Singulärwertzerlegung

Birte Drewes

26. Juni 2002

Proseminar Numerische Mathematik SoSe 2002
bei Prof. Dr. J. Struckmeier

0 Einleitung

Der Titel des Vortrags lautet „Singulärwertzerlegung“. Kurz gefaßt versteht man darunter die Transformation einer Matrix auf eine Diagonalform.
Der Vortrag behandelt das Thema in zwei Schritten: Der erste Abschnitt umfaßt die mathematischen Grundlagen der Singulärwertzerlegung, d.h. deren Existenz und Eindeutigkeit. Im zweiten Teil wird ein numerisches Verfahren zur Berechnung der Singulärwertzerlegung vorgestellt.
Das Skript steht im Internet zum Download bereit unter
$\longrightarrow$ `www.bonu.de`

1 Mathematische Grundlagen

1.1 Existenz einer Singulärwertzerlegung

Die formale Definition der Singulärwertzerlegung [engl: *singular value decomposition*] ergibt sich aus dem folgenden

Satz[1]**.** *Gegeben sei eine Matrix $A \in \mathbb{R}^{m\times n}$. Dann existieren zwei orthogonale Matrizen*

$$U \in \mathbb{R}^{m\times m} \quad \text{und} \quad V \in \mathbb{R}^{n\times n},$$

so daß

$$U^T A V = \operatorname{diag}(\sigma_1, ..., \sigma_{\min(m,n)}) \in \mathbb{R}^{m\times n}$$

[1] [Werner1, Satz 6.2]

ist, d.h. ausgeschrieben:

$$U^TAV = \left(\begin{array}{ccc|ccc} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_r & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\1\\1\end{array}}\right\} r := Rg(A) \\ \left.\vphantom{\begin{array}{c}1\\1\\1\end{array}}\right\} m-r \end{array}$$

$$\underbrace{\qquad\qquad}_{r} \qquad \underbrace{\qquad\qquad}_{n-r}$$

mit $\sigma_1 \geq ... \geq \sigma_r > \underbrace{\sigma_{r+1} = ... = \sigma_{\min(m,n)}}_{=0}$

Definition. *Die* $\sigma_1 \geq ... \geq \sigma_r$ *heißen die* ***singulären Werte*** *von* A.
Wie sich im Beweis zum Satz herausstellen wird, sind sie

1. *eindeutig durch* A *bestimmt und*

2. *notwendig die Wurzeln aus den positiven Eigenwerten der Matrizen*

$$A^TA \quad bzw. \quad AA^T \quad (\text{\it übereinstimmend}).$$

Der zweite Punkt wird in der Literatur[2] *häufig als Definition des Begriffs „Singulärwerte" verwendet.*

1.2 Anwendung auf lineare Ausgleichsprobleme

Bevor der Satz nun bewiesen wird, soll seine Bedeutung in der numerischen Anwendung motiviert werden durch die folgende

Bemerkung. Gegeben sei das lineare Ausgleichsproblem

$$||Ax - b||_2 \stackrel{!}{=} min, \tag{*}$$

dabei ist $A \in \mathbb{R}^{m\times n}$ eine Matrix (wobei $m \geq n$), $b \in \mathbb{R}^m$ ein Vektor und gesucht wird der Vektor $x \in \mathbb{R}^n$.
Sei, wie im Satz angegeben,

$$U^TAV = \mathrm{diag}(\sigma_i)$$

die Singulärwertzerlegung von A mit orthogonalen Matrizen

$$U = (u_1, ..., u_m) \in \mathbb{R}^{m\times m} \quad und \quad V = (v_1, ..., v_n) \in \mathbb{R}^{n\times n} \quad (Spaltenvektoren)$$

[2] z.B. [Opfer, Def. 9.18]

Für beliebiges $x \in \mathbb{R}^n$ gilt dann

$$\begin{aligned} \|Ax-b\|_2^2 &= \|U^T(Ax-b)\|_2^2 \quad \textit{(orth. Matrizen sind normerhaltend)} \\ &= \|U^TAV(V^Tx) - U^Tb\|_2^2 \\ &= \|\operatorname{diag}(\sigma_i)V^Tx - U^Tb\|_2^2 \\ &= \sum_{i=1}^{r}[\sigma_i(V^Tx)_i - u_i^Tb]^2 + \sum_{i=r+1}^{m}(u_i^Tb)^2. \end{aligned}$$

Hieraus folgt: $x \in \mathbb{R}^n$ ist genau dann eine Lösung des linearen Ausgleichsproblems, wenn

$$V^Tx = (\frac{u_1^Tb}{\sigma_1}, ..., \frac{u_r^Tb}{\sigma_r}, \alpha_{r+1}, ..., \alpha_m)^T \quad mit\ \alpha_i \in \mathbb{R}.$$

Dies wiederum liefert: Ist $x \in \mathbb{R}^n$ eine Lösung von (*), so ist

$$\|x\|_2^2 = \|V^Tx\|_2^2 = \sum_{i=1}^{r}\left(\frac{u_i^Tb}{\sigma_i}\right)^2 + \sum_{i=r+1}^{m}\alpha_i^2$$

Es existiert eine eindeutige Lösung $\tilde{x}$ des linearen Ausgleichsproblems mit minimaler Norm. Sie ist gegeben durch

$$\tilde{x} = V(\frac{u_1^Tb}{\sigma_1}, ..., \frac{u_r^Tb}{\sigma_r}, 0, ..., 0)^T = \sum_{i=1}^{r}\frac{u_i^Tb}{\sigma_i}v_i$$

Zusammenfassend gilt also: Kennt man eine Singulärwertzerlegung von $A \in \mathbb{R}^{m\times n}$, so kann man für beliebiges $b \in \mathbb{R}^m$ die Menge aller Lösungen des zugehörigen linearen Ausgleichsproblems explizit angeben.
Die Singulärwertzerlegung hat gegenüber der in der Numerik-Vorlesung vorgestellten QR-Zerlegung (Householder) den Vorteil, dass die Voraussetzung über den vollen Rang von A ($\operatorname{Rg} A = n$) fallengelassen wird[3]

