PCA

探索PCA的数学思想

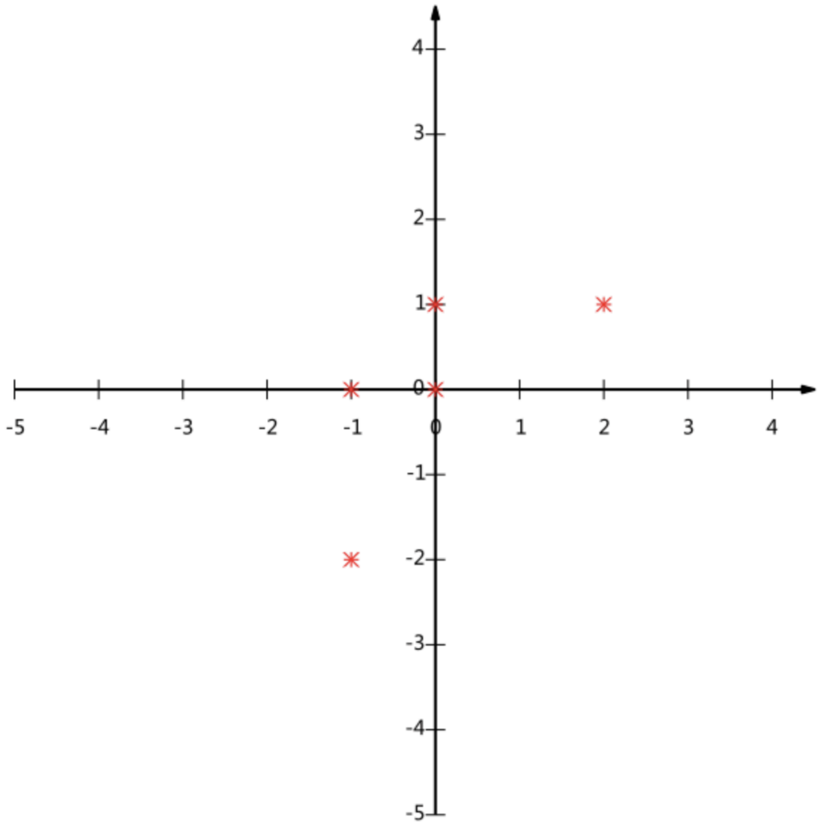
有一个样本集X，行表示样本，列表示属性，有m个样本，n个属性：

$$\begin{bmatrix} 1,1,3,2,5,2 \\ 1,3,3,2,9,5 \\ 2,3,6,4,5,3 \\ 4,4,9,8,4,7 \\ 2,4,6,4,8,9 \end{bmatrix}_{m\times n}$$

