机器学习算法系列（22）：主成分分析

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。

一、数据的向量表示及降维问题

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，格式如下：

> (日期, 浏览量, 访客数, 下单数, 成交数, 成交金额)

其中“日期”是一个记录标志而非度量值，而数据挖掘关心的大多是度量值，因此如果我们忽略日期这个字段后，我们得到一组记录，每条记录可以被表示为一个五维向量，其中一条看起来大约是这个样子：

$$(500, 240, 25, 13, 2312.15)^T$$

注意这里用了转置，因为习惯上使用列向量表示一条记录（后面会看到原因），本文后面也会遵循这个准则。不过为了方便有时会省略转置符号，但我们说到向量默认都是指列向量。

我们当然可以对这一组五维向量进行分析和挖掘，不过我们知道，很多机器学习算法的复杂度和数据的维数有着密切关系，甚至与维数呈指数级关联。当然，这里区区五维的数据，也许还无所谓，但是实际机器学习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。

举个例子，假如某学籍数据有两列M和F，其中M列的取值是如何此学生为男性取值1，为女性取值0；而F列是学生为女性取值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。

当然上面是一个极端的情况，在现实中也许不会出现，不过类似的情况还是很常见的。例如上面淘宝店铺的数据，从经验我们可以知道，“浏览量”和“访客数”往往具有较强的相关关系，而“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）”。后面的章节中我们会给出相关性的严格数学定义。

这种情况表明，如果我们删除浏览量或访客数其中一个指标，我们应该期待并不会丢失太多信息。因此我们可以删除一个，以降低机器学习算法的复杂度。

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，我们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。而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。下面我不会直接描述PCA，而是通过逐步分析问题，让我们一起重新“发明”一遍PCA。

二、向量的表示及基变换

既然我们面对的数据被抽象为一组向量，那么下面有必要研究一些向量的数学性质。而这些数学性质将成为后续导出PCA的理论基础。

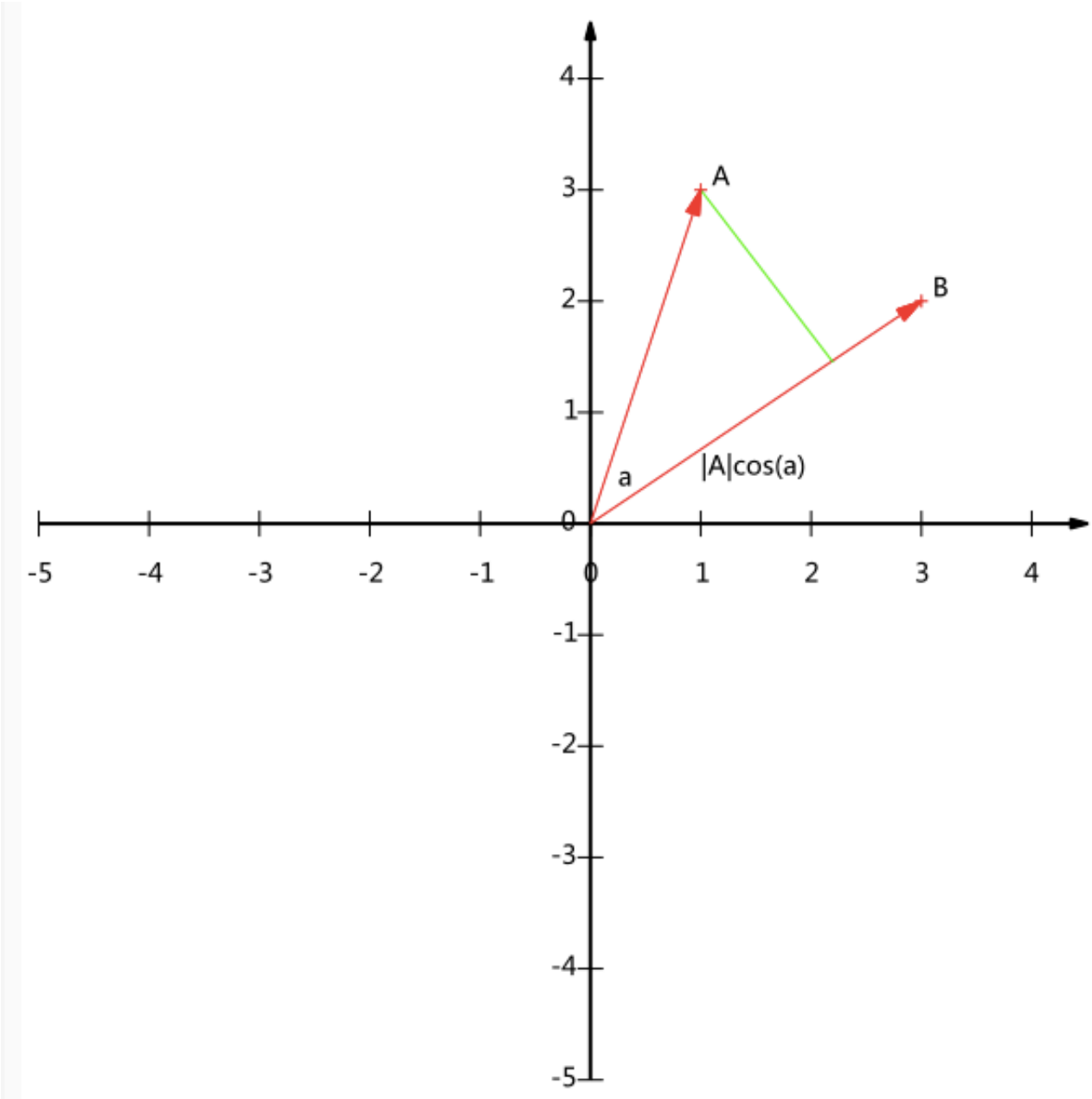
2.1 内积与投影

下面先来看一个高中就学过的向量运算：内积。两个维数相同的向量的内积被定义为：

$$(a_1, a_2, \cdots, a_n)^{\mathsf{T}} \cdot (b_1, b_2, \cdots, b_n)\mathsf{T} = a_1b_1 + a_2b_2 + \cdots + a_nb_n$$