1.3 Beweis des Existenzsatzes

Exkurs: Lineare Algebra

Es werden einige Konzepte und Definitionen aus der Linearen Algebra benötigt, unter anderem:

- Man nennt zwei Matrizen $A, B \in \mathbb{R}^{m\times m}$ **ähnlich**, wenn es eine invertierbare Matrix $W \in \mathbb{R}^{m\times m}$ gibt mit $W^{-1}AW = B$.
  Ähnlichkeit ist eine Äquivalenzrelation auf $\mathbb{R}^{m\times m}$, wegen

  $$\det(B - xE_m) = \det(W^{-1}(A - xWE_mW^{-1})W) = \det(A - xE_n)$$

  bleiben die Eigenwerte einer Matrix unter Ähnlichkeitstransformationen erhalten, d.h. A und B haben dieselben EW.

[3][Schwarz, S. 360]

- $A \in \mathbb{R}^{n\times n}$ heißt **symmetrisch**, wenn $A = A^T$. Die Eigenwerte einer symmetrischen Matrix sind reell[4]. Sind $\lambda_1, ..., \lambda_n \in \mathbb{R}$ die EW der symmetrischen Matrix A, so existiert hierzu eine *Orthonormalbasis* $\{u_1, ..., u_n\} \subset \mathbb{R}^n$ von Eigenvektoren, d.h. es ist

$$u_j^T u_k = \delta_{jk} := \begin{cases} 1 & f\ddot{u}r\ j = k \\ 0 & f\ddot{u}r\ j \neq k \end{cases}$$

  und $Au_j = \lambda_j u_j$ für $j = 1, ..., n$.
  Sei nun eine Matrix $S \in \mathbb{R}^{n\times n}$ definiert durch $S := (u_1, ..., u_n)$. Dann ist S orthogonal, da $S^T S = E_n$. Außerdem gilt: $S^{-1}AS = S^T AS = \mathrm{diag}(\lambda_1, ..., \lambda_n)$.
  Eine symmetrische Matrix kann also durch eine *Ähnlichkeitstransformation* mit einer orthogonalen Matrix auf Diagonalgestalt transformiert werden.
  Spezielle symmetrische Matrizen, die aus einer beliebigen Matrix $A \in \mathbb{R}^{m\times n}$ entstehen, sind $A^T A$ und AA^T. Sie besitzen noch die zusätzliche Eigenschaft, dass beide *positiv semidefinit* sind. Das heißt

$$\forall x \in \mathbb{R}^n : x^T(A^T A)x \geq 0 \Leftrightarrow \textit{Alle EW von } A^T A \textit{ sind nichtnegativ.}$$

  (Entsprechendes gilt für AA^T).

- Ferner wird im Beweis die sog. **Dimensionsformel** aus der linearen Algebra[5] verwendet: Sei die Matrix $A \in \mathbb{R}^{m\times n}$ gegeben.

$$\mathrm{Rg}(A) = \dim Im\ A = \dim \mathbb{R}^n - \dim \mathrm{Ker}\, A = n - \dim \mathrm{Ker}\, A$$

  Dabei gilt die Beziehung:

$$\mathrm{Ker}\, A = \mathrm{Ker}\, A^T A$$

  Bew.:

$$\begin{array}{lrl} „\subset“: & x \in \mathrm{Ker}\, A & \Leftrightarrow Ax = 0 \Rightarrow A^T Ax = 0 \\ & & \Leftrightarrow x \in \mathrm{Ker}\, A^T A \\ „\supset“: & x \in \mathrm{Ker}\, A^T A & \Leftrightarrow A^T Ax = 0 \Rightarrow < A^T Ax, x >= 0 \\ & & \Rightarrow < Ax, Ax >= \|Ax\|^2 = 0 \\ & & \Rightarrow Ax = 0 \Leftrightarrow x \in \mathrm{Ker}\, A \end{array}$$

Beweis. Der Beweis zerfällt in vier Teile.

1. Die positiven Eigenwerte von $A^T A \in \mathbb{R}^{n\times n}$ und $AA^T \in \mathbb{R}^{m\times m}$ stimmen überein und haben dieselbe geometrische Vielfachheit, d.h. es ist

$$\dim \mathrm{Ker}(A^T A - \lambda E_n) = \dim \mathrm{Ker}(AA^T - \lambda E_m)$$

   für alle EW $\lambda > 0$ von $A^T A$ bzw. AA^T.

[4] [Fischer, Le. 5.6.2]
[5] [Fischer, Satz 2.2.4]

2. Es ist $r := \mathrm{Rg}(A) = \mathrm{Rg}(A^T) = \mathrm{Rg}(A^TA) = \mathrm{Rg}(AA^T)$ gleich der Anzahl der positiven EW von A^TA bzw. AA^T (entsprechend ihrer algebraischen Vielfachheit gezählt).

