机器学习基础

主成分分析

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 基本思想
- 总体主成分分析
- 样本主成分分析
- 应用：基因组分析

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 基本思想
- 总体主成分分析
- 样本主成分分析
- 应用：基因组分析

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

- 采集了房屋的价格和面积，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

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

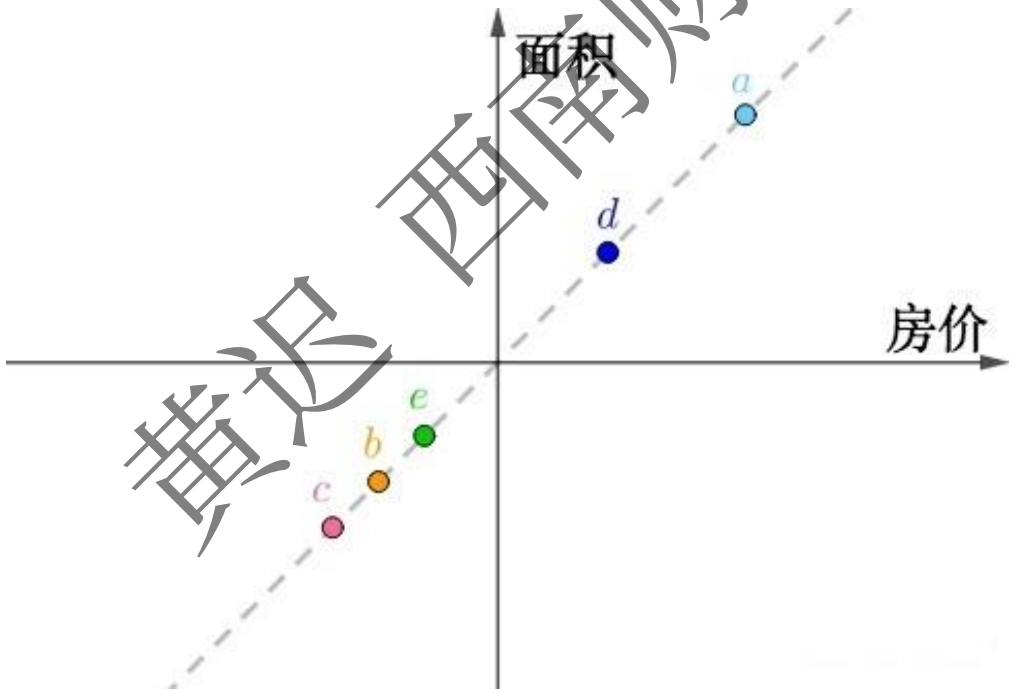
- 求出房价样本、面积样本的均值，分别对房价样本、面积样本进行“中心化”后得到：