我们想把n维的初始样本转换到一个k维$(k < n)$的低维空间。

思路：选择k个基，最大程度保留原有的信息。

选择方差最大的基



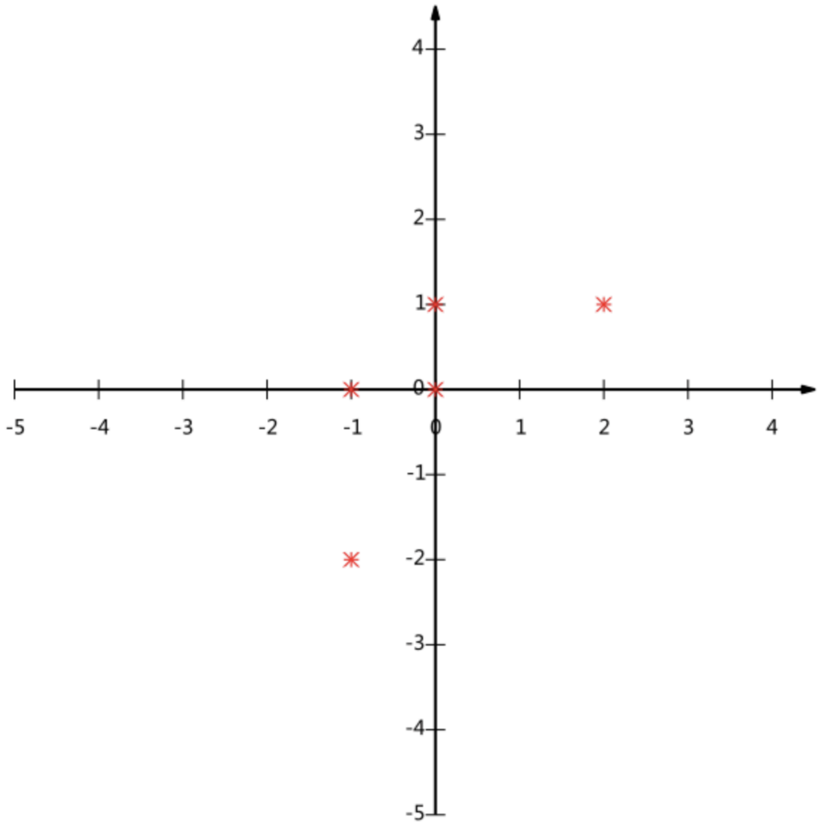
选择x轴，有2对样本重叠。

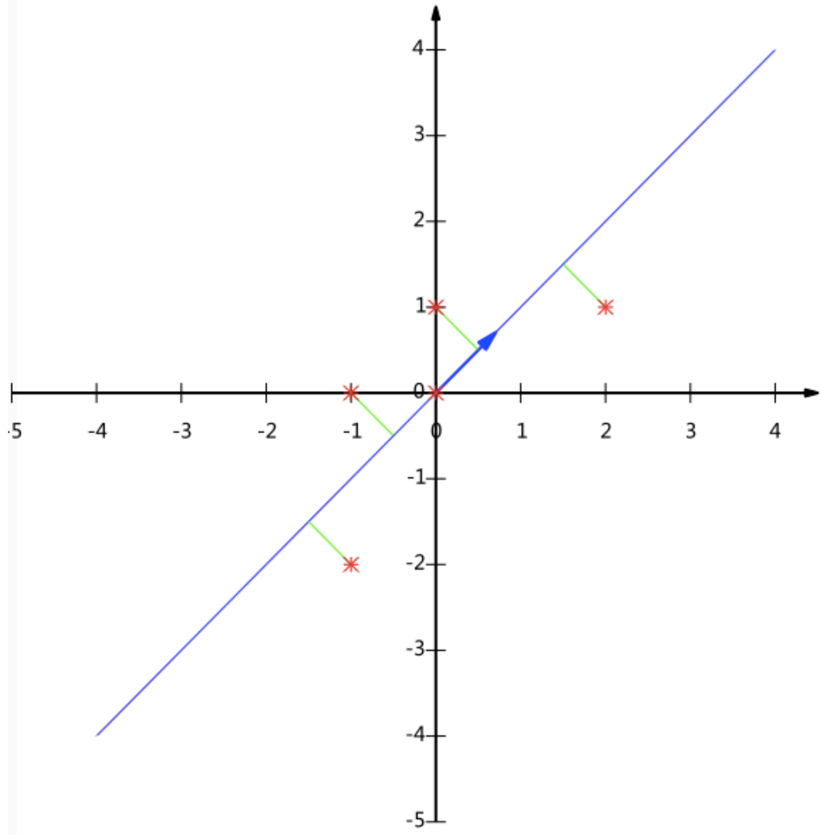
选择y轴，有2对样本重叠。

选择$y = x$，无样本重叠(投影后可区分)。

思路：选择离散程度(方差)最大的方向作为基

选择方差最大的基





选择协方差为0的基

问题：对高维空间，若每次都选择方差最大的方向，则每次选择的方向相同。

期望：每个基尽可能表示更多的原始信息，不存在相关性，相关性意味着重复表示。（正交基）

协方差：

数学上使用协方差表示相关性，若样本均值为0，则协方差：

$$Cov(a,b) = \frac{1}{m}\sum_{i=1}^{m} a_i b_i$$

当协方差为0时，两个字段独立，第二个基只能在与第一个基正交的方向上选择。

思路：选择相互正交的基(协方差为0)

协方差矩阵

假设只有a和b两个属性，每个属性均值为0的矩阵X为：

$$X = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \cdots & \cdots \\ a_m & b_m \end{pmatrix}_{m\times 2}$$

协方差矩阵：

$$C = \frac{1}{m}X^T X = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} a_i^2 & \frac{1}{m}\sum_{i=1}^{m} a_i b_i \\ \frac{1}{m}\sum_{i=1}^{m} a_i b_i & \frac{1}{m}\sum_{i=1}^{m} b_i^2 \end{pmatrix}_{2\times 2}$$