内积运算将两个向量映射为一个实数。其计算方式非常容易理解，但是其意义并不明显。下面我们分析内积的几何意义。

假设A和B是两个n维向量，我们知道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，为了简单起见我们假设A和B均为二维向量，则$A = (x_1, y_1)$，$B = (x_2, y_2)$。则在二维平面上A和B可以用两条发自原点的有向线段表示，见下图：



现在我们从A点向B所在直线引一条垂线。我们知道垂线与B的交点叫做A在B上的投影，再设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。

注意这里我们专门区分了矢量长度和标量长度，标量长度总是大于等于0，值就是线段的长度；而矢量长度可能为负，其绝对值是线段长度，而符号取决于其方向与标准方向相同或相反。

到这里还是看不出内积和这东西有什么关系，不过如果我们将内积表示为另一种我们熟悉的形式：

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

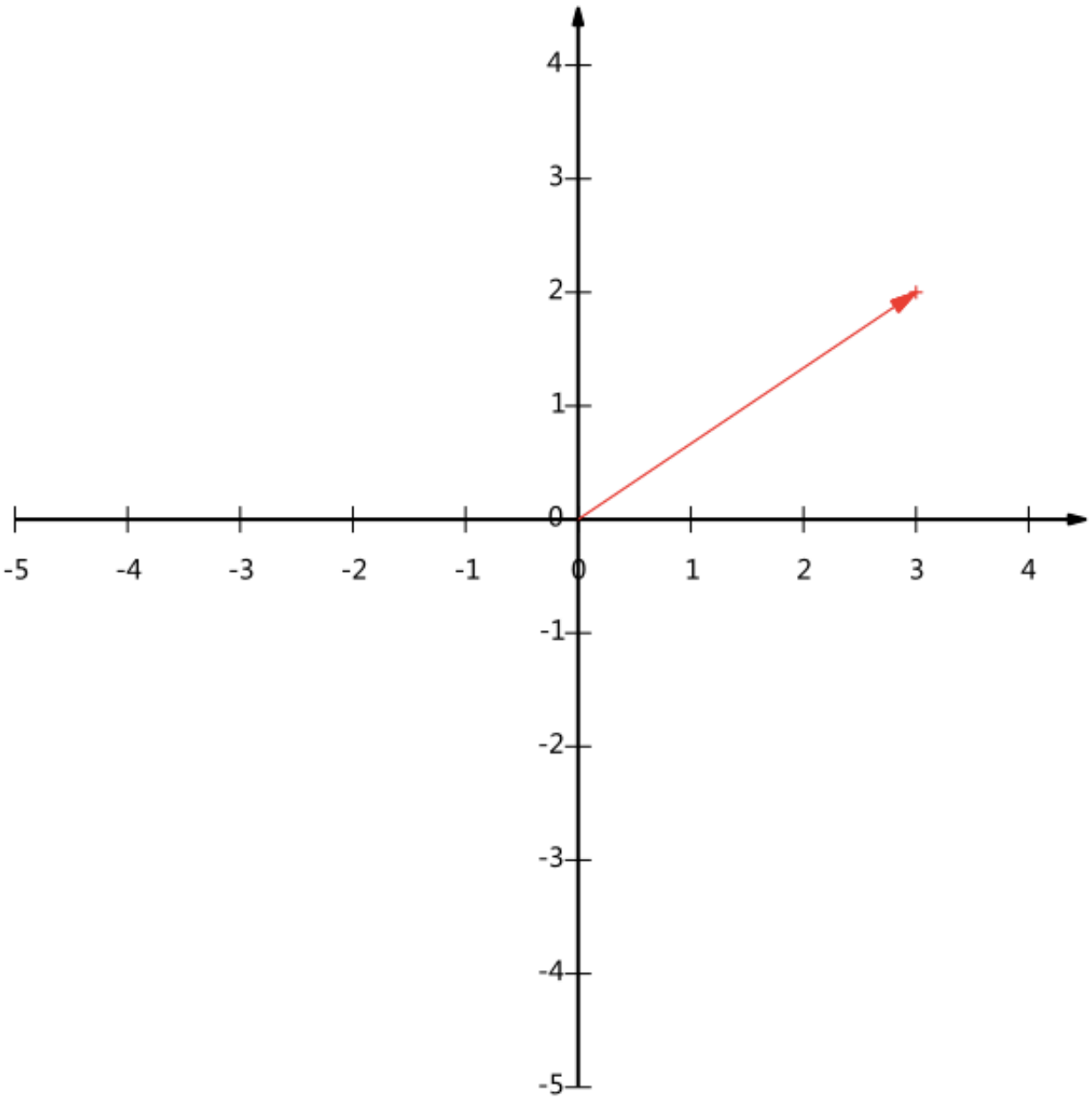
A与B的内积等于A到B的投影长度乘以B的模。再进一步，如果我们假设B的模为1，即让$|B| = 1$，那么就变成了：

