如何理解主元分析（PCA）？

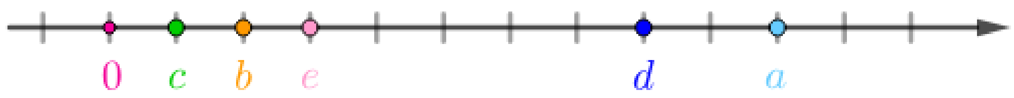
主元分析也就是PCA，主要用于数据降维。

1 什么是降维?

比如说有如下的房价数据：

	房价(百万元)
a	10
b	2
c	1
d	7
e	3

这种一维数据可以直接放在实数轴上：



不过数据还需要处理下，假设房价样本用 X 表示，那么均值为：

$$\overline{X}=\frac{X_1+X_2+X_3+X_4+X_5}{5}=\frac{10+2+1+7+3}{5}=4.6$$

然后以均值 $\overline{X}$ 为原点：

以 $\overline{X}$ 为原点的意思是，以 $\overline{X}$ 为0，那么上述表格的数字就需要修改下：

	房价(百万元)
a	$10-\overline{X}=5.4$
b	$2-\overline{X}=-2.6$
c	$1-\overline{X}=-3.6$
d	$7-\overline{X}=2.4$
e	$3-\overline{X}=-1.6$

这个过程称为“中心化”。“中心化”处理的原因是，这些数字后继会参与统计运算，比如求样本方差，中间就包含了
$X_i-\overline{X}$：

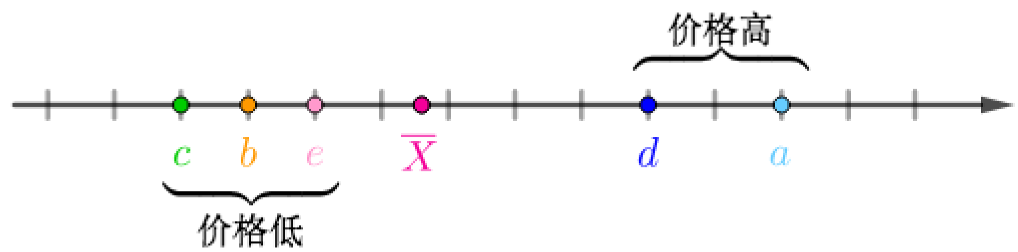
$$Var(X)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

说明下，虽然样本方差的分母是应该为n-1，这里分母采用n是因为这样算出来的样本方差$Var(X)$为一致估计量，不会太影响计算结果并且可以减小运算负担。

用“中心化”的数据就可以直接算出“房价”的样本方差：

$$Var(X)=\frac{1}{n}\left(5.4^2+(-2.6)^2+(-3.6)^2+2.4^2+(-1.6)^2\right)$$

“中心化”之后可以看出数据大概可以分为两类：



现在新采集了房屋的面积，可以看出两者完全正相关，有一列其实是多余的：

	房价(百万元)	面积(百平米)
a	10	10
b	2	2
c	1	1
d	7	7
e	3	3

求出房屋样本、面积样本的均值，分别对房屋样本、面积样本进行“中心化”后得到：

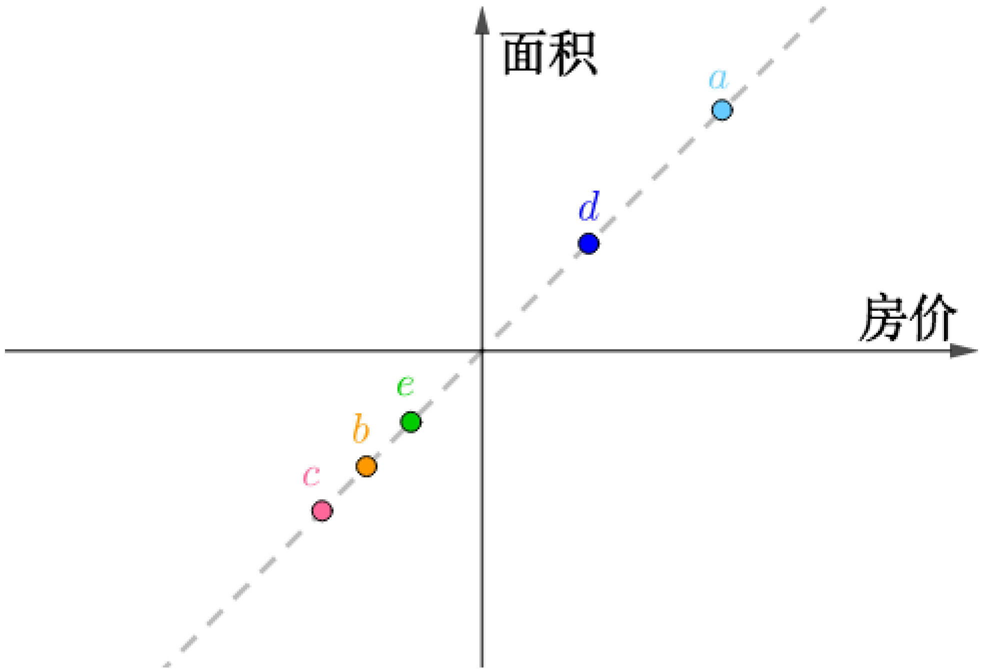
	房价(百万元)	面积(百平米)
a	5.4	5.4
b	-2.6	-2.6
c	-3.6	-3.6
d	2.4	2.4
e	-1.6	-1.6

