计算方法

刘景铖

计算机软件新技术国家重点实验室
南京大学

回顾：线性迭代方法的收敛性

我们看到了迭代法解线性方程组的例子

解$Ax = b$时：**Jacobi**和**Gauss-Seidel**迭代都通过写出不动点方程，使用不动点迭代

解$Ax = \lambda x$：同时有两个变量：λ和x。严格来说，它们的乘积并不是线性关系，但（线性）迭代法依然适用

Power method: 收敛到最大特征值的方向

Inverse Power method

幂迭代告诉我们如何找到λ_1与v_1
其它特征值和特征向量呢？

给定方阵A. 假设A有n个特征值$\lambda_1, \lambda_2, \dots, \lambda_n$和线性无关的特征向量为$v_1, v_2, \dots, v_n$

A^{-1}的特征值?

$$\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_n}$$

特征向量？

$$v_1, v_2, \dots, v_n$$

迭代需要使用到A^{-1}？只需要解线性方程：**LU**分解

Inverse Power method

幂迭代告诉我们如何找到$\boldsymbol{\lambda_1}$与$\boldsymbol{v_1}$
其它特征值和特征向量呢？

给定方阵$\boldsymbol{A}$. 假设$\boldsymbol{A}$有$\boldsymbol{n}$个特征值$\boldsymbol{\lambda_1, \lambda_2, \dots, \lambda_n}$和线性无关的特征向量为$\boldsymbol{v_1, v_2, \dots, v_n}$

$(\boldsymbol{A} - \boldsymbol{qI})^{-1}$的特征值?

$$\frac{1}{\lambda_1 - q}, \frac{1}{\lambda_2 - q}, \dots, \frac{1}{\lambda_n - q}$$

特征向量？

$$\boldsymbol{v_1, v_2, \dots, v_n}$$

Inverse Power method

给定方阵$\boldsymbol{A}$. 假设$\boldsymbol{A}$有$\boldsymbol{n}$个特征值$\boldsymbol{\lambda_1, \lambda_2, \ldots, \lambda_n}$和线性无关的特征向量为$\boldsymbol{v_1, v_2, \ldots, v_n}$

$(\boldsymbol{A}-\boldsymbol{q}\boldsymbol{I})^{-1}$的特征值?

$$\frac{1}{\lambda_1 - q}, \frac{1}{\lambda_2 - q}, \ldots, \frac{1}{\lambda_n - q}$$

特征向量?

$$v_1, v_2, \ldots, v_n$$

如果对$(\boldsymbol{A}-\boldsymbol{q}\boldsymbol{I})^{-1}$使用幂迭代呢?
即$\boldsymbol{x_k} = (\boldsymbol{A}-\boldsymbol{q}\boldsymbol{I})^{-1}\boldsymbol{x_{k-1}}$
注意：可以通过高斯消元，先求出矩阵$\boldsymbol{B} = (\boldsymbol{A}-\boldsymbol{q}\boldsymbol{I})^{-1}$的LU分解，再进行迭代；对于迭代次数远远比$\boldsymbol{n}$小的稀疏矩阵，也可以尝试每一次迭代都解一次线性方程组

收敛到$\frac{1}{\lambda_k - q} = max_{1\le j\le n}\frac{1}{|\lambda_j - q|}$

收敛速度？取决于第二大特征值与最大特征值的比值——如果$\boldsymbol{q}$越接近一个特征值，越好

如何选择$\boldsymbol{q}$? 可以结合Rayleigh quotient ⇒ Rayleigh quotient iteration

Rayleigh Quotient Iteration

$$\sigma_k = \frac{x_{k-1}^\top A x_{k-1}}{x_{k-1}^\top x_{k-1}},$$
$$w_k = (A - \sigma_k I)^{-1} x_{k-1},$$
$$x_k = \frac{w_k}{\|w_k\|}$$

