1 主成分分析

1.1 基础知识与符号约定

设有样本 $\mathbf{x}$ 和 $\mathbf{y}$

1. 样本均值

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

2. 样本方差, 这里和概率论中的定义不同, 但在本章讨论中是无关紧要的

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

3. 样本 $\mathbf{x}$ 和 $\mathbf{y}$ 的协方差

$$Cov(\mathbf{x},\mathbf{y})=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

4. 设有 p 个样本 $\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_p$, 记 s_{ij} 为 $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 的协方差, 即

$$s_{ij}=\frac{1}{n}<\mathbf{x}_i-\mu_i\mathbf{1},\mathbf{x}_j-\mu_j\mathbf{1}>=\frac{1}{n}(\mathbf{x}_i-\mu_i\mathbf{1})^\top(\mathbf{x}_j-\mu_j\mathbf{1})$$

   则有协方差矩阵

$$\mathbf{S}=(s_{ij})_{p\times p}=\begin{bmatrix}s_{11}&s_{12}&\cdots&s_{1p}\\s_{21}&s_{22}&\cdots&s_{2p}\\\vdots&\vdots&\ddots&\vdots\\s_{p1}&s_{p2}&\cdots&s_{pp}\end{bmatrix}$$

5. 设有 p 个样本 $\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_p$, 则记

$$\mathbf{X}=(\mathbf{1},\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_p)$$

$$\mathbf{X}_0=(\mathbf{x}_1-\mu_1\mathbf{1},\mathbf{x}_2-\mu_2\mathbf{1},\dots,\mathbf{x}_p-\mu_p\mathbf{1})$$

   结合协方差与协方差矩阵的定义, 我们有

$$\mathbf{S}=\left(\frac{1}{n}(\mathbf{x}_i-\mu_i\mathbf{1})^\top(\mathbf{x}_j-\mu_j\mathbf{1})\right)_{p\times p}$$

即

$$\mathbf{S}=\frac{1}{n}(\mathbf{x}_1-\mu_1\mathbf{1},\mathbf{x}_2-\mu_2\mathbf{1},\dots,\mathbf{x}_p-\mu_p\mathbf{1})^\top(\mathbf{x}_1-\mu_1\mathbf{1},\mathbf{x}_2-\mu_2\mathbf{1},\dots,\mathbf{x}_p-\mu_p\mathbf{1})$$

即

$$\mathbf{S}=\frac{1}{n}\mathbf{X}_0^\top\mathbf{X}_0$$

这一方面印证了协方差矩阵是对称阵, 另一方面表明其特征值非负.

6. $\mathbf{x}$ 的标准化向量

$$\mathbf{x}'=\frac{\mathbf{x}-\bar{x}\mathbf{1}}{\|\mathbf{x}-\bar{x}\mathbf{1}\|}$$

7. $\mathbf{x}$ 与 $\mathbf{y}$ 的相关系数

$$r(\mathbf{x},\mathbf{y})=<\mathbf{x}',\mathbf{y}'>$$

经过计算, 我们可以发现

$$r(\mathbf{x},\mathbf{y})=r(\mathbf{x}+k_1\mathbf{1},\mathbf{y}+k_2\mathbf{1})$$

即相关性 (或相关系数) 与加减 $\mathbf{1}$ 向量的倍数无关.

8. $\mathbf{x}$ 与 $\mathbf{y}$ 的夹角余弦

$$\cos\theta=\frac{<\mathbf{x},\mathbf{y}>}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

根据相关系数的定义

$$\begin{aligned} r(\mathbf{x},\mathbf{y})=<\mathbf{x}',\mathbf{y}'>&=<\frac{\mathbf{x}-\bar{x}\mathbf{1}}{\|\mathbf{x}-\bar{x}\mathbf{1}\|},\frac{\mathbf{y}-\bar{y}\mathbf{1}}{\|\mathbf{y}-\bar{y}\mathbf{1}\|}>\\ &=\frac{<\mathbf{x}-\bar{x}\mathbf{1},\mathbf{y}-\bar{y}\mathbf{1}>}{\|\mathbf{x}-\bar{x}\mathbf{1}\|\|\mathbf{y}-\bar{y}\mathbf{1}\|}\\ &=\cos\gamma \end{aligned}$$

γ 为 $\mathbf{x}'$ 与 $\mathbf{y}'$ 的夹角, 与 θ 一般是不同的.

9. 设有 p 个样本 $\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_p$, 记 $r_{ij}=r(\mathbf{x}_i,\mathbf{x}_j)$, 则有相关矩阵

$$\mathbf{R}=(r_{ij})_{p\times p}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p}\\ r_{21} & r_{22} & \cdots & r_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

1.2 主成分分析

1.2.1 概念

假定现有 n 个样本的 p 个属性 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p(\mathbf{x}_i$ 是 n 维的), 由于各个属性之间的相关性, p 个属性间存在信息冗余.

现在的目标是通过正交变换, 得到 $k(1 \leq k < p)$ 个零均值的正交向量, 使得这些向量的方差的和最大 (最大信噪比), 正交的向量组能够方便的计算坐标且无冗余信息.

1.2.2 计算方法

假设我们有有 n 个样本的 p 个属性, 构成矩阵

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

则主成分分析 (S 型) 步骤如下:

1. 求样本均值 $\bar{\mathbf{x}} = (\mu_1, \mu_2, \ldots, \mu_p)$ 和样本的协方差矩阵 $\mathbf{S}$;
2. 求解特征方程 $|\lambda \mathbf{I} - \mathbf{S}| = 0$, 得到 p 个降序的特征根 $\lambda_1, \lambda_2, \ldots, \lambda_p$;
3. 求对应的各个特征向量 $\omega_1, \omega_2, \ldots, \omega_p$, 组成特征矩阵 $\mathbf{\Lambda} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_p)$;
4. 使用 Gram-Schmidt 正交化方法将 $\mathbf{\Lambda}$ 正交化, 并单位化得到 $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p)$. 在实际操作中, 由于单位正交化计算较为复杂, $\boldsymbol{\alpha}$ 应该是能通过观察特征矩阵 $\mathbf{\Lambda}$ 得到的;
5. 选取对应特征值较大的 k 个 $\boldsymbol{\alpha}_i$, 得到 k 个主成分
$$\boldsymbol{f}_i = \mathbf{X}_0 \boldsymbol{\alpha}_i = \sum_{j=1}^{p} \alpha_{ij} (\mathbf{x}_j - \mu_j \mathbf{1})$$
6. 计算累计贡献率, λ_i 也代表对应主成分的方差
$$\frac{\sum\limits_{i=1}^{k} \lambda_i}{\sum\limits_{i=1}^{p} \lambda_i}$$

以及主成分 i 的贡献率

$$\mathrm{PC}_i: \quad \frac{\lambda_i}{\sum\limits_{j=1}^{p} \lambda_j}$$

1.2.3 R 型分析

为了消除各属性量纲的影响, 先将各属性向量标准化再进行主成分分析.

由于对于标准化向量有 $\mathbf{R} = n\mathbf{S}$(或 $\mathbf{R} = (n-1)\mathbf{S}$), 同时在矩阵前乘上一个常系数不会影响最后的 $\boldsymbol{\alpha}$ 与累计贡献率, 且标准化不会影响相关矩阵, 所以这样的分析方法等价于从原数据的相关矩阵 $\mathbf{R}$ 出发进行主成分分析.