	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

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

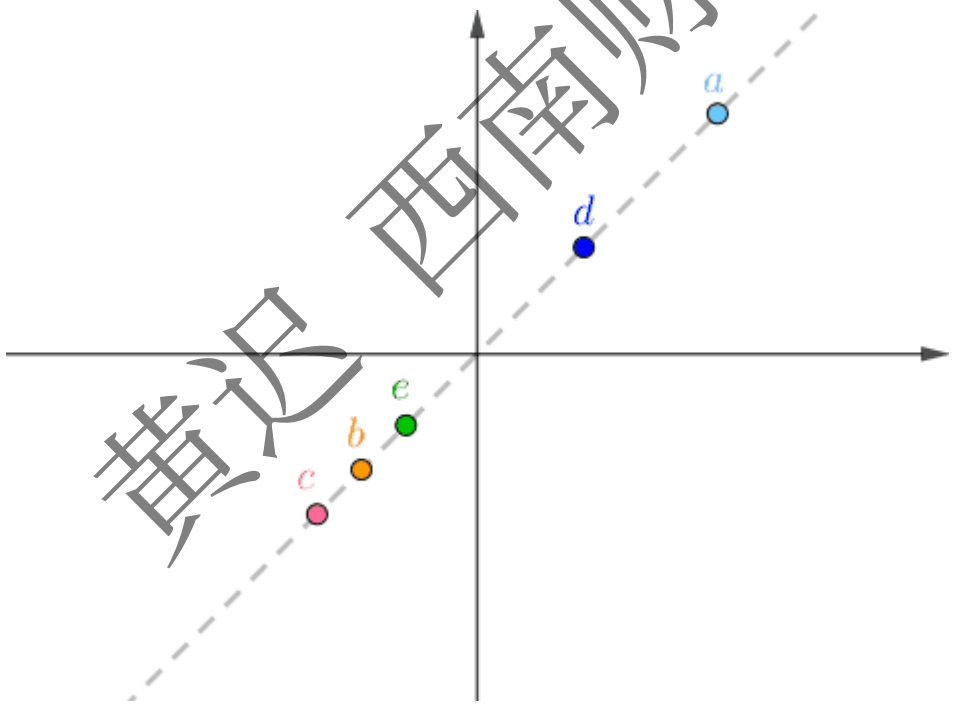
- 把样本数据画在坐标系上



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

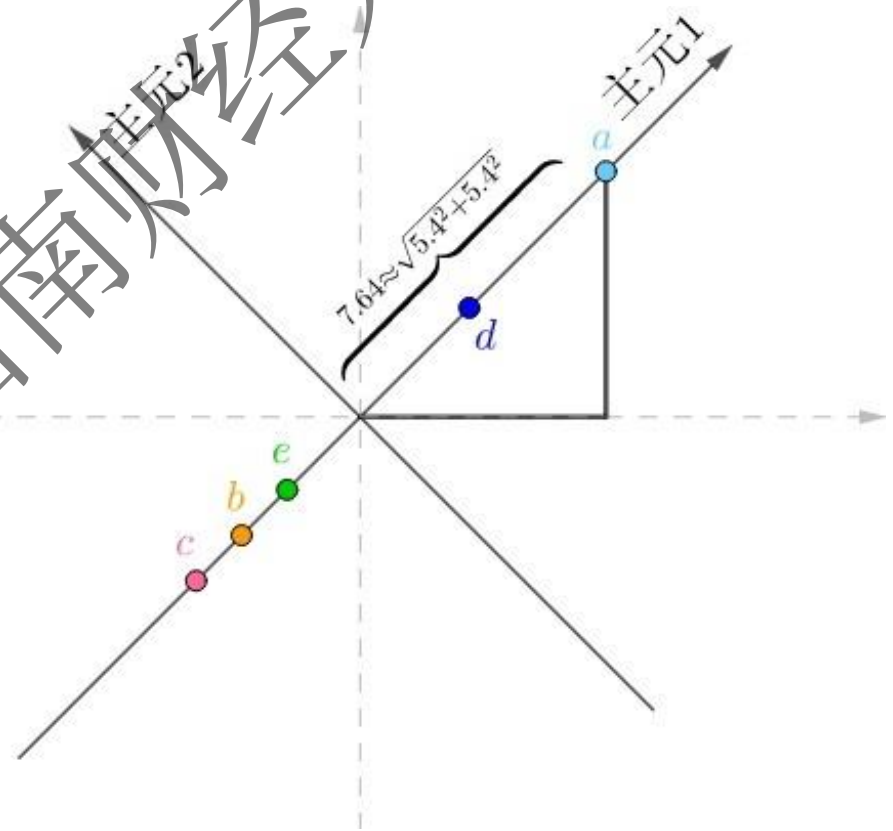
- 旋转坐标系，让横坐标和这条直线重合



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

- 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（线性组合），即“主元1”、“主元2”