注意$(A - \sigma_k I)$不再是固定的
迭代次数更少，但是每一次迭代的计算代价更多

思考：当σ_k很接近一个特征值的时候，$A - \sigma_k I$也非常接近不可逆(ill conditioning)。这会带来问题吗？

Power method

幂迭代告诉我们如何找到λ_1与v_1

其它特征值和特征向量呢？

$$x = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$$

如果$\alpha_1 \neq 0$，随着$k \to \infty, A^k x \to \lambda_1^k \alpha_1 v_1$

如果$\alpha_1 = 0$呢？

- 如果$\alpha_2 \neq 0$，随着$k \to \infty, A^k x \to \lambda_2^k \alpha_2 v_2$
- 如果$\alpha_2 = 0, \alpha_3 \neq 0$，随着$k \to \infty, A^k x \to \lambda_3^k \alpha_3 v_3$

Power method

幂迭代告诉我们如何找到λ_1与v_1
其它特征值和特征向量呢？

如果$\alpha_1 \neq 0$，随着$k \to \infty, A^k x \to \lambda_1^k \alpha_1 v_1$
如果$\alpha_1 = 0$呢？

每一次迭代，通过投影确保与已知的特征向量都是正交的

QR分解与同时迭代

如何同时找出所有的特征值与特征向量？

假设A是实数对称矩阵。同时进行幂迭代，并保持向量两两正交。
回忆：正交化，即QR分解

最开始

$$\left[A\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} \middle| A\begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix} \middle| \cdots \middle| A\begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix} \right] = \left[\overline{q}_1^1|\cdots|\overline{q}_m^1\right]\begin{bmatrix} r_{11}^1 & r_{12}^1 & \cdots & r_{1m}^1 \\ & r_{22}^1 & & \vdots \\ & & \ddots & \vdots \\ & & & r_{mm}^1 \end{bmatrix}$$

$$A\overline{Q}_1 = \left[A\overline{q}_1^1|A\overline{q}_2^1|\cdots|A\overline{q}_m^1\right]$$

$$= \left[\overline{q}_1^2|\overline{q}_2^2|\cdots|\overline{q}_m^2\right]\begin{bmatrix} r_{11}^2 & r_{12}^2 & \cdots & r_{1m}^2 \\ & r_{22}^2 & & \vdots \\ & & \ddots & \vdots \\ & & & r_{mm}^2 \end{bmatrix}$$

$$= \overline{Q}_2 R_2.$$

QR分解与同时迭代

如何同时找出所有的特征值与特征向量？

假设A是实数对称矩阵。同时进行幂迭代，并保持向量两两正交。

$$
\begin{aligned}
A\overline{Q}_1 &= \left[A\overline{q}_1^1 | A\overline{q}_2^1 | \cdots | A\overline{q}_m^1\right] \\
&= \left[\overline{q}_1^2 | \overline{q}_2^2 | \cdots | \overline{q}_m^2\right]
\begin{bmatrix}
r_{11}^2 & r_{12}^2 & \cdots & r_{1m}^2 \\
 & r_{22}^2 & & \vdots \\
 & & \ddots & \vdots \\
 & & & r_{mm}^2
\end{bmatrix} \\
&= \overline{Q}_2 R_2.
\end{aligned}
$$

算法：

Set $\overline{Q}_0 = I$

for $j = 1, 2, 3, \ldots$

$A\overline{Q}_j = \overline{Q}_{j+1} R_{j+1}$

end

相似变换

定义：若矩阵$\boldsymbol{A}$与矩阵$\boldsymbol{B}$满足$\boldsymbol{B} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$，则说它们是相似的

相似矩阵有着相同的特征值

特征向量也能互相表示

QR分解给出一个特殊的相似变换

$$\boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}^{\top}\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}^{\top}\boldsymbol{Q}\boldsymbol{R}\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{Q}$$