统计学上称这种方法为 R 型分析, 而称从原数据的协方差矩阵 $\mathbf{S}$ 出发进行主成分分析的方法为 S 型分析. 两种分析的结果通常是不同的. 一般情况下, 如果各属性的量纲不同, 通常采用 R 型分析.

1.3 例题

由于我们的考试是不允许使用计算器的, 应该不会让我们对太复杂的数据进行主成分分析, 下面给出一个简单的例题.

1. **题目**

   假设我们有一个包含 4 个样本和 3 个具有相同量纲的属性的数据集. 4 个样本点的数据如下:

样本	$\mathbf{x}_1$	$\mathbf{x}_1$	$\mathbf{x}_1$
1	3	5	1
2	2	6	1
3	2	5	2
4	1	4	0

   问题

   (a) 计算样本协方差矩阵 $\mathbf{S}$.

   (b) 求 $\mathbf{S}$ 的特征值 $(\lambda_1 \geq \lambda_2 \geq \lambda_3)$ 和对应的单位特征向量.

   (c) 计算各主成分的方差贡献率.

   (d) 若需保留至少 80% 的总方差, 应保留至少几个主成分?

(e) 解释原始数据是否需要标准化, 并说明原因.

解答

(a) 计算协方差矩阵 S

将数据矩阵 0 均值化:

$$\mathbf{X}_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{bmatrix}$$

计算 $\mathbf{X}_0^\top \mathbf{X}_0$:

$$\mathbf{X}_0^\top \mathbf{X}_0 = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

样本量 $n = 4$, 故协方差矩阵:

$$\boxed{\mathbf{S} = \frac{1}{n} \mathbf{X}_0^\top \mathbf{X}_0 = \frac{1}{4} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}}$$

(b) 求特征值与特征向量

矩阵 $\mathbf{S}$ 的特征多项式为:

$$\det(\mathbf{S} - \lambda \mathbf{I}) = -(\lambda - 1) \left(\lambda - \frac{1}{4}\right)^2$$

因此特征值为:

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_3 = \frac{1}{4}$$

对 $\lambda_1 = 1$ 解齐次方程组 $(\mathbf{S} - \mathbf{I})\mathbf{x} = 0$, 得到特征向量

$$\boldsymbol{\omega}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

单位化得

$$\mathbf{u}_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

对 $\lambda_2 = \lambda_3 = \frac{1}{4}$ 解齐次方程组 $(\mathbf{S} - \frac{1}{4}\mathbf{I})\mathbf{x} = 0$, 得到两个线性无关的特征向量

$$\boldsymbol{\omega}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\omega}_3 = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

它们已经是正交的了, 单位化得

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_3 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

综上

$$\boxed{\begin{array}{l} \text{特征值:} \quad \lambda_1 = 1, \quad \lambda_2 = \lambda_3 = \dfrac{1}{4} \\ \text{单位正交特征向量:} \\ \mathbf{u}_1 = \dfrac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{u}_2 = \dfrac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_3 = \dfrac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \end{array}}$$

(c) 方差贡献率

总方差:

$$\lambda_1 + \lambda_2 + \lambda_3 = \frac{3}{2}$$

各主成分贡献率:

$$\boxed{\begin{array}{ll} \text{PC}_1: & \dfrac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \dfrac{2}{3} \approx 66.67\% \\ \text{PC}_2: & \dfrac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \dfrac{1}{6} \approx 16.67\% \\ \text{PC}_3: & \dfrac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = \dfrac{1}{6} \approx 16.67\% \end{array}}$$

(d) 保留主成分数

设保留 k 个主成分.

$k=1$ 时, 累计贡献率为

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} \approx 66.67\% < 80\%$$

$k=2$ 时, 累计贡献率为

$$\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3} \approx 83.33\% \geq 80\%$$

$\boxed{\text{故需至少保留两个主成分.}}$

(e) 标准化必要性

$\boxed{\text{不需要标准化, 因为三个属性具有相同的量纲, 且数据范围接近.}}$

2 奇异值分解

2.1 特征分解

学习过线性代数, 我们知道如果任意实对称矩阵 $\mathbf{A}$ 都有 n 个线性无关的特征向量, $\mathbf{A}$ 可以被特征分解为

$$\mathbf{A} = \mathbf{W}\mathbf{\Sigma}\mathbf{W}^{-1}$$

其中, $\mathbf{W}$ 是这 n 个特征向量组成的矩阵, $\mathbf{\Sigma}$ 为对应特征值按顺序组成的对角阵.

如果我们将特征向量组正交单位化, 那么 $\mathbf{W}$ 就会成为酉矩阵 (正交矩阵), 即满足 $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$, 也即 $\mathbf{W}^\top = \mathbf{W}^{-1}$.

那么矩阵 $\mathbf{A}$ 的特征分解就可以写成

$$\mathbf{A} = \mathbf{W}\mathbf{\Sigma}\mathbf{W}^\top$$

事实上, 若有特征值 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ 对应特征向量 $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_n$, 上式还可写为

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top$$

2.2 奇异值分解

奇异值分解 (SVD) 是特征分解在任意矩阵上的推广.

假设矩阵 $\mathbf{A}$ 是一个 $m \times n$ 的矩阵, 那么定义矩阵 $\mathbf{A}$ 的 SVD 为:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

其中 $\mathbf{U}$ 是一个 $m \times m$ 的矩阵,$\mathbf{\Sigma}$ 是一个 $m \times n$ 的矩阵, 并且除了主对角线上元素以外全为 0, 主对角线上的每个元素都称为**奇异值**,$\mathbf{V}$ 是一个 $n \times n$ 的矩阵.$\mathbf{U}$ 和 $\mathbf{V}$ 都是酉矩阵, 即满足:

$$\mathbf{U}^\top\mathbf{U} = \mathbf{I}, \quad \mathbf{V}^\top\mathbf{V} = \mathbf{I}$$

经过推导, 如果设 $\mathbf{A}$ 的秩为 r, 则 $\mathbf{A}$ 的 SVD 会有如下形式

$$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m}\mathbf{\Sigma}_{m\times n}\mathbf{V}^\top_{n\times n} = \mathbf{U}_{m\times m}\begin{bmatrix}\mathbf{\Lambda}_{r\times r} & \mathbf{O} \\ \mathbf{O} & \mathbf{O}\end{bmatrix}_{m\times n}\mathbf{V}^\top_{n\times n}$$

其中 $\mathbf{\Lambda}_{r\times r} = \begin{bmatrix}\sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_r}\end{bmatrix}_{r\times r}$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ 为 $\mathbf{A}^\top\mathbf{A}$ 的 r 个非零特征值. $\mathbf{A}^\top\mathbf{A}$ 的 r 个非零特征值开根号 $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_r}$ 称为 $\mathbf{A}$ 的**正奇异值**, $\mathbf{A}^\top\mathbf{A}$ 的 n 个特征值开根号 $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_r}, 0_1, 0_2, \ldots, 0_{n-r}$ 即 $\mathbf{A}$ 的**奇异值**.

和特征分解类似地, 若记 $\mathbf{A}$ 的所有正奇异值按降序排序为 $\sigma_1, \sigma_2, \ldots, \sigma_r$, 则有

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

其中 $\mathbf{u}_i$ 与 $\mathbf{v}_i$ 分别为 $\mathbf{U}$ 和 $\mathbf{V}$ 的第 i 列. 似乎此式在实践中应用更多?