房价（X）和面积（Y）的样本协方差是这样的（这里也是用的一致估计量）：

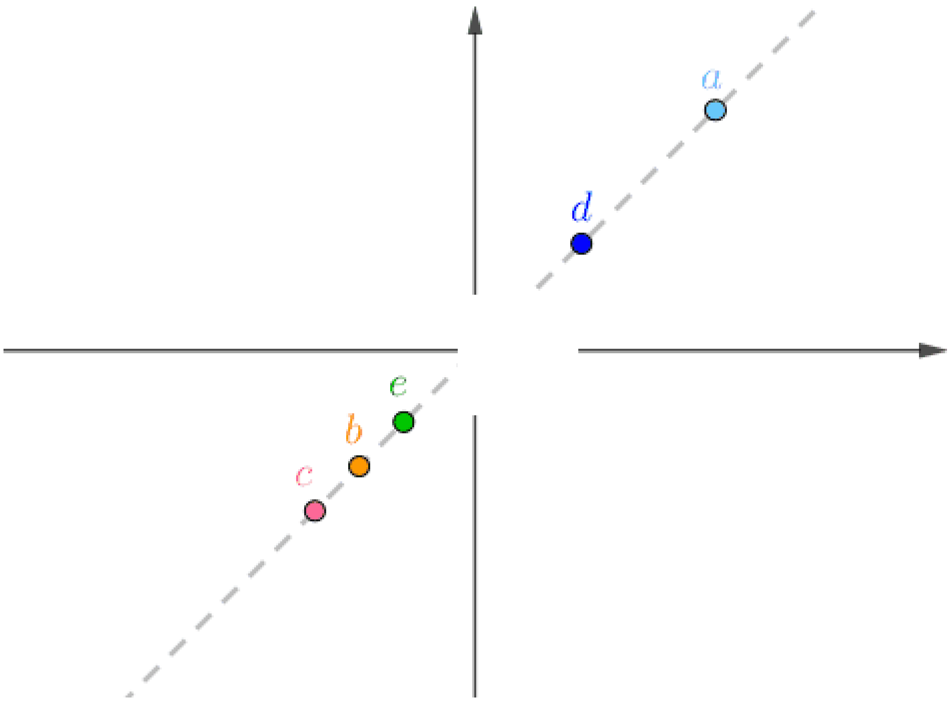
$$Cov(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$$

可见“中心化”后的数据可以简化上面这个公式，这点后面还会看到具体应用。

把这个二维数据画在坐标轴上，横纵坐标分别为“房价”、“面积”，可以看出它们排列为一条直线：

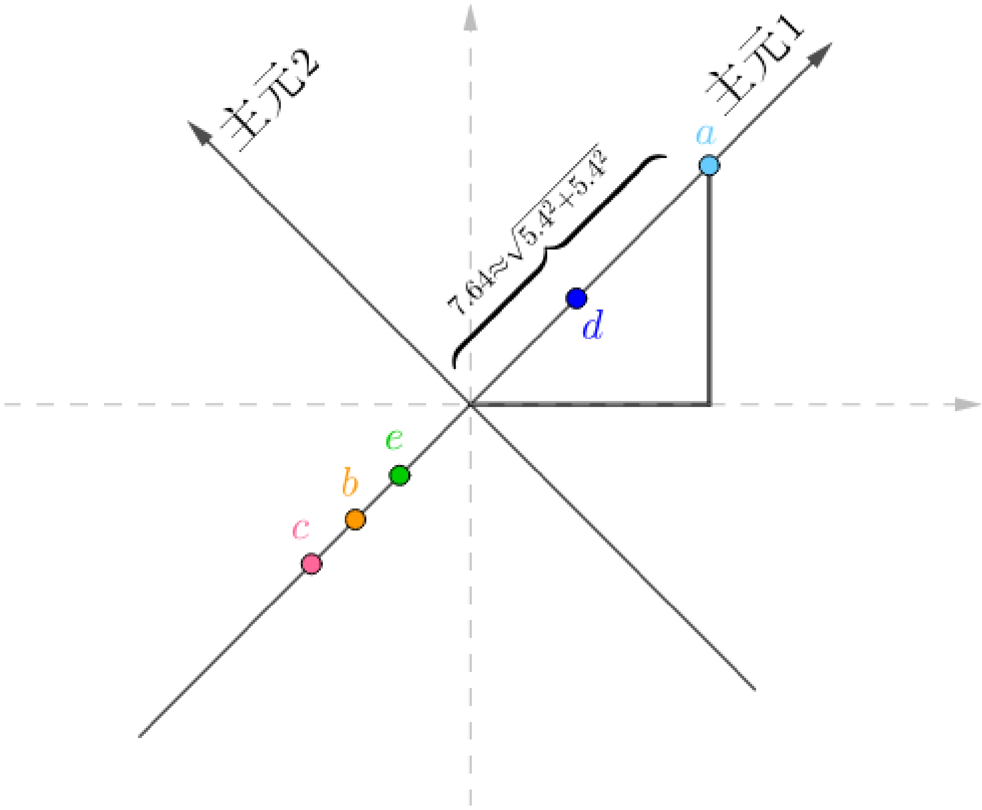


如果旋转坐标系，让横坐标和这条直线重合：



旋转后的坐标系，横纵坐标不再代表“房价”、“面积”了，而是两者的混合（术语是线性组合），这里把它们称作“主元1”、“主元2”，坐标值很容易用勾股定理计算出来，比如