3. Die Matrizen $V \in \mathbb{R}^{n\times n}$ und $U \in \mathbb{R}^{m\times m}$ werden aus den Eigenvektoren von A^TA und AA^T konstruiert und erfüllen
$$U^TAV = \mathrm{diag}(\sigma_i)$$

4. Die Singulärwertzerlegung ist eindeutig bestimmt.

Zu 1) Sei $x \in \mathbb{R}^n \backslash \{0\}$ Eigenvektor zum EW $\lambda > 0$ von A^TA, also $A^TAx = \lambda x$. Dann ist $Ax \in \mathbb{R}^m \backslash \{0\}$ und es folgt
$$(AA^T)Ax = A(A^TA)x = A\lambda x = \lambda(Ax)$$
somit ist $Ax \in \mathbb{R}^m \backslash \{0\}$ Eigenvektor zum EW λ von AA^T.
Die andere Richtung — $\lambda > 0$ ist EW von $AA^T \Rightarrow \lambda$ ist EW von A^TA — ist entsprechend zu zeigen. Damit stimmen die positiven EW von A^TA und AA^T überein. Es bleibt die geometrische Vielfachheit zu zeigen:
Sei $\lambda > 0$ ein EW von A^TA bzw. AA^T und $k := \dim \mathrm{Ker}(A^TA - \lambda E_n)$ die geometrische Vielfachheit von λ in A^TA.
Sei $\{x_1, ..., x_k\}$ ein Orthonormalsystem von $\mathrm{Ker}(A^TA - \lambda E_n)$ aus Eigenvektoren zu λ von A^TA. Wie eben gezeigt, ist dann $Ax_i \in \mathbb{R}^m \backslash \{0\}$ für $i = 1, ..., k$ ein Eigenvektor zu $\lambda > 0$ von AA^T. Wegen
$$(Ax_i)^T(Ax_j) = x_i^T \underbrace{A^TAx_j}_{\lambda x_j} = x_i^T \lambda x_j = \lambda\delta_{ij}$$
ist $\{Ax_1, ..., Ax_k\}$ ein Orthogonalsystem von $\mathrm{Ker}(AA^T - \lambda E_m)$ aus Eigenvektoren von AA^T, also
$$\dim \mathrm{Ker}(A^TA - \lambda E_n) \leq \dim \mathrm{Ker}(AA^T - \lambda E_m).$$
Entsprechend zeigt man $\dim \mathrm{Ker}(AA^T - \lambda E_m) \leq \dim \mathrm{Ker}(A^TA - \lambda E_n)$. Daraus folgt dann die Gleichheit
$$\dim \mathrm{Ker}(A^TA - \lambda E_n) = \dim \mathrm{Ker}(AA^T - \lambda E_m).$$

Zu 2) Es ist $r := \mathrm{Rg}(A) = \mathrm{Rg}(A^T)$, da Zeilen- und Spaltenrang einer Matrix übereinstimmen. Weiter gilt
$$\begin{aligned} \mathrm{Rg}(A) &= n - \dim \mathrm{Ker}\, A && \scriptstyle (Dimensionsformel) \\ &= n - \dim \mathrm{Ker}\, A^TA && \scriptstyle (\mathrm{Ker}\, A = \mathrm{Ker}(A^TA)) \\ &= \mathrm{Rg}(A^TA) \end{aligned}$$

Entsprechend zeigt man $\mathrm{Rg}(A^T) = \mathrm{Rg}(AA^T)$.
Da man eine symmetrische Matrix M durch Ähnlichkeitstransformation auf Diagonalform

$$S^{-1}MS = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

bringen kann, und die Ränge beider Matrizen gleich sind, also

$$\mathrm{Rg}(m) = \mathrm{Rg}(S^{-1}MS) = \mathrm{Rg}(\mathrm{diag}(\lambda_i)) = \sharp\{\lambda_i \neq 0\},$$

ist der Rang einer symmetrischen Matrix durch die Anzahl ihrer von Null verschiedenen EW gegeben.
Da A^TA und AA^T positiv semidefinit sind, sind ihre EW alle nichtnegativ, ihr Rang ist also durch die Anzahl ihrer *positiven* EW (übereinstimmend) gegeben.

Zu 3) Seien $\lambda_1 \geq ... \geq \lambda_r > \lambda_{r+1} = ... = \lambda_n = 0$ die EW von $A^TA \in \mathbb{R}^{n\times n}$ und sei $\{v_1, ..., v_n\} \subset \mathbb{R}^n$ eine zugehörige Orthonormalbasis aus Eigenvektoren (zur Existenz: s.o.).
Man definiere

$$\sigma_i := \begin{cases} \sqrt{\lambda_i} & für\ i = 1, ..., r \\ 0 & für\ i = r+1, ..., \min(m, n) \end{cases}$$

und $u_i := \frac{1}{\sigma_i} A v_i$ für $i = 1, ..., r$.
Aus

$$\begin{aligned} AA^T u_i &= AA^T(\frac{1}{\sigma_i} A v_i) = \frac{1}{\sigma_i} A \underbrace{(A^TA)}_{\lambda_i v_i} v_i \\ &= \frac{1}{\sigma_i} \lambda_i A v_i = \lambda_i u_i \qquad für\ i = 1, ..., r \end{aligned}$$

und

$$\begin{aligned} u_i^T u_j &= (\frac{1}{\sigma_i} A v_i)^T (\frac{1}{\sigma_j} A v_j) \\ &= \frac{1}{\sigma_i} \frac{1}{\sigma_j} v_i^T \underbrace{A^T A v_j}_{\lambda_j v_j} = \lambda_j \frac{1}{\sigma_i \sigma_j} \underbrace{v_i^T v_j}_{\delta_{ij}} \\ &= \lambda_j \frac{1}{\sigma_i \sigma_j} \delta_{ij} = \lambda_i \frac{1}{\sigma_i^2} \delta_{ij} \\ &= \delta_{ij} \qquad für\ i, j = 1, ..., r \end{aligned}$$

folgt, daß $\{u_1, ..., u_r\} \subset \mathbb{R}^m$ ein Orthonormalsystem von Eigenvektoren von AA^T ist. Falls $r < m$, läßt sich $\{u_1, ..., u_r\}$ zu einer Orthonormalbasis $\{u_1, ..., u_m\} \subset \mathbb{R}^m$

aus Eigenvektoren von AA^T ergänzen.
Mit den orthogonalen Matrizen

$$U = (u_1, ..., u_m) \in \mathbb{R}^{m\times m} \quad \textit{und} \quad V = (v_1, ..., v_n) \in \mathbb{R}^{n\times n}$$

gilt dann für die Einträge der Matrix U^TAV:

