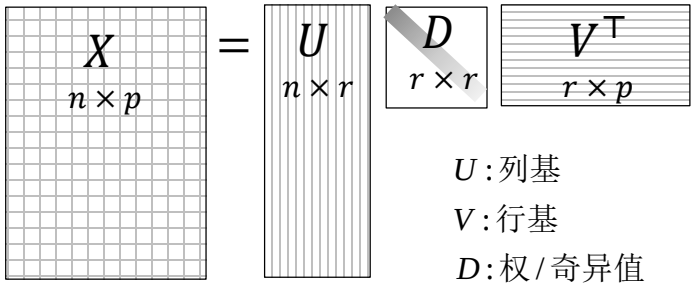
第十五讲 奇异值分解

2025.4.28

$$X = UDV^\top$$



U : 列基

V : 行基

D : 权 / 奇异值

奇异值分解(SVD)

列空间

对任一$n \times p$矩阵X，其列向量张成的空间（简称列空间）记作

$$C(X) = \{X\mathbf{t}\colon\ \mathbf{t} \in R^p\}$$

引理1.

(1) $C(X) = C(XX^\top)$，$C(X^\top) = C(X^\top X)$.

(2) $X^\top X$与$XX^\top$有相同的非0特征根，特征向量相互关联：若$\mathbf{a}$是$X^\top X$的特征向量，则$\mathbf{b} = X\mathbf{a}$是$XX^\top$的特征向量；若$\mathbf{b}$是$XX^\top$的特征向量，则$\mathbf{a} = X^\top\mathbf{b}$ 是$X^\top X$的特征向量。

证: (1) $C(XX^\top) = \{XX^\top\mathbf{t}\colon\ \mathbf{t} \in R^p\} = \{X\mathbf{s}\colon\ \mathbf{s} = X^\top\mathbf{t} \in R^p\} \subset C(X)$

若$\mathbf{x} \in C(XX^\top)^\perp, XX^\top\mathbf{x} = 0 \Rightarrow \mathbf{x}^\top XX^\top\mathbf{x} = 0 \Rightarrow X^\top\mathbf{x} = 0 \Rightarrow\ \mathbf{x} \in C(X)^\perp$

$\Rightarrow C(XX^\top)^\perp \subset C(X)^\perp, C(X) \subset C(XX^\top)$.

(2). $X^\top\mathbf{b} = X^\top X\mathbf{a} = \mathbf{a}\lambda\ \ \Rightarrow XX^\top(X\mathbf{a}) = (X\mathbf{a})\lambda$.

注：

$XX^\top$的特征向量刻画矩阵$C(XX^\top) = C(X)$；

$X^\top X$的特征向量刻画矩阵$C(X^\top X) = C(X^\top)$.

奇异值分解

定理1(奇异值分解，SVD). 任一秩为r的$n\times p$矩阵X可表示为

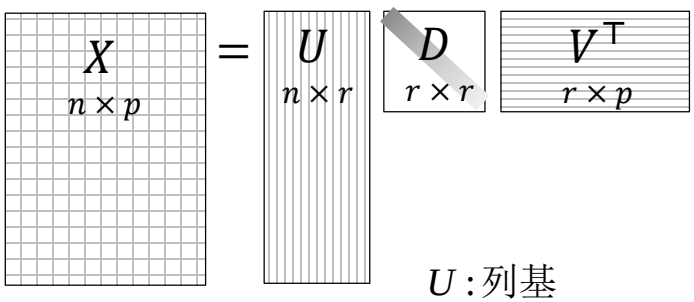
$$X = U_{n\times r} D_{r\times r} V^{\top}{}_{r\times p} = \sqrt{\lambda_1}\mathbf{u}_1\mathbf{v}_1^{\top} + \ldots + \sqrt{\lambda_r}\mathbf{u}_r\mathbf{v}_r^{\top}$$

其中$U^{\top}U = V^{\top}V = I_r$, $D = \mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_r})$的对角元称为奇异值,

其中$\lambda_1 \geq \ldots \geq \lambda_r > 0$为$X^{\top}X$或$XX^{\top}$的$r$个正特征根，$U = (\mathbf{u}_1,\ldots,\mathbf{u}_r)$,

$V = (\mathbf{v}_1,\ldots,\mathbf{v}_r)$的各列分别是$XX^{\top}$，$X^{\top}X$的特征向量。

SVD: $X = UDV^{\top}$

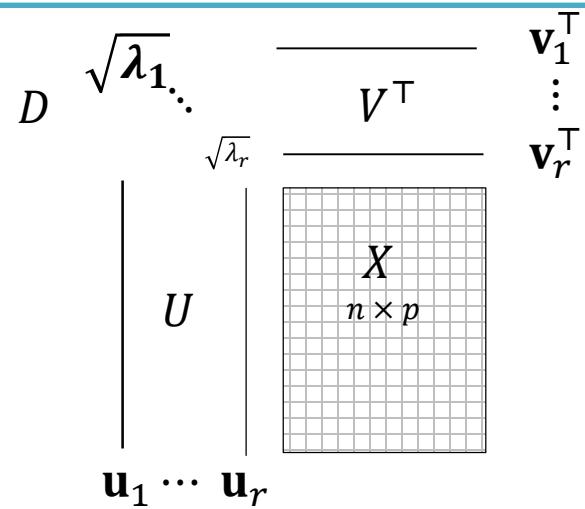


U : 列基

V : 行基

D : 权 / 奇异值

$$X = \sqrt{\lambda_1}\mathbf{u}_1\mathbf{v}_1^{\top} + \cdots + \sqrt{\lambda_r}\mathbf{u}_r\mathbf{v}_r^{\top}$$



