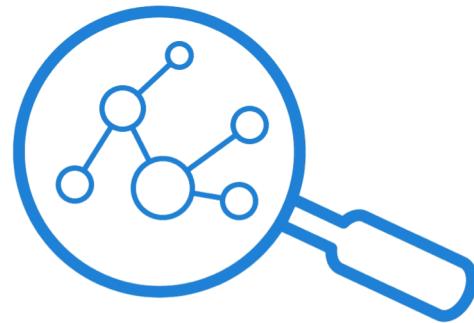


# 数据科学与大数据技术 的数学基础

## 第七讲

计算机学院  
余皓然  
2023/5/16



# 课程内容

---

## Part1 随机化方法

一致性哈希 布隆过滤器 CM Sketch方法 最小哈希  
欧氏距离下的相似搜索 Jaccard相似度下的相似搜索

## Part2 谱分析方法

主成分分析 奇异值分解 谱图论

## Part3 最优化方法

压缩感知

# 随机化方法

---

## ➤ 问题与随机化方法回顾

分布式缓存	一致性哈希
从属判断	布隆过滤器
高频元素寻找	CM Sketch方法
不同元素统计	最小哈希
欧氏距离下的相似搜索	JR转换
Jaccard相似度下的搜索	局部敏感哈希

## ➤ 理解要点 (takeaways)

- 当数据规模巨大时，看似简单的问题可能变得复杂，线性复杂度不可接受
- 随机化方法以较小的空间/时间复杂度近似解决问题，以精确度换空间/时间
- 可以用马尔可夫不等式、切比雪夫不等式等工具刻画方法达到性能的上下界

# 课程内容

---

## Part1 随机化方法

一致性哈希 布隆过滤器 CM Sketch方法 最小哈希

欧氏距离下的相似搜索 Jaccard相似度下的相似搜索

## Part2 谱分析方法

主成分分析 奇异值分解 谱图论

## Part3 最优化方法

压缩感知

# 主成分分析

## 主成分分析的背景



# 背景

---

- 与JL转换类似，主成分分析（Principal Component Analysis）也是一种数据降维的方法
  - JL转换：在降维过程中保护数据间的欧几里得距离
  - 主成分分析：增强降维后数据的可解释性

"principal component analysis"

About 2,730,000 results (0.11 sec)

**[PDF] Principal component analysis**  
H Abdi, LJ Williams - Wiley interdisciplinary reviews ..., 2010 - Wiley Online Library  
Principal component analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent ...  
☆ Save ⚡ Cite Cited by 8306 Related articles All 10 versions

**Principal component analysis**  
S Wold, K Esbensen, P Geladi - Chemometrics and intelligent laboratory ..., 1987 - Elsevier  
Principal component analysis of a data matrix extracts the dominant patterns in the matrix in terms of a complementary set of score and loading plots. It is the responsibility of the data ...  
☆ Save ⚡ Cite Cited by 10809 Related articles All 14 versions

主成分分析是一种主要的数据降维方式

# 背景

---

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2

可将每行记为一个四维向量，能否对这四个四维向量降维并做数据可视化？可否更好地解释四个数据之间的区别？



# 背景

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2

可将每行记为一个四维向量，能否对这四个四维向量降维并做数据可视化？可否更好地解释四个数据之间的区别？

令  $\bar{x} = (5.5, 4.5, 5, 5.5)$ ,  $v_1 = (3, -3, -3, 3)$ ,  $v_2 = (1, -1, 1, -1)$ ，可以将任一行对应的向量近似成：

$$\bar{x} + a_1 v_1 + a_2 v_2$$

其中，张三的数据对应  $(a_1, a_2) = (1, 1)$ ，即  $\bar{x} + v_1 + v_2 = (9.5, 0.5, 3, 7.5)$

李四的数据对应  $(a_1, a_2) = (1, -1)$ ，即  $\bar{x} + v_1 - v_2 = (7.5, 2.5, 1, 9.5)$

王五的数据对应  $(a_1, a_2) = (-1, -1)$ ，赵六的数据对应  $(a_1, a_2) = (-1, 1)$

# 背景

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2

可将每行记为一个四维向量，能否对这四个四维向量降维并做数据可视化？可否更好地解释四个数据之间的区别？

令  $\bar{x} = (5.5, 4.5, 5, 5.5)$ ,  $v_1 = (3, -3, -3, 3)$ ,  $v_2 = (1, -1, 1, -1)$ ，可以将任一行对应的向量近似成：

$$\bar{x} + a_1 v_1 + a_2 v_2$$

其中，张三的数据对应  $(a_1, a_2) = (1, 1)$ ，即  $\bar{x} + v_1 + v_2 = (9.5, 0.5, 3, 7.5)$

李四的数据对应  $(a_1, a_2) = (1, -1)$ ，即  $\bar{x} + v_1 - v_2 = (7.5, 2.5, 1, 9.5)$

王五的数据对应  $(a_1, a_2) = (-1, -1)$ ，赵六的数据对应  $(a_1, a_2) = (-1, 1)$

可用二维向量  $(a_1, a_2)$  近似表示四个数据

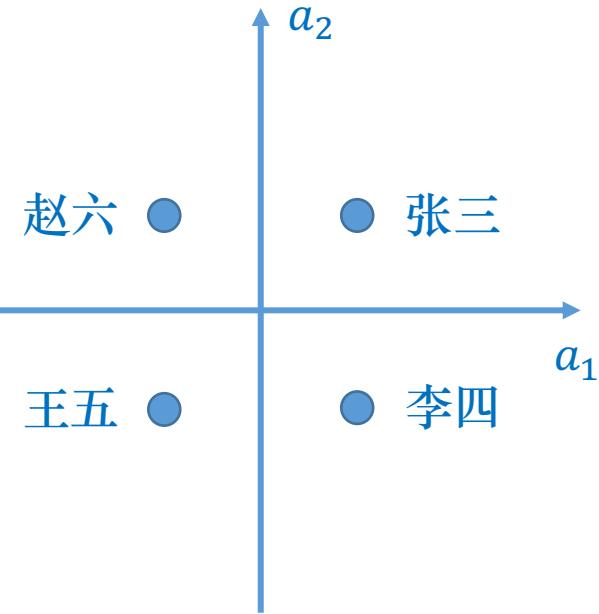
## 背景

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2

四个四维数据



四个二维数据



令  $\bar{x} = (5.5, 4.5, 5, 5.5)$ ,  $v_1 = (3, -3, -3, 3)$ ,  $v_2 = (1, -1, 1, -1)$ , 可以将任一行对应的向量近似成：

$$\bar{x} + a_1 v_1 + a_2 v_2$$

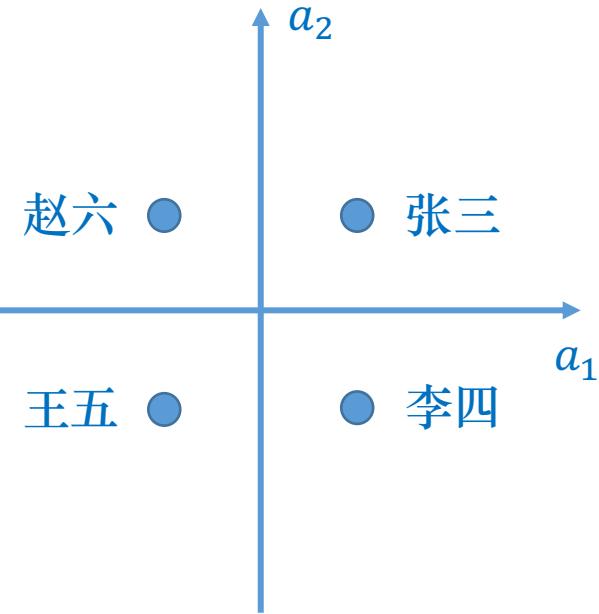
## 背景

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2

四个四维数据



四个二维数据



令  $\bar{x} = (5.5, 4.5, 5, 5.5)$ ,  $v_1 = (3, -3, -3, 3)$ ,  $v_2 = (1, -1, 1, -1)$ , 可以将任一行对应的向量近似成：

$$\bar{x} + a_1 v_1 + a_2 v_2$$

降维的作用：

一、方便对数据进行可视化（若有更多人的数据，可以看到有哪些类、哪些人之间更相似）

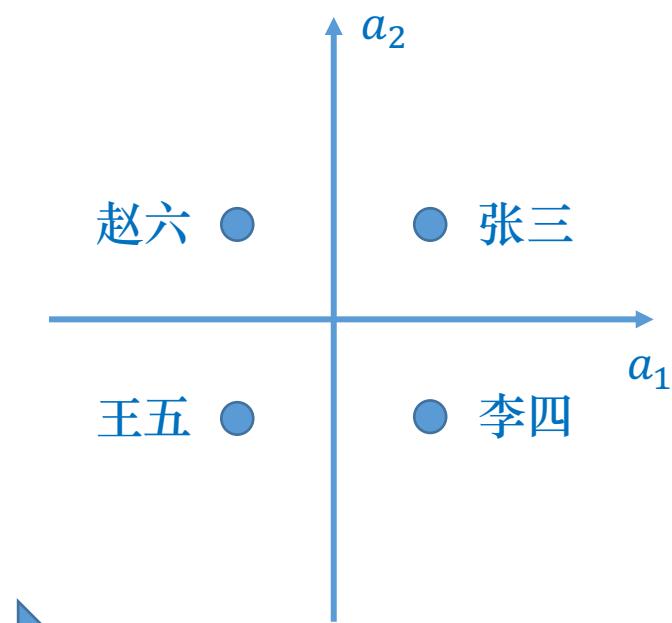
## 背景

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2

四个四维数据



四个二维数据



令  $\bar{x} = (5.5, 4.5, 5, 5.5)$ ,  $v_1 = (3, -3, -3, 3)$ ,  $v_2 = (1, -1, 1, -1)$ , 可以将任一行对应的向量近似成：

$$\bar{x} + a_1 v_1 + a_2 v_2$$

降维的作用：

- 一、方便对数据进行可视化（若有更多人的数据，可以看到有哪些类、哪些人之间更相似）
- 二、通过解释  $v_1, v_2$ , 解释数据（如  $v_1$  对应是否喜欢吃素、 $v_2$  对应是否注重健康，而这些人在是否吃素方面的差异比在是否注重健康方面的差异更大）

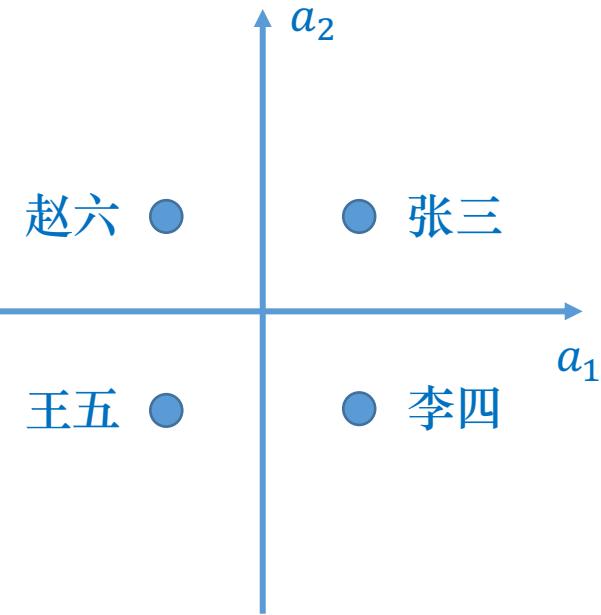
## 背景

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2

四个四维数据



四个二维数据



令  $\bar{x} = (5.5, 4.5, 5, 5.5)$ ,  $v_1 = (3, -3, -3, 3)$ ,  $v_2 = (1, -1, 1, -1)$ , 可以将任一行对应的向量近似成：

$$\bar{x} + a_1 v_1 + a_2 v_2$$

(标准化后的)  $v_1$  和  $v_2$  可称为数据的“主成分”，主成分分析关注如何由数据找到  $\bar{x} + a_1 v_1 + a_2 v_2$

## 背景

---

主成分分析：给定  $m$  个  $n$  维向量  $x_1, \dots, x_m \in \mathbb{R}^n$ ，分别将它们近似为  $k$  个  $n$  维向量  $v_1, \dots, v_k \in \mathbb{R}^n$  的线性组合，即  $x_i \approx \sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



## 背景

---

主成分分析：给定  $m$  个  $n$  维向量  $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ ，分别将它们近似为  $k$  个  $n$  维向量  $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$  的线性组合，即  $\mathbf{x}_i \approx \sum_{j=1}^k a_{ij} \mathbf{v}_j, i = 1, \dots, m$ 。

在前例中，是将  $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$  近似为  $\bar{\mathbf{x}} + a_{i1} \mathbf{v}_1 + a_{i2} \mathbf{v}_2 (i = 1, \dots, m)$  的形式

令  $\bar{\mathbf{x}} = (5.5, 4.5, 5, 5.5)$ ,  $\mathbf{v}_1 = (3, -3, -3, 3)$ ,  $\mathbf{v}_2 = (1, -1, 1, -1)$ ，可以将任一行对应的向量近似成：

$$\bar{\mathbf{x}} + a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2$$

由于  $\bar{\mathbf{x}} = \frac{1}{4}(\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4)$ ，可等价于将  $\mathbf{x}_1 - \bar{\mathbf{x}}, \mathbf{x}_2 - \bar{\mathbf{x}}, \mathbf{x}_3 - \bar{\mathbf{x}}, \mathbf{x}_4 - \bar{\mathbf{x}}$  近似为  $a_{i1} \mathbf{v}_1 + a_{i2} \mathbf{v}_2$  形式

## 背景

---

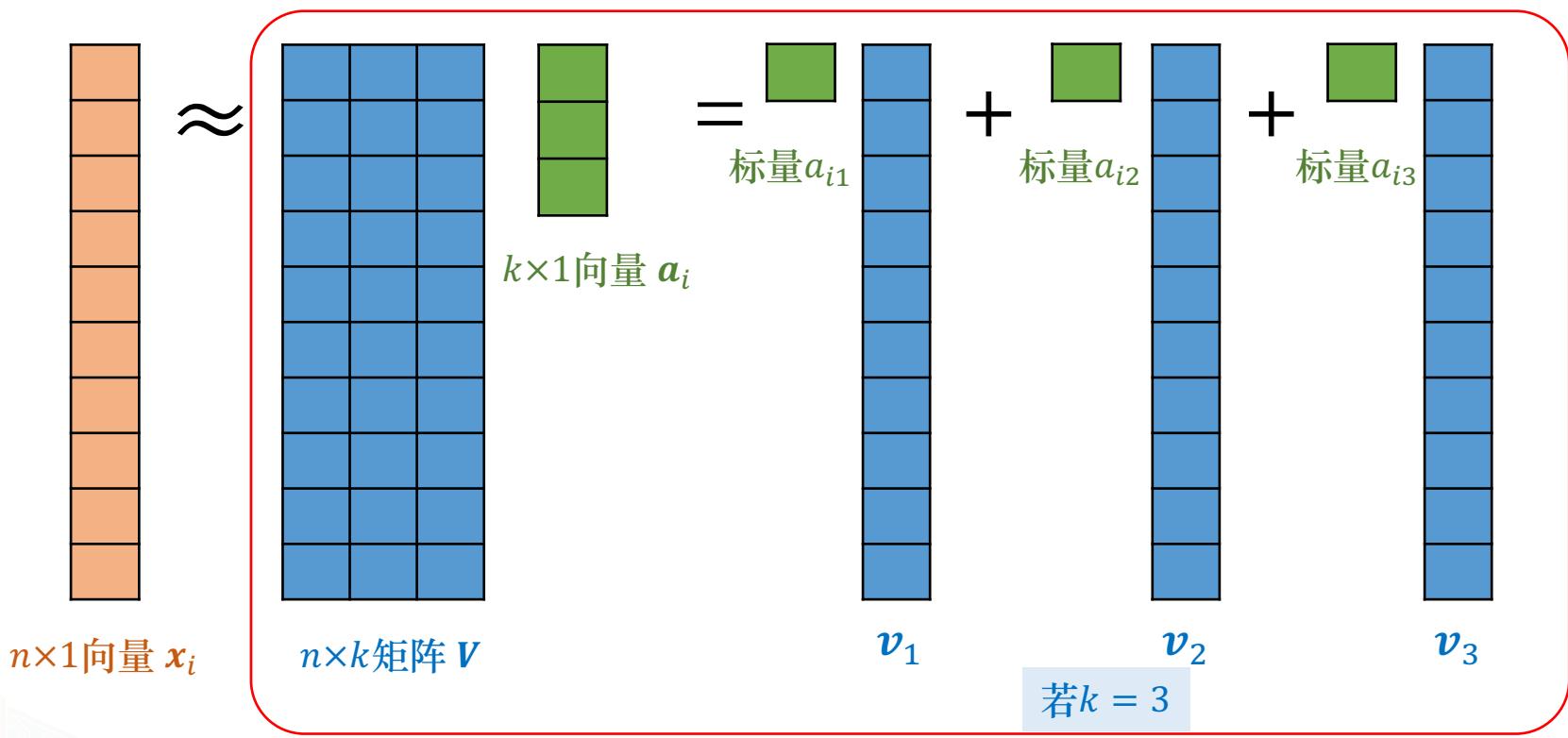
主成分分析：给定  $m$  个  $n$  维向量  $x_1, \dots, x_m \in \mathbb{R}^n$ ，分别将它们近似为  $k$  个  $n$  维向量  $v_1, \dots, v_k \in \mathbb{R}^n$  的线性组合，即  $x_i \approx \sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。

如何把列向量  $x_i$  写成矩阵与向量的乘积形式？



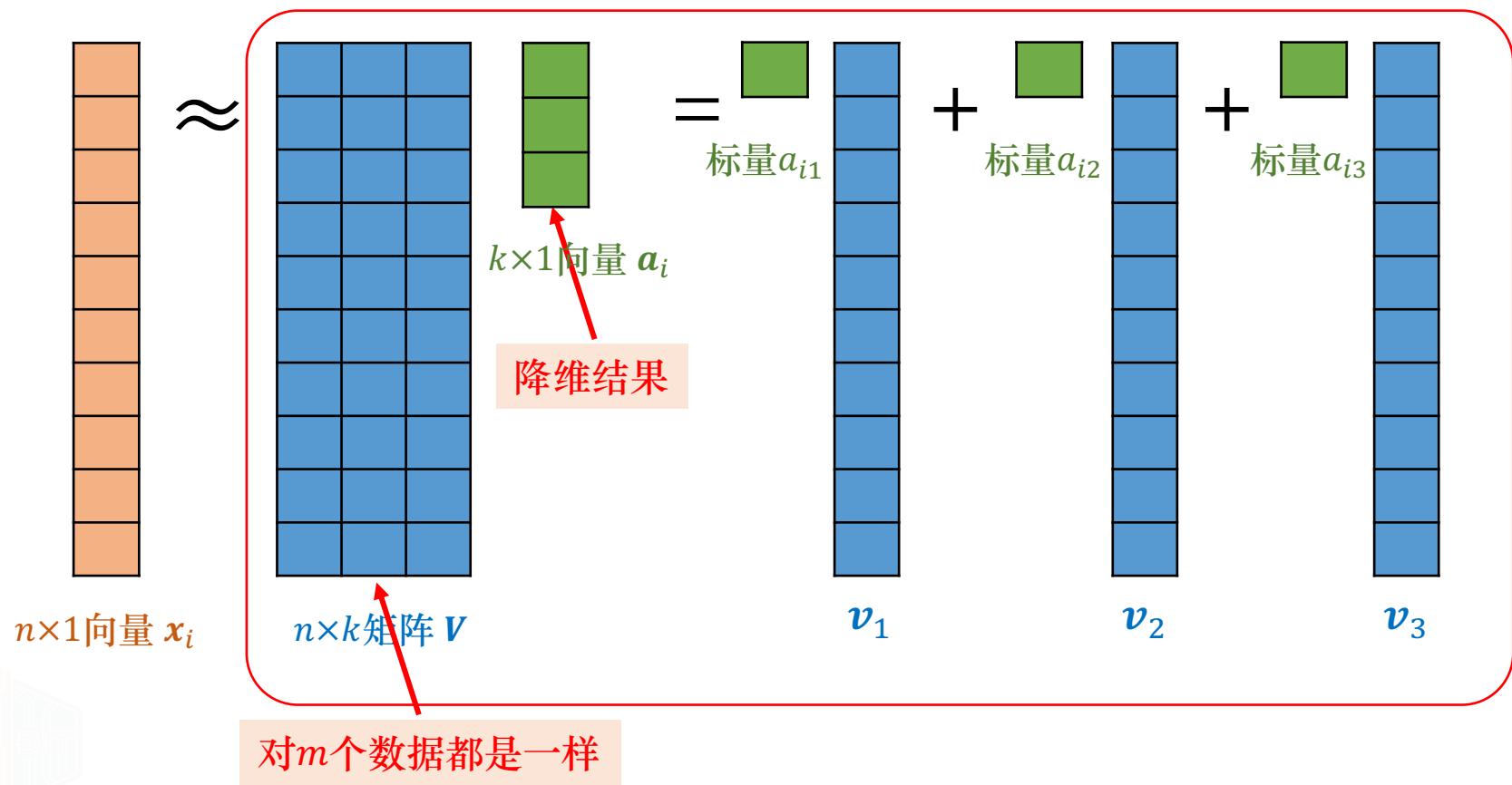
# 背景

主成分分析：给定  $m$  个  $n$  维向量  $x_1, \dots, x_m \in \mathbb{R}^n$ ，分别将它们近似为  $k$  个  $n$  维向量  $v_1, \dots, v_k \in \mathbb{R}^n$  的线性组合，即  $x_i \approx \sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



