. Kurz:

$$\boldsymbol{A(\vec{u}_1^{\perp}) \subseteq (A\vec{u}_1)^{\perp}.}$$

(Andere Begründung auf S. 8.) Wir wählen nun ein $\vec{u}_2 \in \mathbb{R}^n$ mit $\vec{u}_2 \perp \vec{u}_1$, $|\vec{u}_2| = 1$ und

$$\boldsymbol{|A\vec{u}_2| = \max_{|\vec{u}|=1, \vec{u}\perp\vec{u}_1} |A\vec{u}|}$$

und setzen

$$\boldsymbol{\sigma_2 := |A\vec{u}_2|,\ \vec{v}_2 := \frac{1}{\sigma_2} A\vec{u}_2, \quad \text{falls } \sigma_2 > 0.}$$

Insbesondere $\vec{v}_2 \perp \vec{v}_1$. Nun $\vec{u}_3 \perp \vec{u}_1, \vec{u}_2$, usw.

Schrittweise erhalten wir Singulärwerte

$$\boldsymbol{\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_r > 0}$$

von A und zugehörige Singulärvektoren

$$\boldsymbol{\vec{u}_1, \ldots, \vec{u}_r \in \mathbb{R}^n,\ \vec{v}_1, \ldots, \vec{v}_r \in \mathbb{R}^m,}$$

bis $A\vec{u} = \vec{0}$ für $\vec{u} \perp \vec{u}_1, \ldots, \vec{u}_r$ eintritt.

Dabei offenbar $r \leq \min(m,n)$. Man kann nun die Singulärvektoren durch Vektoren $\vec{u}_{r+1}, \ldots, \vec{u}_n$ sowie $\vec{v}_{r+1}, \ldots, \vec{v}_m$ zu Orthonormalbasen des $\mathbb{R}^n$ bzw. des $\mathbb{R}^m$ vervollständigen.

Insgesamt ergibt sich

$$\boldsymbol{A(x_1\,\vec{u}_1 + \cdots + x_n\,\vec{u}_n) = x_1\,\sigma_1\,\vec{v}_1 + \cdots + x_r\,\sigma_r\,\vec{v}_r}$$

oder – dasselbe etwas anders formuliert –

$$\boldsymbol{A\vec{x} = (\vec{x}, \vec{u}_1)\,\sigma_1\,\vec{v}_1 + \cdots + (\vec{x}, \vec{u}_r)\,\sigma_r\,\vec{v}_r}$$

für beliebige $\vec{x} = x_1\,\vec{u}_1 + \cdots + x_n\,\vec{u}_n \in \mathbb{R}^n$.

Und in Matrizen-Schreibweise:

$$\boldsymbol{A\vec{x} = (\vec{v}_1, \ldots, \vec{v}_r) \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix} (\vec{u}_1, \ldots, \vec{u}_r)^T\,\vec{x}}$$

oder auch $AU = V\Sigma$ und

$$\boldsymbol{A = V\,\Sigma\,U^T = \sigma_1\,\vec{v}_1\vec{u}_1^T + \cdots + \sigma_r\,\vec{v}_r\vec{u}_r^T}$$

mit $U = (\vec{u}_1, \ldots, \vec{u}_n) \in \mathbb{R}^{n\times n}$, $V = (\vec{v}_1, \ldots, \vec{v}_m)$ $\in \mathbb{R}^{m\times m}$, $\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ & & \sigma_r \\ 0 & & \end{pmatrix} \in \mathbb{R}^{m\times n}$.

Voilà die Singulärwertzerlegung von A !

Fazit: Eine lineare Abbildung ist *immer* eine Rotation, der eine Achs-Skalierung und eine weitere Rotation und/oder Spiegelung folgt. So wird die lineare Abbildung *faktorisiert* mittels einfacher Abbildungs-Grundtypen.

3. Einige unmittelbare Folgerungen

Da $A\vec{x}$ eine Linearkombination der Spalten von A ist, ist klar:
Die Dimension des Bildraumes $\mathcal{R}(A)$, des von den Spalten von A aufgespannten Raumes, ist r, und $\{\vec{v}_1, \ldots, \vec{v}_r\}$ ist eine Orthonormalbasis des Bildraumes. Ebenso ist $\{\vec{u}_{r+1}, \ldots, \vec{u}_n\}$ eine ONB des Nullraumes (Kerns) $\mathcal{N}(A)$.

Kern, Bild und Dimensionsformel

$$\boldsymbol{\mathcal{N}(A) = \ll \vec{u}_{r+1}, \ldots, \vec{u}_n \gg,}$$
$$\boldsymbol{\mathcal{R}(A) = \ll \vec{v}_1, \ldots, \vec{v}_r \gg,}$$
$$\boldsymbol{n = \dim \mathcal{R}(A) + \dim \mathcal{N}(A).}$$

Durch Transponieren ergibt sich ferner, dass $\{\vec{u}_1, \ldots, \vec{u}_r\}$ eine ONB des von den Zeilen von A aufgespannten Raumes ist; insbesondere:
Zeilenrang = Spaltenrang $= r = \dim \mathcal{R}(A)$;
$n = \dim \mathcal{R}(A) + \dim \mathcal{N}(A)$ ist der *Rang-Satz*.

Die Vertauschung der Rollen durch Transponieren kann man auch so ausdrücken:

$$\boldsymbol{\mathcal{N}(A^T) = \mathcal{R}(A)^\perp = \ll \vec{v}_{r+1}, \ldots, \vec{v}_m \gg,}$$
$$\boldsymbol{\mathcal{R}(A^T) = \mathcal{N}(A)^\perp = \ll \vec{u}_1, \ldots, \vec{u}_r \gg .}$$

4. Algebraisch-analytischer Beweis

Wir formulieren den allgemeinen SVD-Sachverhalt und geben einen von bisherigen Überlegungen unabhängigen Beweis.

Satz Singulärwertzerlegung

Zu jeder Matrix $A \in \mathbb{K}^{m\times n}$ gibt es ONB-Matrizen $U \in \mathbb{K}^{n\times n}$ und $V \in \mathbb{K}^{m\times m}$ sowie eine Diagonalmatrix $\Sigma \in \mathbb{R}^{m\times n}$, so dass

$$\boldsymbol{A = V\,\Sigma U^* = V} \begin{pmatrix} \boldsymbol{\sigma_1} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{\sigma_r} \end{pmatrix} \boldsymbol{U^*} = \sum_{i=1}^{r} \boldsymbol{\sigma_i \vec{v}_i \vec{u}_i^*}$$

mit den Singulärwerten

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \ (= \sigma_{r+1} = \cdots = \sigma_{\min(m,n)})$$

und $r = \mathrm{rg}(A)$ (*Rang* der Matrix).
Bei mehrfachen Singulärwerten sind nicht die ihnen gemäß $A\vec{u}_i = \sigma_i\,\vec{v}_i$ entsprechenden einzelnen U- und V-Spalten, wohl aber die jeweils von diesen aufgespannten Teilräume des $\mathbb{K}^n$ bzw. $\mathbb{K}^m$ *eindeutig bestimmt*.

Dabei gilt $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, „ONB-Matrizen“ bezeichnet *orthogonale* bzw. *unitäre* Matrizen.

Charakterisiert sind ONB-Matrizen U durch $\boldsymbol{U^{-1} = U^* := \overline{U}^T}$ (die adjungierte Matrix, im reellen Falle einfach die *transponierte*). Äquivalent dazu:

U unitär/orthogonal $\Leftrightarrow$ $|U\vec{x}| = |\vec{x}|$ $(\vec{x} \in \mathbb{K}^n)$.

Beweis des Satzes:
Gegeben $A \in \mathbb{K}^{m\times n}$, sei $\vec{u}_1 \in \mathbb{K}^n$ so gewählt, dass

$$|\vec{u}_1| = 1, |A\vec{u}_1| = \max_{|\vec{u}|=1} |A\vec{u}| =: \sigma_1.$$

Wir setzen $\vec{v}_1 := \frac{1}{\sigma_1} A\vec{u}_1$. Gilt $\vec{u} \perp \vec{u}_1$ sowie $|\vec{u}| = 1$, folgt mit beliebigem Skalar α

$$\begin{aligned} \sigma_1^2(1+|\alpha|^2) &\geq |A(\vec{u}_1 + \alpha\,\vec{u})|^2 \\ &= \sigma_1^2 + |\alpha|^2|A\vec{u}|^2 + 2\sigma_1 \Re\,\bar{\alpha}(\vec{v}_1, A\vec{u}), \end{aligned}$$

und letzteres ist $\geq \sigma_1^2(1+2|\alpha|^2)$ für $\sigma_1\alpha = (\vec{v}_1, A\vec{u})$. Also $(\vec{v}_1, A\vec{u}) = 0$, $A(\vec{u}_1^{\perp}) \subseteq (A\vec{u}_1)^{\perp}$. Per Induktion über n folgt die Behauptung.

Zur Eindeutigkeit: Gilt $\sigma_1 = \cdots = \sigma_k > \sigma_{k+1}$, spannen $\vec{u}_1, \ldots, \vec{u}_k$ genau den Unterraum aller $\vec{x} \in \mathbb{K}^n$ mit $|A\vec{x}| = \sigma_1|\vec{x}|$ auf – *dass* es ein Unterraum ist, zeigt der Beweis –, und $\vec{v}_1, \ldots, \vec{v}_k$ spannen sein Bild unter A auf. ■

5. Einige Anwendungen der SVD

1) Eine robuste, wenn auch aufwendige Methode zur *numerischen Rangbestimmung*, evtl. auch zur Lösung eines Gleichungssystems, im singulären oder nahezu singulären Fall ist die Berechnung der SVD.

2) Die eindeutig bestimmte Lösung des linearen Ausgleichsproblems

$$|\boldsymbol{A}\vec{\boldsymbol{x}} - \vec{\boldsymbol{b}}| \stackrel{!}{=} \mathbf{min} \quad \text{und dabei } |\vec{x}| \stackrel{!}{=} \min$$

mit beliebiger Matrix $A \in \mathbb{K}^{m\times n}$ und $\vec{b} \in \mathbb{K}^m$ definiert die Pseudo-Inverse A^+ von A: Die Lösung des Problems ist genau $\vec{\boldsymbol{x}} = \boldsymbol{A}^+\vec{\boldsymbol{b}}$.

Man kann die Pseudo-Inverse charakterisieren durch gewisse Kalkül-Eigenschaften, z.B. die *Moore/Penrose-Bedingungen* $AXA = A$, $XAX = X, (AX)^* = AX, (XA)^* = XA$; *genau ein* X erfüllt sie: $X = A^+$.

Mit der Singulärwertzerlegung lässt sich die Pseudo-Inverse einer Matrix A ganz einfach explizit ausdrücken:

$$A = V\Sigma U^* \quad \Rightarrow \quad A^+ = U\Sigma^+ V^*$$

mit

$$\Sigma^+ = \begin{pmatrix} \sigma_1^{-1} & & 0 \\ & \ddots & \\ & & \sigma_r^{-1} \\ 0 & & \end{pmatrix}.$$

Drückt man $|\boldsymbol{A}\vec{\boldsymbol{x}} - \vec{\boldsymbol{b}}|$ durch die SVD aus, ergibt sich sofort diese (eindeutige!) Darstellung von A^+:

$$\begin{aligned} |A\vec{x} - \vec{b}|^2 &= |\sigma_1(\vec{u}_1 \cdot \vec{x})\vec{v}_1 + \cdots + \sigma_r(\vec{u}_r \cdot \vec{x})\vec{v}_r \\ &\qquad - (\vec{v}_1 \cdot \vec{b})\vec{v}_1 - \cdots - (\vec{v}_m \cdot \vec{b})\vec{v}_m|^2 \\ &= \sum_{i=1}^{r} \left|\sigma_i\, \vec{u}_i \cdot \vec{x} - \vec{v}_i \cdot \vec{b}\right|^2 + \sum_{i=r+1}^{m} \left|\vec{v}_i \cdot \vec{b}\right|^2; \end{aligned}$$

dies wird offenbar minimal für alle $\vec{x}$ mit

$$\vec{u}_i \cdot \vec{x} = (\vec{v}_i \cdot \vec{b})/\sigma_i \ (1 \le i \le r).$$

3) Man kann die Singulärwertzerlegung zur *Datenkompression* verwenden: Ist A beispielsweise eine Graustufen-Matrix, so kann man – analog zur DFT – die SVD von A bilden und alle hinreichend kleinen Singulärwerte verwerfen und dann mit einem r_0 deutlich kleiner als r nur $\vec{u}_i, \vec{v}_i, \sigma_i$ $(1 \le i \le r_0)$ als Bilddaten verwenden und wieder zurücktransformieren.

Im ergänzenden Skriptum *Singulärwertzerlegung mit SCILAB* (28 Seiten) wird die Bilddatenkompression anhand mehrerer Beispiele detailliert demonstriert; siehe auch S. 47.

4) Auch im praktisch bedeutsamen Bereich des automatisierten *Information Retrieval (IR)* wird die SVD benutzt: Die Inhalte eines Dokumenten-Bestands werden durch eine große „Term-Dokument-Matrix“ repräsentiert. Wie die Dokumente werden auch Anfragen durch Begriffs-Vektoren dargestellt und mit einer per SVD rangreduzierten Matrix (Stichwort: *Latent Semantic Indexing*, LSI) durch Skalarprodukt-Bildung abgeglichen. Effiziente *Update*-Techniken sind dabei sehr wichtig, da Datenbestände sich ständig ändern.

Der Artikel *Using linear algebra for intelligent information retrieval* von Michael W. Berry, Susan T. Dumais und Gavin W. O'Brian, SIAM Review 37 (1995), pp. 573-595, gibt einen guten Überblick. (Im selben Review-Heft auch ein fundamentaler Artikel von Chu, Funderlic und Golub über Rang-1-Reduktionen und Matrix-Faktorisierungen, relevant für das Verständnis der SVD.)

Zur Rang-Reduktion: Mit $U_k := (\vec{u}_1, \ldots, \vec{u}_k)$, $V_k := (\vec{v}_1, \ldots, \vec{v}_k)$ ist $A_k := V_k \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_k \end{pmatrix} U_k^T$ optimale Rang-k-Approximation an A. Denn $\|A - A_k\|_F \leq \|A - B\|_F$ für alle anderen Matrizen B mit Rang k. Dabei (Frobenius-Norm)

$$\|A\|_F^2 := \sum |a_{i\,k}|^2 = \text{Spur}\, A^T A = \sum \sigma_i^2.$$

Weiter unten wird die Rang-k-Approximation noch genauer diskutiert.

5) *Polarzerlegung* quadratischer Matrizen.

Ist $A \in \mathbb{R}^{n\times n}$, ergibt die Singulärwertzerlegung $\boldsymbol{A = V\Sigma U^T = (V\Sigma V^T)(VU^T) =: PQ}$. Dabei ist $P = V\Sigma V^T$ offenbar positiv semidefinit und $Q = VU^T$ orthogonal. Diese Zerlegung gleicht der Polarform $z = r\, e^{i\,\varphi}$ komplexer Zahlen und wird deshalb „Polarzerlegung“ genannt. (Da $AA^T = P^2$, ist P, da positiv semidefinit, *eindeutig bestimmt*; dies zeigt der erste Beweis in Abschnitt 7.)

Variante: $\boldsymbol{A = (VU^T)(U\Sigma U^T) = Q\tilde{P}}$.

6) *Rechenpraktisch* ist die SVD viel brauchbarer als die *Jordansche Normalform*. Letztere ist höchstens im Falle der Diagonalisierbarkeit nicht *sehr schlecht konditioniert*, während die SVD – wenn auch so gut wie nie von Hand ausrechenbar – mit robusten Verfahren bestimmt werden kann, z.B. mit dem Golub/Reinsch-Algorithmus.

Beschrieben ist dieser in **J.Stoer/R.Bulirsch**, *Numerische Mathematik 2*, Springer; C-Programme findet man bei **Press, Teukolsky, et al.**, *Numerical Recipes, 3rd Ed.*, Cambridge UP.

Die Jordan-Form fußt auf der den Eigenwert-Begriff verallgemeinernden Unterraum-Invarianz und kommt *ohne metrische Strukturen* aus, zerlegt aber nur *quadratische* Matrizen.

6. Singulärwerte und Eigenwerte

Nur quadratische Matrizen haben *Eigenwerte*; dennoch haben die Singulärwerte von A etwas mit Eigenwerten zu tun: Aus $A = V\Sigma U^*$ folgt $A^*A = U\Sigma^*V^*\,V\Sigma U^* = U\Sigma^*\Sigma U^*$ mit

$$\Sigma^*\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ & & \sigma_r^2 \\ 0 & & \end{pmatrix} \in \mathbb{K}^{n\times n}.$$

Also $A^*A\,\vec{u}_i = \sigma_i^2\vec{u}_i$ ($\vec{u}_i$ die Spalten von U); d.h.:

Die Singulärwerte von A sind die Quadratwurzeln der Eigenwerte von A^*A.

Analog kann man auch $A\,A^*$ betrachten, wobei dann die $\vec{v}_k$ anstelle der $\vec{u}_i$ die Rolle der Eigenvektoren übernehmen.

Nebenbei halten wir fest:

$$\sigma_1 = \|A\|_2 = \max_{|\vec{x}|=1} |A\vec{x}|, \quad \sigma_1^2 = \max_{|\vec{x}|=1} (A^*A\vec{x},\vec{x})\,.$$

Die erste Herleitung der SVD (E. Beltrami 1873) ging aus von der heuristischen Überlegung, dass bei der Gleichung $S = U^TAV$ mit gegebener Matrix $A \in \mathbb{R}^{n\times n}$ durch n^2+n Bedingungen U und V als orthogonal festgelegt werden ($\vec{u}_i\cdot\vec{u}_k = \vec{v}_i\cdot\vec{v}_k = \delta_{ik}$ $(i\le k)$) und durch n^2-n weitere Bedingungen S zu einer Diagonalmatrix wird, genau entsprechend den $2n^2$ Parametern, den Elementen von U,V. *Ausgehend* von einer solchen Darstellung $S = U^TAV$ ergibt sich, dass die s_{ii}^2 die Eigenwerte sowohl von AA^T als auch von A^TA sind, mit den Eigenvektoren $\vec{u}_i$ bzw. $\vec{v}_i$; der Schlüssel zur Berechnung der Faktorisierung.

Auch die symmetrische $(m+n)\times(m+n)$-Matrix

$$\boldsymbol{B} := \begin{pmatrix} \boldsymbol{0} & \boldsymbol{A^*} \\ \boldsymbol{A} & \boldsymbol{0} \end{pmatrix}$$

hat eine einfache Beziehung zu den Singulärwerten und Singulärvektoren von $A = V\Sigma U^*$. Mit $AU = V\Sigma$ und $A^*V = U\Sigma$ folgt

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}\begin{pmatrix} U & U \\ -V & V \end{pmatrix} = \begin{pmatrix} -U\Sigma & U\Sigma \\ V\Sigma & V\Sigma \end{pmatrix} \quad !!$$

- *beinahe* jedenfalls. Wir müssen die Eigenvektor-„Matrix“ $\begin{pmatrix} U & U \\ -V & V \end{pmatrix}$ noch ein wenig korrigieren: Sei r der Rang von A; dann hat B den Rang $2r$, und es gilt für $1 \le i \le r$

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}\begin{pmatrix} \vec{u}_i \\ -\vec{v}_i \end{pmatrix} = -\sigma_i\begin{pmatrix} \vec{u}_i \\ -\vec{v}_i \end{pmatrix}, \quad \begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}\begin{pmatrix} \vec{u}_i \\ \vec{v}_i \end{pmatrix} = \sigma_i\begin{pmatrix} \vec{u}_i \\ \vec{v}_i \end{pmatrix}.$$

Daraus folgt: Es gibt $2r$ untereinander orthogonale Eigenvektoren zu den Eigenwerten $\pm\sigma_i$ $(1 \le i \le r)$ von B, und diese können – nach Normierung mit Faktor $1/\sqrt{2}$ – durch $m+n-2r$ Vektoren zu einer ONB von $\mathbb{R}^{m+n}$ bzw. $\mathbb{C}^{m+n}$ ergänzt werden, etwa durch die Vektoren (eine ONB des Nullraums von B)

$$\begin{pmatrix} \vec{u}_{r+1} \\ \vec{0} \end{pmatrix}, \ldots, \begin{pmatrix} \vec{u}_n \\ \vec{0} \end{pmatrix} \text{ und } \begin{pmatrix} \vec{0} \\ \vec{v}_{r+1} \end{pmatrix}, \ldots, \begin{pmatrix} \vec{0} \\ \vec{v}_m \end{pmatrix}.$$

7. Äquivalenz zum Spektralsatz

Der Spektralsatz für symmetrische Matrizen ist „äquivalent“ zur Singulärwertzerlegung in dem Sinne, dass beide auseinander durch relativ kurze elementaralgebraische Überlegungen folgen. Man braucht keine analytischen Hilfsmittel: keine Existenz von Maxima stetiger Funktionen auf kompakten Mengen oder von Eigenwerten (= Polynom-Nullstellen).

Satz Reeller Spektralsatz

Zu einer symmetrischen Matrix A gibt es stets eine *orthogonale* Matrix Q, deren Spalten Eigenvektoren zu den – sämtlich *reellen* – Eigenwerten von A sind, so dass also mit $Q = (\vec{b}_1, \ldots, \vec{b}_n)$ gilt $AQ = (\lambda_1 \vec{b}_1, \ldots, \lambda_n \vec{b}_n)$ und daher

$$A = Q \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} Q^T .$$

Geometrisch: Bezogen auf ein geeignet verdrehtes rechtwinkliges Koordinatensystem entspricht einer symmetrischen linearen Abbildung eine reelle Diagonalmatrix.

Aus der SVD folgt der Spektralsatz:

Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und o.B.d.A. positiv definit. (Durch Übergang von A zu $A+\lambda E$ folgt der Spektralsatz für beliebige symmetrische Matrizen aus dem für positiv definite.)

Es gibt eine SVD $A = V\Sigma U^T$, also, da A symmetrisch, auch $A = U\Sigma V^T$ und daher $A^2 = U\Sigma^2 U^T$. D.h.: Die Spalten $\vec{u}_1, \ldots, \vec{u}_n$, eine ONB des $\mathbb{R}^n$, sind Eigenvektoren von A^2 zu den Eigenwerten $\sigma_1^2, \ldots, \sigma_n^2$.