证明： 考虑$X^\top X$的非0特征根λ_i，特征向量$\mathbf{v}_i$： $X^\top X\mathbf{v}_i = \mathbf{v}_i\lambda_i, i = 1, \dots, r.$

特征方程： $X^\top XV = VD^2$

其中$D^2 = \mathrm{diag}(\lambda_1, \dots, \lambda_r)$，
$V = (\mathbf{v}_1, \dots, \mathbf{v}_r)$

两边左乘X

$$XX^\top \underline{XV} = \underline{XV}D^2$$

令 $Y = XV$

Y的列向量是$XX^\top$的特征向量

$Y^\top Y = V^\top X^\top XV = D^2$
$\Rightarrow D^{-1}Y^\top YD^{-1} = I_r$

$$XX^\top Y = YD^2$$

令$U = YD^{-1}$

U的列向量是$XX^\top$正交单位化特征向量。$U^\top U = I_r$

特征方程： $XX^\top U = UD^2$

$$Y = XV = UD$$

右乘$V^\top$

$$XVV^\top = UDV^\top$$

$XVV^\top = X$

$$X = UDV^\top.$$

$C(V) = C(X^\top X) = C(X^\top)$
$\Rightarrow VV^\top = P_V = P_{X^\top} \Rightarrow XVV^\top = XP_{X^\top} = X$

SVD评注

- 若X已经中心化，则 SVD $\Leftrightarrow$ PCA
  $X^\top X = (n-1)S$，$Y = XV = UD$ 是所有非0奇异值对应的主成分。
- 正交展开： $X = UDV^\top = \sqrt{\lambda_1}\mathbf{u}_1\mathbf{v}_1^\top + \cdots + \sqrt{\lambda_r}\mathbf{u}_r\mathbf{v}_r^\top.$
- V、U分别是X的行、列特征： $XX^\top U = UD^2$，$X^\top XV = VD^2$
- 对偶方程： $XV = UD$，$X^\top U = VD$

对U, V的特定一列$\mathbf{u}, \mathbf{v}$ (对应的特征根λ):

$$X\mathbf{v} = \mathbf{u}\sqrt{\lambda}, \qquad X^\top\mathbf{u} = \mathbf{v}\sqrt{\lambda}$$

$$\mathbf{v} \in C(X^\top) \underset{X^\top}{\overset{X}{\rightleftarrows}} \mathbf{u} \in C(X)$$

虽然X不是方阵，没有特征向量，但$\{\mathbf{v}, \mathbf{u}\}$可看作是$X$及其伴随$X^\top$的“特征向量”/不变量

例1.

$$X=\begin{pmatrix}0&1&0&0\\0&0&2&0\\0&0&0&3\\0&0&0&0\end{pmatrix},\quad XX^{\mathsf{T}}=\begin{pmatrix}1&0&0&0\\0&4&0&0\\0&0&9&0\\0&0&0&0\end{pmatrix},\quad X^{\mathsf{T}}X=\begin{pmatrix}0&0&0&0\\0&1&0&0\\0&0&4&0\\0&0&0&9\end{pmatrix},$$

XX^{T}的特征根：$\lambda_1=9$，$\lambda_2=4$，$\lambda_3=1$。X的奇异值：3,2,1

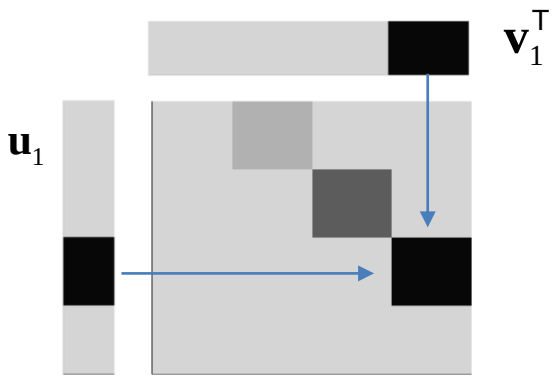
$$X=UDV^{\mathsf{T}}$$

其中$U=\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\\0&0&0\end{pmatrix}$，$V=\begin{pmatrix}0&0&0\\0&0&1\\0&1&0\\1&0&0\end{pmatrix}$，$D=\mathrm{diag}(3,2,1)$

XX^{T}或$X^{\mathsf{T}}X$的最大特征根$\lambda_1 = 9$对应的特征向量

- $\mathbf{u}_1 = (0,0,1,0)^{\mathsf{T}}$：$X$的第3行最重要，$(XX^{\mathsf{T}})_{33}$最大
- $\mathbf{v}_1 = (0,0,0,1)^{\mathsf{T}}$：$X$的第4列最重要，$(X^{\mathsf{T}}X)_{44}$最大

如下图，$\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}}$定位矩阵$X$的最重要的位置：(3,4)



类似地，第二大特征值4对应的$\mathbf{u}_2 = (0,1,0,0)^{\mathsf{T}}$，$\mathbf{v}_2 = (0,0,1,0)^{\mathsf{T}}$，

$\mathbf{u}_2\mathbf{v}_2^{\mathsf{T}}$定位矩阵$A$的（2,3）位置，$\mathbf{u}_3\mathbf{v}_3^{\mathsf{T}}$定位矩阵$A$的（1,2）

1阶逼近： $3\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

2阶逼近：$3\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}} + 2\mathbf{u}_2\mathbf{v}_2^{\mathsf{T}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