特点：矩阵对角线上为方差，其他元素是协方差，两者被统一到了一个矩阵内。

思路：协方差矩阵的对角化

协方差矩阵对角化

我们现在需要一个“转移矩阵”P，它的作用是将原样本X映射到新的低维空间坐标X'。

思路：特征值分解，求特征向量P和特征值矩阵Λ，其中Λ是对角阵。

求得P还不够，使用P转换到新空间的X'大小依然是$m \times n$。

思路：降维

开始降维

降维：将特征向量按对应特征值大小从左到右按列排列成矩阵，取前k列组成矩阵P。假设X维度$m\times n$，此时P的大小为$n\times k$，则新空间X'大小为$m\times k$ 。转换式为：

$$X' = XP$$

物理意义：协方差矩阵C'的对角线为方差，可视为信号的“能量”。以图像处理为例，“能量”高的地方聚集了图像的大部分特征，而“能量”低的地方往往是噪声，因此保留“能量”高的特征，舍去“能量”低的特征，既可以较为完整的保留原信号，又可以过滤噪声。

至此，PCA的流程就结束了。

重构（对比用）：$X_{\text{reconstruct}} = X'P^T = (m\times k)\cdot(k\times n) = (m\times n)$

PCA算法

假设X是按照“行表示样本，列表示属性”排列的，有m条n维属性，PCA算法流程如下：

1) 将X的每一列属性进行零均值化，即减去该列的均值。

2) 求出协方差矩阵

$$C = \frac{1}{m} X^T X$$

3) 求出协方差矩阵的特征值及对应的特征向量。

