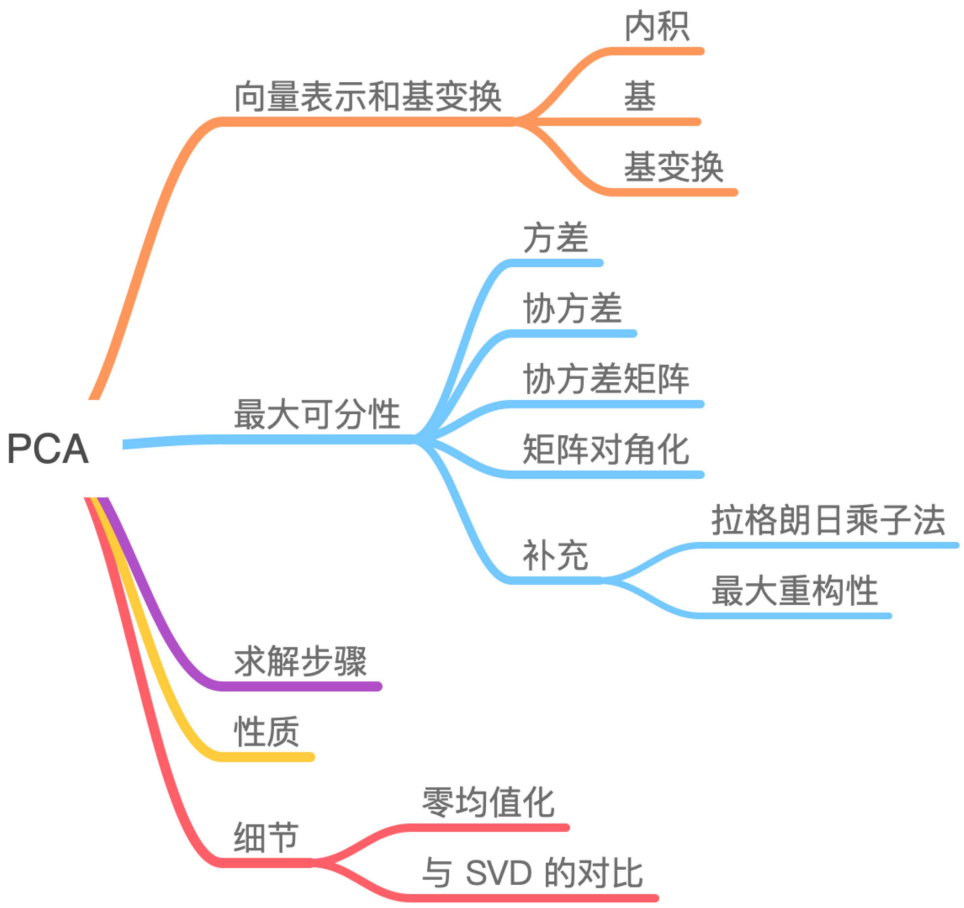


【机器学习】降维——PCA（非常详细）

阿泽
复旦大学 计算机技术硕士

PCA（Principal Component Analysis） 是一种常见的数据分析方式，常用于高维数据的降维，可用于提取数据的主要特征分量。

PCA 的数学推导可以从最大可分型和最近重构性两方面进行，前者的优化条件为划分后方差最大，后者的优化条件为点到划分平面距离最小，这里我将从最大可分性的角度进行证明。