Es gilt auch $A\vec{u}_k = \sigma_k \vec{u}_k$ $(1 \le k \le n)$ (nur für $\sigma_k = 0$ von vornherein klar), da

$$\begin{aligned} &\big(A(A\vec{u}_k - \sigma_k\vec{u}_k), A\vec{u}_k - \sigma_k\vec{u}_k\big) \\ =& \big(\sigma_k^2\vec{u}_k - \sigma_k A\vec{u}_k, A\vec{u}_k - \sigma_k\vec{u}_k\big) \\ =& -\sigma_k\,|A\vec{u}_k - \sigma_k\vec{u}_k|^2 \ge 0 \end{aligned}$$

wegen positiver Definitheit von A. ■

Aus dem Spektralsatz folgt die SVD:

Sei $A \in \mathbb{R}^{m\times n}$. Dann ist A^TA symmetrisch, und gemäß Spektralsatz gibt es eine ONB $\{\vec{u}_1, \ldots, \vec{u}_n\}$ des $\mathbb{R}^n$, so dass $A^TA\vec{u}_k = \lambda_k\vec{u}_k$ für $1 \le k \le n$ mit $\lambda_1, \ldots, \lambda_n \ge 0$, da $(A^TA\vec{x}, \vec{x})$ $= |A\vec{x}|^2 \ge 0$ für alle $\vec{x}$.

O.B.d.A. nehmen wir $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ an und setzen $A\vec{u}_k =: \sigma_k \vec{v}_k$ mit $|\vec{v}_k| = 1$, $\sigma_k > 0$ für alle k mit $\lambda_k > 0$. Je zwei solcher Vektoren $\vec{v}_k$ und $\vec{v}_l$ sind orthogonal zueinander, da

$$\sigma_k \, \sigma_l \, (\vec{v}_k, \vec{v}_l) = (A\vec{u}_k, A\vec{u}_l) = (A^T A \vec{u}_k, \vec{u}_l) = 0$$

für $k \neq l$. Wir ergänzen die $\vec{v}_k$ (falls es nicht schon m sind) zu einer Basis des $\mathbb{R}^m$. Damit ergibt sich, mit $V = (\vec{v}_1, \ldots, \vec{v}_m)$, die SVD $AU = V\Sigma$. Insbesondere $\lambda_k = \sigma_k^2$. ■

Direkter Beweis des Spektralsatzes:

Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch.
Die stetige Funktion $\boldsymbol{q : \vec{x} \mapsto (\vec{x}, A\,\vec{x})}$ $(\vec{x} \in \mathbb{R}^n)$ nimmt auf der Menge $|\vec{x}| = 1$ irgendwo, etwa an der Stelle $\vec{x}_1$, ihr absolutes Maximum an. Wir setzen $\boldsymbol{\lambda_1 := q(\vec{x}_1)}$ und wählen ein $\vec{x} \perp \vec{x}_1$ mit $|\vec{x}| = 1$, so dass gemäß Definition von $\vec{x}_1$

$$|\vec{x}_1 + t\,\vec{x}|^2 = 1 + t^2, \;\; q(\vec{x}_1 + t\,\vec{x}) \leq (1 + t^2)\lambda_1,$$

andererseits (beachte $(x_1, A\vec{x}) = (\vec{x}, A\vec{x}_1)$)

$$q(\vec{x}_1 + t\,\vec{x}) = \lambda_1 + 2t\,(\vec{x}_1, A\vec{x}) + t^2 q(\vec{x}).$$

Aus $2t\,(\vec{x}_1, A\vec{x}) + t^2 q(\vec{x}) \leq t^2\,\lambda_1$ für alle $t \in \mathbb{R}$ folgt $(x_1, A\vec{x}) = (\vec{x}, A\vec{x}_1) = 0$, also $A\vec{x} \perp \vec{x}_1$, $A\vec{x}_1 \perp \vec{x}$. D.h.: Der Unterraum $\vec{x}_1^{\perp} := \{\, \vec{x} \in \mathbb{R}^n \mid \vec{x} \perp \vec{x}_1 \,\}$ ist „invariant unter A“, und $A\vec{x}_1$, da senkrecht auf diesem, ist Vielfaches von $\vec{x}_1$: $A\vec{x}_1 = c\vec{x}_1$. Sofort klar: $c = \lambda_1$; m.a.W.: $\vec{x}_1$ ist *Eigenvektor* zum *Eigenwert* λ_1 von A. Per Induktion folgt der Spektralsatz. ■

Im *komplexen* Fall verlaufen die Beweise völlig analog.

8. Einige Singulärwert-Ungleichungen

Unmittelbar aus der schrittweisen Herleitung der SVD folgt per Induktion über k die wichtige Minimax-Eigenschaft der Singulärwerte:

$$\begin{aligned}
\boldsymbol{\sigma_k} &= \min_{\vec{x}_1,\ldots,\vec{x}_{k-1}\in\mathbb{R}^n} \; \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{k-1}}} |\boldsymbol{A\vec{x}}| \\
&= \max_{\vec{x}_1,\ldots,\vec{x}_{n-k}\in\mathbb{R}^n} \; \min_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{n-k}}} |\boldsymbol{A\vec{x}}| \\
&= \max_{\boldsymbol{\dim(V)=k}} \; \min_{\boldsymbol{\vec{x}\in V,\ |\vec{x}|=1}} |\boldsymbol{A\vec{x}}| \quad \boldsymbol{(1 \leq k \leq n).}
\end{aligned}$$

Dabei ist mit V ein Unterraum des $\mathbb{R}^n$ gemeint. Im komplexen Fall gilt alles ganz genauso.
Im folgenden bezeichnet, absteigend gezählt, $\sigma_i(A)$ den i-ten Singulärwert von A.

Satz Singulärwert-Abschätzungen
Für $A, B \in \mathbb{C}^{m\times n}$ gilt

$$\sigma_{i+k+1}(A + B) \leq \sigma_{i+1}(A) + \sigma_{k+1}(B)$$

sowie

$$\boldsymbol{|\sigma_i(A) - \sigma_i(B)| \leq \|A - B\|_2.}$$

Beweis:

$$\sigma_{i+k+1}(A+B) = \min_{\vec{x}_1,\ldots,\vec{x}_{i+k}\in\mathbb{C}^n} \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i+k}}} |(A+B)\vec{x}|$$

$$\leq \min_{\vec{x}_1,\ldots,\vec{x}_{i+k}\in\mathbb{C}^n} \left(\max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i+k}}} |A\vec{x}| + \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i+k}}} |B\vec{x}| \right)$$

$$\leq \min_{\vec{x}_1,\ldots,\vec{x}_{i+k}\in\mathbb{C}^n} \left(\max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i}}} |A\vec{x}| + \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_{i+1},\ldots,\vec{x}_{i+k}}} |B\vec{x}| \right)$$

$$= \min_{\vec{x}_1,\ldots,\vec{x}_{i}\in\mathbb{C}^n} \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i}}} |A\vec{x}| + \min_{\vec{x}_1,\ldots,\vec{x}_{k}\in\mathbb{C}^n} \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{k}}} |B\vec{x}|\,,$$

ferner mit dem soeben Bewiesenen

$\sigma_i(A) \leq \sigma_i(B) + \sigma_1(A-B) = \sigma_i(B) + \|A-B\|_2$

sowie

$\sigma_i(B) \leq \sigma_i(A) + \sigma_1(B-A) = \sigma_i(A) + \|B-A\|_2$.

■

Insbesondere die zweite Abschätzung zeigt, dass die Singulärwerte gut konditioniert sind. Eine analoge Produkt- Abschätzung:

Satz Produkt-Singulärwert-Abschätzungen

Für $A \in \mathbb{C}^{l\times m}$, $B \in \mathbb{C}^{m\times n}$ gilt

$$\sigma_{i+k+1}(AB) \leq \sigma_{i+1}(A)\,\sigma_{k+1}(B)$$

sowie

$$\boldsymbol{\sigma_i(AB) \leq \min\big(\sigma_i(A)\cdot\|B\|_2, \|A\|_2\cdot\sigma_i(B)\big).}$$

Beweis:

Seien $\vec{u}_1,\ldots,\vec{u}_m \in \mathbb{C}^m$ Singulärvektoren zu A.

$$\sigma_{i+k+1}(AB) = \min_{\vec{x}_1,\ldots,\vec{x}_{i+k}\in\mathbb{C}^n} \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i+k}}} |AB\vec{x}|$$

$$\leq \min_{\vec{x}_1,\ldots,\vec{x}_{k}\in\mathbb{C}^n} \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp B^*\vec{u}_1,\ldots,B^*\vec{u}_i \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{k}}} |AB\vec{x}|$$

$$\leq \min_{\vec{x}_1,\ldots,\vec{x}_{k}\in\mathbb{C}^n} \sigma_{i+1}(A) \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{k}}} |B\vec{x}|$$

$$= \sigma_{i+1}(A)\,\sigma_{k+1}(B)\,.$$

(Man beachte: $B\vec{x} \perp \vec{u}_i \Leftrightarrow \vec{x} \perp B^*\vec{u}_i$.)

Indem wir jeweils einen der Parameter i,k gleich 0 setzen, erhalten wir die zweite Abschätzung. ■

Bemerkenswert ist folgende Majorisierungs-Eigenschaft der Singulärwerte:

Satz Majorisierungsungleichungen

Ist $A = (a_{ij})$ eine $m\times n$-Matrix mit den Singulärwerten σ_i, gilt

$$|a_{11}| + \cdots + |a_{kk}| \leq \sigma_1 + \cdots + \sigma_k$$

für $1 \leq k \leq \min(m,n)$.

Beweis:
Sei $A = V\Sigma U^*$ Singulärwertzerlegung von A. Dann gilt mit $l := \min(m,n)$ insbesondere

$$\begin{pmatrix} a_{11} \\ \vdots \\ a_{ll} \end{pmatrix} = \begin{pmatrix} v_{11}u_{11} & \cdots & v_{1l}u_{1l} \\ \vdots & & \vdots \\ v_{l1}u_{l1} & \cdots & v_{ll}u_{ll} \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_l \end{pmatrix}$$

und daher für $1 \le k \le l$

$$|a_{11}|+\cdots+|a_{kk}| \le \Big(|v_{11}u_{11}|+\cdots+|v_{k1}u_{k1}|\Big)\,\sigma_1 + \cdots + \Big(|v_{1l}u_{1l}|+\cdots+|v_{kl}u_{kl}|\Big)\,\sigma_l\,.$$

Wegen Orthogonalität von U und V gelten nach Cauchy/Schwarz Beziehungen wie

$$|v_{1i}u_{1i}|+\cdots+|v_{ki}u_{ki}| \le \sqrt{\sum_j |v_{ji}|^2 \cdot \sum_j |u_{ji}|^2} = 1\,,$$

so dass die Koeffizienten von σ_1 bis σ_l (= die eingeklammerten Summen) sämtlich ≤ 1 sind und eine Gesamt-Summe $\le k$ besitzen. Also folgt wegen $\sigma_1 \ge \cdots \ge \sigma_l$, dass die rechte Seite $\le \sigma_1 + \cdots + \sigma_k$ ist. (Man kann die Koeffizienten von $\sigma_{k+1}, \sigma_{k+2}, \ldots$ „verbrauchen“, um die von $\sigma_1, \ldots, \sigma_k$ jeweils bis höchstens zum Wert 1 „aufzufüllen“.) ■

Im folgenden benutzen wir die Frobenius-Norm

$$\|\boldsymbol{A}\|_{\boldsymbol{F}} := \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} |\boldsymbol{a_{ik}}|^{\boldsymbol{2}}} = \sqrt{\mathbf{Spur}(\boldsymbol{A^*A})}\,.$$

Sie ist die *euklidische* Norm, angewandt auf den *Vektor*raum der Matrizen. Georg Frobenius, der auch den Rang-Begriff einführte, nannte diese Norm die „Spanne“ einer Matrix. Es ist *keine induzierte* Norm im Sinne von $\|A\| = \max \|A\vec{x}\|/\|\vec{x}\|$, dennoch eine *submultiplikative* Norm (Cauchy/Schwarz!):

$$\|AB\|_F \le \|A\|_F \cdot \|B\|_F\,;$$

ferner gilt

$$\|\boldsymbol{A}\|_{\boldsymbol{F}} = \sqrt{\boldsymbol{\sigma_1^2} + \cdots + \boldsymbol{\sigma_r^2}} \ge \|A\|_2 = \sigma_1.$$

Denn da $\|A\|_F^2$ Summe der Längenquadrate der *Spalten* wie der *Zeilen* von A ist, folgt

$$\|VA\|_F = \|AU\|_F = \|A\|_F$$

für orthogonale bzw. unitäre U und V und daher $\|A\|_F = \|V\Sigma U^*\|_F = \|\Sigma\|_F$.

Nun beweisen wir die *stetige Abhängigkeit* der Singulärwerte von den Matrix-Elementen, eine „Wielandt/Hoffman“-Ungleichung:

Satz Globaler Singulärwert-Vergleich

Für $A, B \in \mathbb{C}^{m\times n}$ gilt

$$\sum_{i=1}^{\min(m,n)} \Big(\sigma_i(A) - \sigma_i(B)\Big)^2 \le \|A-B\|_F^2.$$

Beweis:
Für zwei $m \times n$-Matrizen A und B gilt

$$\|A - B\|_F = \|U(A - B)V\|_F$$

mit beliebigen unitären Matrizen $U \in \mathbb{C}^{m\times m}$ und $V \in \mathbb{C}^{n\times n}$. U und V so gewählt, dass UBV die Diagonalmatrix der SVD von B ergibt, setzen wir $\tilde{A} = (\tilde{a}_{ik}) := UAV$, $\sigma_i := \sigma_i(A)$, $s_i := \sigma_i(B)$ und erhalten

$$\begin{aligned}\|A - B\|_F^2 &= \sum_{i\neq k} |\tilde{a}_{ik}|^2 + \sum_i |\tilde{a}_{ii} - s_i|^2 \\ &= \sum_{i,k} |\tilde{a}_{ik}|^2 + \sum_i s_i^2 - 2\sum_i \Re\tilde{a}_{ii}\, s_i \\ &\geq \sum_i \Big(\sigma_i - s_i\Big)^2,\end{aligned}$$

da $\sum_{i,k} |\tilde{a}_{ik}|^2 = \|\tilde{A}\|_F^2 = \sum_i \sigma_i^2(\tilde{A}) = \sum_i \sigma_i^2$ und, mit $l := \min(m, n)$, $A_j := \sum_{i=1}^{j} |\tilde{a}_{ii}|$,

$$\begin{aligned}\sum_{i=1}^{l} |\tilde{a}_{ii}|\, s_i &= \sum_{i=1}^{l} (A_i - A_{i-1})\, s_i \\ &= \sum_{i=1}^{l-1} A_i\,(s_i - s_{i+1}) + A_l\, s_l \\ &\leq \sum_{i=1}^{l-1} S_i\,(s_i - s_{i+1}) + S_l\, s_l = \sum_{i=1}^{l} \sigma_i\, s_i\end{aligned}$$

wegen $S_j := \sum_{i=1}^{j} \sigma_i \geq A_j$ (Majorisierung!). ■

Zum Vergleich formulieren wir den eigentlichen Satz von Wielandt/Hoffman, aus dem der soeben direkt bewiesene Satz auch gefolgert werden könnte, etwa durch Übergang von einer Matrix A zur hermiteschen Matrix $\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}$. (Am Ende dieses Abschnitts ein *sehr kurzer* Beweis des Wielandt/Hoffman-Satzes, der Anhang A benutzt.)

Satz Wielandt/Hoffman
Für zwei normale $n \times n$-Matrizen A, B mit den Eigenwerten λ_i bzw. μ_k gilt

$$\sum_{i=1}^{n} |\lambda_i - \mu_i|^2 \leq \|A - B\|_F^2$$

bei geeigneter Numerierung der Eigenwerte.

Beweis:
Seien U und V unitäre Matrizen, so dass

$$U^*AU = \Lambda, \quad V^*BV = M$$

mit Diagonalmatrizen Λ und M, deren Diagonalelemente die Eigenwerte λ_i und μ_i von A bzw. B sind. Wegen Unitärinvarianz der Frobeniusnorm folgt

$$\|A-B\|_F^2 = \|U\Lambda U^* - VMV^*\|_F^2 = \|\Lambda W - WM\|_F^2$$

mit $W = (w_{ij}) := U^*V$, also

$$\|A-B\|_F^2 = \sum_{i,j=1}^{n} |\lambda_i w_{ij} - w_{ij}\mu_j|^2 = \sum_{i,j=1}^{n} c_{ij}|\lambda_i - \mu_j|^2,$$

wobei $c_{ij} := |w_{ij}|^2$ und daher

$$0 \le c_{ij} \le 1,\ \sum_{i=1}^{n} c_{ij} = 1,\ \sum_{j=1}^{n} c_{ij} = 1\ (1 \le i,j \le n).\ \text{(DS)}$$

D.h.: $C = (c_{ij})$ ist eine „doppeltstochastische“ Matrix.

Gibt es ein echt zwischen 0 und 1 liegendes c_{ij}, kann man eine gerade Anzahl solcher c_{ij} auswählen, derart dass in jeder an der Auswahl beteiligten Zeile und Spalte genau *zwei* Elemente ausgewählt sind. (Begründung später, am Ende des Beweises.) Ist nun $S = (s_{ij})$ die Matrix, die in jeder Zeile an den Auswahlstellen die Werte 1 und -1, an allen anderen Stellen den Wert 0 annimmt, so gibt es $\varepsilon_1, \varepsilon_2 > 0$, so daß die Matrizen $C - \varepsilon_1 S$ und $C + \varepsilon_2 S$ weiterhin (DS) erfüllen, aber mindestens ein Element 0 oder 1 mehr besitzen als C. Gilt $\sum s_{ij}|\lambda_i - \mu_j|^2 \ge 0$, setzen wir

$$C^{(1)} = C - \varepsilon_1 S,$$

im Falle $\sum s_{ij}|\lambda_i - \mu_j|^2 < 0$ hingegen

$$C^{(1)} = C + \varepsilon_2 S.$$

Mit $C^{(1)} = (c_{ij}^{(1)})$ folgt dann

$$\sum c_{ij}^{(1)}|\lambda_i - \mu_j|^2 \le \sum c_{ij}|\lambda_i - \mu_j|^2 .$$

Gibt es ein $c_{ij}^{(1)}$, das $\neq 0$ und $\neq 1$ ist, wiederholen wir den Auswahl- und Umformungsschritt, bilden also eine Matrix $C^{(2)}$; usw. Nach endlich vielen Schritten gelangen wir zu einer Matrix $C^{(N)} = (c_{ij}^{(N)})$, die immer noch (DS) erfüllt, aber nur noch Elemente 0 und 1 enthält, also in jeder Zeile und Spalte genau eine Eins und ansonsten Nullen. Also gilt

$$\sum c_{ij}^{(N)}|\lambda_i - \mu_j|^2 = |\lambda_1 - \mu_{j_1}|^2 + \cdots + |\lambda_n - \mu_{j_n}|^2 \le \|A-B\|_F^2$$

mit einer Permutation $(j_1, \ldots, j_n)$ von $(1, \ldots, n)$. Analog ergibt eine modifizerte $C^{(i)}$-Definition eine Permutation $(k_1, \ldots, k_n)$ mit $|\lambda_1 - \mu_{k_1}|^2 + \cdots + |\lambda_n - \mu_{k_n}|^2 \ge \|A-B\|_F^2$.

Auswahl der c_{ij}: Man wähle ein erstes $c_{ij} \in (0,1)$, dann in derselben Zeile ein zweites, dann in der Spalte des zweiten ein drittes, dann in der Zeile des dritten ein viertes, usw. Sobald eine zuvor gewählte Zeile oder Spalte *nicht unmittelbar danach* erneut gewählt wird, endet damit die Wahl. Ist die Zeile bzw. Spalte nun *dreifach* belegt, streicht man alle anfänglichen bis zum *erstgewählten* Element in dieser Zeile bzw. Spalte. Ergebnis: In allen beteiligten Zeilen und Spalten sind je zwei c_{ij} ausgewählt. ■

Nun zwei Ungleichungen, welche Singulär- und Eigenwerte *quadratischer* Matrizen vergleichen. Dazu formulieren wir zunächst eine der einfachsten und grundlegendsten Faktorisierungsaussagen für quadratische Matrizen:

Satz Schur-Zerlegung

Zu jeder Matrix $A \in \mathbb{C}^{n\times n}$ gibt es eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$ mit

$$A = U \begin{pmatrix} \lambda_1 & \cdots & \star \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix} U^*.$$

Beweis:

Es gibt ein $\lambda_n \in \mathbb{C}$ mit $\mathrm{rg}(A-\lambda_n) < n$ (Eigenwert), also ein $\vec{u} \in \mathbb{C}^n$ mit $|\vec{u}| = 1$, $(A-\lambda_n)^*\vec{u} = \vec{0}$. Mit einer ONB $\{\vec{b}_1, \ldots, \vec{b}_{n-1}, \vec{u}\}$ folgt (LK=Linearkombination):
$A\vec{b}_i$ $(1 \le i < n)$, $(A-\lambda_n)\vec{u}$ sind LKen von $\vec{b}_1, \ldots, \vec{b}_{n-1}$.
D.h.: $(\vec{b}_1, \ldots, \vec{b}_{n-1}, \vec{u})^* A (\vec{b}_1, \ldots, \vec{b}_{n-1}, \vec{u}) = \left(\begin{array}{c:c} A_0 & * \\ \hdashline 0\cdots 0 & \lambda_n \end{array}\right)$
mit $A_0 \in \mathbb{C}^{(n-1)\times(n-1)}$, $\det(A-\lambda) = (\lambda_n - \lambda)\det(A_0 - \lambda)$.
Nun kann dieselbe Überlegung und Transformation auf A_0 angewendet werden, usw.; nach $n-1$ Schritten ergibt sich die Behauptung. Man nutzt dabei *geänderte* Matrizen wie $\left(\begin{array}{c:c} B & \begin{smallmatrix}0\\ \vdots \\ 0\end{smallmatrix} \\ \hdashline 0\cdots 0 & 1 \end{array}\right)$, $\left(\begin{array}{c:c} B & \begin{smallmatrix}0&0\\ \vdots&\vdots\\ 0&0\end{smallmatrix} \\ \hdashline \begin{smallmatrix}0\cdots 0\\ 0\cdots 0\end{smallmatrix} & \begin{smallmatrix}1&0\\ 0&1\end{smallmatrix} \end{array}\right)$, usw. ■

Bem.: Die schrittweise Konstruktion ermöglicht *jede Eigenwert-Reihenfolge* in der Diagonalen der oberen Dreiecksmatrix.