## 背景

主成分分析：给定  $m$  个  $n$  维向量  $x_1, \dots, x_m \in \mathbb{R}^n$ ，分别将它们近似为  $k$  个  $n$  维向量  $v_1, \dots, v_k \in \mathbb{R}^n$  的线性组合，即  $x_i \approx \sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。

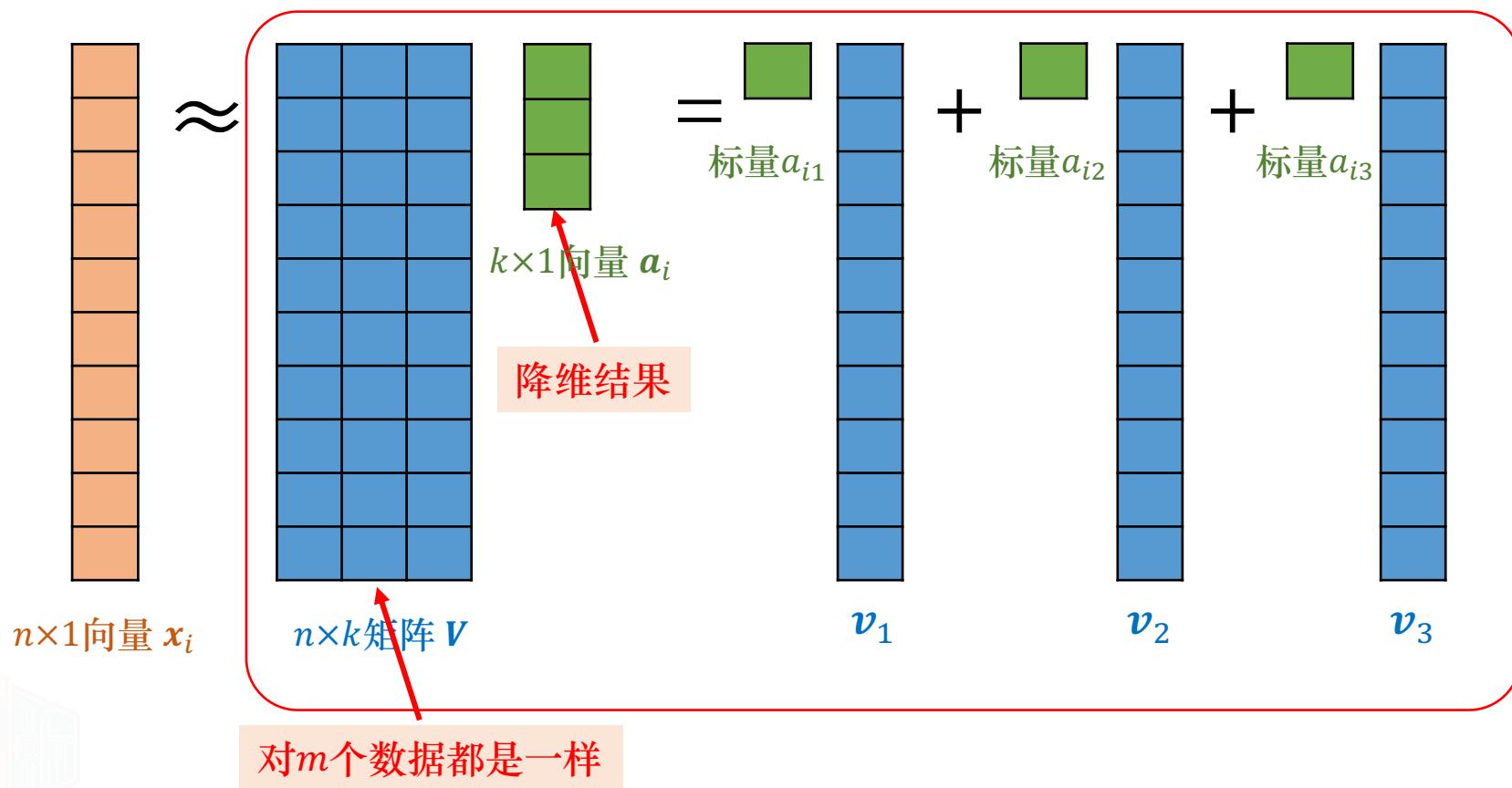


## 背景

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2

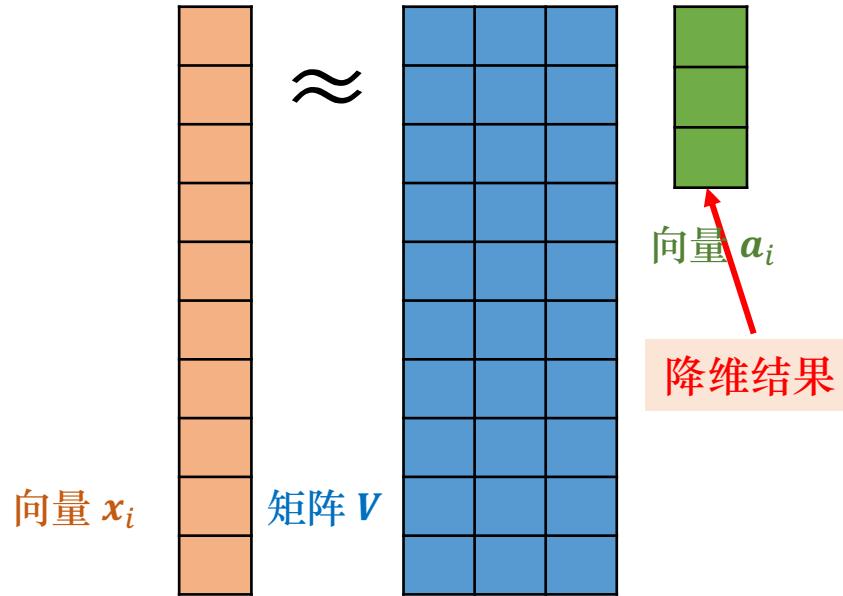
此例中,  $m = 4, n = 4, k = 2$

主成分分析: 给定  $m$  个  $n$  维向量  $x_1, \dots, x_m \in \mathbb{R}^n$ , 分别将它们近似为  $k$  个  $n$  维向量  $v_1, \dots, v_k \in \mathbb{R}^n$  的线性组合, 即  $x_i \approx \sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。

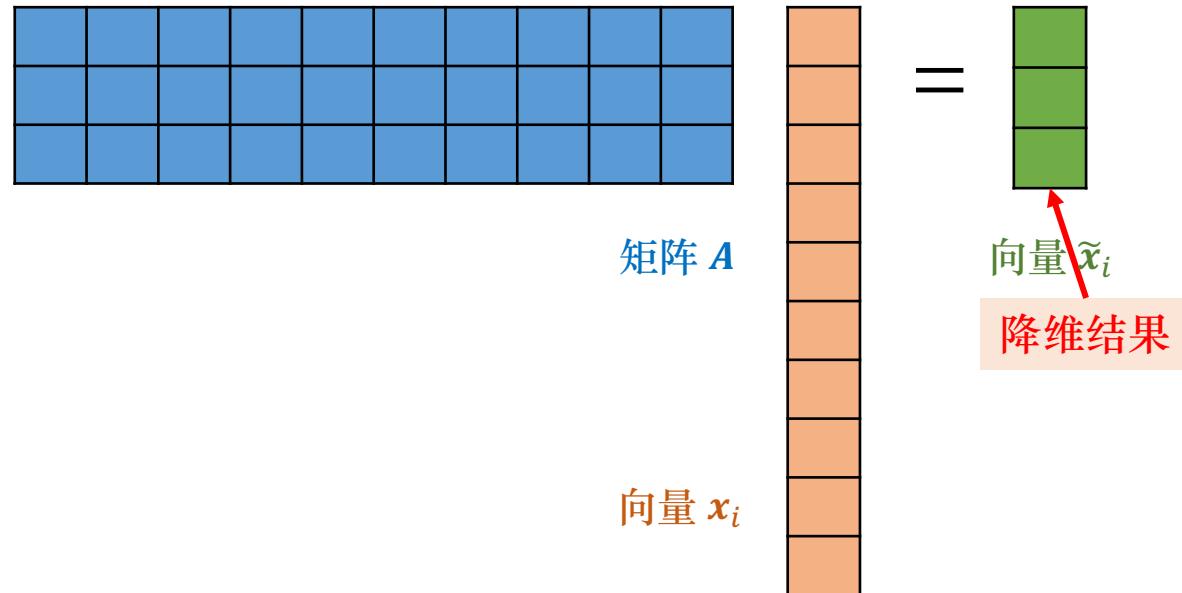


# PCA与JL转换对比

主成分分析 (PCA)



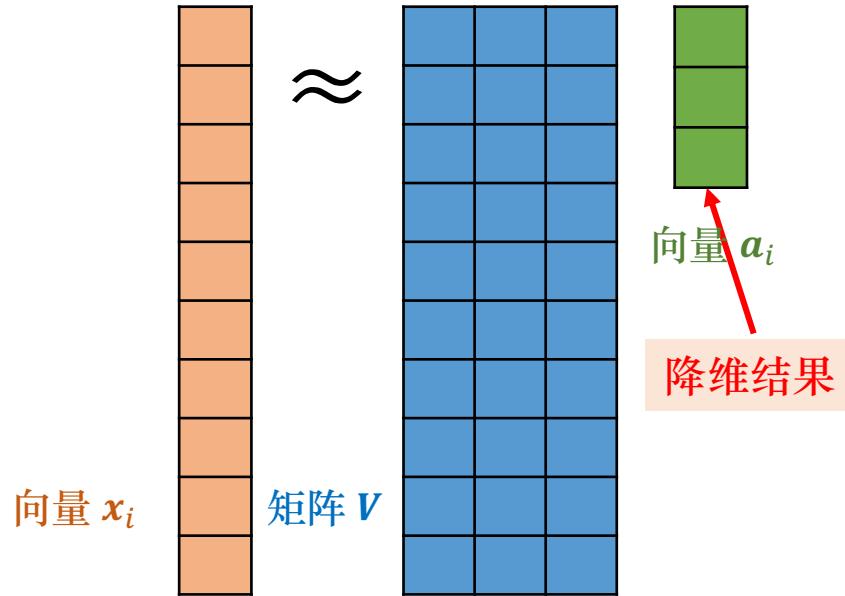
JL转换



# PCA与JL转换对比

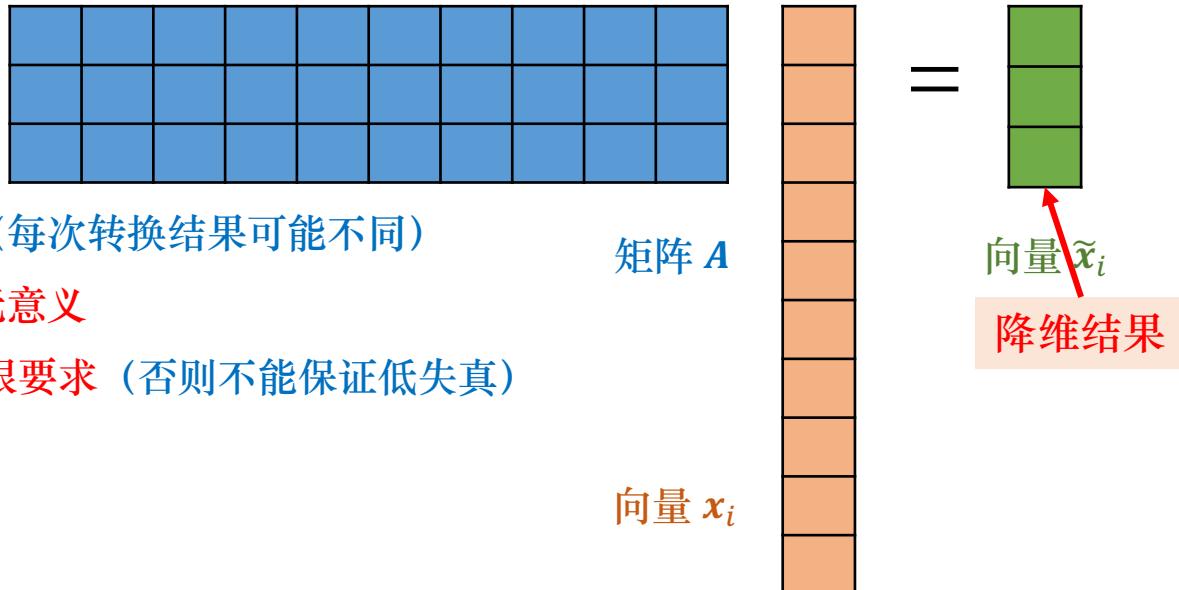
## 主成分分析 (PCA)

1. 降维后可能无法保持原数据间的欧氏距离
2. 矩阵  $V$  与数据有关, 根据数据得到  $V$  从而解释数据
3. 矩阵  $V$  的  $k$  个列向量对应的坐标一般**有意义**
4. 可能降至  $k = 1$  或  $k = 2$  依然得到**有意义**的结果



## JL转换

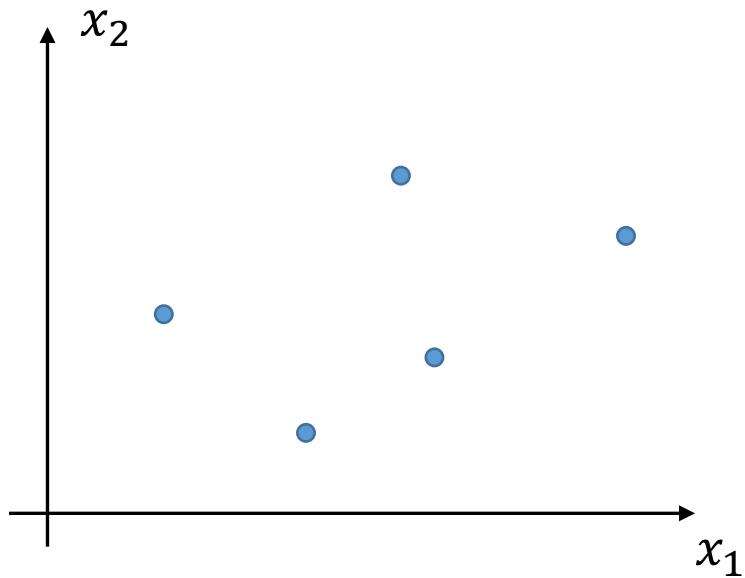
1. 降维后**保持**原数据间的欧氏距离
2. 矩阵  $A$  与数据无关, 是**随机生成** (每次转换结果可能不同)
3. 矩阵  $A$  的行向量对应的坐标一般**无意义**
4. 根据JL引理, 降维后的维度**有下限要求** (否则不能保证低失真)



