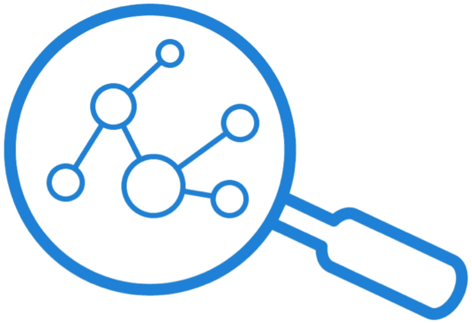
数据科学与大数据技术的数学基础

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	JL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

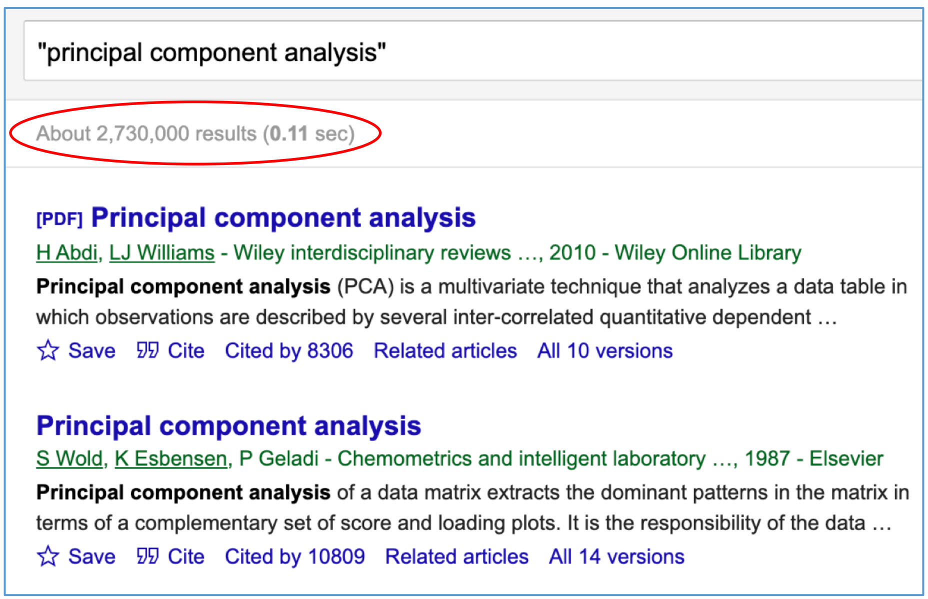
Part1 随机化方法
一致性哈希 布隆过滤器 CM Sketch方法 最小哈希 欧氏距离下的相似搜索 Jaccard相似度下的相似搜索
Part2 谱分析方法
主成分分析 奇异值分解 谱图论
Part3 最优化方法
压缩感知

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



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？

例子修改自Tim Roughgarden and Gregory Valiant <CS 168 - The Modern Algorithmic Toolbox>

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

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

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

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

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

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

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

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

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

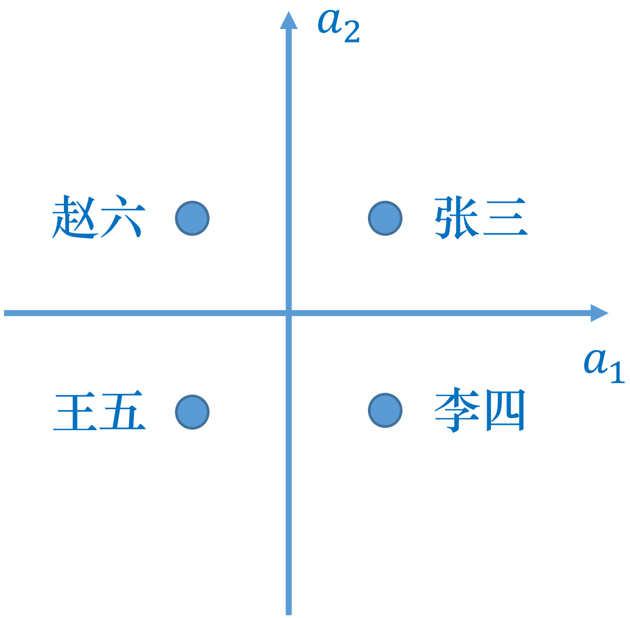
可用二维向量(a_1, a_2)近似表示四个数据

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

降维



四个二维数据

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

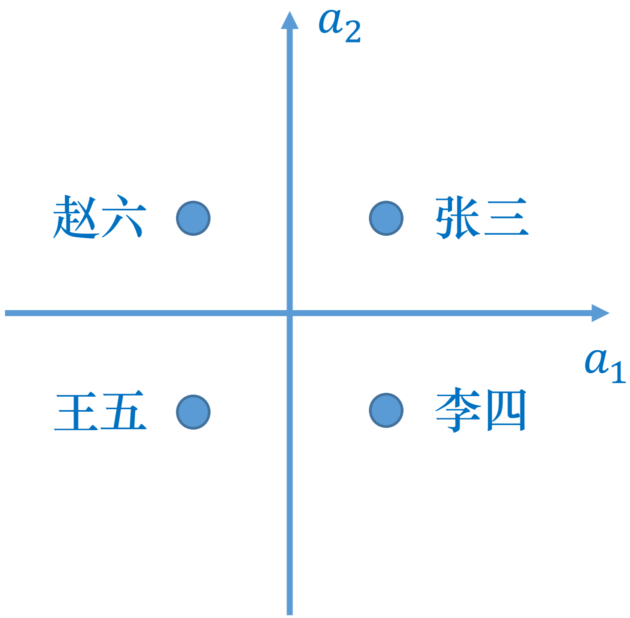
$$\overline{\boldsymbol{x}} + a_1\boldsymbol{v_1} + a_2\boldsymbol{v_2}$$

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

降维



四个二维数据

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

$$\overline{\boldsymbol{x}} + a_1\boldsymbol{v_1} + a_2\boldsymbol{v_2}$$

降维的作用：

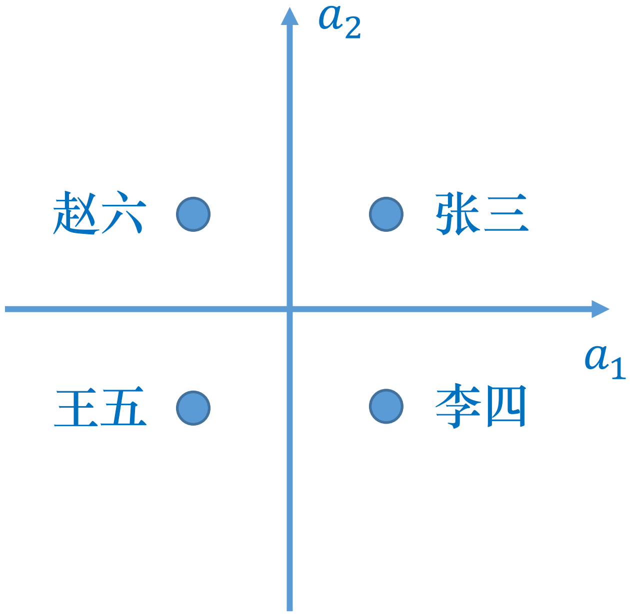
一、方便对数据进行可视化（若有更多人的数据，可以看到有哪些类、哪些人之间更相似）

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

降维



四个**二**维数据

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

$$\overline{\boldsymbol{x}} + a_1\boldsymbol{v_1} + a_2\boldsymbol{v_2}$$

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）

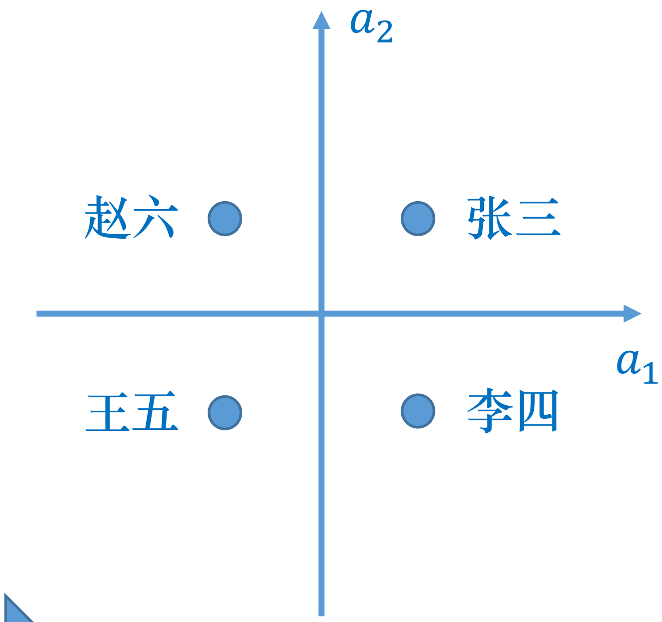
二、通过解释$\boldsymbol{v_1}, \boldsymbol{v_2}$，解释数据（如$\boldsymbol{v_1}$对应是否喜欢吃素、 $\boldsymbol{v_2}$对应是否注重健康，而这些人在是否吃素方面的差异比在是否注重健康方面的差异更大）

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

降维



四个二维数据

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

$$\overline{\boldsymbol{x}} + a_1\boldsymbol{v_1} + a_2\boldsymbol{v_2}$$

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

背景

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

背景

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

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

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

$$\overline{\boldsymbol{x}} + a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2$$

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

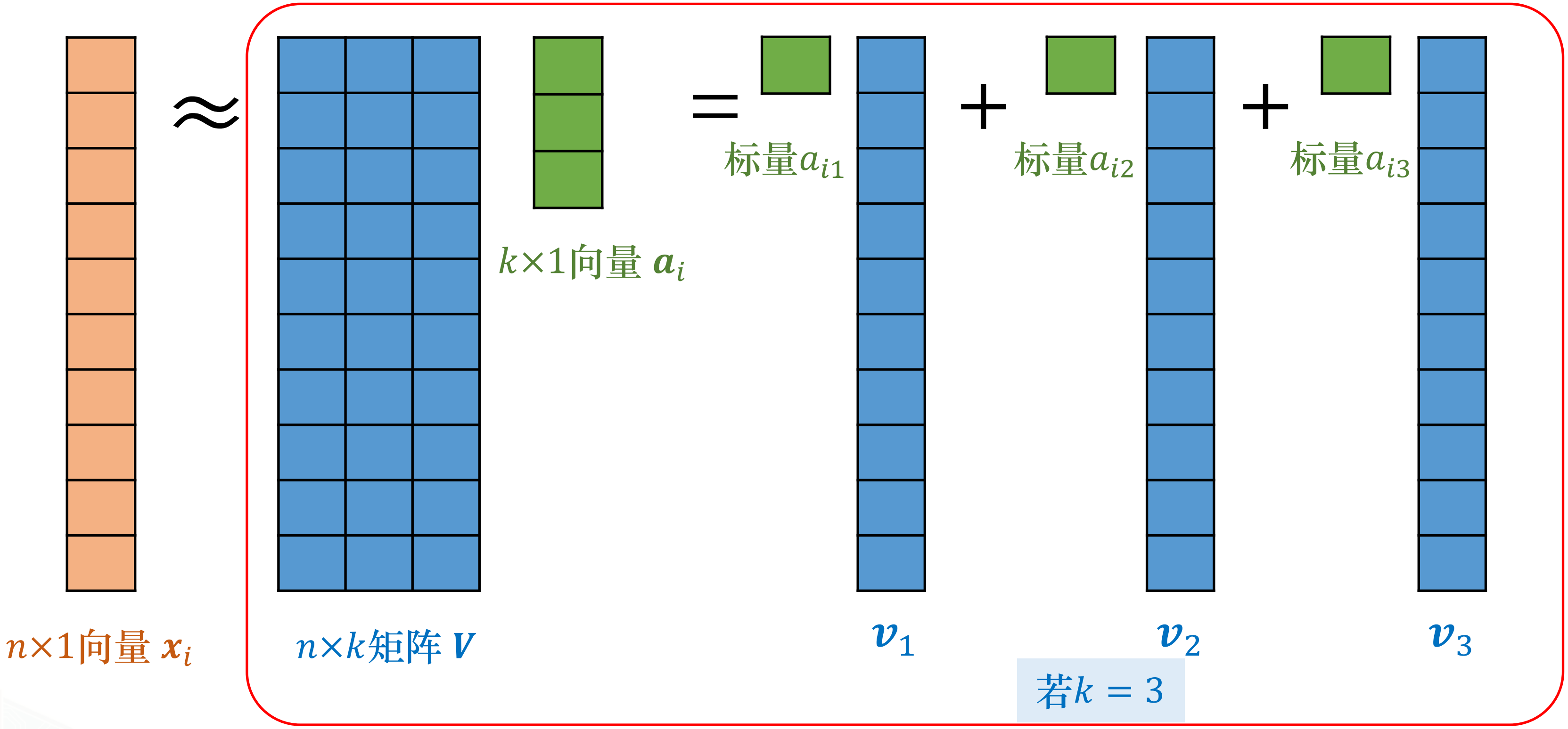
背景

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

如何把列向量$\boldsymbol{x}_i$写成矩阵与向量的乘积形式?