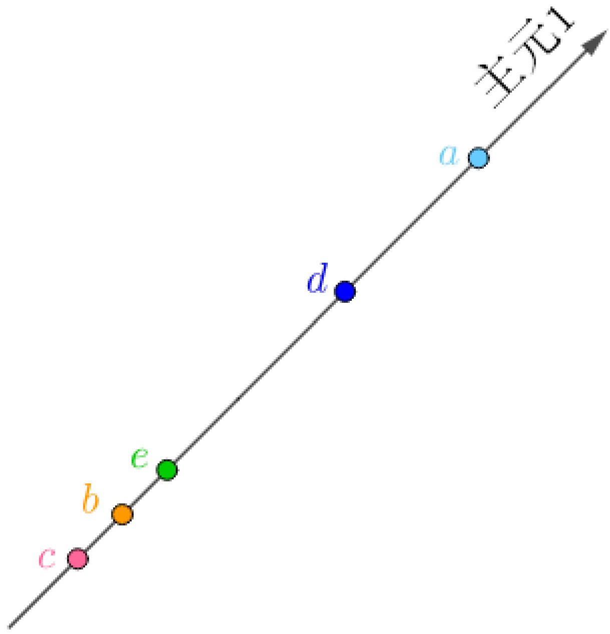
$\boldsymbol{a}$在“主元1”的坐标值为：



很显然$\boldsymbol{a}$在“主元2”上的坐标为0，把所有的房间换算到新的坐标系上：

	主元1	主元2
a	7.64	0
b	-3.68	0
c	-5.09	0
d	3.39	0
e	-2.26	0

因为“主元2”全都为0，完全是多余的，我们只需要“主元1”就够了，这样就又把数据降为了一维，而且没有丢失任何信息：



2 非理想情况如何降维?

