第四讲 矩阵的奇异值分解

李宇峰

liyf@nju.edu.cn

人工智能学院

矩阵的奇异值分解

奇异值分解(singular Value decomposition, 简称SVD)是一种矩阵分解方法。

它是科学工程计算和数值代数中的最有用和最有效的工具之一。

它在统计分析、物理和应用科学(如信号与图像处理、系统理论和控制、通信、计算机视觉等)中被广泛地应用。

奇异值分解的概念在理论讨论中同样十分重要：在很多数值代数问题中，我们首先会思考：如果把奇异值分解用进去，会得到什么结果？这往往会使我们解决问题的思路得到开拓。

4.1奇异值分解

奇异值分解已经有一百多年的历史。1873年Beltrami从双线性函数$f(x,y)=x^TAy, A\in R^{n\times n}$出发，引入线性变换$x=U\xi$，$y=V\eta$, 这样双线性函数变为$f(x,y)=\xi^TS\eta$, 其中

$$S=U^TAV.$$

Beltrami观测到，如果约束U和V为正交矩阵，则它们的选择各存在n^2-n个自由度。他提出利用这些自由度使矩阵S的对角线以外的元素全部为零，即矩阵

$$S=\Sigma=diag(\sigma_1,\sigma_2,\cdots,\sigma_n)$$

为对角矩阵。于是用U和V^T分别左乘和右乘式$S=U^TAV$

立即得到

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

这就是Beltrami在1873年得到的实正方矩阵的奇异值分解。1874年Jordan也独立地推导出了实正方矩阵的奇异值分解。

后来，Autonne于1902年把奇异值分解推广到复正方矩阵；Eckart与Young于1936年又进一步把奇异值分解推广到一般的复长方形矩阵。

完全SVD

定理4.1 (实矩阵的奇异值分解) 令$A \in R^{m\times n}$，则存在正交矩阵$U \in R^{m\times m}$和$V \in R^{n\times n}$使得

$$A = U\Sigma V^T \qquad (1)$$

其中$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}_{m\times n}$, 且$\Sigma_1 = diag(\sigma_1, \sigma_2, \cdots, \sigma_r)$, 对角线元素按照从大到小的次序排列，即

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \qquad r = rank(A)$$

数值$\sigma_1, \sigma_2, \cdots, \sigma_r$连同$\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_{\min(m,n)} = 0$称为$A$的奇异值。

上面是$m \times n$的实矩阵A的奇异值分解定理，对于复数矩阵有

定理4.2 (复数矩阵的奇异值分解) 令$A \in C^{m \times n}$，则存在酉矩阵$U \in C^{m \times m}$和$V \in C^{n \times n}$使得

$$A = U\Sigma V^H \qquad (2)$$

其中$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}$, 且$\Sigma_1 = diag(\sigma_1, \sigma_2, \cdots, \sigma_r)$, 对角线元素按照从大到小的次序排列，即

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \qquad r = rank(A)$$

数值$\sigma_1, \sigma_2, \cdots, \sigma_r$连同$\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_{\min(m,n)} = 0$称为$A$的奇异值。