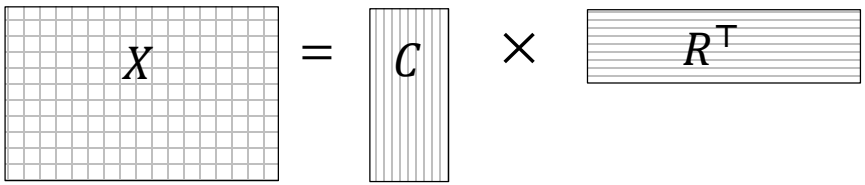
3阶即为原数据：$3\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}} + 2\mathbf{u}_2\mathbf{v}_2^{\mathsf{T}} + \mathbf{u}_3\mathbf{v}_3^{\mathsf{T}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix} = X$

矩阵分解和低秩逼近

矩阵分解

矩阵分解(matrix factorization)将矩阵分解成两个或多个矩阵的乘积，在新的基下重新表示矩阵的行或列向量：

$$\underset{n\times p}{X} = \underset{n\times k}{C}\,\underset{k\times p}{R^\top},\quad C\text{：列基，}\ R\text{：行基}\qquad rank(X)\le k$$



矩阵分解有任意多种，我们可以构造如下：

假设$X=(\mathbf{x}_1,\dots,\mathbf{x}_p)$, $rank(X)\le k$，假设$\mathbf{c}_1,\dots,\mathbf{c}_k$是$C(X)$的一组基（有冗余），$\mathrm{C}=(\mathbf{c}_1,\dots,\mathbf{c}_k)$，则存在$\mathbf{t}_i\in R^k$使得$\mathbf{x}_i=C\mathbf{t}_i$,

$$X=(\mathbf{x}_1,\dots,\mathbf{x}_p)=(C\mathbf{t}_1,\dots,C\mathbf{t}_p)=C(\mathbf{t}_1,\dots,\mathbf{t}_p)\triangleq CR^\top$$

其中$R^\top=(\mathbf{t}_1,\dots,\mathbf{t}_p)$是$k\times p$矩阵, R的列向量是X行空间的基。

当 $k<\mathrm{rank}(X)$ 时, $CR^\top$是X的秩k逼近(low rank approximation)

列基
$X = C\ R^\top$
行基

由上页讨论知，矩阵分解$X = C\,R^\top$将X的列向量表示成C的列向量的组合，同理，X的行向量表示为$R^\top$的行向量的线性组合，因此我们称C的列为列空间的基（虽然它们可能有冗余），R的列为行空间的基。

例1（PCA）假设$X_{n\times p}$已中心化, 样本协方差矩阵 $S = X^\top X/(n-1)$的谱分解$S = V\Lambda V^\top$, 主成分矩阵$Y = XV$, 则有矩阵分解（主成分变换的反变换）：

$$X = YV^\top$$

注意到Y是列正交的，即 $Y^\top Y = V^\top X^\top XV = (n-1)\Lambda \triangleq D^2$, 因此矩阵分解$X = YV^\top$的行基和列基都是正交基 – 这很难得，我们有理由相信该分解一定具有某些优良性质。

令$U = YD^{-1}$, 则 $U^\top U = I_p$,

$$\Rightarrow X = YV^\top = UDV^\top,$$

这正是中心化矩阵X的SVD, U或者$Y = UD$的列向量是列空间$C(X)$的正交基，V或者VD的列向量是行空间$C(X^\top)$的正交基。

我们后面将看到SVD/PCA在LS意义上具有优良逼近性质。

C,R的列:基

$$C_{n\times k}=(\mathbf{c}_1,\dots,\mathbf{c}_k)=\begin{pmatrix}\mathbf{s}_1^\top\\ \vdots\\ \mathbf{s}_n^\top\end{pmatrix},R_{p\times k}=(\mathbf{r}_1,\dots,\mathbf{r}_k)=\begin{pmatrix}\mathbf{t}_1^\top\\ \vdots\\ \mathbf{t}_p^\top\end{pmatrix}$$

C和R的列向量表示：$C_{n\times k}=(\mathbf{c}_1,\dots,\mathbf{c}_k),R_{p\times k}=(\mathbf{r}_1,\dots,\mathbf{r}_k)$，它们分别是$X$的列空间、行空间的一组基，$X=CR^\top$是$X$的秩1展开：

$$X=CR^\top=\mathbf{c}_1\mathbf{r}_1^\top+\cdots+\mathbf{c}_k\mathbf{r}_k^\top$$

其中$\{\mathbf{c}_i\mathbf{r}_i^\top,i=1,\dots,k\}$是$X$的基. 当$k\le rank(X)$较小时, $\mathbf{c}_1\mathbf{r}_1^\top+\cdots+\mathbf{c}_k\mathbf{r}_k^\top$是$X$的秩$k$逼近。