# PCA与线性回归对比

## 主成分分析（PCA）

若要把若干个二维数据降维成一维数据



## 线性回归

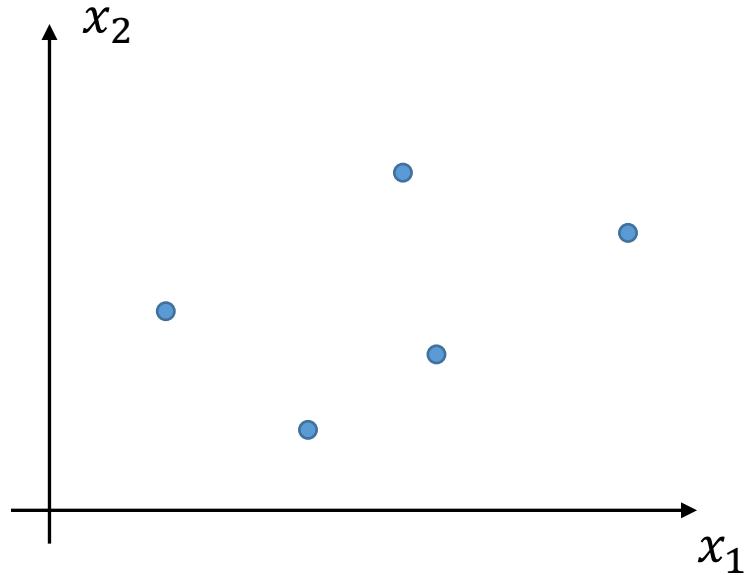


# PCA与线性回归对比

## 主成分分析 (PCA)

若要把若干个二维数据降维成一维数据

即  $k = 1$ , 有  $x_i \approx \sum_{j=1}^k a_{ij} v_j = a_i v$



## 线性回归

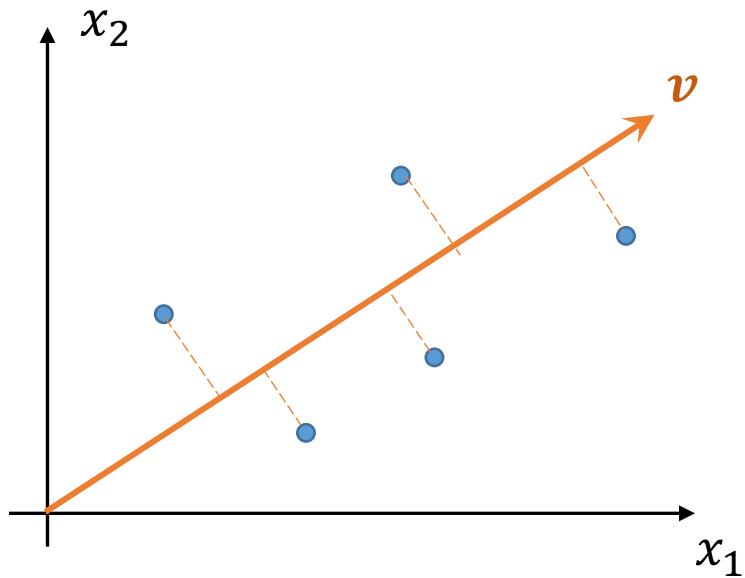


# PCA与线性回归对比

## 主成分分析 (PCA)

若要把若干个二维数据降维成一维数据

即  $k = 1$ , 有  $x_i \approx \sum_{j=1}^k a_{ij} v_j = a_i v$



## 线性回归

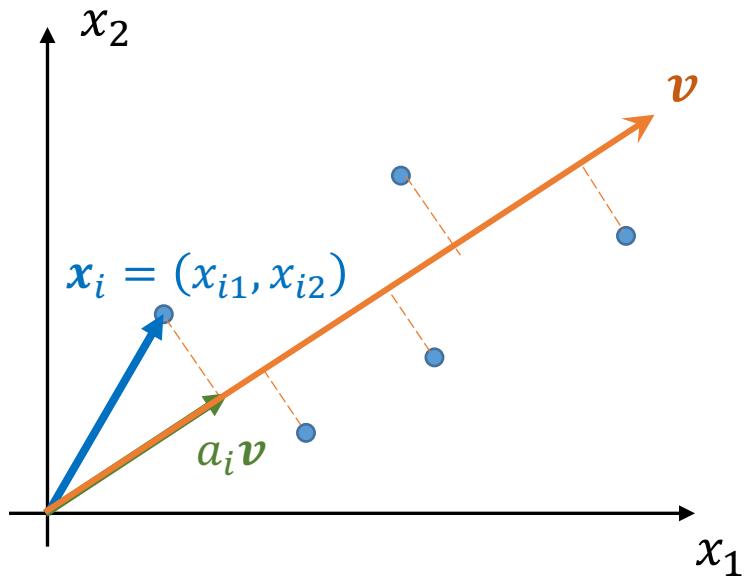


# PCA与线性回归对比

## 主成分分析 (PCA)

若要把若干个二维数据降维成一维数据

即  $k = 1$ , 有  $x_i \approx \sum_{j=1}^k a_{ij} v_j = a_i v$



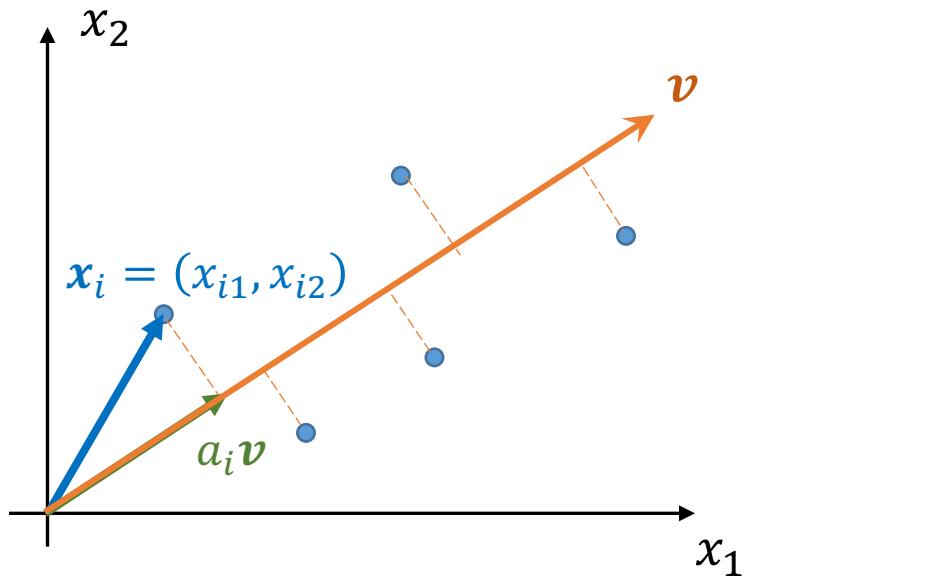
## 线性回归

# PCA与线性回归对比

## 主成分分析 (PCA)

若要把若干个二维数据降维成一维数据

即  $k = 1$ , 有  $x_i \approx \sum_{j=1}^k a_{ij} v_j = a_i v$



向量  $v$  需要满足：最小化各个点（原数据）到  $v$  的垂直距离的平方和

不希望  $x_i$  与  $a_i v$  欧式距离过大

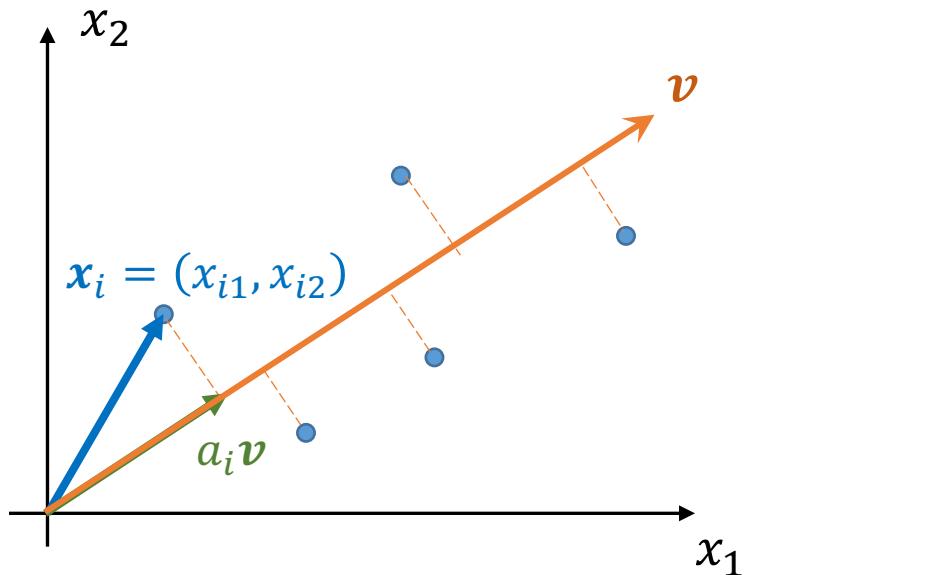
## 线性回归

# PCA与线性回归对比

## 主成分分析 (PCA)

若要把若干个二维数据降维成一维数据

即  $k = 1$ , 有  $x_i \approx \sum_{j=1}^k a_{ij} v_j = a_i v$



向量  $v$  需要满足：最小化各个点（原数据）到  $v$  的垂直距离的平方和

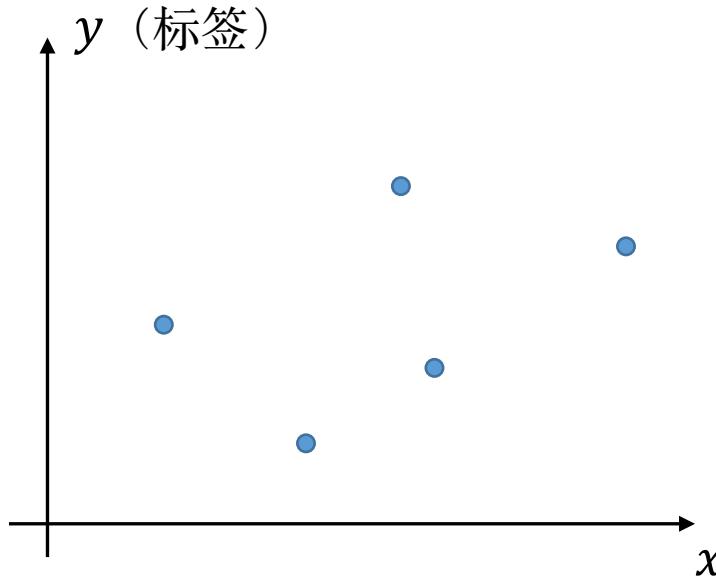
不希望  $x_i$  与  $a_i v$  欧式距离过大

## 线性回归 (属于机器学习中的监督学习)

若采集到五个商品房的面积  $x$  与售价  $y$  的数据

希望用直线  $y = \theta x$  近似未知函数  $f: x \rightarrow y$

使得  $\theta x_i$  接近于  $y_i$

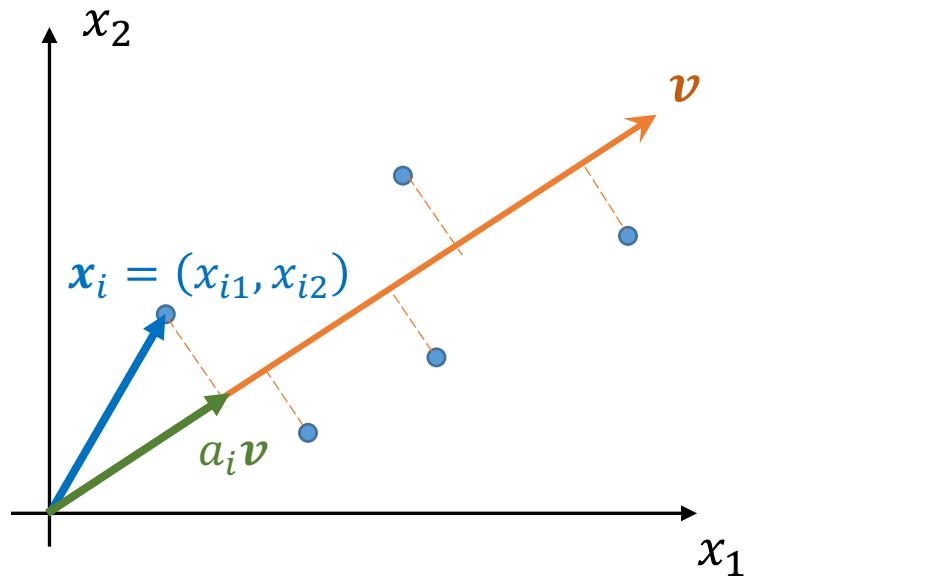


# PCA与线性回归对比

## 主成分分析 (PCA)

若要把若干个二维数据降维成一维数据

即  $k = 1$ , 有  $x_i \approx \sum_{j=1}^k a_{ij} v_j = a_i v$



向量  $v$  需要满足：最小化各个点（原数据）到  $v$  的垂直距离的平方和

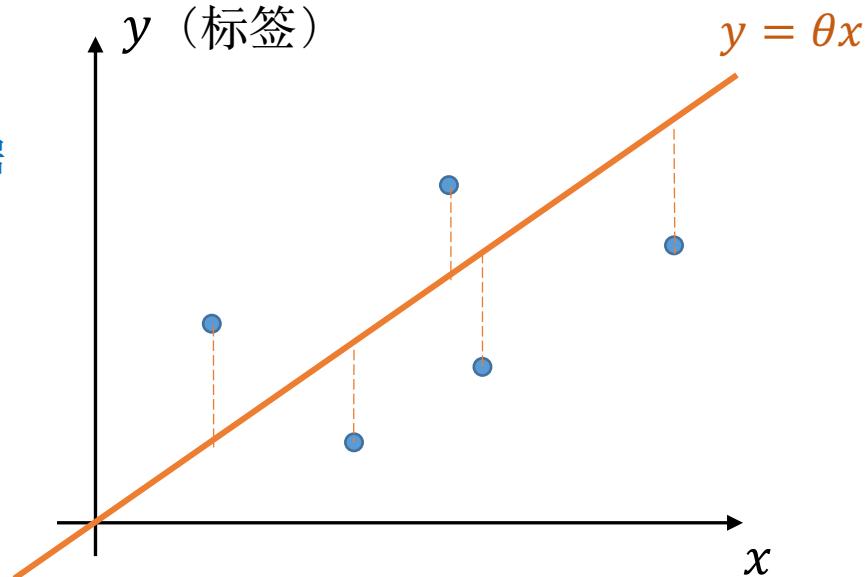
不希望  $x_i$  与  $a_i v$  欧式距离过大

## 线性回归 (属于机器学习中的监督学习)

若采集到五个商品房的面积  $x$  与售价  $y$  的数据

希望用直线  $y = \theta x$  近似未知函数  $f: x \rightarrow y$

使得  $\theta x_i$  接近于  $y_i$

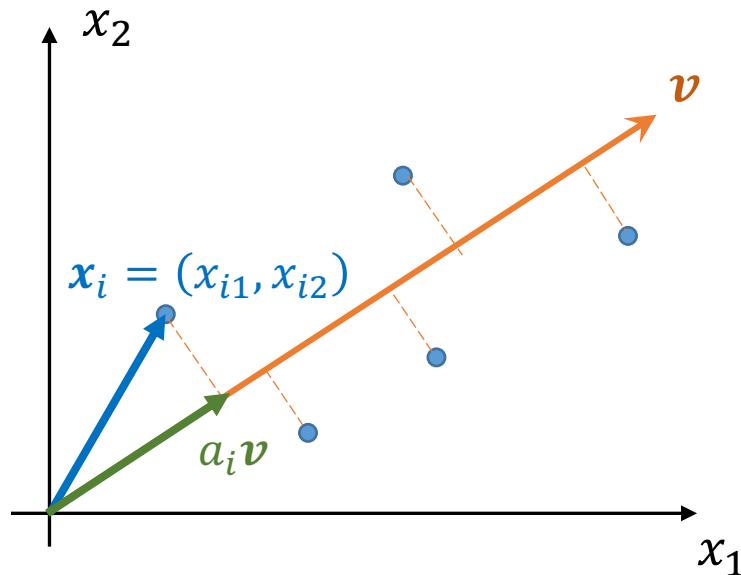


# PCA与线性回归对比

## 主成分分析 (PCA)

若要把若干个二维数据降维成一维数据

即  $k = 1$ , 有  $x_i \approx \sum_{j=1}^k a_{ij} v_j = a_i v$



向量  $v$  需要满足：最小化各个点（原数据）到  $v$  的垂直距离的平方和

不希望  $x_i$  与  $a_i v$  欧式距离过大

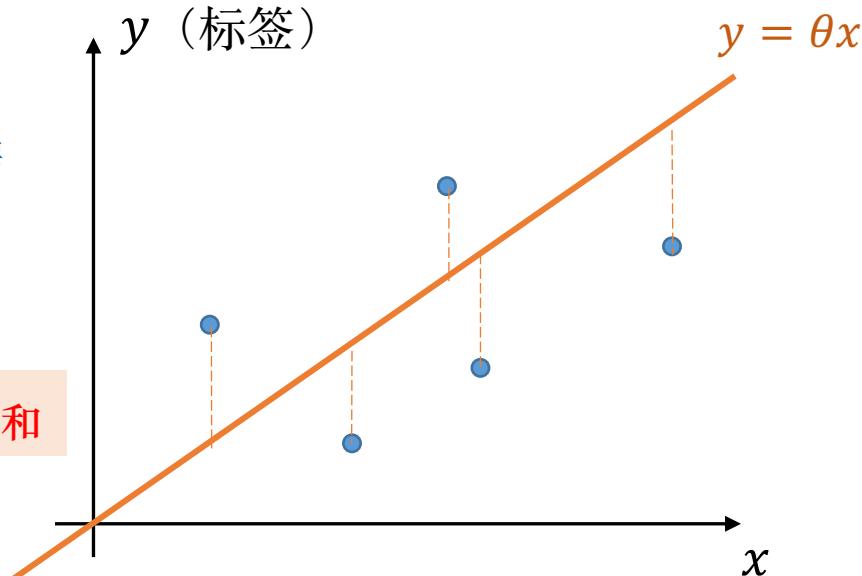
## 线性回归 (属于机器学习中的监督学习)

若采集到五个商品房的面积  $x$  与售价  $y$  的数据

希望用直线  $y = \theta x$  近似未知函数  $f: x \rightarrow y$

使得  $\theta x_i$  接近于  $y_i$

斜率  $\theta$  需要满足：最小化  $\theta x_i$  到  $y_i$  的差的平方和

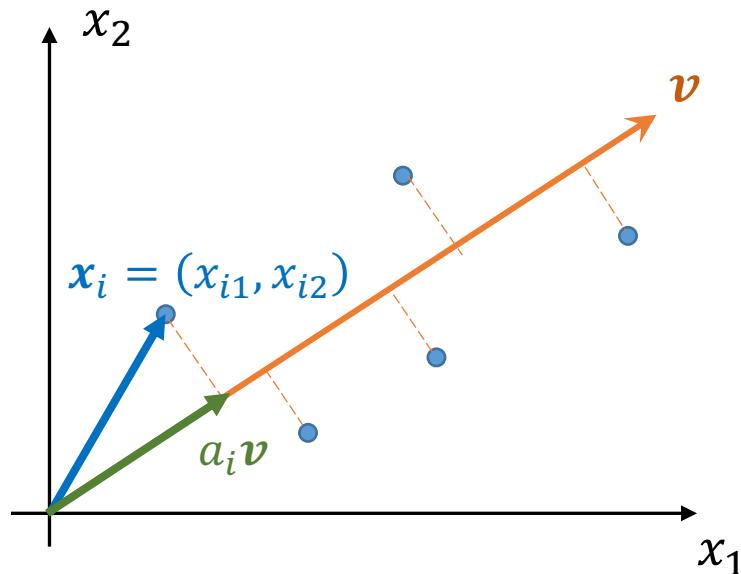


# PCA与线性回归对比

## 主成分分析 (PCA)

若要把若干个二维数据降维成一维数据

即  $k = 1$ , 有  $x_i \approx \sum_{j=1}^k a_{ij} v_j = a_i v$



向量  $v$  需要满足：最小化各个点（原数据）到  $v$  的垂直距离的平方和

不希望  $x_i$  与  $a_i v$  欧式距离过大

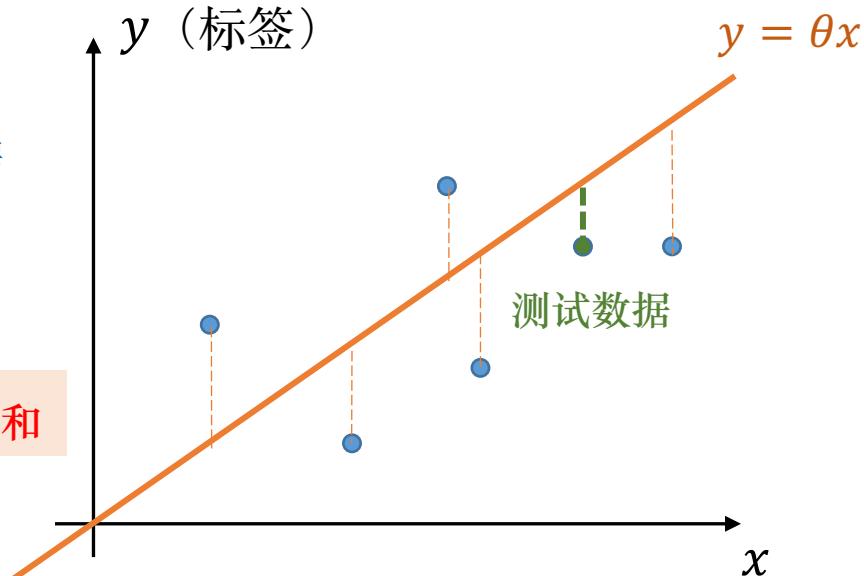
## 线性回归 (属于机器学习中的监督学习)

若采集到五个商品房的面积  $x$  与售价  $y$  的数据

希望用直线  $y = \theta x$  近似未知函数  $f: x \rightarrow y$

使得  $\theta x_i$  接近于  $y_i$

斜率  $\theta$  需要满足：最小化  $\theta x_i$  到  $y_i$  的差的平方和



# 主成分分析

## 主成分分析问题定义



# 数据预处理

---

## 一、将所有数据“平移”，令数据的中点为原点

回顾前例，令  $\bar{x} = (5.5, 4.5, 5, 5.5)$ ,  $v_1 = (3, -3, -3, 3)$ ,  $v_2 = (1, -1, 1, -1)$ ,

可以将  $x_1, x_2, x_3, x_4$  近似为  $\bar{x} + a_{i1}v_1 + a_{i2}v_2$  的形式，

由于  $\bar{x} = \frac{1}{4}(x_1 + x_2 + x_3 + x_4)$ , 等价于将  $x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, x_4 - \bar{x}$  近似为  $a_{i1}v_1 + a_{i2}v_2$  形式



# 数据预处理

## 一、将所有数据“平移”，令数据的中点为原点

回顾前例，令  $\bar{x} = (5.5, 4.5, 5, 5.5)$ ,  $v_1 = (3, -3, -3, 3)$ ,  $v_2 = (1, -1, 1, -1)$ ,

可以将  $x_1, x_2, x_3, x_4$  近似为  $\bar{x} + a_{i1}v_1 + a_{i2}v_2$  的形式，

由于  $\bar{x} = \frac{1}{4}(x_1 + x_2 + x_3 + x_4)$ , 等价于将  $x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, x_4 - \bar{x}$  近似为  $a_{i1}v_1 + a_{i2}v_2$  形式

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2



	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	4.5	-3.5	-3	1.5
李四	1.5	-2.5	-4	4.5
王五	-3.5	4.5	2	-2.5
赵六	-2.5	1.5	5	-3.5

每行数据减去  $\bar{x} = (5.5, 4.5, 5, 5.5)$

每个维度下所有数据之和为0

# 数据预处理

---

## 一、将所有数据“平移”，令数据的中点为原点

计算  $\bar{x} = \frac{1}{m}(x_1 + \dots + x_m)$ ，为每个数据  $x_i$  计算  $\tilde{x}_i = x_i - \bar{x}$ ，然后对  $\tilde{x}_1, \dots, \tilde{x}_m$  做降维  
即有  $\tilde{x}_i \approx \sum_{j=1}^k a_{ij} v_j$  以及  $x_i \approx \bar{x} + \sum_{j=1}^k a_{ij} v_j$



# 数据预处理

一、将所有数据“平移”，令数据的中点为原点

二、对数据的各维度缩放，减小各维度对应的数字量级不同带来的影响

	身高	体重	年龄
张三	170	65	32
李四	180	70	25
王五	185	80	28
赵六	165	55	35



张三	-5	-2.5	2
李四	5	2.5	-5
王五	10	12.5	-2
赵六	-10	-12.5	5

	身高	体重	年龄
张三	1.7	65	32
李四	1.8	70	25
王五	1.85	80	28
赵六	1.65	55	35



张三	-0.05	-2.5	2
李四	0.05	2.5	-5
王五	0.1	12.5	-2
赵六	-0.1	-12.5	5

# 数据预处理

一、将所有数据“平移”，令数据的中点为原点

二、对数据的各维度缩放，减小各维度对应的数字量级不同带来的影响

	身高	体重	年龄
张三	170	65	32
李四	180	70	25
王五	185	80	28
赵六	165	55	35



张三	-5	-2.5	2
李四	5	2.5	-5
王五	10	12.5	-2
赵六	-10	-12.5	5

	身高	体重	年龄
张三	1.7	65	32
李四	1.8	70	25
王五	1.85	80	28
赵六	1.65	55	35



张三	-0.05	-2.5	2
李四	0.05	2.5	-5
王五	0.1	12.5	-2
赵六	-0.1	-12.5	5

PCA将最小化数据 $\tilde{x}_i$ 到 $\sum_{j=1}^k a_{ij} v_j$ 的距离的平方和，若某维度的数字量级大、对距离的计算影响大

# 数据预处理

一、将所有数据“平移”，令数据的中点为原点

二、对数据的各维度缩放，减小各维度对应的数字量级不同带来的影响

把数据  $\tilde{x}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{in})$  缩放为  $\left( \frac{1}{\sqrt{\sum_{i=1}^m \tilde{x}_{i1}^2}} \tilde{x}_{i1}, \dots, \frac{1}{\sqrt{\sum_{i=1}^m \tilde{x}_{in}^2}} \tilde{x}_{in} \right)$