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

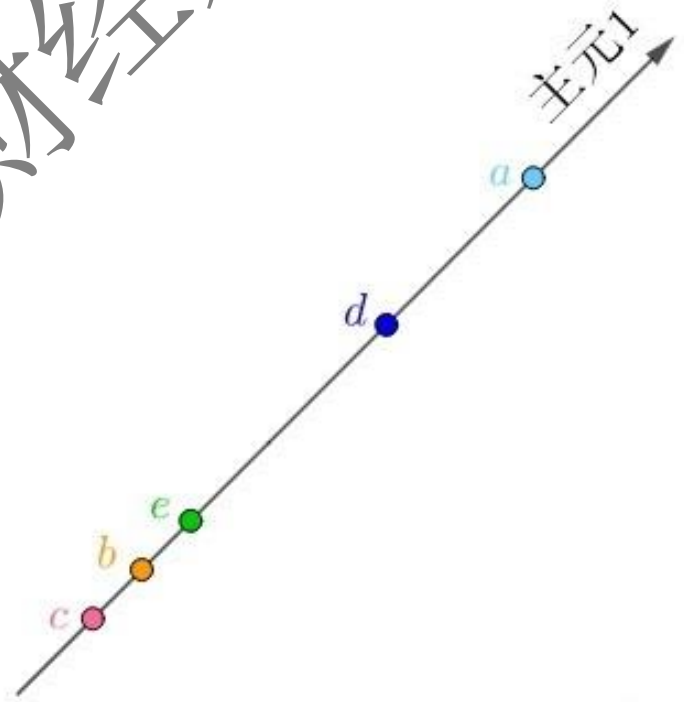
- 重新计算样本在新坐标系下的坐标有

	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

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

- 所有样本在主元2上的坐标都是0，所以只需要保留主元1，这样就把数据降为了一维，而且没有丢失任何信息



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

更加真实的例子

- 样本的房价和面积不再完全相同

	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

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

更加真实的例子

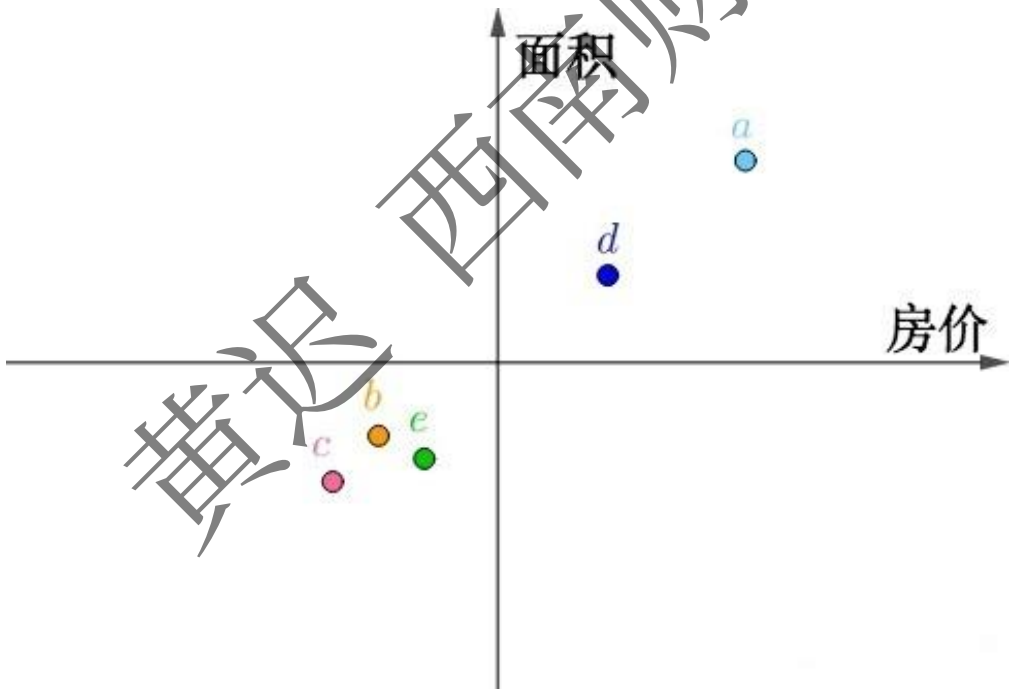
- 中心化

	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

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

更加真实的例子

- 把样本数据画在坐标系上



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

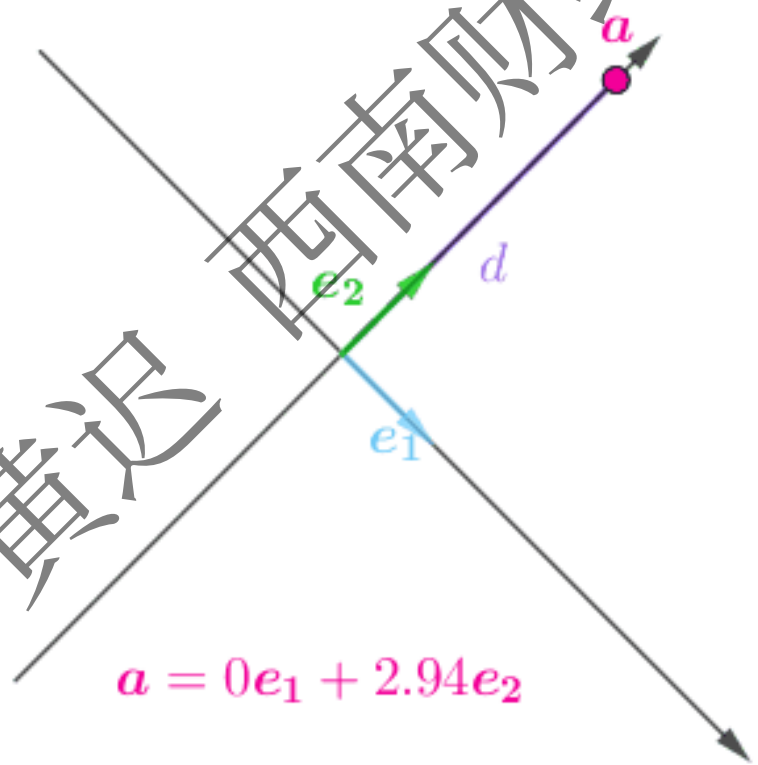
更加真实的例子

- 如何降维？
  - 假设 $\boldsymbol{e}_1, \boldsymbol{e}_2$ 是二维平面中任意一组标准正交基　为什么要求正交？
  - 点 $\boldsymbol{a}$ 在 $\boldsymbol{e}_1, \boldsymbol{e}_2$ 下的坐标为 (x, y)，即 $\boldsymbol{a} = x\boldsymbol{e}_1 + y\boldsymbol{e}_2$
  - 当 $\boldsymbol{e}_1, \boldsymbol{e}_2$ 变化时，坐标会发生改变，但点 $\boldsymbol{a}$ 到坐标原点的距离不会发生改变，即 $d = x^2 + y^2$ 是常数
  - 若 x 变大，y 会变小；反之亦然

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

如何降维？

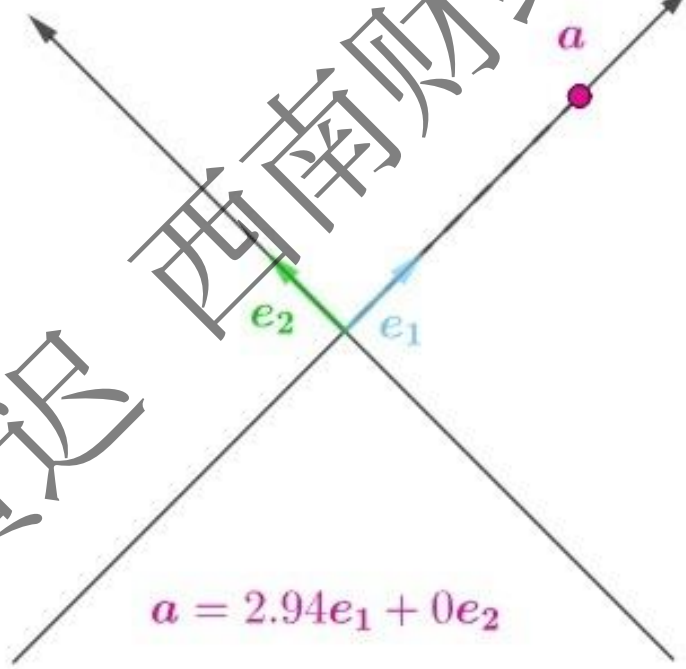
- x 和 y 的变化



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

如何降维？

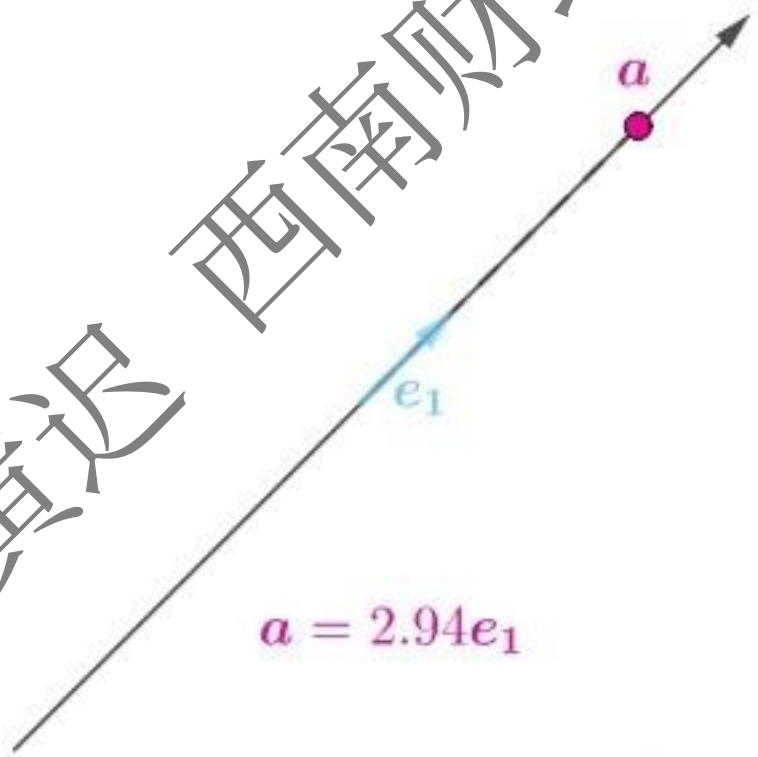
- 极端情况下，$y=0$



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

如何降维？

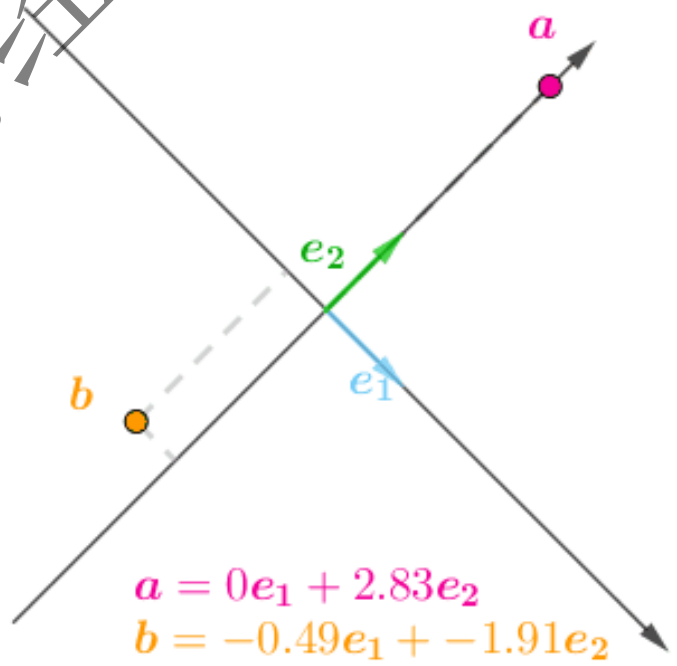
- 因此去掉 e_2 不会丢失信息，实现了降维



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

如何降维？

- 如果是两个点 $a=(x_1,y_1)$, $b=(x_2,y_2)$, 则很难使 y_1, y_2 同时为 0
- 为了降维，应该尽量增大 x_1, x_2，相应的 y_1, y_2 就会减小