1. 向量表示与基变换

我们先来介绍些线性代数的基本知识。

1.1 内积

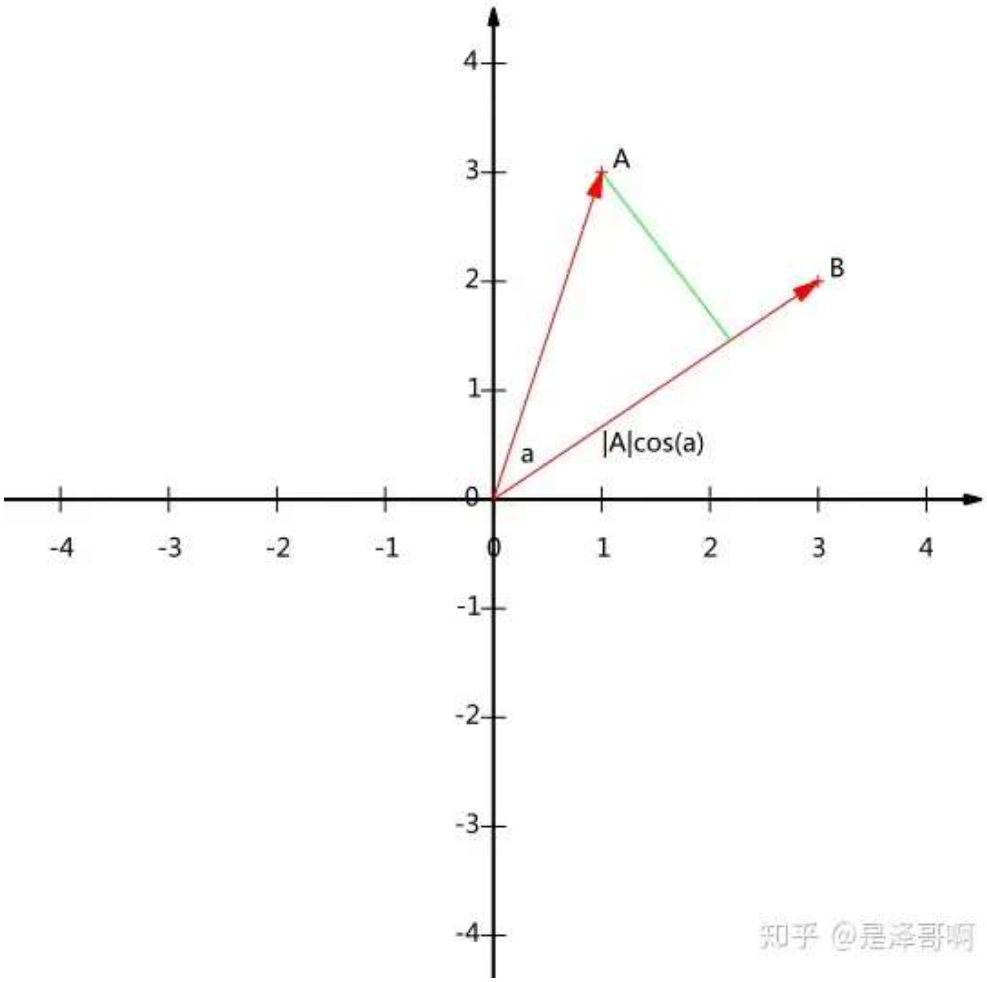
两个向量的 A 和 B 内积我们知道形式是这样的：

$$(a_1, a_2, \cdots, a_n) \cdot (b_1, b_2, \cdots, b_n)^{\mathrm{T}} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

内积运算将两个向量映射为实数，其计算方式非常容易理解，但我们无法看出其物理含义。接下来我们从几何角度来分析，为了简单起见，我们假设 A 和 B 均为二维向量，则：

$$A = (x_1, y_1),\ B = (x_2, y_2)\ A \cdot B = |A||B|cos(\alpha)$$

其几何表示见下图：



我们看出 A 与 B 的内积等于 A 到 B 的投影长度乘以 B 的模。

如果假设 B 的模为 1，即让 $|B| = 1$，那么就变成了：

$$A \cdot B = |A|cos(a)$$

也就是说，**A 与 B 的内积值等于 A 向 B 所在直线投影的标量大小。**

这就是内积的一种几何解释，也是我们得到的第一个重要结论。在后面的推导中，将反复使用这个结论。

1.2 基

在我们常说的坐标系种，向量 (3,2) 其实隐式引入了一个定义：以 x 轴和 y 轴上正方向长度为 1 的向量为标准。向量 (3,2) 实际是说在 x 轴投影为 3 而 y 轴的投影为 2。**注意投影是一个标量，所以可以为负。**