$$(U^TAV)_{ij} = u_i^TAv_j = \frac{1}{\sigma_i}v_i^T \underbrace{A^TAv_j}_{\lambda_j v_j} = \frac{\lambda_j}{\sigma_i}v_i^Tv_j = \sigma_i\delta_{ij} \qquad \textit{für } 1 \le i, j \le r,$$

d.h. der linke obere Teil von U^TAV ist eine Diagonalmatrix. Für die restlichen Einträge folgt mit

$$AA^Tu_i = 0 \overset{\text{Ker } A^T = \text{Ker } AA^T}{\Leftrightarrow} \underset{\textit{für } i=r+1,...,m \ (u_i \textit{ ist EV zum EW } 0)}{A^Tu_i = 0 \quad \Leftrightarrow \quad u_i^TA = 0}$$

und

$$A^TAv_j = 0 \overset{\text{Ker } A = \text{Ker } A^TA}{\Leftrightarrow} Av_j = 0 \qquad \textit{für } i=r+1,...,n$$

daß diese verschwinden, d.h.

$$(U^TAV)_{ij} = u_i^TAv_j = 0 \qquad \textit{für } i\in\{r+1,...,m\} \textit{ oder } j\in\{r+1,...,n\}.$$

Damit ergibt sich für die Matrix U^TAV die Darstellung

$$U^TAV = \begin{pmatrix} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

und die Existenz der Singulärwertzerlegung einer Matrix $A \in \mathbb{R}^{m\times n}$ ist somit gezeigt.

Zu 4) Es bleibt die Eindeutigkeit der Singulärwertzerlegung zu zeigen.
Sei $U^TAV = \Sigma$ mit $\Sigma := \text{diag}(\sigma_i)$. Dann ist

$$(U^TAV)^T = V^TA^TU = \Sigma^T$$

Daraus folgt

$$V^TA^TAV = V^TA^TUU^TAV = \Sigma^T\Sigma = \text{diag}(\sigma_1^2, ..., \sigma_r^2, 0, ..., 0) \in \mathbb{R}^{n\times n}$$

und

$$U^TAA^TU = U^TAVV^TA^TU = \Sigma\Sigma^T = \text{diag}(\sigma_1^2, ..., \sigma_r^2, 0, ..., 0) \in \mathbb{R}^{m\times m}$$

Also sind A^TA und AA^T orthogonal ähnlich zu $\Sigma^T\Sigma$ bzw. $\Sigma\Sigma^T$. Die σ_i^2 sind damit den positiven EW von A^TA bzw. AA^T äquivalent. Da $\sigma_i > 0$ gilt, sind die σ_i somit eindeutig durch A festgelegt.
Das gleiche gilt für die Transformationsmatrizen U, V: Die Spalten von V bestehen beispielsweise aus den zu den σ_i^2 und 0 gehörenden Eigenvektoren von A^TA. Sie müssen gleichzeitig, da V orthogonal ist, jeweils die Länge 1 haben und sind somit eindeutig bestimmt. □

2 Numerisches Verfahren

In diesem Abschnitt wird das Verfahren von GOLUB-REINSCH vorgestellt, das die Singulärwerte bzw. die Singulärwertzerlegung einer Matrix $A \in \mathbb{R}^{m\times n}$ numerisch günstig berechnet.
Im folgenden sei $m \geq n$, ansonsten ersetze man A durch A^T. Im Prinzip könnte man die Singulärwerte bestimmen, indem man das Eigenwertproblem für die Matrix A^TA bzw. AA^T löst. Dieses Vorgehen ist jedoch nicht ratsam, da die explizite Berechnung von A^TA bzw. AA^T besonders bei einer schlecht konditionierten Matrix A mit einem Genauigkeitsverlust verbunden ist (vgl. Normalgleichungen $A^TAx = A^Tb$).
Hierzu ein

Beispiel Sei

$$A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix} \quad mit\ |\varepsilon| < \sqrt{g},\ g:\ Maschinengenauigkeit.$$

Die Matrix A^TA ist gegeben durch

$$A^TA = \begin{pmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{pmatrix}$$

und A besitzt die Singulärwerte $\sigma_1 = \sqrt{2+\varepsilon^2}$ und $\sigma_2 = |\varepsilon|$.
Wegen der Gleitpunktarithmetik erhält man statt A^TA die Matrix

$$gl(A^TA) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad mit\ den\ EW\ \tilde{\sigma}_1 = 2\ und\ \tilde{\sigma}_2 = 0.$$

Der zweite Singulärwert geht bei der Produktbildung verloren.
Aus diesem Grund haben GOLUB und REINSCH ein Verfahren entwickelt, das nicht mit der Matrix A^TA arbeitet. Es besteht aus mehreren Schritten.

1. Transformiere $A \in \mathbb{R}^{m\times n}$ orthogonal auf obere Bidiagonalform $\begin{pmatrix} B \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times n}$.

2. Berechne die Singulärwertzerlegung von B durch sukzessive

(a) Umformung von B auf eine Bidiagonalmatrix B_+, wobei ihr letztes Superdiagonalelement wesentlich kleiner ist als das von B.

(b) Ist dieses Superdiagonalelement kleiner als eine vorgegebene Toleranz geworden, so kann man es „zu Null erklären“, und das Verfahren für die übrige Bidiagonalmatrix fortsetzen. Man hat dann das Problem sozusagen um eine Dimension reduziert.

(c) Man erhält eine Diagonalmatrix B_+, d.h. $B_+^T B_+ = (B_+)^2$, aus der man die singulären Werte von B_+ direkt ablesen kann.