2.3 计算方法

SVD 的计算有两种方法, 可以从 $\mathbf{A}^\top\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top$ 出发计算, 也可以从 $\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top$ 出发计算. 设 $\mathbf{A}^\top\mathbf{A}$ 与 $\mathbf{A}^\top\mathbf{A}$ 的秩为 r.

2.3.1 第一种计算方法

从 $\mathbf{A}^\top\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top$ 出发计算 SVD 可以按照以下步骤.

1. 特征分解 $\mathbf{A}^\top\mathbf{A}$, 得到对角阵 $\begin{bmatrix} \mathbf{\Lambda}_{r\times r}^2 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$ 与酉矩阵 $\mathbf{V}$, 使得

$$\mathbf{A}^\top\mathbf{A} = \mathbf{V}_{n\times n} \begin{bmatrix} \mathbf{\Lambda}_{r\times r}^2 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \mathbf{V}_{n\times n}^\top$$

2. 将 $\mathbf{V}$ 拆分, 记

$$\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2), \quad \mathbf{V}_1 \in \mathbb{R}^{n\times r}, \mathbf{V}_2 \in \mathbb{R}^{n\times(n-r)}$$

3. 得到 $\mathbf{U}_1 = \mathbf{A}\mathbf{V}_1\mathbf{\Lambda}^{-1}$, 则有 $\mathbf{U}_1 \in \mathbb{R}^{m\times r}$, 这是我们要求的 $\mathbf{U}$ 的一部分, 可以设 $\mathbf{U}_1 = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r)$.

4. 将 $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ 扩充为标准正交基, 即将 $\mathbf{U}_1$ 扩充为酉矩阵 $\mathbf{U}$, 这有很多办法:

   (a) 观察法, 直接看出 $\mathbf{U}$;

   (b) 任意补充 $m-r$ 个向量 $\mathbf{u}'_{r+1}, \mathbf{u}'_{r+2}, \ldots, \mathbf{u}'_m$, 使得 $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r, \mathbf{u}'_{r+1}, \mathbf{u}'_{r+2}, \ldots, \mathbf{u}'_m$ 这 m 个向量线性无关, 使用 Gram-Schmidt 正交化方法正交化, 最后单位化;

   (c) 求 $\mathbf{U}_1^\top\mathbf{x} = \mathbf{0}$ 的基础解系, 再将基础解系单位正交化.

   据汪建基老师讲, 对于考试的题目应该可以使用观察法 (

5. 得到 $\mathbf{A}$ 的 SVD

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \mathbf{\Lambda} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}_{m\times n} \mathbf{V}^\top$$

   一定不要忘记 $\mathbf{V}$ 的转置符号哦, 或者直接写出 $\mathbf{V}$ 的转置也可以; 还有就是 $\mathbf{\Sigma}$ 要补成 $m \times n$.

2.3.2 第二种计算方法

从 $\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top$ 出发计算 SVD 可以按照以下步骤.

1. 特征分解 $\mathbf{AA}^\top$, 得到对角阵 $\begin{bmatrix}\mathbf{\Lambda}_{r\times r}^2 & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{bmatrix}$ 与酉矩阵 $\mathbf{U}$, 使得

$$\mathbf{AA}^\top = \mathbf{U}_{m\times m}\begin{bmatrix}\mathbf{\Lambda}_{r\times r}^2 & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{bmatrix}\mathbf{U}_{m\times m}^\top$$

2. 将 $\mathbf{U}$ 拆分, 记

$$\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2), \quad \mathbf{U}_1 \in \mathbb{R}^{m\times r}, \mathbf{U}_2 \in \mathbb{R}^{m\times(m-r)}$$

3. 得到 $\mathbf{V}_1 = \mathbf{A}^\top\mathbf{U}_1\mathbf{\Lambda}^{-1}$, 则有 $\mathbf{V}_1 \in \mathbb{R}^{n\times r}$, 这是我们要求的 $\mathbf{V}$ 的一部分, 可以设 $\mathbf{V}_1 = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r)$.

4. 将 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ 扩充为标准正交基, 即将 $\mathbf{V}_1$ 扩充为酉矩阵 $\mathbf{V}$.

5. 得到 $\mathbf{A}$ 的 SVD

$$\mathbf{A} = \mathbf{U}\begin{bmatrix}\mathbf{\Lambda} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{bmatrix}_{m\times n}\mathbf{V}^\top$$

2.4 例题

1. **题目**

用第一种方法求 $\mathbf{A} = \begin{bmatrix}4 & 11 & 14\\ 8 & 7 & -2\end{bmatrix}$ 的一个奇异值分解.

解答

首先计算

$$\mathbf{A}^\top\mathbf{A} = \begin{bmatrix}80 & 100 & 40\\ 100 & 170 & 140\\ 40 & 140 & 200\end{bmatrix}$$

解 $|\mathbf{A}^\top\mathbf{A} - \lambda\mathbf{I}| = 0$ 得其特征值为 $\lambda_1 = 360, \lambda_2 = 90, \lambda_3 = 0$, 对应有酉矩阵

$$\mathbf{V} = \begin{bmatrix}\frac{1}{3} & -\frac{2}{3} & \frac{2}{3}\\ \frac{2}{3} & -\frac{1}{3} & -\frac{2}{3}\\ \frac{2}{3} & \frac{2}{3} & \frac{1}{3}\end{bmatrix}$$

由于 $\mathbf{A}^\top\mathbf{A}$ 的秩为 2, 将 $\mathbf{V}$ 拆分得到

$$\mathbf{V}_1 = \begin{bmatrix}\frac{1}{3} & -\frac{2}{3}\\ \frac{2}{3} & -\frac{1}{3}\\ \frac{2}{3} & \frac{2}{3}\end{bmatrix}, \quad \mathbf{V}_2 = \begin{bmatrix}\frac{2}{3}\\ -\frac{2}{3}\\ \frac{1}{3}\end{bmatrix}$$

根据特征值, 我们得到

$$\mathbf{\Lambda} = \begin{bmatrix} 6\sqrt{10} & 0 \\ 0 & 3\sqrt{10} \end{bmatrix}$$

补充为 2×3 的矩阵

$$\mathbf{\Sigma} = \begin{bmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{bmatrix}$$

接下来求解 $\mathbf{U}$, 首先有

$$\begin{aligned} \mathbf{U}_1 &= \mathbf{A}\mathbf{V}_1\mathbf{\Lambda}^{-1} \\ &= \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 6\sqrt{10} & 0 \\ 0 & 3\sqrt{10} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}} \end{bmatrix} \end{aligned}$$

本应将 $\mathbf{U}_1$ 扩充, 但其已经是 2×2 的酉矩阵, 无需扩充, 故

$$\mathbf{U} = \mathbf{U}_1 = \begin{bmatrix} \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}} \end{bmatrix}$$

故 $\mathbf{A}$ 的 SVD 为

$$\boxed{\begin{aligned} \mathbf{A} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \\ &= \begin{bmatrix} \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}} \end{bmatrix} \begin{bmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{bmatrix}^\top \end{aligned}}$$

2. **题目**

用第二种方法求 $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix}$ 的一个奇异值分解.