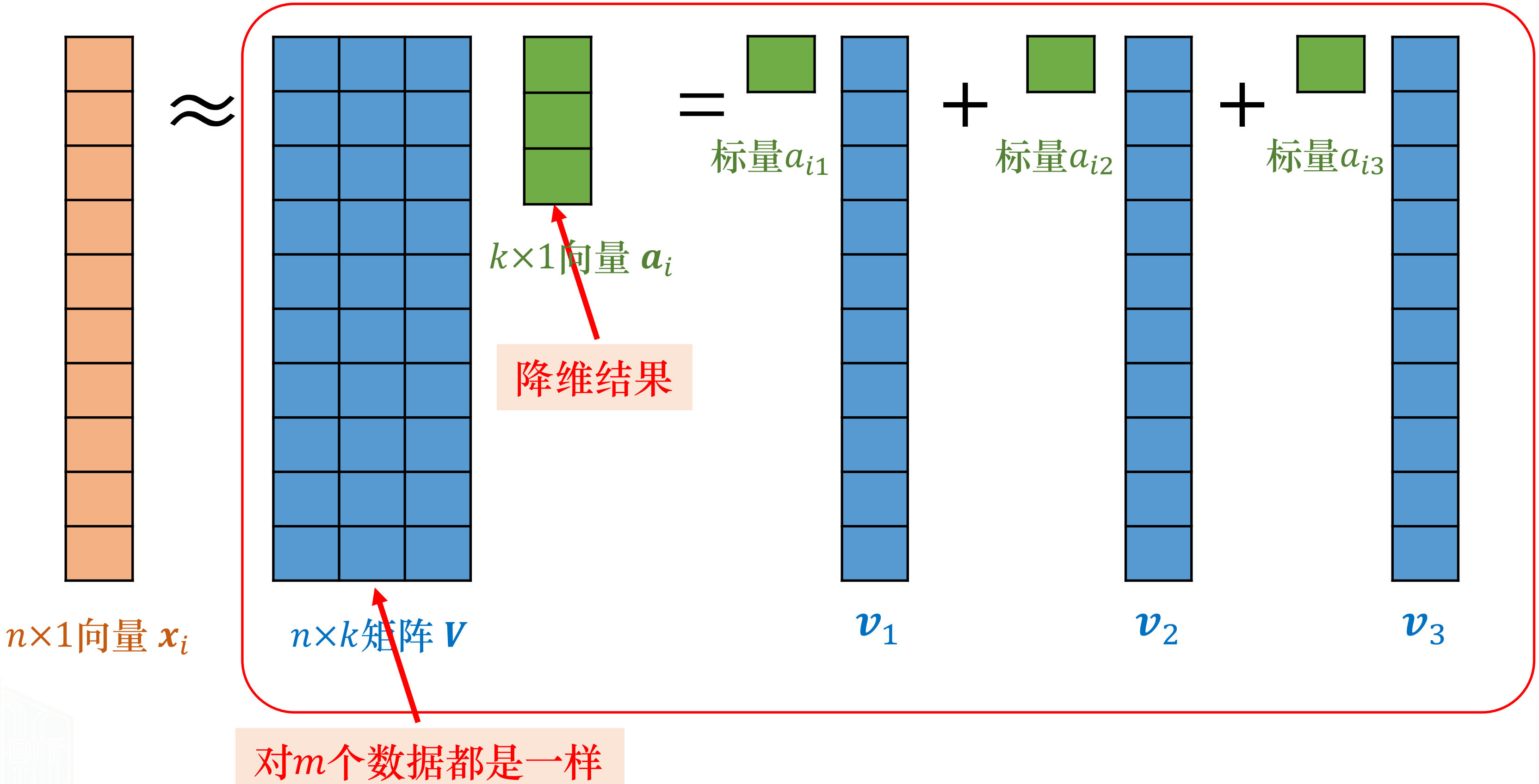
背景

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



背景

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

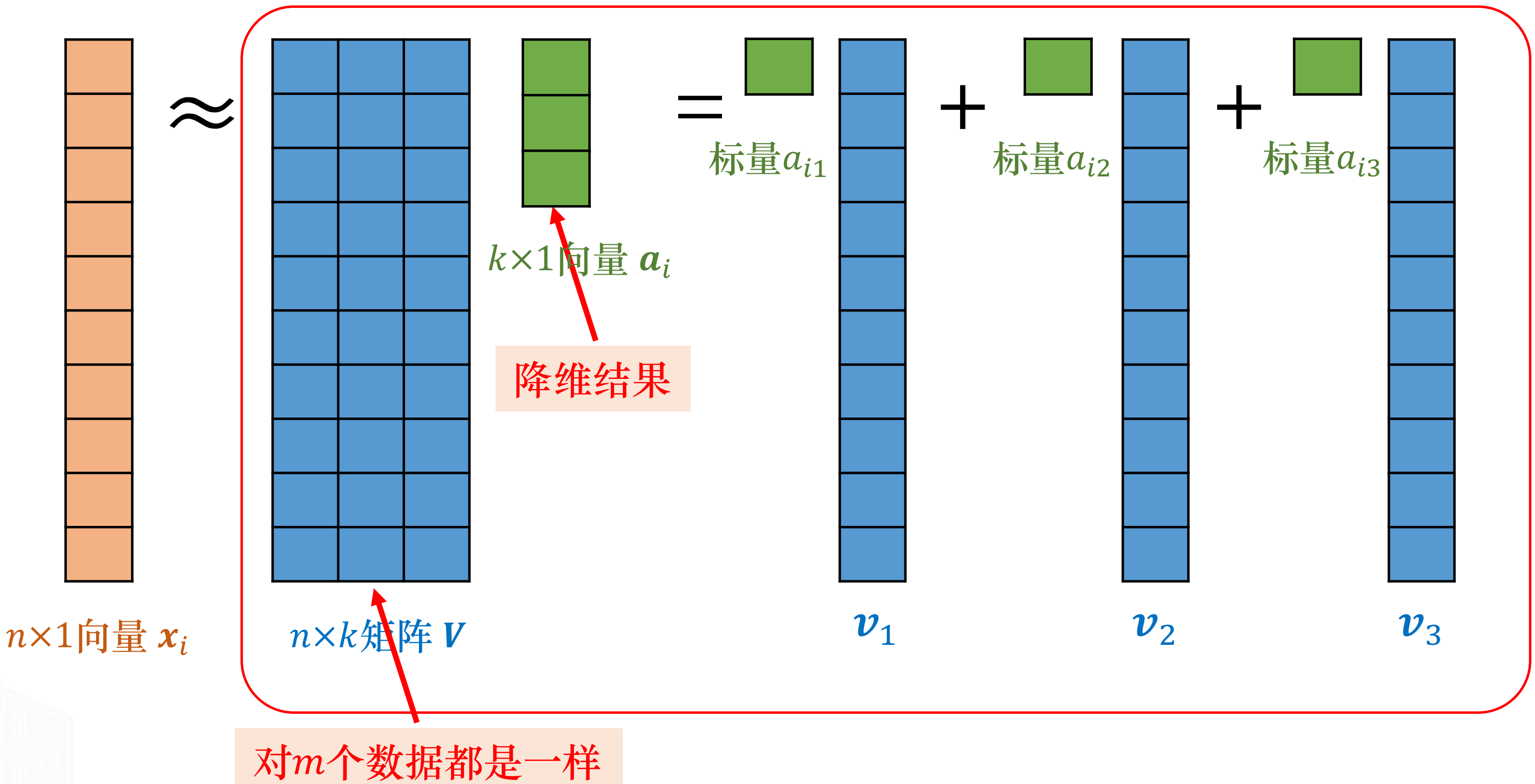


背景

	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

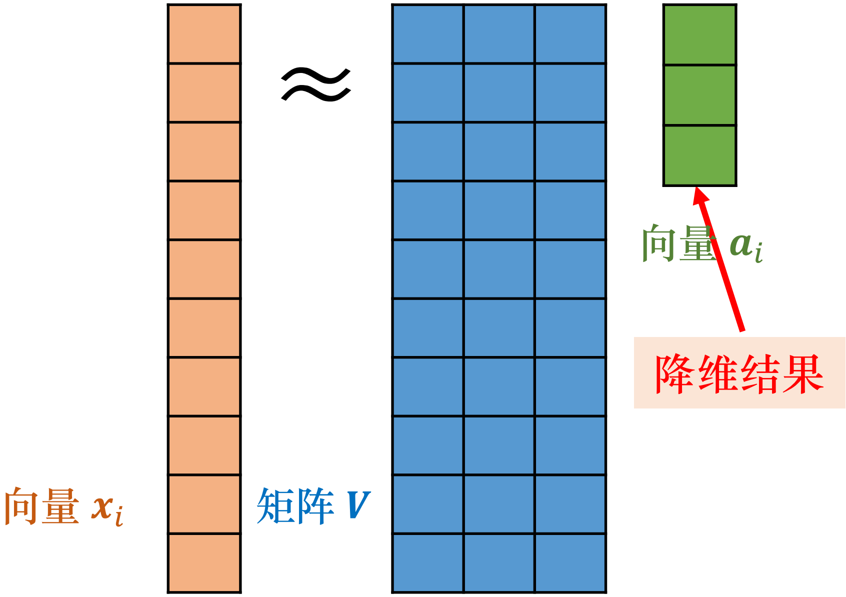
此例中，$m = 4, n = 4, k = 2$

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

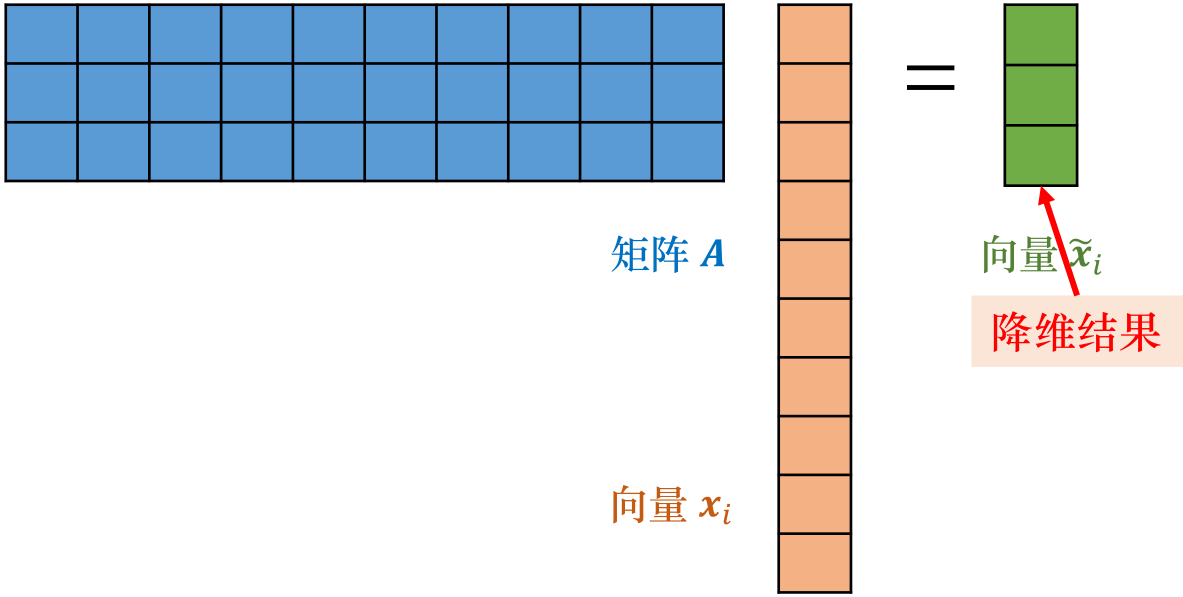


PCA与JL转换对比

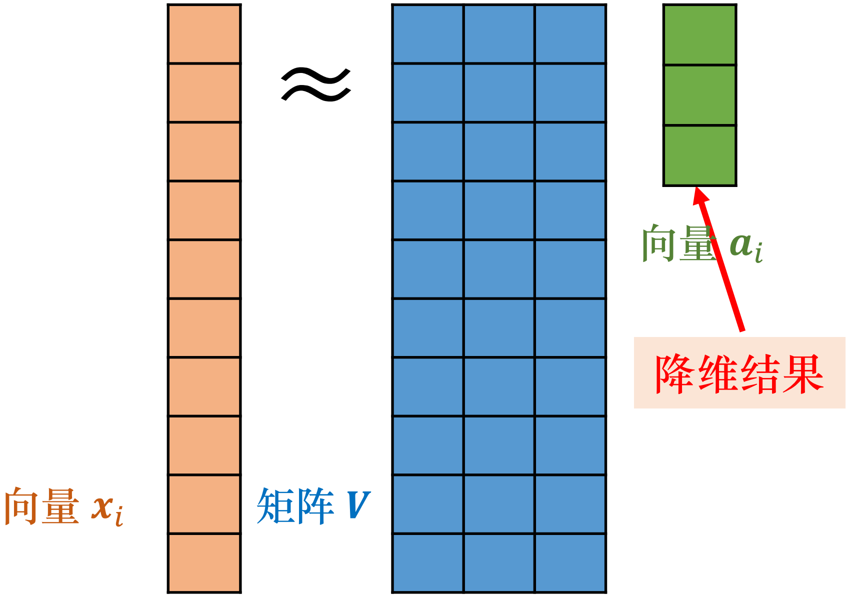
主成分分析（PCA）



JL转换



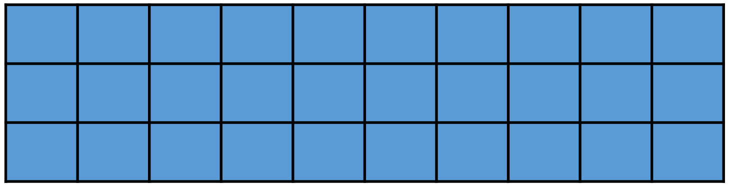
PCA与JL转换对比



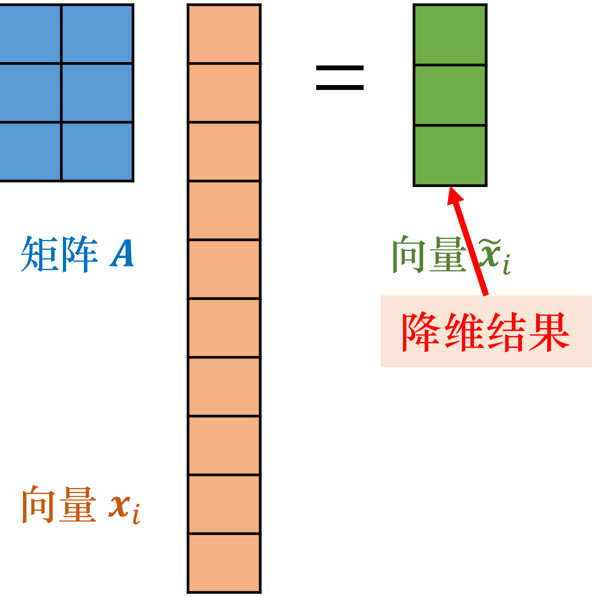
主成分分析（PCA）

1. 降维后可能无法保持原数据间的欧氏距离
2. 矩阵$\boldsymbol{V}$与数据有关，根据数据得到$\boldsymbol{V}$从而解释数据
3. 矩阵$\boldsymbol{V}$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



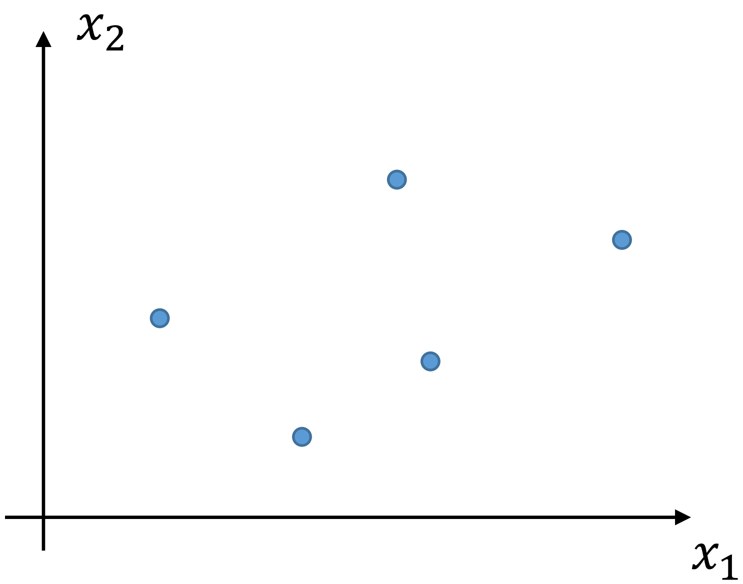
1. 降维后保持原数据间的欧氏距离
2. 矩阵$\boldsymbol{A}$与数据无关，是随机生成（每次转换结果可能不同）
3. 矩阵$\boldsymbol{A}$的行向量对应的坐标一般无意义
4. 根据JL引理，降维后的维度有下限要求（否则不能保证低失真）



PCA与线性回归对比

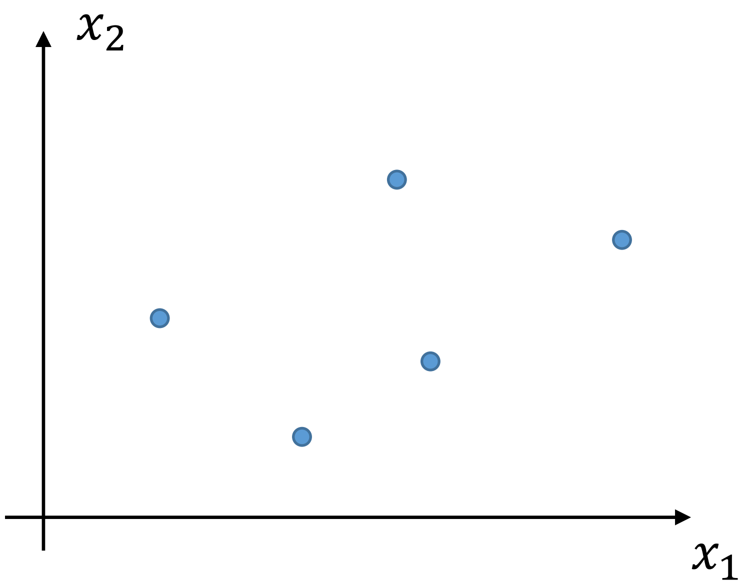
主成分分析（PCA）

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$\boldsymbol{x}_i \approx \sum_{j=1}^{k} a_{ij}\boldsymbol{v}_j = a_i\boldsymbol{v}$

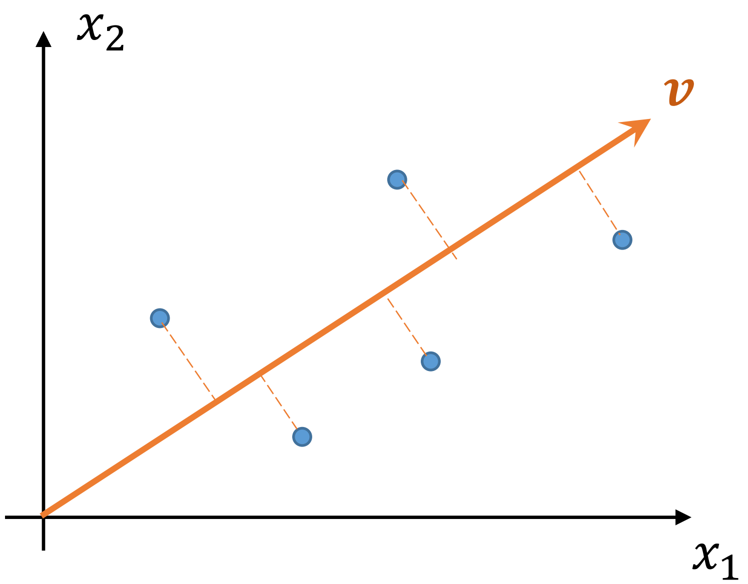
线性回归

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$\boldsymbol{x}_i \approx \sum_{j=1}^{k} a_{ij}\boldsymbol{v}_j = a_i\boldsymbol{v}$



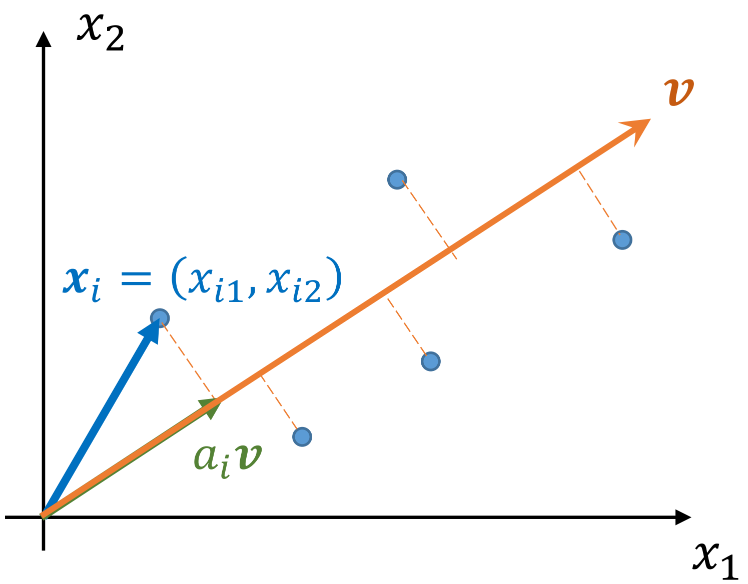
线性回归

PCA与线性回归对比

主成分分析（PCA）

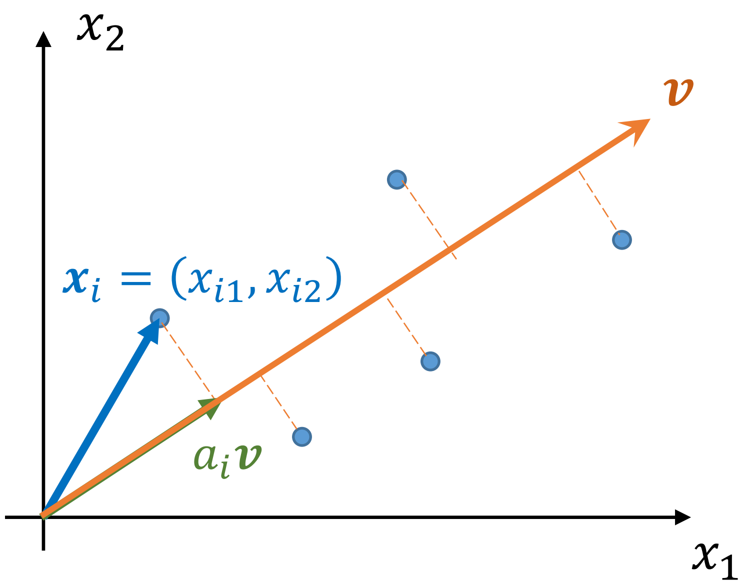
若要把若干个二维数据降维成一维数据

即$k = 1$，有$\boldsymbol{x}_i \approx \sum_{j=1}^{k} a_{ij}\,\boldsymbol{v}_j = a_i\boldsymbol{v}$



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$\boldsymbol{x}_i \approx \sum_{j=1}^{k} a_{ij}\,\boldsymbol{v}_j = a_i\boldsymbol{v}$