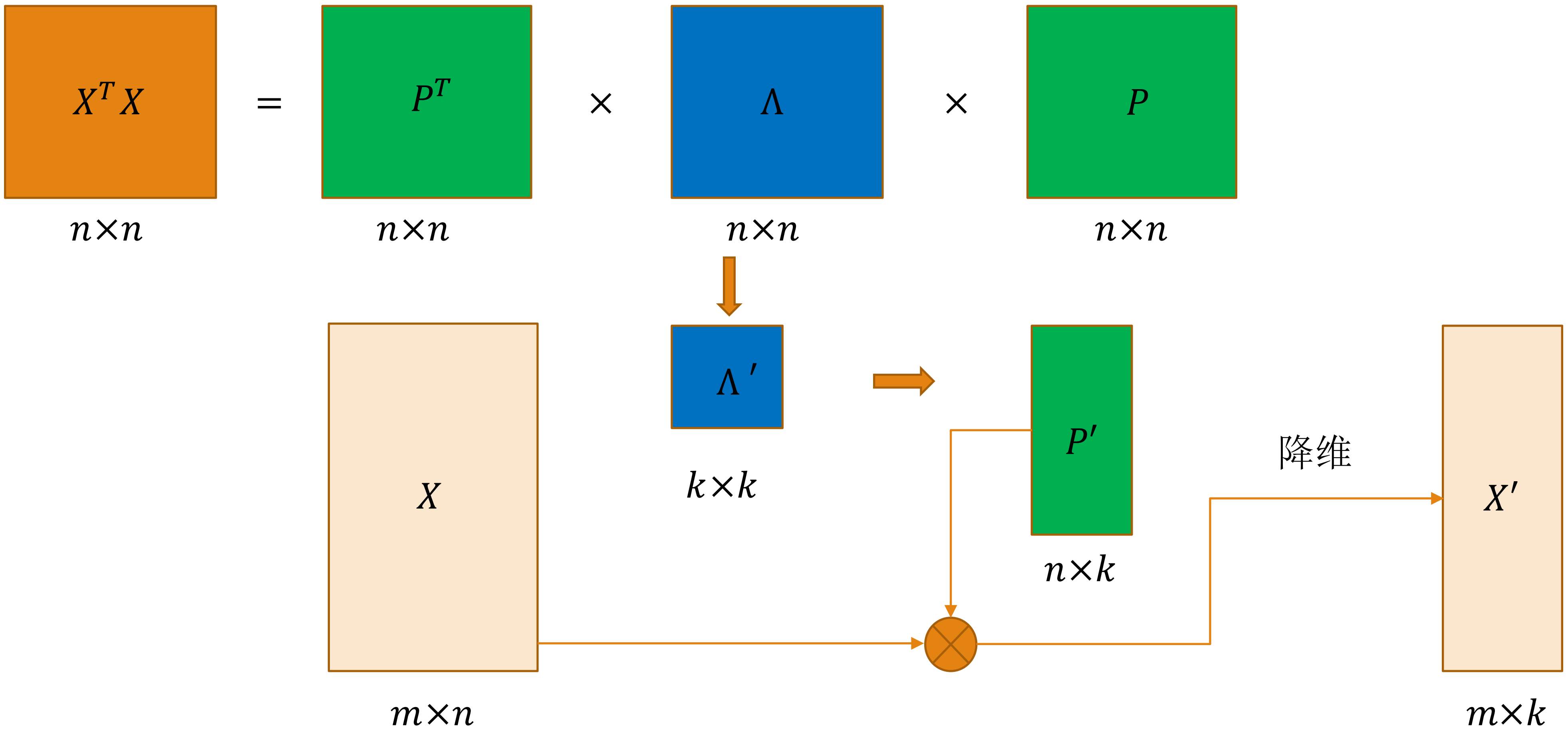
4) 将特征向量按对应特征值大小从左到右按列排列成矩阵，取前k列组成矩阵P。

5) 降维转换：

$$X' = XP$$

特征值分解与降维流程

注：特征值由大到小排列



SVD

奇异值分解

注：奇异值由大到小排列

奇异值分解

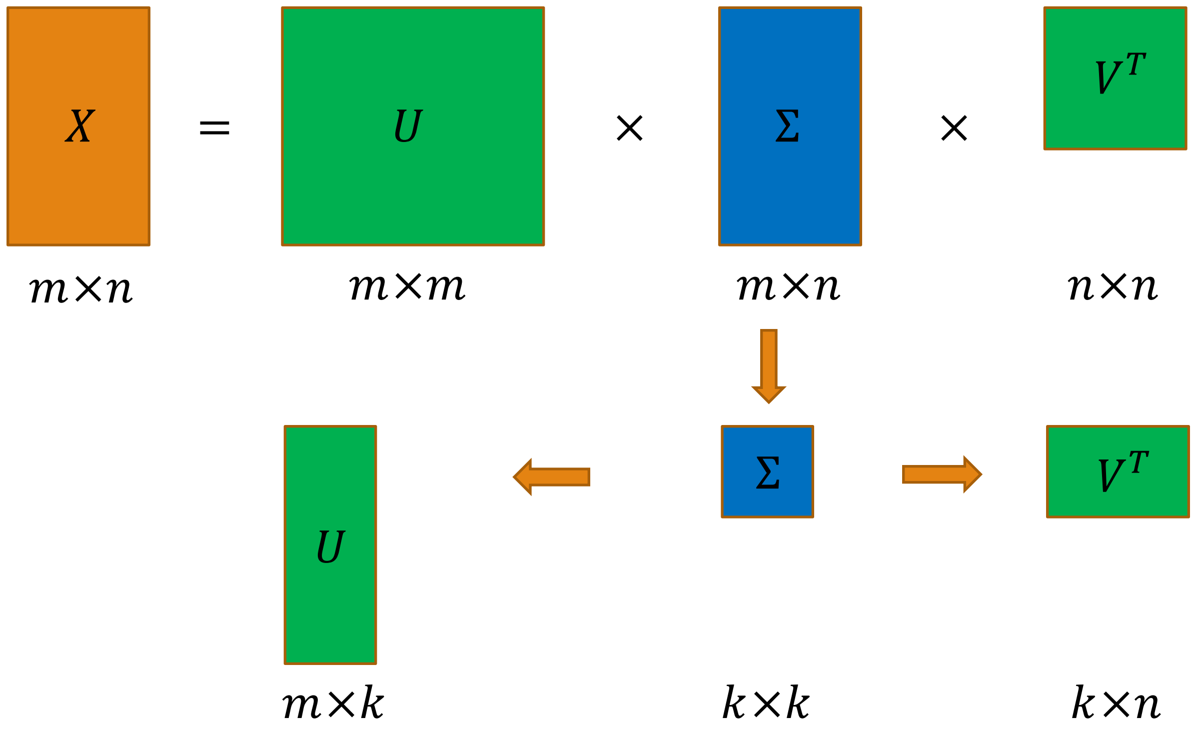
左奇异向量　　奇异值向量　　右奇异向量

$$\underset{m\times n}{X} = \underset{m\times m}{U} \times \underset{m\times n}{\Sigma} \times \underset{n\times n}{V^T}$$

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \\ 0 & 2 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{5}} & 0 & \frac{-2}{\sqrt{5}} & 0 \\ 0 & \frac{-1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{2}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{5} & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^{H}$$

