
Least Squares Method: A Brief Introduction

inlmouse
Glassix Research
March 3, 2020
inlmouse@glassix.com

Abstract

在众多科学与工程学科中，如物理、化工、统计学、经济学、生物学、信号处理、自动控制、系统理论、医学和军事工程中等，许多问题都可以用数学建模成矩阵方程 $\mathbf{Ax} = \mathbf{b}$ 。根据数据向量 $\mathbf{b} \in \mathbb{R}^{m \times 1}$ 和数据矩阵 $\mathbf{A} \in \mathbb{R}^{m \times n}$ 的不同可以分为三种主要类型：

1. 超定矩阵方程： $m > n$ ，并且数据向量 \mathbf{b} 和数据矩阵 \mathbf{A} 已知，其中之一或者二者可能存在误差；
2. 盲矩阵方程：仅数据向量 \mathbf{b} 已知，数据矩阵 \mathbf{A} 未知；
3. 欠定稀疏矩阵方程： $m < n$ ，并且数据向量 \mathbf{b} 和数据矩阵 \mathbf{A} 已知，但未知向量 \mathbf{x} 为稀疏向量；

这里我们主要讨论第一种情况。

1 普通最小二乘法 (Ordinary Least Squares Method)

考虑矩阵方程 $\mathbf{Ax} = \mathbf{b}$ ，其中 $\mathbf{b} \in \mathbb{R}^{m \times 1}$ 为数据向量， $\mathbf{A} \in \mathbb{R}^{m \times n}$ 为数据矩阵，并且 $m > n$ 。

假定数据向量存在加性观测误差或者噪声，即 $\mathbf{b} = \mathbf{b}_0 + \mathbf{e}$ ，其中 \mathbf{b}_0 和 \mathbf{e} 分别是无误差数据向量和误差向量。

为了抵消误差对矩阵方程求解的影响，我们引入一个矫正向量 $\Delta \mathbf{b}$ ，并用它是“扰动”有误差的数据向量 \mathbf{b} 。我们的目标是，是矫正项 $\Delta \mathbf{b}$ “尽可能小”。同时令 $\mathbf{Ax} = \mathbf{b} + \Delta \mathbf{b}$ 补偿存在于数据向量中的不确定性，使得 $\mathbf{b} + \Delta \mathbf{b} = \mathbf{b}_0 + \mathbf{e} + \Delta \mathbf{b} \rightarrow \mathbf{b}_0$ ，从而实现：

$$\mathbf{Ax} = \mathbf{b} + \Delta \mathbf{b} \implies \mathbf{Ax} = \mathbf{b}_0 \quad (1)$$

的转换¹。也即，如果直接选择矫正向量 $\Delta \mathbf{b} = \mathbf{Ax} - \mathbf{b}$ ，并且使其“尽可能小”，则能实现对方程 $\mathbf{Ax} = \mathbf{b}_0$ 的最小误差求解。

这一求解思路使用优化理论的角度²进行描述即是：

$$\min_{\mathbf{x}} \|\Delta \mathbf{b}\|^2 = \|\mathbf{Ax} - \mathbf{b}\|_2^2 = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) \quad (2)$$

这一方法称普通最小二乘法 (Ordinary Least Squares Method)，简称最小二乘法。

于是矩阵方程 $\mathbf{Ax} = \mathbf{b}$ 的最小二乘解为：

$$\hat{\mathbf{x}}_{LS} = \arg \min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 \quad (3)$$

¹ 上式也称线性模型的正规方程，线性模型的正规方程必有解。

² 这里值得注意的是优化理论的角度上，截至现在的描述使用不同的范数实际可以达成同一目的，L2 范数不是必须的。除了 L1 范数不可导而外，其导致的稀疏解对超定任务并不友好。至于选取其他范数导致的不同结果不在此次讨论范围内。

为考察上式解析解，展开式2并对 \mathbf{x} 求导³，并令其结果等于零：

$$\frac{d(\mathbf{x}^T \mathbf{A}^T \mathbf{A} \mathbf{x} - \mathbf{x}^T \mathbf{A}^T \mathbf{b} - \mathbf{b}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{b})}{d\mathbf{x}} = 2\mathbf{A}^T \mathbf{A} \mathbf{x} - 2\mathbf{A}^T \mathbf{b} = 0 \quad (4)$$