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

不希望$\boldsymbol{x}_i$与$a_i\boldsymbol{v}$欧式距离过大

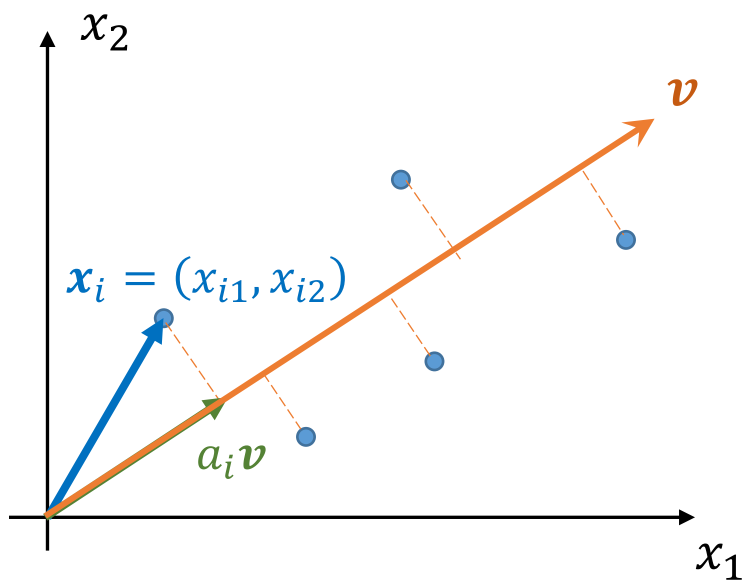
线性回归

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$\boldsymbol{x}_i \approx \sum_{j=1}^{k} a_{ij}\,\boldsymbol{v}_j = a_i\boldsymbol{v}$



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

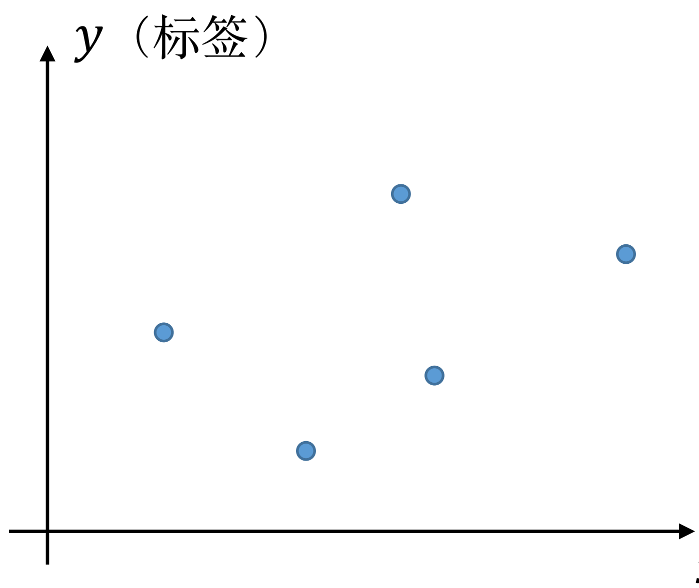
不希望$\boldsymbol{x}_i$与$a_i\boldsymbol{v}$欧式距离过大

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$f: x \to y$

使得θx_i接近于y_i

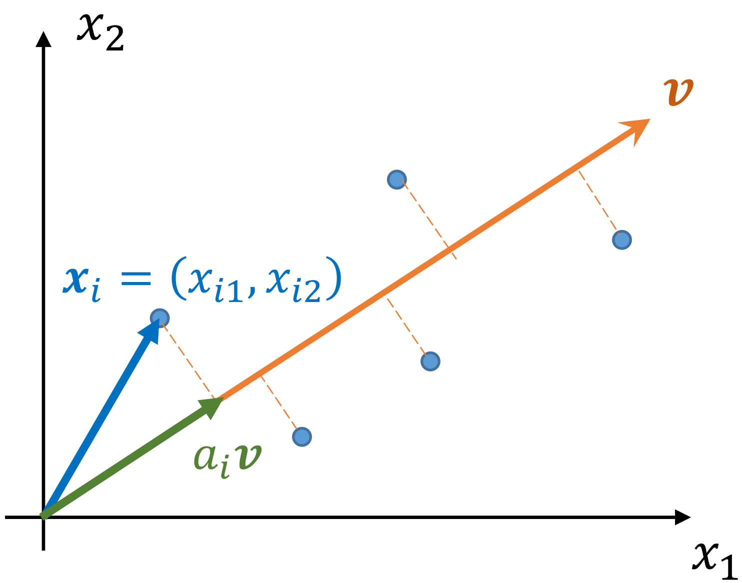


PCA与线性回归对比

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$\boldsymbol{x}_i \approx \sum_{j=1}^{k} a_{ij}\,\boldsymbol{v}_j = a_i\boldsymbol{v}$



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

不希望$\boldsymbol{x}_i$与$a_i\boldsymbol{v}$欧式距离过大

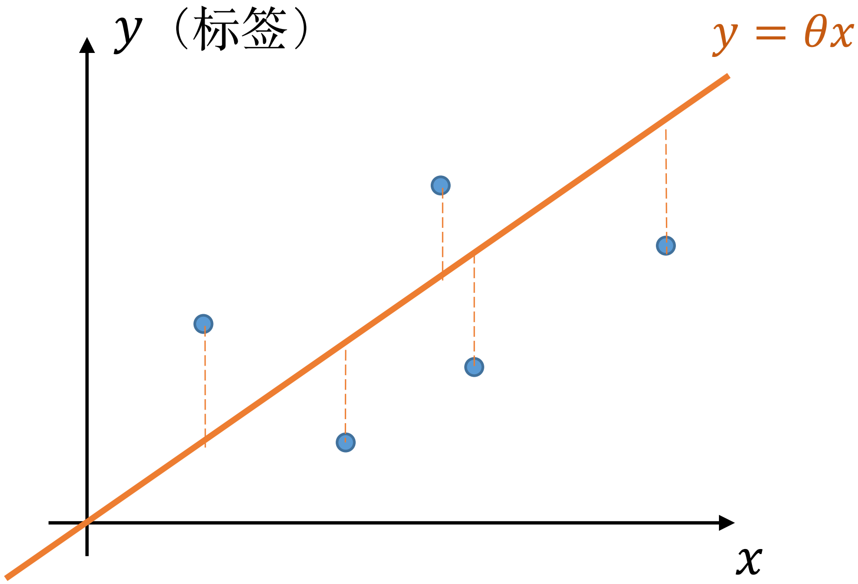
线性回归

（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$f\colon x\rightarrow y$

使得θx_i接近于 y_i

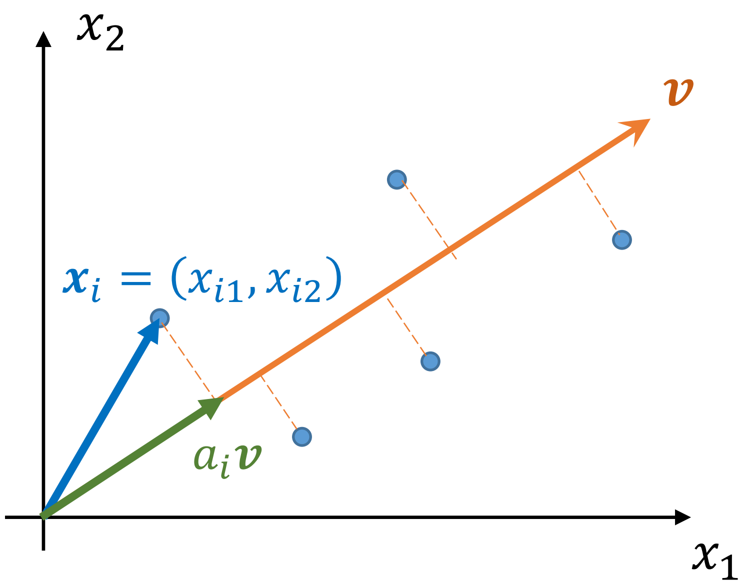


PCA与线性回归对比

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$\boldsymbol{x}_i \approx \sum_{j=1}^{k} a_{ij}\,\boldsymbol{v}_j = a_i\boldsymbol{v}$



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

不希望$\boldsymbol{x}_i$与$a_i\boldsymbol{v}$欧式距离过大

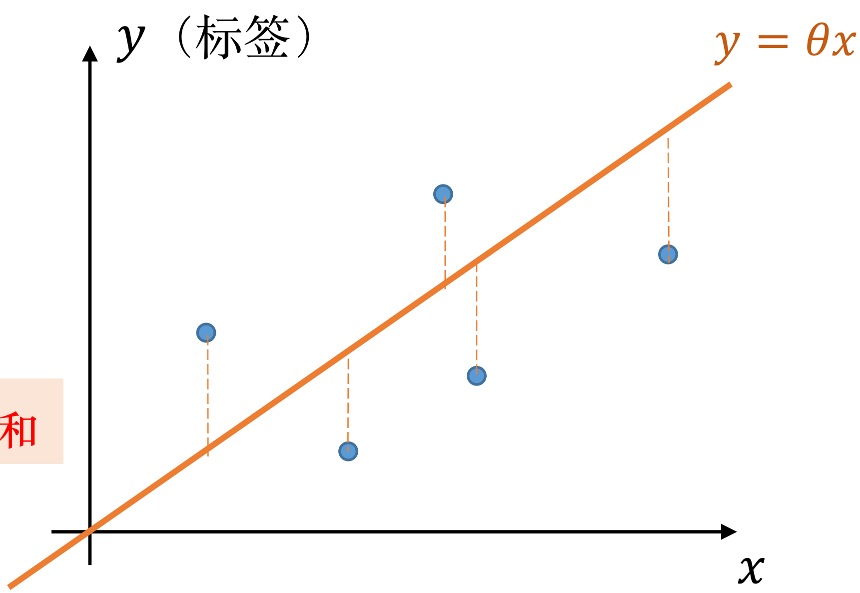
线性回归

（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$f: x \to y$

使得θx_i接近于 y_i



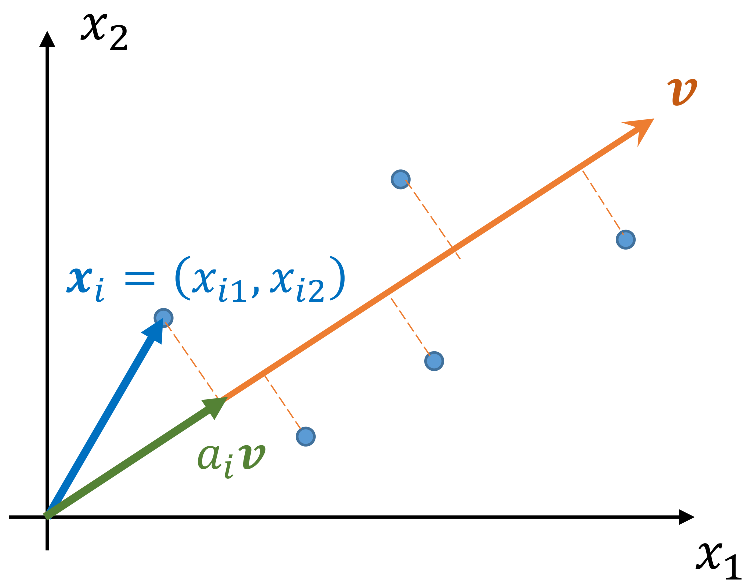
斜率θ需要满足：最小化θx_i到y_i的差的平方和

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$\boldsymbol{x}_i \approx \sum_{j=1}^{k} a_{ij}\,\boldsymbol{v}_j = a_i\boldsymbol{v}$



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

不希望$\boldsymbol{x}_i$与$a_i\boldsymbol{v}$欧式距离过大

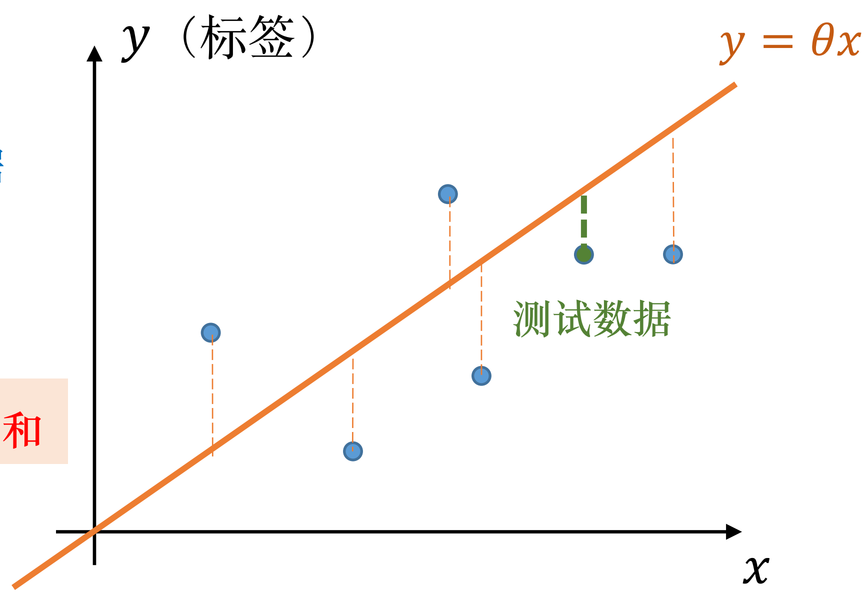
线性回归

（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θx_i接近于y_i



斜率θ需要满足：最小化θx_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$\overline{\boldsymbol{x}} = (5.5,4.5,5,5.5), \boldsymbol{v_1} = (3,-3,-3,3), \boldsymbol{v_2} = (1,-1,1,-1)$，

可以将$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$近似为$\overline{\boldsymbol{x}} + a_{i1}\boldsymbol{v_1} + a_{i2}\boldsymbol{v_2}$的形式，

由于$\overline{\boldsymbol{x}} = \frac{1}{4}(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3 + \boldsymbol{x}_4)$，等价于将$\boldsymbol{x}_1 - \overline{\boldsymbol{x}}, \boldsymbol{x}_2 - \overline{\boldsymbol{x}}, \boldsymbol{x}_3 - \overline{\boldsymbol{x}}, \boldsymbol{x}_4 - \overline{\boldsymbol{x}}$近似为$a_{i1}\boldsymbol{v_1} + a_{i2}\boldsymbol{v_2}$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$\overline{\boldsymbol{x}} = (5.5,4.5,5,5.5), \boldsymbol{v_1} = (3,-3,-3,3), \boldsymbol{v_2} = (1,-1,1,-1)$，

可以将$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$近似为$\overline{\boldsymbol{x}} + a_{i1}\boldsymbol{v_1} + a_{i2}\boldsymbol{v_2}$的形式，

由于$\overline{\boldsymbol{x}} = \frac{1}{4}(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3 + \boldsymbol{x}_4)$，等价于将$\boldsymbol{x}_1 - \overline{\boldsymbol{x}}, \boldsymbol{x}_2 - \overline{\boldsymbol{x}}, \boldsymbol{x}_3 - \overline{\boldsymbol{x}}, \boldsymbol{x}_4 - \overline{\boldsymbol{x}}$近似为$a_{i1}\boldsymbol{v_1} + a_{i2}\boldsymbol{v_2}$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

平移

	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$\overline{\boldsymbol{x}} = (5.5,4.5,5,5.5)$

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$\overline{\boldsymbol{x}} = \frac{1}{m}(\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_m)$，为每个数据$\boldsymbol{x}_i$计算$\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}$，然后对$\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_m$做降维

即有$\widetilde{\boldsymbol{x}}_i \approx \sum_{j=1}^{k} a_{ij}\, \boldsymbol{v}_j$以及$\boldsymbol{x}_i \approx \overline{\boldsymbol{x}} + \sum_{j=1}^{k} a_{ij}\, \boldsymbol{v}_j$

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

平移

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

平移

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

平移

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

平移

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$\tilde{\boldsymbol{x}}_i$到$\sum_{j=1}^{k} a_{ij}\, \boldsymbol{v}_j$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$\tilde{\boldsymbol{x}}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{in})$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

缩放

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

缩放

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$\tilde{\boldsymbol{x}}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{in})$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

缩放

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

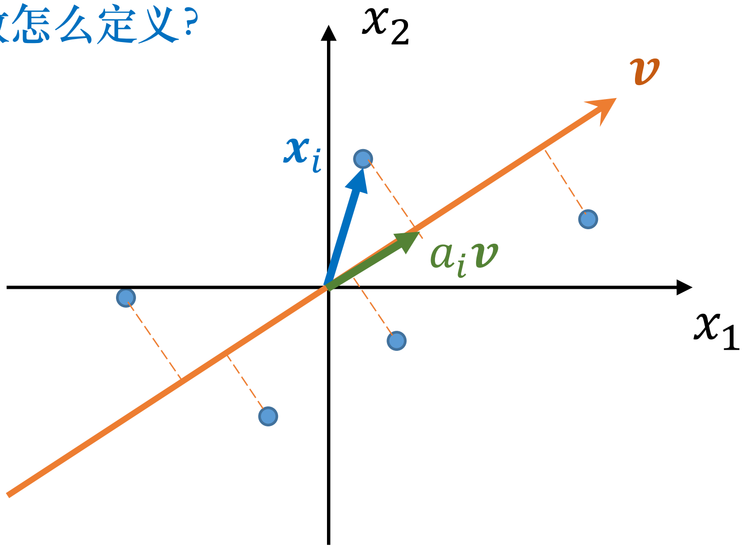
缩放

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$\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in \mathbb{R}^n$近似为$a_i\boldsymbol{v}, i = 1, \dots, m$。

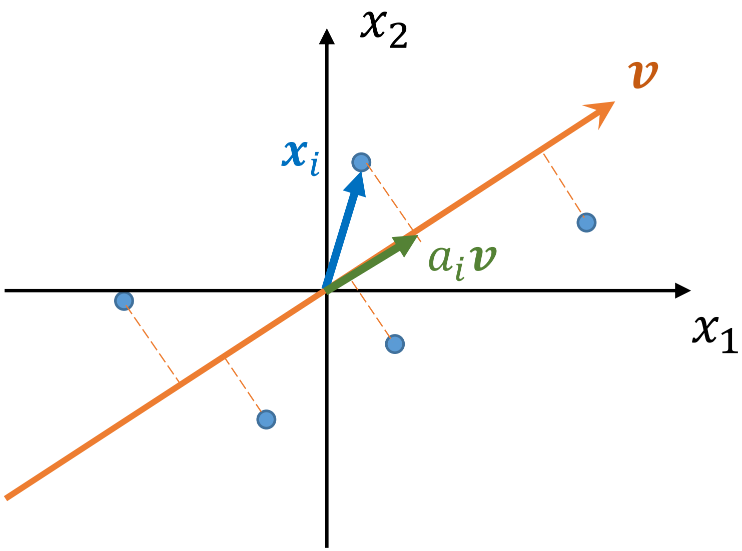
如何选择向量$\boldsymbol{v}$？ 目标函数怎么定义?