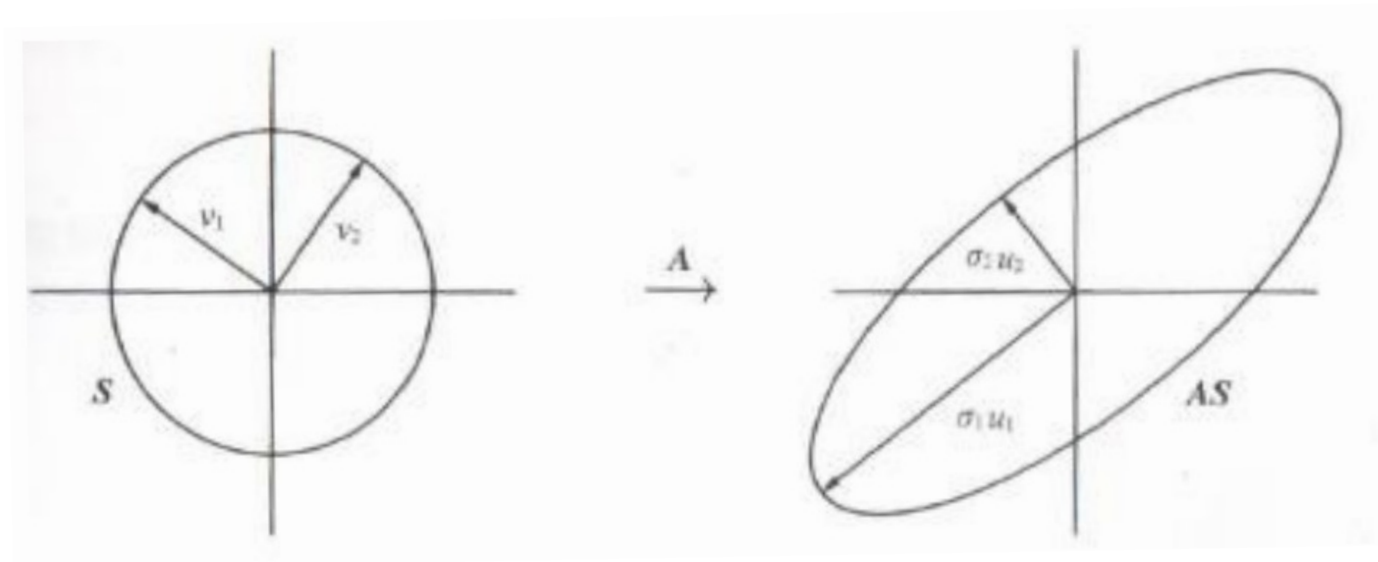
SVD的几何解释

我们也可以从几何的观点来看SVD，这里其实用到一个基本事实，就是任意一个$m\times n$矩阵将n维空间的单位球面映射为m维空间一个超球面。

令S为R^n中的单位球面，取任意的$A\in R^{m\times n}$，其中$m\geq n$，为方便起见，现设$rank(A)=n$，这时A的像AS是R^m中的一个超椭圆，这样可以根据AS的形状确定A的性质。

我们以下面的示意图说明有关概念。

SVD的几何解释



2×2矩阵的SVD

这是2×2矩阵的SVD，对于2维空间的单位球面S，给定2×2矩阵后，S经A的映射后AS就是一个2维椭球，这样长轴$\sigma_1 u_1$和短轴$\sigma_2 u_2$就确定了，而$\|u_1\|_2 = \|u_2\|_2 = 1$，而$\sigma_1 u_1$对应在$S$上的

SVD的几何解释

原像为v_1, $\sigma_2 u_2$对应在S上的原像为v_2，也就是

$$Av_1 = \sigma_1 u_1, Av_2 = \sigma_2 u_2.$$

推广到高维空间，设$A \in R^{n\times n}$, $rank(A) = n$, S是R^n中的单位球面，S经A的映射后，其像AS在R^n中为一椭球面，它的n个主半轴的长度记成$\sigma_1, \sigma_2, \cdots, \sigma_n$, 习惯上假设奇异值以降序编号，$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$.

其次我们定义AS的n个主半轴方向的单位向量$u_1, u_2, \cdots, u_n$, 为左奇异向量，其编号对应于奇异值。向量$\sigma_i u_i$就是AS的第i位大的主半轴。

SVD的几何解释

最后，我们定义S中的单位向量$v_1, v_2, \cdots, v_n \in S$，它是$AS$的$n$个主半轴的原像，称为$A$的右奇异向量，编号使得

$$Av_j = \sigma_j u_j.$$

当然我们也可以将矩阵推广到$A \in R^{m \times n}$的情况。

关于SVD

关于矩阵的奇异值分解我们有几点解释和注记：

(1) $n \times n$矩阵V为酉矩阵，用V右乘$A = U\Sigma V^H$就得到$AV = U\Sigma$, 这个列向量形式就是

$$Av_i = \begin{cases} \sigma_i u_i & i = 1,2,\cdots,r \\ 0 & i = r+1, r+2,\cdots,\min(m,n) \end{cases} \quad (3)$$