C,R 的行:潜变量

假设C和R的行向量表示为：$C_{n\times k}=\begin{pmatrix}\mathbf{s}_1^\top\\ \vdots\\ \mathbf{s}_n^\top\end{pmatrix},R_{p\times k}=\begin{pmatrix}\mathbf{t}_1^\top\\ \vdots\\ \mathbf{t}_p^\top\end{pmatrix}$，则

$$X=CR^\top=(\mathbf{s}_i^\top\mathbf{t}_j),\ x_{ij}=\mathbf{s}_i^\top\mathbf{t}_j$$

$$\mathbf{s}_i,\mathbf{t}_j\in R^k,1\le i\le n,1\le j\le p$$

$\mathbf{s}_i$代表第i个个体（X的第i行），$\mathbf{t}_j$代表第j个变量（X的第j列）。

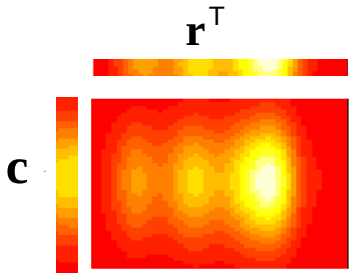
当$k<\min(n,p)$时，$\mathbf{t}_j$是X的第j列$\mathbf{x}_j$的某种压缩或潜变量表示，$\mathbf{s}_i$是X的第i行的某种压缩（类似于主成分的潜变量）。当$\mathbf{s}_i$与$\mathbf{t}_j$相似，$x_{ij}=\mathbf{s}_i^\top\mathbf{t}_j$较大（参见例3）。

例2. 若$X_{n\times p}$的秩为1，则X是两个向量的外积：

$$X = \mathbf{c}\mathbf{r}^\top$$

其中$\mathbf{c} \in R^n$, $\mathbf{r} \in R^p$分别是列空间和行空间的基，x_{ij}的大小由c_i，r_j的乘积决定（乘积模型）：

$$x_{ij} = c_i r_j$$



例3. 2006年Netflix推荐算法比赛的优胜者使用了矩阵分解推荐算法。假设基于用户U_i对已看电影M_j的打分评价为x_{ij}，记

$$X = (x_{ij})$$

每个用户只看少数电影并给出评价，因此X矩阵大部分位置是空的。我们的任务是填上这些空格，即预测用户对未看过的电影的评价，并对高评价预测的电影进行推荐。

	M1	M2	M3	M4
U1	5	3	-	1
U2	4	-	-	1
U3	1	1	-	5
U4	1	-	-	4
U5	-	1	5	4

来源：http://www.albertauyeung.com/post/python-matrix-factorization/

经典算法：协同过滤

协同过滤算法 (collaborative filtering) 假设对某些电影评价类似/相反的用户对其它电影也应有类似/相反的评价，是推荐系统中最常见的假设。

	M1	M2	M3	M4
U1	5	3	?	1
U2	4	?		1
U3	1	1		5
U4	1			4
U5		1	5	4

U2对M2的评价？

U1，U2都评价了M1和M4，打分类似。U1对M2的评价为3，我们推测U2对M2的评价大概为3

U1对M3的评价？

U1似乎与U3，4，5相反，U5的评价为5，故推测U1对M3的评价较小，比如1.

	M1	M2	M3	M4
U1	5	3	1	1
U2	4	2	1	1
U3	1	1	5	5
U4	1	1	4	4
U5	1	1	5	4

Netflix算法：矩阵分解

一个人对某电影评价高是因为其某些特征恰好符合该用户的喜好。

- 假设电影M_j有k个特征刻画(潜变量): $\mathbf{t}_j = (t_{j1}, \dots, t_{jk})^\top$；
- 假设用户U_i有k个特征（潜变量）：$\mathbf{s}_i = (s_{i1}, \dots, s_{ik})^\top$；
- 假设$\mathbf{s}_i$，$\mathbf{t}_j$的契合程度（内积）决定了U_i对M_j的评价：

$$x_{ij} \approx \sum_{f=1}^{k} (U_i\text{对特征}f\text{的喜好程度}) \times (M_j\text{中特征}f\text{的强度})$$
$$= \sum_{f=1}^{k} s_{if}\, t_{jf} = \mathbf{s}_i^\top \mathbf{t}_j$$

问题：

给定$X_{n\times p} = (x_{ij})$,给定$k \le rank(X)$，求解$\{\mathbf{s}_i, \mathbf{t}_j \in R^k\}$，极小化误差平方和：

$$J(C,R) = \sum_{i,j} (x_{ij} - \mathbf{s}_i^\top \mathbf{t}_j)^2 = \|X - CR^\top\|^2$$

其中记$C_{n\times k} = (\mathbf{s}_1, \dots, \mathbf{s}_n)^\top$, $R_{p\times k} = (\mathbf{t}_1, \dots, \mathbf{t}_p)^\top$。

- 如果矩阵X没有空值，则最优解为k阶SVD；
- 如果有空值，$J(C,R)$只对非空位置求和，梯度下降法求解。

k阶SVD: 矩阵的最佳秩k逼近

低秩逼近问题

给定$n\times p$矩阵X,求解秩为k的$n\times p$矩阵A，使得误差平方和最小:

$$\min_{rank(A)=k} \| X-A\|_F^2 = \min_{\substack{C_{n\times k},R_{p\times k}\\ \text{列满秩}}} \| X-CR^{\mathsf{T}} \|_F^2$$

其中矩阵的Frobenius 模定义为$\| A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{tr(A^{\mathsf{T}}A)}.$

- 上述低秩逼近的最优解是k-阶SVD（下页定理2）
- 对特定的问题，矩阵分解施加一些限制将会得到不同于SVD的其它方法。例如
  - 如果限制R的元素为0，1，代表k个类别的哑变量表示，得到的解将会把数据点聚集为k类（聚类分析）。
  - 若X元素非负，要求C，R元素也是非负，那么得到非负矩阵分解（non-negative matrix factorizaton，NMF），该方法因为容易解释且具有聚类效果而颇为流行。

SVD的最优性

定理2（Eckart - Young - Mirsky）.