所以，对于向量 (3, 2) 来说，如果我们想求它在 $(1, 0), (0, 1)$ 这组基下的坐标的话，分别内积即可。当然，内积完了还是 (3, 2)。

所以，我们大致可以得到一个结论，我们**要准确描述向量，首先要确定一组基，然后给出在基所在的各个直线上的投影值，就可以了。**为了方便求坐标，我们希望这组基向量模长为 1。因为向量的内积运算，当模长为 1 时，内积可以直接表示投影。然后还需要这组基是线性无关的，我们一般用正交基，非正交的基也是可以的，不过正交基有较好的性质。

1.3 基变换的矩阵表示

这里我们先做一个练习：对于向量 (3,2) 这个点来说，在 $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ 和 $\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ 这组基下的坐标是多少？

我们拿 (3,2) 分别与之内积，得到 $\left(\frac{5}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ 这个新坐标。

我们可以用矩阵相乘的形式简洁的表示这个变换：

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 5/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

左边矩阵的两行分别为两个基，乘以原向量，其结果刚好为新基的坐标。推广一下，如果我们有 m 个二维向量，只要将二维向量按列排成一个两行 m 列矩阵，然后用“基矩阵”乘以这个矩阵就可以得到了所有这些向量在新基下的值。例如对于数据点 $(1,1)$，$(2,2)$，$(3,3)$ 来说，想变换到刚才那组基上，则可以这样表示：

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 2/\sqrt{2} & 4/\sqrt{2} & 6/\sqrt{2} \\ 0 & 0 & 0 \end{pmatrix}$$

我们可以把它写成通用的表示形式：

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix} \begin{pmatrix} a_1 & a_2 & \cdots & a_M \end{pmatrix} = \begin{pmatrix} p_1 a_1 & p_1 a_2 & \cdots & p_1 a_M \\ p_2 a_1 & p_2 a_2 & \cdots & p_2 a_M \\ \vdots & \vdots & \ddots & \vdots \\ p_R a_1 & p_R a_2 & \cdots & p_R a_M \end{pmatrix}$$

其中 p_i 是一个行向量，表示第 i 个基，a_j 是一个列向量，表示第 j 个原始数据记录。实际上也就是做了一个向量矩阵化的操作。

上述分析给矩阵相乘找到了一种物理解释：**两个矩阵相乘的意义是将右边矩阵中的每一列向量 a_i 变换到左边矩阵中以每一行行向量为基所表示的空间中去。**也就是说一个矩阵可以表示一种线性变换。

2. 最大可分性

上面我们讨论了选择不同的基可以对同样一组数据给出不同的表示，如果基的数量少于向量本身的维数，则可以达到降维的效果。

但是我们还没回答一个最关键的问题：如何选择基才是最优的。或者说，如果我们有一组 N 维向量，现在要将其降到 K 维（K 小于 N），那么我们应该如何选择 K 个基才能最大程度保留原有的信息？

一种直观的看法是：希望投影后的投影值尽可能分散，因为如果重叠就会有样本消失。当然这个也可以从熵的角度进行理解，熵越大所含信息越多。

2.1 方差

我们知道数值的分散程度，可以用数学上的方差来表述。一个变量的方差可以看做是每个元素与变量均值的差的平方和的均值，即：

$$Var(a) = \frac{1}{m} \sum_{i=1}^{m} (a_i - \mu)^2$$

为了方便处理，我们将每个变量的均值都化为 0 ，因此方差可以直接用每个元素的平方和除以元素个数表示：

$$Var(a) = \frac{1}{m}\sum_{i=1}^{m} a_i^2$$

于是上面的问题被形式化表述为：**寻找一个一维基，使得所有数据变换为这个基上的坐标表示后，方差值最大。**

2.2 协方差

在一维空间中我们可以用方差来表示数据的分散程度。而对于高维数据，我们用协方差进行约束，协方差可以表示两个变量的相关性。为了让两个变量尽可能表示更多的原始信息，我们希望它们之间不存在线性相关性，因为相关性意味着两个变量不是完全独立，必然存在重复表示的信息。