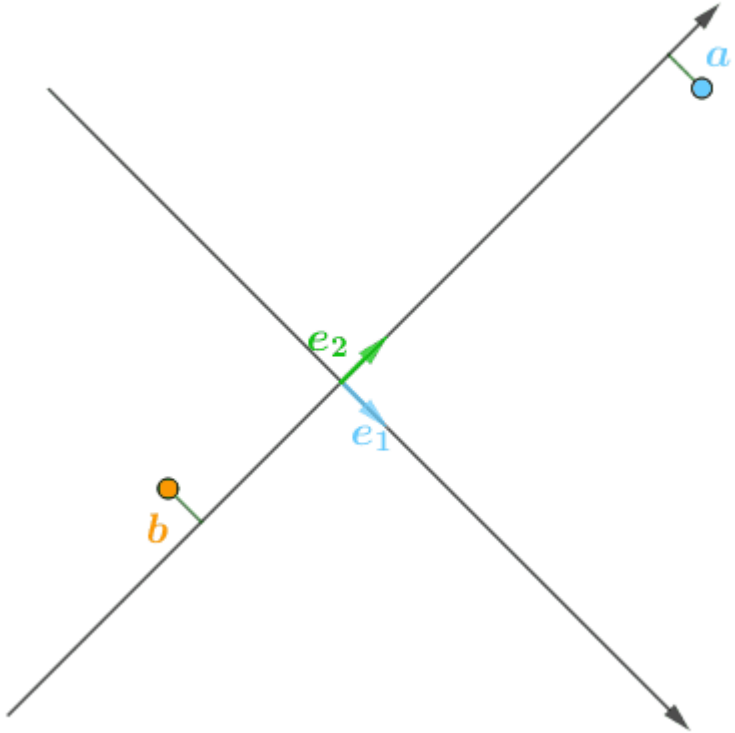
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

如何降维？

- 令样本在不同的标准正交基下坐标为

	$\boldsymbol{X}=(1,0)$	$\boldsymbol{Y}=(0,1)$	$\boldsymbol{e_1}$	$\boldsymbol{e_2}$
$\boldsymbol{a}$	a_x	a_y	x_1	y_1
$\boldsymbol{b}$	b_x	b_y	x_2	y_2

- 随着坐标系的变化，x_1, x_2 也在不断变化



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

如何降维？

- 为了实现降维，应该要让 x_1, x_2 尽量大，即实现 $\max\ x_1^2 + x_2^2$
- 由于 $x_1 = \boldsymbol{a} \cdot \boldsymbol{e_1}$, $x_2 = \boldsymbol{b} \cdot \boldsymbol{e_1}$，有

$$x_1^2 + x_2^2 = \boldsymbol{e_1}^\top \boldsymbol{a}^\top \boldsymbol{a} \boldsymbol{e_1} + \boldsymbol{e_1}^\top \boldsymbol{b}^\top \boldsymbol{b} \boldsymbol{e_1}$$

$$= \boldsymbol{e_1}^\top \begin{pmatrix} a_x^2 + b_x^2 & a_x a_y + b_x b_y \\ a_x a_y + b_x b_y & a_y^2 + b_y^2 \end{pmatrix} \boldsymbol{e_1} \triangleq \boldsymbol{e_1}^\top P \boldsymbol{e_1}$$

- 矩阵 P 是二次型矩阵，存在正交分解 $P = U \Sigma U^\top = U \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} U^\top$

  其中 σ_i 是特征值（奇异值），且 $\sigma_1 \geq \sigma_2$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

如何降维？

- 将分解代回，有 $x_1^2+x_2^2=\boldsymbol{e_1}^\top U\Sigma U^\top \boldsymbol{e_1}$
- 令 $\boldsymbol{n}=\begin{pmatrix}n_1\\n_2\end{pmatrix}=U^\top \boldsymbol{e_1}$，则 $\boldsymbol{n}$ 也是单位向量，即 $n_1^2+n_2^2=1$
- 继续回代有 $x_1^2+x_2^2=\boldsymbol{n}^\top\Sigma\boldsymbol{n}=\boldsymbol{n}^\top\begin{pmatrix}\sigma_1&0\\0&\sigma_2\end{pmatrix}\boldsymbol{n}=\sigma_1 n_1^2+\sigma_2 n_2^2$
- 最终优化问题变为

$$\max\ x_1^2+x_2^2 \quad\Rightarrow\quad \begin{aligned}\max\quad & \sigma_1 n_1^2+\sigma_2 n_2^2\\ s.t.\quad & n_1^2+n_2^2=1\\ & \sigma_1\geq\sigma_2\end{aligned}$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

如何降维？

- 当 $n_1=1$，$n_2=0$ 时，该优化问题达到最优，此时有

$$\boldsymbol{n}=\begin{pmatrix}1\\0\end{pmatrix}=U^{\top}\boldsymbol{e_1}\ \Rightarrow \boldsymbol{e_1}=U\begin{pmatrix}1\\0\end{pmatrix}=Col_1(U)$$

  即 $\boldsymbol{e_1}$ 是最大特征值 σ_1 的特征向量

- 由于 $\boldsymbol{e_1}$，$\boldsymbol{e_2}$ 相互正交，因此 $\boldsymbol{e_2}=Col_2(U)$

  即 $\boldsymbol{e_2}$ 是第二特征值 σ_2 的特征向量

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

统计分析

- 回顾样本数据：

	$\boldsymbol{X}=(1,0)$	$\boldsymbol{Y}=(0,1)$	$\boldsymbol{e_1}$	$\boldsymbol{e_2}$
$\boldsymbol{a}$	a_x	a_y	x_1	y_1
$\boldsymbol{b}$	b_x	b_y	x_2	y_2