解答

计算 $\mathbf{A}\mathbf{A}^\top$:

$$\mathbf{A}\mathbf{A}^\top = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 1 & -2 & 2 \\ -1 & 2 & -2 \end{bmatrix} = \begin{bmatrix} 2 & -4 & 4 \\ -4 & 8 & -8 \\ 4 & -8 & 8 \end{bmatrix}.$$

易见其秩为 1. 特征分解得到酉矩阵

$$\mathbf{U} = \begin{bmatrix} \frac{1}{3} & \frac{2}{\sqrt{5}} & -\frac{2}{3\sqrt{5}} \\ -\frac{2}{3} & \frac{1}{\sqrt{5}} & \frac{4}{3\sqrt{5}} \\ \frac{2}{3} & 0 & \frac{5}{3\sqrt{5}} \end{bmatrix}$$

与

$$\mathbf{\Lambda} = \begin{bmatrix} 3\sqrt{2} \end{bmatrix}$$

补充成 3×2 的矩阵

$$\mathbf{\Sigma} = \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

分割 $\mathbf{U}$ 得到

$$\mathbf{U}_1 = \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{2}{3} \end{bmatrix}$$

计算 $\mathbf{V}_1$

$$\begin{aligned} \mathbf{V}_1 &= \mathbf{A}^\top \mathbf{U}_1 \mathbf{\Lambda}^{-1} \\ &= \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix}^\top \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{2}{3} \end{bmatrix} \begin{bmatrix} 3\sqrt{2} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \end{aligned}$$

可以通过观察将其补充为酉矩阵

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

故 $\mathbf{A}$ 的 SVD 为

$$\boxed{\mathbf{A} = \begin{bmatrix} \frac{1}{3} & \frac{2}{\sqrt{5}} & -\frac{2}{3\sqrt{5}} \\ -\frac{2}{3} & \frac{1}{\sqrt{5}} & \frac{4}{3\sqrt{5}} \\ \frac{2}{3} & 0 & \frac{5}{3\sqrt{5}} \end{bmatrix} \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^\top}$$

3　函数矩阵求导

3.1　函数矩阵的定义

以变量 x 的函数 $a_{ij}(x)$ 为元素构成的矩阵称为函数矩阵. 函数矩阵的运算性质 (加法、数乘、乘法、转置等) 与常数矩阵的运算性质相同.

3.2　函数矩阵的逆矩阵

设 $\mathbf{A}(x)$ 为 n 阶函数矩阵, 如果存在 n 阶函数矩阵 $\mathbf{B}(x)$, 对于任意的 $x \in \mathbb{D}$ 都有

$$\mathbf{A}(x)\mathbf{B}(x) = \mathbf{B}(x)\mathbf{A}(x) = \mathbf{E}_n$$

则称 $\mathbf{A}(x)$ 在区间 $\mathbb{D}$ 上可逆,$\mathbf{B}(x)$ 是 $\mathbf{A}(x)$ 的逆矩阵, 记为 $\mathbf{A}^{-1}(x)$.

求函数矩阵逆矩阵的方法与常数矩阵类似:

$$\mathbf{A}^{-1}(x) = \frac{\mathbf{A}^{*}(x)}{|\mathbf{A}(x)|}$$

3.3　函数矩阵的导数

若函数矩阵 $\mathbf{A}(x) = [a_{ij}(x)]_{n\times m}$ 的所有元素 $a_{ij}(x)$ 在区间 $\mathbb{D}$ 上对 x 处处可导, 则称函数矩阵 $\mathbf{A}(x)$ 在区间 $\mathbb{D}$ 上对 x 可导, 其导数记为

$$\mathbf{A}'(x) = \frac{d\mathbf{A}(x)}{dx} = \left[\frac{da_{ij}(x)}{dx}\right]_{n\times m}$$

函数矩阵的导数为一同型函数矩阵, 其每个元素都是原函数矩阵对应元素的导数.

3.4　函数矩阵导数运算的常用性质

1. 常数矩阵的导数为零矩阵.
2. $[\lambda\mathbf{A}(x) + \mu\mathbf{B}(x)]' = \lambda\mathbf{A}'(x) + \mu\mathbf{B}'(x)$.
3. $[\mathbf{A}(x)\mathbf{B}(x)]' = \mathbf{A}'(x)\mathbf{B}(x) + \mathbf{A}(x)\mathbf{B}'(x)$.
4. 若 $\mathbf{A}(x)$,$\mathbf{A}^{-1}(x)$ 都可导, 则

$$[\mathbf{A}^{-1}(x)]' = -\mathbf{A}^{-1}(x)\mathbf{A}'(x)\mathbf{A}^{-1}(x)$$

3.5 逆矩阵求导的计算方法

逆矩阵的导数可以通过两种方法计算.

1. 先计算出逆矩阵, 然后分别对每个元素求导.
2. 利用性质 (4)进行求导:$[\mathbf{A}^{-1}(x)]' = -\mathbf{A}^{-1}(x)\mathbf{A}'(x)\mathbf{A}^{-1}(x)$.

例题: 函数矩阵 $\mathbf{A}(x) = \begin{bmatrix} x & 2x-1 \\ 1 & x \end{bmatrix}$, $x \neq 1$, 求 $\mathbf{A}^{-1}(x)$ 的导数.

解: 基础的计算过程在此省略, 使用两种方法均可求出

$$\mathbf{A}^{-1}(x) = -\frac{1}{(x-1)^3}\begin{bmatrix} x+1 & -2x \\ -2 & x+1 \end{bmatrix}$$

4 矩阵的偏导与梯度

4.1 定义与关系

Jacobian 矩阵 $1 \times n$ 行向量偏导算子记为

$$D_{\mathbf{x}} \triangleq \frac{\partial}{\partial \mathbf{x}^\top} = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]$$

当实值标量函数 $f(\mathbf{X})$ 的变元为实值矩阵 $\mathbf{X} \in \mathbb{R}^{n\times m}$ 时, 其 Jacobian 矩阵 $D_{\mathbf{X}} f(\mathbf{X})$ 定义为

$$D_{\mathbf{X}} f(\mathbf{X}) \triangleq \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}^\top} = \left[\frac{\partial f(\mathbf{X})}{\partial x_{ji}}\right]_{m\times n}$$

梯度矩阵 $n \times 1$ 列向量偏导算子记为

$$\nabla_{\mathbf{x}} \triangleq \frac{\partial}{\partial \mathbf{x}} = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^\top$$

当实值标量函数 $f(\mathbf{X})$ 的变元为实值矩阵 $\mathbf{X} \in \mathbb{R}^{n\times m}$ 时, 其梯度矩阵 $\nabla_{\mathbf{X}} f(\mathbf{X})$ 定义为

$$\nabla_{\mathbf{X}} f(\mathbf{X}) \triangleq \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \left[\frac{\partial f(\mathbf{X})}{\partial x_{ij}}\right]_{n\times m}$$

二者关系 实值标量函数 $f(\mathbf{X})$ 的梯度矩阵等于其 Jacobian 矩阵的转置:

$$\nabla_{\mathbf{X}} f(\mathbf{X}) = D_{\mathbf{X}}^\top f(\mathbf{X})$$

4.2 偏导和梯度计算的常用法则

计算实值函数对向量或矩阵的偏导数的常用法则与标量函数类似:

1. **常数法则** 若 $f(\mathbf{X}) = c$ 为常数, 则 $\frac{\partial c}{\partial \mathbf{X}} = \mathbf{O}$.