协方差公式为：

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

由于均值为 0，所以我们的协方差公式可以表示为：

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

当样本数较大时，不必在意其是 m 还是 m-1，为了方便计算，我们分母取 m。

当协方差为 0 时，表示两个变量完全独立。为了让协方差为 0，我们选择第二个基时只能在与第一个基正交的方向上进行选择，因此最终选择的两个方向一定是正交的。

（**2020 年 12 月 15 日补充**：协方差为 0 时，两个变量只是线性不相关。完全独立是有问题的，才疏学浅，还望见谅。）

至此，我们得到了降维问题的优化目标：**将一组 N 维向量降为 K 维，其目标是选择 K 个单位正交基，使得原始数据变换到这组基上后，各变量两两间协方差为 0，而变量方差则尽可能大（在正交的约束下，取最大的 K 个方差）。**

2.3 协方差矩阵

针对我们给出的优化目标，接下来我们将从数学的角度来给出优化目标。

我们看到，最终要达到的目的与**变量内方差及变量间协方差**有密切关系。因此我们希望能将两者统一表示，仔细观察发现，两者均可以表示为内积的形式，而内积又与矩阵相乘密切相关。于是我们有：

假设我们只有 a 和 b 两个变量，那么我们将它们按行组成矩阵 X：

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

然后：

$$\frac{1}{m}XX^{\mathsf{T}} = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} a_i^2 & \frac{1}{m}\sum_{i=1}^{m} a_i b_i \\ \frac{1}{m}\sum_{i=1}^{m} a_i b_i & \frac{1}{m}\sum_{i=1}^{m} b_i^2 \end{pmatrix} = \begin{pmatrix} Cov(a,a) & Cov(a,b) \\ Cov(b,a) & Cov(b,b) \end{pmatrix}$$

我们可以看到这个矩阵对角线上的分别是两个变量的方差，而其它元素是 a 和 b 的协方差。两者被统一到了一个矩阵里。

我们很容易被推广到一般情况：

设我们有 m 个 n 维数据记录，将其排列成矩阵 $X_{n,m}$ ，设 $C=\frac{1}{m}XX^T$ ，则 C 是一个对称矩阵，其对角线分别对应各个变量的方差，而第 i 行 j 列和 j 行 i 列元素相同，表示 i 和 j 两个变量的协方差。

2.4 矩阵对角化

根据我们的优化条件，**我们需要将除对角线外的其它元素化为 0，并且在对角线上将元素按大小从上到下排列（变量方差尽可能大）**，这样我们就达到了优化目的。这样说可能还不是很明晰，我们进一步看下原矩阵与基变换后矩阵协方差矩阵的关系。

设原始数据矩阵 X 对应的协方差矩阵为 C，而 P 是一组基按行组成的矩阵，设 Y=PX，则 Y 为 X 对 P 做基变换后的数据。设 Y 的协方差矩阵为 D，我们推导一下 D 与 C 的关系：

$$\begin{aligned} D &= \frac{1}{m}YY^T \\ &= \frac{1}{m}(PX)(PX)^T \\ &= \frac{1}{m}PXX^TP^T \\ &= P(\frac{1}{m}XX^T)P^T \\ &= PCP^T \end{aligned}$$

这样我们就看清楚了，我们要找的 P 是能让原始协方差矩阵对角化的 P。换句话说，优化目标变成了**寻找一个矩阵 P，满足 PCP^T 是一个对角矩阵，并且对角元素按从大到小依次排列，那么 P 的前 K 行就是要寻找的基，用 P 的前 K 行组成的矩阵乘以 X 就使得 X 从 N 维降到了 K 维并满足上述优化条件。**

至此，我们离 PCA 还有仅一步之遥，我们还需要完成对角化。

由上文知道，协方差矩阵 C 是一个是对称矩阵，在线性代数中实对称矩阵有一系列非常好的性质：

1. 实对称矩阵不同特征值对应的特征向量必然正交。
2. 设特征向量 λ 重数为 r，则必然存在 r 个线性无关的特征向量对应于 λ ，因此可以将这 r 个特征向量单位正交化。

由上面两条可知，一个 n 行 n 列的实对称矩阵一定可以找到 n 个单位正交特征向量，设这 n 个特征向量为 $e_1,e_2,\cdots,e_n$ ，我们将其按列组成矩阵： $E=(e_1,e_2,\cdots,e_n)$ 。