2.1 Transformation auf Bidiagonalform

Satz. *Sei $A \in \mathbb{R}^{m\times n}, m \geq n$, gegeben und seien*

$$P \in \mathbb{R}^{m\times m} \quad \text{und} \quad Q \in \mathbb{R}^{n\times n}$$

orthogonale Matrizen mit

$$P^T AQ = \begin{pmatrix} B \\ 0 \end{pmatrix} \begin{matrix} \}\ n \\ \}\ m-n \end{matrix} \quad \in \mathbb{R}^{m\times n}$$

Dann gilt:

1. *$A \in \mathbb{R}^{m\times n}$ und $B \in \mathbb{R}^{n\times n}$ haben dieselben singulären Werte.*

2. *Sei $B = U_B \operatorname{diag}(\sigma_1, ..., \sigma_n)V_B^T$ mit orthogonalen Matrizen $U_B, V_B \in \mathbb{R}^{n\times n}$ eine Singulärwertzerlegung von B.*
   Dann ist durch

$$A = U \operatorname{diag}(\sigma_1, ..., \sigma_n)V^T$$

   mit $U = P \begin{pmatrix} U_B & 0 \\ 0 & E_{n-m} \end{pmatrix}$ und $V = QV_B$ eine Singulärwertzerlegung von A gegeben.

Beweis.

1. Es ist

$$\begin{aligned} B^T B &= \begin{pmatrix} B \\ 0 \end{pmatrix}^T \begin{pmatrix} B \\ 0 \end{pmatrix} = (P^T AQ)^T (P^T AQ) \\ &= Q^T A^T \underbrace{PP^T}_{E} AQ = Q^T A^T AQ, \end{aligned}$$

   also sind die Matrizen $B^T B$ und $A^T A$ orthogonal ähnlich, was bedeutet, daß ihre EW und damit auch die Singulärwerte von A und B übereinstimmen.

2. Wegen

$$\begin{aligned} A &= P \begin{pmatrix} B \\ 0 \end{pmatrix} Q^T = P \begin{pmatrix} U_B & 0 \\ 0 & E_{n-m} \end{pmatrix} \cdot \operatorname{diag}(\sigma_1, ..., \sigma_n) \cdot V_B^T Q^T \\ &= P \begin{pmatrix} U_B & 0 \\ 0 & E_{n-m} \end{pmatrix} \cdot \operatorname{diag}(\sigma_i) \cdot (QV_B)^T \\ &= U \operatorname{diag}(\sigma_i) V^T \end{aligned}$$

erhält man aus einer Singulärwertzerlegung von B eine Singulärwertzerlegung von A. □

Im folgenden geht es darum, wie genau man die Matrix $A \in \mathbb{R}^{m\times n}$ auf obere Bidiagonalgestalt $\begin{pmatrix} B \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times n}$ mit $B \in \mathbb{R}^{n\times n}$ transformieren kann.
Es wird das Bidiagonalisierungsverfahren angesprochen, dessen Idee und Durchführung von Golub und Kahan stammt.
Dabei ähnelt das Vorgehen sehr dem bei einer QR-Zerlegung (Householder-Verfahren), die Grundidee besteht darin, einen vom Nullvektor verschiedenen Vektor durch Multiplikation mit einer geeigneten Householder-Matrix in ein Vielfaches des ersten Einheitsvektors zu überführen.
Im Unterschied zur QR-Zerlegung werden jedoch Spalten *und Zeilen* mittels Householder-Transformationen behandelt. Formal bedeutet das rechts- und linksseitige Multiplikationen im Wechsel:
Man bestimme also Householder-Matrizen

$$P_1, ..., P_n \in \mathbb{R}^{m\times m} \quad und \quad Q_1, ..., Q_{n-2} \in \mathbb{R}^{n\times n}$$

so daß

$$P_n ... P_1 A Q_1, ..., Q_{n-2}$$

obere Bidiagonalmatrix ist, also die Form

$$P_n ... P_1 A Q_1, ..., Q_{n-2} = \left(\begin{array}{cccc} d_1 & f_2 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & f_n \\ 0 & & & d_n \\ \hline & & 0 & \end{array}\right) =: \begin{pmatrix} B \\ 0 \end{pmatrix}$$

hat.
Im Einzelnen sind die P_i, $i = 1, ..., n$, und Q_i, $i = 1, ..., n-2$, so zu bestimmen, daß P_i die Einträge unterhalb des Elements $a_{ii}^{(i)}$ annulliert, d.h. durch Multiplikation mit P_i von links werden die $(m-i)$ letzten Einträge in der i-ten Spalte von A_i zu null gemacht.
Die Multiplikation von rechts mit Q_i annulliert alle Einträge rechts des Elementes $a_{i,i+1}^{(i)}$, also in der i-ten Zeile von A_i, $i = 1, ..., n-2$.

Dieses Vorgehen läßt sich veranschaulichen am

Beispiel einer 4×3-Matrix.
(Punkte $\bullet$ bezeichnen festbleibende Elemente, Sterne $*$ sich verändernde.)

$$\begin{pmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \end{pmatrix} \xrightarrow{P_1} \begin{pmatrix} * & * & * \\ & * & * \\ & * & * \\ & * & * \end{pmatrix} \xrightarrow{Q_1} \begin{pmatrix} \bullet & * & \\ & * & * \\ & * & * \\ & * & * \end{pmatrix} \xrightarrow{P_2} \begin{pmatrix} \bullet & \bullet & \\ & * & * \\ & & * \\ & & * \end{pmatrix} \xrightarrow{P_3} \begin{pmatrix} \bullet & \bullet & \\ & \bullet & \bullet \\ & & * \\ \hline 0 & 0 & 0 \end{pmatrix}$$

Das Problem eine Singulärwertzerlegung von $A \in \mathbb{R}^{m \times n}$ zu finden wurde nun darauf reduziert, eine Singulärwertzerlegung der Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$ anzugeben.

Bemerkung. Da keine besonderen Voraussetzungen an die Matrix A gestellt wurden, ist es nicht zwangsläufig gegeben, daß die Tridiagonalmatrix $B^T B$ **unreduziert** ist, d.h. sämtliche Subdiagonalelemente von Null verschieden sind.
In der Bidiagonalmatrix B können daher Diagonal- oder Superdiagonalelemente gleich Null sein. Diese Sonderfälle werden auf „vollständige" Bidiagonalmatrizen zurückgeführt; dies geschieht durch besondere Multiplikationen, auf die im Einzelnen hier nicht eingegangen wird[6].