若$X_{n\times p}$的奇异值分解为$X=UDV^{\mathsf{T}}=\sum_{i=1}^{r}\sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^{\mathsf{T}}$，$r=rank(X)$,则对任何$k\le r$,

$$\min_{\substack{A\in R^{n\times p}\\ rank(A)=k}} \|X-A\|_F^2=\|X-\sum_{i=1}^{k}\sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^{\mathsf{T}}\|_F^2=\sum_{i=k+1}^{r}\lambda_i$$

即X的最优秩k逼近为$A_{\text{opt}}=\sum_{i=1}^{k}\sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^{\mathsf{T}}=U_kD_kV_k^{\mathsf{T}}$，其中$U_k=(\mathbf{u}_1,\ldots,\mathbf{u}_k)$,

$V_k=(\mathbf{v}_1,\ldots,\mathbf{v}_k), D_k=diag(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_k})$。

证明：首先,对任何给定的$p\times k$矩阵R,不妨假设R是列正交且单位化的，即$R^{\mathsf{T}}R=I_k$,则必定有

$$\min_{C\in R^{n\times k}}\|X-CR^{\mathsf{T}}\|_F^2=\|X-XRR^{\mathsf{T}}\|_F^2 \qquad (*)$$

如果R不是列正交的，那么我们可以把R的列向量Schmidt正交化：$R=\tilde{R}T,\tilde{R}^{\mathsf{T}}\tilde{R}=I_k$,$T$是上三角，则$CR^{\mathsf{T}}=CT^{\mathsf{T}}\tilde{R}^{\mathsf{T}}\overset{\Delta}{=}\tilde{C}\tilde{R}^{\mathsf{T}}$

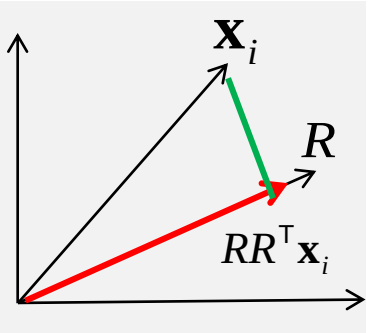
这是因为对于给定的R，$\|X-CR^{\mathsf{T}}\|_F^2=\sum_{i=1}^{n}\|\mathbf{x}_i-R\mathbf{c}_i\|^2$,其中

$X=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{\mathsf{T}}, C=(\mathbf{c}_1,\ldots,\mathbf{c}_n)^{\mathsf{T}}$。

由最小二乘性质，每一项$\|\mathbf{x}_i - R\mathbf{c}_i\|^2$ 在 $R\mathbf{c}_i = P_R\mathbf{x}_i$

$= R\left(R^{\mathsf{T}}R\right)^{-1}R^{\mathsf{T}}\mathbf{x}_i = RR^{\mathsf{T}}\mathbf{x}_i$时达到最小。

$$\Rightarrow \min\sum_{i=1}^{n}\|\mathbf{x}_i - R\mathbf{c}_i\|^2 = \sum_{i=1}^{n}\|\mathbf{x}_i - RR^{\mathsf{T}}\mathbf{x}_i\|^2 = \|X - XRR^{\mathsf{T}}\|_F^2 \text{。}$$



其次，注意到:

$$\begin{aligned}\|X - XRR^{\mathsf{T}}\|_F^2 &= \mathrm{tr}\left(X - XRR^{\mathsf{T}}\right)^{\mathsf{T}}\left(X - XRR^{\mathsf{T}}\right)\\ &= \mathrm{tr}\left(X^{\mathsf{T}}X - X^{\mathsf{T}}XRR^{\mathsf{T}} - RR^{\mathsf{T}}X^{\mathsf{T}}X + RR^{\mathsf{T}}X^{\mathsf{T}}XRR^{\mathsf{T}}\right)\\ &= \mathrm{tr}(X^{\mathsf{T}}X) - \mathrm{tr}(R^{\mathsf{T}}X^{\mathsf{T}}XR) = \|X\|_F^2 - \|XR\|_F^2\end{aligned}$$

所以

$$\min\|X - XRR^{\mathsf{T}}\|_F^2 \Leftrightarrow \max \quad \|XRR^{\mathsf{T}}\|_F^2 = \max\|XR\|_F^2 = \max tr(R^{\mathsf{T}}X^{\mathsf{T}}XR)$$

记R的各列$R = (\mathbf{r}_1,\ldots,\mathbf{r}_k)$,则 $R^{\mathsf{T}}X^{\mathsf{T}}XR = \left(\mathbf{r}_i^{\mathsf{T}}X^{\mathsf{T}}X\mathbf{r}_j^{\mathsf{T}}, 1 \le i, j \le k\right)$,所以

$$\|XR\|_F^2 = tr(R^{\mathsf{T}}X^{\mathsf{T}}XR) = \mathbf{r}_1^{\mathsf{T}}X^{\mathsf{T}}X\mathbf{r}_1 + \ldots + \mathbf{r}_k^{\mathsf{T}}X^{\mathsf{T}}X\mathbf{r}_k$$

因为$R^{\mathsf{T}}R = I_k$,所以$\mathbf{r}_1,\ldots,\mathbf{r}_k$是相互正交的单位长向量。

我们只需极大化$\| XR \|_F^2$ 的各个加项：

$$\max_{\|\mathbf{r}_1\|=1} \mathbf{r}_1^{\top} X^{\top} X \mathbf{r}_1,\ \ldots,\ \max_{\|\mathbf{r}_k\|=1, \mathbf{r}_k \perp \mathbf{r}_1,\ldots,\mathbf{r}_{k-1}} \mathbf{r}_k^{\top} X^{\top} X \mathbf{r}_k \qquad (**)$$