QR分解与同时迭代

$\boldsymbol{A}$和$\boldsymbol{Q^{-1}AQ}$有相同的特征值

$$\boldsymbol{A = QR,}$$
$$\boldsymbol{Q^{-1}A\,Q = RQ}$$

QR迭代的等价形式：

- $\boldsymbol{A_1 = A}$
- 分解 $\boldsymbol{A_k = Q_k R_k}$
- $\boldsymbol{A_{k+1} = R_k\, Q_k}$

如果收敛，则 $\boldsymbol{A_\infty = Q_\infty R_\infty = R_\infty\, Q_\infty}$

因此$\boldsymbol{A_\infty Q_\infty = Q_\infty A_\infty}$

QR分解与同时迭代

假设矩阵$\boldsymbol{A}$的特征向量张满$\mathbb{R}^{\boldsymbol{n}}$，而且特征值互不相同。则矩阵$\boldsymbol{A}$和$\boldsymbol{B}$有相同的特征向量当且仅当

$$\boldsymbol{AB} = \boldsymbol{BA}$$

证明: (=>)

$$\boldsymbol{BAy} = \boldsymbol{BA}\sum_{i}\boldsymbol{\alpha_i x_i} = \boldsymbol{B}\sum_{i}\boldsymbol{\lambda_i^A \alpha_i x_i} = \sum_{i}\boldsymbol{\lambda_i^A \lambda_i^B x_i}$$

$$\boldsymbol{ABy} = \boldsymbol{AB}\sum_{i}\boldsymbol{\alpha_i x_i} = \boldsymbol{A}\sum_{i}\boldsymbol{\lambda_i^B \alpha_i x_i} = \sum_{i}\boldsymbol{\lambda_i^A \lambda_i^B x_i}$$

由于$\boldsymbol{y}$是任意的，因此$\boldsymbol{BA} = \boldsymbol{AB}$

QR分解与同时迭代

假设矩阵$\boldsymbol{A}$的特征向量张满$\mathbb{R}^n$，而且特征值互不相同。则矩阵$\boldsymbol{A}$和$\boldsymbol{B}$有相同的特征向量当且仅当

$$\boldsymbol{AB} = \boldsymbol{BA}$$

证明: (<=)

令$\boldsymbol{x}$为$\boldsymbol{A}$的一个特征向量，满足$\boldsymbol{Ax} = \boldsymbol{\lambda x}$，则有

$$\boldsymbol{A(Bx) = (AB)x = (BA)x = \lambda(Bx)}$$

- 如果$\boldsymbol{Bx} \neq \boldsymbol{0}$，则$\boldsymbol{Bx}$也为$\boldsymbol{A}$的一个特征向量，特征值为$\boldsymbol{\lambda}$。
  由于$\boldsymbol{A}$的特征值互不相同，则存在实数$\boldsymbol{c}$满足$\boldsymbol{Bx = cx}$，因此$\boldsymbol{x}$为$\boldsymbol{B}$的一个特征向量
- 如果$\boldsymbol{Bx = 0}$，则$\boldsymbol{x}$为$\boldsymbol{B}$的一个特征向量，特征值为0

因此$\boldsymbol{A}$的特征向量均为$\boldsymbol{B}$的特征向量，而$\boldsymbol{A}$的特征向量张满$\mathbb{R}^n$，因此它们有着相同的特征向量。

奇异值 (singular values)

对于一般的$\boldsymbol{m}\times\boldsymbol{n}$实数矩阵$\boldsymbol{A}$，可以有奇异值分解：

$$\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{S}\,\boldsymbol{V}^{\top}$$

$\boldsymbol{A}$ 的形状：$\boldsymbol{m}\times\boldsymbol{n}$

$\boldsymbol{U}$的形状：$\boldsymbol{m}\times\boldsymbol{m}$

$\boldsymbol{S}$的形状：$\boldsymbol{m}\times\boldsymbol{n}$

$\boldsymbol{V}$的形状：$\boldsymbol{n}\times\boldsymbol{n}$

$\boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{I}_{\boldsymbol{m}},\qquad \boldsymbol{V}^{\top}\boldsymbol{V} = \boldsymbol{I}_{\boldsymbol{n}}$