上面是比较极端的情况，就是房价和面积完全正比，所以二维数据会在一条直线上。

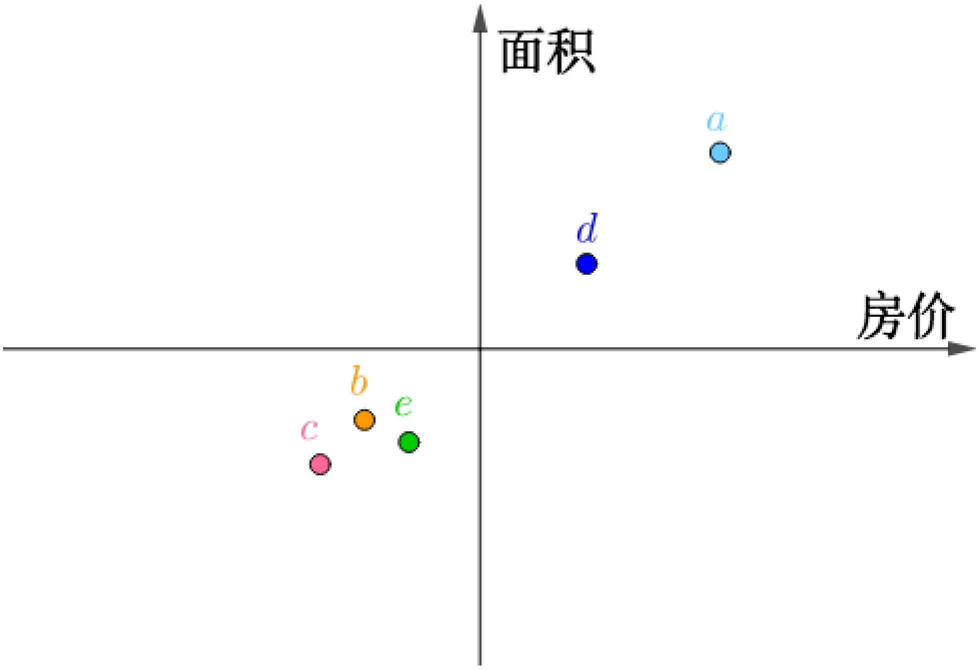
现实中虽然正比，但总会有些出入：

	房价(百万元)	面积(百平米)
a	10	9
b	2	3
c	1	2
d	7	6.5
e	3	2.5

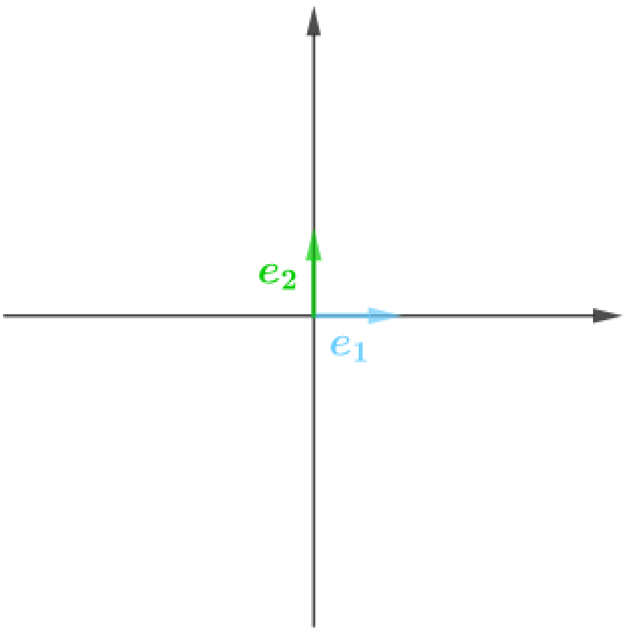
$\xrightarrow{\text{中心化}}$

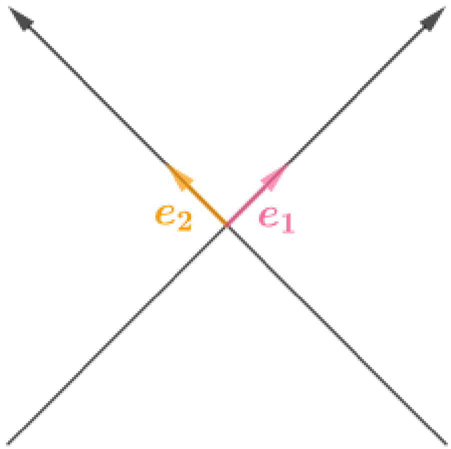
	房价(百万元)	面积(百平米)
a	5.4	4.4
b	-2.6	-1.6
c	-3.6	-2.6
d	2.4	1.9
e	-1.6	-2.1

把这个二维数据画在坐标轴上，横纵坐标分别为“房价”、“面积”，虽然数据看起来很接近一条直线，但是终究不在一条直线上：



那么应该怎么降维呢？分析一下，从线性代数的角度来看，二维坐标系总有各自的标准正交基（也就是两两正交、模长为1的基），
$\boldsymbol{e_1}, \boldsymbol{e_2}$：

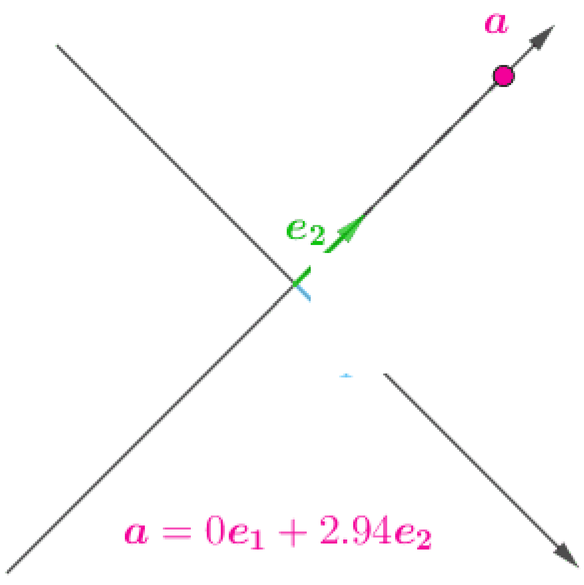




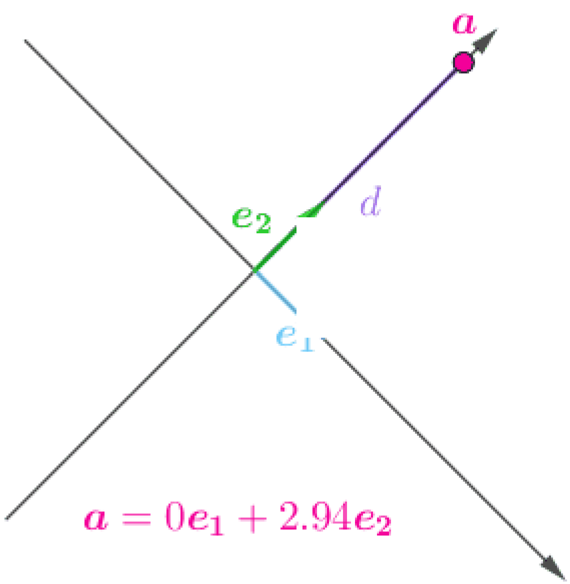
在某坐标系有一个点，$\boldsymbol{a} = \begin{pmatrix} x \\ y \end{pmatrix}$，它表示在该坐标系下标准正交基 $\boldsymbol{e_1}, \boldsymbol{e_2}$ 的线性组合：

$$\boldsymbol{a} = \begin{pmatrix} x \\ y \end{pmatrix} = x\boldsymbol{e_1} + y\boldsymbol{e_2}$$