$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？



注意是距离平方和

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

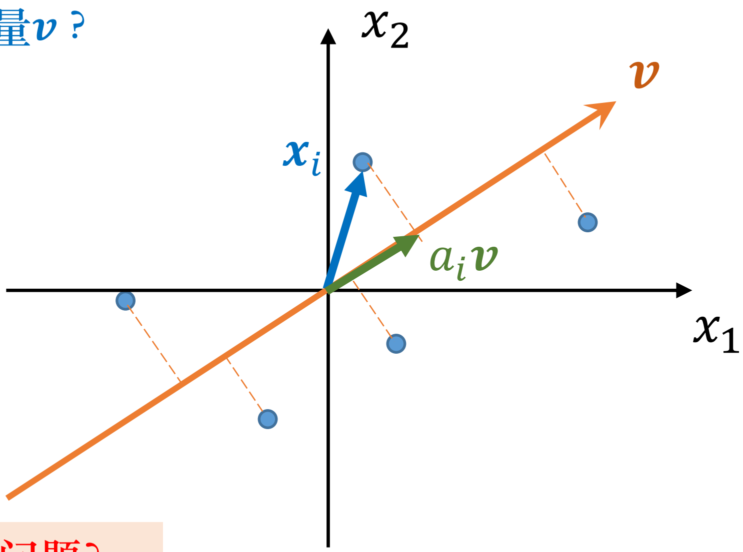
单位向量

公式取自Tim Roughgarden and Gregory Valiant <CS 168 - The Modern Algorithmic Toolbox>

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？



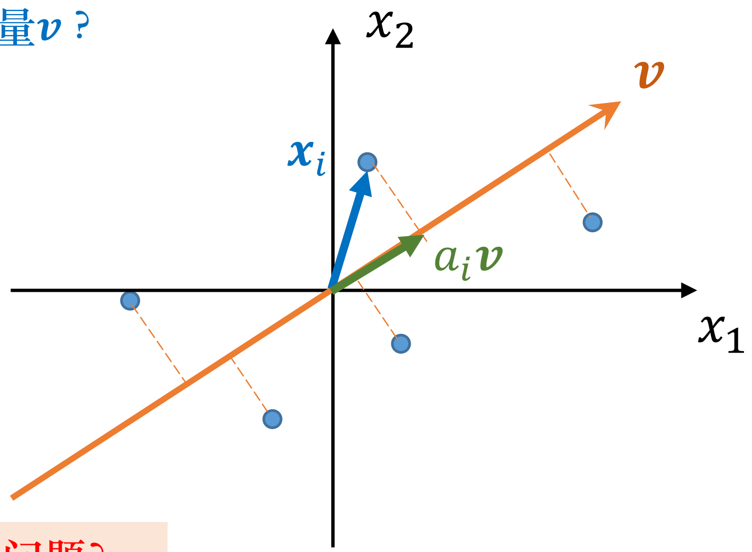
可否进一步改写问题?

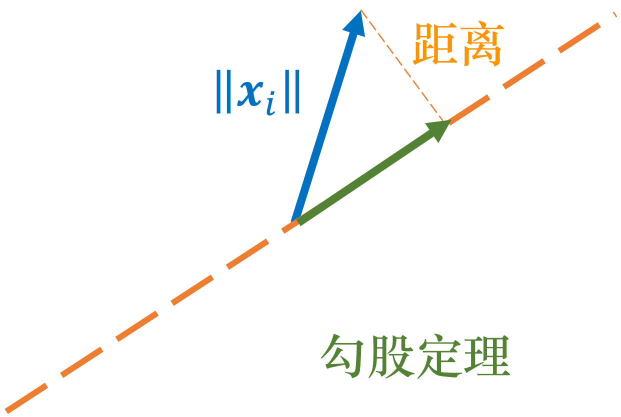
$$\underset{\mathbf{v}\,:\,\|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m}\left((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2\right)$$

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？





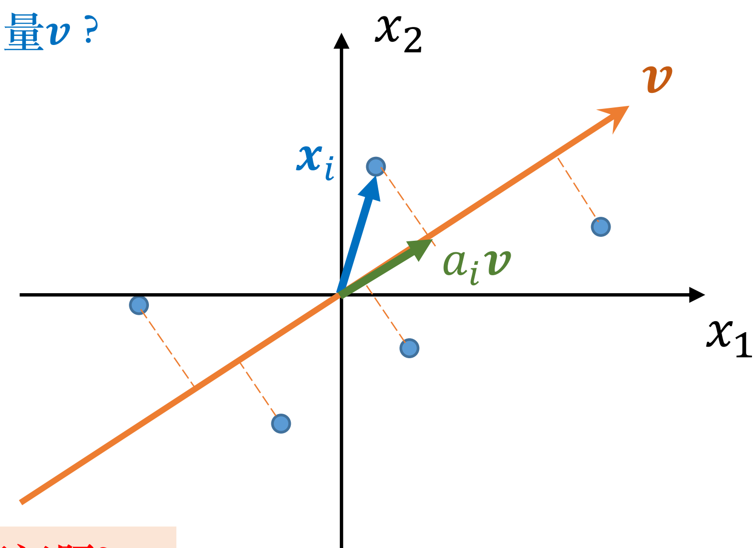
可否进一步改写问题?

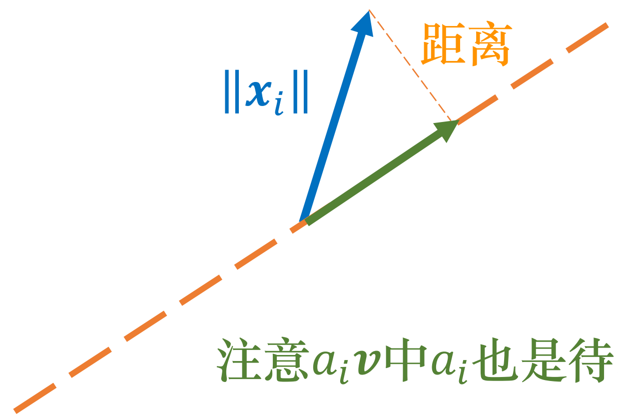
$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m}\left((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2\right)$$

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？





注意$a_i\boldsymbol{v}$中a_i也是待求量

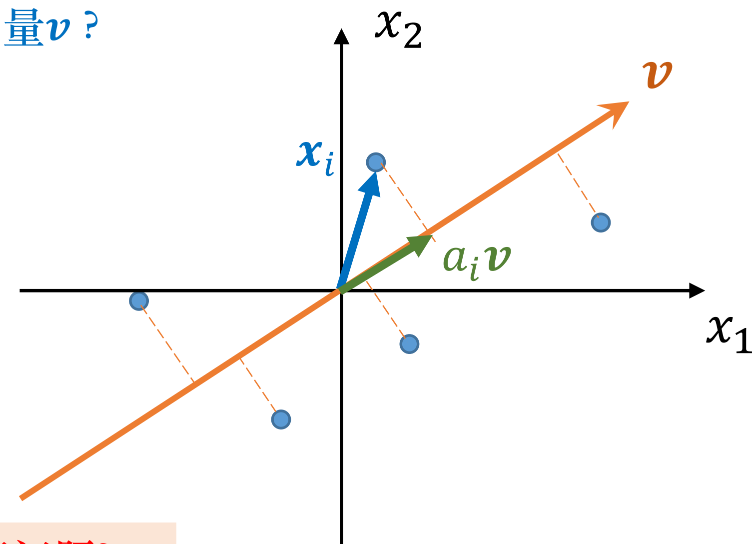
可否进一步改写问题?

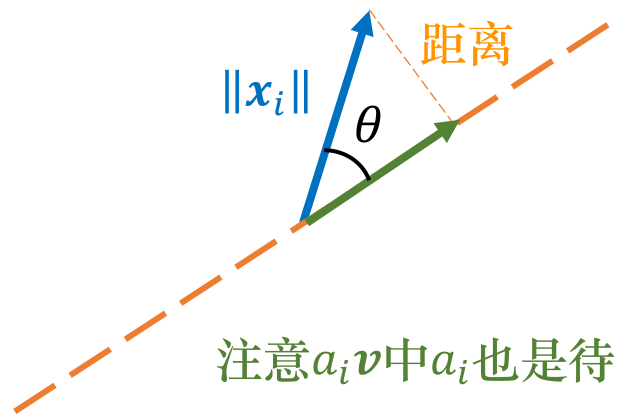
$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m}\left((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2\right)$$

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？





注意$a_i\boldsymbol{v}$中a_i也是待求量

$$a_i = \|\boldsymbol{x}_i\| \cos\theta = \|\boldsymbol{x}_i\| \, \|\boldsymbol{v}\| \cos\theta$$

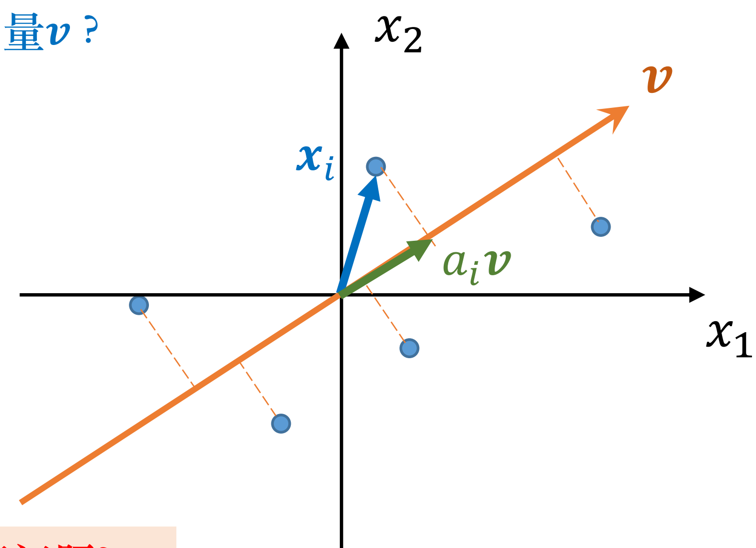
可否进一步改写问题?

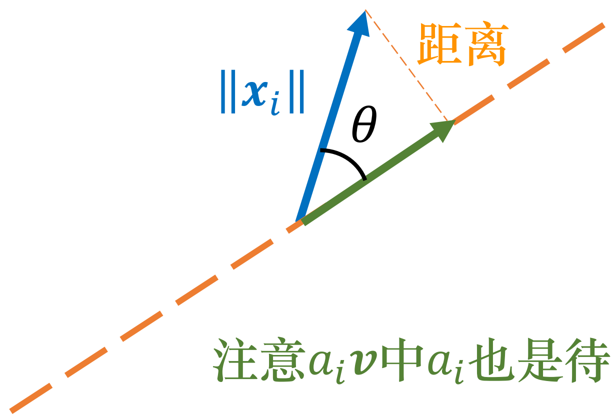
$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \left((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2 \right)$$

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？





注意$a_i\boldsymbol{v}$中a_i也是待求量

$$a_i = \|\boldsymbol{x}_i\|\cos\theta = \|\boldsymbol{x}_i\|\,\|\boldsymbol{v}\|\cos\theta$$
$$= \langle \boldsymbol{x}_i, \boldsymbol{v} \rangle = x_{i1}v_1 + \cdots + x_{in}v_n$$

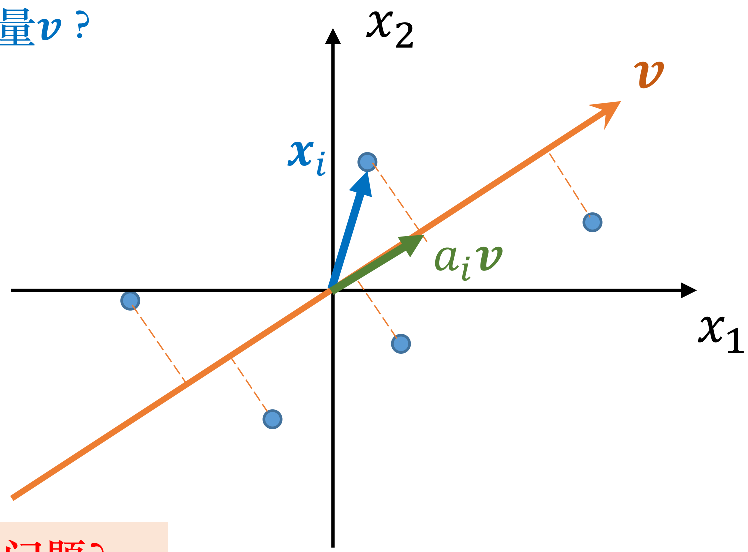
可否进一步改写问题?

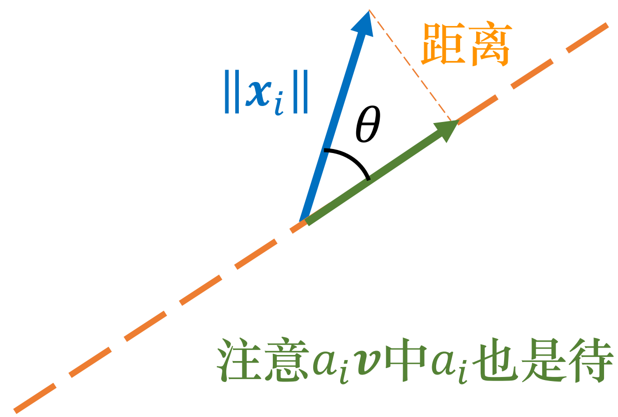
$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m}\left((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2\right)$$

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？





注意$a_i\boldsymbol{v}$中a_i也是待求量

$$a_i = \|\boldsymbol{x}_i\| \cos\theta = \|\boldsymbol{x}_i\|\,\|\boldsymbol{v}\| \cos\theta$$
$$= \langle \boldsymbol{x}_i, \boldsymbol{v} \rangle = x_{i1}v_1 + \cdots + x_{in}v_n$$

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

可否进一步改写问题?

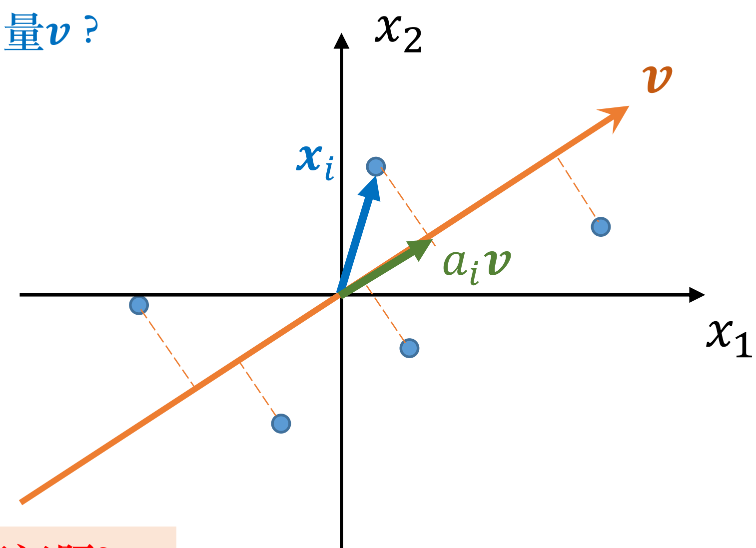
$$\boxed{\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\text{argmin}} \frac{1}{m} \sum_{i=1}^{m} \left((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2 \right)}$$

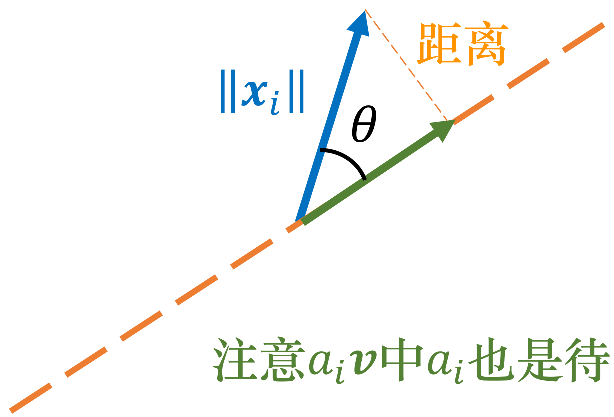
公式取自Tim Roughgarden and Gregory Valiant <CS 168 - The Modern Algorithmic Toolbox>

$k=1$情况

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

如何选择向量$\boldsymbol{v}$？





注意$a_i\boldsymbol{v}$中a_i也是待求量

$$a_i=\|\boldsymbol{x}_i\|\cos\theta=\|\boldsymbol{x}_i\|\,\|\boldsymbol{v}\|\cos\theta$$
$$=\langle\boldsymbol{x}_i,\boldsymbol{v}\rangle=x_{i1}v_1+\cdots+x_{in}v_n$$

可否进一步改写问题?

$$\operatorname*{argmin}_{\mathbf{v}:\|\mathbf{v}\|=1}\frac{1}{m}\sum_{i=1}^{m}\left((\text{distance between }\mathbf{x}_i\text{ and line spanned by }\mathbf{v})^2\right)\iff\operatorname*{argmax}_{\mathbf{v}:\|\mathbf{v}\|=1}\frac{1}{m}\sum_{i=1}^{m}\langle\mathbf{x}_i,\mathbf{v}\rangle^2$$

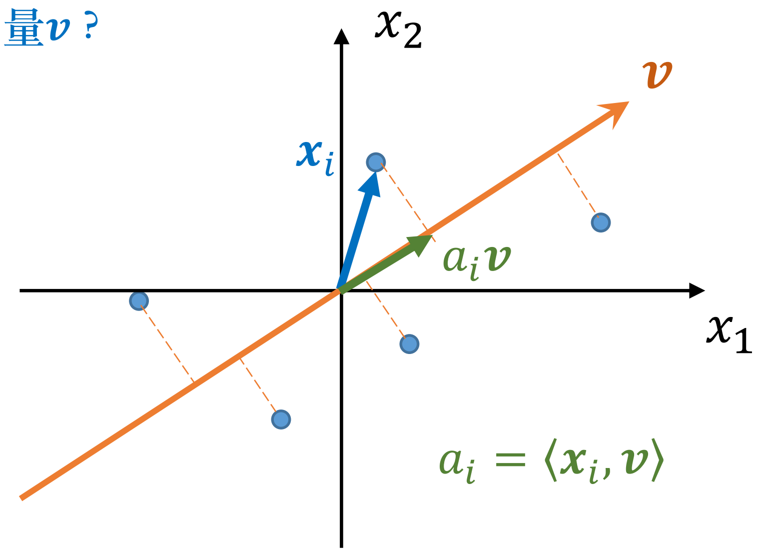
因为$\|\boldsymbol{x}_i\|^2$不受$\boldsymbol{v}$选择的影响，可以从求和中移去

