PCA数学原理解析

数据仓库与数据挖掘

What is PCA?

- PCA（Principal Component Analysis）是一种常用的数据分析方法。
- PCA通过线性变换将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维。

数据的向量表示

- 一般情况下，在数据挖掘和机器学习中，数据被表示为向量。
  - 例如，某个淘宝店2012年全年的流量及交易情况可以看成一组记录的集合，其中每一天的数据是一条记录，列向量格式如下：
  - (浏览量, 访客数, 下单数, 成交数, 成交金额) ^{T}
  - **(500,240,25,13,2312.15)^{T}**

降维问题：目的和原则

- 机器学习算法的复杂度通常与数据的维数有着密切关系，甚至与维数呈指数级关联。
- 机器学习在实际中处理成千上万甚至几十万维的情况也并不罕见，在这种情况下，机器学习的资源消耗是不可接受的，因此我们必须对数据进行降维。
- 降维当然意味着信息的丢失，不过鉴于实际数据本身常常存在的相关性，我们可以想办法在降维的同时将信息的损失尽量降低。

降维与相关性举例（1）

- 某学籍数据有两列M和F，其中，
  - M列的取值是：学生为男性取值1，为女性取值0；
  - F列的取值是：学生为女性取值1，男性取值0。
- 此时，如果我们统计全部学籍数据，会发现对于任何一条记录来说，当M为1时F必定为0，反之当M为0时F必定为1。
- 在这种情况下，我们将M或F去掉实际上没有任何信息的损失，因为只要保留一列就可以完全还原另一列。

降维与相关性举例（2）

- 例如上面淘宝店铺的数据，从经验可知：
  - “浏览量”和“访客数”往往具有较强的相关关系，
  - “下单数”和“成交数”也具有较强的相关关系。
  - 这里我们非正式的使用“相关关系”这个词，可以直观理解为“当某一天这个店铺的浏览量较高（或较低）时，我们应该很大程度上认为这天的访客数也较高（或较低）”。
- 表明：如果删除浏览量或访客数其中一个指标，数据集并不会丢失太多信息。因此，将相关的指标删除一个，以降低机器学习算法的复杂度。

降维方法：从直观描述到数学支撑

- 上面给出的是降维的朴素思想描述，可以有助于直观理解降维的动机和可行性，但并不具有操作指导意义。
  - 例如，到底删除哪一列损失的信息才最小？亦或根本不是单纯删除几列，而是通过某些变换将原始数据变为更少的列但又使得丢失的信息最小？到底如何度量丢失信息的多少？如何根据原始数据决定具体的降维操作步骤？
- 要回答上面的问题，就要对降维问题进行数学化和形式化的讨论。PCA是一种具有严格数学基础并且已被广泛采用的降维方法。

降维的数学支撑

（1）向量的表示及基变换

向量的表示及基变换：内积和投影

- 两个维数相同的向量的内积被定义为：

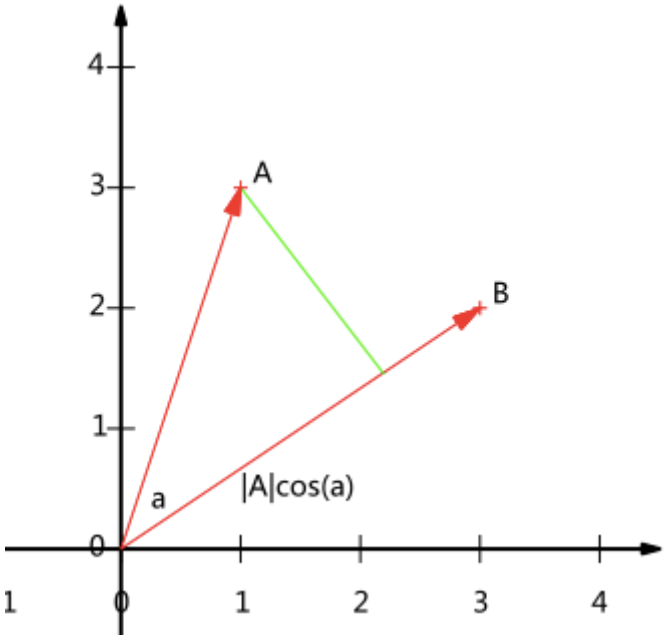
$$(a_1, a_2, \cdots, a_n)^{\mathsf{T}} \cdot (b_1, b_2, \cdots, b_n)^{\mathsf{T}} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