观察：

$$\boldsymbol{A}\boldsymbol{A}^{\top} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{S}^{\top}\boldsymbol{U}^{\top}$$
$$\boldsymbol{A}^{\top}\boldsymbol{A} = \boldsymbol{V}\boldsymbol{S}^{\top}\boldsymbol{S}\boldsymbol{V}^{\top}$$

$\boldsymbol{A}\boldsymbol{A}^{\top}$和$\boldsymbol{A}^{\top}\boldsymbol{A}$都是实数对称阵，可做特征值分解！

SVD

定理： $\boldsymbol{A}\boldsymbol{A}^\top$和$\boldsymbol{A}^\top\boldsymbol{A}$有着相同的非零特征值。

证明：考虑$\boldsymbol{A}\boldsymbol{A}^\top$的特征值$\boldsymbol{\lambda} \neq \mathbf{0}$和特征向量$\boldsymbol{u}$：$\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u} = \boldsymbol{\lambda}\boldsymbol{u}$

$$\boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{u}) = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}) = \boldsymbol{A}^\top(\boldsymbol{\lambda}\boldsymbol{u}) = \boldsymbol{\lambda}(\boldsymbol{A}^\top\boldsymbol{u})$$

可见，只要$\boldsymbol{A}^\top\boldsymbol{u} \neq \mathbf{0}$，则$\boldsymbol{A}^\top\boldsymbol{u}$为$\boldsymbol{A}^\top\boldsymbol{A}$的一个特征向量，特征值也是$\boldsymbol{\lambda}$。不妨假设$\boldsymbol{A}^\top\boldsymbol{u} = \mathbf{0}$，则$\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{u}) = \mathbf{0}$，与$\boldsymbol{\lambda} \neq \mathbf{0}$矛盾。

类似地，考虑$\boldsymbol{A}^\top\boldsymbol{A}$的特征值$\boldsymbol{\lambda} \neq \mathbf{0}$和特征向量$\boldsymbol{v}$：$\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v} = \boldsymbol{\lambda}\boldsymbol{v}$，也有

$$\boldsymbol{A}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{v}) = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}) = \boldsymbol{A}(\boldsymbol{\lambda}\boldsymbol{v}) = \boldsymbol{\lambda}(\boldsymbol{A}\boldsymbol{v}).$$

同理有 $\boldsymbol{A}\boldsymbol{v} \neq \mathbf{0}$，因此$\boldsymbol{A}\boldsymbol{v}$为 $\boldsymbol{A}\boldsymbol{A}^\top$的一个特征向量，特征值也是$\boldsymbol{\lambda}$。

SVD

定理： $\boldsymbol{A}\boldsymbol{A}^\top$**和**$\boldsymbol{A}^\top\boldsymbol{A}$**有着相同的非零特征值**

把它们记作$\{\lambda_i\}$，另外对$\boldsymbol{A}\boldsymbol{A}^\top$和$\boldsymbol{A}^\top\boldsymbol{A}$进行谱分解，可得到正规正交的向量$\{\boldsymbol{u}_i\}, \{\boldsymbol{v}_i\}$作为它们的特征向量：

$$\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i$$
$$\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i$$

把$\{\boldsymbol{u}_i\}$,组合成矩阵$\boldsymbol{U}$，$\{\boldsymbol{v}_i\}$组合成矩阵$\boldsymbol{V}$，则有$\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_m$, $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_n$

回顾之前的证明：

- 若$\boldsymbol{u}$为$\boldsymbol{A}\boldsymbol{A}^\top$的一个特征向量，则$\boldsymbol{A}^\top\boldsymbol{u}$为$\boldsymbol{A}^\top\boldsymbol{A}$的一个特征向量，且有着相同的特征值；
- 若$\boldsymbol{v}$为$\boldsymbol{A}^\top\boldsymbol{A}$的一个特征向量，则$\boldsymbol{A}\boldsymbol{v}$为$\boldsymbol{A}\boldsymbol{A}^\top$的一个特征向量，且有着相同的特征值