因此，V的列向量v_i称为矩阵A右奇异向量(right singular vector), V称为A的右奇异向量矩阵。

关于SVD

(2) $m \times m$矩阵U为酉矩阵，用U^H左乘$A = U\Sigma V^H$就得到

$U^H A = \Sigma V^H$, 这个列向量形式就是

$$u_i^H A = \begin{cases} \sigma_i v_i^H & i = 1,2,\cdots,r \\ 0 & i = r+1, r+2,\cdots,\min(m,n) \end{cases} \qquad (4)$$

因此，U的列向量u_i称为矩阵A的左奇异向量(left singular vector), 矩阵U称为A的左奇异向量矩阵。

(3) 矩阵A的奇异值分解式$A = U\Sigma V^H$可以改写成向量表达式

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^H \qquad (5)$$

关于SVD

(4) 当矩阵A的秩$r = rank(A) < \min\{m, n\}$时，由于奇异值$0 = \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_{\min\{m,n\}}$，故奇异值分解可以简化为

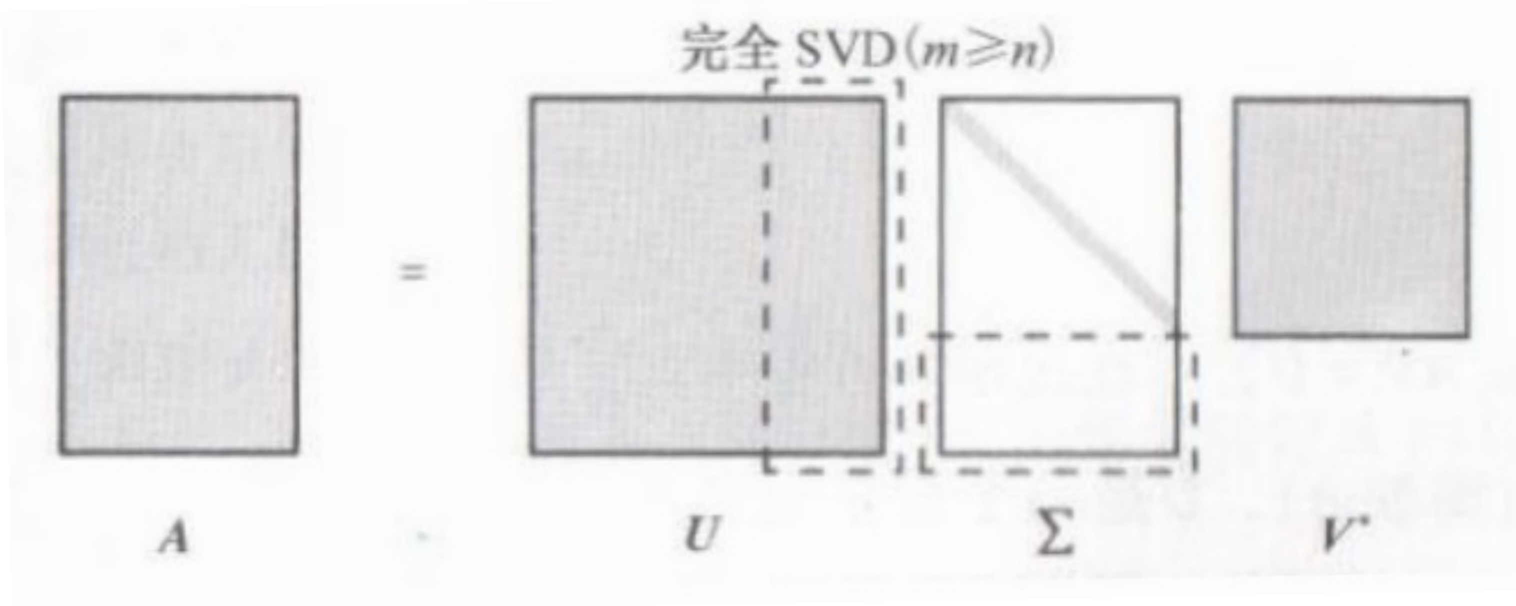
$$A = U_r \Sigma_r V_r^H \qquad (6)$$

其中

$$U_r = [u_1, u_2, \cdots, u_r], V_r = [v_1, v_2, \cdots, v_r], \Sigma_r = diag(\sigma_1, \sigma_2, \cdots, \sigma_r)$$

则(6)式称为A的紧凑奇异值分解（Compact SVD）.