- 内积运算将两个向量映射为一个实数。

- 设A和B是两个n维向量，n维向量可以等价表示为n维空间中的一条从原点发射的有向线段：

$$A = (x_1, y_1)$$

$$B = (x_2, y_2)$$

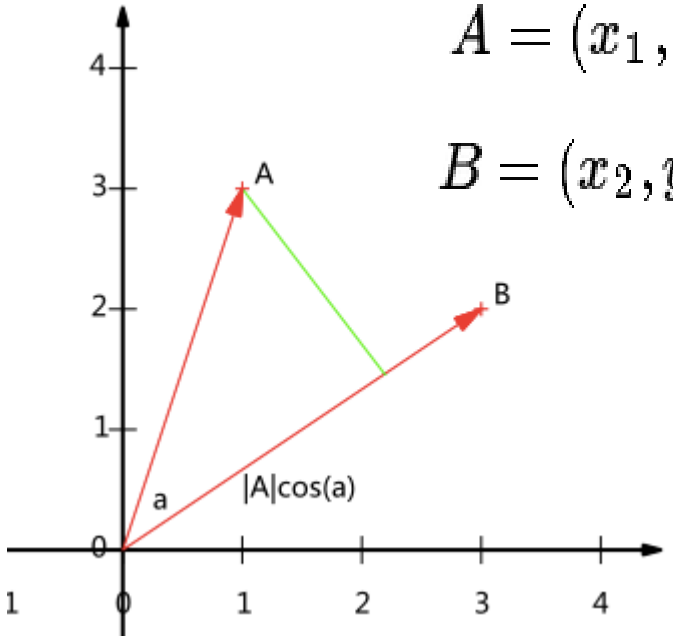


从A点向B所在直线引一条垂线。

垂线与B的交点叫做A在B上的投影，

设A与B的夹角是a，

则投影的矢量长度为$|A|cos(a)$



其中$|A|=\sqrt{x_1^2+y_1^2}$ 是向量A的模，也就是A线段的标量长度。

将内积表示为另一种我们熟悉的形式：$A\cdot B=|A||B|cos(a)$

A与B的内积等于A到B的投影长度乘以B的模。

设B的模为1，即让 $|B|=1$，那么就变成了：

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

设向量B的模为1，则A与B的内积值等于A向B所在直线投影的矢量长度！

基的数学含义

- 一个二维向量可以对应二维笛卡尔直角坐标系中从原点出发的一个有向线段。
- 向量(x,y)实际上表示线性组合：

$$x(1,0)^{\mathsf{T}} + y(0,1)^{\mathsf{T}}$$

  - (1,0)和(0,1)叫做二维空间中的一组基。

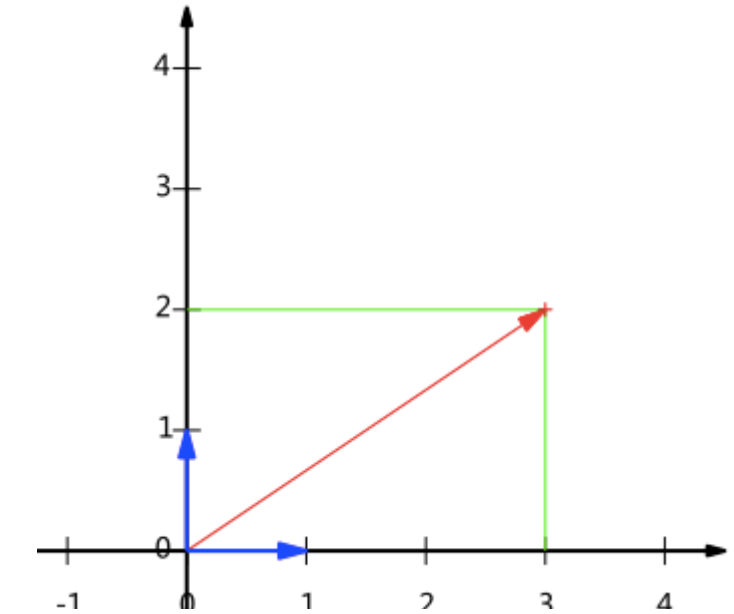
- **要准确描述向量，首先要确定一组基，然后给出在基所在的各个直线上的投影值。**