2.2 Berechnung der Singulärwertzerlegung einer Bidiagonalmatrix nach dem Verfahren von Golub-Reinsch

Die Matrix B wird so verändert, daß man eine Matrix B_+ erhält, von der man sich erhofft, daß ihr letztes Superdiagonalelement wesentlich kleiner ist als das von B.

2.2.1 Transformation von B auf B_+

Sei die Bidiagonalmatrix B gegeben,

$$B = \begin{pmatrix} d_1 & f_2 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & f_n \\ 0 & & & d_n \end{pmatrix} \qquad B^T B = \begin{pmatrix} d_1^2 & d_1 f_2 & & 0 \\ d_1 f_2 & d_2^2 + f_2^2 & \ddots & \\ & \ddots & \ddots & d_{n-1} f_n \\ 0 & & d_{n-1} f_n & d_n^2 + f_n^2 \end{pmatrix}$$

($B^T B$ bitte <u>nicht</u> berechnen!)
Zuerst bestimmt man den Shift-Parameter $\sigma \in \mathbb{R}$ als den EW des unteren 2×2-Blocks

$$\begin{pmatrix} d_{n-1}^2 + f_{n-1}^2 & d_{n-1} f_n \\ d_{n-1} f_n & d_n^2 + f_n^2 \end{pmatrix}$$

[6] vgl. hierzu etwa [Werner2, Lemma 3.8]

von $B^T B$, der näher an $d_n^2 + f_n^2$ liegt.
Anschließend berechnet man eine Givens-Rotation

$$T_{12} = \begin{pmatrix} c & s & & & \\ -s & c & & 0 & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{pmatrix}$$

die die erste Spalte von $B^T B - \sigma E_n$ in ein Vielfaches des ersten Einheitsvektors transformiert, also

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} d_1^2 - \sigma \\ d_1 f_2 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

Exkurs: Givens-Rotationen

Zur Erinnerung: Die sogenannten Givens-Rotationen sind orthogonale $n \times n$-Matrizen der Form

$$G_{ik} = \begin{array}{cc} \begin{pmatrix} 1 & \vdots & & \vdots & \\ \dots & c & \dots & s & \dots \\ & \vdots & & \vdots & \\ \dots & -s & \dots & c & \dots \\ & \vdots & & \vdots & 1 \end{pmatrix} & \begin{matrix} \\ i \\ \\ k \\ \\ \end{matrix} \\ \begin{matrix} \quad & i & \quad\quad & k & \quad \end{matrix} & \end{array}$$

mit $c^2 + s^2 = 1$ und $1 \leq i < k \leq n$. Sie bewirken eine Drehung in der (i, k)-Ebene. Ist $x \in \mathbb{R}^n$ und $y := G_{ik} x$, so ist y gegeben durch

$$y_j = \begin{cases} c x_i + s x_k & \textit{für } j = i \\ -s x_i + c x_k & \textit{für } j = k \\ x_j & \textit{sonst} \end{cases}$$

Multipliziert man eine Matrix $A \in \mathbb{R}^{n \times n}$ von links mit einer Givens-Rotation, so bewirkt dies lediglich eine Veränderung in der i-ten und k-ten Zeile.
Zu einem vorgegebenen Vektor $x \in \mathbb{R}^n$ und Zahlen $1 \leq i < k \leq n$ läßt sich eine Givens-Rotation G_{ik} bestimmen, so daß die k-te Komponente von $G_{ik} x$ zu Null wird, also $(G_{ik} x)_k = 0$. Dabei bleiben außer der i-ten Komponente alle weiteren unverändert.
Genauer sieht die Transformation eines Vektors wie folgt aus:
Sei (α, β) der Vektor, der die i-te und k-te Komponente von einem vorgegeben Vektor $x \in \mathbb{R}^n$ enthält. Dann sind die Konstanten c und s mit $c^2 + s^2 = 1$ und γ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}$$

bestimmt durch

$$c := \pm\frac{\alpha}{\sqrt{\alpha^2+\beta^2}}, \quad s := \pm\frac{\beta}{\sqrt{\alpha^2+\beta^2}}, \quad \gamma := \pm\sqrt{\alpha^2+\beta^2}$$

Beim QR-Verfahren führt man als ersten Schritt eine QR-Zerlegung (mit Givens-Rotationen) auf die Hessenberg-Matrix $B^TB - \sigma E_m$ aus, also

$$B^TB - \sigma E_n = QR \quad \Leftrightarrow \quad Q^T(B^TB - \sigma E_n) = R$$

Dabei werden Givens-Rotationen G_{ik} bestimmt, die die Subdiagonalelemente von $B^TB - \sigma E_n$ eliminieren. Man erhält $Q = G_{12}^T \cdot \ldots \cdot G_{n-1,n}^T$, also
$G_{n-1,n} \cdot \ldots \cdot G_{12}(B^TB - \sigma E_n) = R$.
Die Matrix G_{12} transformiert also die erste Spalte von $B^TB - \sigma E_n$ in eine Vielfaches des ersten Einheitsvektors. Da sie genauso bestimmt wurde wie die Matrix T_{12}, gilt $T_{12} = \pm G_{12}$.
Nun verändert die Rechtsmultiplikation der anderen Givens-Rotationen $G_{i,i+1}^T$, $i > 1$ mit G_{12}^T die erste Spalte von G_{12}^T nicht mehr. Die ersten Spalten der Matrizen Q und T_{12}^T stimmen also evtl. bis auf das Vorzeichen überein. Dies ist ein wesentlicher Punkt zur Rechtfertigung, daß das letzte Superdiagonalelement nach der Transformation klein ist.