张三	-5	-2.5	2
李四	5	2.5	-5
王五	10	12.5	-2
赵六	-10	-12.5	5



张三	-0.05	-2.5	2
李四	0.05	2.5	-5
王五	0.1	12.5	-2
赵六	-0.1	-12.5	5



# 数据预处理

一、将所有数据“平移”，令数据的中点为原点

二、对数据的各维度缩放，减小各维度对应的数字量级不同带来的影响

把数据  $\tilde{x}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{in})$  缩放为  $\left( \frac{1}{\sqrt{\sum_{i=1}^m \tilde{x}_{i1}^2}} \tilde{x}_{i1}, \dots, \frac{1}{\sqrt{\sum_{i=1}^m \tilde{x}_{in}^2}} \tilde{x}_{in} \right)$

张三	-5	-2.5	2
李四	5	2.5	-5
王五	10	12.5	-2
赵六	-10	-12.5	5



张三	-0.316	-0.139	0.263
李四	0.316	0.139	-0.657
王五	0.632	0.693	-0.263
赵六	-0.632	-0.693	0.657

张三	-0.05	-2.5	2
李四	0.05	2.5	-5
王五	0.1	12.5	-2
赵六	-0.1	-12.5	5

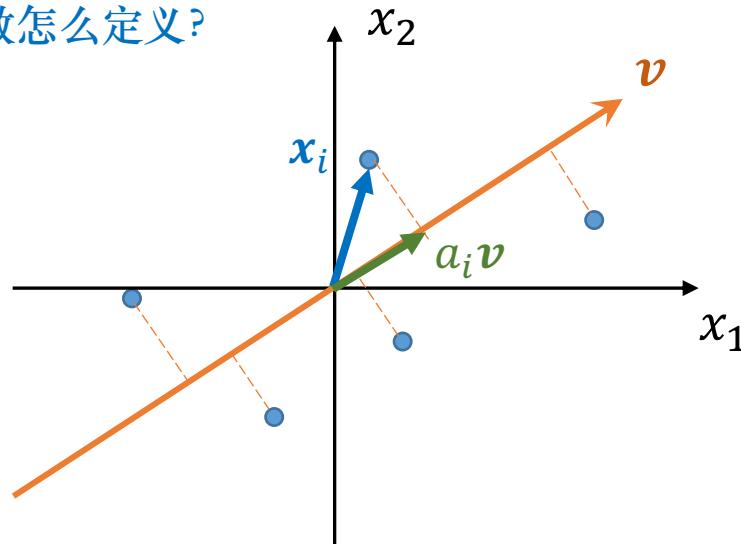


张三	-0.316	-0.139	0.263
李四	0.316	0.139	-0.657
王五	0.632	0.693	-0.263
赵六	-0.632	-0.693	0.657

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

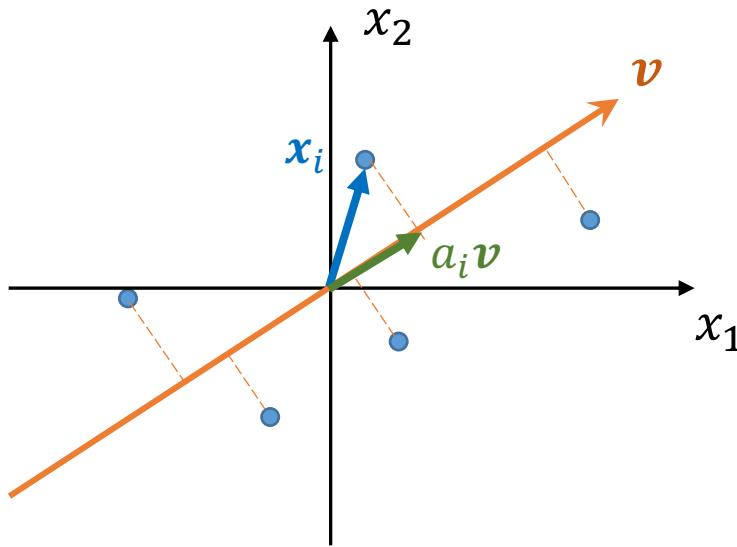
如何选择向量 $v$ ? 目标函数怎么定义?



# $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



注意是距离平方和

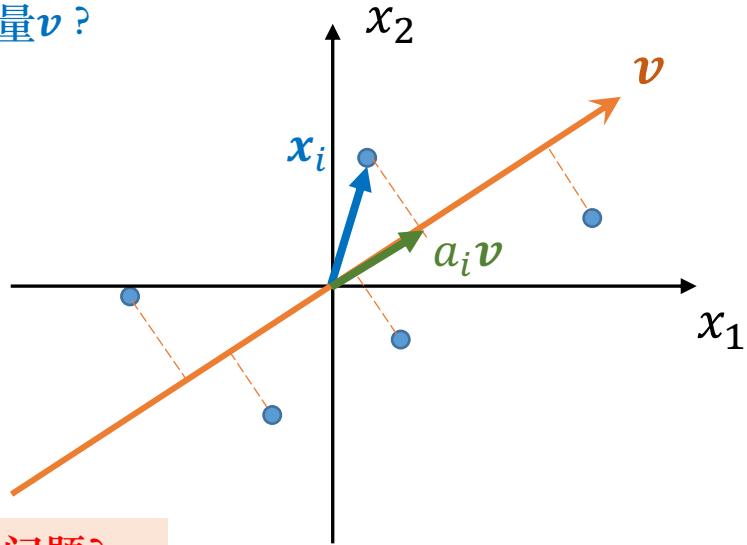
$$\operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2)$$

单位向量

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



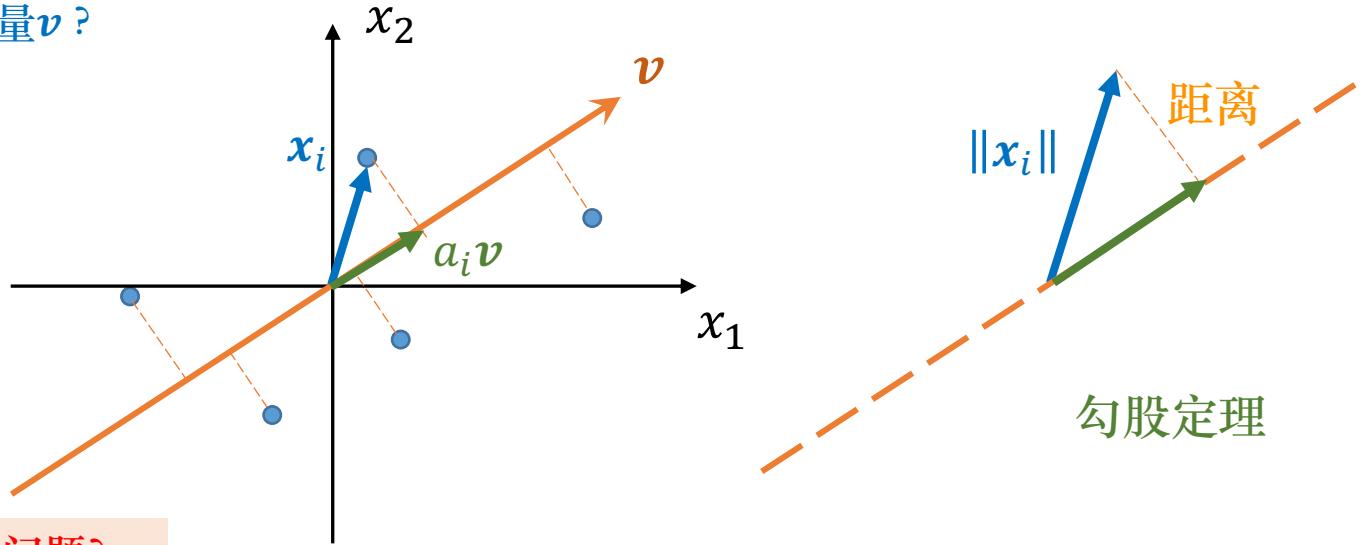
可否进一步改写问题？

$$\operatorname{argmin}_{v: \|v\|=1} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } x_i \text{ and line spanned by } v)^2)$$

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



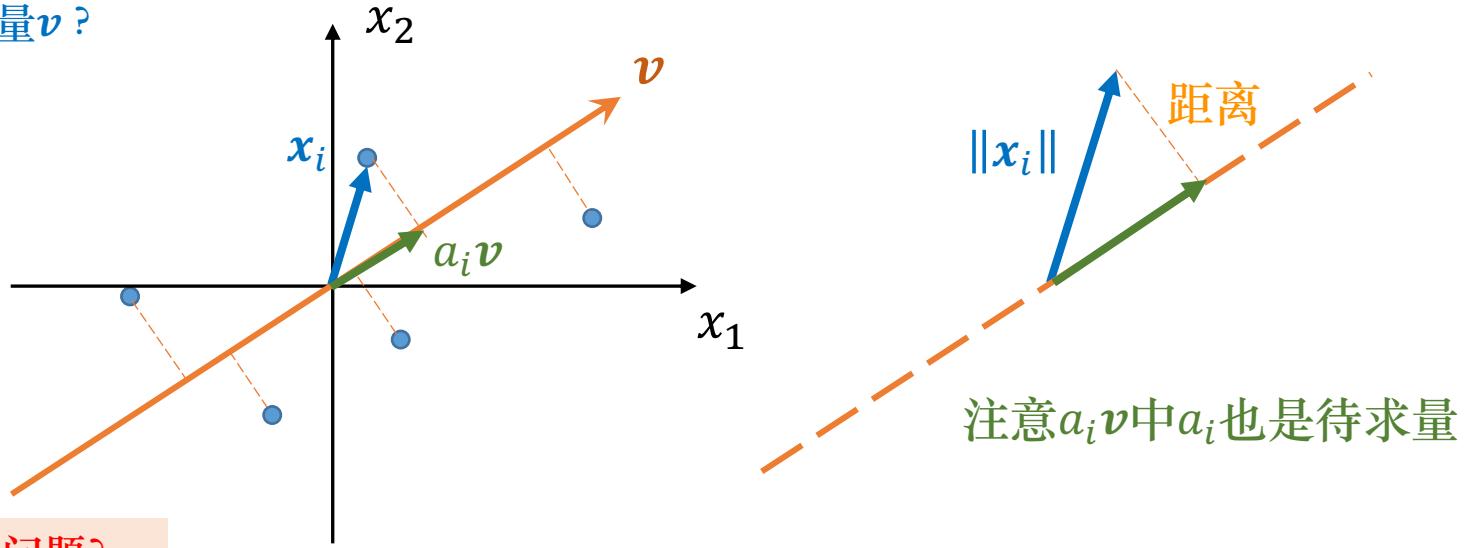
可否进一步改写问题？

$$\operatorname{argmin}_{v: \|v\|=1} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } x_i \text{ and line spanned by } v)^2)$$

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ 近似为 $a_i \mathbf{v}, i = 1, \dots, m$ 。

如何选择向量 $\mathbf{v}$ ？



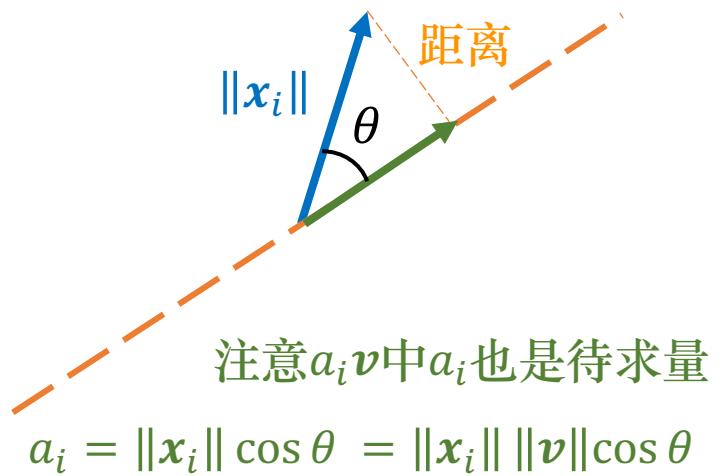
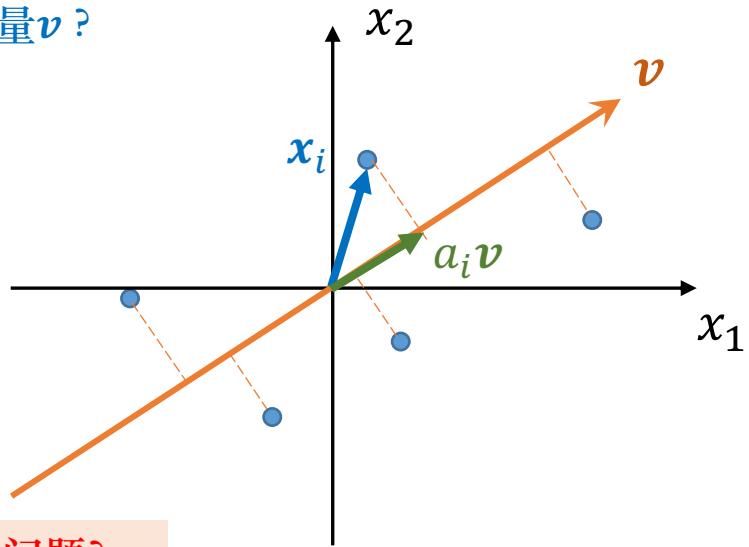
可否进一步改写问题？

$$\operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2)$$

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ 近似为 $a_i \mathbf{v}, i = 1, \dots, m$ 。

如何选择向量 $\mathbf{v}$ ？



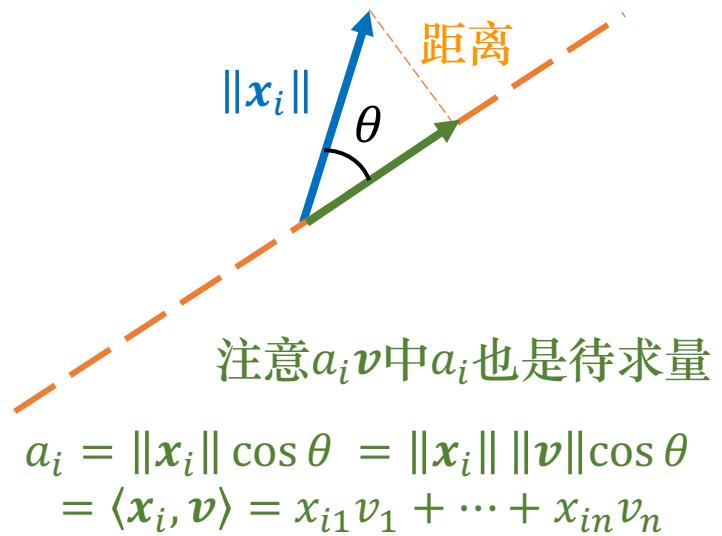
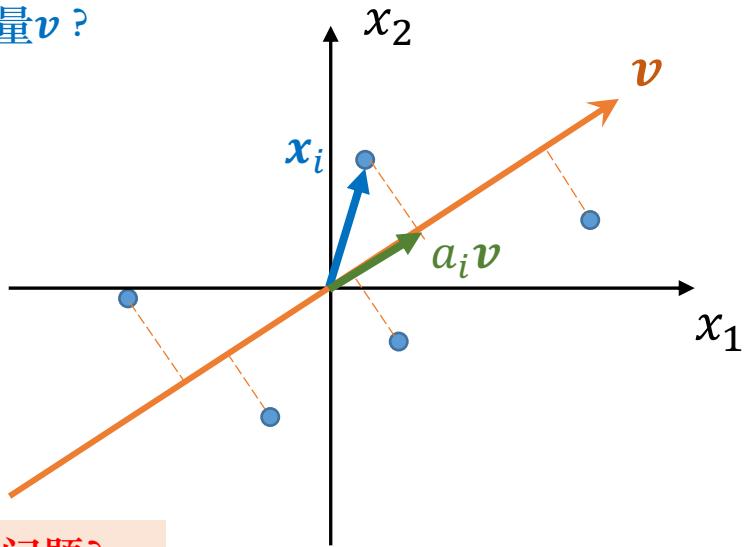
可否进一步改写问题？

$$\operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2)$$

# $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ 近似为 $a_i \mathbf{v}, i = 1, \dots, m$ 。

如何选择向量 $\mathbf{v}$ ？



注意 $a_i \mathbf{v}$ 中 $a_i$ 也是待求量

$$\begin{aligned} a_i &= \|\mathbf{x}_i\| \cos \theta = \|\mathbf{x}_i\| \|\mathbf{v}\| \cos \theta \\ &= \langle \mathbf{x}_i, \mathbf{v} \rangle = x_{i1} v_1 + \dots + x_{in} v_n \end{aligned}$$

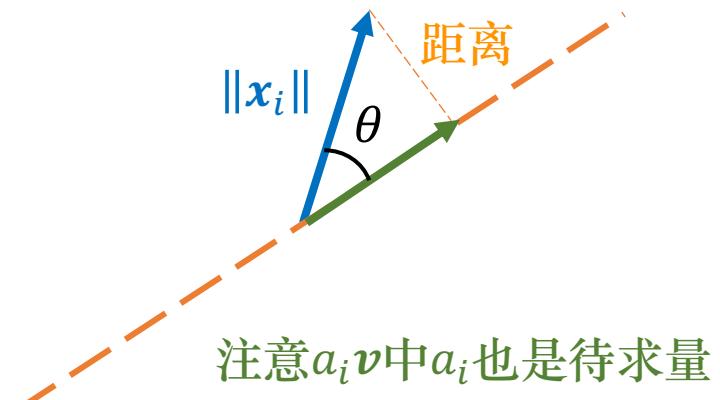
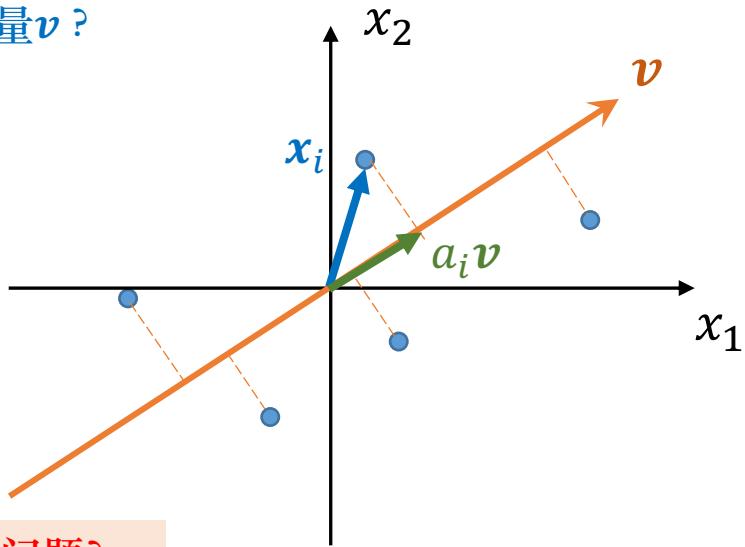
可否进一步改写问题？

$$\operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2)$$

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ 近似为 $a_i \mathbf{v}, i = 1, \dots, m$ 。

如何选择向量 $\mathbf{v}$ ？



注意 $a_i \mathbf{v}$ 中 $a_i$ 也是待求量

$$\begin{aligned} a_i &= \|\mathbf{x}_i\| \cos \theta = \|\mathbf{x}_i\| \|\mathbf{v}\| \cos \theta \\ &= \langle \mathbf{x}_i, \mathbf{v} \rangle = x_{i1} v_1 + \dots + x_{in} v_n \end{aligned}$$