关于SVD



图中的虚线指出了U中“不起作用”的列和Σ中“不起作用”部分的行。

关于SVD

(5) 当u_i^H左乘(3)式，又由于$u_i^H u_i = 1$, 因此得到

$$u_i^H A v_i = \sigma_i,\ i = 1,2,\cdots,\min\{m,n\}$$

表示成矩阵的形式就是

$$U^H A V = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix},\quad \Sigma_1 = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \quad (7)$$

(6) 由(2)式容易得到

$$AA^H = U\Sigma\Sigma^T U^H \qquad (8)$$

这说明了一个重要事实，就是$m \times n$矩阵A的奇异值σ_i就是矩阵乘积AA^H的第i个特征值的算术平方根。

关于SVD

奇异值分解与矩阵A的有关空间有密切关系。也就是有

（7）如果矩阵$A_{m\times n}$的秩为r，那么对应于A的奇异值分解

$A = U\Sigma V^H$, 那么就有下列结果：

1）$m\times n$酉矩阵U的前r列向量组成了矩阵A的列空间

$R(A) = \{y \in C^m \mid Ax = y, \forall x \in C^n\}$的标准正交基底；

2）U的后m-r列向量组成了A^H的零空间

$N(A^H) = \{x \in C^m \mid A^H x = 0\}$的标准正交基底；

3）$n\times n$酉矩阵V的前r列向量组成矩阵A的行空间$R(A^H)$的标准正交基底；

4）V的后n-r列组成矩阵A的零空间$N(A)$的标准正交基底。

关于SVD

其实以上这些性质都是很有用的，也就是说，酉矩阵U和V的列向量分别给出了相关空间$R(A), R(A^H), N(A)$和$N(A^H)$的标准正交基底。

其实当A是一个$m \times n$的矩阵时，A的非零奇异值就是$A^H A$或者AA^H的非零特征值的正平方根，$A^H A$和AA^H有相同的非零特征值，所不同的是零特征值的重数不一样。而U的m个列向量是AA^H的特征向量；V的n个列向量是$A^H A$的特征向量。

关于SVD

这样如果A是一个$n \times n$的正方矩阵，那么只要A有一个奇异值为零，那么A一定是奇异矩阵，事实上如果A的奇异值按照从大到小的次序排列，那么A的最小奇异值就反映了矩阵A接近奇异矩阵的距离，如果其最小奇异值为零，那么A就是一个奇异矩阵；如果最小奇异值不为零但很小，就说明A尽管不是奇异矩阵，但它接近于奇异矩阵。

这里我们再给出《计算方法》中的一个重要概念，就是矩阵A的条件数，所谓A的条件数的定义是：

定义 设非奇异矩阵$A \in R^{m \times m}$, 则定义$\|A\| \cdot \|A^{-1}\|$为A的条件数，记为$cond(A) = \|A\| \cdot \|A^{-1}\|$.

关于SVD

显然我们这里仅是定义了非奇异矩阵的条件数，以后引进了广义逆矩阵的概念后可以对一般的$m \times n$矩阵定义条件数。

这里需要指出两点：

1. 矩阵的条件数是由矩阵A和A^{-1}的范数来确定的，所用范数不同条件数也不同，如果用矩阵的2-范数，则A的条件数是

$$cond_2(A) = \frac{\sigma_1(A\text{的最大奇异值})}{\sigma_n(A\text{的最小奇异值})}$$

2. 矩阵的条件数是矩阵A的固有性质，它对解的精度会产生重大的影响。

事实上矩阵的奇异值与矩阵的范数、行列式、特征值都有密切的关系：

1.奇异值与范数的关系

(1) 矩阵A的2-范数(谱范数)等于A的最大奇异值，即

$$\|A\|_2 = \sigma_1 = \sigma_{max}$$

这是因为A的2-范数$\|A\|_2 = \sqrt{\lambda_{max}(A^H A)} = \sigma_1 = \sigma_{max}$.