„Reparatur": Bidiagonalgestalt

Zurück zur Transformation von B.
Hat man die Matrix T_{12} bestimmt, so berechne man anschließend die Matrix

$$B := BT_{12}^T$$

In der transformierten Matrix BT_{12}^T verändert sich nur der obere 2×2-Block, dieser ist nach der Transformation gegeben durch:

$$\begin{pmatrix} d_1 & f_2 \\ 0 & d_2 \end{pmatrix}\begin{pmatrix} c & -s \\ s & c \end{pmatrix} = \begin{pmatrix} d_1c + f_2s & -d_1s + f_2c \\ d_2s & d_2c \end{pmatrix}$$

Die Bidiagonalgestalt wird also nur durch das Element in der Position $(2,1)$ gestört. Dieses Element soll nun durch Multiplikation von *links* mit einer geeigneten Givens-Rotation S_{12} annulliert werden. Dabei entsteht ein die Bidiagonalgestalt störendes Element in der Position $(1,3)$. Dieses wiederum soll nun durch Multilplikation mit einer Givens-Rotation T_{23}^T von *rechts* auf die Position $(3,2)$ gedrängt werden.
Insgesamt wird also das die Bidiagonalform störende Element durch abwechselnde Multiplikation von links und rechts mit geeigneten Givens-Rotationen „in Rössel-Sprüngen" aus der Matrix gedrängt. Man veranschauliche sich das Vorgehen an folgendem

Beispiel einer 4×4-Matrix.

$$\begin{pmatrix} \bullet & \bullet & & \\ & \bullet & \bullet & \\ & & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{T_{12}^T} \begin{pmatrix} * & * & & \\ * & * & \bullet & \\ & & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{S_{12}} \begin{pmatrix} * & * & * & \\ & * & * & \\ & & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{T_{23}^T} \begin{pmatrix} \bullet & * & & \\ & * & * & \\ & * & * & \bullet \\ & & & \bullet \end{pmatrix}$$

$$\xrightarrow{S_{23}} \begin{pmatrix} \bullet & \bullet & & \\ & * & * & * \\ & & * & * \\ & & & \bullet \end{pmatrix} \xrightarrow{T_{34}^T} \begin{pmatrix} \bullet & \bullet & & \\ & \bullet & * & \\ & & * & * \\ & & * & * \end{pmatrix} \xrightarrow{S_{34}} \begin{pmatrix} \bullet & \bullet & & \\ & \bullet & \bullet & \\ & & * & * \\ & & & * \end{pmatrix}$$

Genauer sehen die weiteren Schritte folgendermaßen aus:
Für $k = 1, ..., n-1$ bestimme man Givens-Rotationen $S_{k,k+1}$ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} b_{kk} \\ b_{k+1,k} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

und berechne $B := S_{k,k+1}B$.
Ist $k < n-1$, so bestimme man Givens-Rotationen $T_{k+1,k+2}$ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} b_{k,k+1} \\ b_{k,k+2} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

und berechne $B := BT_{k+1,k+2}^T$.
Durch die Transformationen erhält man dann die obere Bidiagonalmatrix B_+

$$B_+ := S_{n-1,n} \cdot ... \cdot S_{12} B T_{12}^T \cdot ... \cdot T_{n-1,n}^T$$

Damit ist $B_+^T B_+$ gegeben durch

$$B_+^T B_+ = T_{n-1,n} \cdot ... \cdot T_{12} B^T B (T_{n-1,n} \cdot ... \cdot T_{12})^T$$

2.2.2 Der implizite QR-Schritt

Die Matrix $B_+^T B_+$ hat die Eigenschaft, daß ihr letztes Subdiagonalelement kleiner ist, als das der Matrix $B^T B$. Dies beruht auf der Tatsache, daß das Verfahren zwar nur mit der Matrix B arbeitet, bei Durchführung der einzelnen Schritte auf B aber implizit ein Schritt des QR-Verfahrens auf $B^T B$ durchgeführt wird.

QR-Verfahren mit Shift-Strategie

Sei $B^T B$ die Matrix, auf die das QR-Verfahren angewandt wird. Zuerst bestimme man den Shift $\sigma \in \mathbb{R}$ als den EW des unteren 2×2-Blocks

$$\begin{pmatrix} d_{n-1}^2 + f_{n-1}^2 & d_{n-1} f_n \\ d_{n-1} f_n & d_n^2 + f_n^2 \end{pmatrix}$$

von B^TB, der $f_n^2 + d_n^2$ am nächsten liegt.
Anschließend führe man eine QR-Zerlegung auf $B^TB - \sigma E_n$ durch, also

$$B^TB - \sigma E_n = QR$$

und berechne $(B^TB)_{neu}$ durch

$$(B^TB)_{neu} = RQ + \sigma E_n = Q^TB^TBQ$$

Der QR-Schritt und insbesondere die Shift-Wahl bewirken, daß das letzte Subdiagonalelement der Matrix $(B^TB)_{neu}$ kleiner als das von B^TB. Bei weiteren Durchführungen von QR-Schritten sinkt es sogar unter eine vorgegebene Toleranzgrenze. Dabei wird die Konvergenz wesentlich von der Wahl des Shifts (er ist in jedem Schritt neu zu bestimmen) beeinflußt. Schließlich ist der Übergang zu einem $(n-1)$-dimensionalen Problem möglich.