- 令 $\boldsymbol{x}=\begin{pmatrix}a_x\\b_x\end{pmatrix}$, $\boldsymbol{y}=\begin{pmatrix}a_y\\b_y\end{pmatrix}$，分别表示两个特征在样本上的分布，有

$$P=\begin{pmatrix}a_x^2+b_x^2 & a_xa_y+b_xb_y\\ a_xa_y+b_xb_y & a_y^2+b_y^2\end{pmatrix}=\begin{pmatrix}\boldsymbol{x}\cdot\boldsymbol{x} & \boldsymbol{x}\cdot\boldsymbol{y}\\ \boldsymbol{x}\cdot\boldsymbol{y} & \boldsymbol{y}\cdot\boldsymbol{y}\end{pmatrix}$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

统计分析

- 由于数据已经中心化，所以有
  - 样本方差 $Var(\boldsymbol{x})=\boldsymbol{x}\cdot\boldsymbol{x}$, $Var(\boldsymbol{y})=\boldsymbol{y}\cdot\boldsymbol{y}$
  - 样本协方差 $Cov(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}\cdot\boldsymbol{y}$
- 矩阵 P 就是样本协方差矩阵，即

$$P=\begin{pmatrix}\boldsymbol{x}\cdot\boldsymbol{x} & \boldsymbol{x}\cdot\boldsymbol{y}\\ \boldsymbol{x}\cdot\boldsymbol{y} & \boldsymbol{y}\cdot\boldsymbol{y}\end{pmatrix}=\frac{1}{n-1}\begin{pmatrix}Var(\boldsymbol{x}) & Cov(\boldsymbol{x},\boldsymbol{y})\\ Cov(\boldsymbol{x},\boldsymbol{y}) & Var(\boldsymbol{y})\end{pmatrix}$$

- 主元是样本协方差矩阵的单位特征向量

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

回到例子

	房价	面积
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

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

$$P=\frac{1}{n-1}\begin{pmatrix}Var(\boldsymbol{x}) & Cov(\boldsymbol{x},\boldsymbol{y})\\Cov(\boldsymbol{x},\boldsymbol{y}) & Var(\boldsymbol{y})\end{pmatrix}$$

$$=\frac{1}{4}\begin{pmatrix}57.2 & 45.2\\45.2 & 36.7\end{pmatrix}$$

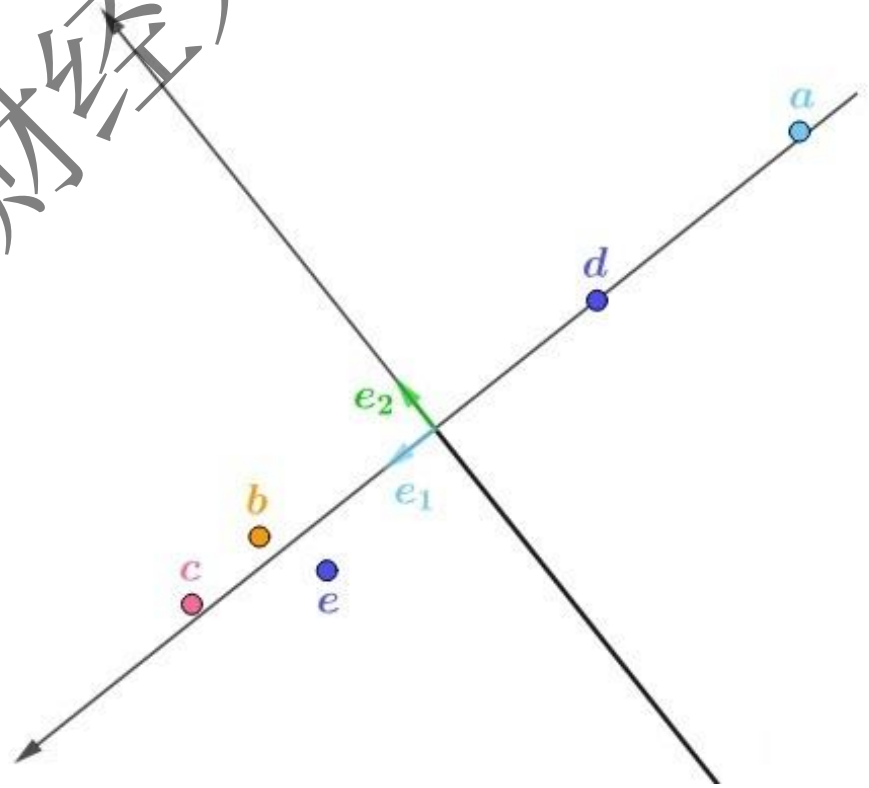
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

回到例子

- 求得样本协方差矩阵的特征值，即主成分分别为

$$\boldsymbol{e}_1=\begin{pmatrix}-0.78\\-0.62\end{pmatrix},\ \boldsymbol{e}_2=\begin{pmatrix}-0.62\\0.78\end{pmatrix}$$

- 在主成分 $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ 下，样本如图所示：



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

回到例子

- 样本在主成分下的坐标为

	主成分 1	主成分 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

- 样本在主成分2中的数值很小，损失信息也很小，降维效果好

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 基本思想
- 总体主成分分析
- 样本主成分分析
- 应用：基因组分析

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

定义

- 假设$\boldsymbol{x}=(x_1,x_2,\cdots,x_m)^\top$是$m$维随机变量（原始坐标）
  - 均值$\boldsymbol{\mu}=\mathbb{E}(\boldsymbol{x})$；协方差矩阵$\Sigma=Cov(\boldsymbol{x},\boldsymbol{x})=\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top]$
- 定义m维随机变量$\boldsymbol{y}=(y_1,y_2,\cdots,y_m)^\top$（旋转后的坐标，即主成分）：

$$y_i=\alpha_i^\top\boldsymbol{x}=\alpha_{1i}x_1+\alpha_{2i}x_2+\cdots+\alpha_{mi}x_m$$

  - $\mathbb{E}(y_i)=\mathbb{E}(\alpha_i^\top\boldsymbol{x})=\alpha_i^\top\boldsymbol{\mu},\ Var(y_i)=\alpha_i^\top\Sigma\alpha_i$

    $Cov(y_i,y_j)=\alpha_i^\top\Sigma\alpha_j,\ i=1,2,\cdots,m,\ j=1,2,\cdots,m$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