只是在不同坐标系中，$\boldsymbol{x}, \boldsymbol{y}$ 的值会有所不同（旋转的坐标表示不同的坐标系）：



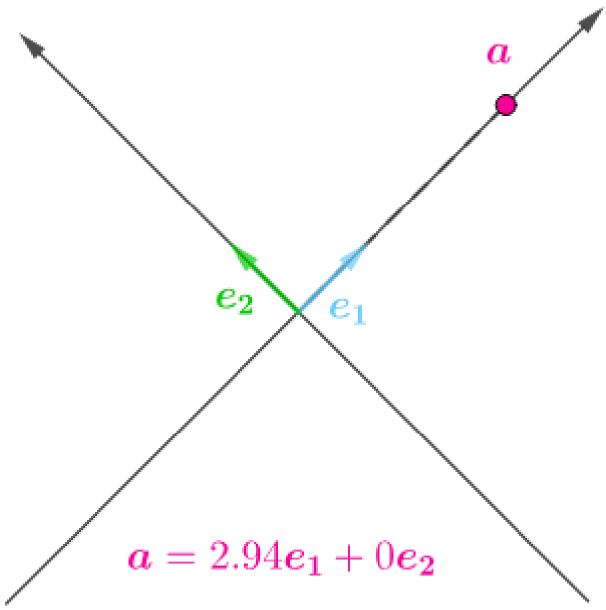
因为$\boldsymbol{a}$到原点的距离d不会因为坐标系改变而改变：



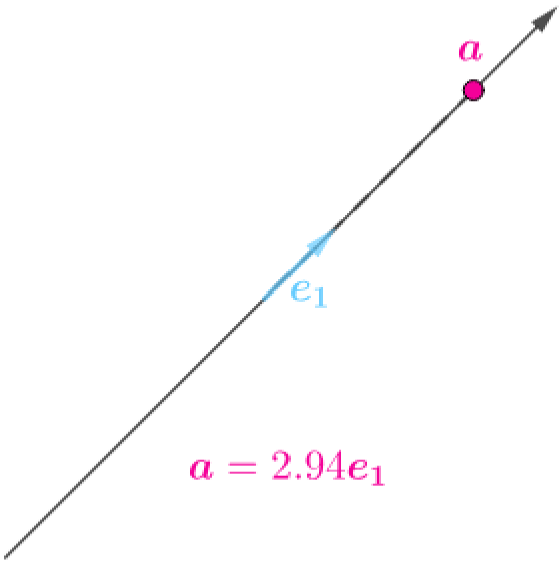
而：

$$d^2 = x^2 + y^2$$

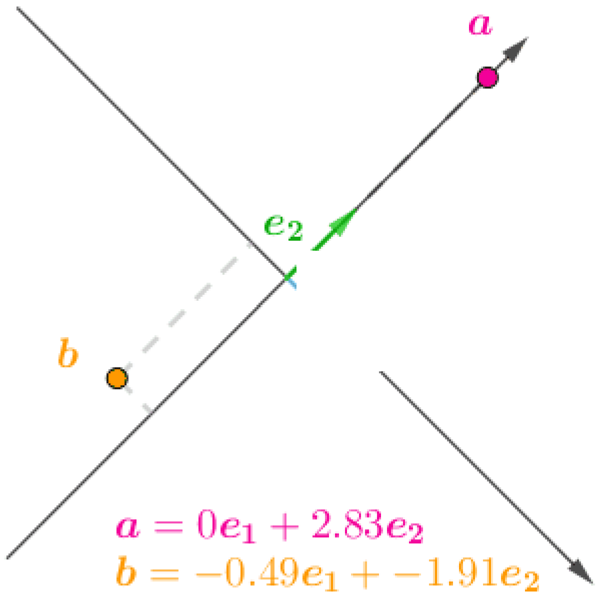
所以，在某坐标系下分配给x较多，那么分配给y
的就必然较少，反之亦然。最极端的情况是，在某个坐标系下，全部分配给了x，使得$y = 0$：



那么在这个坐标系中，就可以降维了，去掉 $\boldsymbol{e_2}$ 并不会丢失信息：



如果是两个点 $\boldsymbol{a} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$，情况就复杂一些：