如果A是实对称矩阵，则$A^H A = A^2$所以

$$\|A\|_2 = \sqrt{\lambda_{max}(A^2)} = |\lambda_{max}(A)|$$

(2) 由于A的F-范数$\|A\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|a_{ij}\right|^2}$是酉不变范数，

即$\|U^HAV\|_F=\|A\|_F$, 因此就有

$$\|A\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|a_{ij}\right|^2}=\|U^HAV\|_F=\|\Sigma\|_F=\sqrt{\sigma_1^2+\sigma_2^2+\cdots+\sigma_r^2}$$

也就是说任何一个矩阵的F-范数等于该矩阵所有奇异值的平方和开根号。从这里也可以看出$\|A\|_2\le\|A\|_F$.

2.奇异值与行列式的关系

如果A是一个$n\times n$的正方矩阵，由于酉矩阵的行列式的绝对值等于1，所以

$$|\det(A)| = |\det(U\Sigma V^H)| = |\det(U)|\times|\det(V^H)|\times|\det(\Sigma)|$$

$$= |\det(\Sigma)| = \sigma_1\sigma_2\cdots\sigma_n$$

这就是说，在A的n个奇异值中只要有一个奇异值为零，则A就是奇异矩阵，即$|\det(A)| = 0$，若奇异值都不等于零，则A为非奇异矩阵。

3.奇异值与特征值的关系

讲特征值只能是正方矩阵。

(1) 设A是$n \times n$的实对称矩阵，且其特征值为$\lambda_1, \lambda_2, \cdots, \lambda_n (|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|)$, 奇异值为$\sigma_1, \sigma_2, \cdots, \sigma_n (|\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_n|)$, 则$\sigma_i = |\lambda_i|, i = 1,2,\cdots,n.$

(2) 如果A为$m \times n$矩阵，$p = \min\{m, n\}$, 且$\sigma_1, \sigma_2, \cdots, \sigma_p$是$A$的奇异值，则

$$tr(A^H A) = \sum_i^p \sigma_i^2$$

(3) A的谱范数$\|A\|_2 = \sigma_{max}$, 矩阵A的谱条件数

$$cond_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{max}}{\sigma_{min}}$$

4.2 矩阵的低秩逼近(low rank approximation)

矩阵的低秩逼近是大数据处理中的一个十分重要的课题。

假如A是一个$m \times n$矩阵，设$rank(A) = p$, 低秩逼近问题就是要在所有的$m \times n$矩阵类中，确定一个$rank$为$k < p$的$m \times n$矩阵A_k, 使其是在一定的度量意义下与A最接近。

$$\min_{B \in R^{m \times n}} \|A - B\|_2$$
$$s.t.\ rank(B) = k < p$$

定理 (The Eckart-Young Theorem): 设$A \in R^{m\times n}$, $rank(A) = p$, A的奇异值分解为

$$A = \sum_{i=1}^{p} \sigma_i u_i v_i^T$$

令$k < p$, 则其低秩逼近矩阵为

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

逼近质量为

$$\min_{B\in R^{m\times n}, rank(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}$$

证明 由于$U^TA_kV = diag(\sigma_1, \sigma_2, \cdots, \sigma_k, 0, \cdots, 0)$, 故推得 $rank(A_k) = k$, 而$U^T(A - A_k)V = diag\left(0, \cdots, 0, \sigma_{k+1}, \cdots, \sigma_p, 0, \cdots, 0\right)$. 因此$\|A - A_k\|_2 = \sigma_{k+1}$.

下面证明对任何$B \in R^{m\times n}$, $rank(B) = k$, 都有$\|A - B\|_2 \geq \sigma_{k+1}$.