总体主成分

- 对 $y_i,\ i=1,2,\cdots,m$，满足以下条件
  - 系数向量 $\alpha_i^\top$ 是单位向量，$\alpha_i^\top\alpha_i=1,\ i=1,2,\cdots,m$
  - y_i 与 y_j 互不相关，$Cov(y_i,y_j)=0,\ i\neq j$
  - 第一主成分 y_1：是 $\boldsymbol{x}$ 所有线性组合中方差最大的，即

$$y=arg\max_y Var(y)=arg\max_{\alpha_1}\alpha_1^\top\Sigma\alpha_1$$
$$s.t.\ \ \alpha_1^\top\alpha_1=1$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

总体主成分

- 第二主成分 y_2：是与 y_1 不相关的 $\boldsymbol{x}$ 所有线性组合中方差最大的，即

$$y = arg\max_y Var(y) = arg\max_{\alpha_2} \alpha_2^\top \Sigma \alpha_2$$
$$s.t.\ \ \alpha_1^\top \Sigma \alpha_2 = 0,\ \alpha_2^\top \alpha_2 = 1$$

- 第 i 主成分 y_i：是与 $y_1, y_2, \cdots, y_{i-1}$ 都不相关的 $\boldsymbol{x}$ 所有线性组合中方差最大的，即

$$y = arg\max_y Var(y) = arg\max_{\alpha_i} \alpha_i^\top \Sigma \alpha_i$$
$$s.t.\ \ \alpha_1^\top \Sigma \alpha_i = 0, \cdots,\ \alpha_{i-1}^\top \Sigma \alpha_i = 0,\ \alpha_i^\top \alpha_i = 1$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

主要性质

- 设 Σ 的特征值分别是 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$，特征值对应的单位特征向量分别是 $\alpha_1, \alpha_2, \cdots, \alpha_m$，则 $\boldsymbol{x}$ 第 k 主成分是

$$y_k = \alpha_k^\top \boldsymbol{x}$$

$\boldsymbol{x}$ 的第 k 主成分的方差是

$$Var(y_k) = \alpha_k^\top \Sigma \alpha_k = \lambda_k$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

主要性质

- 令 $A=[\alpha_1,\alpha_2,\cdots,\alpha_m]$，则 A 为正交矩阵，且 $\boldsymbol{y}=(y_1,y_2,\cdots,y_m)^\top=A^\top\boldsymbol{x}$
- $Cov(\boldsymbol{y})=\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_m)$
- 总体主成分 $\boldsymbol{y}$ 的方差之和等于随机变量 $\boldsymbol{x}$ 的方差之和 $\sum_{i=1}^{m}\lambda_i=\sum_{i=1}^{m}\sigma_{ii}$
- 因子负荷量 $\rho(y_k,x_i)$ 表示主成分 y_k 与变量 x_i 之间的相关关系

$$\rho(y_k,x_i)=\frac{Cov(y_k,x_i)}{\sqrt{Var(y_k)Var(x_i)}}=\frac{\sqrt{\lambda_k}\,\alpha_{ik}}{\sqrt{\sigma_{ii}}},\ k,i=1,2,\cdots,m$$

  - 一般的，小于0.4是低，大于0.6是高

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

主要性质

- 主成分 y_k 与 m 个变量的因子负荷量满足

$$\sum_{k=1}^{m}\sigma_{ii}\rho^2(y_k,x_i)=\sum_{k=1}^{m}\lambda_{ik}\alpha_{ik}^2=\lambda_k\alpha_k^\top\alpha_k=\lambda_k$$

- m 个主成分与变量 x_i 的因子负荷量满足

$$\sum_{k=1}^{m}\rho^2(y_k,x_i)=\sum_{k=1}^{m}\frac{\lambda_{ik}\alpha_{ik}^2}{\sigma_{ii}}=1$$

  - $\sigma_{ii}=[\Sigma]_{ii}=[A\Lambda A^\top]_{ii}=\sum_{k=1}^{m}\lambda_k\alpha_{ik}^2$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

主成分的个数

- **定理16.2**：对任意 $q,\ 1 \leq q \leq m$，考虑正交线性变化 $\boldsymbol{y} = B^{\top}\boldsymbol{x}$，其中 $B^{\top}$ 是 $q \times m$ 矩阵，令 $\boldsymbol{y}$ 的协方差矩阵为 $\Sigma_{\boldsymbol{y}} = B^{\top}\Sigma B$，则 $trace(\Sigma_{\boldsymbol{y}})$ 在 $B = A_q$ 时取得最大值，其中矩阵 A_q 由正交矩阵 A 的前 q 列组成
- 定理结论表明选择前 q 个主成分能保留原有变量的最多信息
  - 方差越大保留的信息越多
- **定理16.3**：$trace(\Sigma_{\boldsymbol{y}})$ 在 $B = A_p$ 时取得最小值，其中矩阵 A_p 由正交矩阵 A 的后 p 列组成

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

方差贡献率

- 主成分 y_k 的方差贡献率是 y_k 的方差与所有方差之和的比

$$\eta_k = \frac{\lambda_k}{\sum_{i=1}^{m} \lambda_i}$$

- 前 k 个主成分的累计方差贡献率是前 k 个方差之和与所有方差之和的比

$$\sum_{i=1}^{k} \eta_i = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i}$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

规范化变量的总体主成分

- 对变量 x_i 进行规范化 $x_i^* = \frac{x_i - \mathbb{E}(x_i)}{\sqrt{Var(x_i)}}$（均值为0，方差为1）
- 规范化随机变量的协方差矩阵就是相关矩阵
- 规范化变量主成分的协方差矩阵 $\Lambda^* = \mathrm{diag}(\lambda_1^*, \lambda_2^*, \cdots, \lambda_m^*)$
  - $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_m^* \geq 0$，且 λ_i^* 是相关矩阵 R 的特征值
- 协方差矩阵的特征值之和为 m

$$\sum_{k=1}^{m} \lambda_k^* = \sum_{i=1}^{m} \sigma_{ii}^* = m$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

规范化变量的总体主成分

- 规范化变量与主成分的因子负荷量是

$$\rho(y_k^*, x_i^*) = \frac{Cov(y_k^*, x_i^*)}{\sqrt{Var(y_k^*)Var(x_i^*)}} = \frac{\sqrt{\lambda_k^*}e_{ik}^*}{\sqrt{\sigma_{ii}^*}} = \sqrt{\lambda_k^*}\, e_{ik}^*,\ k,i = 1,2,\cdots,m$$

  - $e_k^* = (e_{1k}^*, e_{2k}^*, \cdots, e_{mk}^*)^\top$是相关矩阵对应于特征值$\lambda_k^*$ 的特征向量