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

也就是说，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！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。

2.2 基

下面我们继续在二维空间内讨论向量。上文说过，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。例如下面这个向量：



在代数表示方面，我们经常用线段终点的点坐标表示向量，例如上面的向量可以表示为(3,2)，这是我们再熟悉不过的向量表示。

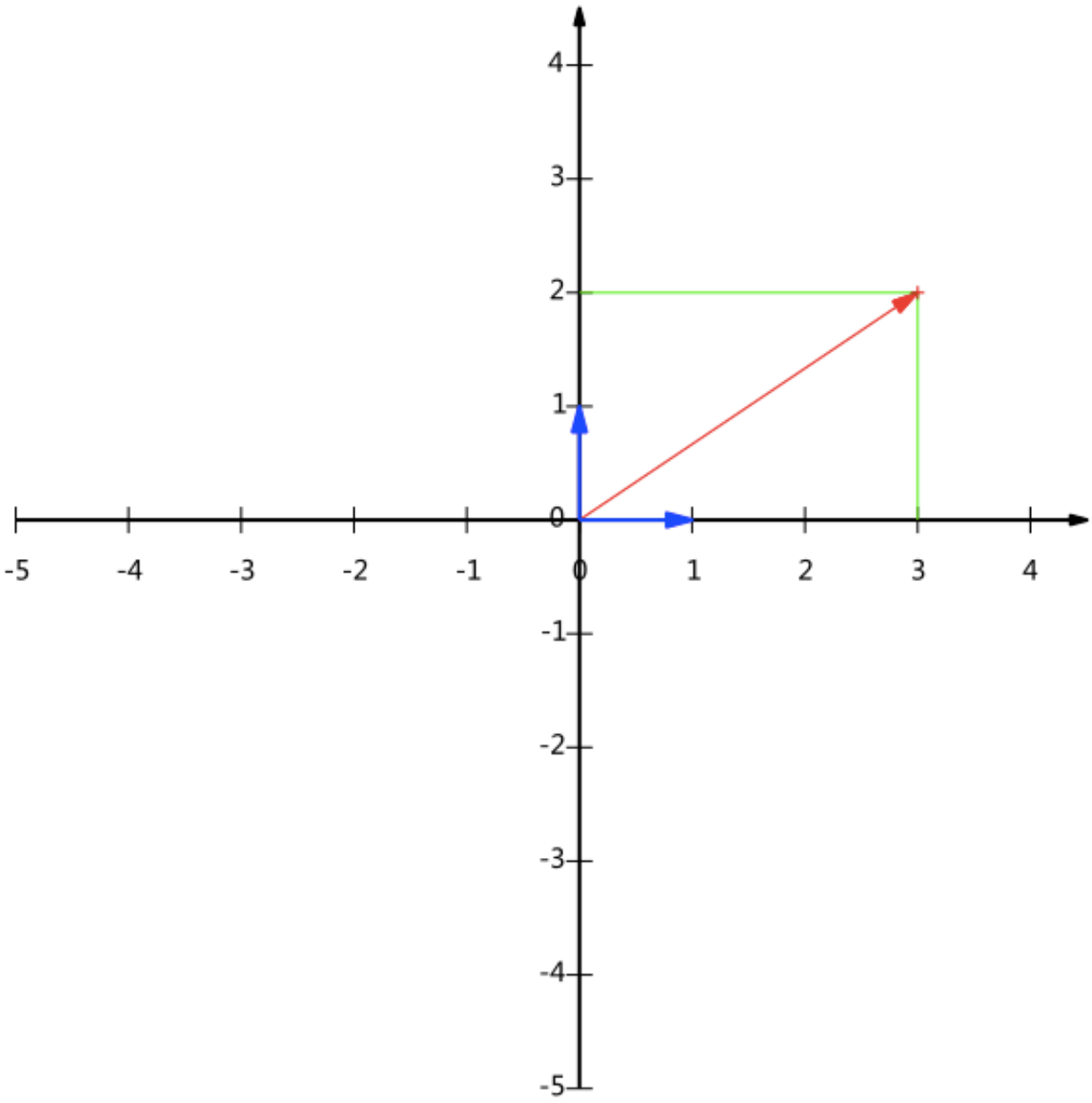
不过我们常常忽略，只有一个(3,2)本身是不能够精确表示一个向量的。我们仔细看一下，这里的3实际表示的是向量在x轴上的投影值是3，在y轴上的投影值是2。也就是说我们其实隐式引入了