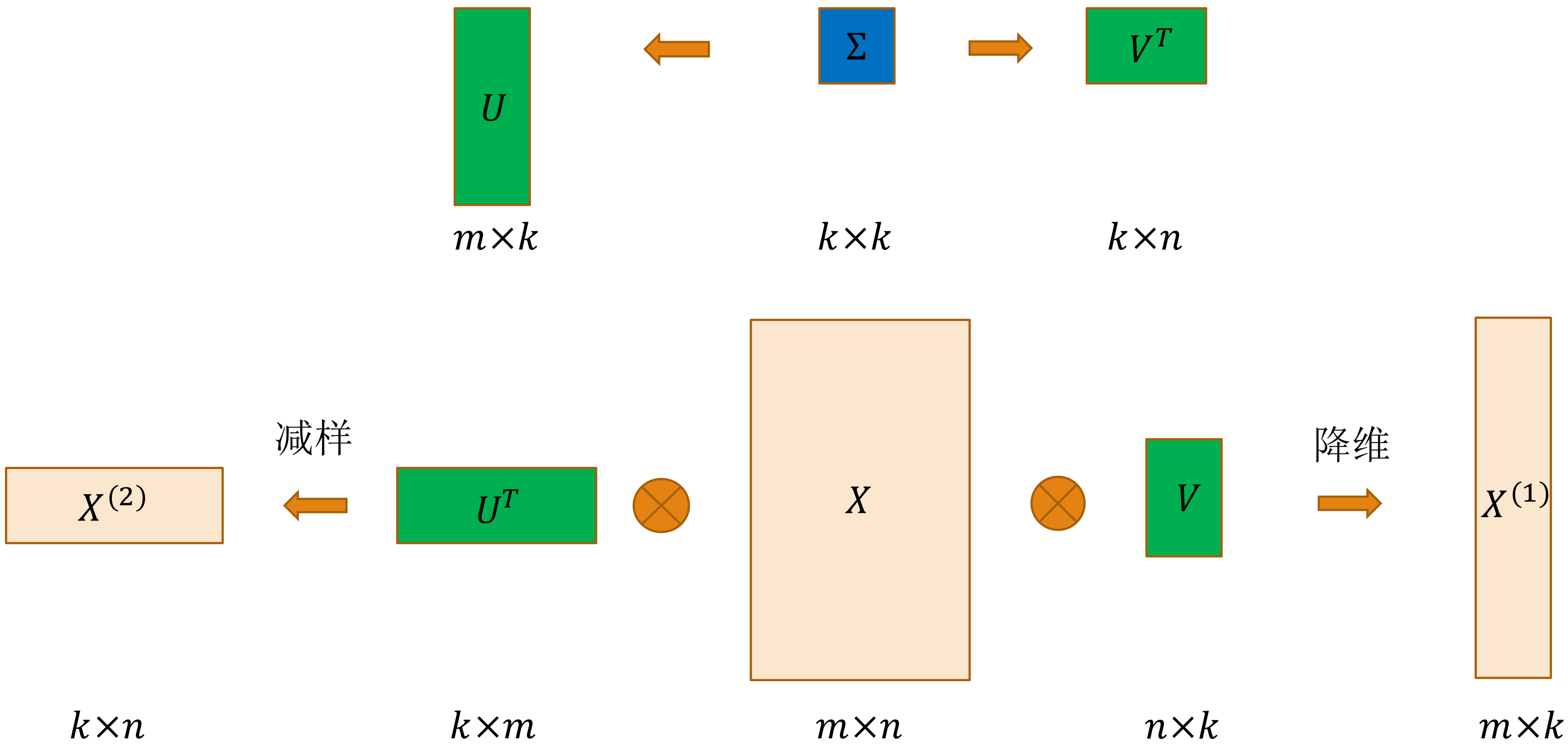
奇异值降维

注：奇异值由大到小排列



奇异值降维

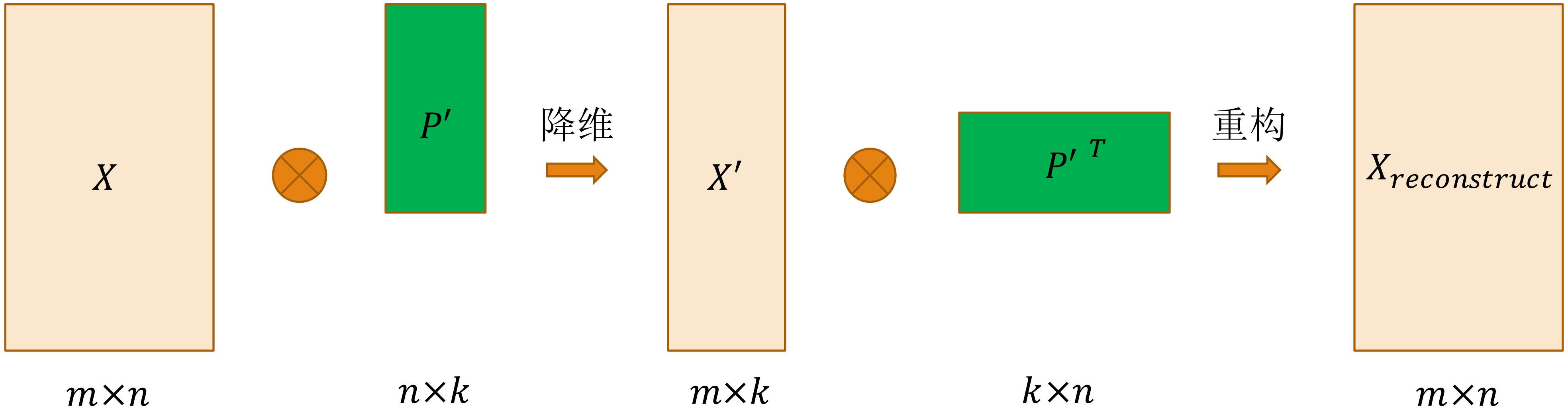
注：奇异值由大到小排列



PCA应用——图像压缩

特征值重构

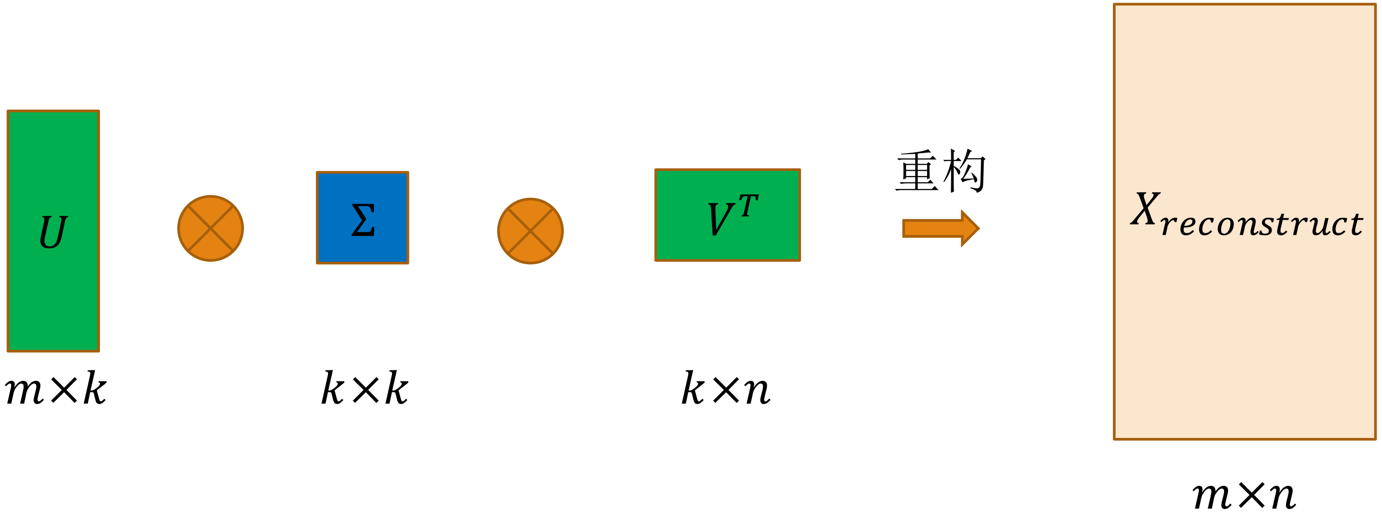
$$X \approx X_{reconstruct}$$



原存储空间：mn　　压缩存储空间：$(m+n)k$　　压缩比：$\frac{k}{m}+\frac{k}{n}$　　$k \ll m, n$

奇异值重构

$$X \approx X_{reconstruct}$$



原存储空间：mn　　压缩存储空间：$(m+n+k)k$　　压缩比：$\frac{k}{m}+\frac{k}{n}+\frac{k^2}{mn}$　　$k \ll m, n$