公式取自Tim Roughgarden and Gregory Valiant <CS 168 - The Modern Algorithmic Toolbox>

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？



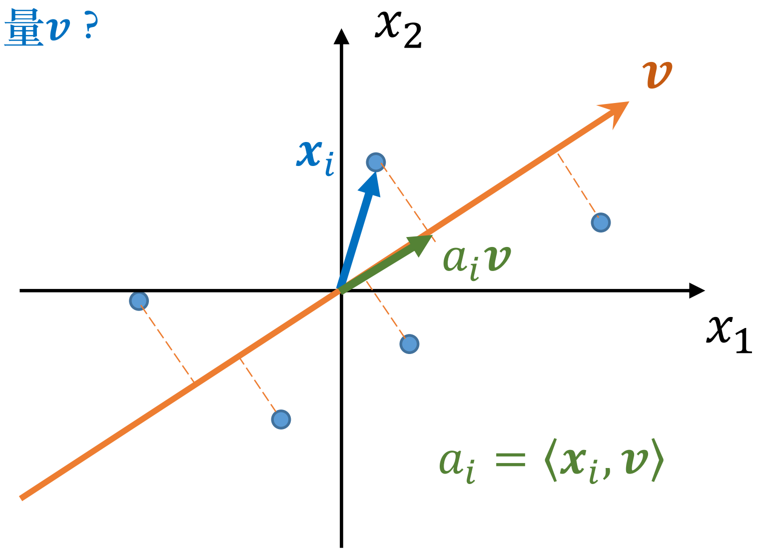
分析新问题

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

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？



分析新问题

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

即

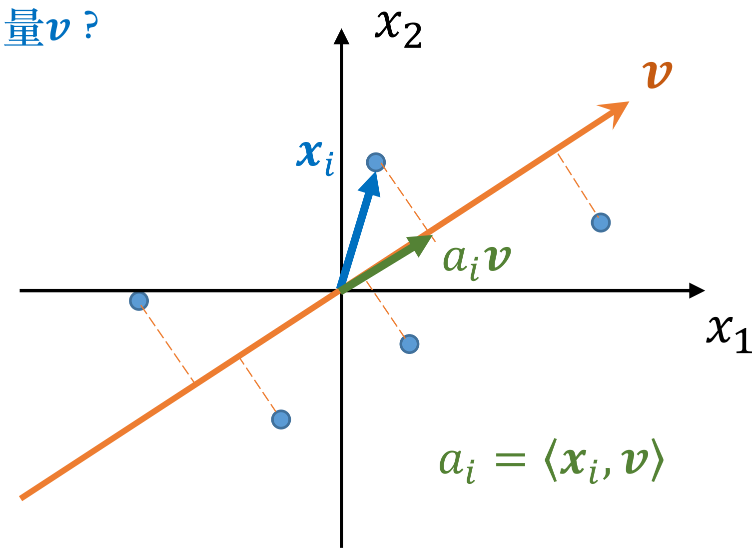
$$\underset{\boldsymbol{v}:\|\boldsymbol{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} a_i{}^2$$

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$\boldsymbol{x}_1,\dots,\boldsymbol{x}_m\in\mathbb{R}^n$近似为$a_i\boldsymbol{v}, i=1,\dots,m$。

如何选择向量$\boldsymbol{v}$？



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

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

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

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

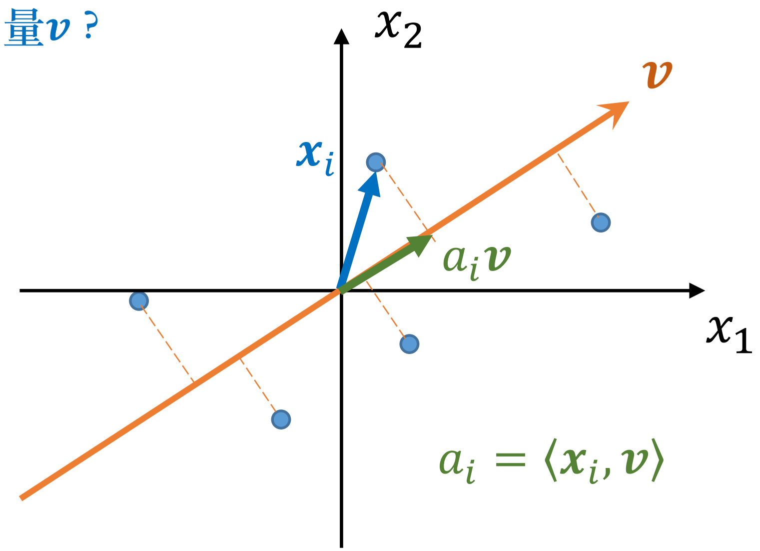
即

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

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？



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

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

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

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

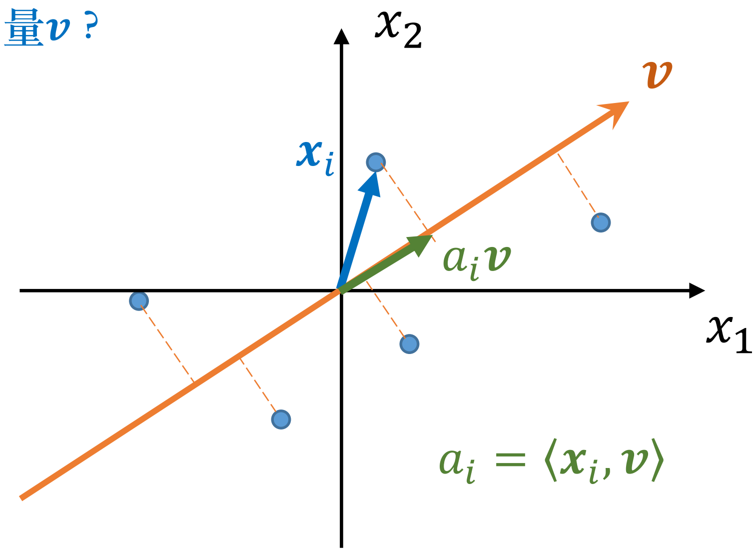
即

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

$k=1$情况

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

如何选择向量$\boldsymbol{v}$？



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

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

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

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

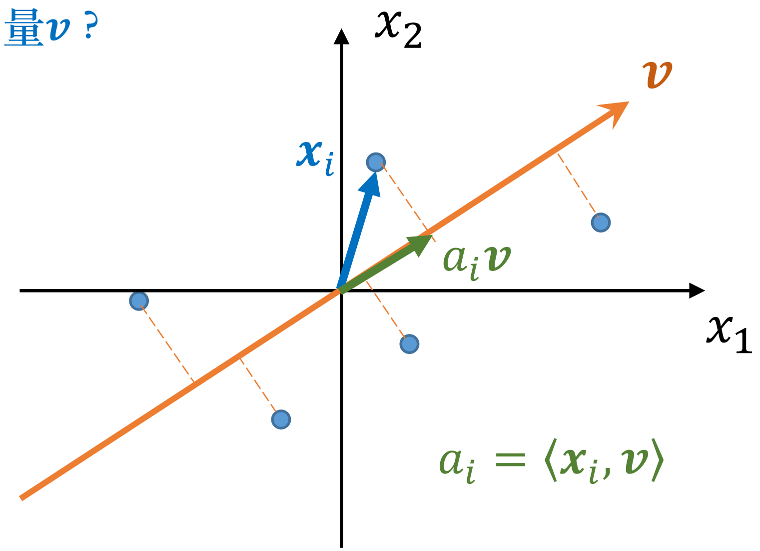
即

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

$k = 1$情况

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

如何选择向量$\boldsymbol{v}$？



分析新问题

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

即

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

新目标函数说明，需选择$\boldsymbol{v}$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

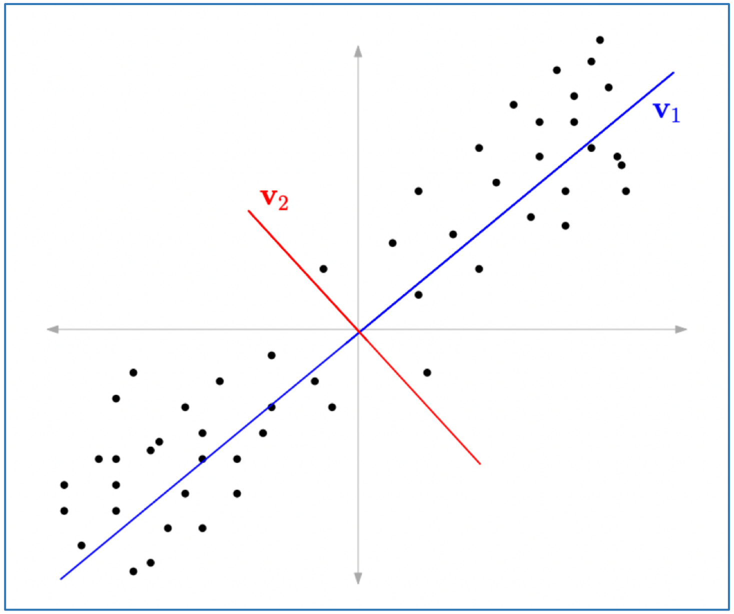
将（已完成预处理的）m个n维向量$\boldsymbol{x}_1,\dots,\boldsymbol{x}_m\in\mathbb{R}^n$近似为$a_i\boldsymbol{v}, i=1,\dots,m$。

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

即

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

需选择$\boldsymbol{v}$使降维后的数据的“方差”尽量大



选择$\boldsymbol{v}_1$好还是$\boldsymbol{v}_2$好?

图例取自Imdad Ullah Khan <Lecture Notes for Big Data Analytics - Principle Component Analysis>

$k = 1$情况

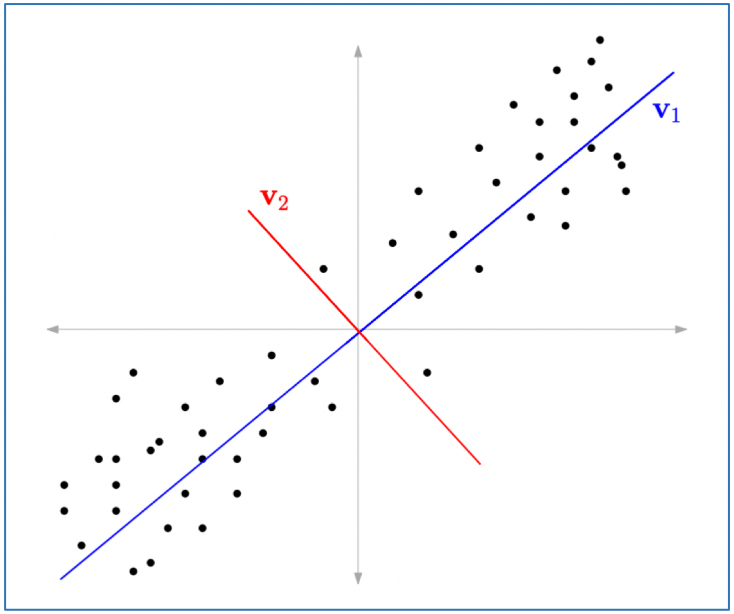
将（已完成预处理的）m个n维向量$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^n$近似为$a_i\boldsymbol{v}, i = 1, \ldots, m$。

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

即

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

需选择$\boldsymbol{v}$使降维后的数据的“方差”尽量大



选择$\boldsymbol{v}_1$更好

新目标函数角度：降维后数据“方差”更大

$k=1$情况

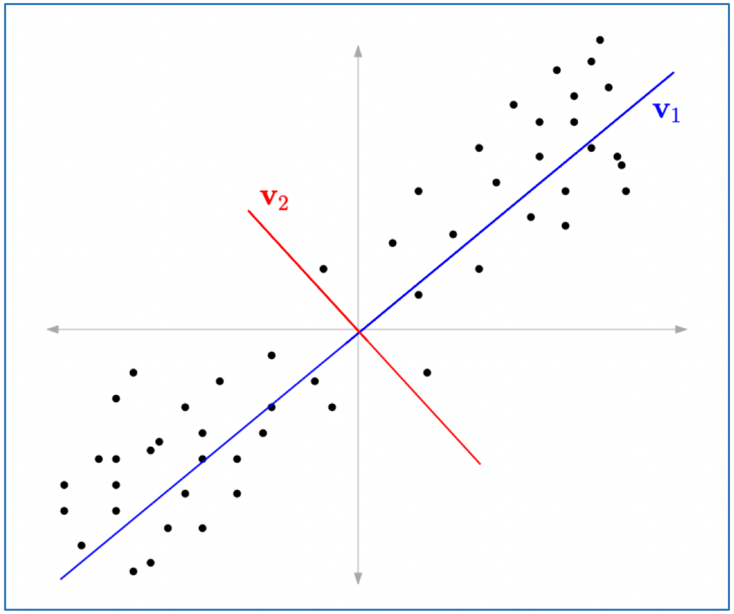
将（已完成预处理的）m个n维向量$\boldsymbol{x}_1,\dots,\boldsymbol{x}_m\in\mathbb{R}^n$近似为$a_i\boldsymbol{v}, i=1,\dots,m$。