则对协方差矩阵 C 有如下结论：

$$E^TCE=\Lambda=\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

其中 Λ 为对角矩阵，其对角元素为各特征向量对应的特征值（可能有重复）。

到这里，我们发现我们已经找到了需要的矩阵 P： $P=E^T$ 。

P 是协方差矩阵的特征向量单位化后按行排列出的矩阵，其中每一行都是 C 的一个特征向量。如果设 P 按照 $\mathbf{\Lambda}$ 中特征值的从大到小，将特征向量从上到下排列，则用 P 的前 K 行组成的矩阵乘以原始数据矩阵 X，就得到了我们需要的降维后的数据矩阵 Y。

2.5 补充

(1) 拉格朗日乘子法

在叙述求协方差矩阵对角化时，我们给出希望变化后的变量有：**变量间协方差为 0 且变量内方差尽可能大**。然后我们通过实对称矩阵的性质给予了推导，此外我们还可以把它转换为最优化问题利用拉格朗日乘子法来给予推导。

我们知道样本点 $\boldsymbol{x_i}$ 在基 w 下的坐标为： $(\boldsymbol{x_i}, \boldsymbol{w}) = \boldsymbol{x_i^T w}$ ，于是我们有方差：

$$\begin{aligned} D(x) &= \frac{1}{m}\sum_{i=1}^{m}(x_i^T w)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}(x_i^T w)^T(x_i^T w) \\ &= \frac{1}{m}\sum_{i=1}^{m} w^T x_i x_i^T w \\ &= w^T(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T)w \end{aligned}$$

我们看到 $\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x_i x_i^T}$ 就是原样本的协方差，我们另这个矩阵为 $\mathbf{\Lambda}$ ，于是我们有：

$$\begin{cases} max\{w^T \Lambda w\} \\ s.t. w^T w = 1 \end{cases}$$

.

然后构造拉格朗日函数：

$$L(w) = w^T \Lambda w + \lambda(1 - w^T w)$$

对 w 求导：

$$\Lambda w = \lambda w$$

此时我们的方差为：

$$D(x) = w^T \Lambda w = \lambda w^T w = \lambda$$

于是我们发现，x 投影后的方差就是协方差矩阵的特征值。我们要找到最大方差也就是协方差矩阵最大的特征值，最佳投影方向就是最大特征值所对应的特征向量，次佳就是第二大特征值对应的特征向量，以此类推。

至此我们完成了基于最大可分性的 PCA 数学证明

(2) 最近重构性

以上的证明思路主要是基于最大可分性的思想，**通过一条直线使得样本点投影到该直线上的方差最大**。除此之外，我们还可以**将其转换为线型回归问题，其目标是求解一个线性函数使得对应直线能够更好地拟合样本点集合**。这就**使得我们的优化目标从方差最大转化为平方误差最小**，因为映射距离越短，丢失的信息也会越小。区别于最大可分性，这是从最近重构性的角度进行论证。

3. 求解步骤

总结一下 PCA 的算法步骤：

设有 m 条 n 维数据。

1. 将原始数据按列组成 n 行 m 列矩阵 X；
2. 将 X 的每一行进行零均值化，即减去这一行的均值；
3. 求出协方差矩阵 $C = \frac{1}{m}XX^{\mathrm{T}}$ ；
4. 求出协方差矩阵的特征值及对应的特征向量；
5. 将特征向量按对应特征值大小从上到下按行排列成矩阵，取前 k 行组成矩阵 P；
6. $Y = PX$ 即为降维到 k 维后的数据。

4. 性质

1. **缓解维度灾难**：PCA 算法通过舍去一部分信息之后能使得样本的采样密度增大（因为维数降低了），这是缓解维度灾难的重要手段；
2. **降噪**：当数据受到噪声影响时，最小特征值对应的特征向量往往与噪声有关，将它们舍弃能在一定程度上起到降噪的效果；
3. **过拟合**：PCA 保留了主要信息，但这个主要信息只是针对训练集的，而且这个主要信息未必是重要信息。有可能舍弃了一些看似无用的信息，但是这些看似无用的信息恰好是重要信息，只是在训练集上没有很大的表现，所以 PCA 也可能加剧了过拟合；
4. **特征独立**：PCA 不仅将数据压缩到低维，它也使得降维之后的数据各特征相互独立；

