浅析主成分分析方法的数学原理
Principles of PCA Method: A Survey

2190400401 陈瀚

摘要:

主成分分析，或者称为 PCA ，是数理统计和机器学习中被广泛使用的技术，应用的领域包括维度降低、有损数据压缩、特征抽取、数据可视化(Jolliffe, 2002)。它也被称 Karhunen-Loève 变换。PCA 可以被定义为数据在低维线性空间上的正交投影的方差最大化情况，也可定义为投影误差的平方和的最小值情况。本文从数学角度将探讨该方法的原理。

1. 概要

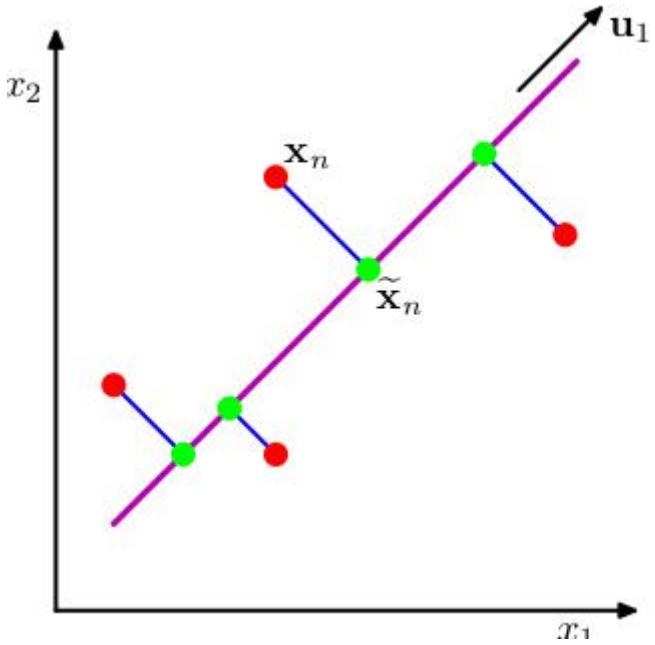

图 1

PCA 可以被定义为数据在低维线性空间上的正交投影，这个线性空间被称为主子空间(principal subspace)，使得投影数据的方差被最大化(Hotelling, 1933)。等价

地，它也可以被定义为使得平均投影代价最小的线性投影。平均投影代价是指数据点和它们的投影之间的平均平方距离(Pearson, 1901)。

正交投影的过程如图 1 所示。

主成分分析寻找一个低维空间，被称为主子平面，用紫色的线表示，使得数据点(红点)在子空间上的正交投影能够最大化投影点(绿点)的方差。 PCA 的另一个定义基于的是投影误差的平方和的最小值，用蓝线表示。

2. 形式 1：最大方差

考虑一组观测数据集x_n，其中 n=1, . . . , N，因此X_n是一个 D 维欧几里得空间中的变量。我们的目标是将数据投影到维度 M < D 的空间中，同时最大化投影数据的方差。现阶段，假设 M 的值是给定的。

首先，考虑在一维空间$M = 1$ 上的投影。我们可以使用 D 维向量u_1定义这个空间的方向。为了方便(并且不失一般性)，我们假定选择一个单位向量，从而${u_1}^T u_1 = 1$ 。我们只对u_1的方向感兴趣，而对 u1 本身的大小不感兴趣)。这样，每个数据点x_n被投影到一个标量${u_1}^T x_n$上。其中$\overline{x}$是样本集合的均值。如下：

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

数据的协方差矩阵，为：

$$\mathrm{S} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \overline{x})(x_n - \overline{x})^T$$

投影数据的方差，如下：

$$u_1^T S u_1 = \frac{1}{N}\sum_{n=1}^{N} \{u_1^T x_n - u_1^T \overline{x}\}^2$$

现在的工作是将上面的$u_1^T S u_1$相对u_1求得最大值。有什么限制？明显：要避免$\|u_1\| \to \infty$。我们通过归一化条件$u_1{}^T u_1 = 1$ 达到这一点。引入拉格朗日乘数，记作λ_1，然后对下式进行最大化：

$$u_1^T S u_1 + \lambda_1(1 - u_1^T u_1)$$

求解关于u_1的导数等于零的情况，可知驻点满足：

$$S u_1 = \lambda_1 u_1$$

表明u_1一定是的S一个特征向量。如果我们左乘u_1^T，我们看到方差为：

$$u_1^T S u_1 = \lambda_1$$

因此将u_1设置为与具有最大的特征值λ_1的特征向量相等时，方差会达到最大值。大部分文献将这个特征向量称为第一主成分(First Principal Component， 或 PC1)。