一个定义：以x轴和y轴上正方向长度为1的向量为标准。那么一个向量(3,2)实际是说在x轴投影为3而y轴的投影为2。注意投影是一个矢量，所以可以为负。

更正式的说，向量(x,y)实际上表示线性组合：

$$x(1,0)^T + y(0,1)^T$$

不难证明所有二维向量都可以表示为这样的线性组合。此处（1，0）和（0，1）叫做二维空间中的一组基。



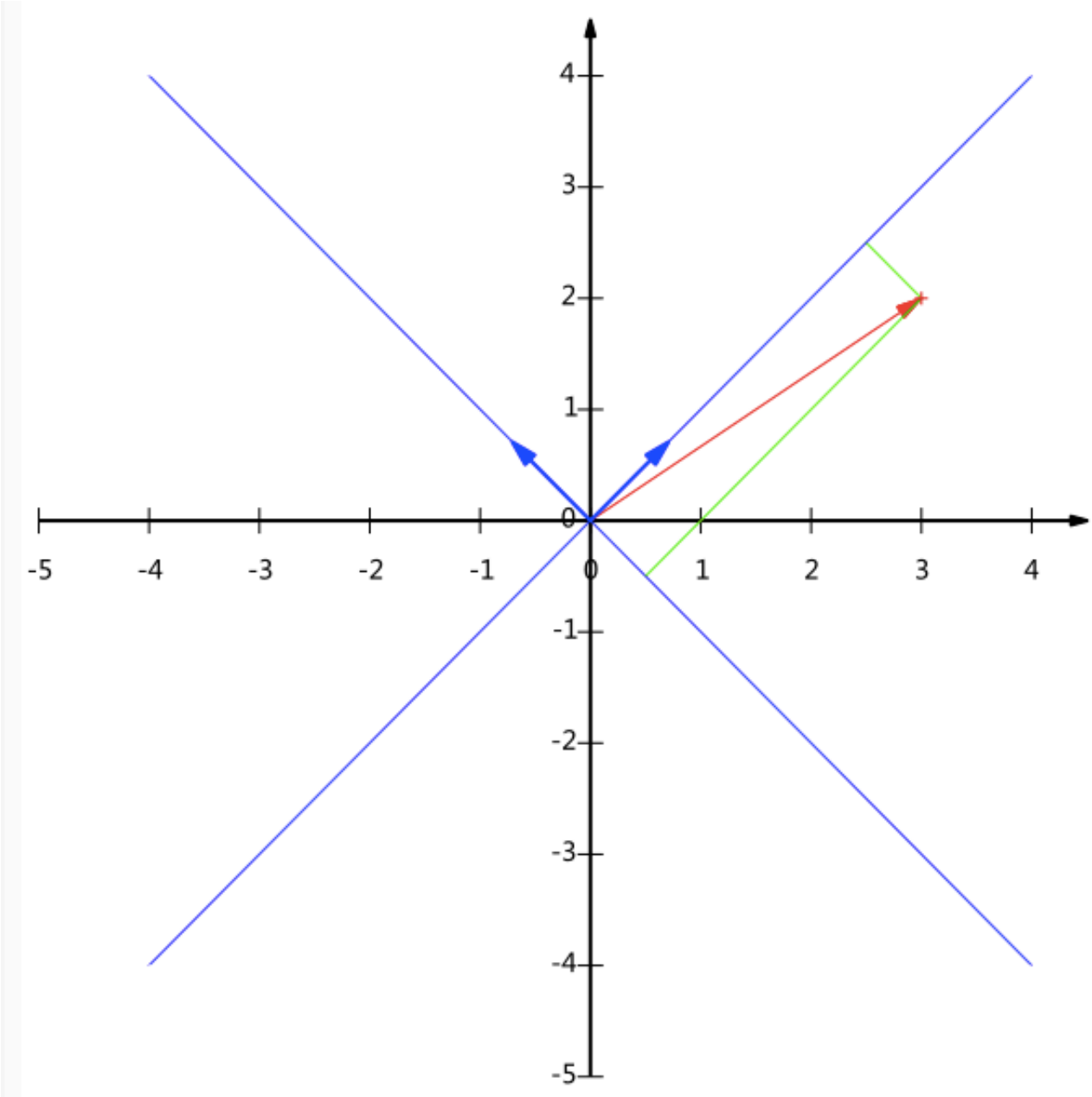
所以，要准确描述向量，首先要确定一组基，然后给出在基所在的各个直线上的投影值，就可以了。只不过我们经常省略第一步，而默认以(1,0)和(0,1)为基。

我们之所以默认选择$(1,0)$和$(0,1)$为基，当然是比较方便，因为它们分别是x和y轴正方向上的单位向量，因此就使得二维平面上点坐标和向量一一对应，非常方便。但实际上任何两个线性无关的二维向量都可以成为一组基，所谓线性无关在二维平面内可以直观认为是两个不在一条直线上

的向量。

例如，$(1, 1)$和$(-1, 1)$也可以成为一组基。一般来说，我们希望基的模是1，因为从内积的意义可以看到，如果基的模是1，那么就可以方便的用向量点乘基而直接获得其在新基上的坐标了！实际上，对应任何一个向量我们总可以找到其同方向上模为1的向量，只要让两个分量分别除以模就好了。例如，上面的基可以变为$(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$和$-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}$。

现在，我们想获得$(3, 2)$在新基上的坐标，即在两个方向上的投影矢量值，那么根据内积的几何意义，我们只要分别计算(3,2)和两个基的内积，不难得到新的坐标为$(\frac{5}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$。下图给出了新的基以及(3,2)在新基上坐标值的示意图：



另外这里要注意的是，我们列举的例子中基是正交的（即内积为0，或直观说相互垂直），但可以成为一组基的唯一要求就是线性无关，非正交的基也是可以的。不过因为正交基有较好的性质，所以一般使用的基都是正交的。

2.3 基变换的矩阵表示

下面我们找一种简便的方式来表示基变换。还是拿上面的例子，想一下，将$(3, 2)$变换为新基上的坐标，就是用$(3, 2)$与第一个基做内积运算，作为第一个新的坐标分量，然后用(3,2)与第二个基做内积运算，作为第二个新坐标的分量。实际上，我们可以用矩阵相乘的形式简洁的表示这个变换：