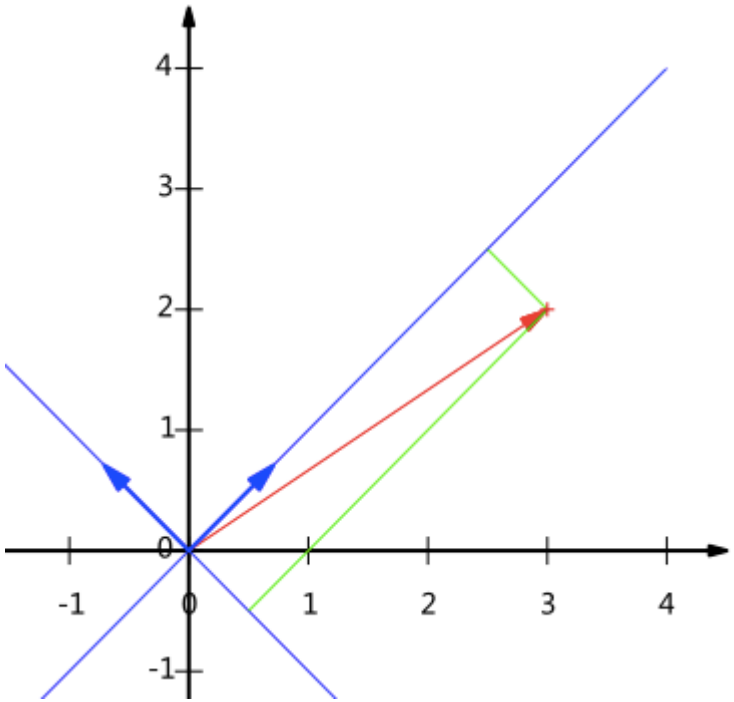
- 例如，(1,1)和(-1,1)也可以成为一组基，标准化后变成

$$(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}) \text{ 和} (-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$$

- (3,2)在新基上的坐标，

$$(\frac{5}{\sqrt{2}},-\frac{1}{\sqrt{2}})$$





基变换的矩阵表示

- （3，2）的基变换：

$(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$ 和$(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$

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

- 例如(1,1)，(2,2)，(3,3)的基变换：

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

- **一般的，如果有M个N维向量，想将其变换为由R个N维向量表示的新空间中，那么首先将R个基按行组成矩阵A，然后将向量按列组成矩阵B，那么两矩阵的乘积AB就是变换结果，其中AB的第m列为A中第m列变换后的结果。**

数学表示为：

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix}\begin{pmatrix} a_1 & a_2 & \cdots & a_M \end{pmatrix}=\begin{pmatrix} p_1a_1 & p_1a_2 & \cdots & p_1a_M \\ p_2a_1 & p_2a_2 & \cdots & p_2a_M \\ \vdots & \vdots & \ddots & \vdots \\ p_Ra_1 & p_Ra_2 & \cdots & p_Ra_M \end{pmatrix}$$

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。

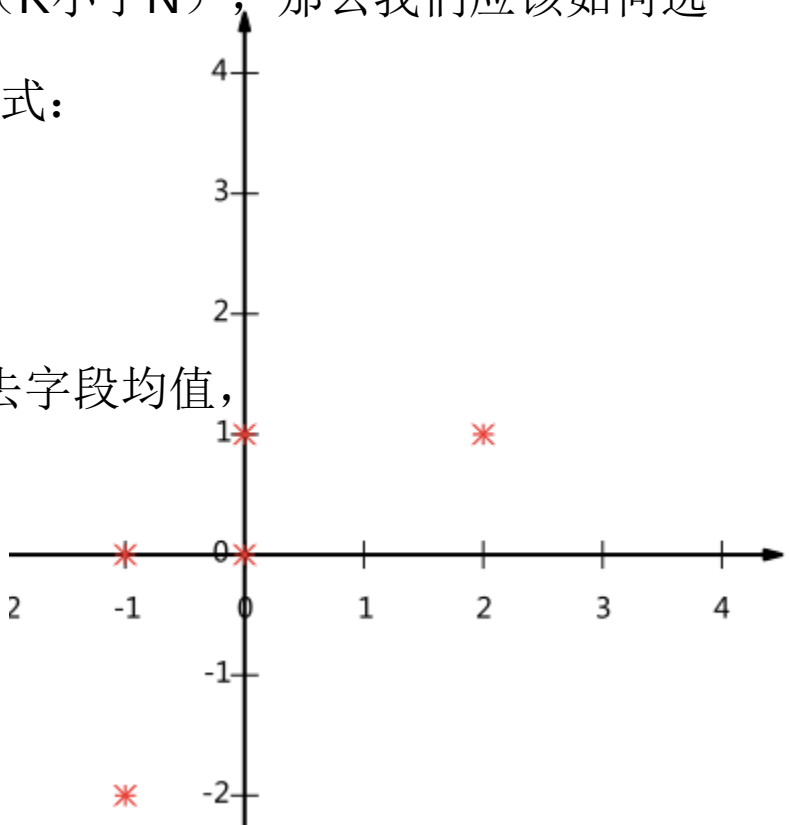
协方差矩阵及优化目标

- 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？
- 假设数据由五条记录组成，将它们表示成矩阵形式：