可以用一种增量的方式定义额外的主成分，方法为：在所有与那些已经考虑过的方向正交的所有可能的方向中，将新的方向选择为最大化投影方差的方向。如果我们考虑 M 维投影空间的一般情形，那么最大化投影数据方差的最优线性投影由数据协方差矩阵 S 的 M 个特征向量 $u_1, \ldots, u_M$定义，对应于 M 个最大的特征值λ_1,...,λ_M。可以通过归纳法很容易地证明出来。

总结而言，主成分分析寻找均值和协方差矩阵，而后求解 S 的对应于 M 个最大特征值的 M 个特征向量。在实际应用中，鉴于 D × D 矩阵的完整的特征向量分解的代价为$O(D^3)$，时间复杂度过高。如能将数据投影到前 M 个主成分中，那么我们只需寻找前 M 个特征值和特征向量。这可以使用更高效的方法得到，例如幂方法、 EM(期望最大化)算法。

3. 形式 2：最小误差

PCA 也可解释为基于误差最小化的投影技术。为了完成这一点，我们引入 D 维基向量的一个完整的单位正交集合 $\{u_i\}$，其中 i= 1,...,D ，满足：

$$u_i^T u_j = \delta_{ij}$$

由于基是完整的，因此每个数据点可以精确地表示为基向量的一个线性组合:

$$x_n = \sum_{i=i}^{D} a_{ni} \boldsymbol{u_i}$$

其中，系数a_{ni}对于不同的数据点来说是不同的。这对应于将坐标系旋转到了一个由$\{u_j\}$定义的新坐标系，原始的 D 个分量$\{x_{n1,\ldots,} x_{nD}\}$被替换为一个等价的集合$\{a_{n1,\ldots,} a_{nD}\}$。与$\{u_i\}$做内积，由单位正交性质有$a_{nj} = x_n^T u_j$。不失一般性，有：

$$x_n = \sum_{i=1}^{D} (x_n^T u_j) \boldsymbol{u_i}$$

我们的目标是使用限定数量 M < D 个变量的一种表示方法来近似数据点，这对应在低维子空间上的一个投影。不失一般性， M 维线性子空间可以用前 M 个基向量表示，因此我们可以用下式来近似每个数据点x_n：

$$\tilde{x}_n = \sum_{i=1}^{M} z_{ni} \boldsymbol{u_i} + \sum_{j=M+1}^{D} b_j \boldsymbol{u_j}$$

其中$\{z_{ni}\}$依赖于特定的数据点，而b_i是常数，对于所有数据点都相同。我们可以任意选择$\{u_i\}$, $\{z_{ni}\}$和$\{b_i\}$，从而最小化由维度降低所引入的失真。作为失真的度量，我们使用原始数据点与它的近似点$\tilde{x}_n$之间的平方距离，在数据集上取平均。因此我们的目标是最小化：

$$J = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \|\tilde{x}_n - x_n\|^2$$

首先考虑关于z_{ni}的最小化。消去$\tilde{x}_n$，令它关于z_{nj}的导数为零，由单位正交的条件，有：

$$z_{nj} = x_n^T u_j$$

其中 $\mathrm{j} = 1, \ldots, \mathrm{M}$。类似地，令 J 关于$b_i$的导数等于零，再次使用单位正交的关系，我们有：

$$\mathrm{b_j} = \widetilde{x}^T u_j$$

其中 $\mathrm{j} = \mathrm{M}+1, \ldots, \mathrm{D}$。如果我们消去$z_{nj}$式子中的$z_{ni}$和$b_i$，使用一般的$x_n$展开式，有：

$$x_n - \widetilde{x}_n = \sum_{k=M+1}^{D} \{(x_n - \overline{x})^T u_k\} u_k$$

从中我们看到，从x_n到$\widetilde{x}_n$的位移向量位于与主子空间垂直的空间中，因为它是u_i的线性组合，其中 $\mathrm{i} = \mathrm{M}+1, \ldots, \mathrm{D}$，如图 1 所示。这与预期相符，因为投影点$\widetilde{x}_n$一定位于主子空间内，但是我们可以在那个子空间内自由移动投影点，因此最小的误差由正交投影给出。