Reelle Schur-Zerlegungs-Varianten im Falle $\boldsymbol{A} \in \mathbb{R}^{n\times n}$:

a) Hat A einen reellen Eigenwert, gibt es dazu auch einen reellen Eigenvektor. Sind *alle* Eigenwerte von A reell, kann die bisherige Argumentation wortwörtlich übernommen werden, und man erhält eine *reelle* obere Dreiecksmatrix Δ und eine reell-orthogonale Matrix Q, so dass $A = Q^T \Delta Q$.

b) Es gibt reelle Matrizen *ohne* reelle Eigenwerte, z.B. Drehmatrizen. Im allgemeinen kommen reelle *und* komplexe Eigenwerte vor. Sei $\lambda = \alpha + \beta\mathrm{i}$ mit $\beta \neq 0$ Eigenwert von A, so dass also $A(\vec{u} + \mathrm{i}\vec{v}) = (\alpha + \beta\mathrm{i})(\vec{u} + \mathrm{i}\vec{v})$ mit $\vec{u}, \vec{v} \in \mathbb{R}^n$. Komplexe Konjugation ergibt die Gleichung $A(\vec{u} - \mathrm{i}\vec{v}) = (\alpha - \beta\mathrm{i})(\vec{u} - \mathrm{i}\vec{v})$. Da $\vec{u} + \mathrm{i}\vec{v}$ und $\vec{u} - \mathrm{i}\vec{v}$ linear unabhängig sind, als Eigenvektoren zu den verschiedenen Eigenwerten $\alpha + \beta\mathrm{i}$ und $\alpha - \beta\mathrm{i}$, sind auch $\vec{u}$ und $\vec{v}$ linear unabhängig. Ausmultiplizieren ergibt

$$\boldsymbol{A\vec{u} = \alpha\vec{u} - \beta\vec{v},\ A\vec{v} = \beta\vec{u} + \alpha\vec{v}.} \qquad (*)$$

Beide Eigenvektorgleichungen sind *jeweils* äquivalent zu $(*)$. Ergänzen wir $\vec{u}, \vec{v}$ durch $\vec{b}_3, \ldots, \vec{b}_n$ zu einer Basis des $\mathbb{R}^n$, ergibt sich $(\vec{u}, \vec{v}, \vec{b}_3, \ldots, \vec{b}_n)^{-1} A (\vec{u}, \vec{v}, \vec{b}_3, \ldots, \vec{b}_n) = \left(\begin{array}{cc:c} \alpha & \beta & * \cdots * \\ -\beta & \alpha & * \cdots * \\ \hdashline 0 & 0 & \\ \vdots & \vdots & A_0 \\ 0 & 0 & \end{array}\right)$. So erzielt man schrittweise eine obere Block-Dreiecksform mit 1x1- oder 2x2-Diagonal-Kästchen, den reellen Eigenwerten bzw. konjugiert komplexen Eigenwert-Paaren entsprechend.

c) Eine *orthogonale* Transformation: Durch Skalieren mit $\sigma + \mathrm{i}\tau$ wird $(\sigma\vec{u} - \tau\vec{v}) + \mathrm{i}(\tau\vec{u} + \sigma\vec{v})$ aus dem Eigenvektor, und $(\sigma\vec{u} - \tau\vec{v}) \cdot (\tau\vec{u} + \sigma\vec{v}) = \sigma\tau(|\vec{u}|^2 - |\vec{v}|^2) - (\sigma^2 - \tau^2)\vec{u}\cdot\vec{v} = 0$ bei passender Wahl von σ, τ. Für $\vec{u}' = \sigma\vec{u} - \tau\vec{v}$, $\vec{v}' = \tau\vec{u} + \sigma\vec{v}$ gilt dann $(*)$, zusätzlich $\vec{u}' \perp \vec{v}'$ und o.B.d.A. $|\vec{u}'| = 1$ (Faktor bei σ, τ). Normierung von $\vec{v}'$: In $(*)$ ersetzt man $-\beta$ durch $-\tilde{\beta}/|\vec{v}'|$, β durch $\hat{\beta}\,|\vec{v}'|$. Nun ergänzt man zu einer reellen ONB $\{\vec{u}', \vec{v}', \vec{b}_3, \ldots, \vec{b}_n\}$; usw. Jedem Eigenwert-Paar $\alpha \pm \beta\mathrm{i}$ entspricht so ein 2x2-Kästchen $\left|\begin{smallmatrix} \alpha & \beta/\varepsilon \\ -\beta\varepsilon & \alpha \end{smallmatrix}\right|$.

Da die Singulärwerte einer Matrix unitärinvariant sind, folgt aus den früher formulierten Ungleichungen (die σ_i wie immer *absteigend*)

$$|a_{11}| + \cdots + |a_{kk}| \le \sigma_1 + \cdots + \sigma_k$$

und der Schur-Zerlegung sofort die erste der folgenden Ungleichungen.

Satz Singulär- und Eigenwert-Vergleich

Die Matrix $A \in \mathbb{C}^{n\times n}$ habe die Eigenwerte $\lambda_1, \ldots, \lambda_n$ und die Singulärwerte $\sigma_1, \ldots, \sigma_n$. Dann gelten die Majorisierungsbeziehungen (mit ebenfalls absteigend angeordneten $|\lambda_i|$)

$$|\lambda_1| + \cdots + |\lambda_k| \le \sigma_1 + \cdots + \sigma_k\,, \qquad (*)$$
$$|\lambda_1 \cdots \lambda_k| \le \sigma_1 \cdots \sigma_k\,, \qquad (**)$$

jeweils für $1 \le k \le n$; bei $(**)$ gilt für $k = n$ das Gleichheitszeichen.

Die Ungleichungen $(**)$ wurden erstmals von Hermann Weyl 1949 formuliert. Alfred Horn zeigte 1954, dass bei gegebenen $\lambda_1, \ldots, \lambda_n$ sowie $\sigma_1, \ldots, \sigma_n$, welche $(**)$ erfüllen, es stets eine $n \times n$-Matrix mit diesen Eigenwerten und Singulärwerten gibt. (Horn/Johnson [2], S. 217ff.; **Marshall/Olkin/Arnold**, *Inequalities: Theory of Majorization and Its Applications, 2nd Edition*, Springer 2010, S. 322f.)

Beweis:

Nur $(**)$ ist noch zu beweisen. Wir benutzen folgenden

Hilfssatz

Sei $M \in \mathbb{C}^{m\times n}$. Sind $\vec{u}_1, \ldots, \vec{u}_k$ mit $k \le \min(m,n)$ untereinander orthogonale Einheitsvektoren aus $\mathbb{C}^n$ und $\vec{v}_1, \ldots, \vec{v}_k$ ebensolche aus $\mathbb{C}^m$, so gilt für $U := (\vec{u}_1, \ldots, \vec{u}_k)$, $V := (\vec{v}_1, \ldots, \vec{v}_k)$ und $N := V^*MU$

$$\sigma_i(M) \ge \sigma_i(N) \quad (1 \le i \le k).$$

Der Beweis des Hilfssatzes folgt weiter unten. Ist nun $U\Delta U^*$ eine Schur-Zerlegung der gegebenen Matrix $A \in \mathbb{C}^{n\times n}$, nennen wir U_k die aus den ersten k Spalten von U gebildete Matrix und wenden auf die $k \times k$-Matrix $\Delta_k := U_k^* A U_k$ den Hilfssatz an. Es folgt

$$\sigma_i(\Delta_k) \le \sigma_i \quad (1 \le i \le k);$$

und da die Determinante der (quadratischen) oberen Dreiecksmatrix Δ_k einerseits gleich $\lambda_1 \cdots \lambda_k$ beträgt, andererseits aufgrund der SVD $|\det \Delta_k| = \sigma_1(\Delta_k) \cdots \sigma_k(\Delta_k)$ gilt, folgen die Ungleichungen $(**)$. Der Gleichheitsfall $k = n$ ergibt sich durch Betrachtung von $\det A$ und der SVD von A. ■

Nun der *Beweis des Hilfssatzes:*

Sei $\mathcal{U}_k$ der von $\vec{u}_1, \ldots, \vec{u}_k$, den Spalten von U, aufgespannte Unterraum des $\mathbb{C}^n$. Mit dem Minimax-Prinzip ergibt sich

$$\begin{aligned}
\sigma_i(M) &= \min_{\vec{x}_1,\ldots,\vec{x}_{i-1}\in\mathbb{C}^n} \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i-1}}} |M\vec{x}| \\
&\geq \min_{\vec{x}_1,\ldots,\vec{x}_{i-1}\in\mathbb{C}^n} \max_{\substack{|\vec{x}|=1,\,\vec{x}\in\mathcal{U}_k \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i-1}}} |M\vec{x}| \\
&\underset{(*)}{=} \min_{\vec{x}_1,\ldots,\vec{x}_{i-1}\in\mathbb{C}^k} \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i-1}}} |MU\vec{x}| \\
&\geq \min_{\vec{x}_1,\ldots,\vec{x}_{i-1}\in\mathbb{C}^k} \max_{\substack{|\vec{x}|=1 \\ \vec{x}\perp\vec{x}_1,\ldots,\vec{x}_{i-1}}} |V^*MU\vec{x}| \\
&= \sigma_i(N)
\end{aligned}$$

für $1 \leq i \leq k$. ($(*)$: Beachte $\vec{x}\perp\vec{x}_1 \Leftrightarrow U\vec{x}\perp U\vec{x}_1$.) ■

Durch analoge Schlussweisen wie beim Beweis des Hilfssatzes kann man auch zeigen, dass für eine durch **Streichen einer Zeile oder Spalte** bei einer Matrix A entstehende Matrix A' für $k = 1, 2, \ldots$ gilt:

$$\boldsymbol{\sigma_k(A) \geq \sigma_k(A') \geq \sigma_{k+1}(A)\,.}$$

Aus $(**)$ folgen (*Majorisierung*; siehe Horn/Johnson [2] oder Bhatia [8]; eine kurze Einführung in Anhang B) die allgemeinen **Weylschen Ungleichungen**

$$\varphi(|\lambda_1|)+\cdots+\varphi(|\lambda_k|) \leq \varphi(\sigma_1)+\cdots+\varphi(\sigma_k) \;\; (1 \leq k \leq n)$$

für alle φ mit der Eigenschaft, dass $x \mapsto \varphi(\mathrm{e}^x)$ wachsend und konvex ist; insbesondere

$$|\lambda_1|^s+\cdots+|\lambda_k|^s \leq \sigma_1^s+\cdots+\sigma_k^s \;\; (s>0,\, 1 \leq k \leq n).$$

Satz John von Neumanns Spur-Ungleichung

Seien $A, B \in \mathbb{C}^{m\times n}$, $p = \min(m,n)$, und $\sigma_1(A), \ldots, \sigma_p(A)$ sowie $\sigma_1(B), \ldots, \sigma_p(B)$ seien die fallend angeordneten Singulärwerte der beiden Matrizen. Dann gilt

$$\Re\ \mathrm{Spur}(AB^*) \leq \sum_{i=1}^{p} \sigma_i(A)\sigma_i(B)$$

Beweis:

O.B.d.A. kann man von quadratischen Matrizen ausgehen, da man im Falle $m > n$ beiden Matrizen rechtsseitig $m-n$ Null-Spalten hinzufügen kann, ohne AB^* und die Singulärwerte zu ändern. Außerdem gilt generell:

$$\mathrm{Spur}(M_1M_2) = \mathrm{Spur}(M_2M_1) \quad (*),$$

wenn beide Matrizenprodukte sinnvoll sind.

Seien also $A, B \in \mathbb{C}^{n\times n}$, und mit $A = V_1\Sigma_1U_1^*$ sowie $B = V_2\Sigma_2U_2^*$ seien Singulärwertzerlegungen der Matrizen gegeben. Mit $U := U_1^*U_1$, $V := V_2^*V_1$ gilt dann wegen $(*)$ $\mathrm{Spur}(AB^*) = \mathrm{Spur}(\Sigma_1U\Sigma_2V) = \sum_{i=1}^{n}\sum_{k=1}^{n} \sigma_i(A)u_{ik}\sigma_k(B)v_{ki}$.

Also $\Re\ \mathrm{Spur}(AB^*) = \sum_{i,k=1}^{n} \sigma_i(A)\sigma_k(B)\Re\, u_{ik}v_{ki}$.

Nun betrachten wir die Matrix $\left(|u_{ik}v_{ki}|\right)_{\substack{1\le i\le n\\ 1\le k\le n}}$.

Klar ist ($|ab| \le \frac{1}{2}(|a|^2+|b|^2)$), dass die Zeilen- und Spaltensummen jeweils ≤ 1 sind. Unter dieser Voraussetzung gilt offenbar: Sind alle Zeilensummen genau gleich 1, folgt das auch für die Spaltensummen.

Umgekehrt: Ist bei Zeile i die Summe < 1, gibt's eine Spalte k, für die dies auch gilt.

Wir nennen die Matrix von jetzt an einfach $C = \left(c_{ik}\right)$.

Das Element c_{ik} wird so weit angehoben, bis die i-te Zeilensumme oder die k-te Spaltensumme $= 1$ ist, aber beide noch ≤ 1. So lange nicht alle Zeilen- und Spaltensummen $= 1$ sind, finden wir weiter solche Zeilen-Spalten-Paare; aber nach endlich vielen Anhebungsschritten sind wir fertig und haben eine doppeltstochastische Matrix $\widehat{C} = \left(\widehat{c}_{ik}\right)$, die gliedweise die ursprüngliche Matrix majorisiert.

Also $\Re$ Spur$(AB^*) \le \sum_{i,k=1}^{n} \sigma_i(A)\sigma_k(B)\widehat{c}_{ik}$. Nach dem Satz von Birkhoff/von Neumann (Anhang A) ist $\widehat{C}$ Konvexkombination von Permutationsmatrizen.

Sei $\widehat{C} = \sum_{j=1}^{N} \lambda_j P_j$ mit $\lambda_1, \ldots, \lambda_N > 0$ und $\lambda_1 + \cdots + \lambda_N = 1$.

Die Permutationsmatrizen: $P_j = \left(p_{ik}^{(j)}\right)_{\substack{1\le i\le n\\ 1\le k\le n}}$.

Damit

$$\sum_{i,k=1}^{n} \sigma_i(A)\sigma_k(B)\widehat{c}_{ik} = \sum_{j=1}^{N} \lambda_j \sum_{i,k=1}^{n} \sigma_i(A)\sigma_k(B)p_{ik}^{(j)}.$$

Es gibt (mindestens) eine P-Matrix, also ein j, für welches die i-k-Doppelsumme am größten ausfällt. Mit so einem j, etwa j_0, gilt dann

$$\sum_{i,k=1}^{n} \sigma_i(A)\sigma_k(B)\widehat{c}_{ik} \le \sum_{i,k=1}^{n} \sigma_i(A)\sigma_k(B)p_{ik}^{(j_0)}.$$

Es gibt eine Permutation $(\pi_1, \ldots, \pi_n)$ der Zahlen $1, 2, \ldots, n$, so dass

$\sum_{i,k=1}^{n} \sigma_i(A)\sigma_k(B)p_{ik}^{(j_0)} = \sum_{i=1}^{n} \sigma_i(A)\sigma_{\pi_i}(B)$. Bringt man die $\sigma_{\pi_i}(B)$ schrittweise durch Paarvertauschungen in die natürliche absteigende Reihenfolge, vergrößert man die Summe, wie exemplarisch durch die triviale Ungleichung

$$AB + ab \ge Ab + aB \;\Leftrightarrow\; (A-a)(B-b) \ge 0$$

klar wird.

Fazit: $\Re$ Spur$(AB^*) \le \sum_{i=1}^{n} \sigma_i(A)\sigma_i(B)$. ■

Eine unmittelbare Folgerung:

$$\sum_{i=1}^{n}(\sigma_i(A)-\sigma_i(B))^2 = \sum_{i=1}^{n}(\sigma_i^2(A)+\sigma_i^2(B)) - 2\sum_{i=1}^{n}\sigma_i(A)\sigma_i(B)$$
$$\le \|A\|_F^2 + \|B\|_F^2 - 2\Re \text{ Spur}(AB^*) = \|A-B\|_F^2.$$

Dies zeigt, dass die weiter oben „globaler Singulärwert-Vergleich“ genannte Ungleichung durch einfache Äquivalenzumformung aus der von Neumannschen Spur-Ungleichung hervorgeht. Der dortige *andere* Beweis (S. 29) ist also auch ein Beweis der Spur-Ungleichung.

Implizit haben wir bei der Umformung das *Frobenius-Skalarprodukt* für Matrizen verwandt. Dazu ein paar Bemerkungen:

$\Re(A,B)_F = \frac{1}{4}(\|A+B\|_F^2 - \|A-B\|_F^2) = \sum_{i,k} \frac{1}{4}(|a_{ik}+b_{ik}|^2 - |a_{ik}-b_{ik}|^2)$ $= \Re \sum_{i,k} a_{ik}\overline{b}_{ik} = \Re\ \mathrm{Spur}(AB^*)$; gilt auch für $A, B \in \mathbb{K}^{m\times n}$. Ferner $\Im(A,B)_F = \Re(A, \mathrm{i}B)_F = \Re\ \mathrm{Spur}(A(\mathrm{i}B)^*) = \Re\ -\mathrm{i}\,\mathrm{Spur}(AB^*) = \Im\ \mathrm{Spur}(AB^*)$. Fazit also: $(A,B)_F = \mathrm{Spur}(AB^*)$.

Insbesondere (man beachte die auftretende 1×1-Matrix):

$(\vec{v}_1\vec{u}_1^*, \vec{v}_2\vec{u}_2^*)_F = \mathrm{Spur}(\vec{v}_1\vec{u}_1^*\vec{u}_2\vec{v}_2^*) = \mathrm{Spur}(\vec{u}_1^*\vec{u}_2\vec{v}_2^*\vec{v}_1) = (\vec{u}_2,\vec{u}_1)(\vec{v}_1,\vec{v}_2)$.

Der *Satz von Wielandt/Hoffman* kann mittels des Anhangs A deutlich kürzer bewiesen werden: Zunächst erhalten wir mit dem Spektralsatz für normale Matrizen (der übrigens *ganz einfach* aus der Schur-Zerlegung folgt) $\|A-B\|_F^2 = \sum_{i,k} c_{ik}|\lambda_i - \mu_k|^2$ mit einer doppeltstochastischen Matrix $C = (c_{ik})$ und den Eigenwerten λ_i bzw. μ_k zweier normaler Matrizen A, B gleicher Reihenzahl n. Also folgt $\|A-B\|_F^2 = \sum_{j=1}^N t_j \sum_{i,k} p_{ik}^{(j)}|\lambda_i - \mu_k|^2$ mit gewissen Permutationsmatrizen $P_j = (p_{ik}^{(j)})$ $(1 \le j \le N)$ und positiven t_j mit $t_1 + \cdots + t_N = 1$. Indem man einmal alle i-k-Summen durch die kleinste, einmal alle durch die größte ersetzt, erhält man:

$$\sum_{i=1}^n |\lambda_i - \mu_{\pi_i}|^2 \le \|A-B\|_F^2 \le \sum_{i=1}^n |\lambda_i - \mu_{\tau_i}|^2$$

mit gewissen Permutationen $(\pi_1,\ldots,\pi_n)$ und $(\tau_1,\ldots,\tau_n)$ der Zahlen $1,\ldots,n$.

9. Die beste Rang-k-Approximation

Die Formel $\sum_i(\sigma_i(A) - \sigma_i(B))^2 \le \|A-B\|_F^2$ („globaler Singulärwert-Vergleich“(S. 28), äquivalent zur von Neumannschen Spur-Formel) zeigt: Hat die Matrix B den Rang k, gilt

$$\|A-B\|_F^2 \ge \sum_{i=k+1}^{\min(m,n)} \sigma_i^2(A)\,.$$

Diese untere Schranke wird exakt angenommen von einer Matrix, die sich unmittelbar aus einer SVD von A ergibt:

Satz Beste Rang-k-Näherung

Die bestmögliche Rang-k-Näherung an eine Matrix A mit der SVD $A = V\Sigma U^*$ ist gegeben durch

$$\mathbf{Rg_k}(A) = \boldsymbol{\sigma_1}\,\vec{\boldsymbol{v}}_1\,\vec{\boldsymbol{u}}_1^* + \cdots + \boldsymbol{\sigma_k}\,\vec{\boldsymbol{v}}_k\,\vec{\boldsymbol{u}}_k^*.$$

„Bestmöglich“ in folgendem Sinne:

$$\|A - \mathrm{Rg}_k(A)\|_F = \min_{\mathrm{rg}(B)=k} \|A-B\|_F\,,$$

und *nur* mittels einer SVD von A gebildete Matrizen vom Typ $B = \mathrm{Rg}_k(A)$ erzielen diesen Minimal-Abstand.

Hat A lauter verschiedene Singulärwerte, sind alle Rang-k-Näherungen eindeutig bestimmt.

Strenggenommen ist nur im Fall lauter verschiedener Werte σ_k $(1 \le k \le r)$ der Ausdruck $\mathrm{Rg}_k(A)$ eine wohldefinierte Funktion von A und k; anderenfalls ist $\mathrm{Rg}_k(A)$ immer auf eine vorher gewählte SVD zu beziehen.

Bei SCILAB (siehe S. 47) gibt es einen Befehl `sva(A,k)`, bei OCTAVE hingegen muss man eine eigene kleine Prozedur schreiben.

Die Singulärwerte bilden so etwas wie das *Rang-Spektrum* einer Matrix: Man kann die Matrix additiv zusammensetzen aus Rang-Eins-Matrizen vom Typ $\vec{v}\,\vec{u}^*$, versehen mit den Singulärwerten als Gewichtsfaktoren.

Matrizen $A = \vec{v}_1\vec{u}_1^* + \cdots + \vec{v}_k\vec{u}_k^*$ haben den Rang k, wenn die Vektoren $\vec{u}_1, \ldots, \vec{u}_k$ und die Vektoren $\vec{v}_1, \ldots, \vec{v}_k$ jeweils linear unabhängig sind, weil dann $A\vec{u} = \vec{0} \Leftrightarrow \vec{u} \perp \vec{u}_1, \ldots, \vec{u}_k$.