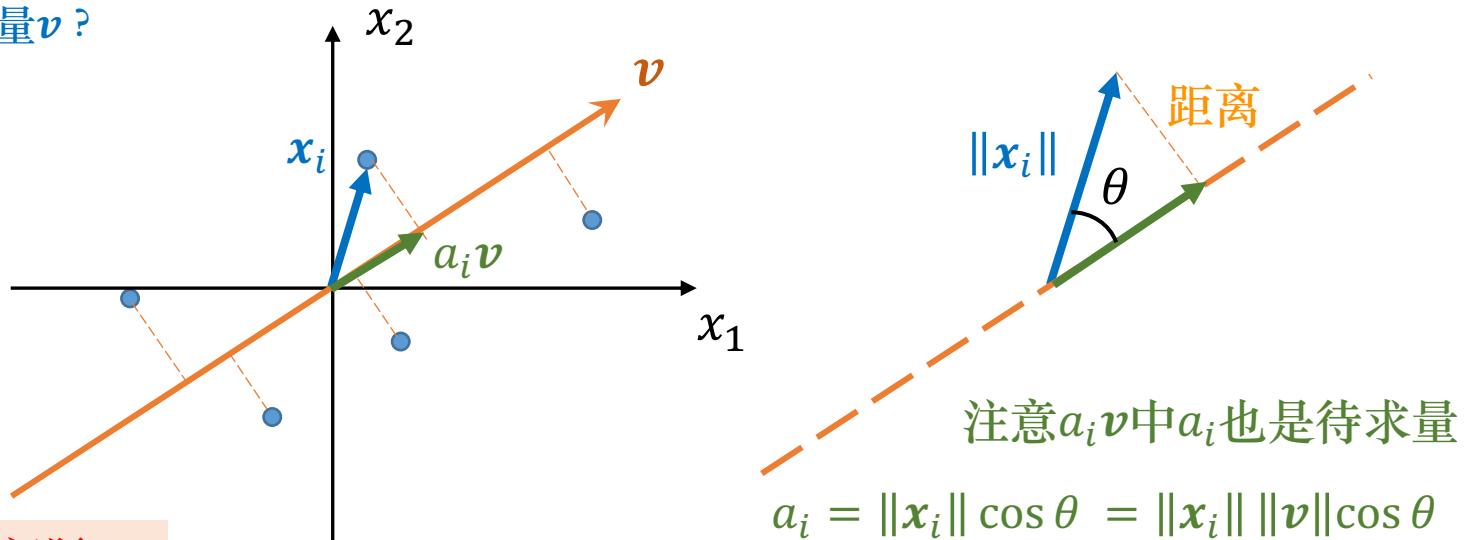
$$(\text{dist}(\mathbf{x}_i \leftrightarrow \text{line}))^2 + \langle \mathbf{x}_i, \mathbf{v} \rangle^2 = \|\mathbf{x}_i\|^2$$

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v}))^2$$

## $k = 1$ 情况

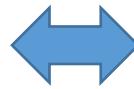
将（已完成预处理的） $m$ 个 $n$ 维向量 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ 近似为 $a_i \mathbf{v}, i = 1, \dots, m$ 。

如何选择向量 $\mathbf{v}$ ？



可否进一步改写问题？

$$\operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2)$$



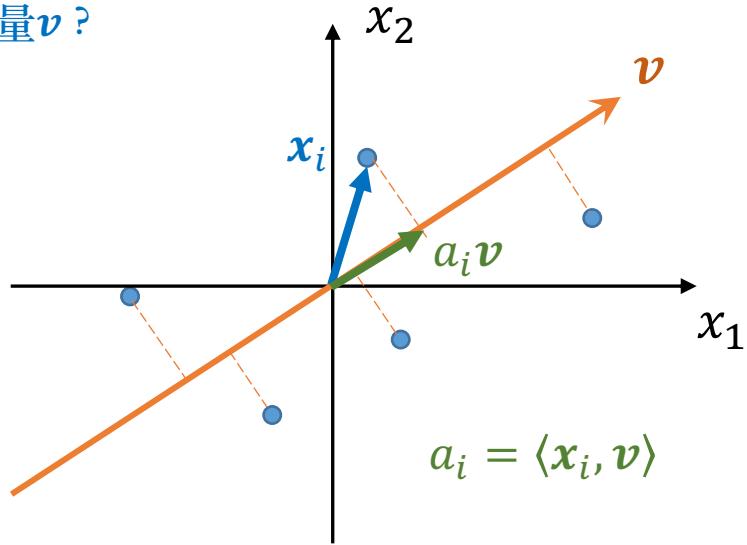
$$\operatorname{argmax}_{\mathbf{v}: \|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

因为 $\|\mathbf{x}_i\|^2$ 不受 $\mathbf{v}$ 选择的影响，可以从求和中移去

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



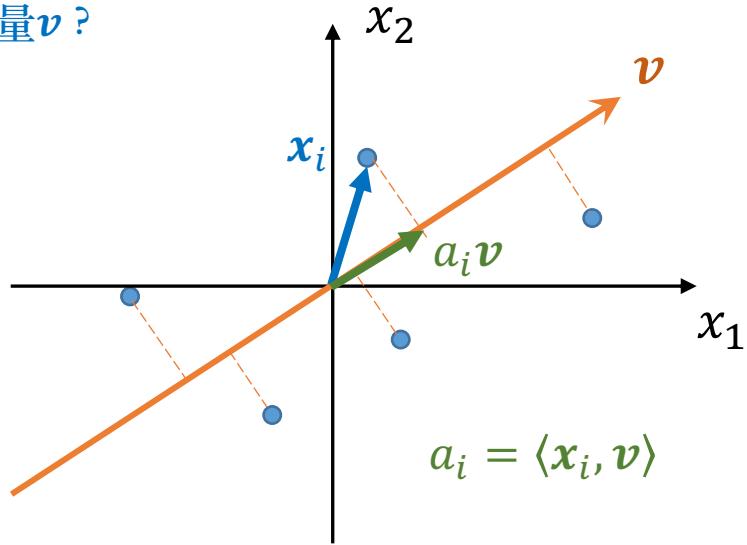
分析新问题

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



分析新问题

$$\operatorname{argmax}_{v: \|v\|=1} \frac{1}{m} \sum_{i=1}^m \langle x_i, v \rangle^2$$

即

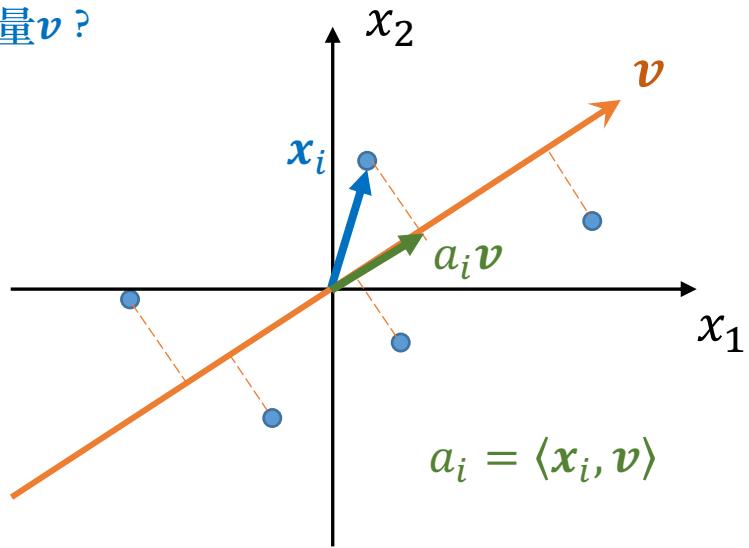
$$\operatorname{argmax}_{v: \|v\|=1} \frac{1}{m} \sum_{i=1}^m a_i^2$$

如何理解新目标函数的含义

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



可以证得  $\bar{a} = \frac{1}{m} \sum_{i=1}^m a_i = 0$

因为  $\frac{1}{m} \sum_{i=1}^m a_i = \frac{1}{m} \sum_{i=1}^m (x_{i1} v_1 + \dots + x_{in} v_n)$

由数据已经预处理得  $\frac{1}{m} v_1 \sum_{i=1}^m x_{i1} = 0$  等

分析新问题

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

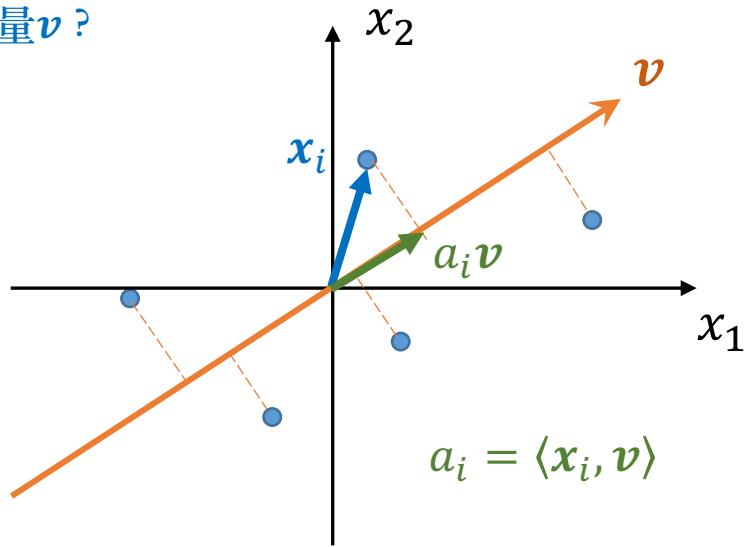
即

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m a_i^2$$

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



可以证得  $\bar{a} = \frac{1}{m} \sum_{i=1}^m a_i = 0$

因为  $\frac{1}{m} \sum_{i=1}^m a_i = \frac{1}{m} \sum_{i=1}^m (x_{i1} v_1 + \dots + x_{in} v_n)$

由数据已经预处理得  $\frac{1}{m} v_1 \sum_{i=1}^m x_{i1} = 0$  等

此时，降维后的数据均值为0

分析新问题

$$\operatorname{argmax}_{\mathbf{v}: \|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

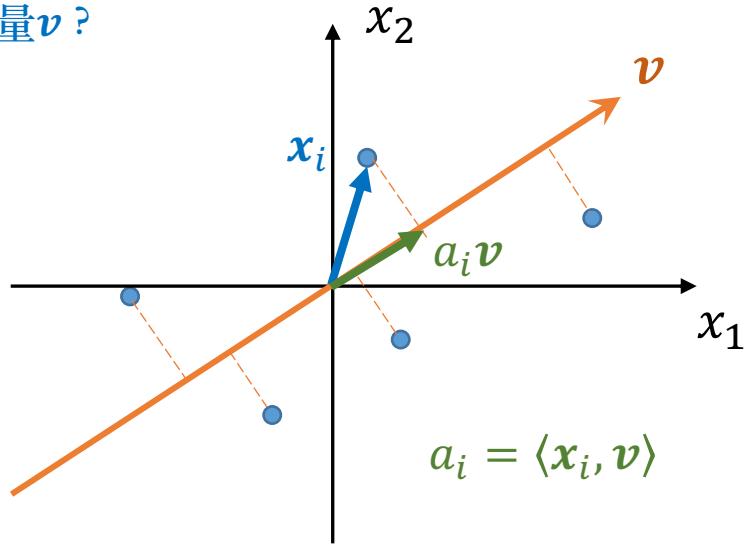
即

$$\operatorname{argmax}_{\mathbf{v}: \|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^m a_i^2$$

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



可以证得  $\bar{a} = \frac{1}{m} \sum_{i=1}^m a_i = 0$

因为  $\frac{1}{m} \sum_{i=1}^m a_i = \frac{1}{m} \sum_{i=1}^m (x_{i1} v_1 + \dots + x_{in} v_n)$

由数据已经预处理得  $\frac{1}{m} v_1 \sum_{i=1}^m x_{i1} = 0$  等

此时，降维后的数据均值为0

分析新问题

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

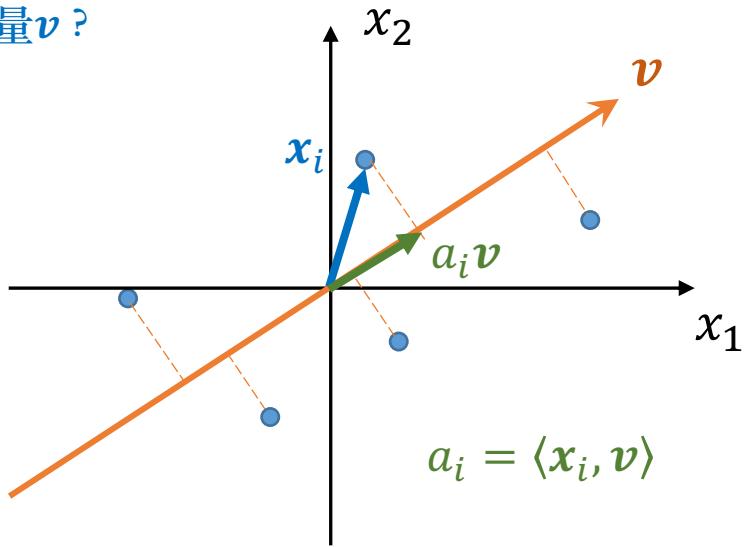
即

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m (a_i - \bar{a})^2$$

## $k = 1$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

如何选择向量 $v$ ？



$$a_i = \langle x_i, v \rangle$$

分析新问题

$$\operatorname{argmax}_{v: \|v\|=1} \frac{1}{m} \sum_{i=1}^m \langle x_i, v \rangle^2$$

即

$$\operatorname{argmax}_{v: \|v\|=1} \frac{1}{m} \sum_{i=1}^m (a_i - \bar{a})^2$$

新目标函数说明，需选择 $v$ 使降维后的数据的“方差”尽量大

即：使降维后数据之间依然可以较好区别开

## $k = 1$ 情况

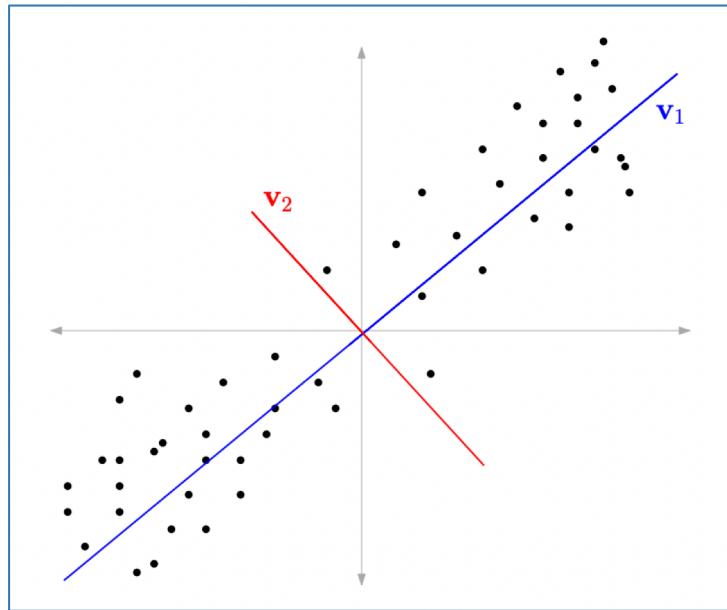
将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

即

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m (a_i - \bar{a})^2$$

需选择 $v$ 使降维后的数据的“方差”尽量大



选择 $v_1$ 好还是 $v_2$ 好?

## $k = 1$ 情况

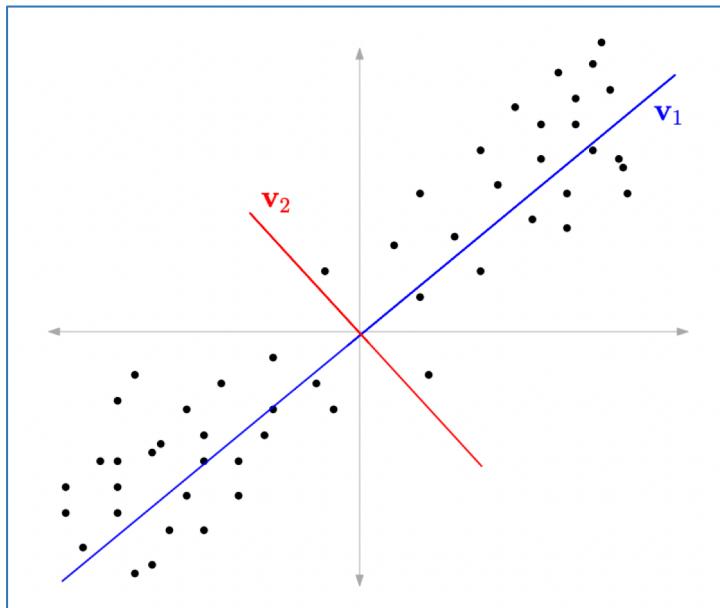
将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $a_i v, i = 1, \dots, m$ 。

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

即

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m (a_i - \bar{a})^2$$

需选择 $v$ 使降维后的数据的“方差”尽量大



选择 $v_1$ 更好

新目标函数角度：降维后数据“方差”更大

## $k = 1$ 情况

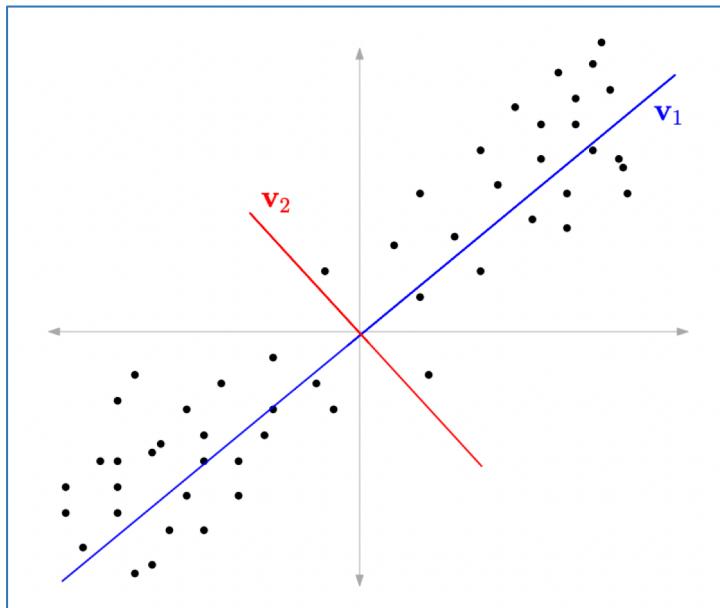
将（已完成预处理的） $m$ 个 $n$ 维向量 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ 近似为 $a_i \mathbf{v}, i = 1, \dots, m$ 。

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

即

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m (a_i - \bar{a})^2$$

需选择 $\mathbf{v}$ 使降维后的数据的“方差”尽量大



选择 $\mathbf{v}_1$ 更好

新目标函数角度：降维后数据“方差”更大

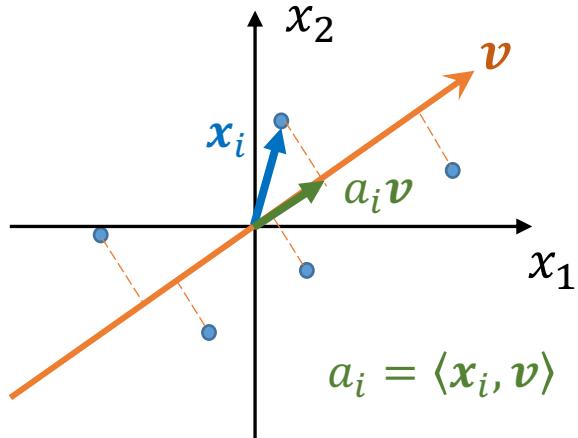
原目标函数角度：近似误差平方和更小

$$\underset{\mathbf{v}: \|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^m ((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v}))^2$$

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。

回顾 $n = 2, k = 1$ 情况

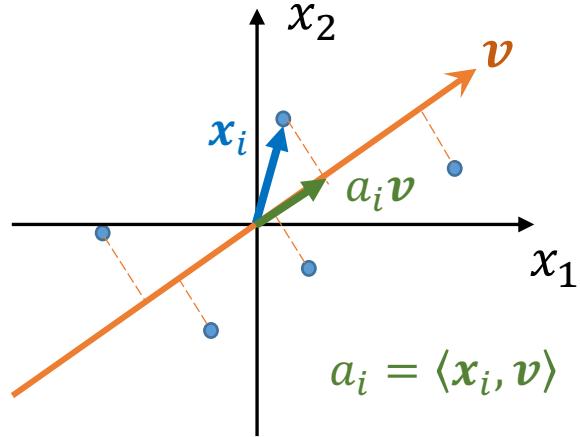


想象 $n = 3, k = 2$ 情况

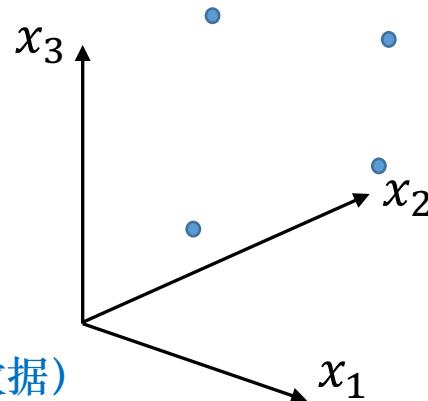
# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。