则 \mathbf{x}_{LS} 分两种情况：

1. 列满秩，即 $\text{Rank}(\mathbf{A}) = n$ ：则 $\mathbf{A}^T \mathbf{A}$ 非奇异，方程唯一解：

$$\mathbf{x}_{LS} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \quad (5)$$

2. 秩亏缺，即 $\text{Rank}(\mathbf{A}) < n$ ：方程有解⁴：

$$\mathbf{x}_{LS} = (\mathbf{A}^T \mathbf{A})^\dagger \mathbf{A}^T \mathbf{b} \quad (6)$$

2 Gauss-Markov Theorem

在数理统计中的参数估计理论中，称参数向量 $\boldsymbol{\theta}$ 的估计 $\hat{\boldsymbol{\theta}}$ 为无偏估计，若它的期望值等于未知参数向量的真实值，即 $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ 。进一步，如果一个无偏估计还具有最小方差⁵，则称这一估计为最优无偏估计。

同理，对于数据向量 \mathbf{b} 含有噪声的超定方程 $\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{e}$ ，若最小二乘解的数学期望等于真实参数向量，则称最小二乘解是无偏的。如果其还具有最小方差，则称最小二乘解是最优无偏的。

定理 1 (Gauss-Markov Theorem[1])。考虑矩阵方程：

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{e} \quad (7)$$

其中随机误差向量 $\mathbf{e} = (e_1, \dots, e_m)^T$ 的均值向量和协方差矩阵分别为：

$$E(\mathbf{e}) = \mathbf{0}, \text{Cov}(\mathbf{e}) = E(\mathbf{e}\mathbf{e}^T) = \sigma^2 \mathbf{I} \quad (8)$$

则，当且仅当 $\text{Rank}(\mathbf{A}) = n$ 时，参数向量 \mathbf{x} 的最优无偏解存在，且由最小二乘解给定：

$$\mathbf{x}_{LS} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \quad (9)$$

其方差：

$$\text{Var}(\mathbf{x}_{LS}) \leq \text{Var}(\tilde{\mathbf{x}}) \quad (10)$$

上式中 $\tilde{\mathbf{x}}$ 为方程7的任一其他解。

Proof. 显然：

$$E(\mathbf{b}) = E(\mathbf{A}\mathbf{x}) - E(\mathbf{e}) = \mathbf{A}\mathbf{x} \quad (11)$$

由于 $\text{Rank}(\mathbf{A}) = n$ ，则 $\mathbf{A}^T \mathbf{A}$ 非奇异⁶，由5式：

$$E(\hat{\mathbf{x}}_{LS}) = E((\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}) = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T E(\mathbf{b}) = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{x}$$

因此 $\hat{\mathbf{x}}_{LS}$ 无偏。下面证明 $\hat{\mathbf{x}}_{LS}$ 具有最小方差。

假定 \mathbf{x} 还有一候补解 $\tilde{\mathbf{x}}$ ，则可表示为：

$$\tilde{\mathbf{x}} = \hat{\mathbf{x}}_{LS} + \mathbf{C}\mathbf{b} + \mathbf{d}$$

其中 \mathbf{C} 和 \mathbf{d} 分别为常数矩阵和向量。首先要求 $\tilde{\mathbf{x}}$ 是无偏的，则下式恒成立：

$$\forall \mathbf{x}, E(\tilde{\mathbf{x}}) = \hat{\mathbf{x}}_{LS} + E(\mathbf{C}\mathbf{b}) + \mathbf{d} = \mathbf{x} + \mathbf{C}\mathbf{A}\mathbf{x} + \mathbf{d} = \mathbf{x}$$