Im Sinne des der Frobeniusnorm entsprechenden Skalarproduktes $(A, B)_F$ bilden die Matrizen $\vec{v}_i\,\vec{u}_k^*$ mit Orthonormalbasen $\vec{u}_1, \ldots, \vec{u}_n$, $\vec{v}_1, \ldots, \vec{v}_m$ des $\mathbb{K}^n$ bzw. des $\mathbb{K}^m$ eine ONB des Vektorraumes $\mathbb{K}^{m\times n}$, da

$$(\vec{v}_1\vec{u}_1^*, \vec{v}_2\vec{u}_2^*)_F = (\vec{v}_1, \vec{v}_2)(\vec{u}_2, \vec{u}_1)\,.$$

Nun der **Beweis** des Satzes:
Nur noch die Eindeutigkeitsaussage ist zu beweisen.
Ist B irgendeine bestapproximierende Rang-k-Matrix, müssen ihre nichtverschwindenden Singulärwerte mit den ersten k Singulärwerten von A übereinstimmen. Stellt man A in der zu B gehörenden SVD-Basis dar, $B = V\Sigma U^*$, $A = V\tilde{A}U^*$, $\tilde{A} = (\tilde{a}_{ij})$, so folgt

$$\|A - B\|_F^2 = \sum_{i\neq j} |\tilde{a}_{ij}|^2 + \sum_i |\tilde{a}_{ii} - \sigma_i|^2.$$

Wäre $\tilde{a}_{ii} \neq \sigma_i$ für ein $i \in \{1, 2, \ldots, k\}$, könnte $\|A - B\|_F^2$ verkleinert werden, indem man den Singulärwert σ_i von B ein wenig variiert. Also muss $\tilde{a}_{ii} = \sigma_i$ gelten für $1 \le i \le k$. Da somit

$$|A\vec{u}_i|^2 = \left|\sum_{j=1}^m \tilde{a}_{ji}\vec{v}_j\right|^2 = \sum_{j=1}^m |\tilde{a}_{ji}|^2 \ge \sigma_i^2,$$

folgt gemäß Maximaleigenschaft der Singulärwerte von A schrittweise, dass

$$A\vec{u}_i = \sigma_i\vec{v}_i \quad (1 \le i \le k).$$

Also stimmen nicht nur die Singulärwerte, sondern auch die zugehörigen Singulärvektoren von B mit Singulärvektoren von A überein, die zu den k größten Singulärwerten gehören. ■

Für die induzierte Matrixnorm $\|\cdot\|_2$ (anstelle von $\|\cdot\|_F$) ist $\mathrm{Rg}_k(A)$ ebenfalls beste Rang-k-Approximation an A, *cum grano salis*:
Erstens gilt

$$\|A - \mathrm{Rg}_k(A)\|_2 = \sigma_{k+1},$$

da σ_{k+1} größter Singulärwert von $A - \mathrm{Rg}_k(A)$ ist.
Zweitens: Gilt $A = \sigma_1\vec{v}_1\vec{u}_1^* + \cdots + \sigma_r\vec{v}_r\vec{u}_r^*$ und $B = s_1\vec{q}_1\vec{p}_1^* + \cdots + s_k\vec{q}_k\vec{p}_k^*$ mit $r > k$, gibt es eine nichtverschwindende Linearkombination

$$\vec{x} = x_1\vec{u}_1 + \cdots + x_{k+1}\vec{u}_{k+1} \perp \vec{p}_1, \ldots, \vec{p}_k\,,$$

da homogene Gleichungssysteme mit mehr Unbekannten als Gleichungen bekanntlich immer nichttrivial lösbar sind; also

$$|A\vec{x} - B\vec{x}| = |A\vec{x}| \geq \sigma_{k+1}|\vec{x}|,$$
$$\|A - B\|_2 \geq \sigma_{k+1}.$$

Fazit: Für jede Rang-k-Matrix B gilt

$$\|A - \mathrm{Rg}_k(A)\|_2 \leq \|A - B\|_2\,.$$

Aber die bzgl. $\|\cdot\|_2$ beste Rang-k-Approximation *ist nicht eindeutig bestimmt.* Z.B. gilt:

$$B_t := (\sigma_1 - t\sigma_{k+1})\vec{v}_1\vec{u}_1^* + \cdots + (\sigma_k - t\sigma_{k+1})\vec{v}_k\vec{u}_k^*$$
$$\Rightarrow |(A - B_t)\vec{x}| \leq \sigma_{k+1}|\vec{x}| \quad (0 \leq t \leq 1);$$

also $\quad \|A - B_t\|_2 = \sigma_{k+1} \quad (0 \leq t \leq 1)\,.$

Der **Abstand einer invertierbaren Matrix von der Gesamtheit aller singulären Matrizen** lässt sich nach diesen Betrachtungen einfach durch den kleinsten Singulärwert ausdrücken – ob man nun die induzierte euklidische oder die Frobenius-Norm zum Maßstab nimmt. (*Kondition*: $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \sigma_1(A)/\sigma_n(A)$.)

Als **Beispiel** betrachten wir die Matrizen

$$A_n := \begin{pmatrix} 1 & -1 & \cdots & -1 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}\,.$$

Sie haben alle die Determinante 1 und machen einen ganz harmlosen nichtsingulären Eindruck. Dass sie aber *fast singulär* sind, zeigt folgende Tabelle ihrer kleinsten Singulärwerte (gerundet) für verschiedene n.

n	10	20	30	40
$\sigma_{n-1}(A_n)$	1.5	1.5	1.5	1.5
$\sigma_n(A_n)$	.003	$3\cdot 10^{-6}$	$2.79\cdot 10^{-9}$	$2.73\cdot 10^{-12}$

Man sieht, dass die Matrizen alle beinahe den Rang $n-1$ haben. Folgende kurze SCILAB-Prozedur diente zur Berechnung:

```
function [s1,s2]=fastsing(n)
A=eye(n,n)-bool2s(matrix([1:(n*n)],n,n)>matrix([1:(n*n)],n,n)')
S=svd(A)
s1=S(n-1)
s2=S(n)
endfunction
```

SCILAB, ein für Privatanwender freies, aber professionellen Ansprüchen genügendes mathematisch-numerisches Software-System, dessen Befehlssprache weitgehend identisch mit der des kommerziellen Systems Matlab ist, ermöglicht insbesondere den effizienten rechnenden Umgang mit Matrizen und Vektoren. Statt SCILAB kann man auch GNU OCTAVE verwenden, dessen Befehlssatz und Funktionalität ebenfalls fast vollständig mit Matlab übereinstimmen.

OCTAVE: `A=eye(5)-(reshape([1:25],5,5)>reshape([1:25],5,5)')`

SCILAB:

```
-->A=eye(5,5)-bool2s(matrix([1:(5*5)],5,5)>matrix([1:(5*5)],5,5)')
 A  =
    1.  - 1.  - 1.  - 1.  - 1.
    0.    1.  - 1.  - 1.  - 1.
    0.    0.    1.  - 1.  - 1.
    0.    0.    0.    1.  - 1.
    0.    0.    0.    0.    1.

-->[U S V]=svd(A)     // Bei Octave genau derselbe Befehl!
 V  =
  - 0.2556536    0.3290435    0.2547217  - 0.1364558    0.8618981
    0.0465883  - 0.3969474  - 0.6424765    0.4900540    0.4328202
    0.3403404  - 0.6440739    0.0808234  - 0.6433686    0.221092
    0.5720713  - 0.1142105    0.6002650    0.5337402    0.1203898
    0.6995524    0.5534327  - 0.3943227  - 0.2059737    0.0801439
 S  =
    2.7363296    0.           0.           0.           0.
    0.           1.6819438    0.           0.           0.
    0.           0.           1.5480531    0.           0.
    0.           0.           0.           1.5094537    0.
    0.           0.           0.           0.           0.0929853
 U  =
  - 0.6995524    0.5534327    0.3943227  - 0.2059737    0.0801439
  - 0.5720713  - 0.1142105  - 0.6002650    0.5337402    0.1203898
  - 0.3403404  - 0.6440739  - 0.0808234  - 0.6433686    0.221092
  - 0.0465883  - 0.3969474    0.6424765    0.4900540    0.4328202
    0.2556536    0.3290435  - 0.2547217  - 0.1364558    0.8618981

-->sum(abs(A-U*S*V'))   //Probe!     =0 bis auf Rundungsfehler
 ans  =
    1.553D-14
```


10. Beste orthogonale und unitäre Approximation

Gegeben eine nichtsinguläre (annähernd orthogonale bzw. unitäre) $n \times n$-Matrix A, suchen wir die im Sinne der Frobeniusnorm beste Approximation B, die exakt orthogonal (unitär) ist. Dies ist eine Fragestellung, die z.B. bei Aufgaben der Bewegungsplanung (in der Robotik) eine Rolle spielen kann.

Die Lösung des Problems ist denkbar einfach: Mit der Singulärwertzerlegung $A = V\Sigma U^*$ und orthogonaler (unitärer) Matrix B gilt

$$\begin{aligned}\|A-B\|_F &= \|V(\Sigma - V^*BU)U^*\|_F \\ &= \|\Sigma - V^*BU\|_F \\ &= \sum_i |\sigma_i - c_{ii}|^2 + \sum_{i\neq k} |c_{ik}|^2 \\ &= \sum_i \sigma_i^2 + \sum_{i,k} |c_{ik}|^2 - 2\sum_i \sigma_i \,\Re c_{ii}\end{aligned}$$

mit $V^*BU =: (c_{ik})$. Da V^*BU orthogonal bzw. unitär, wird der Ausdruck minimal *nur* im Falle $c_{11} = \cdots = c_{nn} = 1$ (gilt auch noch für $\sigma_{n-1} > \sigma_n = 0$), also $V^*BU = E$ und daher

$$\boldsymbol{B = V\,U^*}.$$

11. SVD in Hilberträumen*

Die Herleitung der SVD benutzt implizit die *Kompaktheit* der Einheitssphäre in endlichdimensionalen Räumen und ist daher nicht auf *beliebige* lineare Operatoren ausdehnbar. Sind H_1 und H_2 Hilberträume, heißt ein linearer Operator $T : H_1 \to H_2$ kompakt, wenn für jede beschränkte Folge $(x_n)_{n\in\mathbb{N}}$ in H_1 die Bildfolge $(T\,x_n)_{n\in\mathbb{N}}$ in H_2 eine konvergente Teilfolge besitzt. (Mit anderen Worten: T bildet beschränkte auf relativkompakte Mengen ab.)

Satz SVD kompakter Operatoren

Sind H_1 und H_2 Hilberträume über $\mathbb{R}$ (oder beide über $\mathbb{C}$), so ist ein linearer Operator $T : H_1 \to H_2$ genau dann kompakt, wenn es orthonormale Folgen $(u_n)_{n\in\mathbb{N}}$ in H_1 sowie $(v_n)_{n\in\mathbb{N}}$ in H_2 gibt, ferner eine monoton fallende Nullfolge $(\sigma_n)_{n\in\mathbb{N}}$ reeller Zahlen (Singulärwerte von T), so dass

$$T\,x = \sum_{i=1}^{\infty} \sigma_i\,(x, u_i)\,v_i \quad (x \in H_1)\,.$$

Eine gewisse Verallgemeinerung: *Nukleare* Operatoren und Räume (Grothendieck 1955). Aber ein Analogon für beliebige Banachräume gilt *nicht*, da kompakte Operatoren nicht in allen Banachräumen endlichdimensional approximierbar sind (Enflo 1973).

Für diejenigen, die mit den Grundbegriffen der Hilbertraumtheorie noch nicht vertraut sind, folgen hier zur Einführung einige vorbereitende Bemerkungen (10 Seiten, Voraussetzung nur eine gewisse Kenntnis von Skalarprodukt und Norm, Orthogonalität und Orthonormalbasis (ONB) in den Fällen $\mathbb{R}^n$ und $\mathbb{C}^n$), ehe der Beweis des Satzes dargestellt wird.

Die nächstliegende *unendlich*dimensionale Verallgemeinerung der Vektorräume $\mathbb{R}^n$ und $\mathbb{C}^n$ ist der Übergang zu unendlich vielen Koordinaten, zu ∞-Tupeln, also Folgen $(x_i)_{i\in\mathbb{N}}$ reeller bzw. komplexer Zahlen.

Dabei treten aber Konvergenzfragen auf: Um die euklidische Länge eines Vektors in gewohnter Weise als $\sqrt{\sum |x_i|^2}$ definieren zu können, muss $\sum_{i=1}^{\infty} |x_i|^2 < \infty$ *gefordert* werden.

Der reelle (komplexe) Hilbertsche Folgenraum ist also definiert als die Menge aller reellen (komplexen) Zahlenfolgen $\mathbf{x} = (x_i)_{i\in\mathbb{N}}$ mit $\sum_{i=1}^{\infty} |x_i|^2 < \infty$.

Wegen $|x_i \cdot y_i| \le \frac{1}{2}(|x_i|^2 + |y_i|^2)$ folgt sofort die Existenz des *Skalarprodukts*

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} := \sum_{i=1}^{\infty} x_i \bar{y}_i$$

für Folgen $\mathbf{x} = (x_i)$ und $\mathbf{y} = (y_i)$. Die komplexe Konjugation bei den y_i bezieht sich natürlich nur auf den *komplexen* Folgenraum; reell gilt ja $\bar{y}_i = y_i$.

Auch ist klar, dass mit $\mathbf{x}$ und $\mathbf{y}$ auch $\mathbf{x}+\mathbf{y}$, die *komponentenweise* Summe der beiden Folgen, im Folgenraum liegt: $|x_i + y_i|^2 \leq 2|x_i|^2 + 2|y_i|^2$.

Es ist eine leichte Übung, die Gültigkeit aller Skalarproduktaxiome zu bestätigen. Alle Rechnungen, Umformungen, Argumentationen, die auf diesen Axiomen beruhen, behalten ihre Gültigkeit.

Auch die Definitionen von Norm, Abstand und *Cauchyfolge* bleiben dieselben: Für eine Cauchyfolge $\big(\mathbf{x}_n\big)_{n\in\mathbb{N}}$ gilt $|\mathbf{x}_m - \mathbf{x}_n| < \varepsilon$ $(m,n \geq n_0(\varepsilon))$. Und man kann zeigen, dass in diesem Fall ein Element $\mathbf{y}$ des Folgenraumes existiert, gegen das die „Folgen-Folge“ konvergiert.

Man nennt letzteres die *Vollständigeit* des Hilbertschen Folgenraumes, und weil es ein ganz zentraler Punkt ist, sei die einfache Begründung dargestellt.

Mit $\mathbf{x}_n = (x_{n\,i})_{i\in\mathbb{N}}$ gilt $|x_{m\,i} - x_{n\,i}| \leq |\mathbf{x}_m - \mathbf{x}_n|$, weshalb die Folgen $(x_{n\,i})_{n\in\mathbb{N}}$ reelle (komplexe) Cauchyfolgen, also konvergent sind, d.h. $\lim_{n\to\infty} x_{n\,i} = y_i$ existiert für alle $i \in \mathbb{N}$.

Sei nun n_0 gewählt, so dass $|\mathbf{x}_m - \mathbf{x}_n| < \varepsilon$ für $m,n \geq n_0$. Dann gilt
$\sum_{i=1}^{N} |x_{m\,i} - x_{n\,i}|^2 < \varepsilon^2$ für $m,n \geq n_0$ und beliebiges N.

Mit $m \to \infty$ also auch $\sum_{i=1}^{N} |y_i - x_{n\,i}|^2 \leq \varepsilon^2$ und daher

$$\sum_{i=1}^{\infty} |y_i - x_{n\,i}|^2 \leq \varepsilon^2 \quad (n \geq n_0) \qquad (*)$$

Folglich gehört $\mathbf{y}-\mathbf{x}_n$ und damit auch $\mathbf{y} = (\mathbf{y}-\mathbf{x}_n)+\mathbf{x}_n$ dem Hilbertschen Folgenraum an, und wegen $(*)$ ist $\mathbf{y}$ der Grenzwert der Folge $\big(\mathbf{x}_n\big)_{n\in\mathbb{N}}$.

Allgemein nennen wir nun einen normierten Vektorraum über $\mathbb{R}$ ($\mathbb{C}$) einen reellen (komplexen) *Hilbertraum*, wenn seine Norm durch ein Skalarprodukt bestimmt ist, $\|\mathbf{x}\| = \sqrt{\mathbf{x}\cdot\mathbf{x}}$, und jede Cauchyfolge bzgl. dieser Norm gegen ein Element des Raumes konvergiert; letzteres ist die *Vollständigkeit*. Da ein vollständiger normierter Raum auch *Banachraum* genannt wird, ist also ein Hilbertraum ein Banachraum mit skalarprodukterzeugter Norm. Wenn man verschiedene Normen parallel nutzt, schreibt man auch $\|\mathbf{x}\|_2$, um eine skalarprodukterzeugte (= *euklidische*) Norm zu kennzeichnen. Da wir *nur* euklidische Normen benutzen, schreiben wir stattdessen einfach immer $|\mathbf{x}|$.

David Hilbert (1862-1943) war ein bedeutender deutscher, Stefan Banach (1892-1945) ein bedeutender polnischer Mathematiker; beide haben Bahnbrechendes zur Entstehung der Theorie der unendlichdimensionalen Vektorräume beigetragen.

In jedem Hilbertraum hat man mit dem Skalarprodukt, ganz wie im endlichdimensionalen Spezialfall, *Orthogonalität* und *Orthonormalsysteme* als besonders nützliche linear unabhängige Systeme von Vektoren zur Verfügung.
Sei H ein Hilbertraum, $\mathbf{x} \in H$ und $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ ein orthonormales System von Vektoren in H. Dann gilt $\mathbf{x} - \sum_{i=1}^{k} (\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i \perp \mathbf{u}_j$ $(1 \leq j \leq k)$, also

$$|\mathbf{x}|^2 = \left|\mathbf{x} - \sum_{i=1}^{k} (\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i\right|^2 + \left|\sum_{i=1}^{k} (\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i\right|^2$$

und damit

$$\left|\sum_{i=1}^{k} (\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i\right|^2 = \sum_{i=1}^{k} |(\mathbf{x}, \mathbf{u}_i)|^2 \leq |\mathbf{x}|^2 .$$

Ist also $\big(\mathbf{u}_n\big)_{n\in\mathbb{N}}$ eine *orthonormale Folge* in H (im Folgenraum ist die Folge $\big(\mathbf{u}_n\big)$ mit $u_{nn} = 1$ und $u_{nk} = 0$ $(k \neq n)$ ein simples Beispiel einer orthonormalen Folge), so gilt $\sum_{i=1}^{\infty} |(\mathbf{x}, \mathbf{u}_i)|^2 < \infty$ und $\lim_{n\to\infty} (\mathbf{x}, \mathbf{u}_n) = 0$.

Nun kommen wir zur *Kompaktheit*.
Man nennt im $\mathbb{R}^n$ oder $\mathbb{C}^n$ eine Menge *kompakt*, wenn sie abgeschlossen und beschränkt ist. Im $\mathbb{R}^n$ oder $\mathbb{C}^n$ gilt allgemein:
Eine Menge ist abgeschlossen und beschränkt genau dann, wenn jede Folge von Elementen der Menge eine gegen ein Element der Menge konvergierende Teilfolge besitzt.

Stetige Funktionen nehmen auf kompakten Mengen absolute Maxima und Minima an und sind auf kompakten Mengen *gleichmäßig* stetig; die Beweise sind ganz einfach mittels der angesprochenen Folgeneigenschaft. Für viele weitere Schlüsse ist der Übergang zu konvergenten Teilfolgen bei beschränkten Folgen ein wesentliches Hilfsmittel.

Im *unendlichdimensionalen* Fall ist aber eine abgeschlossene und beschränkte Menge wie die Einheitskugel *nicht mehr kompakt* im Sinne der Folgeneigenschaft! Das zeigt am einfachsten eine orthonormale Folge: Je zwei Glieder einer solchen Folge haben den Abstand $\sqrt{2}$, wie man leicht nachrechnet.

Man macht deshalb in beliebigen normierten Räumen (Vektorräumen mit einer *Norm*, einem Vektorlängen-Maß) und sogar in allgemeinen metrischen Räumen die Folgeneigenschaft zum Kriterium für Kompaktheit:
Eine Menge ist kompakt genau dann, wenn jede Folge in dieser Menge eine gegen ein Element der Menge konvergierende Teilfolge besitzt.

Es zeigt sich: Viele lineare Abbildungen (*Transformationen* oder *Operatoren* nennt man sie meist im unendlichdimensionalen Fall) $T : H_1 \to H_2$ „verdichten“ so das Bild beschränkter Mengen, dass für die Bild-Menge doch die charakteristische Folgeneigenschaft gilt.

Zum Beispiel die Umkehroperatoren zu *Differentialoperatoren* sind oft kompakte *Integraloperatoren*; womit ein kurzer Hinweis auf die Motivation der Erforschung unendlichdimensionaler Räume gegeben ist: Es geht um die mathematischen Eigenschaften von Differentialgleichungsproblemen und ihrer Lösungen, ein anspruchsvolles Gebiet mit vielfältigen Anwendungen.

Daher die Definition:

Ein linearer Operator $T : H_1 \to H_2$ heißt *kompakt*, wenn für eine *beschränkte* Menge $\mathcal{M} \subseteq H_1$ jede Folge in $T(\mathcal{M})$ eine konvergente Teilfolge besitzt.

Die einfachsten Beispiele für kompakte lineare Operatoren sind alle stetigen linearen Operatoren, bei denen einer der beiden Räume H_1, H_2 endlichdimensional ist. Denn dann ist in jedem Fall das Bild $T(\mathcal{M})$ einer beschränkten Menge eine beschränkte Menge in einem *endlich*dimensionalen Raum, weshalb jede Folge eine konvergente Teilfolge besitzt.
Wichtiger Spezialfall: $H_2 = \mathbb{R}$ (bzw. $= \mathbb{C}$, je nachdem, ob H_1 ein reeller oder komplexer Vektorraum ist).
In diesem Fall nennt man die stetigen linearen Operatoren *stetige lineare Funktionale*.

In endlichdimensionalen Räumen sind *lineare* Abbildungen immer stetig, im unendlichdimensionalen Fall keineswegs. Aber *kompakte* Operatoren $T : H_1 \to H_2$ *sind* stetig:

Sei $S_1 := \{\mathbf{x} \in H_1 \mid |\mathbf{x}| = 1\}$ die Einheitssphäre in H_1. Gäbe es eine Folge $(\mathbf{x}_n)$ in S_1, deren Bildfolge $\big(T\mathbf{x}_n\big)$ nicht beschränkt ist, so könnte man o.B.d.A. davon ausgehen, dass für die Bildfolge $|T\mathbf{x}_n| \to \infty$ gilt. Das aber schlösse die Existenz einer konvergenten Teilfolge aus. Die Menge der Werte $|T\mathbf{x}| \quad (\mathbf{x} \in S_1)$ ist folglich beschränkt und damit T ein *beschränkter*, mit anderen Worten *stetiger* linearer Operator; denn mit $|T| := \sup_{\mathbf{x}\in S_1} |T\mathbf{x}|$ gilt für $\mathbf{x} \neq \mathbf{y}$ wegen Linearität
$|T\mathbf{x} - T\mathbf{y}| = |T(\mathbf{x}-\mathbf{y})| = \left|T\left(\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|}\right)\right| \cdot |\mathbf{x}-\mathbf{y}| \leq |T| \cdot |\mathbf{x}-\mathbf{y}|$.
Man nennt $|T|$ die *Norm* des Operators; genauer: die durch die Vektornormen *induzierte* Operatornorm.
(Die Frobenius-Norm bei Matrizen ist das Beispiel einer *nicht* durch Vektornormen induzierten Operatornorm: Die Einheitsmatrix hat nicht die Norm 1. Trotzdem gilt für die Frobenius-Norm wie *für jede induzierte* Operatornorm die bei Abschätzungen wichtige *Submultiplikativitäts-Ungleichung* $|T_1 \circ T_2| \leq |T_1| \cdot |T_2|$.)

Konkrete Beispiele *unstetiger* linearer Funktionale auf Hilberträumen gibt es *nicht*. (Man definiert unstetige lineare Funktionale mittels „Hamel-Basen“; bei manchen unvollständigen *Prä*hilberträumen wie z.B. dem aller Folgen mit nur endlich vielen Gliedern $\neq 0$ ist eine solche linearalgebraische Basis leicht explizit anzugeben.) Aber die *stetigen* linearen Funktionale sind in Hilberträumen *sehr* konkret:

Ist H Hilbertraum und $f : H \to \mathbb{K}$ stetiges lineares Funktional ($\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ Skalarenkörper des Raumes), so gibt es genau ein $\mathbf{y} \in H$ mit $f(\mathbf{x}) = (\mathbf{x}, \mathbf{y})$ $(\mathbf{x} \in H)$. Dabei gilt $|f| = |\mathbf{y}|$.