SVD

回顾之前的证明：

- 若$\boldsymbol{u}$为$\boldsymbol{A}\boldsymbol{A}^\top$的一个特征向量，则$\boldsymbol{A}^\top\boldsymbol{u}$为$\boldsymbol{A}^\top\boldsymbol{A}$的一个特征向量，且有着相同的特征值；
- 若$\boldsymbol{v}$为$\boldsymbol{A}^\top\boldsymbol{A}$的一个特征向量，则$\boldsymbol{A}\boldsymbol{v}$为$\boldsymbol{A}\boldsymbol{A}^\top$的一个特征向量，且有着相同的特征值

$$\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}_i = \boldsymbol{\lambda}_i\boldsymbol{u}_i$$
$$\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}_i = \boldsymbol{\lambda}_i\boldsymbol{v}_i$$

利用上述观察可得：存在常数c_i, d_i使得

$$\boldsymbol{A}^\top\boldsymbol{u}_i = \boldsymbol{c}_i\boldsymbol{v}_i$$
$$\boldsymbol{A}\boldsymbol{v}_i = \boldsymbol{d}_i\boldsymbol{u}_i$$

接下来，利用$\boldsymbol{v}_i$和$\boldsymbol{u}_i$都是单位长度的性质，可以求出c_i, d_i

注意到$\boldsymbol{u}_i^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}_i$可以有两种写法：

$$\boldsymbol{u}_i^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}_i = \boldsymbol{\lambda}_i\boldsymbol{u}_i^\top\boldsymbol{u}_i = \boldsymbol{\lambda}_i$$
$$\boldsymbol{u}_i^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}_i = \left(\boldsymbol{A}^\top\boldsymbol{u}_i\right)^\top\boldsymbol{A}^\top\boldsymbol{u}_i = \boldsymbol{c}_i^2\boldsymbol{v}_i^\top\boldsymbol{v}_i = \boldsymbol{c}_i^2$$

因此$\boldsymbol{c}_i^2 = \boldsymbol{\lambda}_i$，类似可得$\boldsymbol{d}_i^2 = \boldsymbol{\lambda}_i$，因此有：

$$\boldsymbol{A}^\top\boldsymbol{u}_i = \sqrt{\boldsymbol{\lambda}_i}\,\boldsymbol{v}_i$$
$$\boldsymbol{A}\boldsymbol{v}_i = \sqrt{\boldsymbol{\lambda}_i}\,\boldsymbol{u}_i$$

SVD

由于

$$\boldsymbol{A}^\top \boldsymbol{u}_i = \sqrt{\lambda_i}\, \boldsymbol{v}_i$$
$$\boldsymbol{A}\boldsymbol{v}_i = \sqrt{\lambda_i}\, \boldsymbol{u}_i$$

不失一般性地假设$\boldsymbol{m} > \boldsymbol{n}$，则有

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\, \mathbf{diag}\left(\sqrt{\boldsymbol{\lambda}_1}, \dots, \sqrt{\boldsymbol{\lambda}_n}\right)$$
$$\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{V} = \mathbf{diag}\left(\sqrt{\boldsymbol{\lambda}_1}, \dots, \sqrt{\boldsymbol{\lambda}_n}\right)$$
$$\Rightarrow\ \boldsymbol{A} = \boldsymbol{U}\, \boldsymbol{S}\, \boldsymbol{V}^\top$$

$\boldsymbol{S}$的对角线元素被称作奇异值(singular values)，通常记作$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$

从几何上看，矩阵$\boldsymbol{A}$的作用为：

- $\mathbb{R}^{\boldsymbol{n}}$上的旋转变换
- 伸缩变换
- $\mathbb{R}^{\boldsymbol{m}}$上的旋转变换