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{5}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

其中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。可以稍微推广一下，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，就得到了所有这些向量在新基下的值。例如(1,1)，(2,2)，(3,3)，想变换到刚才那组基上，则可以这样表示：

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

于是一组向量的基变换被干净的表示为矩阵的相乘。

一般的，如果我们有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：

$$\begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ p_r \end{bmatrix} \begin{bmatrix} a_1 & a_2 & \cdots & a_M \end{bmatrix} = \begin{bmatrix} p_1a_1 & p_1a_2 & \cdots & p_1a_M \\ p_2a_1 & p_2a_2 & \cdots & p_2a_M \\ \cdots & \cdots & \cdots & \cdots \\ p_ra_1 & p_ra_2 & \cdots & p_ra_M \end{bmatrix}$$

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。

特别要注意的是，这里R可以小于N，而R决定了变换后数据的维数。也就是说，我们可以将一N维数据变换到更低维度的空间中去，变换后的维度取决于基的数量。因此这种矩阵相乘的表示也可以表示降维变换。

最后，上述分析同时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列向量变换到左边矩阵中每一行行向量为基所表示的空间中去。更抽象的说，一个矩阵可以表示一种线性变换。很多同学在学线性代数时对矩阵相乘的方法感到奇怪，但是如果明白了矩阵

相乘的物理意义，其合理性就一目了然了。

三、协方差矩阵及优化目标

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，而且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。但是我们还没有回答一个最最关键的问题：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？

要完全数学化这个问题非常繁杂，这里我们用一种非形式化的直观方法来看这个问题。

为了避免过于抽象的讨论，我们仍以一个具体的例子展开。假设我们的数据由五条记录组成，将它们表示成矩阵形式：

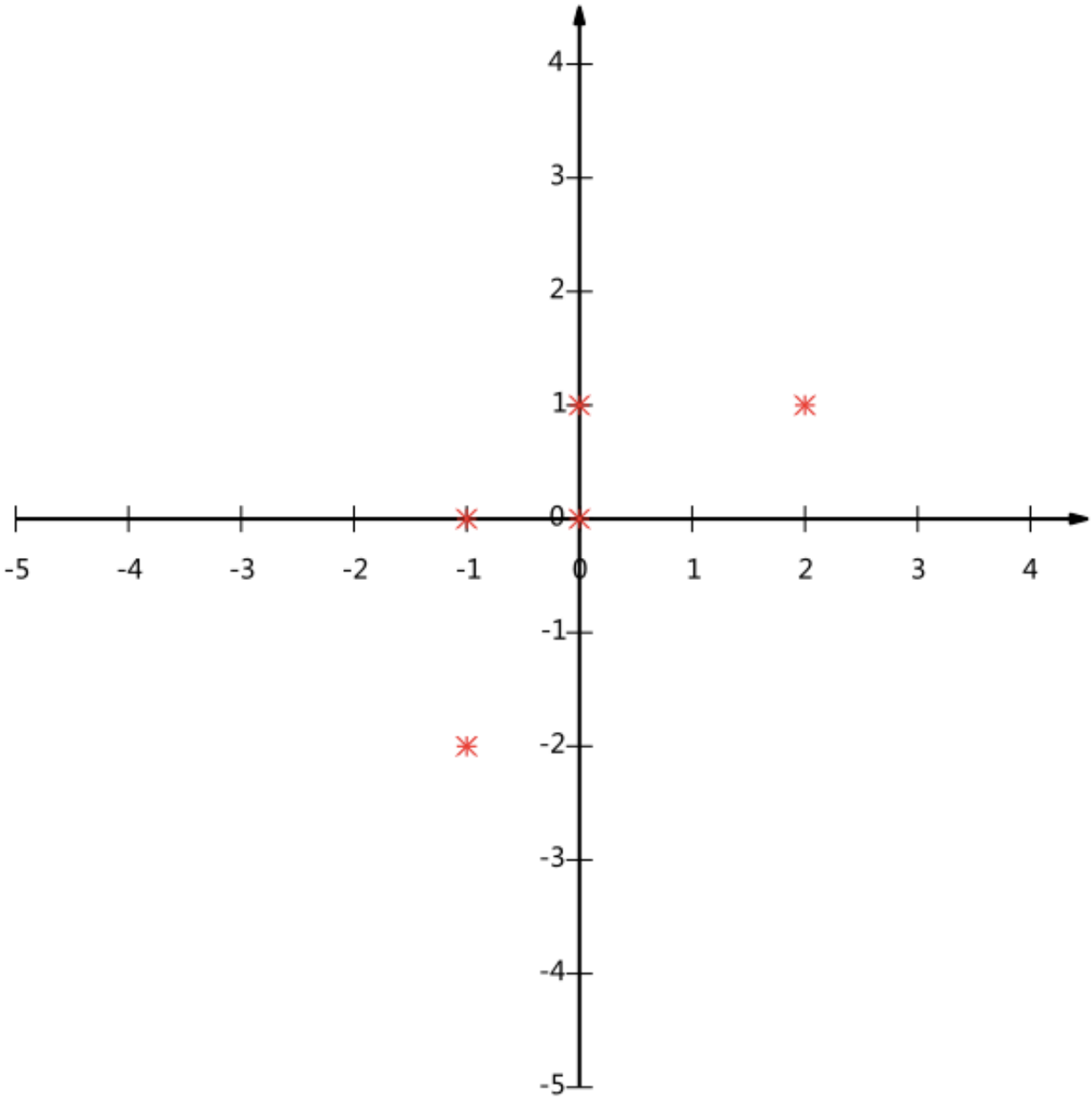
$$\begin{bmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{bmatrix}$$

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，我们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（这样做的道理和好处后面会看到）。

我们看上面的数据，第一个字段均值为2，第二个字段均值为3，所以变换后：

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

我们可以看下五条数据在平面直角坐标系内的样子：



现在问题来了：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?