于是，我们得到了失真度量J的表达式，它是一个纯粹的关于u_i的函数，形式为：

$$J = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{j}=\mathrm{M}+1}^{\mathrm{D}} (x_n^T u_i - \overline{x}^T u_j)^2$$

也即：

$$J = \sum_{\mathrm{i}=\mathrm{M}+1}^{\mathrm{D}} u_1^T S u_1$$

剩下的任务是关于$\{u_i\}$对 J 进行最小化，这必须是具有限制条件的最小化，因为如果不这样，我们会得到$u_i = 0$ 这一没有意义的结果。限制来自于单位正交条件。并且：解可以表示为协方差矩阵的特征向量展开式。考虑一个形式化的解之前应考察一下这个结果。

考虑二维数据空间 $\mathrm{D} = 2$ 以及一维主子空间 $\mathrm{M} = 1$ 的情形。我们必须选择一个方向u_2来最小化$J = u_2^T S u_2$同时满足限制条件${u_2}^T u_2 = 1$。使用拉格朗日乘数λ_1来强制满足这个限制，我们考虑最小化：

$$\widetilde{J} \quad = u2T S u_2 + \lambda_2 (1 - {u_2}^T u_2)$$

令关于u_2 的导数等于零，我们有 $Su_2 = \lambda_2 u_2$，从而u_2是 S 的一个特征向量，且特征值为λ_2。因此任何特征向量都会定义失真度量的一个驻点。为了找到 J 在最小值点处的值，我们将u_2的解代回到失真度量中，得到$J = u_2$。于是，我们通过将u_2选择为对应于两个特征值中较小的那个特征值的特征向量，可以得到J的最小值。因此，我们应该将主子空间与具有较大的特征值的特征向量对齐。这个结果与我们的直觉相符，即为了最小化平均平方投影距离，我们应该将主成分子空间选为穿过数据点的均值并且与最大方差的方向对齐。对于特征值相等的情形，任何主方向的选择都会得到同样的J值。

对于任意的 D 和任意的 M $<$ D，最小化 J 的一般解都可以通过将$\{u_2\}$选择为协方差矩阵的特征向量的方式得到，即：

$$Su_2 = \lambda_2 u_2$$

其中 i $=$ $1, \ldots,$ D，并且与平常一样，特征向量$\{u_i\}$被选为单位正交的。失真度量的对应的值为：

$$J = \sum_{\mathrm{i=M+1}}^{\mathrm{D}} \lambda_i$$

这就是与主子空间正交的特征值的加和。于是，我们可以通过将这些特征向量选择成 D $-$ M 个最小的特征值对应的特征向量，来得到J的最小值，因此定义了主子空间的特征向量是对应于 M 个最大特征值的特征向量。

虽然我们已经考虑了 M $<$ D 的情形，但是 PCA 对于 M $=$ D 的情形仍然成立，这种情况下没有维度的降低，仅仅是将坐标轴旋转，与主成分对齐即可。

最后，值得注意的时，存在一个与此密切相关的线性维度降低的方法，被称为典型相关分析(canonical correlation analysis)， 或 者 CCA(Hotelling, 1936; Bach and Jordan, 2002)。PCA 操作的对象是一个随机变量，而CCA 考虑两个(或者更多)的变量，并且试图找到具有较高的交叉相关性的线性子空间对，从而在一个子空间中的每个分量都与另一个子空间的一个分量具有相关性。它的解可以表示为一般的特征向量问题。

4. 参考资料

[1]. Larry Wasserman. All of Statistcs , A Concies Course in Statistical Inference. *Springer*: 2002. pp 327-348.

[2]. Fotopoulos, Stergios B. All of Nonparametric Statistics. *Springer*: 2007. pp 100-103.

[3]. Jolliffe, I. Principal Component Analysis. 2nd Edition, *Springer*: 2002, New York.

[4]. HOTELLING, Harold, 1933. *Analysis of a Complex of Statistical Variablesinto Principal Components*.*Journal of Educational Psychology*,24(6 & 7),417–441 & 498–520.(一般认为是此文第一个发展了 PCA 方法)

[5]. DeGroot, Morris H, Schervish, Mark J. Probabilty and Statistics. 4th Edition, *Addison-Wesley*: 2002.