5. 细节

5.1 零均值化

当对训练集进行 PCA 降维时，也需要对验证集、测试集执行同样的降维。而**对验证集、测试集执行零均值化操作时，均值必须从训练集计算而来**，不能使用验证集或者测试集的中心向量。

其原因也很简单，因为我们的训练集时可观测到的数据，测试集不可观测所以不会知道其均值，而验证集再大部分情况下是在处理完数据后再从训练集中分离出来，一般不会单独处理。如果真的是单独处理了，不能独自求均值的原因是和测试集一样。

另外我们也需要保证一致性，我们拿训练集训练出来的模型用来预测测试集的前提假设就是两者是独立同分布的，如果不能保证一致性的话，会出现 Variance Shift 的问题。

5.2 与 SVD 的对比

这是两个不同的数学定义。我们先给结论：**特征值和特征向量是针对方阵**才有的，而**对任意形状的矩阵都可以做奇异值分解。**

PCA：方阵的特征值分解，对于一个方针 A，总可以写成： $A = Q\Lambda Q^{-1}$ 。

其中，Q 是这个矩阵 A 的特征向量组成的矩阵， Λ 是一个对角矩阵，每一个对角线元素就是一个特征值，里面的特征值是由小排列的，这些特征值所对应的特征向量就是描述这个矩阵变化方向（从主要的变化到次要的变化排列）。也就是说矩阵 A 的信息可以由其特征值和特征向量表示。

SVD：矩阵的奇异值分解其实就是对于矩阵 A 的协方差矩阵 A^TA 和 AA^T 做特征值分解推导出来的：

$$A_{m,n} = U_{m,m}\Lambda_{m,n}V^T_{n,n} \approx U_{m,k}\Lambda_{k,k}V^T_{k,n}$$

其中：U V 都是正交矩阵，有 $\boldsymbol{U^TU = I_m, V^TV = I_n}$ 。这里的约等于是因为 $\boldsymbol{\Lambda}$ 中有 n 个奇异值，但是由于排在后面的很多接近 0，所以我们可以仅保留比较大的 k 个奇异值。

$$\boldsymbol{A^TA = (U\Lambda V^T)^T U\Lambda V^T = V\Lambda^T U^T U\Lambda V^T = V\Lambda^2 V^T}$$
$$\boldsymbol{AA^T = U\Lambda V^T (U\Lambda V^T)^T = U\Lambda V^T V\Lambda^T U^T = U\Lambda^2 U^T}$$

所以，V U 两个矩阵分别是 $\boldsymbol{A^TA}$ 和 $\boldsymbol{AA^T}$ 的特征向量，中间的矩阵对角线的元素是 $\boldsymbol{A^TA}$ 和 $\boldsymbol{AA^T}$ 的特征值。我们也很容易看出 A 的奇异值和 $\boldsymbol{A^TA}$ 的特征值之间的关系。

PCA 需要对协方差矩阵 $\boldsymbol{C = \frac{1}{m}XX^T}$ 。进行特征值分解； SVD 也是对 $\boldsymbol{A^TA}$ 进行特征值分解。如果取 $\boldsymbol{A = \frac{X^T}{\sqrt{m}}}$ 则两者基本等价。所以 PCA 问题可以转换成 SVD 求解。

而实际上 Sklearn 的 PCA 就是用 SVD 进行求解的，原因有以下几点：

1. 当样本维度很高时，协方差矩阵计算太慢；
2. 方阵特征值分解计算效率不高；
3. SVD 除了特征值分解这种求解方式外，还有更高效更准球的迭代求解方式，避免了 $\boldsymbol{A^TA}$ 的计算；
4. 其实 PCA 与 SVD 的右奇异向量的压缩效果相同。

6. 参考

1. 《机器学习》周志华
2. PCA 的数学原理
3. Singular Value Decomposition (SVD) tutorial
4. 机器学习中的数学（4）——线性判别分析（LDA），主成分分析（PCA）
5. 从SVD到PCA——奇妙的数学游戏