浅析主成分分析方法的数学原理

Principles of PCA Method: A Survey

# 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

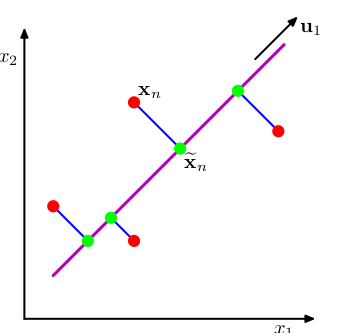


图1

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。

1. 形式1：最⼤⽅差

考虑⼀组观测数据集，其中n=1, . . . , N，因此是⼀个 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 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，考虑在⼀维空间上的投影。我们可以使⽤ D 维向量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 。我们只对的⽅向感兴趣，⽽对 u1 本⾝的⼤⼩不感兴趣)。这样，每个数据点被投影到⼀个标量上。其中是样本集合的均值。如下：

数据的协⽅差矩阵，为：

投影数据的方差，如下：

现在的工作是将上面的相对求得最大值。有什么限制？明显：要避免。我们通过归⼀化条件达到这一点。引⼊拉格朗⽇乘数，记作，然后对下式进⾏最⼤化：

求解关于的导数等于零的情况，可知驻点满⾜：

表明⼀定是的⼀个特征向量。如果我们左乘，我们看到⽅差为：

因此将设置为与具有最⼤的特征值的特征向量相等时，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 定义，对应于 M 个最⼤的特征值,...,。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，鉴于矩阵的完整的特征向量分解的代价为 ，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。

1. 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 ，其中 i= 1,...,D ，满⾜：

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:

其中，系数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定义的新坐标系，原始的 D 个分量被替换为⼀个等价的集合。与做内积，由单位正交性质有。不失⼀般性，有：

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：

其中依赖于特定的数据点，⽽是常数，对于所有数据点都相同。我们可以任意选择, 和，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：

⾸先考虑关于的最⼩化。消去，令它关于的导数为零，由单位正交的条件，有：

其中。类似地，令J关于的导数等于零，再次使⽤单位正交的关系，我们有：

其中。如果我们消去式子中的和，使⽤⼀般的展开式，有：

从中我们看到，从到的位移向量位于与主⼦空间垂直的空间中，因为它是的线性组合，其中，如图1所⽰。这与预期相符，因为投影点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的表达式，它是⼀个纯粹的关于的函数，形式为：

也即：

剩下的任务是关于对进⾏最⼩化，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。

考虑⼆维数据空间以及⼀维主⼦空间的情形。我们必须选择⼀个⽅向来最⼩化同时满⾜限制条件。使⽤拉格朗⽇乘数来强制满⾜这个限制，我们考虑最⼩化：

令关于的导数等于零，我们有 ，从⽽是的⼀个特征向量，且特征值为。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，我们将的解代回到失真度量中，得到。于是，我们通过将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值。

对于任意的和任意的，最⼩化的⼀般解都可以通过将选择为协⽅差矩阵的特征向量的⽅式得到，即：

其中，并且与平常⼀样，特征向量被选为单位正交的。失真度量的对应的值为：

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个最⼩的特征值对应的特征向量，来得到的最⼩值，因此定义了主⼦空间的特征向量是对应于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。

1. 参考资料
2. Larry Wasserman. All of Statistcs , A Concies Course in Statistical Inference. *Springer*: 2002. pp 327-348.
3. Fotopoulos, Stergios B. All of Nonparametric Statistics. *Springer*: 2007. pp 100-103.
4. Jolliffe, I. Principal Component Analysis. 2nd Edition, *Springer*: 2002, New York.
5. 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)
6. DeGroot, Morris H, Schervish, Mark J. Probabilty and Statistics. 4th Edition, *Addison-Wesley*: 2002.