(*Rieszscher Darstellungssatz*, benannt nach F. Riesz, einem Mitbegründer der *Funktionalanalysis*, also der Theorie unendlichdimensionaler Vektorräume und ihrer Operatoren.)

Beweis: Sei $H_0 := \{\mathbf{x} \in H \mid f(\mathbf{x}) = 0\} = \mathcal{N}(f)$, der Nullraum von f, ein *abgeschlossener* Unterraum. Falls $H_0 = H$, leistet $\mathbf{y} = \mathbf{0}$ das Gewünschte und offenbar kein anderes $\mathbf{y}$.

Anderenfalls sei $\mathbf{x}_0 \in H \setminus H_0$. Wir betrachten eine Folge $\big(\mathbf{x}_n\big)$ in H_0 mit $\lim_{n\to\infty} |\mathbf{x}_n - \mathbf{x}_0| = \inf_{\mathbf{x}\in H_0} |\mathbf{x} - \mathbf{x}_0| =: d$.

Aufgrund der *Parallelogramm-Identität* (leicht rechnerisch zu bestätigen) $|\mathbf{u}+\mathbf{v}|^2 + |\mathbf{u}-\mathbf{v}|^2 = 2|\mathbf{u}|^2 + 2|\mathbf{v}|^2$ folgt wegen $\mathbf{x}_m - \mathbf{x}_n = (\mathbf{x}_m - \mathbf{x}_0) - (\mathbf{x}_n - \mathbf{x}_0)$

$$|\mathbf{x}_m - \mathbf{x}_n|^2 = 2\underbrace{|\mathbf{x}_m - \mathbf{x}_0|^2}_{\to d^2} + 2\underbrace{|\mathbf{x}_n - \mathbf{x}_0|^2}_{\to d^2} - 4\underbrace{\left|\frac{\mathbf{x}_m + \mathbf{x}_n}{2} - \mathbf{x}_0\right|^2}_{\geq d^2}$$

Also ist $\big(\mathbf{x}_n\big)$ eine Cauchyfolge, deren Limes $\mathbf{x}^*$ wegen Abgeschlossenheit in H_0 liegt.

Es gilt also $|\mathbf{x}^* - \mathbf{x}_0| = \min_{\mathbf{x}\in H_0} |\mathbf{x} - \mathbf{x}_0|$, und dieses Minimum ist *eindeutig bestimmt*, da wir statt $\mathbf{x}_m$ und $\mathbf{x}_n$ ja auch zwei Minimumsstellen (Grenzwerte) in die spezielle Parallelogramm-Identität einsetzen können.

Die dimensionsunabhängige Überlegung auf Seite 8 zeigt, dass $|\mathbf{x}^* + t\mathbf{y} - \mathbf{x}_0| - |\mathbf{x}^* - \mathbf{x}_0| \approx t(\mathbf{y}, \mathbf{x}^* - \mathbf{x}_0)/|\mathbf{x}^* - \mathbf{x}_0|$ für kleine $|t|$, falls $\mathbf{y} \not\perp \mathbf{x}^* - \mathbf{x}_0$; im komplexen Fall betrachtet man $t\mathrm{e}^{\mathrm{i}\varphi}$ statt t mit passendem φ.

Also folgt $\mathbf{x}_1 := \mathbf{x}_0 - \mathbf{x}^* \perp H_0$, $f(\mathbf{x}_1) = f(\mathbf{x}_0) \neq 0$.

(Anschaulich gesagt: Gäbe es in H_0 eine nicht zu $\mathbf{x}_0 - \mathbf{x}^*$ senkrechte Richtung, könnte man längs dieser Richtung den Abstand zu $\mathbf{x}_0$ noch verkleinern.)

Da nun offenbar $\mathbf{x} - \frac{f(\mathbf{x})}{f(\mathbf{x}_1)}\mathbf{x}_1 \in H_0$ für beliebiges $\mathbf{x} \in H$, folgt $(\mathbf{x} - \frac{f(\mathbf{x})}{f(\mathbf{x}_1)}\mathbf{x}_1, \mathbf{x}_1) = 0$ und damit $f(\mathbf{x}) = (\mathbf{x}, \frac{\overline{f(\mathbf{x}_1)}}{|\mathbf{x}_1|^2}\mathbf{x}_1)$.

Die *Eindeutigkeit* ist klar, da aus $(\mathbf{x}, \mathbf{y}_1) = (\mathbf{x}, \mathbf{y}_2)$, also $(\mathbf{x}, \mathbf{y}_1 - \mathbf{y}_2) = 0$ für alle $\mathbf{x}$ unmittelbar $\mathbf{y}_1 - \mathbf{y}_2 = 0$ folgt. Einerseits gilt $\left|\frac{\overline{f(\mathbf{x}_1)}}{|\mathbf{x}_1|^2}\mathbf{x}_1\right| \leq |f|$, zum anderen $|f| = \sup_{|\mathbf{x}|=1} |f(\mathbf{x})| = \sup_{|\mathbf{x}|=1} |(\mathbf{x}, \mathbf{y})| \leq |\mathbf{y}|$. Das beweist den Rieszschen Darstellungssatz. ■

Implizit wurde übrigens damit bewiesen, dass $H_0^\perp$ *ein*dimensional ist. Außerdem sei erwähnt, dass die Parallelogramm-Identität *charakteristisch* ist für skalarprodukterzeugte Normen; gilt die Identität, wird durch $\Re(\mathbf{x}, \mathbf{y}) := \frac{1}{4}(|\mathbf{x}+\mathbf{y}|^2 - |\mathbf{x}-\mathbf{y}|^2)$ und $\Im(\mathbf{x}, \mathbf{y}) = \Re(\mathbf{x}, \mathrm{i}\mathbf{y})$ ein Skalarprodukt definiert, das genau die gegebene Norm erzeugt (Satz von P.Jordan/J.v.Neumann).

Mit dem Rieszschen Satz ist leicht der *adjungierte Operator* zu definieren, der den Begriff der transponierten bzw. adjungierten Matrix verallgemeinert. Wir beschränken uns dabei auf *stetige* lineare Operatoren.

Ist $A : H_1 \to H_2$ ein stetiger linearer Operator, so ist für festes $\mathbf{y} \in H_2$ durch $f_\mathbf{y} : \mathbf{x} \mapsto (A\mathbf{x}, \mathbf{y})$ $(\mathbf{x} \in H_1)$ ein stetiges lineares Funktional gegeben. Es gibt also genau ein $\mathbf{y}^* \in H_1$ mit $(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y}^*)$ für alle $\mathbf{x} \in H_1$.
Wir definieren:
$A^* : \mathbf{y} \mapsto \mathbf{y}^*$ $(\mathbf{y} \in H_2)$ (zu A *adjungierter* Operator).
Linearität ist klar; denn für alle $\mathbf{x} \in H_1$ gilt
$(A\mathbf{x}, \alpha\mathbf{y}_1 + \beta\mathbf{y}_2) = \bar{\alpha}(A\mathbf{x}, \mathbf{y}_1) + \bar{\beta}(A\mathbf{x}, \mathbf{y}_2) =$
$= \bar{\alpha}(\mathbf{x}, A^*\mathbf{y}_1) + \bar{\beta}(\mathbf{x}, A^*\mathbf{y}_2) = (\mathbf{x}, \alpha A^*\mathbf{y}_1 + \beta A^*\mathbf{y}_2)$.
Beschränktheit: Es gilt $|f_\mathbf{y}(\mathbf{x})| \le |A\mathbf{x}| \cdot |\mathbf{y}| \le |A| \cdot |\mathbf{y}| \cdot |\mathbf{x}|$, also $|\mathbf{y}^*| = |f_\mathbf{y}| \le |A| \cdot |\mathbf{y}|$ und damit $|A^*| \le |A|$.
Ferner für irgendein $\mathbf{x}_0 \neq \mathbf{0}$: $f_{A\mathbf{x}_0}(\mathbf{x}_0) = |A\mathbf{x}_0|^2$, also $|f_{A\mathbf{x}_0}| \ge \frac{|A\mathbf{x}_0|}{|\mathbf{x}_0|}|A\mathbf{x}_0|$, weshalb $|A^*| \ge \frac{|A\mathbf{x}_0|}{|\mathbf{x}_0|}$, da $|A^*A\mathbf{x}_0| = |f_{A\mathbf{x}_0}|$ nach Riesz. ($|A^*| \ge |A|$ folgt auch direkt aus $A^{**} = A$.)
Fazit: $|A^*| = |A|$.

Es gilt $(A^*\mathbf{y}, \mathbf{x}) = (\mathbf{y}, \mathbf{x}^*)$ $(\mathbf{y} \in H_2)$ mit $\mathbf{x}^* = A^{**}\mathbf{x}$.
Andererseits $(A^*\mathbf{y}, \mathbf{x}) = \overline{(\mathbf{x}, A^*\mathbf{y})} = \overline{(A\mathbf{x}, \mathbf{y})} = (\mathbf{y}, A\mathbf{x})$.
Also $A^{**}\mathbf{x} = A\mathbf{x}$ und damit, da $\mathbf{x}$ beliebig, $A^{**} = A$.

Es gilt: *Ist A kompakt, so auch A^*.*
Beweis: Sei $(\mathbf{y}_n)$ eine Folge in H_2, $|\mathbf{y}_n| \le 1$ $(n \in \mathbb{N})$. Dann ist $(A^*\mathbf{y}_n)$ eine beschränkte Folge in H_1. Es gilt
$|A^*\mathbf{y}_m - A^*\mathbf{y}_n|^2 = (\mathbf{y}_m - \mathbf{y}_n, AA^*\mathbf{y}_m - AA^*\mathbf{y}_n) \le$
$\le 2|AA^*\mathbf{y}_m - AA^*\mathbf{y}_n|$.
Da es nach Voraussetzung eine Teilfolge von $(AA^*\mathbf{y}_n)$ gibt, die Cauchyfolge ist, gilt dies auch für die entsprechende Teilfolge von $(A^*\mathbf{y}_n)$.

Nun der **Beweis** des Satzes über die SVD:

Sei $T : H_1 \to H_2$ kompakt. Wir setzen $|T| > 0$ voraus, anderenfalls bleibt nichts zu zeigen. Mit S_1 und S_2 seien im folgenden die Sphären der Einheitsvektoren in H_1 bzw. H_2 bezeichnet.

Im ersten Schritt zeigen wir, dass ein $\mathbf{u}_1 \in S_1$ existiert, so dass $T\mathbf{u}_1 = |T|\mathbf{v}_1$ mit $\mathbf{v}_1 \in S_2$.

(Im Prinzip derselbe Ansatz wie im Matrizen-Fall. Aber da S_1 nicht kompakt ist, braucht man eine zusätzliche Überlegung.)

Wir gehen aus von einer Folge $(\mathbf{y}_n)$ in S_2 mit $T^*\mathbf{y}_n \to \mathbf{x}_1$ mit $|\mathbf{x}_1| = |T|$. Da T^* kompakt und $|T^*| = |T|$, gibt es eine solche Folge.
Aus $(T^*\mathbf{y}_n, T^*\mathbf{y}_n) = (TT^*\mathbf{y}_n, \mathbf{y}_n) \le |TT^*\mathbf{y}_n|$ folgt für $n \to \infty$: $|\mathbf{x}_1|^2 \le |T\mathbf{x}_1|$, und da andererseits $|T\mathbf{x}_1| \le |T| \cdot |\mathbf{x}_1| = |\mathbf{x}_1|^2$, folgt $|T\mathbf{x}_1| = |\mathbf{x}_1|^2$ und daher $\mathbf{u}_1 \in S_1$, $\mathbf{v}_1 \in S_2$ und $T\mathbf{u}_1 = |T|\mathbf{v}_1$ für $\mathbf{u}_1 := \mathbf{x}_1/|T|$, $\mathbf{v}_1 := T\mathbf{x}_1/|T|^2$.

Ganz genau wie im endlichdimensionalen Fall gilt $T\mathbf{u} \perp T\mathbf{u}_1$, falls $\mathbf{u} \perp \mathbf{u}_1$. Das frühere Argument, an das schon im Beweis des Rieszschen Darstellungssatzes erinnert wurde, kann wörtlich übernommen werden.

Mit $T(\mathbf{u}_1^\perp) \subseteq (T\mathbf{u}_1)^\perp$ kann nun das Argument des ersten Schritts iteriert werden, und per Induktion folgt die Existenz einer orthonormalen Folge $\left(\mathbf{u}_n\right)_{n\in\mathbb{N}}$ in H_1, einer orthonormalen Folge $\left(\mathbf{v}_n\right)_{n\in\mathbb{N}}$ in H_2 sowie einer monoton fallenden Folge $(\sigma_n)_{n\in\mathbb{N}}$ positiver Zahlen, so dass $T\mathbf{u}_n = \sigma_n\mathbf{v}_n$ $(n \in \mathbb{N})$ mit $\sigma_1 = |T|$.

Bricht die schrittweise Konstruktion nicht wegen $T(\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}^\perp) = \{\mathbf{0}\}$ nach n Schritten ab, bilden die σ_n eine *Nullfolge*, da anderenfalls die Folge der $T\mathbf{u}_n$ offenbar keine konvergente Teilfolge besäße, im Widerspruch zur Kompaktheit von T; denn $|\sigma_i\mathbf{v}_i - \sigma_k\mathbf{v}_k|^2 = \sigma_i^2 + \sigma_k^2$ $(i \neq k)$.

Da $\mathbf{x} - \sum_{i=1}^k (\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i = \tilde{\mathbf{x}} \perp \mathbf{u}_1, \ldots, \mathbf{u}_k$ und $|\tilde{\mathbf{x}}| \leq |\mathbf{x}|$, folgt $|T\mathbf{x} - \sum_{i=1}^k \sigma_i(\mathbf{x}, \mathbf{u}_i)\mathbf{v}_i| = |T\tilde{\mathbf{x}}| \leq \sigma_{k+1}|\mathbf{x}|$ und mit $k \to \infty$ die Reihen-Darstellung von T, die Singulärwertzerlegung.

Bleibt noch der Umkehrschluss, dass aus dieser Reihendarstellung die Kompaktheit von T folgt. Mit $T_n\mathbf{x} := \sum_{i=1}^n \sigma_i(\mathbf{x}, \mathbf{u}_i)\mathbf{v}_i$ $(n \in \mathbb{N})$ gilt $|T - T_n| = \sigma_{n+1}$, und T_n als stetiger linearer Operator mit endlichdimesionalem Bild *ist* kompakt. Der im Anschluss formulierte Sachverhalt beendet daher den Beweis. ■

Sind $T_n : H_1 \to H_2$ $(n \in \mathbb{N})$ kompakte lineare Operatoren, H_1, H_2 Hilberträume, und ist $T : H_1 \to H_2$ ein linearer Operator mit $|T - T_n| \to 0$, so ist auch T kompakt.

Beweis: Sei $\left(\mathbf{x}_n\right)$ eine durch M beschränkte Folge in H_1. Wir wählen Teilfolgen $\left(\mathbf{x}_n^{(k)}\right)$, so dass $\left(T_k\mathbf{x}_n^{(k)}\right)$ jeweils eine in H_2 konvergente Folge ist. Dabei können wir o.B.d.A. davon ausgehen, dass stets $\left(\mathbf{x}_n^{(k+1)}\right)$ eine Teilfolge von $\left(\mathbf{x}_n^{(k)}\right)$ ist. Die „Diagonal"-Folge $\left(\mathbf{x}_n^{(n)}\right)$ ist dann ab ihrem k-ten Glied Teilfolge von $\left(\mathbf{x}_n^{(k)}\right)$:

$\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \mathbf{x}_3^{(1)}, \mathbf{x}_4^{(1)}, \ldots$ $\quad$ $\mathbf{x}_1^{(2)}, \mathbf{x}_2^{(2)}, \mathbf{x}_3^{(2)}, \mathbf{x}_4^{(4)}, \ldots$

$\mathbf{x}_1^{(3)}, \mathbf{x}_2^{(3)}, \mathbf{x}_3^{(3)}, \mathbf{x}_4^{(3)}, \ldots$ $\quad$ $\mathbf{x}_1^{(4)}, \mathbf{x}_2^{(4)}, \mathbf{x}_3^{(4)}, \mathbf{x}_4^{(4)}, \ldots$

Also ist $\left(T_k\mathbf{x}_n^{(n)}\right)_{n\in\mathbb{N}}$ konvergent für alle k. Zu $\varepsilon > 0$ gibt's ein n_0 mit $|T - T_n| < \varepsilon$ $(n \geq n_0)$, und mit geignet gewähltem n_1 (von n_0 und ε abhängig) gilt

$|T\mathbf{x}_m^{(m)} - T\mathbf{x}_n^{(n)}| \leq |(T-T_{n_0})\mathbf{x}_m^{(m)}| + |(T-T_{n_0})\mathbf{x}_n^{(n)}| + |T_{n_0}\mathbf{x}_m^{(m)} - T_{n_0}\mathbf{x}_n^{(n)}| <$

$(2M+1)\varepsilon$ $(m, n \geq n_1)$. ■

Literatur-Hinweis: Es gibt *viele* Bücher zur *Funktionalanalysis*, darunter gewichtige Klassiker, Meisterwerke, geschrieben von berühmten Mathematikern. Nur ein *schmales* Taschenbuch, dennoch eine gehaltvolle Darstellung wichtiger Teile dieses mittlerweile *sehr umfangreichen* mathematischen Fachgebiets: [12].

Schlussbemerkung: Meist wird die SVD als Anhängsel der Eigenwerttheorie symmetrischer Matrizen bzw. Operatoren behandelt, siehe z.B. Horn/Johnson[1], S. 150f. Stattdessen legte ich - im endlichdimensionalen *und* unendlichdimensionalen Fall - Wert auf eine Herleitung aus *ursprünglichen, eigenständigen* Gesichtspunkten, da die SVD einen begrifflich eigenständigen Platz verdient. (Der Zusammenhang mit Eigenwerten *wird* aber natürlich in den Abschnitten 6 und 7 eingehend behandelt.)

12. Literatur

[1], [2] **Roger A. Horn / Charles R. Johnson**
Matrix Analysis, 2nd Ed.; *Topics in Matrix Analysis*
Cambridge University Press 2013 ([1]1985); 1991/94(corr.)

[3] **Gene H. Golub / Charles F. van Loan**
Matrix Computations, 3rd Ed., Oxford University Press 1996

[4] **Gilbert Strang**, *Lineare Algebra*, Springer 2003

[5] **Rudolf Zurmühl**
Matrizen und ihre technischen Anwendungen, 4. Aufl.
Springer 1964

[6] **Gilbert Strang**
Linear Algebra and Its Applications, 3rd Ed.
Harcourt, Brace, Jovanovich 1988

[7] **Peter D. Lax**, *Linear Algebra and Its Applications, 2nd Ed.*
Wiley 2007

[8] **Rajendra Bhatia**, *Matrix Analysis*, Springer 1997

[9] **Theodor Bröcker**, *Lineare Algebra und Analytische Geometrie*, Birkhäuser 2003

[10] **Bertram Huppert / Wolfgang Willems**, *Lineare Algebra*
Teubner 2006

[11] **Michael W. Berry / Murray Browne**
Understanding Search Engines, 2nd Ed., SIAM 2005

[12] **Friedrich Hirzebruch / Winfried Scharlau**
Einführung in die Funktionalanalysis, Bibl. Institut 1971

[13] **Gilbert W. Stewart**
On the Early History of the Singular Value Decomposition,
SIAM Review 35 (1993), pp. 551–566

[1], [2], [3] informieren umfassend über die Singulärwertzerlegung, [4] ist ein gutes Anfängerlehrbuch. Der ältere Klassiker [5] behandelt knapp die SVD quadratischer Matrizen (S. 215f.). Selbst manche exzellenten neueren Lehrbücher erwähnen die SVD ganz am Rande oder gar nicht ([7], [9], [10]), im Gegensatz zu internationaler Standardliteratur wie [4], [6], [8]. IR-Anwendungen der SVD schildert [11].

Anhang A: Satz von Birkhoff/von Neumann

Eine quadratische Matrix heißt *doppeltstochastisch*, wenn ihre Elemente nichtnegativ sind und jede Zeilen- sowie jede Spalten-Summe 1 beträgt.

Satz (Garrett Birkhoff und John von Neumann):
Jede doppeltstochastische Matrix $M \in \mathbb{R}^{n\times n}$ *ist darstellbar als Konvexkombination* $M = \sum_i \lambda_i P_i$ *mit* Permutationsmatrizen P_i*; die Anzahl der Summanden ist* $\leq n^2-n+1$*. Die doppeltstochastischen sind die* konvexe Hülle *der Permutationsmatrizen.*

Es gilt also $\lambda_i > 0$ (alle i), $\sum_i \lambda_i = 1$; die P_i entstehen aus der Einheitsmatrix durch Permutation der Zeilen (oder Spalten), besitzen also in jeder Zeile und Spalte genau einen Eintrag 1, ansonsten lauter Nullen.

Beweis: Wir benutzen den *Heiratssatz* aus der Kombinatorik, dessen Beweis wir am Schluss nachliefern. Er besagt:
Gegeben sei eine Gruppe von n Frauen und n Männern. Hat jede Teilmenge von k Frauen insgesamt mindestens k Freunde unter den Männern, gibt's eine „Heirat" (Bijektion der Frauen- auf die Männer-Menge), bei der jede Frau einen ihrer Freunde heiratet.

Sei $M \in \mathbb{R}^{n\times n}$ doppeltstochastisch. Zeile i der Matrix heiße mit Spalte k *befreundet*, wenn $m_{ik} > 0$.