2. **线性法则** 若 $f(\mathbf{X})$ 和 $g(\mathbf{X})$ 为 $\mathbf{X}$ 的实值函数,c_1 和 c_2 为实常数, 则
$$\frac{\partial[c_1 f(\mathbf{X}) + c_2 g(\mathbf{X})]}{\partial \mathbf{X}} = c_1 \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} + c_2 \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}}$$

3. **乘积法则** 若 $f(\mathbf{X})$、$g(\mathbf{X})$ 和 $h(\mathbf{X})$ 都是 $\mathbf{X}$ 的实值函数, 则
$$\frac{\partial[f(\mathbf{X})g(\mathbf{X})]}{\partial \mathbf{X}} = g(\mathbf{X})\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} + f(\mathbf{X})\frac{\partial g(\mathbf{X})}{\partial \mathbf{X}}$$

4. **商法则** 若 $g(\mathbf{X}) \neq 0$, 则
$$\frac{\partial[f(\mathbf{X})/g(\mathbf{X})]}{\partial \mathbf{X}} = \frac{1}{g^2(\mathbf{X})}\left[g(\mathbf{X})\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} - f(\mathbf{X})\frac{\partial g(\mathbf{X})}{\partial \mathbf{X}}\right]$$

5. **链式法则** 若 $y = f(\mathbf{X})$ 和 $g(y)$ 分别是以矩阵 $\mathbf{X}$ 和标量 y 为变量的实值函数, 则
$$\frac{\partial g(f(\mathbf{X}))}{\partial \mathbf{X}} = \frac{dg(y)}{dy}\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$$

独立性基本假设 在进行偏导计算时, 通常假设向量变元 $\mathbf{x}$ 或矩阵变元 $\mathbf{X}$ 的元素之间是相互独立的.

4.3 偏导矩阵的计算

几个例题:

1. $$\frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

2. $$\frac{\partial \|\mathbf{x}\|_2^2}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$$

3. $$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}^\top + \mathbf{A})\mathbf{x}$$

证明. 先将函数展开为求和形式:

$$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}} \left(\sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} x_j x_k \right)$$

则梯度向量的第 i 个元素为

$$\left[\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} \right]_i = \sum_{j=1}^{n} a_{ij} x_j + \sum_{k=1}^{n} a_{ki} x_k$$

故其梯度向量为

$$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}^\top + \mathbf{A})\mathbf{x}$$

□

4.

$$\frac{\partial \mathbf{a}^\top \mathbf{X}\mathbf{b}}{\partial \mathbf{x}} = \mathbf{a}\mathbf{b}^\top$$

5.

$$\frac{\partial \mathbf{a}^\top \mathbf{X}\mathbf{X}^\top \mathbf{b}}{\partial \mathbf{x}} = (\mathbf{a}\mathbf{b}^\top + \mathbf{b}\mathbf{a}^\top)\mathbf{X}$$

证明. 先将函数展开为求和形式:

$$\partial \mathbf{a}^\top \mathbf{X}\mathbf{X}^\top \mathbf{b} = \sum_{k=1}^{n} \sum_{l=1}^{n} \sum_{p=1}^{m} a_k b_l x_{kp} x_{lp}$$

则梯度矩阵的第 (i, j) 个元素为

$$\left[\frac{\partial \mathbf{a}^\top \mathbf{X}\mathbf{X}^\top \mathbf{b}}{\partial \mathbf{x}} \right]_{ij} = \sum_{k=1}^{n} b_i a_k x_{kj} + \sum_{l=1}^{n} a_i b_l x_{lj}$$

故其梯度矩阵为

$$\frac{\partial \mathbf{a}^\top \mathbf{X}\mathbf{X}^\top \mathbf{b}}{\partial \mathbf{x}} = (\mathbf{a}\mathbf{b}^\top + \mathbf{b}\mathbf{a}^\top)\mathbf{X}$$

□

6.

$$\frac{\partial \mathrm{tr}(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial \mathrm{tr}(\mathbf{X}\mathbf{A})}{\partial \mathbf{X}} = \mathbf{A}^\top$$

证明. 矩阵乘积的元素为

$$[\mathbf{AX}]_{ij} = \sum_{k=1}^{n} a_{ik}x_{kj}$$

故矩阵乘积的迹为

$$\mathrm{tr}(\mathbf{AX}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_{ji}$$

由此可得

$$\frac{\partial \mathrm{tr}(\mathbf{AX})}{\partial \mathbf{X}} = \frac{\partial \mathrm{tr}(\mathbf{XA})}{\partial \mathbf{X}} = [a_{ji}]_{n\times n} = \mathbf{A}^\top$$

□

7.

$$\frac{\partial \|\mathbf{Wx}\|_2^2}{\partial \mathbf{x}} = 2\mathbf{W}^\top\mathbf{Wx}, \quad \frac{\partial \|\mathbf{Wx}\|_2^2}{\partial \mathbf{W}} = 2\mathbf{Wxx}^\top$$

证明. 先将函数展开:

$$\|\mathbf{Wx}\|_2^2 = (\mathbf{Wx})^\top\mathbf{Wx} = \mathbf{x}^\top\mathbf{W}^\top\mathbf{Wx} = \sum_{k=1}^{n}\sum_{l=1}^{m}\sum_{p=1}^{m} x_l x_p w_{kl} w_{kp}$$

其对 $\mathbf{x}$ 的偏导可以参考例 3计算, 得到

$$\frac{\partial \mathbf{x}^\top\mathbf{W}^\top\mathbf{Wx}}{\partial \mathbf{x}} = 2\mathbf{W}^\top\mathbf{Wx}$$

其对 $\mathbf{W}$ 的梯度矩阵的第 (i,j) 个元素为

$$\left[\frac{\partial \mathbf{x}^\top\mathbf{W}^\top\mathbf{Wx}}{\partial \mathbf{W}}\right]_{ij} = 2x_j \sum_{k=1}^{m} w_{ik}x_k$$

故其对 $\mathbf{W}$ 的梯度矩阵为

$$\frac{\partial \mathbf{x}^\top\mathbf{W}^\top\mathbf{Wx}}{\partial \mathbf{W}} = 2\mathbf{Wxx}^\top$$

□

5 伪逆和广义逆

5.1 前置知识

这里列出一些下面提到的, 我们上学期学过的线性代数知识点. 保证不超纲:

1. 值空间和核. 对于矩阵 $A \in \mathbb{R}^{n\times m}$, 其值空间为 $R(A) = \{Ax \mid x \in \mathbb{R}^n\}$, 核为 $\ker(A) = \{x \mid Ax = 0\}$.
2. 秩-零度和定理:$\text{rank } A + \dim\ker(A) = n$
3. 向量和子空间的正交: 对于向量 x 和子空间 V, 若 $\forall v \in V, v^{\mathrm{T}}x = 0$, 称 $x \perp V$.

5.2 伪逆

5.2.1 定义和性质

伪逆是一种针对**满秩矩阵**的概念. 对于秩为 r, 大小为 $n \times m$ 的矩阵 A, 如果:

- $r = m \leq n$, 则存在唯一的左伪逆 $L \in \mathbb{R}^{m\times n}$ 使得 $LA = I_m$
- $r = n \leq m$, 则存在唯一的左伪逆 $R \in \mathbb{R}^{m\times n}$ 使得 $AR = I_n$

一个自然的点是, 如果 $r = m = n$, 则 L, R 退化到 A^{-1}.

不难注意到一个特点: 无论左伪逆还是右伪逆, 形状都是 A^{T} 的形状. 这在记忆公式的时候非常有用.

左逆和右逆没有本质区别, 因为 $R^{\mathrm{T}}A^{\mathrm{T}} = I_n$, 所以 R^{T} 是 A^{T} 的左逆.