注意如果X已经中心化，
$X^{\top}X = (n-1)S$,
$(**)$ 是PCA的极大化问题。

由定理11.1，上述各项的最大值分别为$\lambda_1,\ldots,\lambda_k$,最大值在$X^{\top}X$的前$k$个特征向量$\mathbf{v}_1,\ldots,\mathbf{v}_k$达到，即最优$\hat{R} = V_k = (\mathbf{v}_1,\ldots,\mathbf{v}_k)$. 所以

$$\max \| X\hat{R} \|_F^2 = \| XV_k \|_F^2 = \lambda_1 + \ldots + \lambda_k.$$

$$\min \| X - XRR^{\top} \|_F^2 = \| X - XV_kV_k^{\top} \|_F^2 = tr(X^{\top}X) - \| XV_k \|_F^2 = \lambda_{k+1} + \ldots + \lambda_r$$

因为$XV_k = U_kD_k$，所以X的最佳逼近为

$$X\hat{R}\hat{R}^{\top} = XV_kV_k^{\top} = U_kD_kV_k^{\top} = \sqrt{\lambda_1}\mathbf{u}_1\mathbf{v}_1^{\top} + \ldots + \sqrt{\lambda_k}\mathbf{u}_k\mathbf{v}_k。$$

总结

低秩逼近问题的转化过程总结如下：

低秩逼近 $\min\limits_{rank(A)=k} \| X - A \|_F^2 \Leftrightarrow$

$$\min \| X - XRR^{\mathsf{T}} \|_F^2 \Leftrightarrow \max \| XR \|_F^2$$

$$\Leftrightarrow \max_{\mathbf{r}_1,\ldots,\mathbf{r}_k\text{正交，模长1}} \left(\mathbf{r}_1^{\mathsf{T}} X^{\mathsf{T}} X\mathbf{r}_1 + \ldots + \mathbf{r}_k^{\mathsf{T}} X^{\mathsf{T}} X\mathbf{r}_k\right)$$

$\Leftrightarrow$ 极大化加和中的每一项：

$$\max_{\|\mathbf{r}_1\|=1}\left(\mathbf{r}_1^{\mathsf{T}} X^{\mathsf{T}} X\mathbf{r}_1\right), \max_{\mathbf{r}_2\perp\mathbf{r}_1,\|\mathbf{r}_2\|=1}\left(\mathbf{r}_2^{\mathsf{T}} X^{\mathsf{T}} X\mathbf{r}_2\right), \ldots, \max_{\mathbf{r}_k\perp(\mathbf{r}_1,\ldots,\mathbf{r}_{k-1}),\|\mathbf{r}_k\|=1}\left(\mathbf{r}_k^{\mathsf{T}} X^{\mathsf{T}} X\mathbf{r}_k\right)$$

- 这些极大化问题的最优解为$X^{\mathsf{T}}X$的最大k个特征值对应的特征向量，最优的秩k矩阵为k阶SVD。
- 当X是中心化矩阵时，$X^{\mathsf{T}}X = (n-1)S, S$为样本协方差矩阵，该优化问题即是PCA问题。

SVD⇔PCA

中心化X的SVD等价于PCA：若X的SVD为$X = UDV^{\mathsf{T}}$，则V的各列是主成分方向，$Y = XV = UD$的第k列是第k主成分

奇异值分解的应用

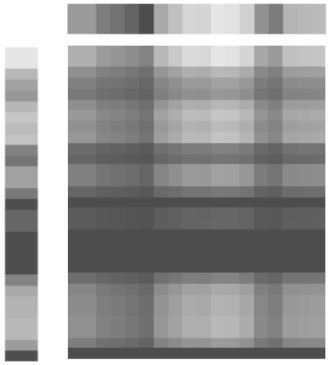
1.图像压缩

图像压缩中很少采用SVD逼近方法，我们这里采用SVD压缩图像数据知识为了演示目的。

例4. 下表29 × 18 矩阵是头像数据，元素为 0-1，1：白，0：黑。

```
111111111111111111
111111111111111111
111110000001110111
111100000000011111
111000000000001111
110000000111001111
110000001111100111
000000001111100111
000000011111100111
000000000010000011
000000000010000001
000000011011100000
000000011011100000
000000001001000000
000000000000000000
000000000001000000
000000000010000000
000000000000000000
000000000000000000
000000000000000000
000000000000000000
000000100110000000
000000111111000000
000000111111100000
000000111111100000
000000111111110000
000000111111110000
000000111110000000
000000000000000000
```

$1: \sqrt{\lambda_1}\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}}$



轮廓

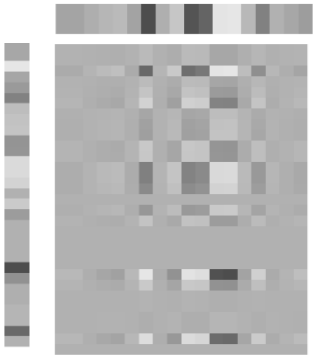
$2: \sqrt{\lambda_2}\mathbf{u}_2\mathbf{v}_2^{\mathsf{T}}$