Gäbe es Zeilen $i_1, i_2, \ldots, i_m$, die insgesamt mit weniger als m Spalten befreundet sind, so hätte man damit weniger als m Spalten,

in denen schon allein durch die Elemente aus den genannten Zeilen die Element*summe* m erreicht würde, im Widerspruch zur Elementsumme 1 in jeder Spalte. *Also* ist die Voraussetzung des Heiratssatzes erfüllt, es gibt eine Heirat. Was bedeutet: Es gibt eine Permutation $(\pi_1, \pi_2, \ldots, \pi_n)$, so dass $m_{i\pi_i} > 0$ $(1 \leq i \leq n)$.
Sei nun P die Permutationsmatrix, bei der genau in den Positionen (i, π_i) Einsen stehen. Dann ist $M_1 = M - \min_{1\leq i\leq n} m_{i\pi_i} \cdot P$ eine Matrix, die mindestens ein Element 0 mehr besitzt als M und das $(1 - \min_{1\leq i\leq n} m_{i\pi_i})$-Fache einer doppeltstochastischen Matrix.
Endlich viele Subtraktions-Schritte ergeben die Nullmatrix. Vor dem letzten $= N$-ten Schritt hat man eine doppeltstochastische Matrix mit $\geq N-1$ Nullen; also $N - 1 \leq n^2 - n$. ■

Beweis des Heiratssatzes: Induktion über die Anzahl der Frauen. Fall n=1 ist trivial. Angenommen, für $\leq n$ Frauen und Männer gebe es eine Heirat. Sei F eine Menge von n+1 Frauen, die bzgl. einer Menge M von n+1 Männern die Voraussetzung erfüllt.
Zwei Fälle: *Entweder* hat jede echte Teilmenge $F_0 \subset F$ sogar einen Freund mehr; dann verheiraten wir *eine* Frau mit einem Freund und wenden auf die restlichen Damen und Herren, für die immer noch die Voraussetzung gilt, die Induktionsannahme an. *Oder aber* es gibt eine Teilmenge F_0 mit genau gleichvielen Freunden, Menge M_0. Dann hat jede Menge $F_1 \subseteq F \setminus F_0$ mindestens $|F_1|$ Freunde *außerhalb* von M_0, da es $\geq |F_1 \cup F_0| = |F_1| + |F_0|$ Freunde für $F_1 \cup F_0$ geben muss. Also ist die Induktionsvoraussetzung auf F_0, M_0 und auf $F \setminus F_0, M \setminus M_0$ getrennt anwendbar. Die demnach möglichen zwei Teilheiraten kann man zu einer Gesamtheirat für F zusammenfügen. ■

Anhang B: Majorisierung

Gelte $x_1 \geq x_2 \geq \cdots \geq x_n$ sowie $x'_1 \geq x'_2 \geq \cdots \geq x'_n$. Man sagt, das reelle n-Tupel $(x_1, \ldots, x_n)$ *majorisiere* das n-Tupel $(x'_1, \ldots, x'_n)$, wenn

$$x'_1 + \cdots + x'_k \leq x_1 + \cdots + x_k \quad (1 \leq k \leq n), \qquad (*)$$
$$x'_1 + \cdots + x'_n = x_1 + \cdots + x_n. \qquad (**)$$

In Zeichen: $(\boldsymbol{x'_1, \ldots, x'_n}) \prec (\boldsymbol{x_1, \ldots, x_n})$.

Gilt nur $(*)$, aber nicht $(**)$, spricht man von *schwacher* Majorisierung: $(\boldsymbol{x'_1, \ldots, x'_n}) \prec_w (\boldsymbol{x_1, \ldots, x_n})$.

Bei *Permutationen*, also Bijektionen $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$, benutzen wir die Schreibweise $\pi(x_1, , \ldots, x_n) = (x_{\pi(1)}, \ldots, x_{\pi(n)})$.

SATZ 1: *Es gilt* $(x'_1, \ldots, x'_n) \prec (x_1, \ldots, x_n)$ *genau dann, wenn*

$$(x'_1, \ldots, x'_n) = \sum_{i=1}^{N} t_i \, \pi_i(x_1, \ldots, x_n) \qquad (***)$$

mit Permutationen π_i, $t_i > 0$ $(1 \leq i \leq N)$ *und* $\sum_{i=1}^{N} t_i = 1$.

D.h.: Majorisierte n-Tupel sind beliebige *Konvexkombinationen* von Permutationen des majorisierenden n-Tupels; bei $(***)$ ist die Reihenfolge der x_i und x'_i durch Wahl der π_i beliebig veränderbar. (Bei der *Definition* der Majorisierung durch $(*), (**)$ hingegen sind fallende x'_i *Voraussetzung*, wie das Gegenbeispiel $(x_1, x_2) = (c, c)$, $(x'_1, x'_2) = (c - \varepsilon, c + \varepsilon)$ zeigt.)

Beweis:

Zuerst Richtung „$\Leftarrow$“. Gilt $x'_k = \sum_{i=1}^{N} t_i \, x_{\pi_i(k)}$ $(1 \leq k \leq n)$, folgt

$$x'_1 + \cdots + x'_k = \sum_{j=1}^{k} \sum_{i=1}^{N} t_i \, x_{\pi_i(j)} = \sum_{i=1}^{N} t_i \sum_{j=1}^{k} x_{\pi_i(j)} \leq \sum_{i=1}^{N} t_i (x_1 + \cdots + x_k).$$

Letzteres ist gleich $x_1 + \cdots + x_k$; und im Falle $k = n$ hat man $\sum_{j=1}^{n} x_{\pi_i(j)} = x_1 + \cdots + x_n$, also $x'_1 + \cdots + x'_n = x_1 + \cdots + x_n$.

Nun Richtung „$\Rightarrow$“. Induktion über n.

$n = 1$: Trivial!

$n \to n + 1$:

Gegeben $(x'_1, \ldots, x'_{n+1}) \prec (x_1, \ldots, x_{n+1})$, beide Tupel fallend durchnumeriert.

Erster Fall: $x'_1 = x_1$. Dann gibt es laut Induktionsannahme Permutationen π_i und positive t_i mit $\sum_{i=1}^{N} t_i = 1$, so dass

$$(x'_2, \ldots, x'_{n+1}) = \textstyle\sum_{i=1}^{N} t_i \, \pi_i(x_2, \ldots, x_{n+1}),$$

da ja auch $(x'_2, \ldots, x'_{n+1}) \prec (x_2, \ldots, x_{n+1})$. Die π kann man auch auffassen als Permutationen von $(1, 2, \ldots, n+1)$, bei denen 1 fest bleibt.

Zweiter Fall: $x'_1 < x_1$. Da $x'_1 + \cdots + x'_{n+1} = x_1 + \cdots + x_{n+1}$, gibt es ein $k \geq 2$ mit $x_1 \geq \cdots \geq x_{k-1} \geq x'_1 > x_k$ und damit ein $t \in (0, 1)$ mit $x'_1 = t \, x_1 + (1 - t) x_k$. Damit auch $t \, x_k + (1 - t) x_1 =: \tilde{x}'_1 > x_k$. Die Konvexkombination

$$t(x_1, \ldots, x_k, \ldots) + (1 - t)(x_k, \ldots, x_1, \ldots) = (x'_1, \ldots, \tilde{x}'_1, \ldots)$$

hat an den Positionen $\neq 1, \neq k$ unveränderte Komponenten x_i. Ordnet man wieder nach fallender Größe, ergibt sich $(x_2, \ldots, x_{k-1}, x'_1, \tilde{x}'_1, x_{k+1}, \ldots)$, falls $x'_1 \geq \tilde{x}'_1$, und z.B. $(\tilde{x}'_1, x_2, \ldots, x_{k-1}, x'_1, x_{k+1}, \ldots)$, falls $\tilde{x}'_1 > x'_1$ ($\tilde{x}'_1$ landet auf *irgendeinem* der Plätze 1 bis $k - 1$). Das bedeutet: Die ersten $k - 1$ Elemente sind $\geq x'_1$, und die Summe der ersten k Elemente hat denselben Wert wie vorher. Also majorisiert das veränderte Tupel weiterhin $(x'_1, \ldots, x'_{n+1})$. Und die Majorisierung *bleibt erhalten*, wenn bei beiden Tupeln das Element x'_1 entfernt wird:

Bei den Vergleichssummen (im Sinne von $(*)$) der Längen $\leq k-2$ wird nur die ohnehin kleinere Summe verkleinert, bei Summenlänge $k-1$ bleiben beim majorisierenden Tupel nur Elemente $\geq x'_1$, und ab Länge k wird in beiden Summen dasselbe gestrichen.

Also kann man die Induktionsannnahme auf die verkürzten Tupel anwenden: $(x'_2, \ldots, x'_{n+1})$ ist Konvexkombination von Permutationen des n-Tupels, das durch Streichen von x'_1 am Platz $k-1$ bzw. k übrigbleibt. Indem man alle Permutationen um die Abbildung des Streichplatzes auf die Position 1 erweitert, hat man die gewünschte Darstellung. ■

Nebenbei sieht man, dass alles schrittweise aus Konvexkombinationen der Identität mit reinen *Paarvertauschungs-Permutationen* (genau ein Paar wird vertauscht, alle anderen Elemente bleiben an ihrem Platz) zusammengesetzt werden kann.

SATZ 2: *Genau dann gilt* $(x'_1, \ldots, x'_n) \prec (x_1, \ldots, x_n)$*, wenn es eine doppeltstochastische Matrix* $P = (p_{ik})$ *gibt mit*

$$(x'_1, \ldots, x'_n) = (x_1, \ldots, x_n)P; \quad \textit{d.h. } x'_k = \sum_{i=1}^{n} p_{ik} x_i \ (1 \leq k \leq n).$$

Dies folgt unmittelbar aus der Tatsache, dass die doppeltstochastischen Matrizen genau die Konvexkombinationen von Permutationsmatrizen sind; siehe Anhang A. ■

Wir betrachten nun eine konvexe monoton wachsende Funktion $\varphi : D \to \mathbb{R}$, deren Definitionsbereich $D \subseteq \mathbb{R}$ die vorkommenden Tupel-Elemente umfasst.

SATZ 3: *Ist* φ *konvex und wachsend, gilt*

$(x'_1, \ldots, x'_n) \prec (x_1, \ldots, x_n) \Rightarrow (\varphi(x'_1), \ldots, \varphi(x'_n)) \prec_w (\varphi(x_1), \ldots, \varphi(x_n))$.

Beweis: Es gibt eine doppeltstochastische Matrix $P = (p_{ij})$, so dass $\varphi(x'_i) = \varphi\big(\sum_{j=1}^{n} p_{ij} x_j\big) \leq \sum_{j=1}^{n} p_{ij} \varphi(x_j)$ für $1 \leq i \leq n$. Dabei gilt $\varphi(x_1) \geq \varphi(x_2) \geq \cdots \geq \varphi(x_n)$, da φ wachsend. Für $1 \leq k \leq n$ folgt $\displaystyle\sum_{i=1}^{k} \varphi(x'_i) \leq \sum_{i=1}^{k} \sum_{j=1}^{n} p_{ij} \varphi(x_j) = \sum_{j=1}^{n} \left(\sum_{i=1}^{k} p_{ij} \right) \varphi(x_j)$.
Die Koeffizienten der $\varphi(x_j)$ sind sämtlich ≥ 0 und ≤ 1 und ergeben insgesamt k Zeilensummen einer doppeltstochastischen Matrix, also die Gesamtsumme k. Wir können den kleineren $\varphi(x_j)$ die Koeffizienten „wegnehmen“ und die Koeffizienten der k größten zum Wert 1 „auffüllen“. Damit $\sum_{i=1}^{k} \varphi(x'_i) \leq \sum_{j=1}^{k} \varphi(x_j)$. ■

Nun die *Weylschen Ungleichungen*. Der Vergleichssatz (Seite 35) sagt aus, dass für die fallend angeordneten Eigenwertbeträge $|\lambda_i|$ sowie Singulärwerte σ_i einer Matrix $A \in \mathbb{C}^{n \times n}$ gilt:
$|\lambda_1| \cdots |\lambda_k| \leq \sigma_1 \cdots \sigma_k$ $(1 \leq k \leq n)$, mit Gleichheit für $k = n$.
Also $\ln|\lambda_1| + \cdots + \ln|\lambda_k| \leq \ln\sigma_1 + \cdots + \ln\sigma_k$ $(1 \leq k \leq n)$, mit Gleichheit wiederum für $k = n$. Mit anderen Worten:
$(\ln|\lambda_1|, \ldots, \ln|\lambda_n|) \prec (\ln\sigma_1, \ldots, \ln\sigma_n)$.
Ist also $\varphi \circ \exp$ wachsend und konvex, so folgt nach Satz 3:
$(\varphi(|\lambda_1|), \ldots, \varphi(|\lambda_n|)) \prec_w (\varphi(\sigma_1), \ldots, \varphi(\sigma_n))$. Das *sind* die Weylschen Ungleichungen. Allerdings ist noch der Fall $\operatorname{rg} A = r < n$ zu diskutieren, da dann $\ln\sigma_n$ und $\ln|\lambda_n|$ nicht definiert sind.
Dann gilt $\lambda_{r+1} = \cdots = \lambda_n = 0 = \sigma_{r+1} = \cdots = \sigma_n$, und mit $c := \min(\ln|\lambda_r|, \ln\sigma_r)$ sowie $d := \sum_{i=1}^{r} (\ln\sigma_i - \ln|\lambda_i|)$ folgt:
$\ln|\lambda_1|, \ldots, \ln|\lambda_r|, c) \prec (\ln\sigma_1, \ldots, \ln\sigma_r, c-d)$, und mit Satz 3 gelten die Weyl-Ungleichungen (inklusive trivialer Summanden $\varphi(0)$).

Anhang C: Geraden auf Geraden

Ausgangspunkt der Überlegungen zur Singulärwertzerlegung war die Feststellung, dass lineare Abbildungen Geraden auf Geraden, Quadrate auf Parallelogramme und Kreise auf Ellipsen abbilden. Die erste dieser Eigenschaften ergibt eine anschaulich-einfache, aber eher unübliche Charakterisierung der Linearität, die hier präzise formuliert werden soll.

Ohne eine Zusatz-Voraussetzung sind Abbildungen von *Geraden auf Geraden* nicht notwendigerweise linear, wie etwa das Beispiel

$$f(\vec{x}) := \vec{a} + |\vec{x}| \cos |\vec{x}| \ \vec{b} \quad (\vec{x} \in \mathbb{R}^2)$$

mit beliebigen $\vec{a}, \vec{b} \in \mathbb{R}^2$ zeigt. Da $\mathbb{R}$ und $\mathbb{R}^2$ gleichmächtig sind, gibt es auch *injektive* f, deren Bild eine einzige Gerade ist, die also Geraden *in* Geraden abbilden. Wir nennen in diesem Zusammenhang eine Abbildung $f : \mathbb{R}^2 \to \mathbb{R}^2$ *nichttrivial*, wenn ihr Bild nicht in einer einzigen Geraden enthalten ist. Ein Beispiel: die x-Achse wird *auf* zwei ihrer Punkte, alle anderen Punkte werden auf einunddenselben Nicht-x-Achsenpunkt abgebildet.

Satz Geraden-Kriterium

Sei $f : \mathbb{R}^2 \to \mathbb{R}^2$. Falls f nichttrivial ist und Geraden *auf* Geraden abbildet, ebenso, falls f surjektiv ist und Geraden *in* Geraden abbildet, ist $f - f(\vec{0})$ *linear*.

Beweis:
Gäbe es ein echtes Dreieck $\vec{a}, \vec{b}, \vec{c}$, so dass $f(\vec{a})$, $f(\vec{b})$ und $f(\vec{c})$ auf einer Geraden lägen, so lägen sämtliche Bildpunkte der Geraden durch zwei Eckpunkte, damit letztlich alle Bilder auf einer einzigen Geraden; f wäre trivial.
Gäbe es zwei Punkte $\vec{a}$ und $\vec{b}$ mit $f(\vec{a}) = f(\vec{b})$, wäre demnach auch f trivial (Hinzunahme eines dritten Punktes).
Also ist f auf jeden Fall *injektiv*, unter beiden Voraussetzungen.
Ausserdem folgt, dass im zweiten Fall auch Geraden *auf* Geraden abgebildet werden, da kein Punkt der Bildgeraden Bild eines außerhalb der Urbildgeraden liegenden Punktes sein kann, weil ja echte Dreiecke auf echte Dreiecke abgebildet werden.
Nun folgt sofort, dass Parallelen auf Parallelen abgebildet werden, da ein Schnittpunkt der Injektivität widerspräche.
Fazit: f ist injektiv und bildet Parallelen auf Parallelen ab.

O.B.d.A. nehmen wir von jetzt an $f(\vec{0}) = \vec{0}$ an.
Sind $\vec{a}$, $\vec{b}$ linear unabhängig, bilden

$$\vec{0}, \quad \vec{a}, \quad \vec{b}, \quad \vec{a} + \vec{b}$$

ein Parallelogramm. Also ist $\overline{f(\vec{a}), f(\vec{a}+\vec{b})}$ parallel zu $\overline{\vec{0}, f(\vec{b})}$ und $\overline{f(\vec{b}), f(\vec{a}+\vec{b})}$ parallel zu $\overline{\vec{0}, f(\vec{a})}$. Es folgt:

$$f(\vec{a} + \vec{b}) = f(\vec{a}) + f(\vec{b});$$

d.h.: f ist *additiv*. (Betrachten von $\vec{a} - \vec{c} + \vec{c}$ und $\vec{a} - \vec{c} + \vec{b} + \vec{c}$ liefert den Nachweis im Falle *abhängiger* $\vec{a}$ und $\vec{b}$.)
Nach Voraussetzung gilt

$$f(\lambda\vec{a}) = \alpha\, f(\vec{a}), \quad f(\lambda\vec{b}) = \beta\, f(\vec{b}).$$

Da $f(\lambda\vec{a} + \lambda\vec{b})$ Vielfaches von $f(\vec{a} + \vec{b}) = f(\vec{a}) + f(\vec{b})$ ist, muss $\alpha = \beta$ gelten; also:

$$f(\lambda\vec{a}) = \varphi(\lambda)\, f(\vec{a})$$

mit von $\vec{a}$ *unabhängiger* Funktion $\varphi : \mathbb{R} \to \mathbb{R}$. Wegen Additivität von f folgt

$$\varphi(\lambda + \mu) = \varphi(\lambda) + \varphi(\mu),$$

und wegen Injektivität $\varphi(\lambda) \neq 0$ für $\lambda \neq 0$.
Nun gilt einerseits

$$f(\lambda\vec{a} + \mu\vec{b}) = \varphi(\lambda) f(\vec{a}) + \varphi(\mu) f(\vec{b}),$$

zum anderen

$$f(\lambda\vec{a} + \mu\vec{b}) = \varphi(\lambda) f(\vec{a} + \frac{\mu}{\lambda}\vec{b}) = \varphi(\lambda)\Big(f(\vec{a}) + \varphi\big(\tfrac{\mu}{\lambda}\big) f(\vec{b})\Big),$$

also $\varphi(\mu) = \varphi(\lambda) \cdot \varphi\big(\frac{\mu}{\lambda}\big)$ für $\lambda \neq 0$. M.a.W.:

$$\varphi(\lambda\mu) = \varphi(\lambda) \cdot \varphi(\mu).$$

Insbesondere $\varphi(1) = 1$. Additivität und Induktion ergeben

$$\varphi(\rho) = \rho \qquad (\rho \in \mathbb{Q});$$

und wegen $\varphi(\lambda) = \varphi(\sqrt{\lambda})^2 > 0$ $(\lambda > 0)$ ist φ *monoton wachsend*, weshalb schließlich

$$\varphi(\lambda) = \lambda \qquad (\lambda \in \mathbb{R}).$$

Mithin ist gezeigt (im Falle $f(\vec{0}) = \vec{0}$):

$$f(\vec{a}+\vec{b}) = f(\vec{a}) + f(\vec{b}) \ (\vec{a}, \vec{b} \in \mathbb{R}^2), \quad f(\lambda\vec{a}) = \lambda f(\vec{a}) \ (\lambda \in \mathbb{R}, \vec{a} \in \mathbb{R}^2).$$

■

Anhang D: Camille Jordans SVD-Herleitung

Ausführliche Angaben zu den frühen Arbeiten zur SVD findet man bei Horn/Johnson [2], inklusive vieler Literaturhinweise. Die *ersten* Herleitungen der SVD stammen, fast gleichzeitig und unabhängig voneinander, von den beiden großen Mathematikern Eugenio Beltrami (1873) und Camille Jordan (1874). Der Italiener Beltrami ist berühmt insbesondere für seine Beiträge zur Differentialgeometrie, und nach dem Franzosen, u.a. Autor eines vielfach bewunderten dreibändigen *Cours d'Analyse de l'École Polytechnique*, ist die von ihm entdeckte *Jordansche Normalform* quadratischer Matrizen benannt. Beide beschränken sich auch bei der SVD auf den reellen *quadratischen* Fall.

Nur *quadratische* Matrizen zu betrachten ist aber keine wesentliche Einschränkung, wie zunächst gezeigt wird.

Sei $A = (\vec{a}_1, \ldots, \vec{a}_n) \in \mathbb{R}^{m\times n}$ mit $m < n$. (Den Fall $m > n$ führt man durch Transponieren auf diesen zurück.) Durch Hinzufügen von $n-m$ Null-Zeilen entsteht aus A die Matrix $\tilde{A} = \left(\begin{array}{c|c} \vec{a}_1 \cdots \vec{a}_n \\ \hline 0 \cdots 0 \\ \vdots \quad \vdots \\ 0 \cdots 0 \end{array}\right)$.

Nun sei $\tilde{A} = U\Sigma V^T$ mit $U = (\vec{u}_1, \ldots, \vec{u}_n)$ und $V = (\vec{v}_1, \ldots, \vec{v}_n)$ eine SVD von $\tilde{A}$, so dass also $\tilde{A}\vec{v}_i = \sigma_i \vec{u}_i$ $(1 \le i \le n)$, wobei $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n \ge 0$. Der Rang von $\tilde{A}$ ist $\le m$, weshalb mit einem $r \le m$ gilt: $\sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$. (Den Trivialfall des Ranges 0 klammern wir natürlich aus.) Die unteren $n-m$ Komponenten der Vektoren $\vec{u}_1$, ..., $\vec{u}_r$ verschwinden jeweils. Deshalb kann man sie auch als orthonormale Vektoren $\vec{u}^o_1, \ldots, \vec{u}^o_r$ des $\mathbb{R}^m$ auffassen und durch geeignete $\vec{u}^o_{r+1}, \ldots, \vec{u}^o_m$ zu einer ONB des $\mathbb{R}^m$ ergänzen; so dass mit $A\vec{v}_i = \sigma_i \vec{u}^o_i$ $(1 \le i \le m)$ eine SVD für A gegeben ist.

Nun Jordans Beweis, ungefähr wie bei G. W. Stewart [13] dargestellt, aber ganz mit heutigen Ausdrucks- und Argumentationsweisen.

Ausgangspunkt ist eine Bilinearform $\vec{x}^T A\vec{y}$ mit $A \in \mathbb{R}^{n\times n}$. Jordan überlegt, für welche Einheitsvektoren $\vec{x}$ und $\vec{y}$ die Bilinearform *maximal* bzw. *minimal* ausfällt. Er drückt dies differentiell aus.

Es geht um die Extrema von $f(\vec{x},\vec{y}) = \vec{x}^T A\vec{y}$ unter den Nebenbedingungen $g_1(\vec{x},\vec{y}) = |\vec{x}|^2 - 1 = 0$, $g_2(\vec{x},\vec{y}) = |\vec{y}|^2 - 1 = 0$. Man kann sagen: ∇f, die Hauptänderungsrichtung von f, hat im Extremum keinen Anteil, der zu beiden Bedingungsflächen tangential ist, also auf beiden Bedingungs-Gradienten $\nabla g_1, \nabla g_2$ senkrecht steht. Also liegt ∇f in der von ∇g_1 und ∇g_2 aufgespannten Ebene.

D.h.: Mit gewissen Koeffizienten λ_1, λ_2 gilt

$$\nabla f(\vec{x},\vec{y}) = \lambda_1 \nabla g_1(\vec{x},\vec{y}) + \lambda_2 \nabla g_2(\vec{x},\vec{y})$$

für Extremumsstellen von f. Dies ist die *Lagrangesche Multiplikatorenregel*, und λ_1, λ_2 heißen Lagrange-Multiplikatoren.