为了降维，应该选择尽量多分配给x_1, x_2，少分配给y_1, y_2的坐标系。

3 主元分析（PCA）

怎么做呢？假设有如下数据：

	X	Y
a	a_1	b_1
b	a_2	b_2

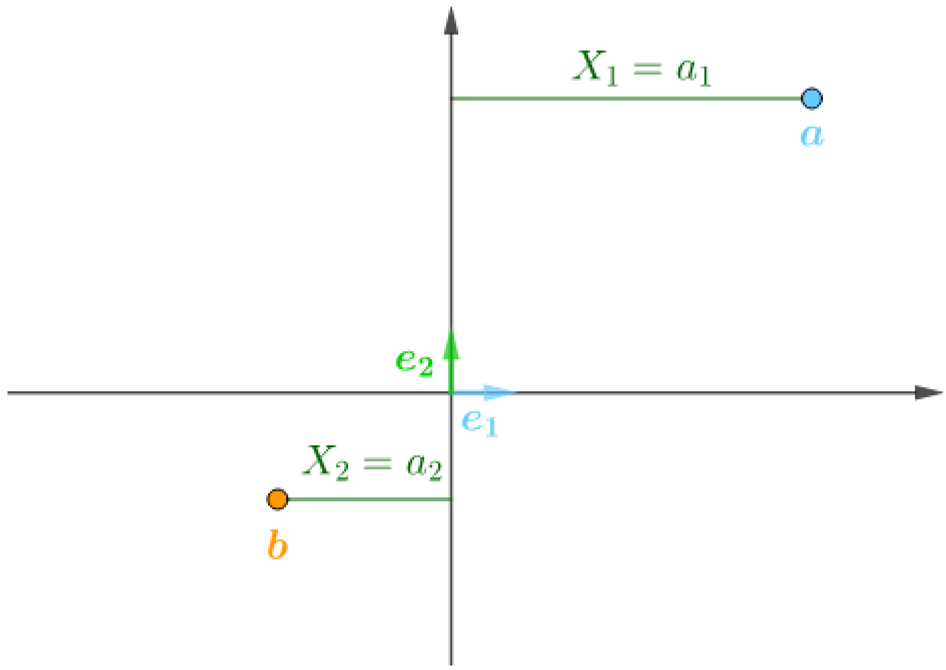
上面的数据这么解读，表示有两个点：

$$\boldsymbol{a} = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}$$

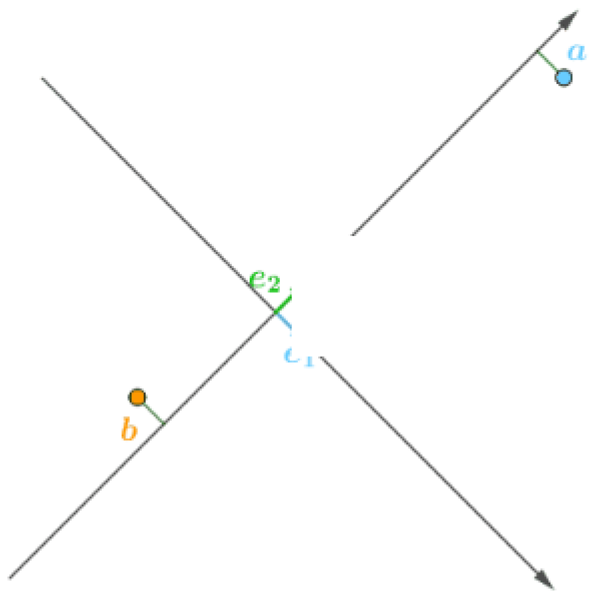
这两个点在初始坐标系下（也就是自然基$\boldsymbol{e_1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \boldsymbol{e_2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$）下坐标值为：

$$\boldsymbol{a} = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix} = \begin{pmatrix} a_2 \\ b_2 \end{pmatrix}$$

图示如下：



随着坐标系的不同，X_1, X_2 的值会不断变化：



要想尽量多分配给 X_1, X_2，借鉴最小二乘法（请参考如何理解最小二乘法）的思想，就是让：

$$X_1^2 + X_2^2 = \sum_{i=0}^{2} X_i^2 \quad 最大$$

要求这个问题，先看看 X_1, X_2 怎么表示，假设：

$$\boldsymbol{e_1} = \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} \quad \boldsymbol{e_2} = \begin{pmatrix} e_{21} \\ e_{22} \end{pmatrix}$$

根据点积的几何意义（如何通俗地理解协方差和点积）有：

$$X_1 = \boldsymbol{a} \cdot \boldsymbol{e_1} = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \cdot \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} = a_1 e_{11} + b_1 e_{12}$$

$$X_2 = \boldsymbol{b} \cdot \boldsymbol{e_1} = \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \cdot \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} = a_2 e_{11} + b_2 e_{12}$$

那么：

$$\begin{aligned} X_1^2 + X_2^2 &= (a_1 e_{11} + b_1 e_{12})^2 + (a_2 e_{11} + b_2 e_{12})^2 \\ &= a_1^2 e_{11}^2 + 2a_1 b_1 e_{11} e_{12} + b_1^2 e_{12}^2 + a_2^2 e_{11}^2 + 2a_2 b_2 e_{11} e_{12} + b_2^2 e_{12}^2 \\ &= (a_1^2 + a_2^2) e_{11}^2 + 2(a_1 b_1 + a_2 b_2) e_{11} e_{12} + (b_1^2 + b_2^2) e_{12}^2 \end{aligned}$$

上式其实是一个二次型（可以参看如何通俗地理解二次型）：

$$X_1^2 + X_2^2 = \boldsymbol{e_1}^{\mathrm{T}} \underbrace{\begin{pmatrix} a_1^2 + a_2^2 & a_1 b_1 + a_2 b_2 \\ a_1 b_1 + a_2 b_2 & b_1^2 + b_2^2 \end{pmatrix}}_{P} \boldsymbol{e_1} = \boldsymbol{e_1}^{\mathrm{T}} P \boldsymbol{e_1}$$

这里矩阵 P 就是二次型，是一个对称矩阵，可以进行如下的奇异值分解（可以参看如何通俗地理解奇异值分解）：

$$P = U \Sigma U^{\mathrm{T}}$$

其中，U 为正交矩阵，即 $UU^{\mathrm{T}} = I$。

而 Σ 是对角矩阵：

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

其中，σ_1, σ_2 是奇异值，$\sigma_1 > \sigma_2$。

将 P 代回去：

$$\begin{aligned} X_1^2 + X_2^2 &= \boldsymbol{e_1}^{\mathrm{T}} P \boldsymbol{e_1} \\ &= \boldsymbol{e_1}^{\mathrm{T}} U \Sigma U^{\mathrm{T}} \boldsymbol{e_1} \\ &= (U^{\mathrm{T}} \boldsymbol{e_1})^{\mathrm{T}} \Sigma (U^{\mathrm{T}} \boldsymbol{e_1}) \end{aligned}$$

因为 U 是正交矩阵，所以令：

$$\boldsymbol{n} = U^{\mathrm{T}}\boldsymbol{e_1}$$

所得的$\boldsymbol{n}$也是单位向量，即：

$$\boldsymbol{n} = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \implies n_1^2 + n_2^2 = 1$$