假设$B \in R^{m\times n}$, $rank(B) = k$. 这就推得B的零空间$N(B)$的维数为n-k, 设$x_1, x_2, \cdots, x_{n-k}$是$N(B)$的一组标准正交基，即

$$N(B) = span\{x_1, x_2, \cdots, x_{n-k}\}$$

由于

$\dim(span\{x_1, x_2, \cdots, x_{n-k}\}) + \dim(span\{v_1, v_2, \cdots, v_{k+1}\}) > \mathrm{n}$,

所以$span\{x_1, x_2, \cdots, x_{n-k}\} \cap span\{v_1, v_2, \cdots, v_{k+1}\} \neq \{0\}$.

设$z \in span\{x_1, x_2, \cdots, x_{n-k}\} \cap span\{v_1, v_2, \cdots, v_{k+1}\}$, 且$\|z\|_2 = 1$.

由于$Bz = 0, (A - B)z = Az$, 又由于$z \in span\{v_1, v_2, \cdots, v_{k+1}\}$

故z可以表示成

$$z = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_{k+1} v_{k+1}$$
$$= (v_1^T z) v_1 + (v_2^T z) v_2 + \cdots + (v_{k+1}^T z) v_{k+1}$$

因此

$$Az = A[(v_1^T z)v_1 + (v_2^T z)v_2 + \cdots + (v_{k+1}^T z)v_{k+1}]$$

$$= \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i$$

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2$$

这就证明了A_k是使得满足定理条件的低秩逼近矩阵。

前面讨论的是2-范数，对于F-范数，我们有

定理 对任意满足$0 \le k \le p = rank(A)$, 上述给出的A_k也满足

$$\|A - A_k\|_F = \min_{B \in R^{m\times n}, rank(B)=k} \|A - B\|_F = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_p^2}$$

这些用低秩的矩阵 A_k 去逼近一个有噪声或者扰动的矩阵A时，其逼近的误差由那些被舍弃的小的奇异值的平方和的平方根来确定。

例 计算以下10×8矩阵A的SVD, $rank(A) = 8$.

A =

4.9304	-0.1289	-0.1778	-0.2162	-0.2443	-0.2620	-0.3057	-0.3494
-0.1197	3.7814	-0.2967	-0.3541	-0.3906	-0.4064	-0.4742	-0.5419
-0.1501	-0.2691	2.6432	-0.4134	-0.4389	-0.4332	-0.5053	-0.5775
-0.1609	-0.2803	-0.3581	1.6056	-0.3891	-0.3422	-0.3993	-0.4563
-0.1522	-0.2524	-0.3006	-0.2969	0.7587	-0.1337	-0.1560	-0.1782
-0.1238	-0.1852	-0.1843	-0.1210	0.0046	0.1926	0.2246	0.2567
-0.1444	-0.2160	-0.2150	-0.1412	0.0053	0.2246	0.2621	0.2995
-0.1650	-0.2469	-0.2457	-0.1613	0.0061	0.2567	0.2995	0.3423
-0.1856	-0.2778	-0.2764	-0.1815	0.0069	0.2888	0.3369	0.3850
-0.2063	-0.3086	-0.3071	-0.2017	0.0077	0.3209	0.3743	0.4278

由SVD分解得到, $A = USV^T$:

U =

0.9948	-0.0104	-0.0156	-0.0208	-0.0260	-0.0312	-0.0364	-0.0416	-0.0468	-0.0519
-0.0104	0.9792	-0.0312	-0.0416	-0.0519	-0.0623	-0.0727	-0.0831	-0.0935	-0.1039
-0.0156	-0.0312	0.9532	-0.0623	-0.0779	-0.0935	-0.1091	-0.1247	-0.1403	-0.1558
-0.0208	-0.0416	-0.0623	0.9169	-0.1039	-0.1247	-0.1455	-0.1662	-0.1870	-0.2078
-0.0260	-0.0519	-0.0779	-0.1039	0.8701	-0.1558	-0.1818	-0.2078	-0.2338	-0.2597
-0.0312	-0.0623	-0.0935	-0.1247	-0.1558	0.8130	-0.2182	-0.2494	-0.2805	-0.3117
-0.0364	-0.0727	-0.1091	-0.1455	-0.1818	-0.2182	0.7455	-0.2909	-0.3273	-0.3636
-0.0416	-0.0831	-0.1247	-0.1662	-0.2078	-0.2494	-0.2909	0.6675	-0.3740	-0.4156
-0.0468	-0.0935	-0.1403	-0.1870	-0.2338	-0.2805	-0.3273	-0.3740	0.5792	-0.4675
-0.0519	-0.1039	-0.1558	-0.2078	-0.2597	-0.3117	-0.3636	-0.4156	-0.4675	0.4805