Da im vorliegenden Fall $\nabla f = \begin{pmatrix} A\vec{y} \\ \vec{x}^T A \end{pmatrix}$, $\nabla g_1 = \begin{pmatrix} 2\vec{x} \\ \vec{0} \end{pmatrix}$, $\nabla g_2 = \begin{pmatrix} \vec{0} \\ 2\vec{y} \end{pmatrix}$, folgt mit $\sigma = 2\lambda_1$, $\tau = 2\lambda_2$: $\quad A\vec{y} = \sigma\vec{x}, \quad \vec{x}^T A = \tau\vec{y}^T. \quad (\star)$

Genau diese Beziehung erhält Camille Jordan durch seine zur hier formulierten im wesentlichen gleichwertige Differential-Überlegung. Also gilt an einer Maximumsstelle

$$f(\vec{x},\vec{y}) = \vec{x}^T(A\vec{y}) = (\vec{x}^T A)\vec{y} = \vec{x}^T\sigma\vec{x} = \tau\vec{y}^T\vec{y} = \sigma = \tau\,.$$

Das Gleichungssystem $(\star)$, das die *Maximumsstelle* von f erfüllt, lautet daher, mit $\sigma = f(\vec{u},\vec{v}) = \max_{|\vec{x}|=|\vec{y}|=1} f(\vec{x},\vec{y})$:

$$\left.\begin{aligned} -\sigma\vec{u} + A\vec{v} &= \vec{0} \\ A^T\vec{u} - \sigma\vec{v} &= \vec{0} \end{aligned}\right\} \;\Leftrightarrow\; \begin{pmatrix} -\sigma I & A \\ A^T & -\sigma I \end{pmatrix}\begin{pmatrix} \vec{u} \\ \vec{v} \end{pmatrix} = \vec{0} \quad (\star\star)$$

Lösungen $\neq \vec{0}$ von $(\star\star)$ mit $\sigma \neq 0$ erfüllen $|\vec{u}| = |\vec{v}|$, da $\vec{u}^T A\vec{v} = \vec{v}^T A^T\vec{u}$. Da es eine Lösung mit $|\vec{u}| = |\vec{v}| = 1$ *gibt*, verschwindet die Determinante, d.h. der Maximalwert σ ist *Eigenwert* der *symmetrischen* Matrix $B := \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$. Alle Eigenwerte von B sind reell; der maximale ist das gesuchte Maximum der Bilinearform.

Sei nun $\sigma_1 = \max_{|\vec{x}|=|\vec{y}|=1} f(\vec{x},\vec{y})$, und $(\vec{u}_1, \vec{v}_1)$ sei die (bzw. eine) Stelle, an der dieses Maximum angenommen wird, so dass also $A\vec{v}_1 = \sigma_1\vec{u}_1$, $A^T\vec{u}_1 = \sigma_1\vec{v}_1$. Ergänzt man zu ONBs $(\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_n) = U$ und $(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n) = V$, folgt mit $A\vec{v}_1 = \sigma_1\vec{u}_1$, $\vec{u}_1^T A = \sigma_1\vec{v}_1^T$:

$U^T A V = \left(\begin{array}{c|c} \sigma_1 & 0 \;\cdots\; 0 \\ \hline 0 & \\ \vdots & A_1 \\ 0 & \end{array}\right)$, da $(\vec{u}_1^T A)\vec{v}_k = 0\,(k>1)$, $\vec{u}_i^T(A\vec{v}_1) = 0\,(i>1)$.

Mit $\xi_i = \vec{u}_i^T\vec{x}, \eta_i = \vec{v}_i^T\vec{y}$ folgt $f(\vec{x},\vec{y}) = \sigma_1\xi_1\eta_1 + (\xi_2, \ldots, \xi_n)A_1\begin{pmatrix} \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}$,

da $\vec{x}^T A\vec{y} = \vec{\xi}^T U^T A V\vec{\eta}$. Nun dieselbe Überlegung zu A_1. So kann man schrittweise die Bilinearform bzw. die Matrix *diagonalisieren*.

Nebenbemerkung zur symmetrischen Matrix B: Wegen

$$\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}\begin{pmatrix} \vec{u} \\ \vec{v} \end{pmatrix} = \sigma\begin{pmatrix} \vec{u} \\ \vec{v} \end{pmatrix} \quad\Leftrightarrow\quad \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}\begin{pmatrix} \vec{u} \\ -\vec{v} \end{pmatrix} = -\sigma\begin{pmatrix} \vec{u} \\ -\vec{v} \end{pmatrix}$$

ist mit σ stets auch $-\sigma$ ein Eigenwert. Das charakteristische Polynom $\det(B - \sigma I)$ ist ein *gerades* Polynom.

Nachwort

Früher, im Rahmen des Lineare-Algebra-Teils meiner einführenden Mathematik-Vorlesungen für Informatiker an der TH Nürnberg, widmete ich etwa eine Vorlesungs-Doppelstunde der SVD.

Allerdings blieb es immer nur *fakultativer* Stoff; es gab ja stets Parallelveranstaltungen mit gemeinsamer Klausur, und die anderen Dozenten waren, so damals mein Eindruck, nicht im selben Maße wie ich von der Wichtigkeit der SVD überzeugt.

Dann hielt ich 2010 bei einer Fortbildungstagung für Gymnasiallehrer an der TH N (mit bundesweiter Teilnehmerschaft) einen Vortrag über die SVD. Dieser Vortrag, später erheblich ausgeweitet, war die Grundlage für den vorliegenden Text.

Im Sommer 2022, nach der Zeit der Prüfungsklausuren, habe ich endlich auch den Abschnitt über die „SVD in Hilberträumen“ verfasst. Dabei versuchte ich, *möglichst elementar und direkt* und auch möglichst fast genauso wie im endlichdimensionalen Fall den Satz über die Singulärwertzerlegung kompakter Operatoren herzuleiten, so dass ein guter Student mit passablen Lineare-Algebra-Kenntnissen es hoffentlich verstehen kann.

Wichtig war mir, die SVD nicht als Anhängsel zur Eigenwerttheorie bzw. zur Spektraltheorie symmetrischer (selbstadjungierter) Operatoren darzustellen. Letzteres stört mich auch bei vielen Darstellungen schon im Matrizen-Fall.

Die Singulärwertzerlegung *ist kein Anhängsel* der Eigenwertlehre! Sondern ein *eigenständiger* Denkansatz bei der Analyse linearer Abbildungen, im endlichdimensionalen wie auch im – allerdings nur kompakten – unendlichdimensionalen Fall.

Und das sollte nach meiner Überzeugung auch durch die Art, wie es entwickelt wird, heraussgestellt werden. Selbstverständlich muss man auch auf die Zusammenhänge mit dem symmetrischen Eigenwertproblem hinweisen, aber erst an zweiter Stelle.

Die SVD ist, denke ich, die *aussagefähigste* Matrix-Zerlegung. Auf einen Schlag sagt sie eigentlich alles aus über die Wirkung einer (endlichdimensionalen) linearen Abbildung und macht auch die Lösungsgesamtheit linearer Gleichungssysteme völlig transparent. Die *Pseudoinverse* ergibt sich dabei im Handumdrehen, ohne jeden Aufwand.

Ich wundere mich daher, dass z.B. in dem *exzellenten* Buch von Theo Bröcker *Lineare Algebra und Analytische Geometrie* (2003) die SVD nur als Übungsaufgabe auf S. 265 vorkommt.

Theo Bröckers Buch ist bemerkenswert durch die ungewöhnliche und vielseitige Stoffauswahl und die Berücksichtigung anspruchsvoller Anwendungen. Wieso wird im Kapitel 4 („Bilinearformen“), in dem euklidische und unitäre Räume sowie die Hauptachsentransformation behandelt werden, nicht wenigstens auf ein, zwei Seiten die SVD dargestellt?

Das Buch gleichen Titels von Max Koecher (2. Aufl. 1985), das Bröcker ausdrücklich als eine seiner Quellen nennt, widmet genau eine Seite der SVD, behandelt sie ganz knapp unter ferner liefen, nennt sie aber „orthogonale Normalform“. Der Begriff „Singulärwert“ kommt *nicht* vor. Was soll das? Wo es doch eine international übliche Begrifflichkeit 1985 längst gab. Jede Menge Forschungsergebnisse über Singulärwerte gab’s da auch schon, zum Teil verbunden mit prominenten Namen.

Man kann nicht erwarten, dass Lernende, die dies nur unter ferner liefen lesen, von selbst *ohne jede Anleitung* merken, wie fruchtbar und aussagefähig diese Matrix-Faktorisierung ist.

Aber auch das Buch *Linear Algebra and Its Applications, 2nd. edition* (2007) von Peter D. Lax, ebenfalls durch seine außergewöhnliche Stoffauswahl beeindruckend, behandelt die SVD nur knapp und auf wenig erhellende Weise ganz am Rande. Mag sein, dass die SVD aus seiner Sicht zu sehr nur die einfachsten Aspekte von linearen Abbildungen und Gleichungssystemen betrifft und er mehr die „höhere Warte“ im Sinn hatte. Der Bedeutung angemessen finde ich die Darstellung der SVD trotzdem *nicht*.

Dass ein *älteres* verbreitetes Lehrbuch wie Serge Langs *Linear Algebra* nichts über die SVD enthält, ist verständlich, da es im wesentlichen in den 1960er Jahren entstanden ist; meine Ausgabe ist die 2. Aufl. von 1970. Aber immerhin die Polarzerlegung quadratischer Matrizen kommt vor, als Übungsaufgabe. Die fundamentale *Schur-Zerlegung* kommt auch vor, wird aber nicht so genannt; Lang spricht nur von „Fan-Basen“.

Weil Gene Golub ein so begeisterter Verfechter der SVD und zugleich ein „angewandter“ und numerischer Mathematiker ist, heißt das nicht, dass die SVD *nur* oder hauptsächlich was für Numeriker ist.

Im Gegenteil: Aus logisch-inhaltlichen Gründen sollte sie meines Erachtens in *jedem* einführenden Lineare-Algebra-Kurs einen wichtigen Platz einnehmen. Dass sie implizit Kompaktheitseigenschaften nutzt, ist kein Gegenargument, da die endlichdimensionalen reellen und komplexen Vektorräume und deren unzählige Anwendungen einen großen Teil der Bedeutung der Linearen Algebra ausmachen.

Natürlich macht Vektorraum- und Matrizentheorie z.B. auch über dem Restklassenkörper $\mathbb{Z}_p$ Sinn, und da tut man sich mit der euklidischen Norm und Orthogonalität schwer. Aber in beinahe allen endlichdimensionalen Vektorräumen *hat* man, sofern man will, eine euklidische Metrik zur Verfügung – und damit auch die SVD als grundlegende Faktorisierung linearer Abbildungen.

Es ist eine historische *Kuriosität*, dass es so lange dauerte, bis die SVD, obwohl schon von Beltrami und Jordan auf bis heute vorbildliche Weise hergeleitet, endlich zum gedanklichen Gemeingut wurde – zumindest in der angelsächsischen Welt. Dass es hierzulande noch deutlich *länger* dauert(e), finde ich nicht kurios, sondern eigentlich unbegreiflich.

EMEW, Nürnberg, Sommer 2022

Abschließend ein kleines Zitat:

Wir wollen Ihnen unsere Meinung ganz offen sagen. Die Singulärwertzerlegung stellt den Höhepunkt dieses Kurses über lineare Algebra dar.

Gilbert Strangs in [4] auf Seite 373 formulierte Meinung
(deutsche Ausgabe, Springer 2003, ca. 650 Seiten)

Die *Schur-Zerlegung* (von Issai Schur (Math. Ann. 1909) angegeben, aber Schur selbst verweist auf L. Stickelberger (Polyt. Schule Zürich 1877)), hier auf S. 33f. behandelt, könnte man als zumindest matrizen*theoretisch* noch fundamentaler ansehen als die SVD: Elegant und einfach per Deflation zu beweisen (nichts als die *Existenz* von Eigenwerten ist vorauszusetzen), folgt aus ihr ganz leicht der Spektralsatz für normale Matrizen und damit die SVD; ebenso folgen viele andere Aussagen, etwa der Satz von Hamilton/Cayley. In Strangs Buch [4], das die SVD so mustergültig darstellt, kommt die Schur-Zerlegung *nicht* vor, wohl, weil er sie für ein eher theoretisches Werkzeug hält und sein Fokus auf den *Anwendungen* der Linearen Algebra liegt. Bemerkenswert, dass schon Camille Jordans SVD-Beweis die Technik der Deflation benutzt.

SINGULÄRWERTZERLEGUNG mit SCILAB

Scilab 6.1.1

Additional information about Scilab is available at:

https://www.scilab.org

Eine (kleine) Matrix von Hand eingeben:

```
--> A=[1 2 3 4; 2 0 -7 1; 1 1 1 0]
A =
   1.   2.   3.   4.
   2.   0.  -7.   1.
   1.   1.   1.   0.
```

Die Singulärwertzerlegung (**svd** = **s**ingular **v**alue **d**ecomposition):

```
--> svd(A)
 ans  =
   7.8783275
   4.8246252
   1.2864478
```

Das ist nur der *Vektor* der Singulärwerte. (Achtung: Keine *Eigen*werte! Die gibt's bei einer nicht *quadratischen* Matrix nicht; eine Singulärwertzerlegung gibt es *immer*!)

Man kann aber auch die *vollständige* Singulärwertzerlegung **A = U * S * V'** ermitteln (V' ist die zu V *adjungierte*, d.h. im reellen Fall die *transponierte* Matrix; also A*V = U*S , da V' die zu V *inverse* Matrix ist).

```
--> [U S V]=svd(A)
U  =
  -0.4389633  -0.8789483  -0.1864432
   0.8904604  -0.4532693   0.0403388
  -0.1199647  -0.148313    0.9816373
S  =
   7.8783275   0.          0.          0.
   0.          4.8246252   0.          0.
   0.          0.          1.2864478   0.
V  =
   0.1551081  -0.4008187   0.6808452  -0.5930731
  -0.1266628  -0.3951      0.473203    0.7771302
  -0.9735667   0.0803642   0.108777   -0.1840572
  -0.1098447  -0.8226676  -0.5483579  -0.102254
```

Nun die Probe!


```
--> U*S*V'  // Probe!
ans  =
   1.   2.          3.   4.
   2.  -3.365D-16  -7.   1.
   1.   1.          1.  -1.110D-16

--> clean(ans)
ans =
   1.   2.   3.   4.
   2.   0.  -7.   1.
   1.   1.   1.   0.
```

Bis auf Rundungsfehler ergibt sich die Ausgangsmatrix.

Die *Spalten* der Matrix V :

```
--> v1=V(:,1), v2=V(:,2), v3=V(:,3), v4=V(:,4)
v1  =
   0.1551081
  -0.1266628
  -0.9735667
  -0.1098447

v2  =
  -0.4008187
  -0.3951
   0.0803642
  -0.8226676

v3  =
   0.6808452
   0.473203
   0.108777
  -0.5483579

v4  =
  -0.5930731
   0.7771302
  -0.1840572
  -0.102254
```

(*Zeilen:* V(1,:), V(2,:), usw.)

Zugriff auf ein Matrixelement `M(i, k)`:

```
--> S(1,1)
ans = 7.8783275
```

Nun A v_i mit den Spalten u_i von U vergleichen!

```
--> U
U =
  -0.4389633  -0.8789483  -0.1864432
   0.8904604  -0.4532693   0.0403388
  -0.1199647  -0.148313    0.9816373

--> A*v1 / S(1,1)
ans =
  -0.4389633
   0.8904604
  -0.1199647

--> A*v2 / S(2,2)
ans =
  -0.8789483
  -0.4532693
  -0.148313

--> A*v3 / S(3,3)
ans =
  -0.1864432
   0.0403388
   0.9816373

--> A*v4
ans =
   2.220D-16
   2.220D-16
   4.441D-16
```

(Ein *Nullvektor*, bis auf Rundungsfehler.)

```
--> B=A'*A
B =
   6.    3.  -10.   6.
   3.    5.   7.    8.
  -10.   7.   59.   5.
   6.    8.   5.    17.

--> det(B)
ans = 3.375D-13
```

Trotz *ganzzahliger* Matrix **B** treten – wie man sieht - bei der *numerischen* Determinantenberechnung Rundungsfehler auf.

(Da bei **A** nur drei Spalten linear unabhängig sind, sind auch bei **B=A'*A** höchstens drei Spalten linear unabhängig. Also hat **B** *nicht* den Rang 4, und daher `det(B)=0`.)


```
--> [M Z N] =svd(B)
 M  =
  -0.1551081  -0.4008187  -0.6808452   0.5930731
   0.1266628  -0.3951     -0.473203   -0.7771302
   0.9735667   0.0803642  -0.108777    0.1840572
   0.1098447  -0.8226676   0.5483579   0.102254
 Z  =
   62.068043   0.          0.         0.
   0.          23.277009   0.         0.
   0.          0.          1.654948   0.
   0.          0.          0.         1.670D-16
 N  =
  -0.1551081  -0.4008187  -0.6808452   0.5930731
   0.1266628  -0.3951     -0.473203   -0.7771302
   0.9735667   0.0803642  -0.108777    0.1840572
   0.1098447  -0.8226676   0.5483579   0.102254
```

Der vierte Singulärwert von **B** ist gleich 0.

```
--> B*N(:,4)    //  Ein Nullvektor !
ans =
  -5.551D-15
  -3.331D-15
   6.106D-15
  -4.885D-15
```

Mit **rand(m,n)** erzeugt man eine (m,n)-Matrix aus gleichverteilten Zufallszahlen zwischen 0 und 1.

```
--> C=10*rand(1000,10); // Hier MUSS die Ausgabe unterdrückt werden!
```

(**;** am Befehlsende verhindert die Anzeige des Resultats.)

```
--> svd(C)
ans =
   509.40610
   99.473987
   95.668827
   94.305278
   92.699693
   90.369592
   87.745184
   87.471793
   86.103673
   84.829436
```

Die Matrix C hat nur den Rang 10, da nur 10-zeilig. Auffallend: Ein Singulärwert ist der mit Abstand größte.

Nun eine noch viel größere Matrix.

```
--> D=10*rand(100000,10);  // D hat 10^6 Elemente bei nur 10 Spalten

--> svd(D)
ans =
   5079.3433
   919.63026
   918.86132
   916.11504
   912.77164
   911.58574
   910.51520
   910.04715
   907.15200
   906.32561

--> E=10*rand(100000,10)-10*rand(100000,10);

--> max(E)  // Unterschied zu Octave, wo man max(max(E)) bilden muss
ans =
   9.9855307

--> min(E)
ans =
  -9.9997768

--> sum(E)
ans =
  -6634.6721

--> mean(E)      // Mittelwert aller Matrixelemente
ans =
  -0.0066347

--> svd(E)
ans =
   1300.7722
   1300.2751
   1295.9286
   1295.4121
   1292.1966
   1290.9977
   1289.7511
   1285.5160
   1284.1762
   1279.2907
```


Nun ein “Graustufen-Bild” in VGA-Größe:

```
--> Bild0=floor(256*rand(480,640));
```

Ein kleiner Ausschnitt der Bildmatrix:

```
--> Bild0(101:110,341:350)
ans =
   238.   200.   132.   104.   212.   64.    59.    107.   145.   132.
   189.   225.   29.    129.   88.    34.    152.   171.   133.   134.
   127.   55.    33.    37.    3.     83.    133.   224.   132.   105.
   160.   134.   41.    18.    192.   171.   66.    237.   11.    244.
   206.   175.   244.   155.   55.    244.   151.   125.   153.   119.
   148.   217.   91.    123.   85.    185.   50.    2.     35.    71.
   124.   35.    250.   166.   35.    199.   62.    142.   18.    127.
   176.   210.   52.    23.    51.    186.   83.    32.    184.   45.
   29.    149.   223.   252.   7.     47.    191.   23.    178.   32.
   234.   117.   4.     90.    115.   129.   239.   98.    178.   244.

--> s_bild0=svd(Bild0);

--> s_bild0(1:10)
ans =
   70726.571
   3445.8155
   3400.9380
   3385.6293
   3363.8518
   3348.7005
   3325.9166
   3313.1862
   3305.9107
   3289.8747
```

Einen Überblick über den Verlauf der Singulärwerte erhält man durch eine Stichproben-Zusammenstellung, bei der auch die letzten Singulärwerte angezeigt werden. Die letzten Singulärwerte sind zwar viel kleiner als der führende Wert, aber längst nicht nahezu 0:

```
--> [s_bild0(1:10) s_bild0(101:110) s_bild0(301:310) s_bild0(471:480)]
ans =
   70726.571   2471.7797   1284.7659   333.15463
   3445.8155   2461.0744   1281.0956   328.03722
   3400.938    2457.0828   1274.4654   320.82294
   3385.6293   2451.6813   1270.9191   315.59797
   3363.8518   2446.2404   1260.7974   302.80878
   3348.7005   2444.764    1260.337    300.26763
   3325.9166   2436.4112   1249.4385   297.93318
   3313.1862   2432.3588   1245.3615   292.55801
   3305.9107   2427.4315   1237.1913   271.0859
   3289.8747   2415.8031   1234.4004   261.78102
```

Bei einem *echten*, nicht aus Zufallszahlen bestehenden Graustufen-Bild klingen die Singulärwerte viel *stärker* ab. Das ergibt eine Methode der **Bilddatenkompression**: Nur die großen Singulärwerte $s_1, \ldots, s_k$ und die entsprechenden U- und V-Spalten werden verwendet. Das Bild kann aus der **rangreduzierten** Graustufenmatrix $G = (s_1\mathbf{u}_1, s_2\mathbf{u}_2, \ldots, s_k\mathbf{u}_k)\,(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)^T$ ohne großen Qualitätsverlust rekonstruiert werden. Weiter unten Beispiele dazu.

So (*anders* als bei Octave) speichert man die Matrix **Bild0** als Textdatei ab:

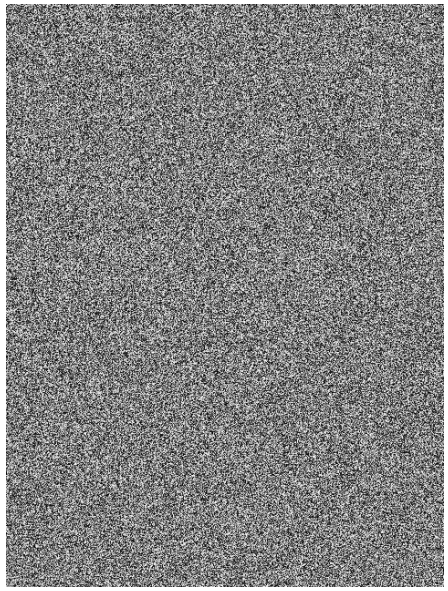
```
--> fprintfMat('/home/emew/Dokumente/SCILAB/Work/SCI_Bild0.txt',Bild0,'%3.0f');
```

Dabei ist das dritte Argument, der Scilab-String **'%3.0f'** eine C-typische Format-Angabe und spezifiziert: dreistellig ohne Nachkomma-Stellen.
(Mit **format(n)** kann man einfach das Zahlen-*Anzeige*format steuern.)