回顾 $n = 2, k = 1$ 情况



想象 $n = 3, k = 2$ 情况

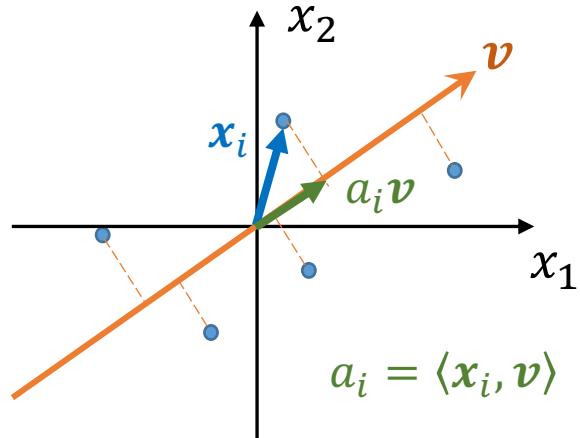


（此例中是未平移的数据）

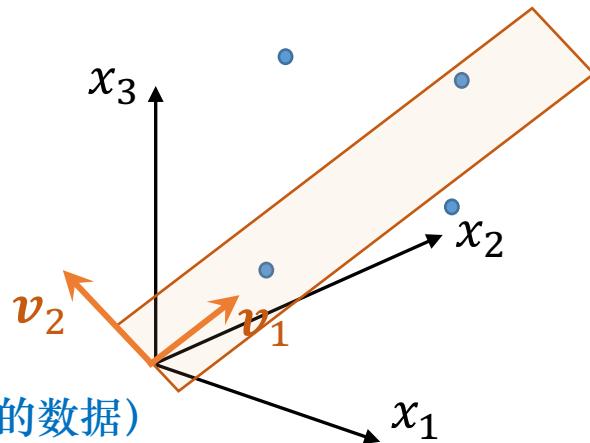
# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。

回顾 $n = 2, k = 1$ 情况



想象 $n = 3, k = 2$ 情况

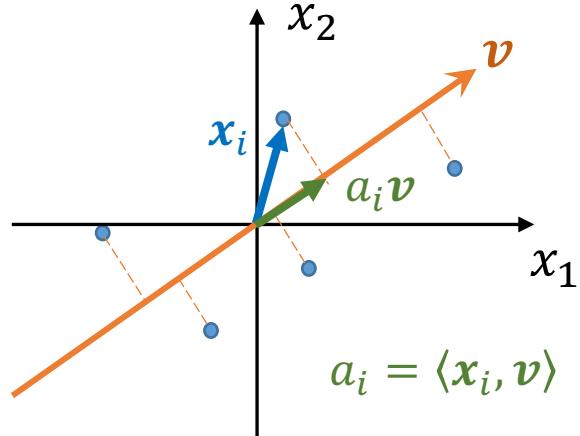


（此例中是未平移的数据）

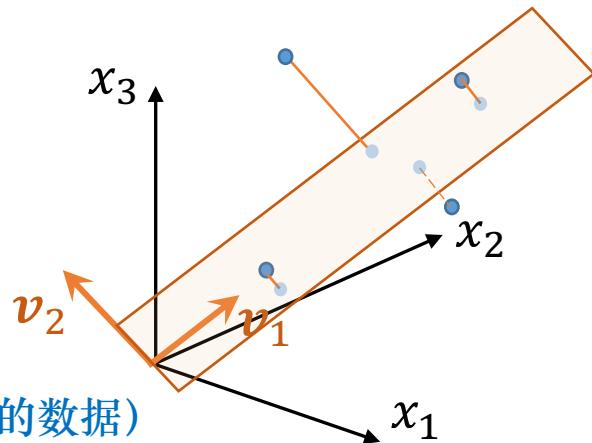
# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。

回顾 $n = 2, k = 1$ 情况



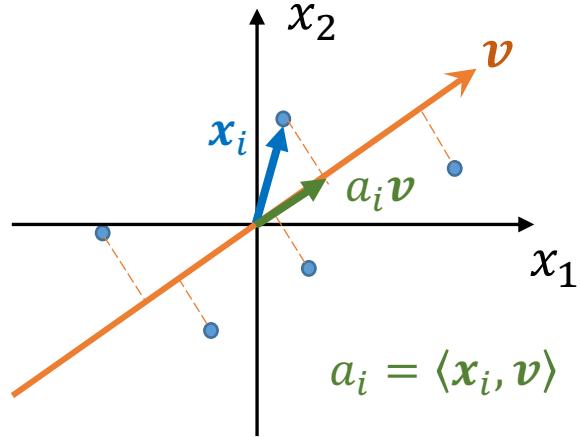
想象 $n = 3, k = 2$ 情况



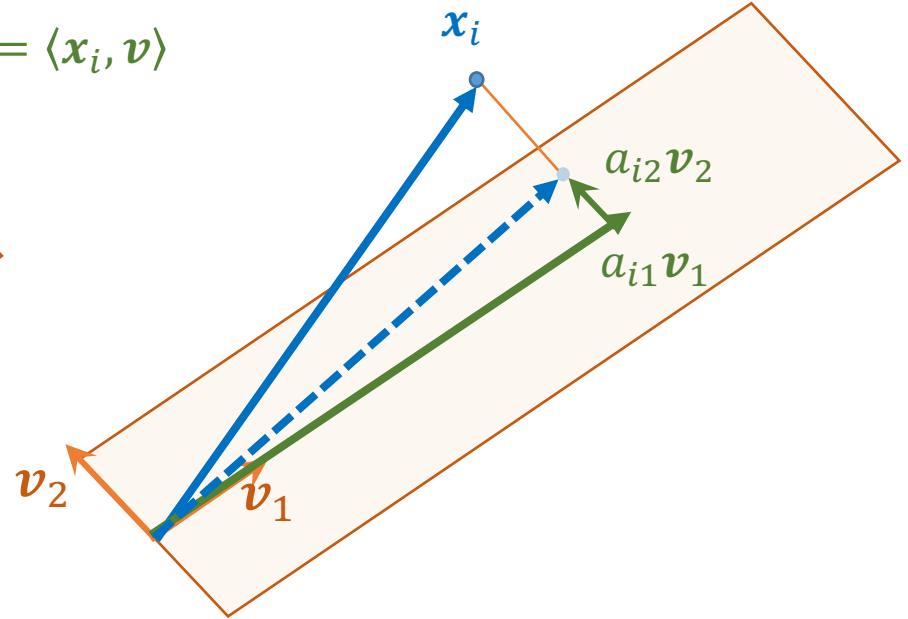
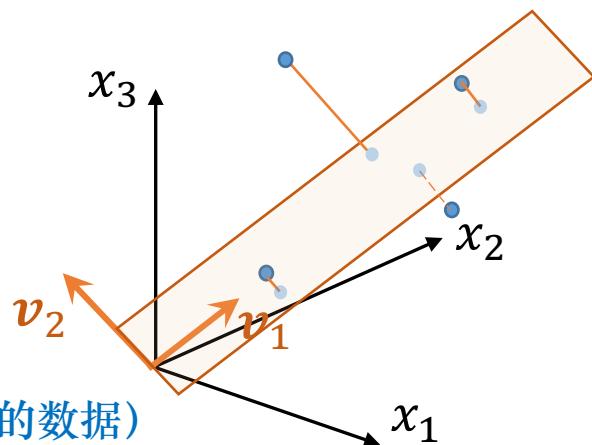
# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。

回顾 $n = 2, k = 1$ 情况

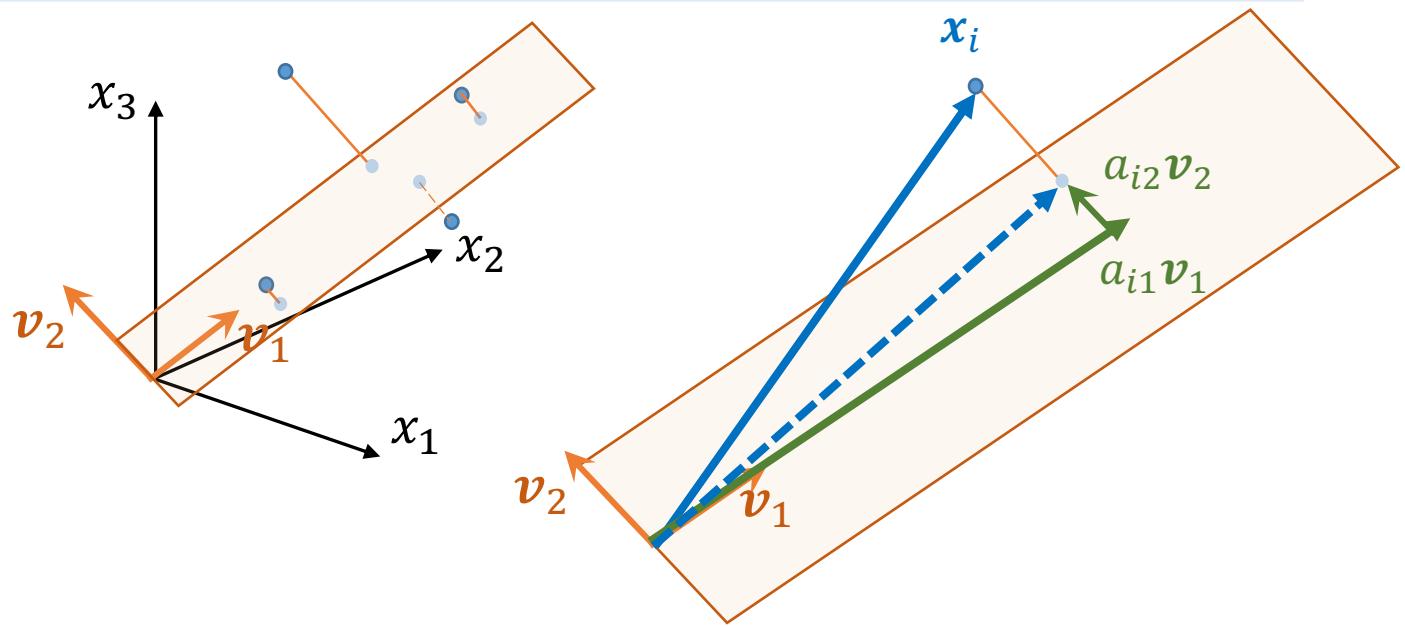


想象 $n = 3, k = 2$ 情况



# 任意 $k$ 情况

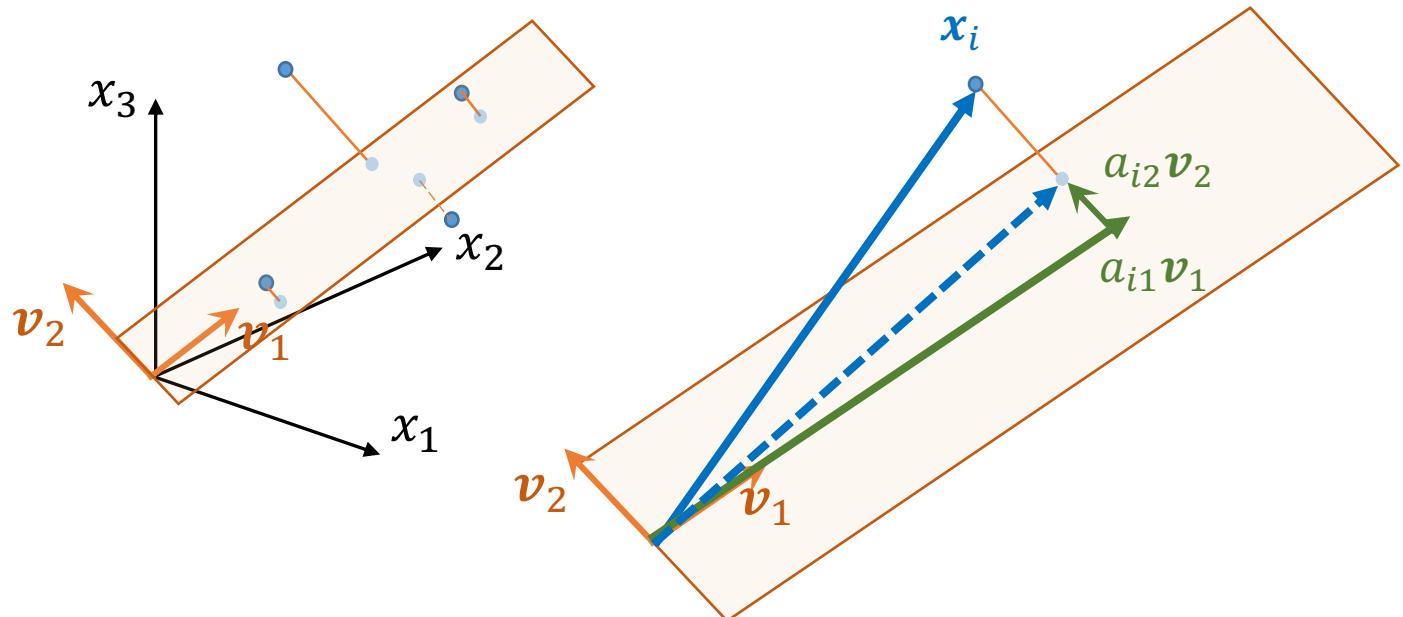
将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



优化问题?

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



## 优化问题

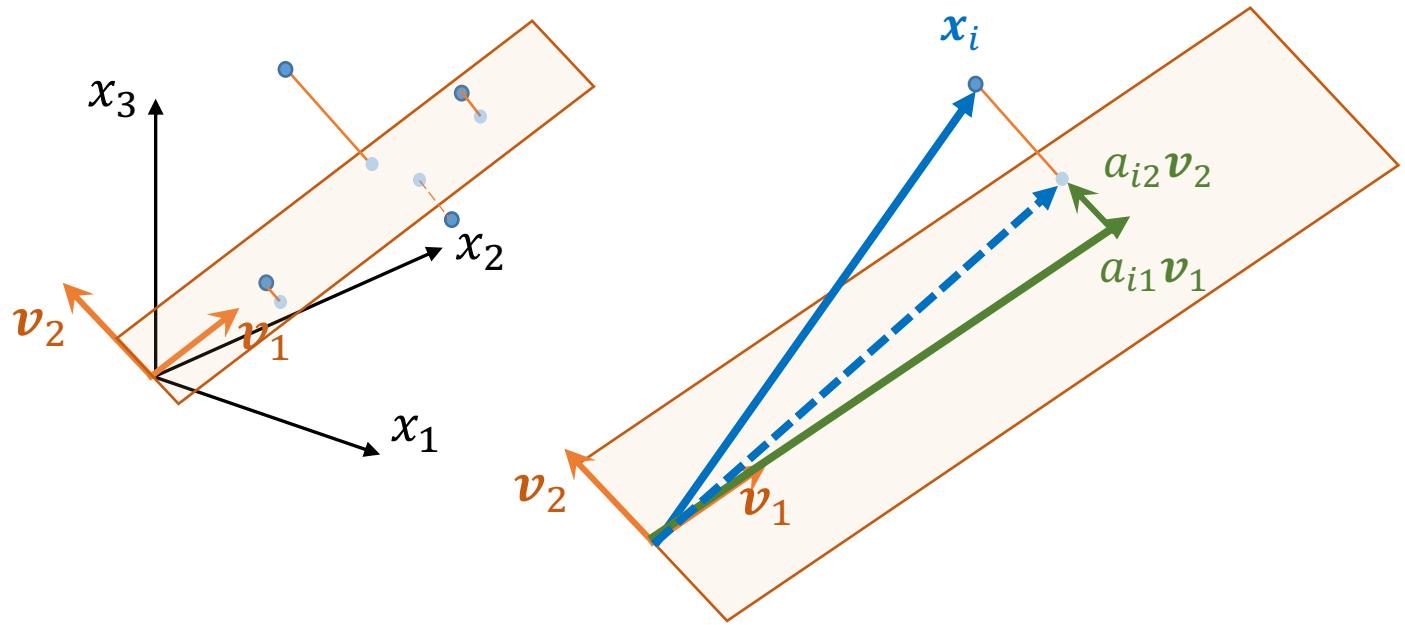
$$\operatorname{argmin} \frac{1}{m} \sum_{i=1}^m (\text{distance between } x_i \text{ and } k - \text{dimensional subspace spanned by } v_1, \dots, v_k)^2$$

当 $k = 1$ 时，即向量 $v$ 确定的直线  
当 $k = 2$ 时，即向量 $v_1, v_2$ 确定的平面

## 对向量 $v_1, \dots, v_k$ 的要求？

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



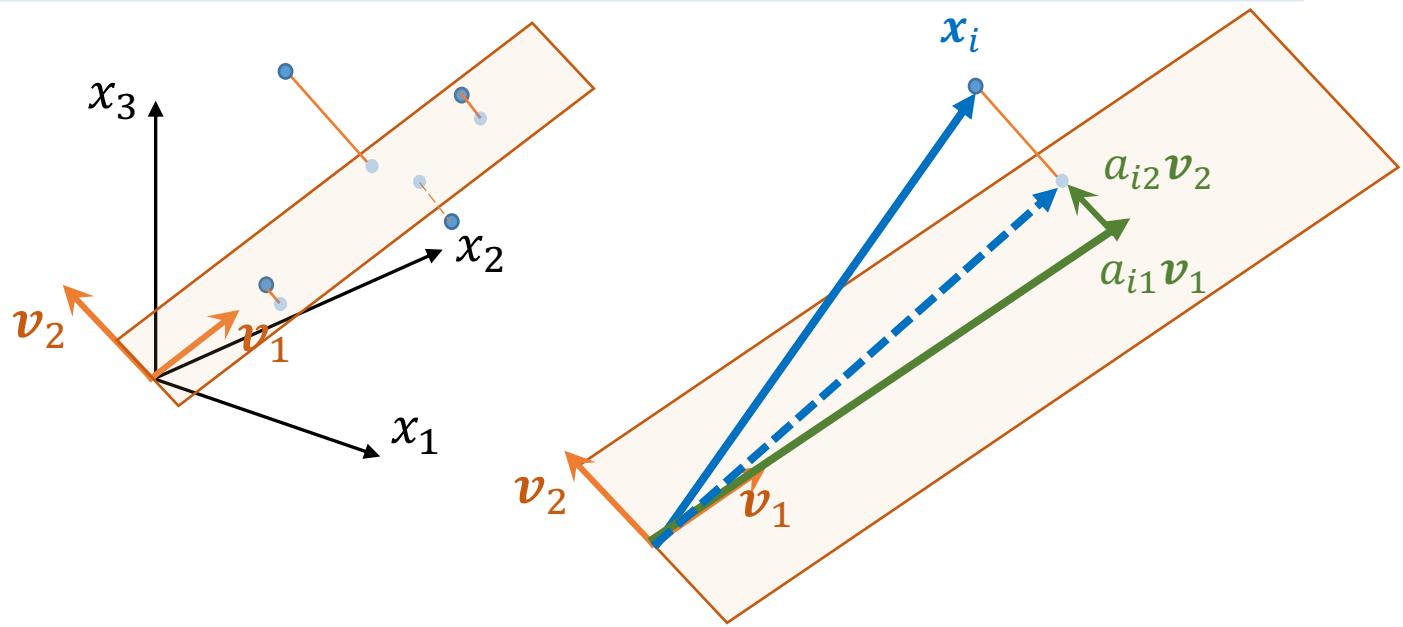
## 优化问题

$$\operatorname{argmin} \frac{1}{m} \sum_{i=1}^m (\text{distance between } x_i \text{ and } k - \text{dimensional subspace spanned by } v_1, \dots, v_k)^2$$

$v_1, \dots, v_k$ 要满足： (1) 都是单位向量，即 $\|v_1\| = \dots = \|v_k\| = 1$ ; (2) 线性独立?