继续回代：

$$\begin{aligned} X_1^2 + X_2^2 &= (U^{\mathrm{T}}\boldsymbol{e_1})^{\mathrm{T}}\Sigma(U^{\mathrm{T}}\boldsymbol{e_1}) \\ &= \boldsymbol{n}^{\mathrm{T}}\Sigma\boldsymbol{n} \\ &= \begin{pmatrix} n_1 & n_2 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \\ &= \sigma_1 n_1^2 + \sigma_2 n_2^2 \end{aligned}$$

最初求最大值的问题就转化为了：

$$X_1^2 + X_2^2 = \sum_{i=0}^{2} X_i^2 \text{ 最大} \iff \begin{cases} \sigma_1 n_1^2 + \sigma_2 n_2^2 \text{ 最大} \\ n_1^2 + n_2^2 = 1 \\ \sigma_1 > \sigma_2 \end{cases}$$

感兴趣可以用拉格朗日乘子法计算上述条件极值（参看如何通俗地理解拉格朗日乘子法以及KKT条件），结果是当$n_1 = 1, n_2 = 0$时取到极值。

因此可以推出要寻找的主元1，即：

$$\boldsymbol{n} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = U^{\mathrm{T}}\boldsymbol{e_1} \implies \boldsymbol{e_1} = U\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

总结下：

$$\boldsymbol{e_1} = \begin{cases} P = U\Sigma U^{\mathrm{T}} \\ \text{最大奇异值}\sigma_1\text{对应的奇异向量} \end{cases}$$

同样的思路可以求出：

$$\boldsymbol{e_2} = \begin{cases} P = U\Sigma U^{\mathrm{T}} \\ \text{最小奇异值}\sigma_2\text{对应的奇异向量} \end{cases}$$

4 协方差矩阵

上一节的数据：

	X	Y
a	a_1	b_1
b	a_2	b_2

我们按行来解读，得到了两个向量$\boldsymbol{a},\boldsymbol{b}$：

		X	Y
$\boldsymbol{a}:$	a	a_1	b_1
$\boldsymbol{b}:$	b	a_2	b_2

在这个基础上推出了矩阵：

$$P=\begin{pmatrix}a_1^2+a_2^2 & a_1b_1+a_2b_2\\ a_1b_1+a_2b_2 & b_1^2+b_2^2\end{pmatrix}$$

这个矩阵是求解主元1、主元2的关键。

如果我们按列来解读，可以得到两个向量$\boldsymbol{X},\boldsymbol{Y}$：

	$\boldsymbol{X}$	$\boldsymbol{Y}$
	X	Y
a	a_1	b_1
b	a_2	b_2

即：

$$\boldsymbol{X}=\begin{pmatrix}a_1\\ a_2\end{pmatrix}\quad \boldsymbol{Y}=\begin{pmatrix}b_1\\ b_2\end{pmatrix}$$

那么刚才求出来的矩阵就可以表示为：

$$P=\begin{pmatrix}a_1^2+a_2^2 & a_1b_1+a_2b_2\\ a_1b_1+a_2b_2 & b_1^2+b_2^2\end{pmatrix}=\begin{pmatrix}X\cdot X & X\cdot Y\\ X\cdot Y & Y\cdot Y\end{pmatrix}$$

之前说过“中心化”后的样本方差（关于样本方差、协方差可以参看这篇文章：

如何通俗地理解协方差和点积）：

$$Var(X)=\frac{1}{n}\sum_{i=1}^{n}X_i^2=\frac{1}{n}X\cdot X$$

样本协方差为：

$$Cov(X,Y)=\frac{1}{n}\sum_{i=1}^{n}X_iY_i=\frac{1}{n}X\cdot Y$$

两相比较可以得到一个新的矩阵，也就是协方差矩阵：

$$Q=\frac{1}{n}P=\begin{pmatrix}Var(X) & Cov(X,Y)\\ Cov(X,Y) & Var(Y)\end{pmatrix}$$

P,Q 都可以进行奇异值分解：

$$P=U\begin{pmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{pmatrix}U^{\mathrm{T}}\quad Q=\frac{1}{n}P=U\begin{pmatrix}\frac{\sigma_1}{n} & 0\\ 0 & \frac{\sigma_2}{n}\end{pmatrix}U^{\mathrm{T}}$$

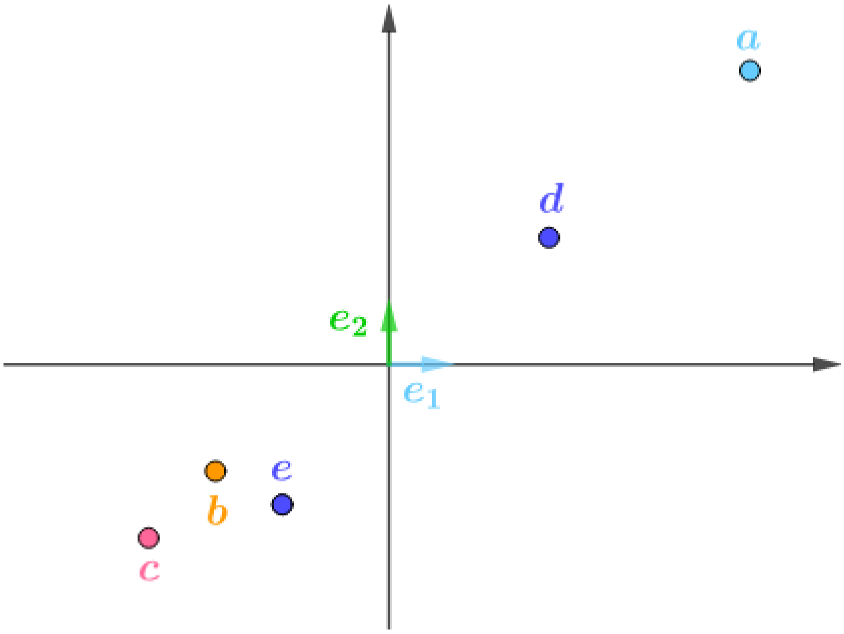
可见，协方差矩阵 Q 的奇异值分解和 P 相差无几，只是奇异值缩小了 n 倍，但是不妨碍奇异值之间的大小关系，所以在实际问题中，往往都是直接分解协方差矩阵 Q 。