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$

其中每一列为一条数据记录，而一行为一个字段。
为了后续处理方便，首先将每个字段内所有值都减去字段均值，其结果是将每个字段都变为均值为0

$$\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$



- 如果我们必须使用一维来表示这些数据，
- 又希望尽量保留原始的信息，你要如何选择？

- 通过基变换的讨论我们知道，这个问题实际上是要在二维平面中选择一个方向，将所有数据都投影到这个方向所在直线上，用投影值表示原始记录。这是一个实际的二维降到一维的问题。

- 如果向x轴投影，则最左边的两个点会重叠在一起，中间的两个点也会重叠在一起，于是本身四个各不相同的二维点投影后只剩下两个不同的值了，这是一种严重的信息丢失，同理，如果向y轴投影最上面的两个点和分布在x轴上的两个点也会重叠。单独选x和y轴都不是最好的投影选择。

方差

- 目标：投影后投影值尽可能分散，而这种分散程度，可以用数学上的方差来表述。
- 此处，一个字段的方差可以看做是每个元素与字段均值的差的平方和的均值，即：

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

- 由于上面已将每个字段的均值都化为0了，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。

协方差

- 目的：从直观上说，让两个字段尽可能表示更多的原始信息，不希望它们之间存在（线性）相关性的，因为相关性意味着两个字段不是完全独立，必然存在重复表示的信息。
- 数学上可以用两个字段的协方差表示其相关性，由于已经让每个字段均值为0，则：

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

  - 可以看到，在字段均值为0的情况下，两个字段的协方差简洁的表示为其内积除以元素数m。

- 当协方差为0时，表示两个字段完全独立。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选择。因此最终选择的两个方向一定是正交的。

- 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（K大于0，小于N），其目标是选择K个单位（模为1）正交基，使得原始数据变换到这组基上后，各字段两两间协方差为0，而字段的方差则尽可能大（在正交的约束下，取最大的K个方差）。**

协方差矩阵

- 目的：降维的指标与字段内方差及字段间协方差有密切关系。因此，可将两者统一表示。

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

- 用X乘以X的转置，并乘上系数1/m：

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

  - 矩阵对角线上的两个元素分别是两个字段的方差，而其它元素是a和b的协方差。
  - 两者被统一到了一个矩阵的。

协方差矩阵推广

设我们有m个n维数据记录，将其按列排成n乘m的矩阵X，设$C = \frac{1}{m} X X^{\mathsf{T}}$，

则C是一个对称矩阵，其对角线分别 各个字段的方差，

而第i行j列和j行i列元素相同，表示i和j两个字段的协方差。

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

协方差矩阵对角化

- 目标：将协方差矩阵对角化，即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
- 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：

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

- P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**。

由上文知道，协方差矩阵C是一个是对称矩阵，在线性代数上，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$e_1, e_2, \cdots, e_n$，将其按列组成矩阵：

$$E = \begin{pmatrix} e_1 & e_2 & \cdots & e_n \end{pmatrix}$$

则对协方差矩阵C有如下结论：

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

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：

$$P = E^{\mathsf{T}}$$

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Λ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。

PCA算法

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（代表一个属性字段）进行零均值化，即减去这一行的均值

3）求出协方差矩阵$C = \frac{1}{m} X X^{\mathrm{T}}$

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

这里以上文提到的

$$\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

为例，我们用PCA方法将这组二维数据其降到一维。

因为这个矩阵的每行已经是零均值，这里我们直接求协方差矩阵：

$$C = \frac{1}{5}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$$

然后求其特征值和特征向量，具体求解方法不再详述，可以参考相关资料。求解后特征值为：

$$\lambda_1 = 2, \lambda_2 = 2/5$$

其对应的特征向量分别是：

$$c_1\begin{pmatrix} 1 \\ 1 \end{pmatrix}, c_2\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

其中对应的特征向量分别是一个通解，c_1 和c_2 可取任意实数。那么标准化后的特征向量为：$\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

矩阵P是：

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

矩阵P是：

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

可以验证协方差矩阵C的对角化：

$$PCP^{\mathsf{T}} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2/5 \end{pmatrix}$$

最后我们用P的第一行乘以数据矩阵，就得到了降维后的表示：

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