通过上一节对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。

那么如何选择这个方向（或者说基）才能尽量保留最多的原始信息呢？一种直观的看法是：希望投影后的投影值尽可能分散。

以上图为例，可以看出如果向x轴投影，那么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。所以看来x和y轴都不是最好的投影选择。我们直观目测，如果向通过第一象限和第三象限的斜线投影，则五个点在投影后还是可以区分的。

下面，我们用数学方法表述这个问题。

3.1 方差

上文说到，我们希望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：

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

由于上面我们已经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。

3.2 协方差

对于上面二维降成一维的问题来说，找到那个使得方差最大的方向就可以了。不过对于更高维，还有一个问题需要解决。考虑三维降到二维问题。与之前相同，首先我们希望找到一个方向使得投影后方差最大，这样就完成了第一个方向的选择，继而我们选择第二个投影方向。

如果我们还是单纯只选择方差最大的方向，很明显，这个方向与第一个方向应该是“几乎重合在一起”，显然这样的维度是没有用的，因此，应该有其他约束条件。从直观上说，让两个字段尽可能表示更多的原始信息，我们是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）。

3.3 协方差矩阵

上面我们导出了优化目标，但是这个目标似乎不能直接作为操作指南（或者说算法），因为它只说要什么，但根本没有说怎么做。所以我们要继续在数学上研究计算方案。

我们看到，最终要达到的目的与字段内方差及字段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。于是我们来了灵感：

假设我们只有a和b两个字段，那么我们将它们按行组成矩阵X：

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

然后我们用X乘以X的转置，并乘上系数$1/m$：

$$\frac{1}{m}XX^T = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m} a_i^2 & \frac{1}{m}\sum_{i=1}^{m} a_i b_i \\ \frac{1}{m}\sum_{i=1}^{m} a_i b_i & \frac{1}{m}\sum_{i=1}^{m} b_i^2 \end{bmatrix}$$

奇迹出现了！这个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。

根据矩阵相乘的运算法则，这个结论很容易被推广到一般情况：设我们有m个n维数据记录，将其按列排成n乘m的矩阵X，设￥$C = \frac{1}{m}XX^{\mathrm{T}}$，则C是一个对称矩阵，其对角线分别个各个字段的方差，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。

3.4 协方差矩阵对角化

根据上述推导，我们发现要达到优化条件，等价于将协方差对角化：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，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：

$$D = \frac{1}{m} YY^T$$
$$= \frac{1}{m}(PX)(PX)^T$$
$$= \frac{1}{m} PXX^T P$$
$$= PCP^T$$

现在事情很明白了，我们要找的P不是别的，而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，并且对角元素按从小到大依次排列，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“发明”PCA还有仅一步之遥！

现在所有焦点都聚焦在了协方差矩阵对角化问题上，有时，我们真应该感谢数学家的先行，因为矩阵对角化在线性代数领域已经属于被玩烂了的东西，所以这在数学上根本不是问题。

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λλ重数为r，则必然存在r个线性无关的特征向量对应于λλ，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：

$$E^T C E = \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_n \end{bmatrix}$$

其中ΛΛ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）。以上结论不再给出严格的数学证明，对证明感兴趣的朋友可以参考线性代数书籍关于“实对称矩阵对角化”的内容。

到这里，我们发现我们已经找到了需要的矩阵P：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ΛΛ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。

至此我们完成了整个PCA的数学原理讨论。在下面的一节，我们将给出PCA的一个实例。

四、算法及实例

为了巩固上面的理论，我们在这一节给出一个具体的PCA实例。

4.1 PCA算法

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（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$C = \frac{1}{m}XX^T$
- 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

4.2 实例

这里以上文提到的

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

为例，我们用PCA方法将这组二维数据降到一维。

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：

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

然后求其特征值和特征向量，具体求解方法不再详述。求解后特征值为：

$$\lambda_1 = 2, \lambda_2 = \frac{2}{5}$$

其对应的特征向量分别是：

$$c_1 \begin{bmatrix} -2 \\ 0 \end{bmatrix}, c_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