5 实战

回到使用之前“中心化”了的数据：

	房价(百万元)	面积(百平米)
a	5.4	4.4
b	-2.6	-1.6
c	-3.6	-2.6
d	2.4	1.9
e	-1.6	-2.1

这些数据按行，在自然基下画出来就是：



按列解读得到两个向量：

$$\boldsymbol{X}=\begin{pmatrix}5.4\\-2.6\\-3.6\\2.4\\-1.6\end{pmatrix}\quad \boldsymbol{Y}=\begin{pmatrix}4.4\\-1.6\\-2.6\\1.9\\-2.1\end{pmatrix}$$

组成协方差矩阵：

$$Q=\begin{pmatrix}Var(X) & Cov(X,Y)\\ Cov(X,Y) & Var(Y)\end{pmatrix}=\frac{1}{5}\begin{pmatrix}X\cdot X & X\cdot Y\\ X\cdot Y & Y\cdot Y\end{pmatrix}=\frac{1}{5}\begin{pmatrix}57.2 & 45.2\\ 45.2 & 36.7\end{pmatrix}$$

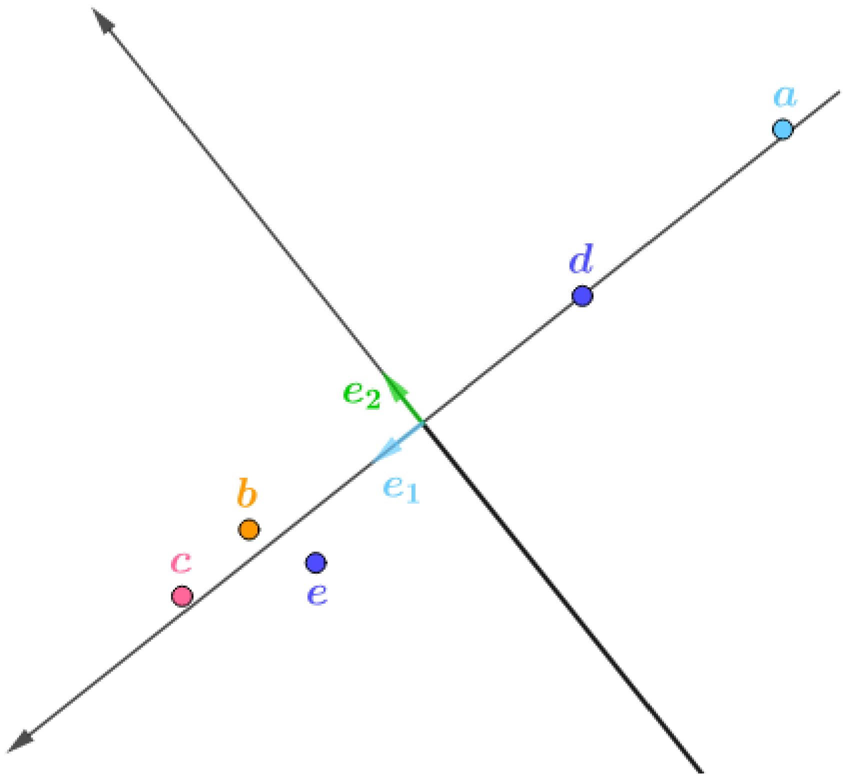
进行奇异值分解：

$$Q\approx\begin{pmatrix}-0.78 & -0.62\\ -0.62 & 0.78\end{pmatrix}\begin{pmatrix}18.66 & 0\\ 0 & 0.12\end{pmatrix}\begin{pmatrix}-0.78 & -0.62\\ -0.62 & 0.78\end{pmatrix}$$

根据之前的分析，主元1应该匹配最大奇异值对应的奇异向量，主元2匹配最小奇异值对应的奇异向量，即：

$$\boldsymbol{e_1}=\begin{pmatrix}-0.78\\-0.62\end{pmatrix}\quad \boldsymbol{e_2}=\begin{pmatrix}-0.62\\0.78\end{pmatrix}$$

以这两个为主元画出来的坐标系就是这样的：



如下算出新坐标，比如对于$\boldsymbol{a}$：

$$X_1=\boldsymbol{a}\cdot\boldsymbol{e_1}=-6.94\quad X_2=\boldsymbol{a}\cdot\boldsymbol{e_2}=0.084$$

以此类推，得到新的数据表：

	主元1	主元2
a	-6.94	0.084
b	3.02	0.364
c	4.42	0.204
d	-3.05	-0.006
e	2.55	-0.646

主元2整体来看，数值很小，丢掉损失的信息也非常少，这样就实现了非理想情况下的降维。

标签与声明

标签：主元分析