关于唯一性的证明不重要, 但还是附上, 以左伪逆为例:

如果存在 $L_1A = L_2A = I_m$, 则:

- A 在 $\mathbb{R}^n$ 到 $R(A)$ 上是一一映射, 则 L_1, L_2 在 $R(A)$ 上是相同的一一映射, 即 A 的逆映射.
- 由于 L_1, L_2 秩为 n, 则对于任何 y 正交于 $R(A)$,$L_1y = L_2y = 0$.

综上, 对于任意 $y \in \mathbb{R}^m$, 将其分解为 $y = y_R + y_N$, 其中 $y_R \in R(A)$,y_N 正交于 $R(A)$, 则 $L_1y = L_2y = Ly_R$. 说明 $L_1 = L_2$.

5.2.2 计算

L, R 计算如下:

$$L = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$$
$$R = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}$$

以 L 为例进行说明:

> 引理: 若 $r=m\le n$, 即 A 列满秩, 则 $A^{\mathrm{T}}A$ 可逆. 证明: 在 $x\neq 0$ 时, 由于 $A^{\mathrm{T}}Ax=0\Rightarrow x^{\mathrm{T}}A^{\mathrm{T}}Ax=0\Leftrightarrow Ax=0$, 但是 A 列满秩, 不存在 $x\neq 0, Ax=0$, 也就不存在 $A^{\mathrm{T}}Ax=0$, 所以 $A^{\mathrm{T}}A$ 列满秩. 而它又是方阵, 则 $A^{\mathrm{T}}A$ 可逆.

既然存在 L, 说明 $r=m\le n$, 由引理 $A^{\mathrm{T}}A\in\mathbb{R}^{m\times m}$ 是可逆, 则

$$LA=(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}A=I_m$$

5.2.3 例题

计算如下矩阵的左伪逆:

$$A=\begin{bmatrix}-1 & 0 & 1\\ 1 & 0 & 0\\ 1 & -1 & 0\\ -1 & 0 & 1\end{bmatrix}$$

答案: 计算得:

$$A^{\mathrm{T}}A=\begin{bmatrix}4 & -1 & -2\\ -1 & 1 & 0\\ -2 & 0 & 2\end{bmatrix}$$

$$(A^{\mathrm{T}}A)^{-1}=\begin{bmatrix}1 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 1.5\end{bmatrix}$$

所以:

$$L=(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}=\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 1 & -1 & 0\\ 0.5 & 1 & 0 & 0.5\end{bmatrix}$$

5.3 普通广义逆

所谓广义逆, 是**放弃了满秩要求**后更一般的逆.

对于一般矩阵 $A\in\mathbb{R}^{n\times m}$, 当 $r<\min(n,m)$ 时, 不可能有任何矩阵 A^- 满足 $AA^-=I$ 或 $A^-A=I$. 此时的广义逆从线性方程组着手.

对于理想的逆, 应该能够做到 $x = A^{-}y \Leftrightarrow Ax = y$. 可惜 A 不可逆. 所以我们退而求其次, 把充要条件换成充分条件: 若 $x = A^{-}y \Rightarrow Ax = y$, 称 A^{-} 是 A 的广义逆矩阵.

但是这个说法并不严谨, 因为没有对 y 的范围做限制, 否则当 A^{-} 不是单射的时候就不可能成立.

严谨的说, 广义逆矩阵要满足:$\forall y \in R(A), A(A^{-}y) = y$. 又因为 $\forall x, Ax \in R(A)$, 所以:

$$\forall x, AA^{-}Ax = Ax$$

即 $AA^{-}A = A$. 这就是广义逆的充要条件. 显然, 伪逆是一种特殊的广义逆.

5.4 穆尔-彭罗斯 (Moore–Penrose) 广义逆

广义逆不是唯一的. 有很多自由度:

1. 我们只规定了 $\forall y \in R(A)$,$A^{-}y$ 应该取什么值. 但是对于 $y \perp R(A)$, 我们没有任何限制.
2. 我们只规定了 $A(A^{-}y) = y$, 但是没有指明 $A^{-}y$ 和 $\ker(A)$ 的关系,$A^{-}y$ 即使不和 $\ker A$ 正交也完全满足要求.

如果我们进一步, 对 $A^{-}y$ 的取值做一些限定, 就会得到要求更强的广义逆.M-P 逆从**最小化 L2 范数**出发, 做出了如下要求:

- 对于 $y \in R(A)$, 我们希望 $A^{-}y \perp \ker(A)$.
- 对于 $y \perp R(A)$, 我们希望 $A^{-}y = 0$.

这样的广义逆, 就是 Moore–Penrose 逆, 我们用 A^{+} 来表示.

> 这种逆的性质非常好, 如下图所示: 左下示意的是 $\ker(A)$ 子空间, 右上示意的是 $R(A)$ 子空间. 左上是行空间 $R(A^{\mathrm{T}})$, 它正交于 $\ker A$, 他们的直和是 $\mathbb{R}^{n}$(注意这个图的 A 是 $m \times n$ 的, 和前文相反); 右下是左零空间 $\ker(A^{\mathrm{T}})$, 它正交于 $R(A)$, 它们的直和是 $\mathbb{R}_{m}$. 这段话是”线性代数基本定理“, 在这里不展开.
>
> M-P 逆是对空间的这种分割下最自然的逆: 在 $R(A)$ 上, 它是 A 的逆映射; 而在 $\ker(A^{\mathrm{T}})$ 上, 它恒为零映射.

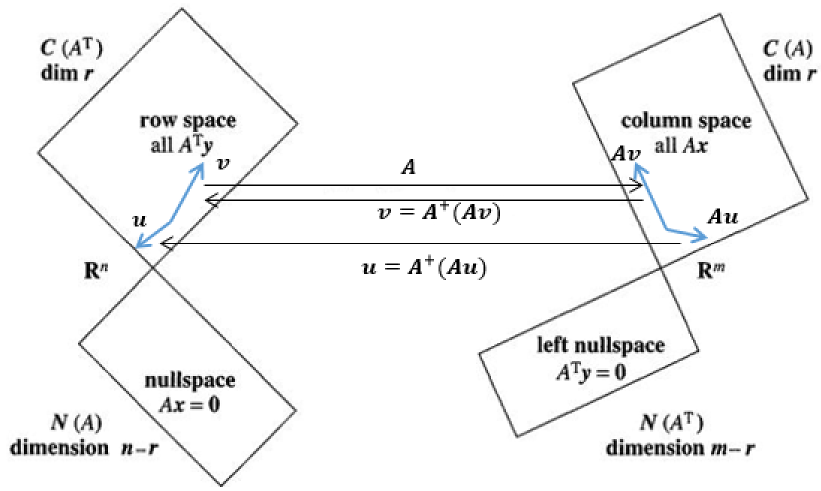

图 1: Moore–Penrose 逆

5.4.1 四条件

上面补充的两个要求, 加上广义逆本身要求的 $\forall y \in R(A), A(A^-y) = y$, 一共三个要求. 这三个要求和 M-P 四条件是等价的. 这四个条件是:

1. $AA^-A = A$
2. $A^-AA^- = A^-$
3. $(AA^+)^\mathrm{T} = AA^+$
4. $(A^+A)^\mathrm{T} = A^+A$

证明:
先证三条加强要求 $\implies$ Penrose 四条件:

1. 条件 1 已显然: 三条要求中的第三条就是 $AA^-A = A$.
2. 条件 2($A^-AA^- = A^-$)
   任取 $y \in \mathbb{R}^m$, 分两种情况讨论.
   (a) 若 $y \in R(A)$, 则 $A^-y \perp \ker A$, 又 $A(A^-y) = y$, 即 A^-y 就是 $Ax = y$ 的唯一最小范数解. 而 $A^-AA^-y = A^-y$.