脸部水平特征：
头发,眼睛,嘴巴。
竖直方向边际

$3: \sqrt{\lambda_3}\mathbf{u}_3\mathbf{v}_3^{\mathsf{T}}$

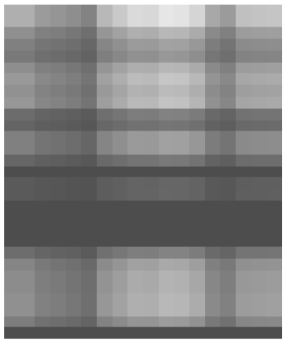
脸部垂直轮廓

$4: \sqrt{\lambda_4}\mathbf{u}_4\mathbf{v}_4^{\mathsf{T}}$

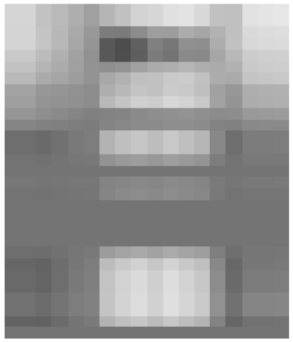


鼻子

1



$1+2$



$1+2+3$

$1+2+3+4$

2.网页搜索排序

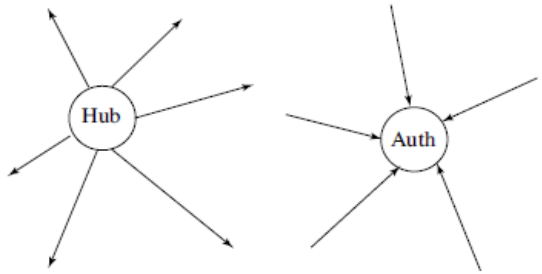
网页排序算法HITS与谷歌的PageRank都出现在1998-1999年，方法也有类似之处，都与SVD有关。

来源：Gilbert Strang (2016) **Introduction to Linear Algebra**, 5th ed

网页之间以超链接建立联系。定义邻接矩阵(adjacency matrix)

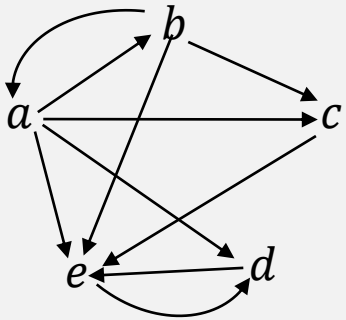
$$A=(a_{ij}),\qquad a_{ij}=\begin{cases}1 & \text{若}i\text{链接到}\ j\ (\text{记作}i\to j)\\ 0 & \text{否则}\end{cases}.$$

定义authority，hub：
一个网页被多个网页链入(link-in)，则其权威性Authority较大;
一个网页如果连接到(out-link)多个其它网页，则其hub较大。



黄页的Hub最大，它收集各种网页供用户查询（以前的门户网站，比如Yahoo!）

例5. 5个网页链接情况如图。网页a指向其它所有4个网页，是链出最多的网页；网页e被4个网页链接，是链入最多的网页。



link-in

Link-out

$$A = \begin{array}{c|ccccc|c} & a & b & c & d & e & \text{合计} \\ \hline a & 0 & 1 & 1 & 1 & 1 & 4 \\ b & 1 & 0 & 1 & 0 & 1 & 3 \\ c & 0 & 0 & 0 & 0 & 1 & 1 \\ d & 0 & 0 & 0 & 0 & 1 & 1 \\ e & 0 & 0 & 0 & 1 & 0 & 1 \\ \hline \text{合计} & 1 & 1 & 2 & 2 & 4 & \end{array}$$

如果以网页链入个数（矩阵的列和）作为authority重要性度量，则在这种度量下，e最受欢迎（4分）。网页c和d的得分都是2， 但网页d被权威性较高的e链入，故d的authority应该大于c。

所以在计算一个网页的权威性计分的时候，不但要考虑链入个数，也要考虑链入网页的权威性。

HITS 算法

HITS基本思想：

一个权威性网页应该有很多网页指向它，指向它的网页的hub重要性越高，该网页的权威性越高；而一个网页的hub重要性与它所指向的网页的权威性成正比:

$$\mathbf{v} \propto A^\top \mathbf{u},\ \mathbf{u} \propto A\mathbf{v}$$

其中$\mathbf{u}, \mathbf{v}$分别代表所有网页的hub和authority计分, A为邻接矩阵

解释：网页 j 的authority计分 v_j 与指向它的网页的hub计分总和成正比：

$$v_j = \alpha \sum_{k \to j} v_k = \alpha [A^\top \mathbf{u}]_j\ ,\ 即\mathbf{v} = \alpha A^\top \mathbf{u}$$

网页 i 的hub计分u_i与所指向的网页的authority计分总和成正比：

$$u_i = \beta \sum_{i \to k} v_k = \beta [A\mathbf{v}]_i,\ 即\mathbf{u} = \beta A\mathbf{v}$$

对偶方程$\mathbf{v} = \alpha A^\top \mathbf{u}, \mathbf{u} = \beta A\mathbf{v}\ \Rightarrow$

$$\mathbf{u} = \beta A\mathbf{v} = \alpha\beta AA^\top \mathbf{u},\ \ \mathbf{v} = \alpha A^\top \mathbf{u} = \alpha\beta A^\top A\mathbf{v}$$

故$\mathbf{u}, \mathbf{v}$分别是$AA^\top, A^\top A$的特征向量，分别是奇异值分解$A = UDV^\top$中U, V的某一列