Nun wird gezeigt, daß die Matrizen $B_+^TB_+$ und Q^TB^TBQ *im wesentlichen übereinstimmen*, d.h. daß jedes ihrer Elemente betragsmäßig gleich ist:

$$(B_+^TB_+)_{ij} = \pm(Q^TB^TBQ)_{ij}$$

Damit ist dann auch das letzte Subdiagonalelement der Matrix $B_+^TB_+$ kleiner als das von B^TB. Deswegen erhofft man sich, daß das letzte Superdiagonalelement von B_+ kleiner ist als das von B.
Die wesentliche Übereinstimmung beweist der nächste

Satz[7]**.** *Seien $Q = (q_1, ..., q_n)$ und $V = (v_1, ..., v_n)$ orthogonale Matrizen, die eine gegebene Matrix $A \in \mathbb{R}^{n \times n}$ jeweils in eine Hessenberg-Matrix*

$$H := Q^TAQ \quad bzw. \quad G := V^TAV$$

transformieren, wobei G unreduziert ist. Stimmen die ersten Spalten von Q und V evtl. bis auf das Vorzeichen überein, ist also $q_1 = \pm v_1$, so ist auch H unreduziert. Ferner sind dann Q und V sowie H und G im wesentlichen gleich, d.h. $D := V^TQ$ ist eine (orthogonale) Diagonalmatrix (besitzt nur +1 oder -1 als Diagonalelemente) und H ist gegeben durch $H = D^TGD$.

Beweis. (*durch „Induktion“:*)
Man definiere die orthogonale Matrix $D := V^TQ = (d_1, ..., d_n)$.
Ind.Anfang: Aus der Voraussetzung $q_1 = \pm v_1$ folgt, daß $d_1 = \pm e_1$, d.h. evtl. bis auf das Vorzeichen der erste Einheitsvektor ist.
Ind.Schluß: Angenommen, es sei $d_i = \pm e_i$ für $i = 1, ..., k$. Aus

$$GD = V^TAVV^TQ = V^TAQ = V^TQQ^TAQ = V^TQH = DH$$

[7] [Werner2, Satz 2.10]

erhält man beim Vergleich der k-ten Spalte

$$\pm g_k = \pm Ge_k = Gd_k = (GD)_k = (DH)_k = h_{k+1,k}d_{k+1} + \sum_{i=1}^{k} h_{ik}d_i$$

Multipliziert man diese Gleichung von links mit $d_i^T = \pm e_i^T$, so erhält man $\pm g_{ik} = h_{ik}$ für $i = 1, ..., k$ und anschließend $\pm g_{k+1,k}e_{k+1} = h_{k+1,k}d_{k+1}$. Da G unreduziert ist, ist $h_{k+1,k} \neq 0$ und damit $d_{k+1} = \pm e_{k+1}$. Daher ist auch H unreduziert und D eine (orthogonale) Diagonalmatrix. Für die Matrix H erhält man schließlich die Darstellung

$$H = Q^T AQ = Q^T(VV^T)A(VV^T)Q = D^T GD$$

□

Jetzt muß gezeigt werden, daß die Matrizen

$$B_+^T B_+ = T_{n-1,n} \cdot ... \cdot T_{12}B^T B(T_{n-1,n} \cdot ... \cdot T_{12})^T$$

und

$$(B^T B)_{neu} = Q^T B^T BQ$$

die Voraussetzungen des Satzes erfüllen und damit *im wesentlichen* übereinstimmen:
Beide Matrizen besitzen Hessenberg-Form und sind orthogonal ähnlich zu $B^T B$. Wegen

$$(T_{n-1,n} \cdot ... \cdot T_{12})^T e_1 = T_{12}^T \cdot ... \cdot T_{n-1,n}^T e_1 = T_{12}^T e_1$$

ist die erste Spalte von $(T_{n-1,n} \cdot ... \cdot T_{12})^T$ genau die erste Spalte von T_{12}^T, die wiederum (evtl. bis auf das Vorzeichen) mit der ersten Spalte von Q übereinstimmt. Somit folgt aus dem Satz, daß $B_+^T B_+$ und $(B^T B)_{neu}$ im wesentlichen gleich sind.

Zusammenfassung

Dementsprechend besteht ein typischer Schritt $B \to B_+$ des Verfahrens von Golub-Reinsch darin, daß man zunächst einen reellen Shift-Parameter σ bestimmt, dann T_{12} wählt und weitere Givens-Matrizen $T_{k,k+1}^T$, $S_{k,k+1}$ so bestimmt, daß

$$B_+ = S_{n-1,n} \cdot ... \cdot S_{12}BT_{12}^T \cdot ... \cdot T_{n-1,n}^T$$

wieder eine Bidiagonalmatrix wird.
$B_+^T B_+$ stimmt im wesentlichen mit der Matrix überein, die das QR-Verfahren angewandt auf $B^T B - \sigma E_n$ liefert. Deshalb hofft man, daß das letzte Superdiagonalelement von B_+ kleiner ist als das von B. Ist es nach weiterer Durchführung des Verfahrens unter eine Toleranzschranke gesunken, so kann man zu einer $(n-1) \times (n-1)$-Matrix übergehen.
Durch die *implizite* Berücksichtigung der Shifts wird insbesondere der Genauigkeitsverlust vermieden, der bei einer expliziten Ausführung der Subtraktion

$B^TB \to B^TB - \sigma E_m$ und der Addition $RQ \to RQ + \sigma E_m$ etwa für große k auftreten würde. Die Konvergenzeigenschaften sind natürlich die gleichen wie beim QR-Verfahren.

Literatur

[Fischer] Gerd Fischer. “Lineare Algebra“.
Vieweg, Braunschweig/Wiesbaden 1975, 2000.

[Opfer] Gerhard Opfer. “Numerische Mathematik für Anfänger“.
Vieweg, Braunschweig/Wiesbaden 1992, 2001.

[Schwarz] Hans R. Schwarz. “Numerische Mathematik“.
Teubner, Stuttgart 1997.

[Stoer2] Josef Stoer, Roland Bulirsch. “Numerische Mathematik 2“.
Springer, Berlin/Heidelberg/New York 1973, 2000.

[Werner1] Jochen Werner. “Numerische Mathematik 1“.
Vieweg, Braunschweig/Wiesbaden 1992.

[Werner2] Jochen Werner. “Numerische Mathematik 2“.
Vieweg, Braunschweig/Wiesbaden 1992.