Aus der abgespeicherten Datei machen wir nun ein **PGM-Bild** (Portable Graymap), indem wir mit einem Editor (z.B. **kate** oder **Emacs** unter Linux) einen **pgm-Header** hinzufügen (Breite 480 Pixel, Höhe 640 Pixel):

```
P2
# Created by Scilab
480 640
255
 14 209 163 205  51  78  48 133 142  37 189  96 177 234
```

Die danach als SCI_Bild0.pgm abgespeicherte Datei kann man mit einem Bildbetrachter ansehen (auch mit dem *Editor* Emacs selbst, denn da kann man zwischen ASCII- und Bild-Ansicht umschalten). Man kann sie auch *als Bild* hier in diesen LibreOffice-Text laden. Das sieht dann so aus:



Von *Zufalls*-Graustufen kann man nichts anderes erwarten...

Umgekehrt importiert man eine grosse Zahlen-Textdatei (ASCII) als Matrix **OHMIM** in die aktuelle SCILAB-Session durch folgenden Befehl:

```
--> OHMIM=fscanfMat('/home/emew/Dokumente/SCILAB/Work/Ohm_22_1z.txt');
```

Das die Ausgabe auf dem Bildschirm verhindernde **;** am Ende des Lade-Befehls ist hier SEHR wichtig!

„**Ohm_22_1z.txt**“ ist eine folgendermaßen erzeugte *einzeilige* Textdatei:

Ein Jpeg-Bild von Georg Simon Ohm wurde nach Grayscale konvertiert und als PGM-Datei abgespeichert (ImageMagick **convert**). Zunächst im Binärformat, aber per **pnmtopnm -plain infile.pgm > outfile.pgm** ins *ASCII-Format* umgewandelt; auch **convert** mit der Option **-compress none** wäre möglich. Mit dem Befehl **tr -d '\n' < infile.txt > outfile.txt** wurden die Zeilenumbrüche weggenommen und damit in eine einzige Zeile umformatiert; *vorher* wurde der PGM-Header entfernt und die Datei als „infile.txt” abgespeichert.

Diese Linux-Kommandozeilen-Befehle sind schnell und leistungsfähig. Man kann zwar auch z.B. mit *kate* Zeilen zusammenführen; mit so großen und später *noch größeren* Dateien tut sich eine grafische Benutzer-Schnittstelle aber schwerer.

(Das Bild ist einige Seiten später dargestellt.)

Unter Windows eignet sich z.B. das exzellente kostenlose Programm **irfanview** zum Umwandeln der Bilddateien in eine Graustufen-Textdatei.

Nun wird OHMIM bearbeitet, zunächst entsprechend seiner ursprünglichen Pixelzahl umskaliert. Da die Matrix *spalten*weise aus der einzeiligen Matrix aufgebaut wird, der Einzeiler aber durch Aneinanderfügen der *Zeilen* des Bildes entstand, muss man zunächst das Format (623, 866) wählen und dann transponieren.

```
--> size(OHMIM)
 ans  =
    1.    539518.
--> OHMIM=matrix(OHMIM,623,866)';
--> size(OHMIM)
 ans  =
    866.    623.
--> s_ohm=svd(OHMIM);
```

Wir schauen uns einige Singulärwerte von OHMIM an:

```
--> s_ohm=svd(OHMIM); // Wieder wichtig: ; am Ende nicht vergessen!

--> size(s_ohm)
ans =
   623.    1.
```

```
--> [s_ohm(1:10) s_ohm(101:110) s_ohm(201:210) s_ohm(401:410) s_ohm(614:623)]
ans =
   118311.89    757.30549    428.27865    108.89484    2.6911775
   29475.996    754.08316    426.57152    107.176      2.1604555
   24193.949    751.11841    425.90071    105.93624    1.663969
   13904.273    745.13701    422.94707    105.61096    7.184D-11
   11173.183    741.75046    422.15175    103.91132    3.657D-11
   8822.0007    736.52445    419.39224    102.26169    1.668D-11
   8297.8676    732.99032    417.00282    101.90654    9.637D-12
   6333.7484    731.60016    415.43045    101.12241    8.391D-12
   6108.8544    725.28057    412.97372    99.780722    6.896D-12
   5515.8651    720.71703    411.47767    98.591829    5.404D-12
```

Man erkennt, dass der erste Singulärwert sehr groß ist (> 118 300) und sechs weitere im Bereich von 30 000 bis ca. 8 000 liegen; aber schon ab Index 200 (sogar schon ab ca. 175) liegen die Werte unter 500, und am Ende sind sie vernachlässigbar klein, die letzten sieben sind sogar = 0 (nur Rundungsfehler).

Darauf beruht die Kompression mittels **bester Rang-k-Approximation**, bezogen auf die **Frobenius-Norm** $\left(||A||_F = \sum |a_{i\,k}|^2\right)^{1/2}$; diese und der fundamentale *Rang*-Begriff gehen zurück auf F. Georg Frobenius.

Die Scilab-Funktion **`sva(A,k)`** liefert diese Approximation. Im Bereich lauter verschiedener Singulärwerte ist die approximierende Matrix *eindeutig bestimmt. Nur* Approximierende vom Typ $\sum_{i=1}^{k} s_i \mathbf{u}_i\, \mathbf{v}_i^T$ sind beste Rang-k-Approximationen *im Sinne der F(robenius)-Norm*. Sie sind auch beste Rang-k-Approximationen bzgl. der induzierten euklidischen Matrixnorm; aber *diese* Eigenschaft ist *unschärfer* als die bezüglich der F-Norm.

```
--> [U,S,V]=sva(OHMIM,312);

--> OHMIM_50=U*S*V';
--> rest_50=svd(OHMIM-OHMIM_50);
--> rest_50(1)
ans = 217.60047
```

Der größte Singulärwert ist die *induzierte euklidische Norm* einer Matrix; bei **`OHMIM`** ist diese Norm ca. 118300, bei **`OHMIM-OHMIM_50`** nur noch knapp 220.

```
--> max(OHMIM_50)
ans = 266.40658

--> min(OHMIM_50)
ans = -9.0768670
```

Als *Graustufen*-Matrix muss die Rang312-Approximation auf *ganzzahlige* Werte zwischen 0 und 255 getrimmt werden:

```
--> o50=OHMIM_50; HI=255*ones(o50);LO=zeros(o50);
--> o50=round(max(min(o50,HI),LO));
```



```
--> [max(o50), min(o50); size(o50)]  // Jetzt passt's!
ans =
   255.   0.
   866.   623.
```

Die rangreduzierte und getrimmte Matrix wird nun als Datei abgespeichert, dann mit einem PGM-Header versehen und kann danach als Bild betrachtet und in andere Bildformate konvertiert werden.

Hier geht es aber nur darum, die Datei als *Bild* sichtbar zu machen, um zu sehen, dass wir nur einen *Bruchteil* der Spalten von U, S und V benötigen, um das Bild fast genau zu rekonstruieren.

```
--> fprintfMat('/home/emew/Dokumente/SCILAB/Work/OHM_50.txt',o50,'%3.0f');
```

Der Anfang nach Hinzufügen eines PGM-Headers:

```
P2
623 866
255
# Created by Scilab 6.1.1, Mo, Aug 29, 2022
255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255
```

Diese Datei kann von Bildbetrachtern als Bild dargestellt und auch in ein LibreOffice-Dokument importiert werden (siehe weiter unten).

```
--> [U,S,V]=sva(OHMIM,156); o25=U*S*V';
    [U,S,V]=sva(OHMIM,62); o10=U*S*V';
    [U,S,V]=sva(OHMIM,31); o5=U*S*V';
--> o25=round(max(min(o25,HI),LO));
--> fprintfMat('/home/emew/Dokumente/SCILAB/Work/OHM_25.txt',o25,'%3.0f');
--> o10=round(max(min(o10,HI),LO));
--> fprintfMat('/home/emew/Dokumente/SCILAB/Work/OHM_10.txt',o10,'%3.0f');
--> o5=round(max(min(o5,HI),LO));
--> fprintfMat('/home/emew/Dokumente/SCILAB/Work/OHM_5.txt',o5,'%3.0f');
```

Auf den folgenden fünf Seiten sind das Original-Bild sowie die mit 50, 25, 10 bzw. 5 Prozent der V-Spalten und ebenso vielen U-Spalten erzeugten Bilder reproduziert.

Die **JPEG-Kompression** arbeitet zwar mit *Fourier*-Methoden; es ist aber eine prinzipiell ähnliche **Reduktion der Datenmenge durch spektrale Analyse**. Während es bei den Singulärwerten um die Rangfolge der Streckungsfaktoren einer Matrix/linearen Abbildung geht, gemessen mit der euklidischen Norm, zerlegt die Fourier-Zerlegung das Bild in zuerst langwellige und dann immer kurzwelligere Hell-Dunkel-Schwingungen (im Schwarz-Weiß-Fall). Und auch die **Wavelet-Transformation** wird auf analoge Weise zur Bilddaten-Kompression angewandt.

Eine wichtige andere Anwendung der Rangreduktion:

Information Retrieval. Dabei wird etwa ein Literaturbestand durch eine große Matrix dargestellt, deren Spalten die einzelnen Dokumente repräsentieren; die einzelnen Einträge stehen für die verschiedenen wesentlichen Inhalte/Themen der einzelnen Dokumente. Eine Anfrage wird entsprechend durch einen Anfrage-Vektor repräsentiert. Durch Skalarprodukt-Bildung wird die Anfrage mit den Dokumenten verglichen. Um den Rechenaufwand in Grenzen zu halten (sprich: eine hohe Abfrage-Performance zu erzielen), vor allem aber auch, um störendes Rauschen in Form begrifflicher Fehlzuordnungen (repräsentiert durch *kleine* Singulärwerte) zu eliminieren, arbeitet man mit einer optimal rangreduzierten Dokumenten-Begriffs-Matrix. Da sich der Bestand ständig ändert, braucht man auch effiziente **Update-Techniken**.

Ein Pionier, der seit den 1990er Jahren einiges zu diesen linearalgebraischen Rechenmethoden des **IR** publiziert hat, ist M. W. Berry. Ein einführendes kleines Buch zum Thema:

M.W.Berry/M.Browne, *Understanding Search Engines – Mathematical Modeling and Text Retrieval, 2nd edition*, SIAM 2005

Auch das Prinzip von **PageRank**, des Webseiten-Bewertungs-Algorithmus der Google-Gründer S. Brin und L. Page, wird in dem Büchlein kurz erläutert. --- Nun folgen die Ohm-Bilder.

Das Original-Bild

SVD-Kompression mit 312 (50% der 623) V-Spalten und 312 U-Spalten

SVD-Kompression mit 156 (25%) V-Spalten und 156 U-Spalten

SVD-Kompression mit 62 (10%) V-Spalten und 62 U-Spalten

SVD-Kompression mit 31 (5%) V-Spalten und 31 U-Spalten

Es wird noch ein *zweites* Bild ausprobiert, die Fotografie einer Frühlingsszene im Englischen Garten mit dem Monopteros.

Da das Pixelformat des Fotos 3341x2212=7 390 292 Pixel umfasst, muss Scilab hier mit *großen* Matrizen umgehen. Kein Problem!

Scilab ist leistungsfähig genug, um auf einem normalen heutigen PC auch „Real-World“-Probleme zu bewältigen.

Da nur dieselben Arbeitsschritte, die beim Ohm-Bild im einzelnen erläutert wurden, auf eine andere Ausgangs-Bilddatei anzuwenden sind, werden einfach alle SCILAB-Befehle kompakt aneinandergereiht und *zu einer Skript-Datei zusammengefasst*.

Anschließend sind dann nur noch mit einem Text-Editor (etwa *kate*) die PGM-Header hinzuzufügen. Auch dies wurde weiter oben beim Ohm-Beispiel erläutert. Mit ImageMagick display werden die Bilder abschließend aufgerufen und erneut als PGM-Bild abgespeichert. Dann werden sie binär (P5-Format) und damit platzsparender gespeichert. (Geht alternativ z.B. mit Gwenview, unter Linux.)

```
--> MONOP=fscanfMat('/home/emew/Dokumente/SCILAB/Work/Monopt_plain_1z.txt');
--> size(MONOP)
ans =   1.   7390292.
--> MO=matrix(MONOP,3341,2212)';
```

Nun das kompakte Skript, das bei gegebener Grauwert-Matrix **MO** die Bild-Kompression durch Rangreduktion durchführt und die Textdateien exportiert.

```
HI=255*ones(MO);LO=zeros(MO);
[U S V]=sva(MO,1106);mo50=round(max(min(U*S*V',HI),LO));
[U S V]=sva(MO,553);mo25=round(max(min(U*S*V',HI),LO));
[U S V]=sva(MO,221);mo10=round(max(min(U*S*V',HI),LO));
[U S V]=sva(MO,111);mo5=round(max(min(U*S*V',HI),LO));
[U S V]=sva(MO,44);mo2=round(max(min(U*S*V',HI),LO));
[U S V]=sva(MO,22);mo1=round(max(min(U*S*V',HI),LO));
fprintfMat('/home/emew/Dokumente/SCILAB/Work/Mono_50.txt',mo50,'%3.0f');
fprintfMat('/home/emew/Dokumente/SCILAB/Work/Mono_25.txt',mo25,'%3.0f');
fprintfMat('/home/emew/Dokumente/SCILAB/Work/Mono_10.txt',mo10,'%3.0f');
fprintfMat('/home/emew/Dokumente/SCILAB/Work/Mono_5.txt',mo5,'%3.0f');
fprintfMat('/home/emew/Dokumente/SCILAB/Work/Mono_2.txt',mo2,'%3.0f');
fprintfMat('/home/emew/Dokumente/SCILAB/Work/Mono_1.txt',mo1,'%3.0f');
```

Das Pixelformat 3341x2212 (Breite mal Höhe) bringt es mit sich, dass SCILAB für die Rang-k-Approximationen, also die sva-Befehle, einige Minuten an Zeit braucht.

```
--> tic();exec('/home/emew/Dokumente/SCILAB/Work/MO_compress.sce');toc()
```

Auf meinem Standard-Fujitsu-PC aus dem Jahr 2015, 4xIntel Core i5-4590 CPU @ 3.30 GHz, 31.2 GiB Arbeitsspeicher, braucht Scilab ca. 6 Minuten, das Skript auszuführen.

```
--> s_mon=svd(MONOP); [min(s_mon(1:2211)-s_mon(2:2212)),max(s_mon),min(s_mon)]
ans =
  0.0078574   329425.8   10.694231
```

Man sieht: Alle Singulärwerte sind *verschieden,* also **svd(MO)** eindeutig.

Auf den folgenden Seiten die Bilder. (Man kann bei der PDF-Datei den Zoom-Faktor 400% wählen, um sich die Bilddetails ganz genau anzusehen.)

Das Original-Foto

Das 50%-Bild (Rang-1106-Approximation)

Das 25%-Bild (Rang-553-Approximation)

Das 10%-Bild (Rang-221-Approximation)

Das 5%-Bild (Rang-111-Approximation)

Das 2%-Bild (Rang-44-Approximation)

Das 1%-Bild (Rang-22-Approximation)

Bemerkenswert: Bis zu Mono_25 bleiben die Unterschiede zum Original kaum wahrnehmbar, und selbst Mono_5 ist noch brauchbar, auf übliche Papierbild-Größe verkleinert. Erst Mono_2 und Mono_1 sind *sehr* unscharf.

Selbst das 1%-Bild zeigt eigentlich noch erstaunlich viel von der ursprünglichen Struktur des Bildes.

Aber bei genauerem Hinsehen fallen auch schon bei Mono_10 deutliche Inhomogenitäten des Himmels und z.B. der unteren hellen Säulenpartien beim Monopteros auf.

Dies ist natürlich bei weitem keine umfassende mathematische Einführung in die Singulärwertzerlegung und ihre Anwendungen. Es sollte nur mit dem Hilfsmittel Scilab sozusagen *zu Fuß Schritt für Schritt* nachvollzogen werden, wie eine Bilddaten-Kompression funktioniert.

Es ging hier nur um eine *prinzipielle* Darstellung, nicht um eine praktisch konkurrenzfähige möglichst *effiziente* Implementation. Übrigens verfügt Scilab über Bildverarbeitungs-Toolboxen, die aber nicht benutzt wurden.

Man kann statt mit Scilab analog mit Octave oder auch Matlab arbeiten – mit *fast identischen* Befehlen. Aber warum sollte man eine teure kommerzielle Software wie Matlab nutzen? Wer sich mit Scilab oder Octave, beide kostenlos erhältlich, vertraut gemacht hat, wird bei Bedarf, etwa in einer industriellen Entwicklungsabteilung (wo insbesondere die Matlab-Ergänzung Simulink viel genutzt wird), ohne Mühe auch mit Matlab zurechtkommen.

Scilab besitzt anstelle von Simulink das Modellierungs-Tool Xcos (früher Scicos), auch sehr leistungsfähig.

Abschließend seien einige Grundeigenschaften und -anwendungen der Singulärwertzerlegung, der aus heutiger Sicht *theoretisch* wie *numerisch wichtigsten* allgemeinen Matrix-Faktorisierung, zur ersten Orientierung zusammengestellt:

- Numerisch robuste **Rangbestimmung.**
- Es gilt $r = \mathrm{rg}(A)$, $\ll \mathbf{u}_1,\ldots,\mathbf{u}_r \gg = \mathcal{R}(A)$, $\ll \mathbf{v}_{r+1},\ldots,\mathbf{v}_n \gg = \mathcal{N}(A)$, $\ll \mathbf{u}_{r+1},\ldots,\mathbf{u}_m \gg = \mathcal{N}(A^*) = \mathcal{R}(A)^\perp$, $\ll \mathbf{v}_1,\ldots,\mathbf{v}_r \gg = \mathcal{R}(A^*) = \mathcal{N}(A)^\perp$.
($\mathcal{R}(A)$=Bildraum, Spaltenraum; $\mathcal{N}(A)$=Nullraum, Kern; $A^* = \bar{A}^T$ = adjungierte Matrix.)
- Lösung von $|A\mathbf{x}-\mathbf{b}| \stackrel{!}{=} \min \ \wedge\ |\mathbf{x}| \stackrel{!}{=} \min$ und **Pseudo-Inverse** A^+ .
- **Beste Approximation** niedrigeren Ranges an eine Matrix.
- **Geometrische Charakterisierung** der Abbildung: Jede lineare Abbildung des Ranges r bildet die Einheitskugel eines gewissen r-dimensionalen Unterraumes $\mathcal{U}$ des Urbildes auf ein *Ellipsoid* eines gewissen r-dimensionalen Unterraumes des Bildes ab, eventuell mit Orientierungsumkehr; die Hauptachsen-*Urbilder* sind zueinander *orthogonal*, und das orthogonale Komplement $\mathcal{U}^\perp$ wird auf $\mathbf{0}$ abgebildet. SVD: A = Hauptachsen × Streckungsfaktoren × Urbilder.
- Analog zu derjenigen bei Eigenwerten symmetrischer Matrizen gilt für Singulärwerte beliebiger Matrizen eine charakteristische **Minimax-Beziehung**:

$$\sigma_k = \max_{\mathbf{v}\perp\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_{k-1}} \frac{|A\mathbf{v}|}{|\mathbf{v}|} = \min_{\dim\mathcal{U}=n-k+1} \max_{\mathbf{v}\in\mathcal{U}\setminus\{\mathbf{0}\}} \frac{|A\mathbf{v}|}{|\mathbf{v}|} = \max_{\dim\mathcal{U}=k} \min_{\mathbf{v}\in\mathcal{U}\setminus\{\mathbf{0}\}} \frac{|A\mathbf{v}|}{|\mathbf{v}|}$$

  Insbesondere ist σ_1 die induzierte *euklidische Norm* von A.
- Es gilt $\sum_{i=1}^{\min(m,n)} \big(\sigma_i(A)-\sigma_i(B)\big)^2 \le \|A-B\|_F$ für zwei Matrizen $A,B\in\mathbb{C}^{m\times n}$; dabei ist $\|A\|_F^2 = \sum_{i,k=1}^{n} |a_{ik}|^2$ die schon auf Seite 9 erwähnte *Frobenius-Norm*. Für einzelne Singulärwerte gilt $|\sigma_i(A)-\sigma_i(B)| \le \|A-B\|_2$ mit $\|A\|_2 = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{|A\mathbf{x}|}{|\mathbf{x}|} = \sigma_1(A)$.

Die SVD nebst vielen ergänzenden Aussagen und Anwendungen werden ausführlich behandelt in den folgenden Standardwerken:

G.H.Golub/Ch.F.van Loan, *Matrix Computations, 3rd ed.*, Johns Hopkins University Press 1996
R.A.Horn/Ch.R.Johnson, *Matrix Analysis , 2nd ed.*, Cambridge University Press 2013
R.A.Horn/Ch.R.Johnson, *Topics in Matrix Analysis*, Cambridge University Press 1991/94

Diese Bücher, denke ich, *muss* kennen, wer sich *intensiv* mit der SVD und überhaupt mit der Matrizen-Mathematik vertraut machen will. Ebenfalls exzellent:

G.Strang, *Lineare Algebra, Springer* 2003;
G.Strang, *Linear Algebra and Its Applications, 3rd ed.*, Hartcourt Brace Jovanovich 1988

Das zweitgenannte Buch ist ein Klassiker.

Gene Golub ist Mitautor des wohl gebräuchlichsten Verfahrens zur *Berechnung* der Singulärwertzerlegung, des *Algorithmus von Golub und Reinsch*.

Noch ein *drittes* Bild auf den nächsten Seiten. Dazu wurde per Irfanview ein Graustufenbild erzeugt und dann analog weiterverarbeitet. Das Scilab-Skript **`MO_compress.sce`** brauchte nur *geringfügig* abgeändert zu werden.


```
--> TOW=fscanfMat('/home/emew/Dokumente/SCILAB/Work/Tower_Bridge_1z.txt');
--> size(TOW)
ans  =  1.   10036224.    // Das Bild hat rund 10 Megapixel
--> TO=matrix(TOW,2592,3872)';
--> tic();exec('/home/emew/Dokumente/SCILAB/Work/TO_compress.sce');toc()
ans  =  497.26854   // Die Ausführungszeit des Skripts in Sekunden
```

50%-Bild (Rang 1296)

25%-Bild (Rang 648)

10%-Bild (Rang 259)

5%-Bild (Rang 130)

Das 2%-Bild (Rang 52)

Das 1%-Bild (Rang 26)

Auch hier ist das 25%-Bild kaum sichtbar beeinträchtigt und das 10%-Bild noch passabel (trotz schon erkennbarer Inhomogenitäten am Himmel).

Für Scilab gibt's eine Bildverarbeitungs-Toolbox, auf deren Verwendung aber hier verzichtet wurde. So sieht man, wie Bild-Dateien auch mit Scilab-*Grundfunktionen* im- und exportiert werden können.

Erstellt mit **LibreOffice** inklusive dem Formel-Editor **TexMaths**