计算$\mathbf{u}, \mathbf{v}$

- 对于大型矩阵A，幂次迭代：给定$\mathbf{v}^{(k)}$,

$$\mathbf{u}^{(k)} \leftarrow A\mathbf{v}^{(k)}, \mathbf{u}^{(k)} \leftarrow \mathbf{u}^{(k)}/\|\mathbf{u}^{(k)}\|;$$
$$\mathbf{v}^{(k+1)} \leftarrow A^{\top}\mathbf{u}^{(k)}, \mathbf{v}^{(k+1)} \leftarrow \mathbf{v}^{(k+1)}/\|\mathbf{v}^{(k+1)}\|$$

- 一般地，直接应用奇异值分解$A = UDV^{\top}$，取U, V的第一列分别为$\mathbf{u}, \mathbf{v}$的最优解。

例5(续) > svd(A)

```
$U     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 0.71 -0.53  0.00 -0.45  0.00
[2,] 0.56  0.53 -0.58  0.26  0.00
[3,] 0.28  0.27  0.58  0.13 -0.71
[4,] 0.28  0.27  0.58  0.13  0.71
[5,] 0.13 -0.53  0.00  0.84  0.00
```

```
$D
[1] 2.56 1.41 1 0.68  0
```

最大奇异值

```
$V
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 0.22  0.38 -0.58  0.38 -0.58
[2,] 0.28 -0.38  0.00 -0.67 -0.58
[3,] 0.50  0.00 -0.58 -0.29  0.58
[4,] 0.33 -0.76  0.00  0.57  0.00
[5,] 0.72  0.38  0.58  0.09  0.00
```

网页hub排序(U的第一列):
$a > b > c = d > e$

网页权威性排序 (V的第一列):
$e > c > d > b > a$

PageRank

HITS算法的缺点是将一个网页的Authority重要性只与到访网页的hub重要性关联，而没有与到访网页的Authority重要性直接关联。谷歌的PageRank算法也是只使用链接情况而不使用网页内容对网页排序， 但不区分auth,hub。

邻接矩阵A将所有网页之间的链接/关联性等同看待，但事实上一个网页对外链接越多（hub性质越强),它对提升被链接的网页的重要性的作用反而越弱。考虑到这一点，Page and Brin (1998) 把每一个链出网页的重要性平均分配到它链接的网页中。

PageRank的基本思想：

重要的网页应该链接重要的网页, 而且被对外链接较少的重要网页链接。这样一个自循环的描述其实就是特征向量在变换下的不变性。

PageRank：假设网页i的重要性计分为x_i,

$$x_i = \sum_{k:k\to i} \frac{x_k}{d_k}, \quad d_k = \text{邻接矩阵}A\text{的第}k\text{行的行和即网页}k\text{链出个数} \qquad (*)$$

网页i的x_i等于与之链接的所有网页的重要性x_k的加权和,权重为$1/d_k$。

记所有网页的重要性计分$\mathbf{x}=(x_1,...,x_n)^{\mathsf{T}}$,设$A$为邻接矩阵,令各个网站对外链接的个数为 $\mathbf{d}=(d_1,...,d_n)^{\mathsf{T}}$,即$\mathbf{d}=A\mathbf{1}_n$ （A的行和),

令 $D=diag(\mathbf{d})$, 则假设（*）为

$$\mathbf{x}=A^{\mathsf{T}}D^{-1}\mathbf{x} \qquad (**)$$

$\mathbf{x}$是$A^{\mathsf{T}}D^{-1}$的特征根1对应的特征向量,1是$A^{\mathsf{T}}D^{-1}$的最大特征根。

注. PageRank对H做了修正: $Google\ matrix\ \ G=\alpha H+(1-\alpha)\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n$

PageRank算法:

邻接矩阵为$A, D=diag(A\mathbf{1}_n), H=D^{-1}A, G=\alpha H+(1-\alpha)\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n$

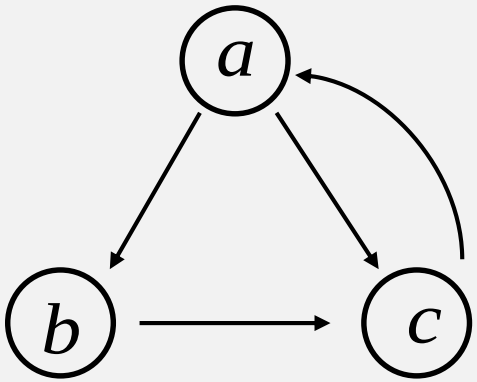
记x_i为网页i的重要性得分$(score)$, $PageRank$假设模型: $\mathbf{x}=G^{\mathsf{T}}\mathbf{x}$

迭代求解$\mathbf{x}$: $\mathbf{x}^{(k+1)}=G^{\mathsf{T}}\mathbf{x}^{(k)}, k=0,1,2,....$

收敛到最大特征根的特征向量。

Ranking page: $rank\ \ x_1,...,x_n$

例6. 假设如下三个网页



$$\text{链出}\quad A = \begin{array}{c|ccc} & \multicolumn{3}{c}{\text{链入}} \\ & a & b & c \\ \hline a & 0 & 1 & 1 \\ b & 0 & 0 & 1 \\ c & 1 & 0 & 0 \end{array}$$

1. HITS算法

权威性得分：$\mathbf{v} = (0,\ 0.53,\ 0.85)$,　权威性排序：$c > b > a$

2. PageRank算法

a, b, c的重要性得分：

$$x_a = x_c = 0.4 > x_b = 0.2,$$

即a, c同等重要：c有两个链入，比较重要；虽然a的链入只有一个，比c少1个，但该唯一的链入是从重要的c链入的，所以a也比较重要，至少比b重要。