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

即

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

需选择$\boldsymbol{v}$使降维后的数据的“方差”尽量大



选择$\boldsymbol{v}_1$更好

新目标函数角度：降维后数据“方差”更大

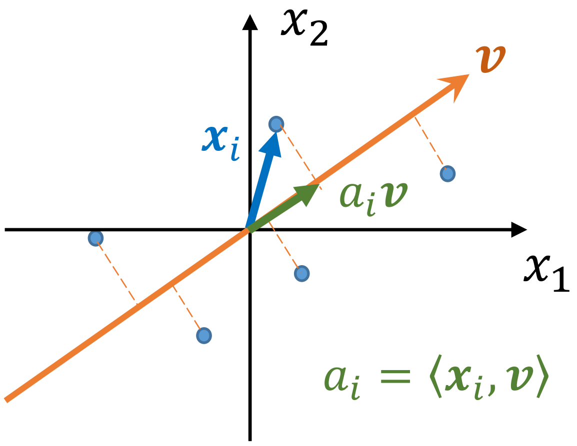
原目标函数角度：近似误差平方和更小

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

任意k情况

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

回顾$n = 2, k = 1$情况

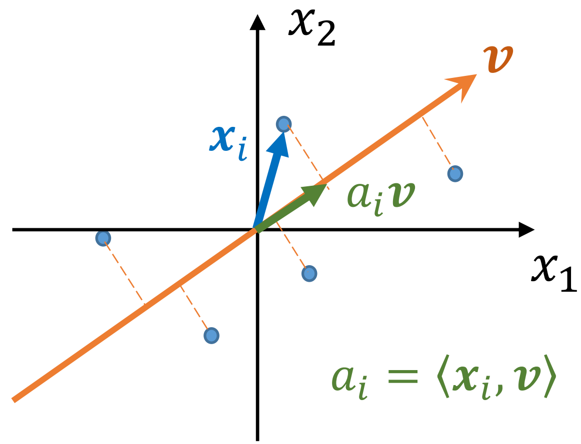


想象$n = 3, k = 2$情况

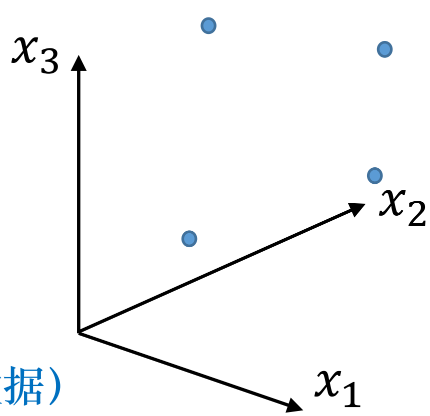
任意k情况

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

回顾$n = 2, k = 1$情况



想象$n = 3, k = 2$情况

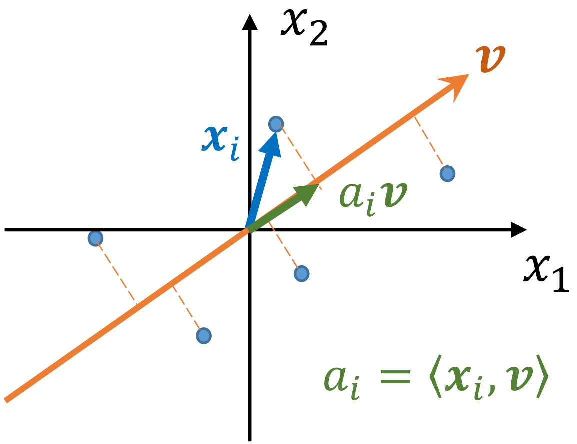


（此例中是未平移的数据）

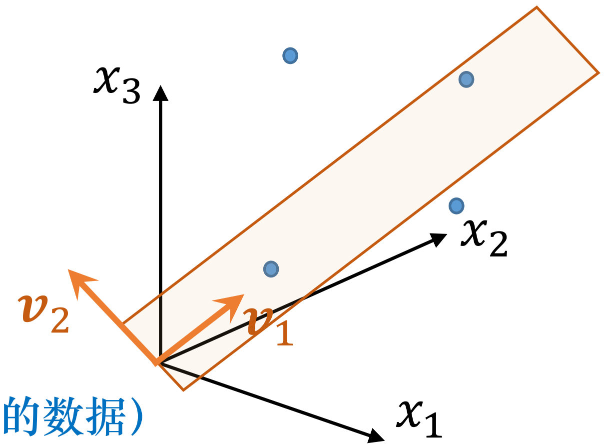
任意k情况

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

回顾$n = 2, k = 1$情况



想象$n = 3, k = 2$情况

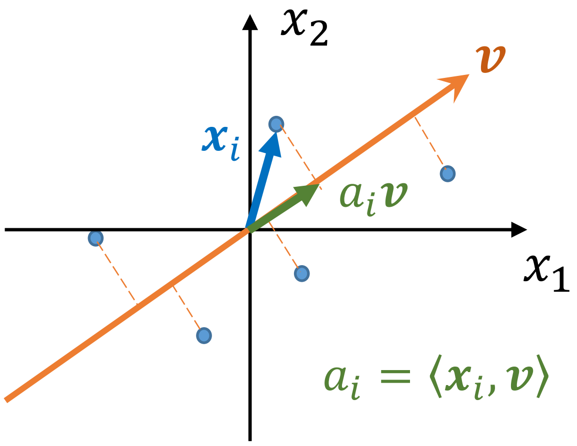


（此例中是未平移的数据）

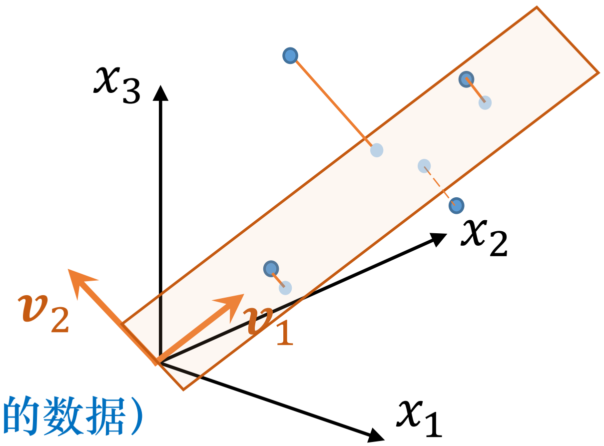
任意k情况

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

回顾$n = 2, k = 1$情况



想象$n = 3, k = 2$情况

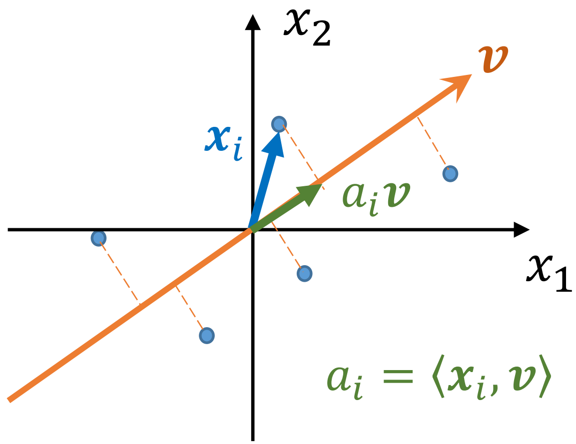


（此例中是未平移的数据）

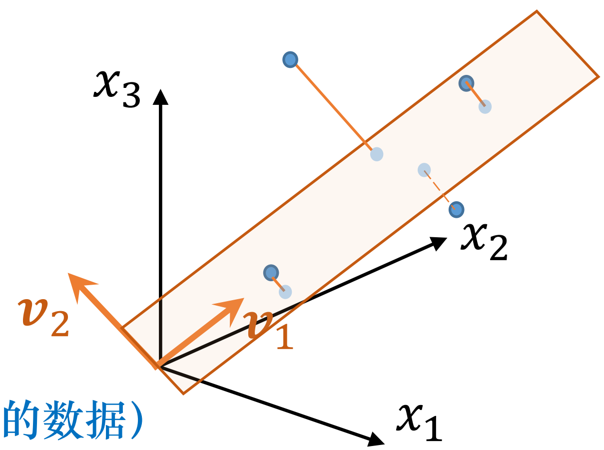
任意k情况

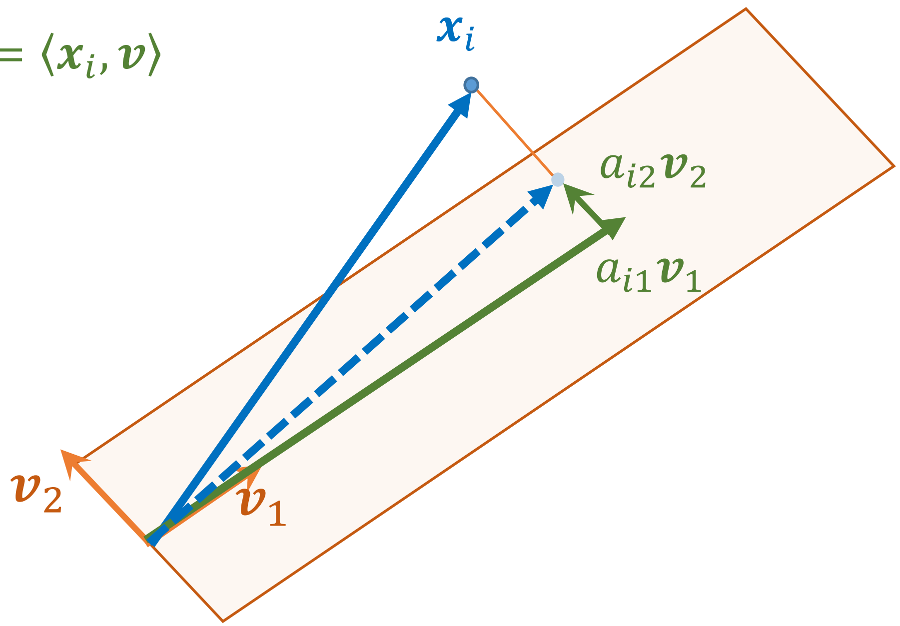
将（已完成预处理的）m个n维向量$\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\in\mathbb{R}^n$近似为$\sum_{j=1}^{k}a_{ij}\,\boldsymbol{v}_j, i=1,\ldots,m$。

回顾$n=2,k=1$情况



想象$n=3,k=2$情况

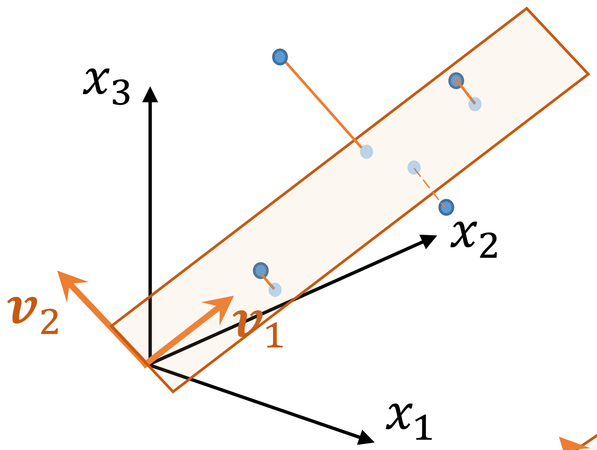


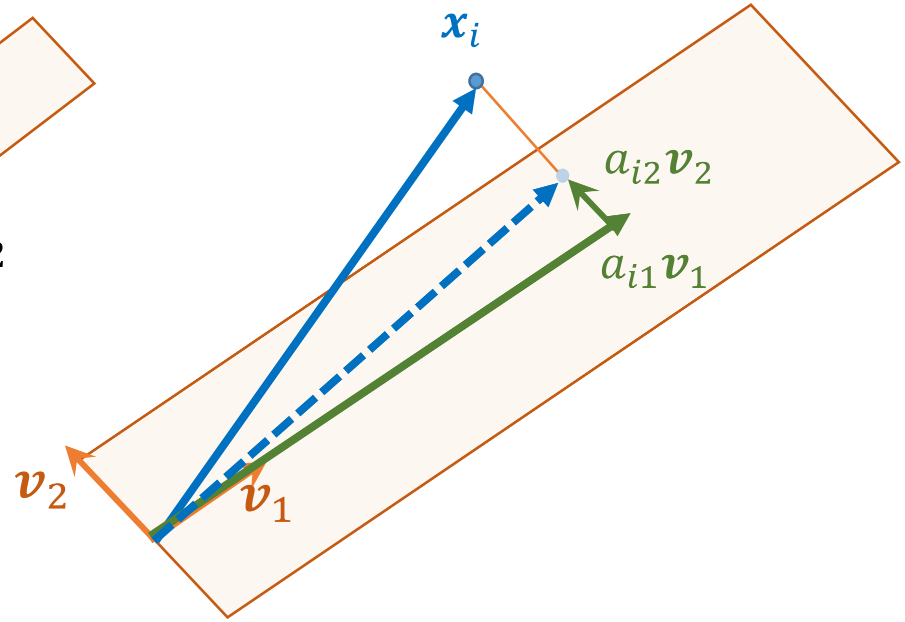


（此例中是未平移的数据）

任意k情况

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

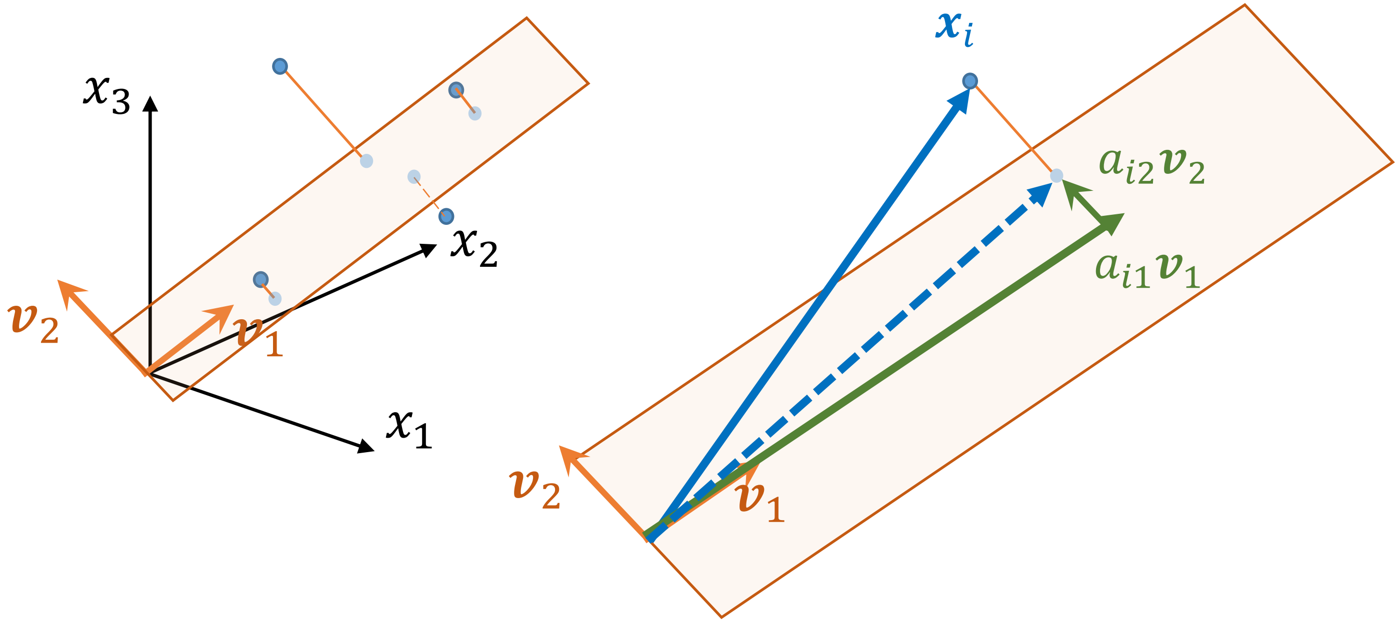




优化问题?

任意k情况

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



优化问题

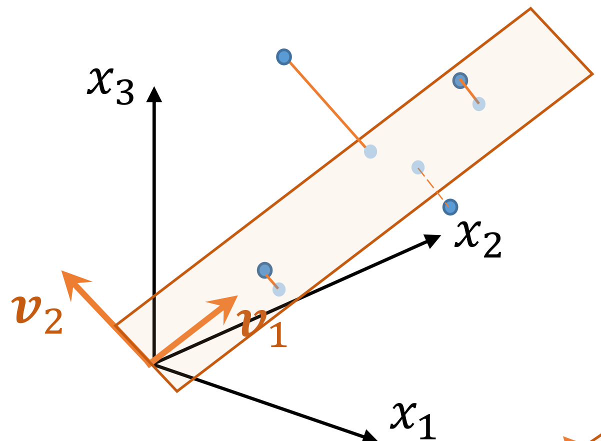
$$\operatorname{argmin} \frac{1}{m}\sum_{i=1}^{m}(distance\ between\ \boldsymbol{x}_i\ and\ k - dimensional\ subspace\ spanned\ by\ \boldsymbol{v}_1, \dots, \boldsymbol{v}_k)^2$$

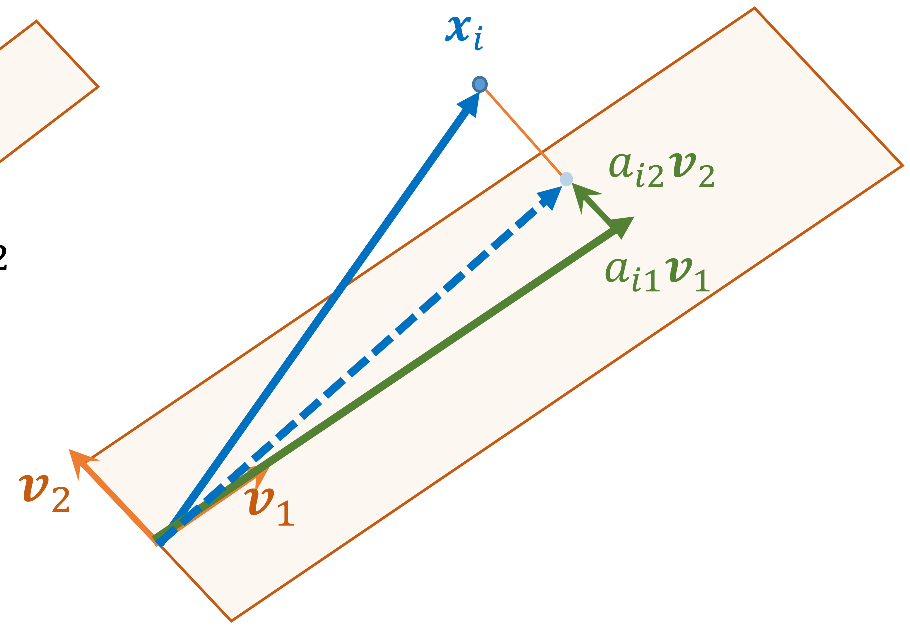
当$k = 1$时，即向量$\boldsymbol{v}$确定的直线
当$k = 2$时，即向量$\boldsymbol{v}_1, \boldsymbol{v}_2$确定的平面

对向量$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$的要求?

任意k情况

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





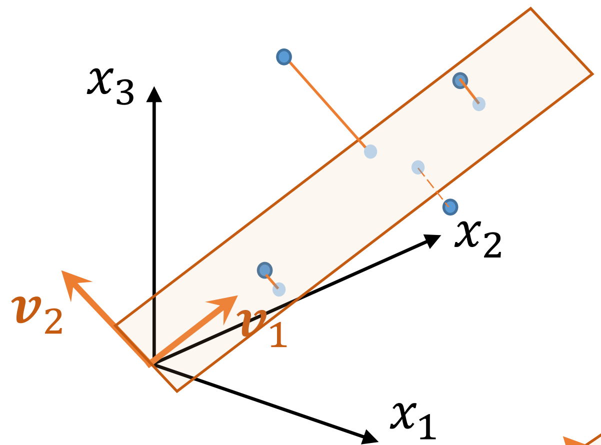
优化问题

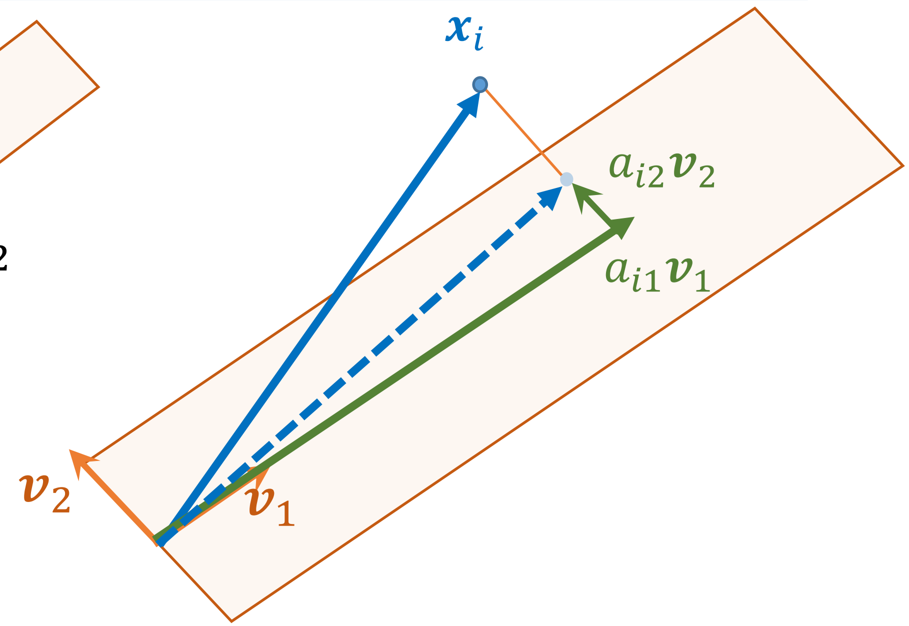
$$\text{argmin} \frac{1}{m} \sum_{i=1}^{m} (distance\ between\ \boldsymbol{x}_i\ and\ k - dimensional\ subspace\ spanned\ by\ \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)^2$$

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

任意k情况

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





优化问题

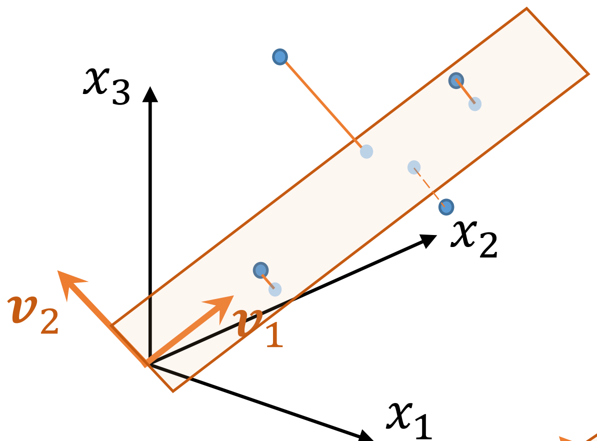
$$\operatorname{argmin} \frac{1}{m}\sum_{i=1}^{m}(distance\ between\ \boldsymbol{x}_i\ and\ k - dimensional\ subspace\ spanned\ by\ \boldsymbol{v}_1, \dots, \boldsymbol{v}_k)^2$$

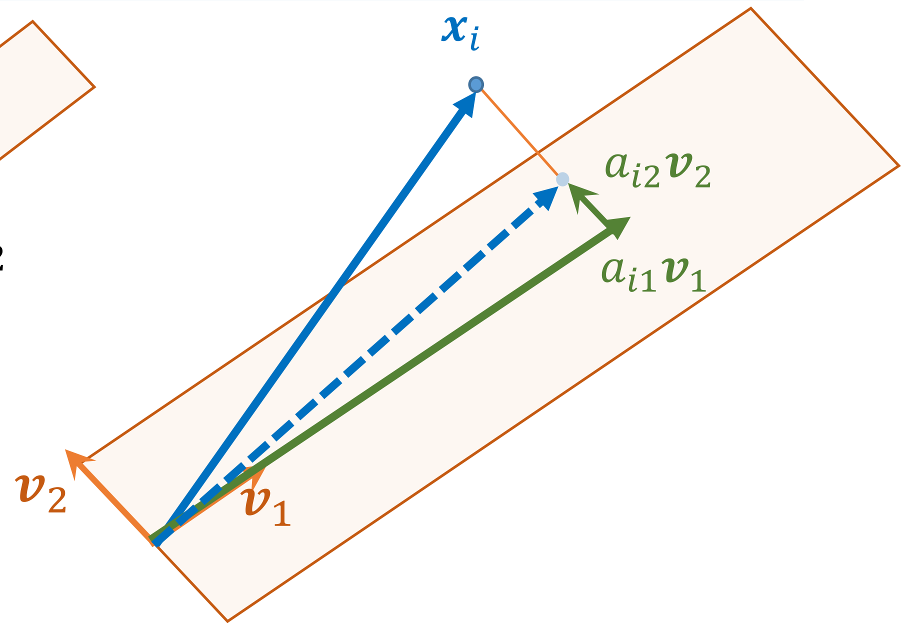
$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$要满足：（1）都是单位向量，即$\|\boldsymbol{v}_1\| = \cdots = \|\boldsymbol{v}_k\| = 1$；（2）线性独立?