SVD与Rayleigh quotient

回顾特征值的min-max刻画，记$\boldsymbol{R_B}(\boldsymbol{x}) := \frac{\boldsymbol{x}^\top \boldsymbol{B}\,\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$

对于对称的矩阵$\boldsymbol{B}$， $\boldsymbol{R_B}(\boldsymbol{x})$在特征值处取到极值

回顾$\|A\|_2 := \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}$， $cond_2(A) := \|A\|_2 \cdot \|A^{-1}\|_2$

对于对称正定的矩阵A，

$$\|A\|_2 = \lambda_{max}(A)$$
$$cond_2(A) = \lambda_{max}(A)/\lambda_{min}(A)$$

对于一般的矩阵呢？

$$\|A\|_2 = \sqrt{\lambda_{max}(A^\top A)} = \sigma_{max}(A)$$
$$cond_2(A) = \frac{\sqrt{\lambda_{max}(A^\top A)}}{\sqrt{\lambda_{min}(A^\top A)}} = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$$

拉格朗日乘数法(选讲)

对于一般的矩阵 A， $\max\limits_{\|x\|_2=1} \|Ax\|_2$什么时候取到极值？

除了通过特征值的min-max刻画，还可以通过拉格朗日乘数法，把带约束的最优化问题转化为不带约束的：

$$\boldsymbol{L(\lambda) := sup_x\ x^\top A^\top A\, x - \lambda(x^\top x - 1)}$$

对$\boldsymbol{x}$求梯度也可得$\boldsymbol{A^\top A\ x = \lambda x}$

因此也可得到 $\max\limits_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{max}(A^\top A)}$

SVD与low rank approximations

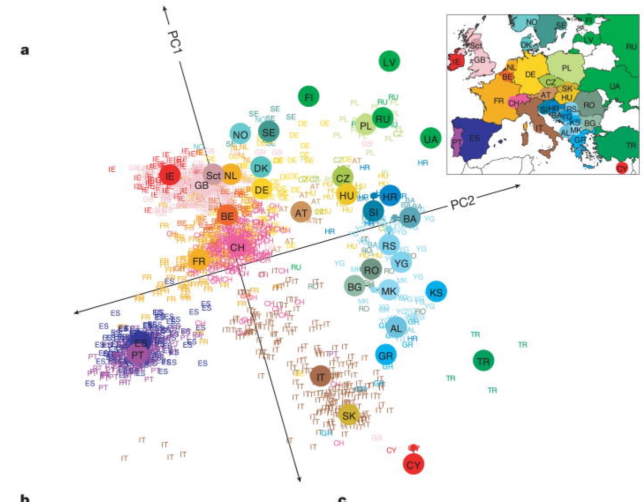
$$A = \sum_{i=1}^{min(m,n)} \sigma_i u_i v_i^\top$$

考虑如下的近似$\tilde{A} \coloneqq \sum_{i=1}^{r} \sigma_i u_i v_i^\top$

Eckart-Young定理： $\tilde{A}$是所有列空间的维度最多为 r的矩阵中，能同时最小化 $\|\tilde{A} - A\|_2$和 $\|\tilde{A} - A\|_F$的矩阵

Many applications:

- Principal component analysis
  - Data visualization
  - Genomes can encode geography
  - Eigenfaces
- Matrix completions
- Recommender systems
- Latent semantic analysis
- Approximation algorithm for Max-cut



SVD与伪逆pseudoinverse （选讲）

$$A = \sum_{i=1}^{min(m,n)} \sigma_i u_i v_i^\top$$

$$A^+ = VS^+U^\top = \sum_{i:\sigma_i \neq 0} \frac{1}{\sigma_i} v_i u_i^\top$$

如果A可逆，则$A^+ = A^{-1}$

如果A是overdetermined，则A^+b为 $Ax = b$的最小二乘解

如果A是underdetermined，则A^+b为 $Ax = b$中2范数最小的最小二乘解