- 规范化主成分与m个规范化变量的因子负荷量满足

$$\sum_{i=1}^{m} \rho(y_k^*, x_i^*) = \sum_{i=1}^{m} \lambda_k^* {e_{ik}^*}^2 = \lambda_k^* \| e_k^* \|^2 = \lambda_k^*,\ k = 1,2,\cdots,m$$

- 规范化变量与规范化主成分的因子负荷量满足$\sum_{k=1}^{m} \rho^2(y_k^*, x_i^*) = 1$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 基本思想
- 总体主成分分析
- 样本主成分分析
- 应用：基因组分析

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

样本定义

- 对随机向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_m)^\top$ 进行 n 次独立观测
  - 得到观测样本 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$，其中 $\boldsymbol{x}_j=(x_{1j},x_{2j},\cdots,x_{mj})^\top$
- 样本矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]$
- 样本矩阵向量 $\bar{x}=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_m)^\top=\dfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$
- 样本协方差矩阵 $S=[s_{ij}]_{m\times m}$，其中 $s_{ij}=\dfrac{1}{n-1}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_i)$
- 样本相关矩阵 $R=[r_{ij}]_{m\times m}$，其中 $r_{ij}=\dfrac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

样本主成分的定义

- 令 m 维向量之间的线性变换为 $\boldsymbol{y}=A^{\top}\boldsymbol{x}$，其中 $A=[a_1\ a_2\ \cdots\ a_m]=[a_{ij}]_{m\times m}$
- 样本 $\boldsymbol{x}_i$ 在变换下为 $\boldsymbol{y}_i=A^{\top}\boldsymbol{x}_i$，变换后的均值为

$$\bar{\boldsymbol{y}}=(\bar{y}_1,\bar{y}_2,\cdots,\bar{y}_m)^{\top}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i=\frac{1}{n}\sum_{i=1}^{n}A^{\top}\boldsymbol{x}_i$$

  - 其中 $\bar{y}_j=\frac{1}{n}\sum_{i=1}^{n}a_j^{\top}\boldsymbol{x}_i=a_j^{\top}\bar{\boldsymbol{x}}$，$\bar{\boldsymbol{x}}$ 为样本均值
- 变换后样本的第 i 个分量 $Y_i=(y_{i1},y_{i2},\cdots,y_{in}),\ i=1,2,\cdots,m$ 的方差为

$$Var(Y_i)=\frac{1}{n-1}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2=\frac{1}{n-1}\sum_{j=1}^{n}(a_i^{\top}\boldsymbol{x}_j-a_i^{\top}\bar{\boldsymbol{x}})^2=a_i^{\top}Sa_i$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

样本主成分的定义

- 变换后两个分量的协方差为

$$
\begin{aligned}
Var(Y_i, Y_k) &= \frac{1}{n-1}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)\left(y_{kj}-\bar{y}_k\right) \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left(a_i^{\top}\boldsymbol{x}_j - a_i^{\top}\bar{\boldsymbol{x}}\right)\left(a_k^{\top}\boldsymbol{x}_j - a_k^{\top}\bar{\boldsymbol{x}}\right) \\
&= a_i^{\top}\left[\frac{1}{n-1}\left(\boldsymbol{x}_j-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_j-\bar{\boldsymbol{x}}\right)^{\top}\right]a_k \\
&= a_i^{\top} S a_k
\end{aligned}
$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

样本主成分

- 给定样本矩阵 $\boldsymbol{X}$
- 样本第一主成分 $y_1 = a_1^\top \boldsymbol{x}$ 是在 $a_1^\top a_1 = 1$ 下使得 $Var(Y_1) = a_1^\top S a_1$ 最大的线性组合
- 样本第二主成分 $y_2 = a_2^\top \boldsymbol{x}$ 是在 $a_2^\top a_2 = 1,\ Var(Y_2, Y_1) = a_1^\top S a_2 = 0$ 下使得 $Var(Y_2)$ 最大的线性组合
- 样本第 i 主成分 $y_i = a_i^\top \boldsymbol{x}$ 是在 $a_i^\top a_i = 1,\ Var(Y_i, Y_j) = a_i^\top S a_j = 0,\ j = 1, 2, \cdots, i-1$ 下使得 $Var(Y_i)$ 最大的线性组合

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

样本主成分与总体主成分

- 总体主成分：在数据总体上进行的主成分分析
- 样本主成分：在有限样本上进行的主成分分析
  - 用样本协方差矩阵 S 替代总体协方差矩阵 Σ
  - 用样本均值 $\bar{\boldsymbol{x}}$ 替代总体均值 $\boldsymbol{\mu}$
- 但结论相同，主成分都是协方差矩阵的特征值所对应的特征向量所构成的线性组合

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

规范后的样本主成分

- 样本数据规范化：$x_{ij}^{*}=\dfrac{x_{ij}-\bar{x}_i}{\sqrt{s_{ii}}}$
  - 其中 $\bar{x}_i=\dfrac{1}{n}\sum_{j=1}^{n}x_{ij},\ s_{ii}=\dfrac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2,\ i=1,2,\cdots,m$
  - 规范化后，样本第 i 个分量 $X_i^{*}=(x_{i1}^{*},x_{i2}^{*},\cdots,x_{in}^{*})$ 的均值为0，方差为1
- 令规范化后的样本矩阵为 $X^{*}=[x_{ij}^{*}]_{m\times m}$，则规范化后的样本协方差矩阵就是样本相关矩阵 $R=\dfrac{1}{n-1}X^{*}(X^{*})^{\top}$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

相关矩阵的特征值分解

- 样本主成分分析是通过求相关矩阵的特征值实现，步骤如下：

1. 将样本进行规范化，得到规范化后的样本矩阵$X \in \mathbb{R}^{m \times n}$
2. 计算相关矩阵 $R=[r_{ij}]_{m \times m}=\frac{1}{n-1} XX^{\top}$，其中 $r_{ij}=\frac{1}{n-1} \sum_{l=1}^{n} x_{il} x_{lj}$
3. 求相关矩阵 R 的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ 和对应的特征向量
4. 求方差贡献率达到预定值的主成分个数 k
5. 求出前 k 个样本主成分 $y_i=a_i^{\top} \boldsymbol{x},\ i=1,2,\cdots,k$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

样本矩阵的奇异值分解

- 传统的主成分分析通过样本的相关矩阵的特征值分解进行，现在常用的方法是通过样本矩阵的奇异值分解进行，步骤如下

1. 构造矩阵 $X' = \frac{1}{\sqrt{n-1}} X^\top \in \mathbb{R}^{n \times m}$
2. 对 X' 进行 k 阶截断奇异值分解，得到 $X' = U \Sigma V^\top$
3. 求前 k 个主成分的矩阵 $Y = V^\top X$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