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



## 优化问题

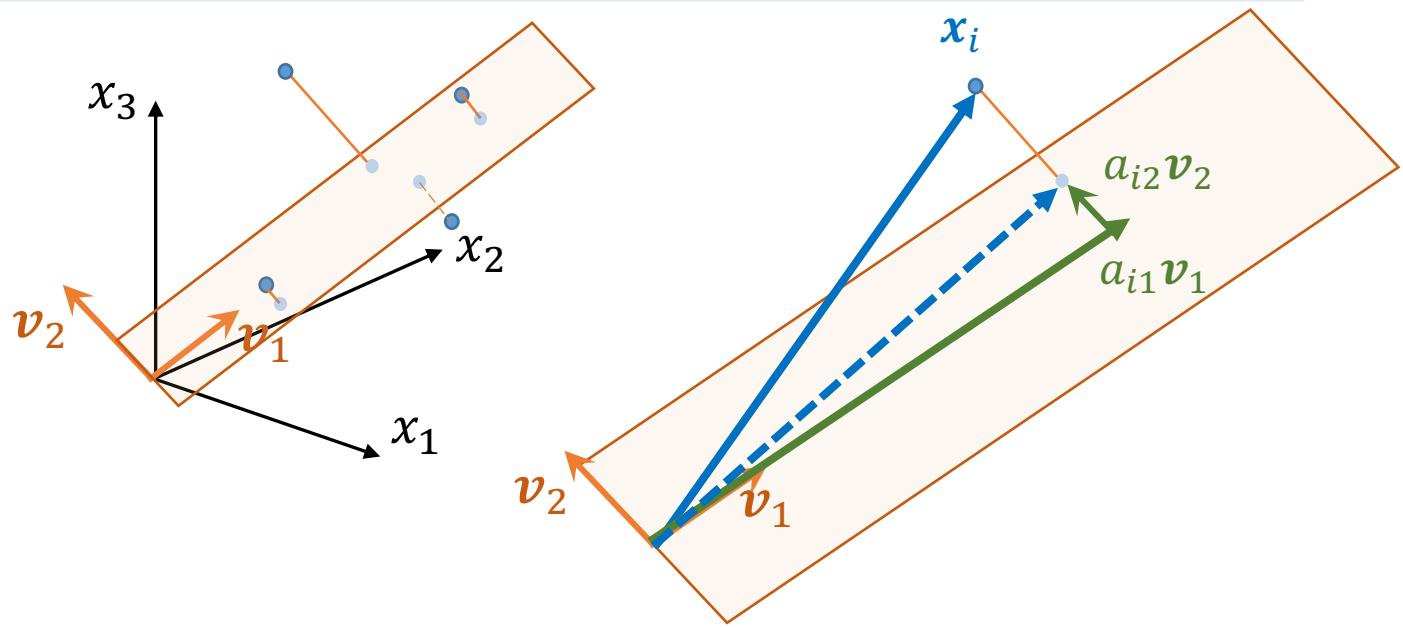
$$\operatorname{argmin} \frac{1}{m} \sum_{i=1}^m (\text{distance between } x_i \text{ and } k - \text{dimensional subspace spanned by } v_1, \dots, v_k)^2$$

$v_1, \dots, v_k$ 要满足： (1) 都是单位向量，即 $\|v_1\| = \dots = \|v_k\| = 1$ ; (2) 线性独立?

如 $k = 2$ 时，若 $v_1$ 和 $v_2$ 线性相关，说明存在常数 $c$ 使得 $v_2 = cv_1$ ，那么 $v_1$ 和 $v_2$ 无法确定一个平面

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} \mathbf{v}_j, i = 1, \dots, m$ 。



## 优化问题

$$\operatorname{argmin} \frac{1}{m} \sum_{i=1}^m (\text{distance between } \mathbf{x}_i \text{ and } k - \text{dimensional subspace spanned by } \mathbf{v}_1, \dots, \mathbf{v}_k)^2$$

$\mathbf{v}_1, \dots, \mathbf{v}_k$ 要满足： (1) 都是单位向量，即 $\|\mathbf{v}_1\| = \dots = \|\mathbf{v}_k\| = 1$ ; (2) 线性独立?

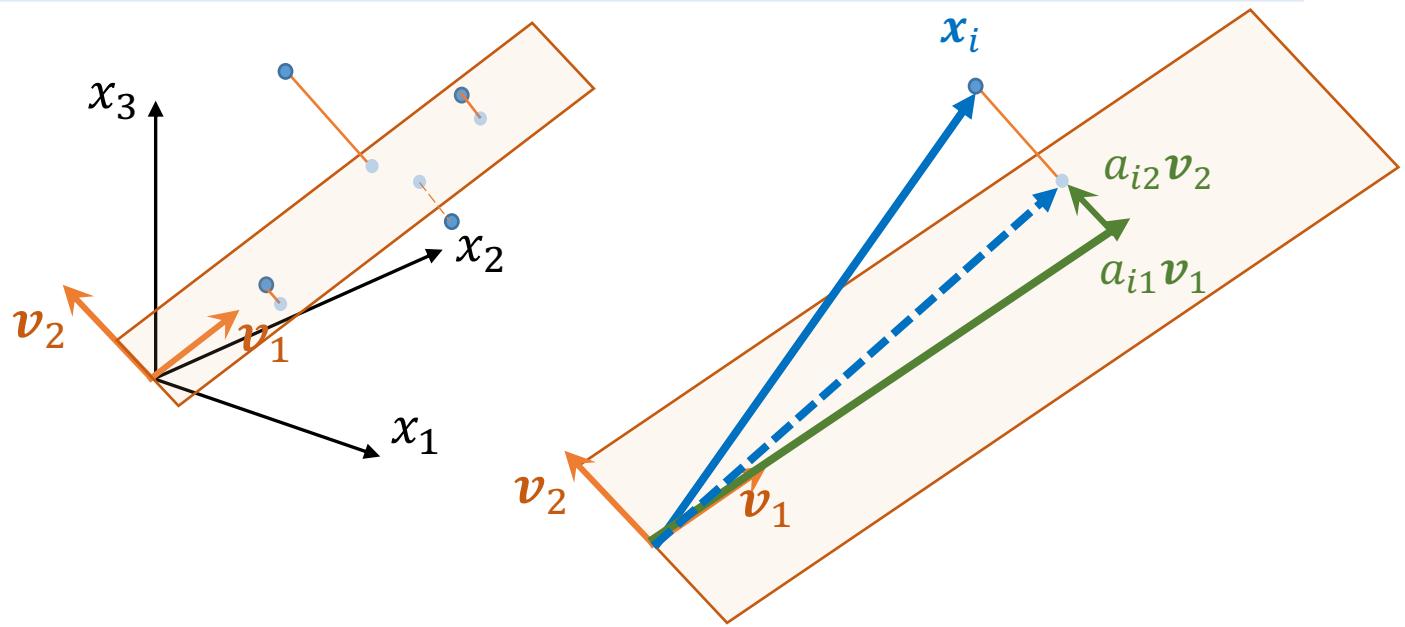
A sequence of vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  is said to be *linearly independent* if it is not linearly dependent, that is, if the equation

$$a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + \dots + a_n \mathbf{v}_n = \mathbf{0},$$

can only be satisfied by  $a_i = 0$  for  $i = 1, \dots, n$ .

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



## 优化问题

$$\operatorname{argmin} \frac{1}{m} \sum_{i=1}^m (\text{distance between } x_i \text{ and } k - \text{dimensional subspace spanned by } v_1, \dots, v_k)^2$$

夹角 $90^\circ$

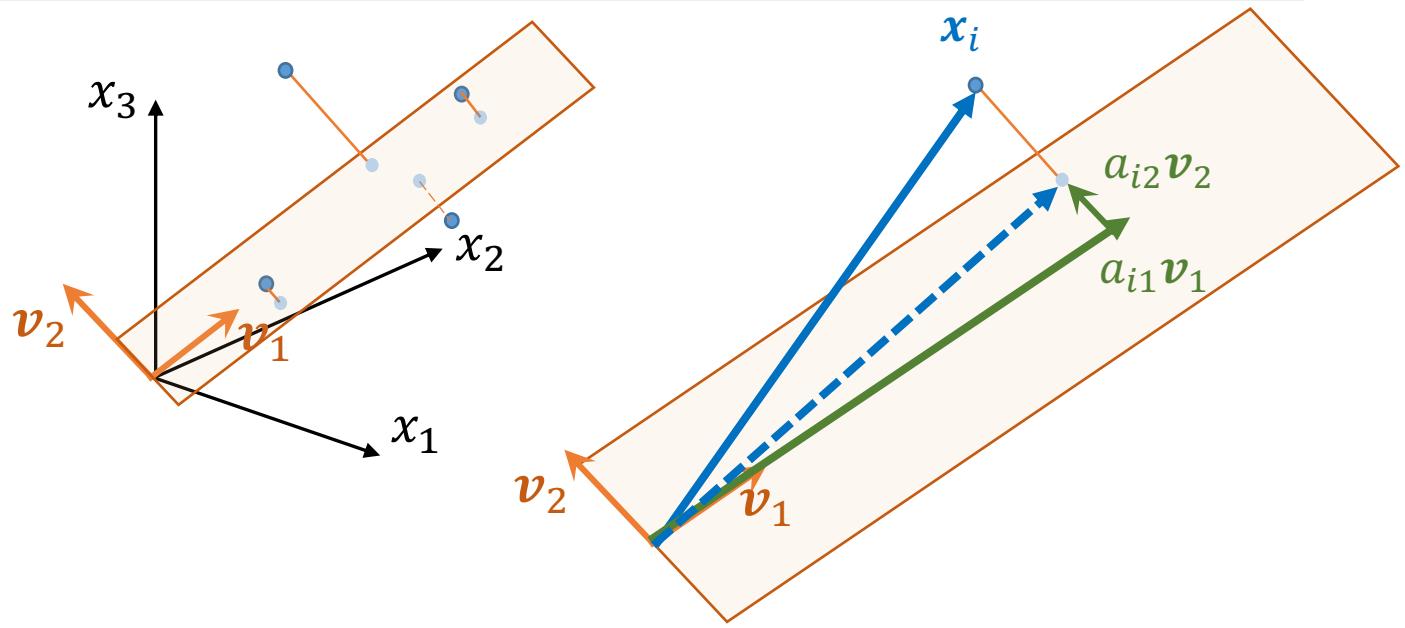
$v_1, \dots, v_k$ 要满足： (1) 都是单位向量，即 $\|v_1\| = \dots = \|v_k\| = 1$ ; (2) 正交，即 $\langle v_i, v_j \rangle = 0, \forall i \neq j$

正交一定线性独立，线性独立不一定正交

正交向量具有很好的性质，方便简化数学问题

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



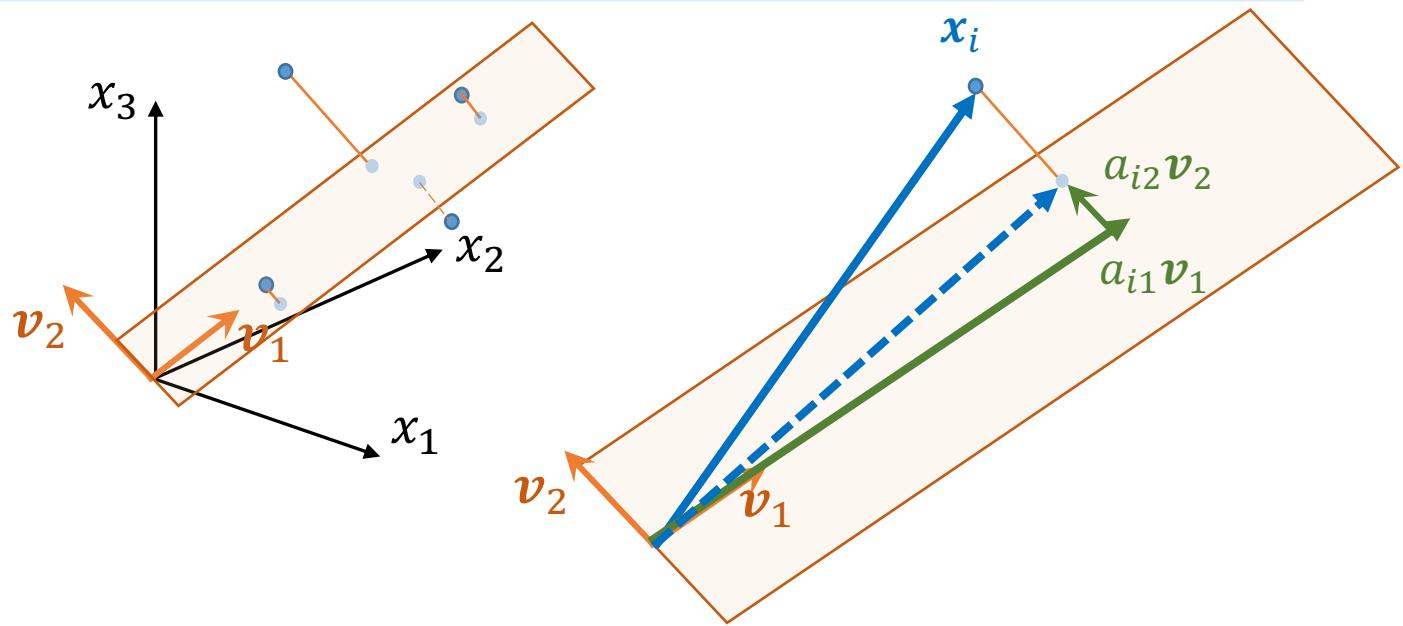
优化问题

$$\operatorname{argmin} \frac{1}{m} \sum_{i=1}^m (\text{distance between } x_i \text{ and } k - \text{dimensional subspace spanned by } v_1, \dots, v_k)^2$$

$v_1, \dots, v_k$ 是一组标准正交向量

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $x_1, \dots, x_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} v_j, i = 1, \dots, m$ 。



## 优化问题

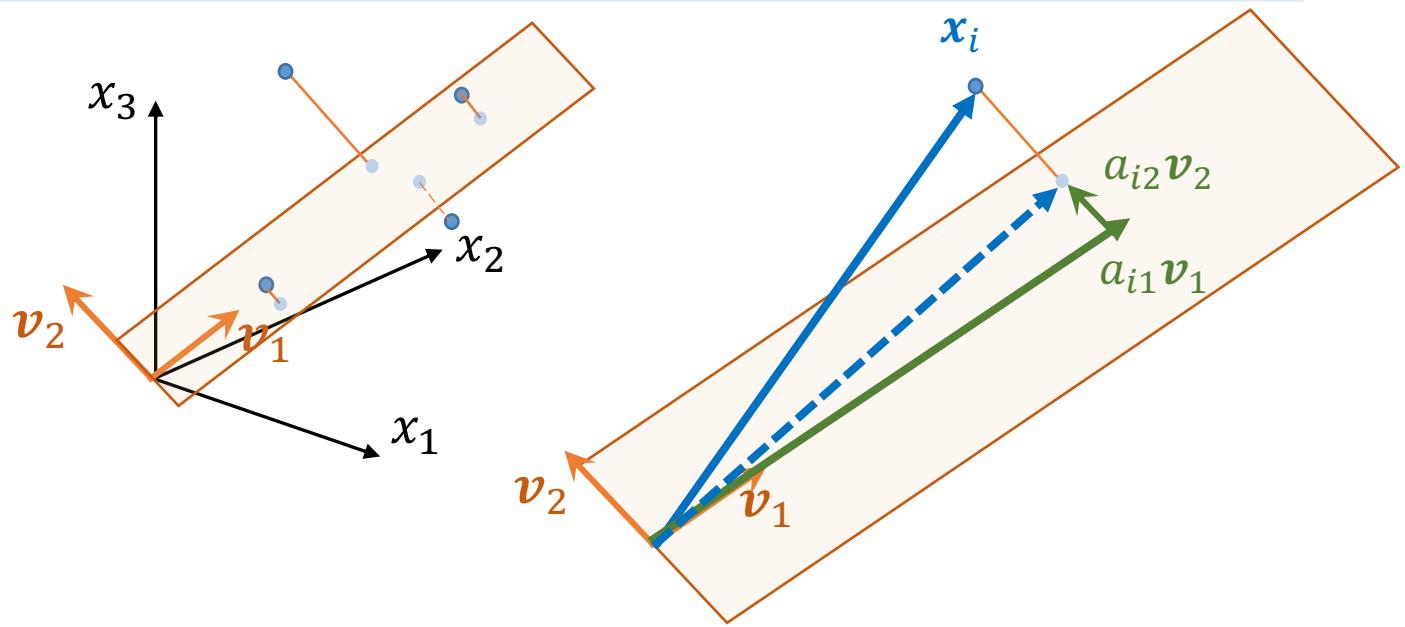
$$\operatorname{argmin} \frac{1}{m} \sum_{i=1}^m (\text{distance between } x_i \text{ and } k\text{-dimensional subspace spanned by } v_1, \dots, v_k)^2$$

$$\operatorname{argmax} \frac{1}{m} \sum_{i=1}^m \left\| \sum_{j=1}^k a_{ij} v_j \right\|^2 \quad \text{即} \quad \operatorname{argmax} \frac{1}{m} \sum_{i=1}^m (a_{i1}^2 + \dots + a_{ik}^2)$$

可用勾股定理以及标准正交向量组的性质证得

# 任意 $k$ 情况

将（已完成预处理的） $m$ 个 $n$ 维向量 $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ 近似为 $\sum_{j=1}^k a_{ij} \mathbf{v}_j, i = 1, \dots, m$ 。



## 优化问题

$$\operatorname{argmin} \frac{1}{m} \sum_{i=1}^m (\text{distance between } \mathbf{x}_i \text{ and } k - \text{dimensional subspace spanned by } \mathbf{v}_1, \dots, \mathbf{v}_k)^2$$

$$\operatorname{argmax} \frac{1}{m} \sum_{i=1}^m (\langle \mathbf{x}_i, \mathbf{v}_1 \rangle^2 + \dots + \langle \mathbf{x}_i, \mathbf{v}_k \rangle^2)$$

可由 $a_{i1} = \langle \mathbf{x}_i, \mathbf{v}_1 \rangle, \dots, a_{ik} = \langle \mathbf{x}_i, \mathbf{v}_k \rangle$ 推得

可以回顾 $k = 1$ 例子进行理解

# 主成分分析

主成分分析 (Principal Component Analysis, PCA)

给定  $x_1, \dots, x_m \in \mathbb{R}^n$  及维数  $k \geq 1$ , 求标准正交向量组  $v_1, \dots, v_k \in \mathbb{R}^n$  从而最大化

$$\frac{1}{m} \sum_{i=1}^m (\langle x_i, v_1 \rangle^2 + \dots + \langle x_i, v_k \rangle^2).$$

求得的  $v_1, \dots, v_k$  称为原数据的  $k$  个主成分



# 主成分分析

## 主成分分析效果



# 数据可视化

---

高维数据难以可视化，可通过主成分分析得到低维数据再可视化

- 步骤一：确定 $k$ （如 $k = 1,2,3$ ），通过主成分分析得到数据的 $k$ 个主成分 $v_1, \dots, v_k$

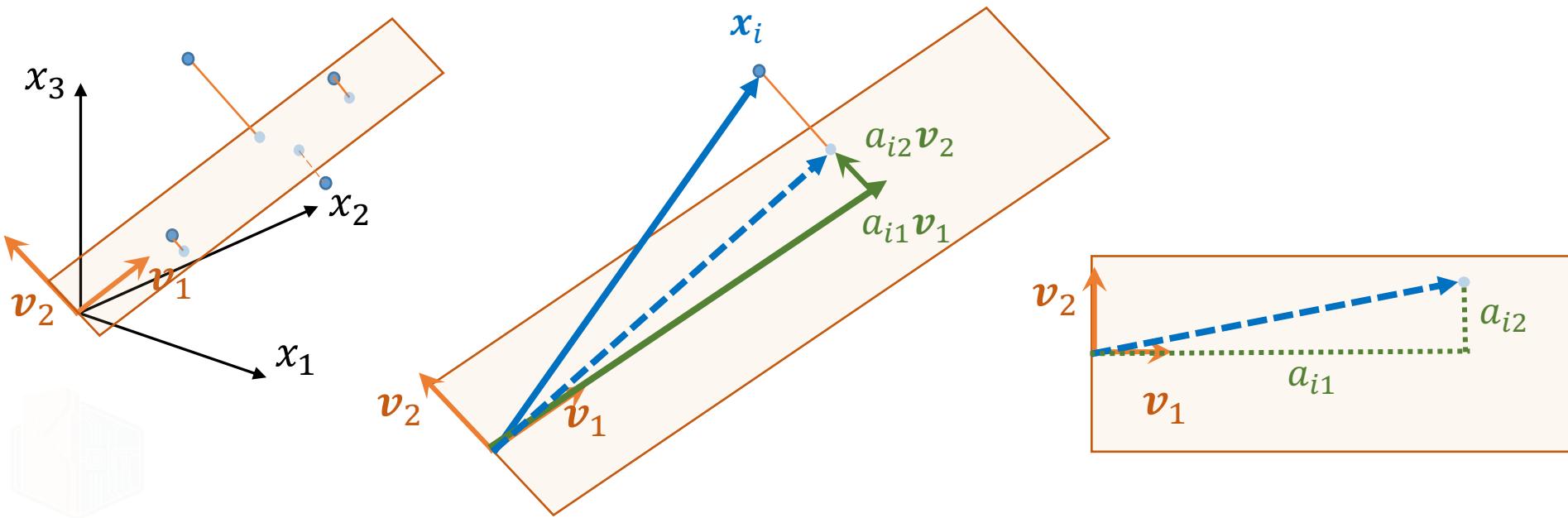


# 数据可视化

高维数据难以可视化，可通过主成分分析得到低维数据再可视化

- 步骤一：确定  $k$ （如  $k = 1, 2, 3$ ），通过主成分分析得到数据的  $k$  个主成分  $v_1, \dots, v_k$
- 步骤二：为每个数据  $x_i$ ，计算  $a_{i1} = \langle x_i, v_1 \rangle, \dots, a_{ik} = \langle x_i, v_k \rangle$