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：

$$\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

因此我们的矩阵P是：

$$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

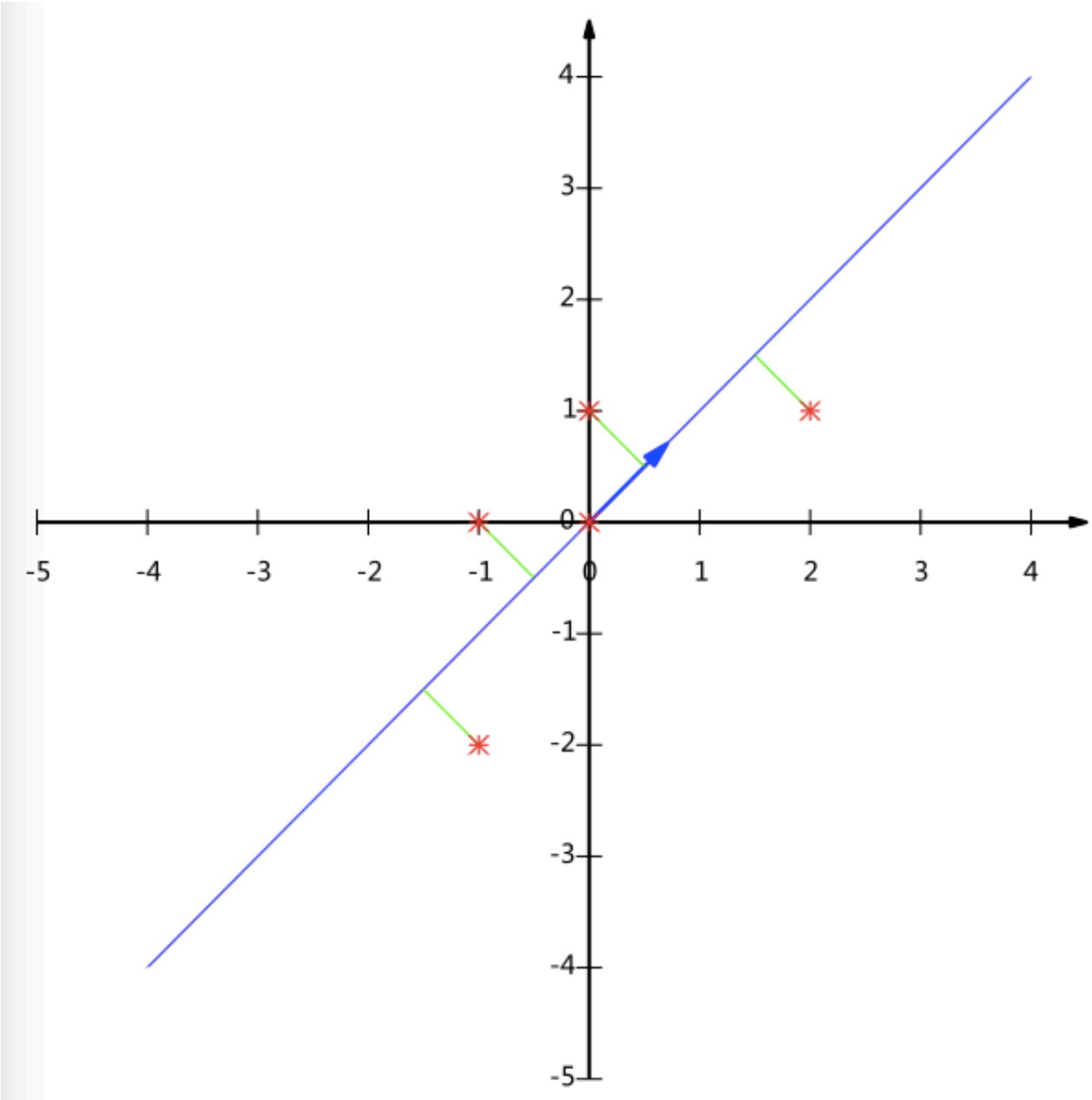
可以验证协方差矩阵C的对角化：

$$PCP^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & \frac{2}{5} \end{bmatrix}$$

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：

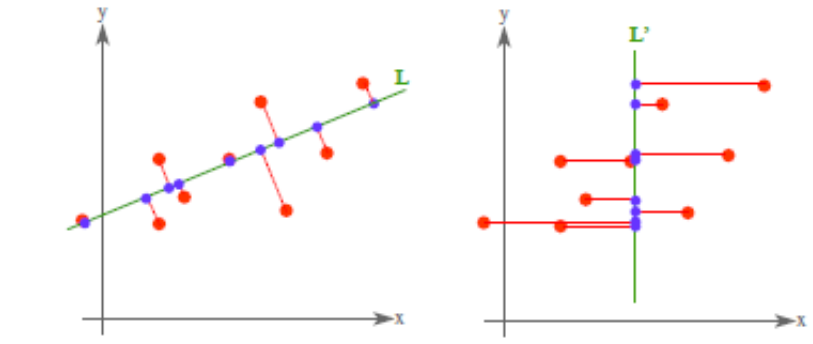
$$Y = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{3}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \frac{3}{\sqrt{2}} & -\frac{1}{2} \end{bmatrix}$$

降维投影结果如下图：



五、理论意义

PCA将n个特征降维到k个，可以用来进行数据压缩，如果100维的向量最后可以用10维来表示，那么压缩率为90%。同样图像处理领域的KL变换使用PCA做图像压缩。但PCA要保证降维后，还要保证数据的特性损失最小。再看回顾一下PCA的效果。经过PCA处理后，二维数据投影到一维上可以有以下几种情况：



我们认为左图好，一方面是投影后方差最大，一方面是点到直线的距离平方和最小，而且直线过样本点的中心点。为什么右边的投影效果比较差？直觉是因为坐标轴之间相关，以至于去掉一个坐标轴，就会使得坐标点无法被单独一个坐标轴确定。