奇异值分解的优点

- 计算样本的相关矩阵 $R=\frac{1}{n-1}XX^{\top}$ 及其特征值需要消耗大量的计算量，特别是当样本量很大时
- 注意到PCA仅仅使用了SVD的右奇异矩阵，没有使用的左奇异矩阵
- 有些SVD的实现算法可以不用求出 R 也能计算出右奇异矩阵 V，从而极大的节约了计算成本

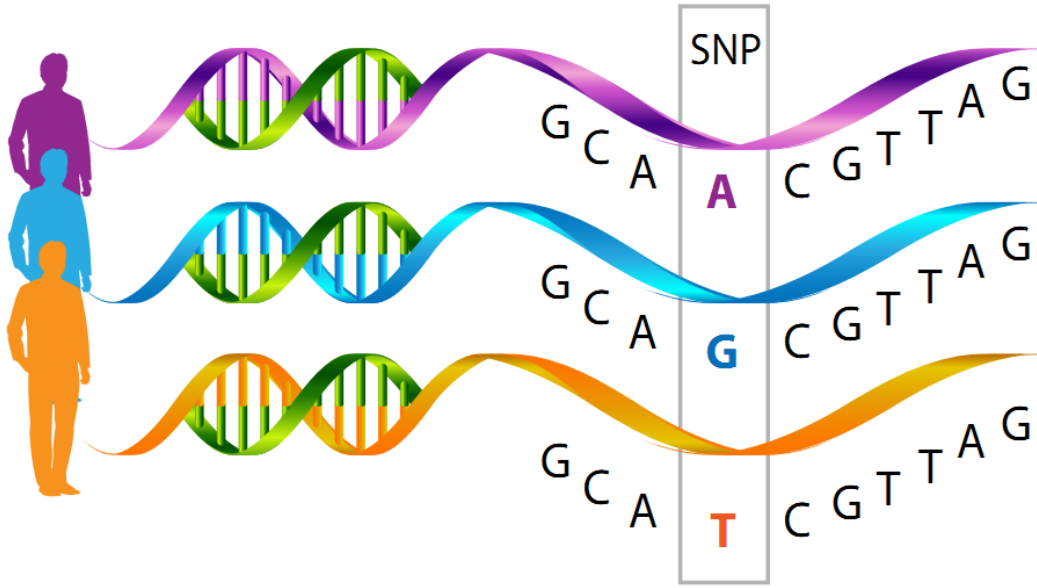
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 基本思想
- 总体主成分分析
- 样本主成分分析
- 应用：基因组分析

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

人类基因组的SNP数据

- 1064个志愿者，并把他们的SNP数据数字化，也就是把每个位置上可能出现的10种碱基对用数字来代表
- SNP （单核苷酸多态性）是人类可遗传的变异中最常见的一种，估计其总数可达300万个甚至更多广泛用于基于群体的基因识别等方面的研究



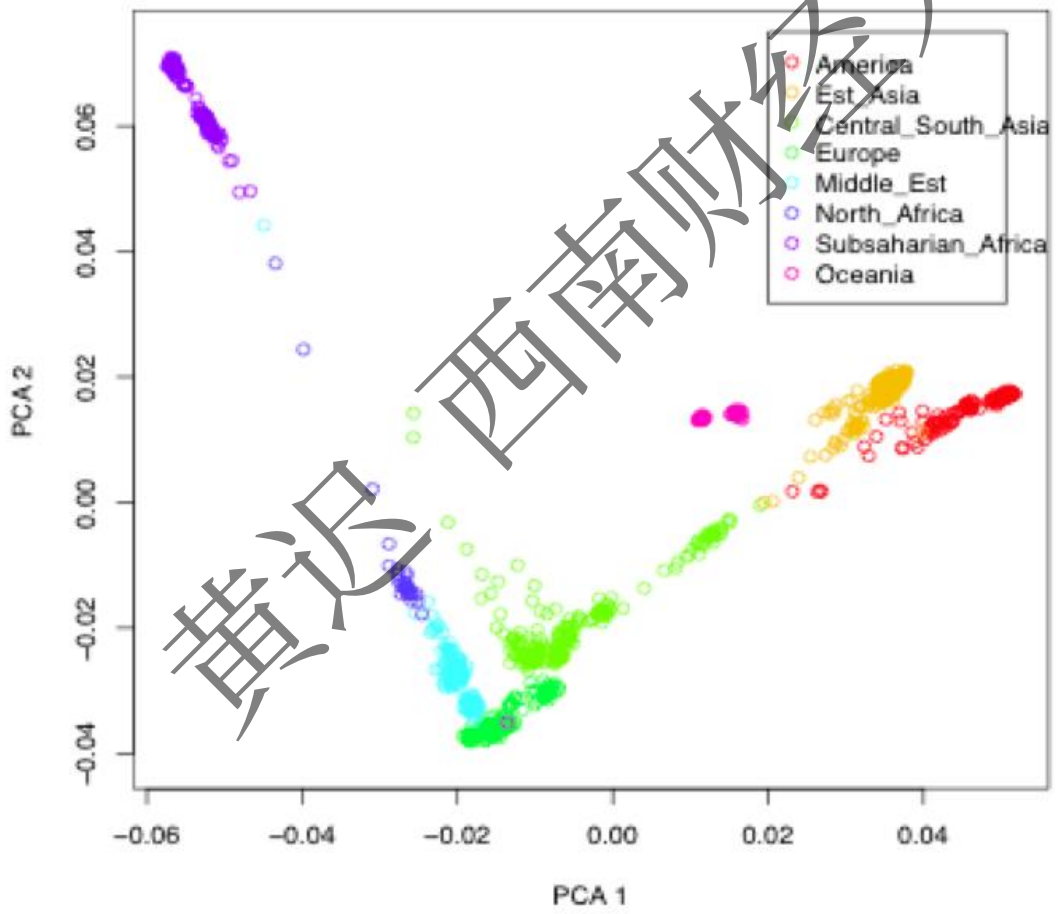
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

基因数据

	SNP 1	SNP 2	...	SNP m
志愿者1	0	1	...	4
志愿者2	1	0	...	2
志愿者3	0	2	...	1
...	...	...	...	...
志愿者n	0	9	...	0

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

主成分分析的结果



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

主成分分析的结果