(b) 若 $y \perp R(A)$, 则 $A^-y = 0$,$A^-AA^-y = A^-A0 = 0 = A^-y$.

综合即对任意 y,$A^-AA^-y = A^-y$, 即 $A^-AA^- = A^-$.

3. 条件 3(AA^- 对称):
先观察 AA^- 是 $R(A)$ 上的正交投影. 对于 $v, w \in \mathbb{R}^m$, 都有 $\langle v, AA^-w\rangle = \langle AA^-v, w\rangle$, 因为 AA^- 映向的是 $R(A)$,A^- 反解回到 $R(A^T)$, 在分解正交的空间结构下 AA^- 自然是对称的.

4. 条件 4(A^-A 对称):
同理,A^-A 是 $R(A^T)$ 上的正交投影, 亦对称.

再证 Penrose 四条件 $\implies$ 这三条加强要求:

1. 首先,$AA^-A = A$ 保证只要 $y \in R(A)$,$A(A^-y) = y$(广义逆定义).
2. $A^-AA^- = A^-$ 保证 A^-y 是所有 $Ax = y$ 解中惟一与 $\ker(A)$ 正交的那个. 这就是"最小范数"要求.
3. AA^-、A^-A 都是对称投影 (正交投影), 这意味着 $A^-y = 0$ 对于 $y \perp R(A)$, 即将正交补映为零.

综上, **这三条加强要求与 Moore–Penrose 四条件充要等价**.

这种逆矩阵是存在且唯一的.

存在性和唯一性的证明几乎是构造性的, 这一节的第一个引用块就几乎说明了一切.

5.4.2 计算

可以通过 SVD 分解计算 M-P 逆矩阵:

- 对于对角矩阵 D, 其 M-P 逆 D^+ 恰为 D^T 中所有非零元素取倒数的结果.
- 对于一般矩阵 A, 设其 SVD 分解为 $A = U\Sigma V^\mathrm{T}$, 则其 M-P 逆为 $A^+ = V\Sigma^+U^\mathrm{T}$.

这两个结果都可以通过带入四条件进行检验.

特别地, 如果 A 可逆, 退化为一般逆; 如果 A 行满秩或列满秩, 退化为对应伪逆.

5.4.3 例题

一、对于如下的 A, B, 验证他们互为 Moore-Penrose 伪逆:

$$A = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}, B = \begin{bmatrix} 0.25 & -0.25 & 0 & 0.5 \\ -0.125 & 0.125 & 0 & 0.25 \\ -0.125 & 0.125 & 0 & 0.25 \end{bmatrix}$$

答案: 代入计算四条件即可:

$$ABA = A$$

$$BAB = B$$

$$AB = \begin{bmatrix} 0.5 & -0.5 & 0 & 0 \\ -0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$BA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \end{bmatrix}$$

二、计算 A 的 Moore-Penrose 伪逆:

(1) $A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$

(2) $A = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & -1 & 0 \end{bmatrix}$

(3) $A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & -1 \end{bmatrix}$

答案:

(1) 是对角阵, 直接得到:$A^+ = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \end{bmatrix}$

(2) 是列满秩阵, 计算其左伪逆得到:

$$(A^{\mathrm{T}}A)^{-1} = \begin{bmatrix} 0.75 & 0.25 & -0.5 \\ 0.25 & 0.75 & -0.5 \\ -0.5 & -0.5 & 1 \end{bmatrix}$$

$$A^+ = \begin{bmatrix} -0.5 & -0.5 & 0 & 0.5 \\ -0.5 & -0.5 & 0 & -0.5 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

(3) 是一般矩阵, 计算其奇异值分解, 再计算其伪逆:

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

$$\Sigma = \begin{bmatrix} 2 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$V^{\mathrm{T}} = \begin{bmatrix} 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 1 & 0 & 0 \end{bmatrix}$$

$$A^+ = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.25 & 0 & 0.5 & -0.25 \\ 0.25 & 0 & -0.5 & -0.25 \end{bmatrix}$$

注意计算时:

- **有没有计算** Σ^+, 是对 Σ **转置后取倒数**.
- 是否算的是 $V\Sigma^+U^{\mathrm{T}}$. **别把转置搞混了**!

6 多元线性回归

6.1 建模

有很多办法可以把问题建模为线性方程组. 比如, 给定散点数据 $X, \mathbf{y}$, 其中:

$$X = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix}, \mathbf{x}_i \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$$

$(\mathbf{x}_i, y_i)$ 是一组数据, 也许是某个复杂现象的自变量-因变量对. 这样的数据一共有 n 组.

我们希望用线性方程来近似拟合这个一般函数. 也即用如下方程:

$$y = \mathbf{x}^{\mathrm{T}}\beta + b_0$$

不妨记 $\mathbf{x}_0 = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}, \beta_0 = \begin{bmatrix} \beta \\ b_0 \end{bmatrix}$, 方程重写如下:

$$y = \mathbf{x}_0^{\mathrm{T}}\beta_0$$

β_0 是参数, 也是我们做拟合需要求出的东西. 这样的话, 记:

$$X_0 = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} & 1 \\ \mathbf{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^{\mathrm{T}} & 1 \end{bmatrix}$$

求 β_0 就是求解线性方程组 $X_0\beta_0 = \mathbf{y}$. 这就是把线性拟合问题建模为线性方程组的过程.

6.2 M-P 逆求解

根据 M-P 逆的内容, 对于任何一般的方程组 $A\mathbf{x} = \mathbf{y}$, 我们总有一个 M-P 逆解:

$$\widehat{\mathbf{x}} = A^{+}\mathbf{y}$$

这个解具有如下性质:

1. 若 $\mathbf{y} \in R(A)$, 即原方程组有至少一个解, 此时 $\widehat{\mathbf{x}} \perp \ker(A)$.
2. 否则可以拆分 $\mathbf{y} = \mathbf{y}_R + \mathbf{y}_N$, 其中 $\mathbf{y}_R \in R(A)$,$\mathbf{y}_N \perp R(A)$, 则此时 $\widehat{\mathbf{x}} = A^{+}(\mathbf{y}_R + \mathbf{y}_N) = A^{+}\mathbf{y}_R$.

6.2.1 有解情况

对于第一条, 我们回顾一下解线性方程组解的性质:

> 如果有解 (此时 $\mathbf{y} \in R(A)$), 通解的格式为 $\mathbf{x} = \mathbf{x} + \mathbf{x}_0$. 其中 $\mathbf{x}$ 满足 $A\mathbf{x} = \mathbf{y}$, 而 $\mathbf{x}_0$ 是任意 $\ker(A)$ 中的向量.

此时可能有很多解, 我们的 $\widehat{\mathbf{x}}$ 有什么优势呢?