如$k = 2$时，若$\boldsymbol{v}_1$和$\boldsymbol{v}_2$线性相关，说明存在常数c使得$\boldsymbol{v}_2 = c\boldsymbol{v}_1$，那么$\boldsymbol{v}_1$和$\boldsymbol{v}_2$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$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^n$近似为$\sum_{j=1}^{k} a_{ij}\,\boldsymbol{v}_j, i = 1, \ldots, m$。





优化问题

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

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

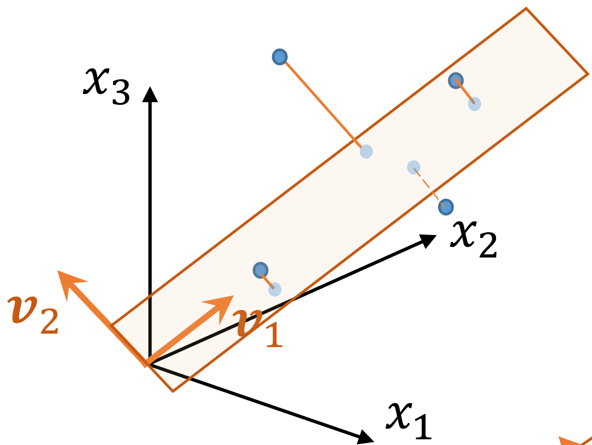
A sequence of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is said to be *linearly independent* if it is not linearly dependent, that is, if the equation

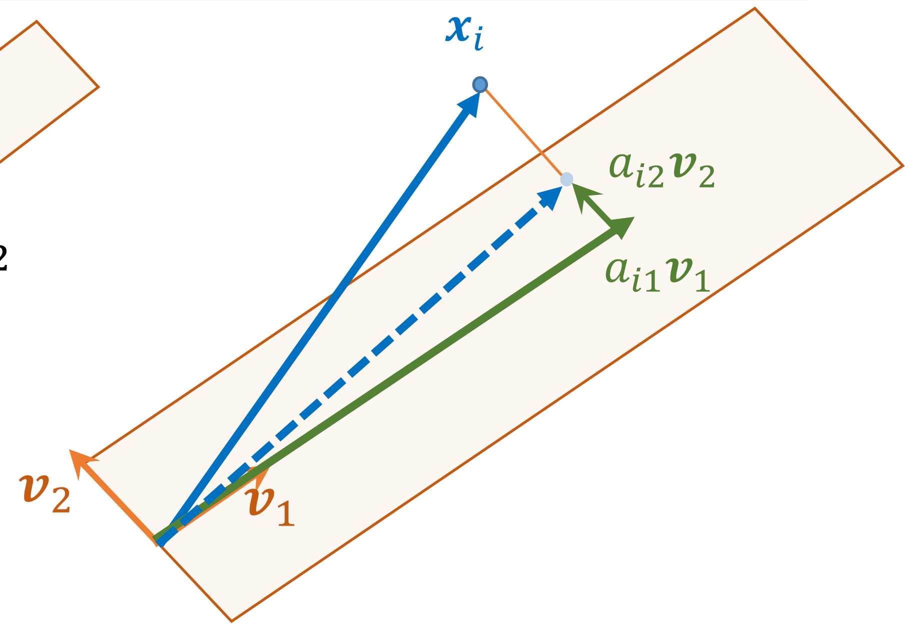
$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n = \mathbf{0},$$

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

任意k情况

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





优化问题

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

夹角90^0

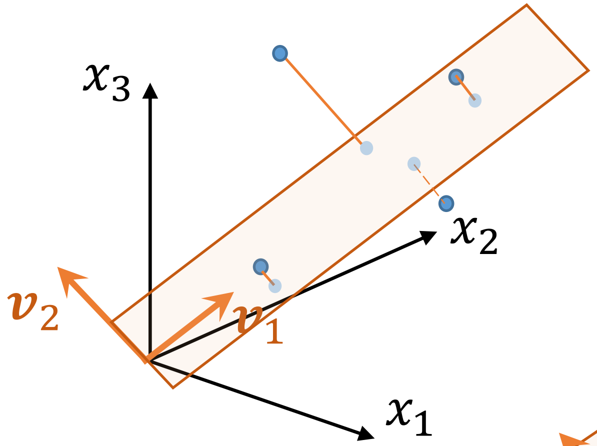
$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$要满足：（1）都是单位向量，即$\|\boldsymbol{v}_1\| = \cdots = \|\boldsymbol{v}_k\| = 1$；（2）正交，即$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = 0, \forall i \neq j$

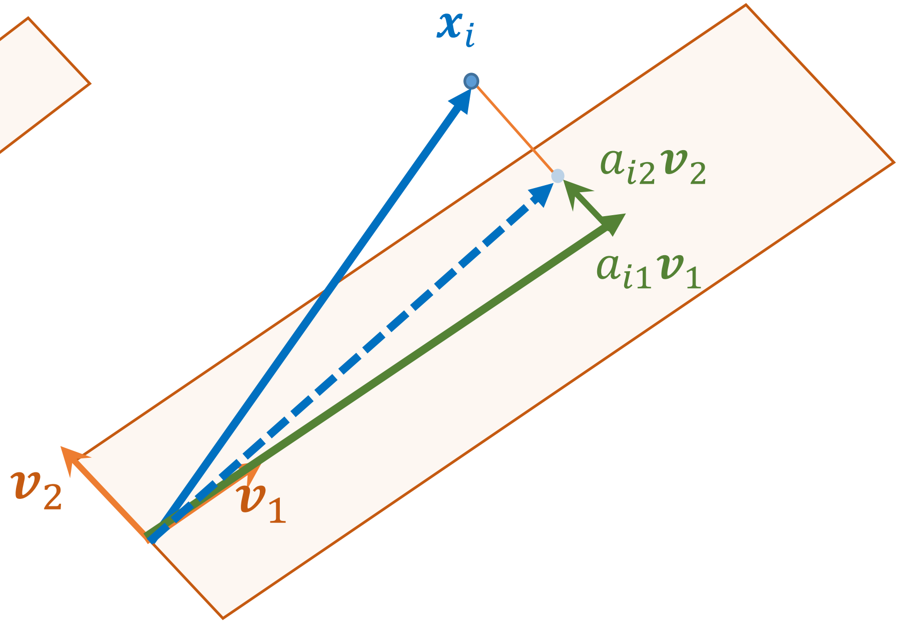
正交一定线性独立，线性独立不一定正交

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$\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in \mathbb{R}^n$近似为$\sum_{j=1}^{k} a_{ij}\, \boldsymbol{v}_j, i = 1, \dots, m$。





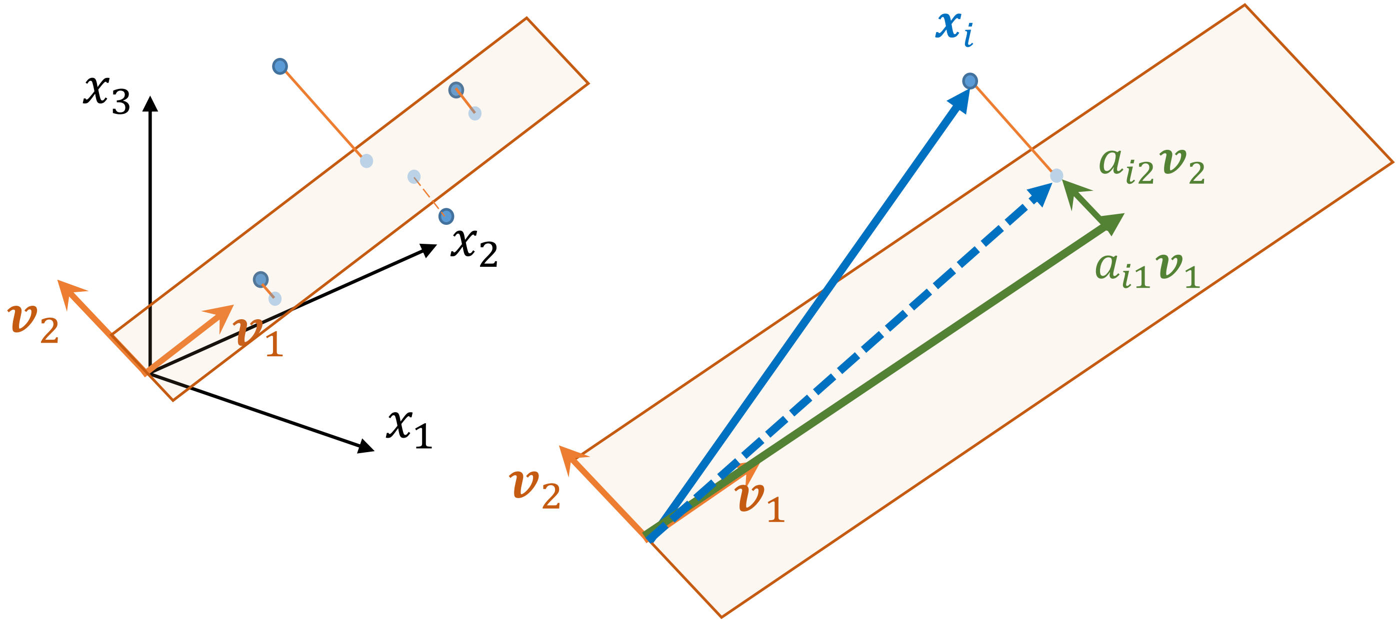
优化问题

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

$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$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$\boldsymbol{x}_1,\dots,\boldsymbol{x}_m\in\mathbb{R}^n$近似为$\sum_{j=1}^{k}a_{ij}\,\boldsymbol{v}_j, i=1,\dots,m$。



优化问题

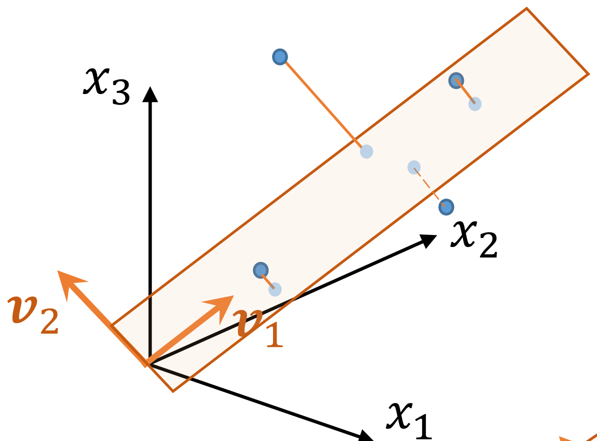
$$\operatorname{argmin}\frac{1}{m}\sum_{i=1}^{m}(\textit{distance between } \boldsymbol{x}_i \textit{ and } k-\textit{dimensional subspace spanned by } \boldsymbol{v}_1,\dots,\boldsymbol{v}_k)^2$$

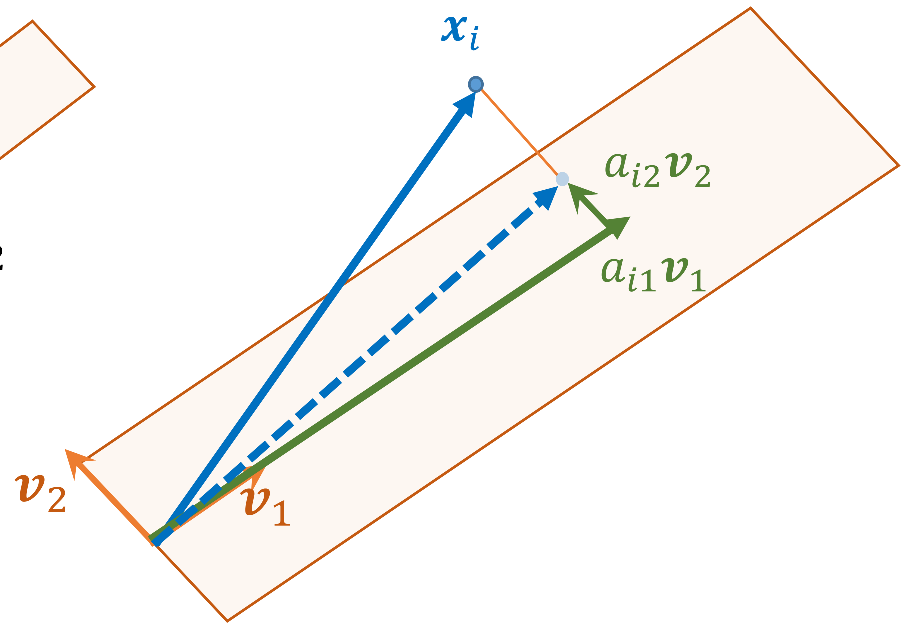
$$\operatorname{argmax}\frac{1}{m}\sum_{i=1}^{m}\left\|\sum_{j=1}^{k}a_{ij}\,\boldsymbol{v}_j\right\|^2 \quad \text{即} \quad \operatorname{argmax}\frac{1}{m}\sum_{i=1}^{m}(a_{i1}^2+\cdots+a_{ik}^2)$$

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$\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in \mathbb{R}^n$近似为$\sum_{j=1}^{k} a_{ij}\, \boldsymbol{v}_j, i = 1, \dots, m$。





优化问题

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

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

可由$a_{i1} = \langle \boldsymbol{x}_i, \boldsymbol{v}_1\rangle, \dots, a_{ik} = \langle \boldsymbol{x}_i, \boldsymbol{v}_k\rangle$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$\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in \mathbb{R}^n$及维数$k \geq 1$，求标准正交向量组$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{R}^n$从而最大化

$$\frac{1}{m}\sum_{i=1}^{m}\left(\langle \boldsymbol{x}_i, \boldsymbol{v}_1\rangle^2 + \cdots + \langle \boldsymbol{x}_i, \boldsymbol{v}_k\rangle^2\right).$$

求得的$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$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

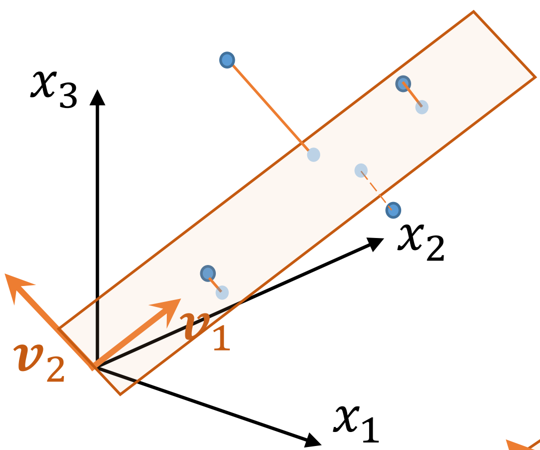
- 步骤一：确定k（如$k = 1,2,3$），通过主成分分析得到数据的k个主成分$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$

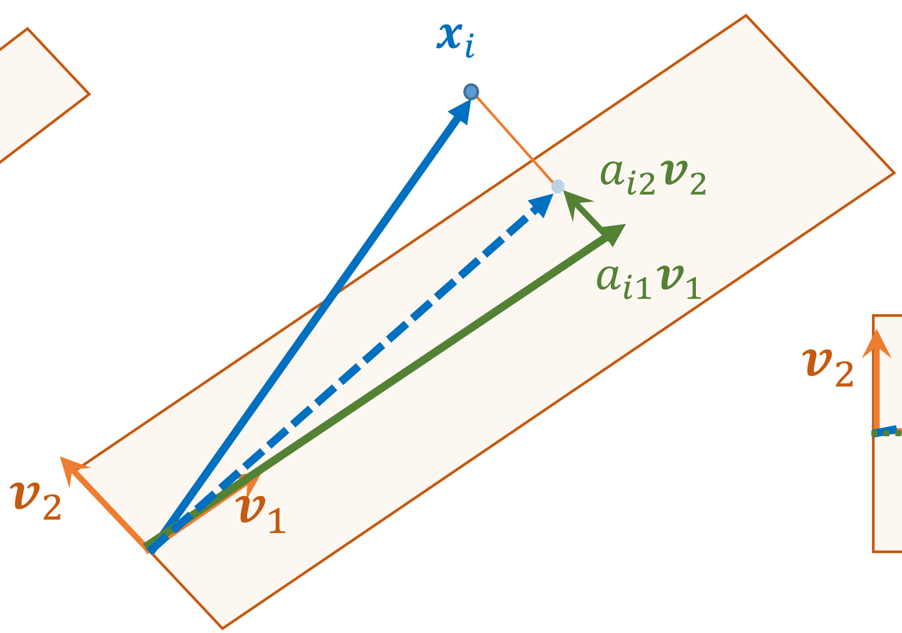
数据可视化

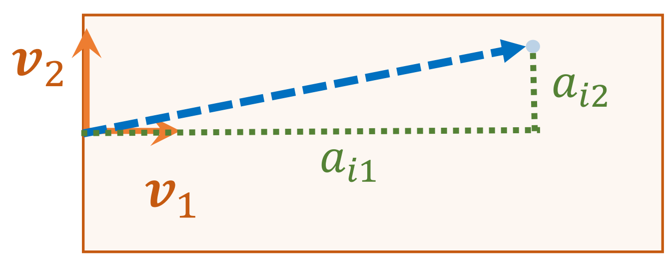
高维数据难以可视化，可通过主成分分析得到低维数据再可视化

- 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$\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$
- 步骤二：为每个数据$\boldsymbol{x}_i$，计算$a_{i1} = \langle \boldsymbol{x}_i, \boldsymbol{v}_1 \rangle, \ldots, a_{ik} = \langle \boldsymbol{x}_i, \boldsymbol{v}_k \rangle$