$a_{i1} \dots a_{ik}$  即数据  $x_i$  在每个主成分向量对应的坐标轴上的数值（可以为负值）

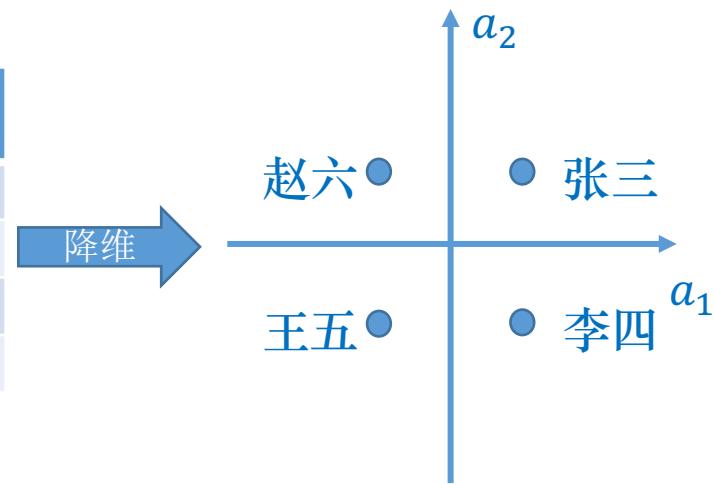


# 数据可视化

高维数据难以可视化，可通过主成分分析得到低维数据再可视化

- 步骤一：确定 $k$ （如 $k = 1, 2, 3$ ），通过主成分分析得到数据的 $k$ 个主成分 $v_1, \dots, v_k$
- 步骤二：为每个数据 $x_i$ ，计算 $a_{i1} = \langle x_i, v_1 \rangle, \dots, a_{ik} = \langle x_i, v_k \rangle$
- 步骤三：在 $k$ 维空间画出数据 $x_i$ 对应的点 $(a_{i1}, \dots, a_{ik})$ ，即 $(\langle x_i, v_1 \rangle, \dots, \langle x_i, v_k \rangle)$

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2



(在前例中未限定单位向量)

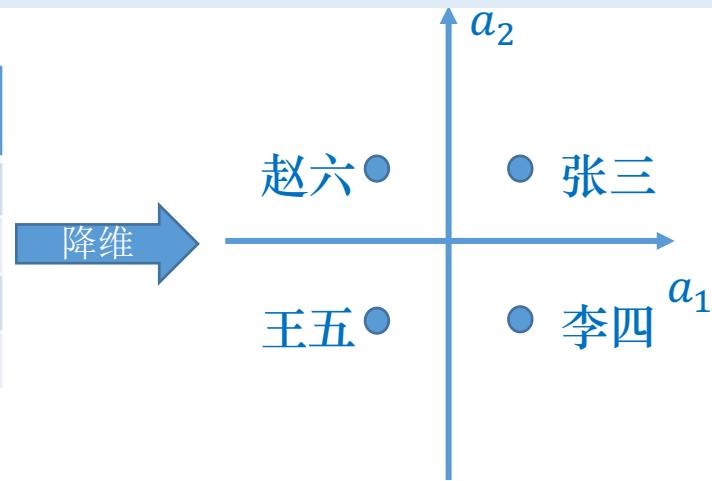
# 数据可视化

高维数据难以可视化，可通过主成分分析得到低维数据再可视化

- 步骤一：确定 $k$ （如 $k = 1, 2, 3$ ），通过主成分分析得到数据的 $k$ 个主成分 $v_1, \dots, v_k$
- 步骤二：为每个数据 $x_i$ ，计算 $a_{i1} = \langle x_i, v_1 \rangle, \dots, a_{ik} = \langle x_i, v_k \rangle$
- 步骤三：在 $k$ 维空间画出数据 $x_i$ 对应的点 $(a_{i1}, \dots, a_{ik})$ ，即 $(\langle x_i, v_1 \rangle, \dots, \langle x_i, v_k \rangle)$

分析数据：（1）解释 $v_1, \dots, v_k$ 坐标轴的含义（如通过查看哪些数据的 $a_{i1}$ 值最大/最小）  
（2）看哪些数据在低维空间距离很近、组成类（clusters）

	沙拉	肉夹馍	蒸鱼	苏打饼干
张三	10	1	2	7
李四	7	2	1	10
王五	2	9	7	3
赵六	3	6	10	2



# 数据可视化

---

可以从一个人的基因推断他的出生地吗?

## [HTML] Genes mirror geography within Europe

[J Novembre](#), [T Johnson](#), [K Bryc](#), [Z Kutalik](#), [AR Boyko](#)... - Nature, 2008 - nature.com

... levels of **genetic** differentiation among **Europeans**, we find a close correspondence between **genetic** and **geographic** distances; indeed, a **geographical** map of **Europe** arises naturally ...

 Save  Cite Cited by 1596 Related articles All 53 versions



# 数据可视化

可以从一个人的基因推断他的出生地吗？

## [HTML] Genes mirror geography within Europe

J Novembre, T Johnson, K Bryc, Z Kutalik, AR Boyko... - Nature, 2008 - nature.com

... levels of **genetic** differentiation among **Europeans**, we find a close correspondence between **genetic** and **geographic** distances; indeed, a **geographical** map of **Europe** arises naturally ...

☆ Save ⌂ Cite Cited by 1596 Related articles All 53 versions

分析**1387**个欧洲人的基因数据，包含每个人的基因组中**20**万个**SNP**位点（易发生突变的位点）

如何利用主成分分析？



# 数据可视化

可以从一个人的基因推断他的出生地吗？

[HTML] Genes mirror geography within Europe

J Novembre, T Johnson, K Bryc, Z Kutalik, AR Boyko... - Nature, 2008 - nature.com

... levels of **genetic** differentiation among **Europeans**, we find a close correspondence between **genetic** and **geographic** distances; indeed, a **geographical** map of **Europe** arises naturally ...

☆ Save ⌂ Cite Cited by 1596 Related articles All 53 versions

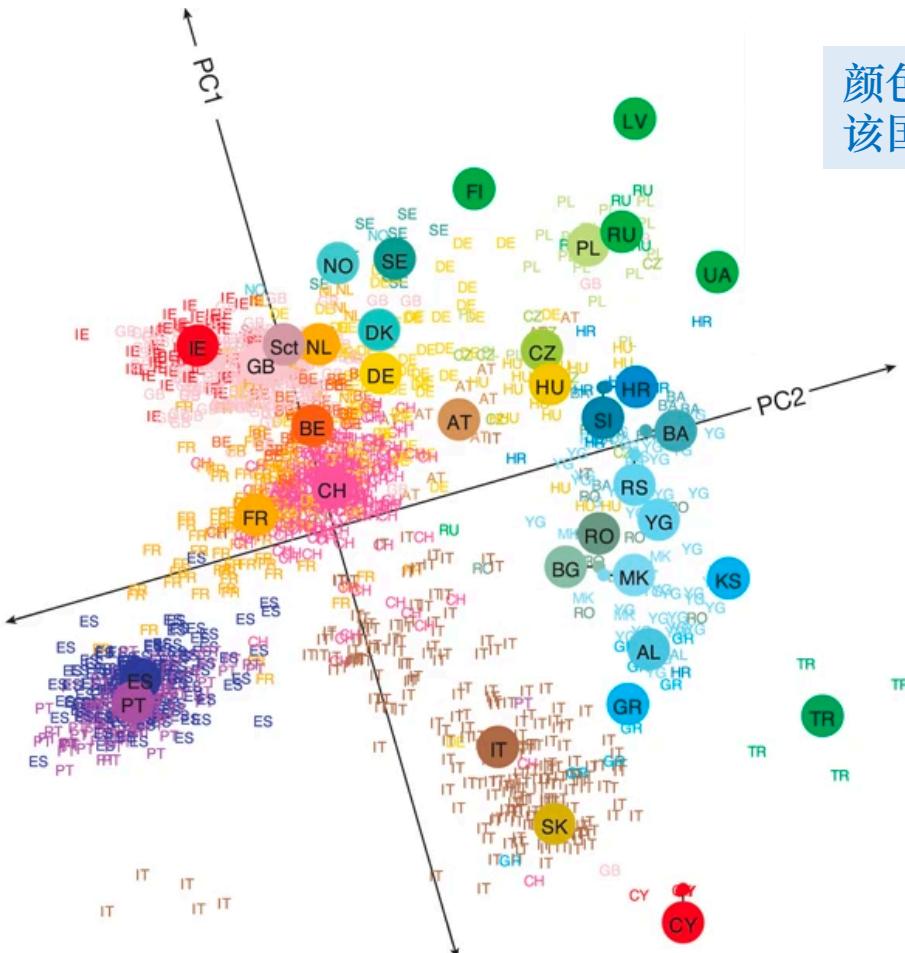
分析**1387**个欧洲人的基因数据，包含每个人的基因组中**20万个SNP位点**（易发生突变的位点）

取 $m = 1387, n = 200000, k = 2$



# 数据可视化

取  $m = 1387, n = 200000, k = 2$

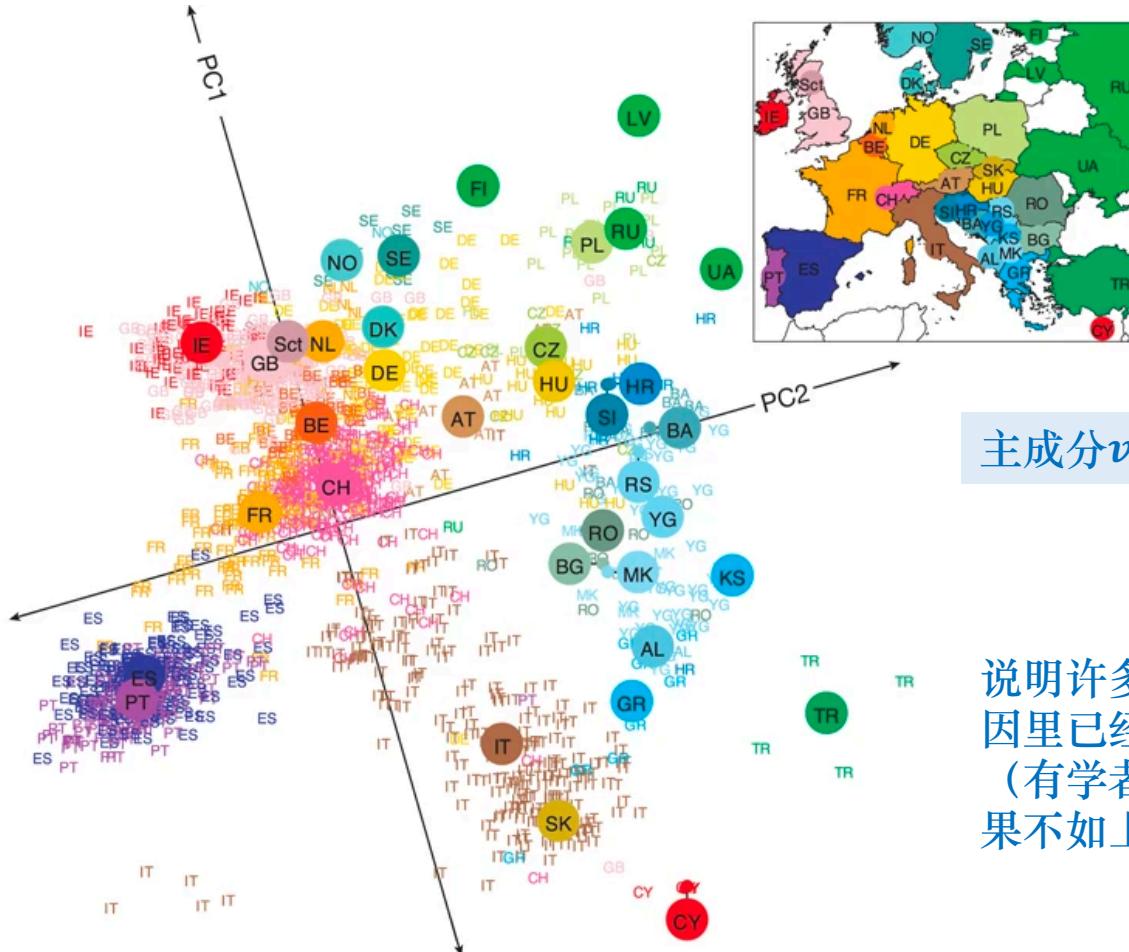


颜色对应每个数据（欧洲人）的原籍国家  
该国籍信息未参与主成分分析

图片取自J. Novembre, et al. <Genes mirror geography within Europe>

# 数据可视化

取  $m = 1387, n = 200000, k = 2$



主成分  $v_1$  与  $v_2$  的含义?

说明许多欧洲人因为世代居住在某国，基因里已经携带了相关地理信息  
(有学者用美国人数据做了类似实验，结果不如上述显著)

图片取自 J. Novembre, et al. <Genes mirror geography within Europe>

# 数据压缩

---

将高维数据进行压缩，方便存储、计算

## Eigenfaces for recognition

[M Turk, A Pentland - Journal of cognitive neuroscience, 1991 - direct.mit.edu](#)

We have developed a near-real-time computer system that can locate and track a subject's head, and then recognize the person by comparing characteristics of the face to those of ...

[☆ Save](#) [⤒ Cite](#) [Cited by 20072](#) [Related articles](#) [All 34 versions](#) [⤓](#)

将人脸图像（6.5万像素，即 $n = 65000$ ）压缩至100~150维数据（即 $k = 100 \sim 150$ ）  
便于存储、做近似查找等



# 数据压缩

Dataset For each face there should be a few training examples



All faces should be centered



R. Grosse @ Uni. of Toronto

08	02	22	97	39	25	00	40	00	78	08	09	07	78	02	12	00			
69	49	99	40	17	81	18	57	40	87	17	40	98	49	63	01	54	42	00	
01	49	31	73	59	79	14	29	93	71	40	87	13	30	03	49	13	34	45	
52	70	95	23	04	60	11	42	63	59	54	01	32	54	71	37	02	36	91	
22	31	16	71	51	37	39	41	92	36	54	22	40	40	28	64	33	13	80	
24	47	13	60	99	03	45	02	44	79	33	59	78	36	04	20	35	17	12	50
02	98	81	28	64	23	67	10	26	38	80	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	95	69	99	39	63	08	60	91	64	89	94	21
24	55	58	05	64	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	14	14	09	53	54	92
16	39	05	42	94	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
06	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	03	94	47	69	28	73	92	13	86	52	17	77	04	89	55	40
04	52	03	83	97	35	99	16	07	97	57	32	16	26	24	79	33	27	98	66
44	44	87	57	42	20	72	03	44	33	67	46	55	12	32	63	93	53	69	
04	42	16	73	01	46	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	85	39	39	49	82	47	59	85	74	04	34	16	
20	73	35	29	78	31	90	01	74	31	49	71	58	74	51	16	23	57	05	51
23	70	54	73	83	51	54	69	16	92	33	45	61	43	52	03	87	21	62	42

What the computer sees

$N \times M$  matrix

$NM \times 1$  vector

## 数据集

## 图片对应的向量

# 数据压缩

Mean face  $\bar{x}$



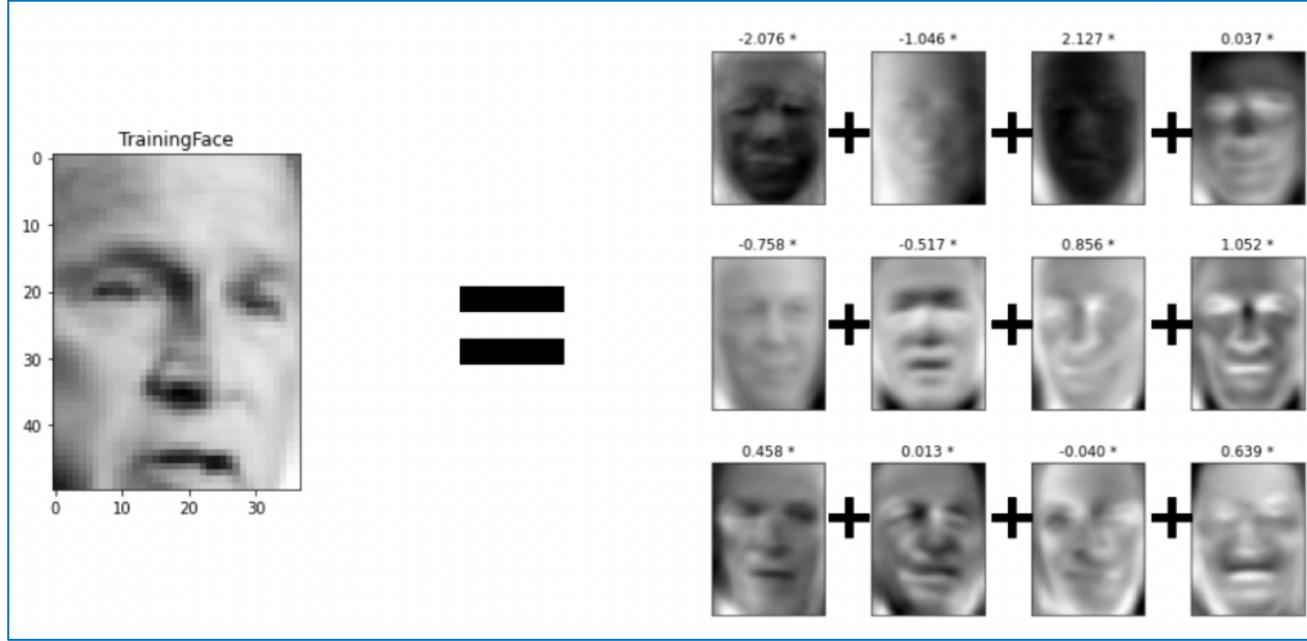
平均向量对应的图片

Top eigenvectors:  $u_1, \dots, u_k$  (visualized as images - eigenfaces)



提取的主成分对应的图片

# 数据压缩



可将数据集中任一图片对应的向量近似为主成分的线性组合

# 数据压缩



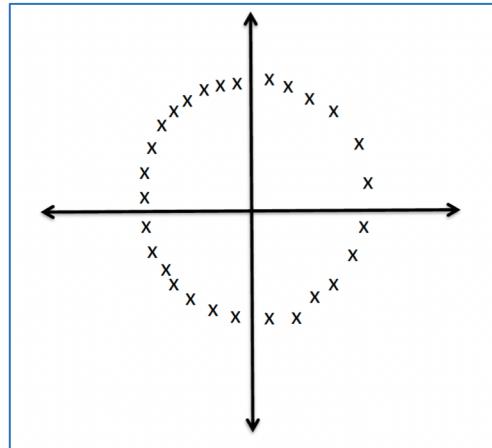
随着主成分个数的增多，近似造成的损失越少

# 讨论

---

什么情况下PCA效果不好?

- 预处理没做好 (如剔除离群数据)
- 数据本身有非线性结构 (PCA擅长捕捉线性结构)
- 主成分之间限定需要正交, 导致后续主成分 (如第三、第四个主成分) 难以解释



数据本身可以很好地近似成一维数据  
(角度) 但PCA难以实现

# 本讲小结

---



主成分分析的背景



主成分分析的问题及应用



# 主要参考资料

Tim Roughgarden and Gregory Valiant <CS 168 - The Modern Algorithmic Toolbox> Lecture Notes  
Cameron Musco <COMPSCI 514 - Algorithms for Data Science> Slides  
Imdad Ullah Khan <Lecture Notes for Big Data Analytics - Principle Component Analysis>  
J. Novembre, et al. <Genes mirror geography within Europe> Paper

# 谢谢！