V =

0.9902	-0.0196	-0.0294	-0.0392	-0.0490	-0.0588	-0.0686	-0.0784
-0.0196	0.9608	-0.0588	-0.0784	-0.0980	-0.1176	-0.1373	-0.1569
-0.0294	-0.0588	0.9118	-0.1176	-0.1471	-0.1765	-0.2059	-0.2353
-0.0392	-0.0784	-0.1176	0.8431	-0.1961	-0.2353	-0.2745	-0.3137
-0.0490	-0.0980	-0.1471	-0.1961	0.7549	-0.2941	-0.3431	-0.3922
-0.0588	-0.1176	-0.1765	-0.2353	-0.2941	0.6471	-0.4118	-0.4706
-0.0686	-0.1373	-0.2059	-0.2745	-0.3431	-0.4118	0.5196	-0.5490
-0.0784	-0.1569	-0.2353	-0.3137	-0.3922	-0.4706	-0.5490	0.3725

$$S = \mathrm{diag}(5, 4, 3, 2, 1, 0.000067, 0.000045, 0.000021)$$

由于奇异值对角矩阵S从第6个奇异值开始突减，故可以认为是有噪音引起，可以用低秩逼近，取$k = 5$，取矩阵U的前5列为$U_{11} \in R^{10\times5}$的列正交矩阵、矩阵V的前5列为$V_{11} \in R^{8\times5}$的列正交矩阵。

V11 =

-0.9948	0.0104	-0.0156	-0.0208	0.0260
0.0104	-0.9792	-0.0312	-0.0416	0.0519
0.0156	0.0312	0.9532	-0.0623	0.0779
0.0208	0.0416	-0.0623	0.9169	0.1039
0.0260	0.0519	-0.0779	-0.1039	-0.8701
0.0312	0.0623	-0.0935	-0.1247	0.1558
0.0364	0.0727	-0.1091	-0.1455	0.1818
0.0416	0.0831	-0.1247	-0.1662	0.2078
0.0468	0.0935	-0.1403	-0.1870	0.2338
0.0519	0.1039	-0.1558	-0.2078	0.2597

S11 =

5.0000	0	0	0	0
0	4.0000	0	0	0
0	0	3.0000	0	0
0	0	0	2.0000	0
0	0	0	0	1.0000

$A_5 = AA = U_{11} \times S_{11} \times V_{11}^T$，这是所有秩为5的10*8的矩阵中欧氏长度意义下$A$的最佳逼近矩阵。

AA =

4.9304	-0.1289	-0.1778	-0.2162	-0.2443	-0.2620	-0.3057	-0.3494
-0.1197	3.7814	-0.2967	-0.3541	-0.3906	-0.4064	-0.4742	-0.5419
-0.1501	-0.2691	2.6432	-0.4134	-0.4389	-0.4332	-0.5053	-0.5775
-0.1609	-0.2803	-0.3581	1.6056	-0.3891	-0.3422	-0.3993	-0.4563
-0.1522	-0.2524	-0.3006	-0.2969	0.7587	-0.1337	-0.1560	-0.1783
-0.1238	-0.1852	-0.1843	-0.1210	0.0046	0.1925	0.2246	0.2567
-0.1444	-0.2160	-0.2150	-0.1412	0.0053	0.2246	0.2620	0.2995
-0.1650	-0.2469	-0.2457	-0.1613	0.0061	0.2567	0.2995	0.3422
-0.1856	-0.2778	-0.2764	-0.1815	0.0069	0.2888	0.3369	0.3850
-0.2063	-0.3086	-0.3071	-0.2017	0.0076	0.3209	0.3743	0.4278