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





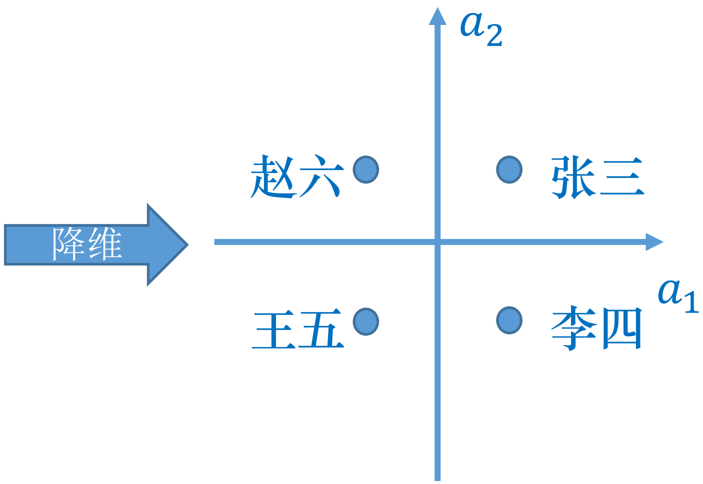


数据可视化

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$\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$
- 步骤二：为每个数据$\boldsymbol{x}_i$，计算$a_{i1} = \langle \boldsymbol{x}_i, \boldsymbol{v}_1 \rangle, \dots, a_{ik} = \langle \boldsymbol{x}_i, \boldsymbol{v}_k \rangle$
- 步骤三：在k维空间画出数据$\boldsymbol{x}_i$对应的点$(a_{i1}, \dots, a_{ik})$，即$(\langle \boldsymbol{x}_i, \boldsymbol{v}_1 \rangle, \dots, \langle \boldsymbol{x}_i, \boldsymbol{v}_k \rangle)$

	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）

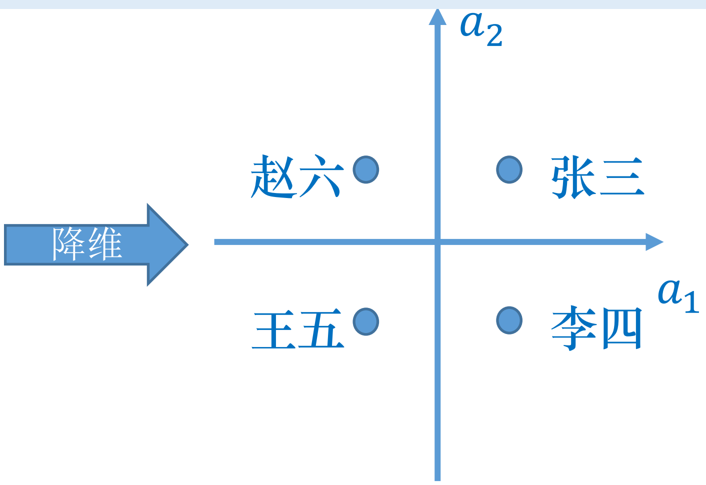
数据可视化

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$\boldsymbol{v}_1,\ldots,\boldsymbol{v}_k$
- 步骤二：为每个数据$\boldsymbol{x}_i$，计算$a_{i1}=\langle \boldsymbol{x}_i,\boldsymbol{v}_1\rangle,\ldots,a_{ik}=\langle \boldsymbol{x}_i,\boldsymbol{v}_k\rangle$
- 步骤三：在k维空间画出数据$\boldsymbol{x}_i$对应的点$(a_{i1},\ldots,a_{ik})$，即$(\langle \boldsymbol{x}_i,\boldsymbol{v}_1\rangle,\ldots,\langle \boldsymbol{x}_i,\boldsymbol{v}_k\rangle)$

分析数据：（1）解释$\boldsymbol{v}_1,\ldots,\boldsymbol{v}_k$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 99 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 …

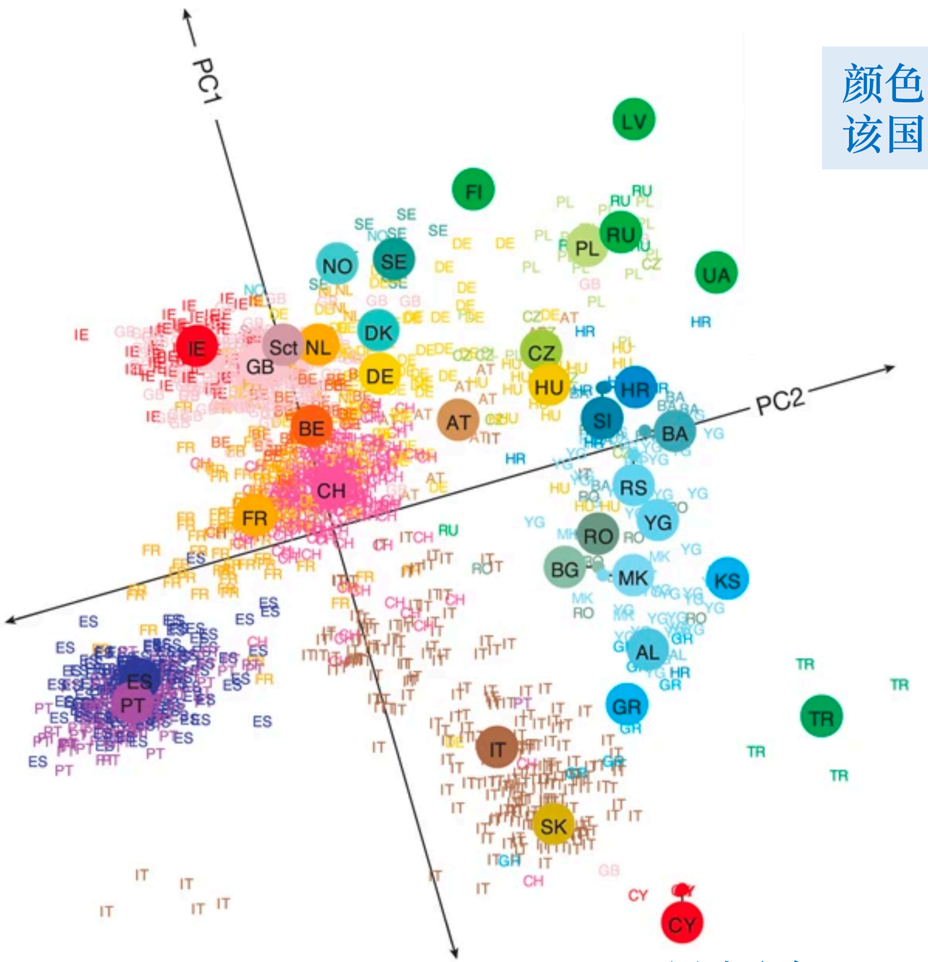
Save Cite Cited by 1596 Related articles All 53 versions

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$m = 1387, n = 200000, k = 2$

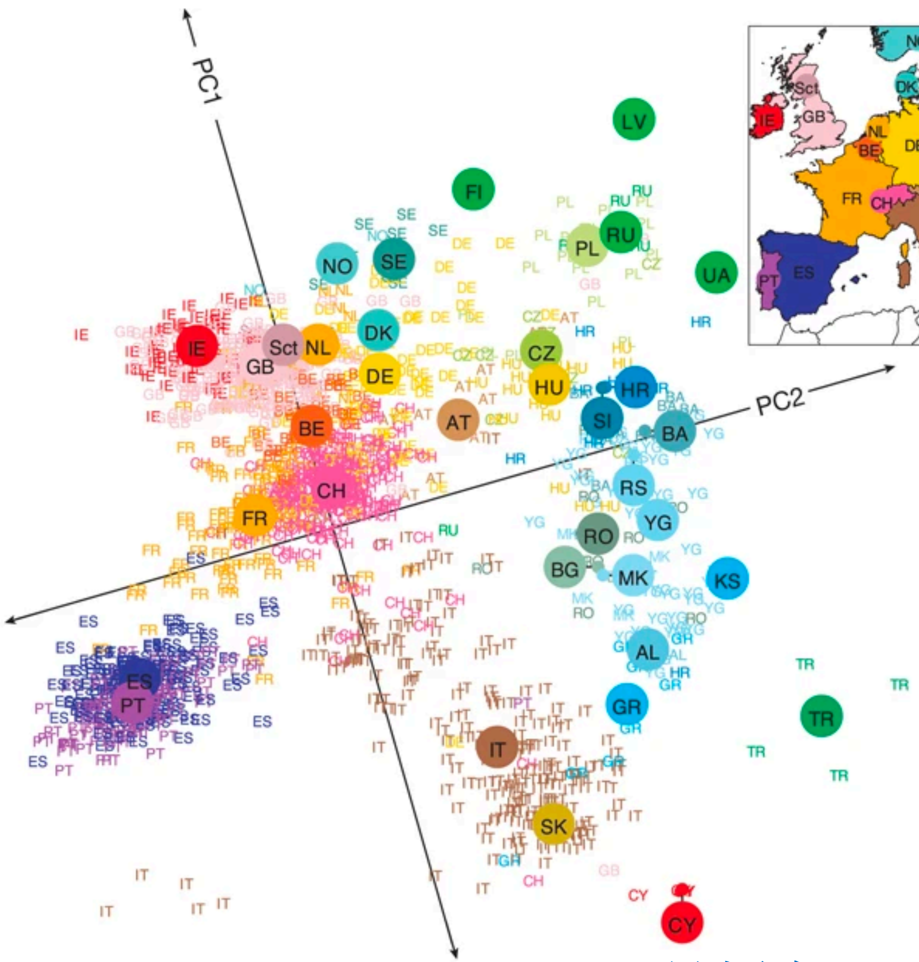


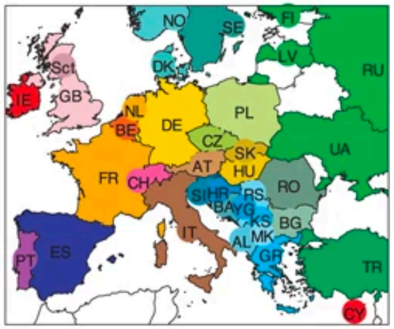
颜色对应每个数据（欧洲人）的原籍国家
该国籍信息未参与主成分分析

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$\boldsymbol{v}_1$与$\boldsymbol{v}_2$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

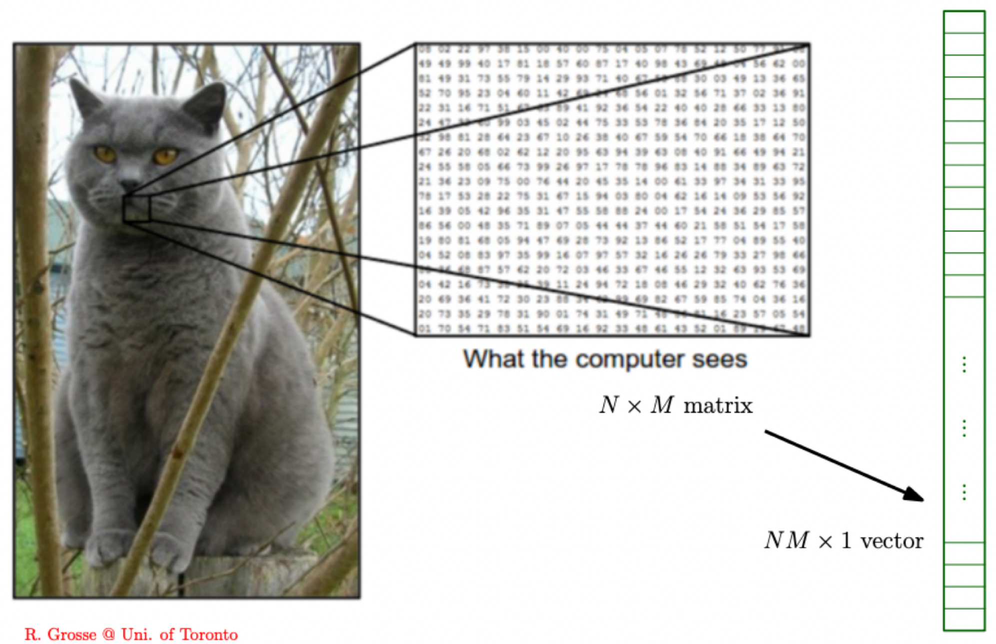
Eigenfaces for **recognition**
M Turk, A Pentland - Journal of cognitive neuroscience, 1991 - direct.mit.edu
We have developed a near-real-time computer system that can locate and track a subject's head, and then recognize the person by comparing characteristics of the face to those of …
Save Cite Cited 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

数据集



图片对应的向量

图例取自Imdad Ullah Khan <Lecture Notes for Big Data Analytics - Principle Component Analysis>及红色标注来源

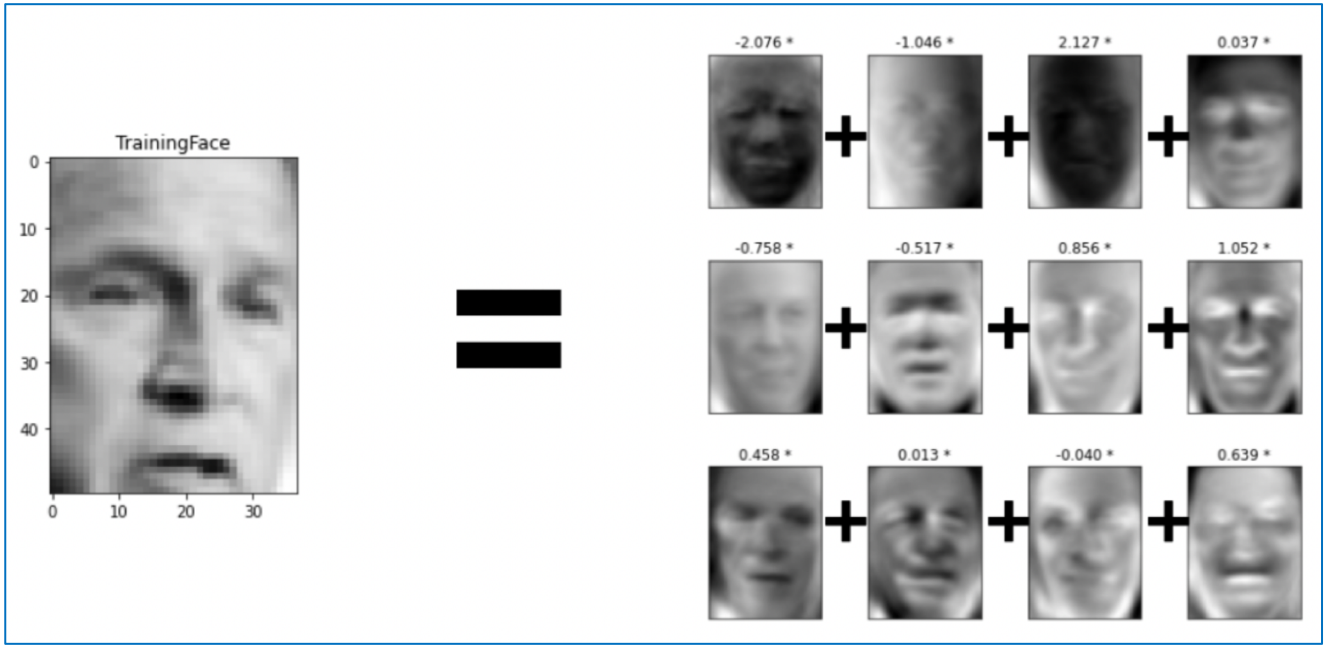
数据压缩

平均向量对应的图片

提取的主成分对应的图片

图例取自Imdad Ullah Khan <Lecture Notes for Big Data Analytics - Principle Component Analysis>

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

图例取自Imdad Ullah Khan <Lecture Notes for Big Data Analytics - Principle Component Analysis>

数据压缩

$K=5$ $K=10$ $K=20$ $K=30$ $K=50$ $K=100$ $K=200$ $K=300$ $K=400$ $K=500$ $K=600$

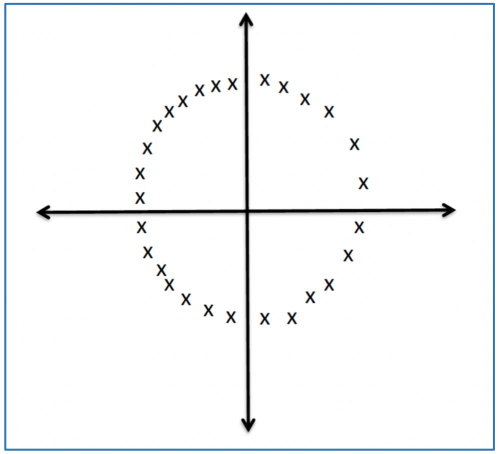
随着主成分个数的增多，近似造成的损失越少

图例取自Imdad Ullah Khan <Lecture Notes for Big Data Analytics - Principle Component Analysis>

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

图片取自Tim Roughgarden and Gregory Valiant <CS 168 - The Modern Algorithmic Toolbox>

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！