M-P 逆的优势在此: **在原方程组有解时,M-P 逆求得的解是所有解中 L2 范数最小的.**

> 证明: 通解中的任何一个解都可以进行拆分:
>
> $$\mathbf{x} = \mathbf{x}_R + \mathbf{x}_N, \mathbf{x}_R \perp \ker(A), \mathbf{x}_N \in \ker(A)$$
>
> 由于通解的格式是 $\mathbf{x} = \mathbf{x} + \mathbf{x}_0$, 那个能变化的 $\mathbf{x}_0 \in \ker(A)$, 所以**对于通解中的任意一个解, 拆分后得到的 $\mathbf{x}_R$ 都是相同的, 变化的部分只有 $\mathbf{x}_N$.**
>
> 进一步, 考虑解的 L2 范数. 因为 $\mathbf{x}_R$ 与 $\ker(A)$ 正交, 自然也与 $\mathbf{x}_N$ 正交, 所以:
>
> $$\|\mathbf{x}\|^2 = \|\mathbf{x}_R\|^2 + \|\mathbf{x}_N\|^2$$
>
> 前者对于通解中的所有解都是固定值, 而后者非负. 因此, 所有通解中 L2 范数最小的那个必然满足 $\mathbf{x}_N = 0$, 也即 $\mathbf{x} = \mathbf{x}_R \perp \ker(A)$. 而我们又知道,M-P 逆求得的解恰好满足 $\widehat{\mathbf{x}} \perp \ker(A)$, 所以我们得到结论: 在原方程组有解时,M-P 逆求得的解是所有解中 L2 范数最小的.

6.2.2 无解情况

此时应用第二条性质. 把 $\mathbf{y}$ 拆分成 $\mathbf{y}=\mathbf{y}_R+\mathbf{y}_N$, 其中 $\mathbf{y}_R\in R(A)$,$\mathbf{y}_N\perp R(A)$.

既然原方程组无解, 我们可以退而求其次, 求一个误差最小的解. 所谓误差, 指的是这个向量:

$$\mathbf{e}=\mathbf{y}-A\widehat{\mathbf{x}}=(\mathbf{y}_R-A\widehat{\mathbf{x}})+\mathbf{y}_N$$

此时 M-P 逆的优势如下: **在原方程无解时,M-P 逆求得的解是误差 L2 范数最小的, 而且是误差范数最小的解中自身 L2 范数最小的.**

> 证明: 运用和有解情况非常相似的分析方法: 因为 $A\widehat{\mathbf{x}}\in R(A)$, $\mathbf{y}_R\in R(A)$, 所以括号项也在 $R(A)$ 中；但是 $\mathbf{y}_N\perp R(A)$, 所以上式的 L2 范数如下:
>
> $$\|\mathbf{e}\|^2=\|\mathbf{y}_R-A\widehat{\mathbf{x}}\|^2+\|\mathbf{y}_N\|^2$$
>
> 第二项是和 $\widehat{\mathbf{x}}$ 无关的, 所以最小化 $\|\mathbf{e}\|$ 就是要最小化 $\|\mathbf{y}_R-A\widehat{\mathbf{x}}\|$. 不难发现这一项是可以取到 0 的, 因为 $\mathbf{y}_R\in R(A)$, 所以 $A\mathbf{x}=\mathbf{y}_R$ 必然有解.
>
> 根据性质二,$\mathbf{y}_R$ 的 M-P 逆解和 $\mathbf{y}$ 的 M-P 逆解是同一个；再根据有解情况的分析,M-P 逆的解 $\widehat{\mathbf{x}}$ 是所有解里范数最小的, 所以综合来说,M-P 逆解是误差范数最小的, 而且是误差范数最小的解中自身范数最小的.

6.2.3 总结

既然 M-P 逆求得的解在任何情况下都使得误差范数与自身范数同时最小, 我们把这个解称作 $A\mathbf{x}=\mathbf{y}$ 的**最小二乘解**.

这个解在原方程有解时是 L2 范数最小的；在原方程无解时是使得误差 L2 范数最小的所有解中, 自身 L2 范数最小的.

6.3 帽子矩阵

我们回到多元线性回归的情况. 我们已经将它建模成了 $X_0\beta_0=\mathbf{y}$ 的形式, 因而可以求出参数 β_0 的最小二乘解:

$$\begin{aligned}\hat{\beta}_0 &= X_0^+\mathbf{y}\\ \widehat{\mathbf{y}} &= X_0X_0^+\mathbf{y}\end{aligned}$$

这和帽子矩阵有什么关系呢?

是这样的, 最小二乘解适用于任何形态的方程, 无论它是过定还是欠定, 秩满还是秩亏. 但是在回归分析的时候, 最常见的情况是 $n\gg m$, 即数据量远大于参数量, 这种过定的情况下, 一般 X_0 都是**列满秩**的, 即 $n\gg m+1=\operatorname{rank} A$.

此时,**M-P 逆退化成左伪逆**, 也即:

$$\begin{aligned}\hat{\beta}_0 &= (X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}\mathbf{y}\\ \widehat{\mathbf{y}} &= X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}\mathbf{y}\end{aligned}$$

其中,$X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}$ 是个 $n\times n$ 的对称矩阵, 被称为帽子矩阵 H.

所以, 所谓帽子矩阵, 其实是 M-P 逆在列满秩时,AA^+ 的特殊情况.

这个矩阵是一个对称幂等阵, 即:

- $H^{\mathrm{T}}=H$
- $H^n=H$
- $(I-H)^n=I-H$. 这一点可以展开后由第二点立刻得到.

而我们关心的另一个量 $\mathbf{e}=(I-H)\mathbf{y}$.

6.4 例题

已知一些测量的自变量-因变量值:

x	-1	0	1	2
y	3	1	0	1

从先验知识知道,y 应该是 x 的二次函数, 求参数的最小二乘解和此时对 y 的最小二乘估计.

答案: 假设 $y = ax^2 + bx + c$, 构造数据矩阵:

$$X_0 = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \end{bmatrix}, \beta_0 = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

计算得到:

$$(X_0^{\mathrm{T}} X_0)^{-1} = \begin{bmatrix} 0.25 & -0.25 & -0.25 \\ -0.25 & 0.45 & 0.15 \\ -0.25 & 0.15 & 0.55 \end{bmatrix}$$

$$H = \begin{bmatrix} 0.95 & 0.15 & -0.15 & 0.05 \\ 0.15 & 0.55 & 0.45 & -0.15 \\ -0.15 & 0.45 & 0.55 & 0.15 \\ 0.05 & -0.15 & 0.15 & 0.95 \end{bmatrix}$$

所以:

$$\beta_0 = \begin{bmatrix} 0.75 \\ -1.45 \\ 0.85 \end{bmatrix}, \widehat{\mathbf{y}} = \begin{bmatrix} 3.05 \\ 0.85 \\ 0.15 \\ 0.95 \end{bmatrix}$$

参考图如下:

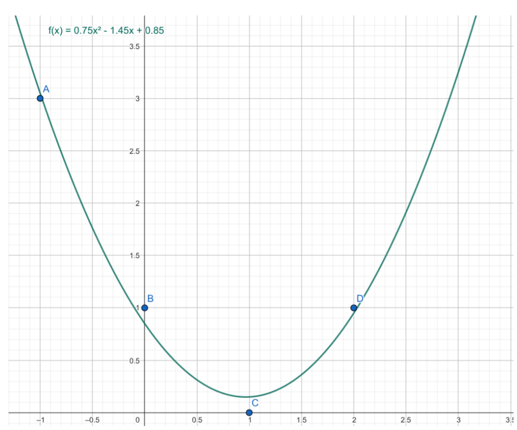

图 2: 最小二乘