PCA得到的k个坐标轴实际上是k个特征向量，由于协方差矩阵对称，因此k个特征向量正交。

得到的新的样例矩阵$Y = PX$就是m个样例到k个特征向量的投影，也是这k个特征向量的线性组合。P中e之间是正交的。从矩阵乘法中可以看出，PCA所做的变换是将原始样本点（n维），投影到k个正交的坐标系中去，丢弃其他维度的信息。举个例子，假设宇宙是n维的（霍金说是11维的），我们得到银河系中每个星星的坐标（相对于银河系中心的n维向量），然而我们想用二维坐标去逼近这些样本点，假设算出来的协方差矩阵的特征向量分别是图中的水平和竖直方向，那么我们建议以银河系中心为原点的x和y坐标轴，所有的星星都投影到x和y上，得到下面的图片。然而我们丢弃了每个星星离我们的远近距离等信息。

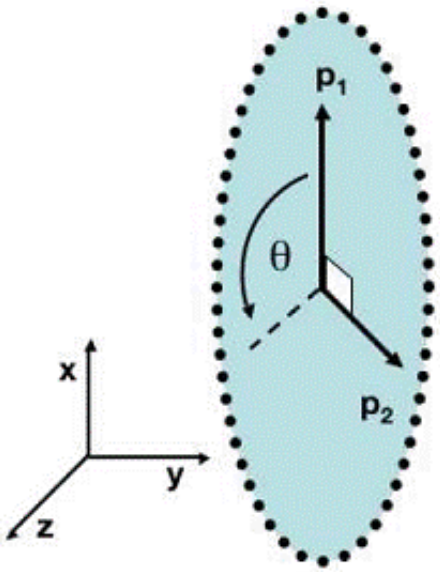
六、进一步讨论

根据上面对PCA的数学原理的解释，我们可以了解到一些PCA的能力和限制。PCA本质上是将方差最大的方向作为主要特征，并且在各个正交方向上将数据“离相关”，也就是让它们在不同正交方向上没有相关性。

因此，PCA也存在一些限制，例如它可以很好的解除线性相关，但是对于高阶相关性就没有办法了，对于存在高阶相关性的数据，可以考虑Kernel PCA，通过Kernel函数将非线性相关转为线性相关，关于这点就不展开讨论了。另外，PCA假设数据各主特征是分布在正交方向上，如果在非正交方向上存在几个方差较大的方向，PCA的效果就大打折扣了。

PCA技术的一个很大的优点是，它是完全无参数限制的。在PCA的计算过程中完全不需要人为的设定参数或是根据任何经验模型对计算进行干预，最后的结果只与数据相关，与用户是独立的。

但是，这一点同时也可以看作是缺点。如果用户对观测对象有一定的先验知识，掌握了数据的一些特征，却无法通过参数化等方法对处理过程进行干预，可能会得不到预期的效果，效率也不高。

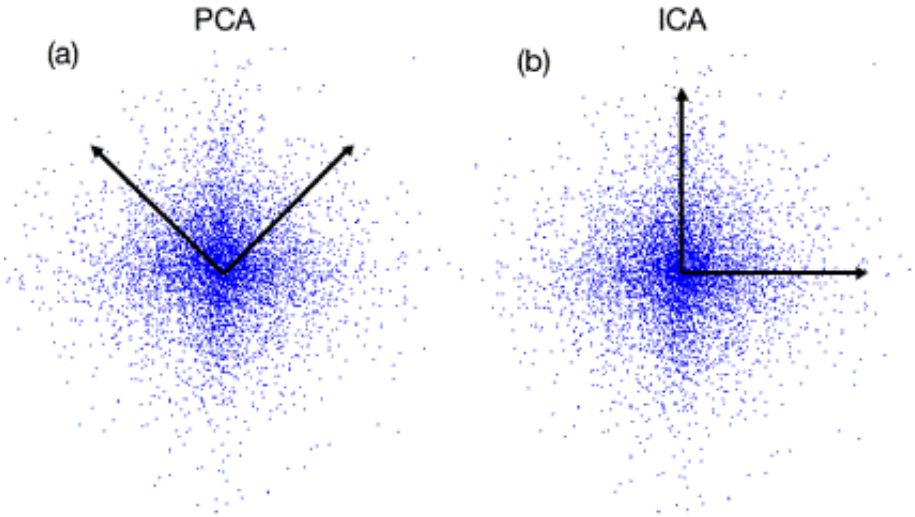


上图中黑色点表示采样数据，排列成转盘的形状。容易想象，该数据的主元是(P_1, P_2)或是旋转角θ。在这里PCA找出的主元将是(P_1, P_2)。但是这显然不是最优和最简化的主元。(P_1, P_2)之间存在着非线性的关系。根据先验的知识可知旋转角θ是最优的主元（类比极坐标）。则在这种情况下，PCA就会失效。但是，如果加入先验的知识，对数据进行某种划归，就可以将数据转化为以θ为线性的空间中。这类根据先验知识对数据预先进行非线性转换的方法就成为kernel-PCA，它扩展了PCA能够处理的问题的范围，又可以结合一些先验约束，是比较流行的方法。

有时数据的分布并不是满足高斯分布。如图表 5所示，在非高斯分布的情况下，PCA方法得出的主元可能并不是最优的。在寻找主元时不能将方差作为衡量重要性的标准。要根据数据的分布情况选择合适的描述完全分布的变量，然后根据概率分布式

$$P(y_1, y_2) = P(y_1)P(y_2)$$

来计算两个向量上数据分布的相关性。等价的，保持主元间的正交假设，寻找的主元同样要使$P(y_1, y_2) = 0$。这一类方法被称为独立主元分解(ICA)。



数据的分布并不满足高斯分布，呈明显的十字星状。这种情况下，方差最大的方向并不是最优主

元方向。另外PCA还可以用于预测矩阵中缺失的元素。