³求导算子： $\frac{d}{d\mathbf{x}} = (\frac{d}{dx_1}, \dots, \frac{d}{dx_n})^T$ 。

⁴† 指 Moore-Penrose 伪逆。这个解也满足非一致方程的最小范数最小二乘解。

⁵即是以概率为 1 地等于 Cramer-Rao 下界。

⁶显然 $\mathbf{A}\mathbf{x} = \mathbf{0}$ 和 $\mathbf{A}^T \mathbf{A}\mathbf{x} = \mathbf{0}$ 同解，也即 $\text{Rank}(\mathbf{A}) = \text{Rank}(\mathbf{A}^T \mathbf{A})$ 。

当且仅当：

$$\mathbf{CA} = \mathbf{O}, d = 0 \quad (12)$$

时成立。于是：

$$\begin{aligned} \text{Cov}(\tilde{\mathbf{x}}) &= \text{Cov}(\hat{\mathbf{x}}_{LS} + \mathbf{Cb}) = E[(\hat{\mathbf{x}}_{LS} - \mathbf{x}) + \mathbf{Cb}][(\hat{\mathbf{x}}_{LS} - \mathbf{x}) + \mathbf{Cb}]^T \\ &= \text{Cov}(\hat{\mathbf{x}}_{LS}) + E((\hat{\mathbf{x}}_{LS} - \mathbf{x})(\mathbf{Cb})^T) + E(\mathbf{Cb}(\hat{\mathbf{x}}_{LS} - \mathbf{x})^T) + E(\mathbf{Cbb}^T \mathbf{C}^T) \end{aligned}$$

其中：

$$E(\mathbf{bb}^T) = E(\mathbf{Axx}^T \mathbf{A}^T) + E(\mathbf{ee}^T) = \mathbf{Axx}^T \mathbf{A}^T + \sigma^2 \mathbf{I}$$

则有：

$$\begin{aligned} E((\hat{\mathbf{x}}_{LS} - \mathbf{x})(\mathbf{Cb})^T) &= E((\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{bb}^T \mathbf{C}^T) - E(\mathbf{x} \mathbf{b}^T \mathbf{C}^T) \\ &= (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T E(\mathbf{bb}^T) \mathbf{C}^T - \mathbf{x} E(\mathbf{b}^T) \mathbf{C}^T \\ &= \mathbf{O} = \mathbf{O}^T \\ &= E(\mathbf{Cb}(\hat{\mathbf{x}}_{LS} - \mathbf{x})^T) \end{aligned}$$

$$E(\mathbf{Cbb}^T \mathbf{C}^T) = \mathbf{CE}(\mathbf{bb}^T) \mathbf{C}^T = \sigma^2 \mathbf{CC}^T$$

故：

$$\text{Cov}(\tilde{\mathbf{x}}) = \text{Cov}(\hat{\mathbf{x}}_{LS}) + \sigma^2 \mathbf{CC}^T \quad (13)$$

对上式取迹⁷，并注意到零均值向量 $\text{tr}(\text{Cov}(\mathbf{x})) = \text{Var}(\mathbf{x})$ ，于是上式改写为：

$$\text{Var}(\tilde{\mathbf{x}}) = \text{Var}(\hat{\mathbf{x}}_{LS}) + \sigma^2 \text{tr}(\mathbf{CC}^T) \geq \text{Var}(\hat{\mathbf{x}}_{LS})$$

即是 $\hat{\mathbf{x}}_{LS}$ 为最优无偏解。 □

注意 GM 定理的条件 $\text{Cov}(\mathbf{e}) = \sigma^2 \mathbf{I}$ 意味着误差 \mathbf{e} 的各个分量互不相关，并且具有相同的方差 σ^2 。只有在这种情况下，最小二乘解才是无偏和最优的，否则即是 PCA 解是最优的。

3 普通最小二乘解与最大似然解的等价性

如果误差向量 \mathbf{e} 各个分量均为 iid 的高斯随机变量，则其概率密度函数：

$$f(\mathbf{e}) = \frac{1}{\pi^m |\boldsymbol{\Sigma}_e|} \exp[-(\mathbf{e} - \boldsymbol{\mu}_e)^T \boldsymbol{\Sigma}_e^{-1} (\mathbf{e} - \boldsymbol{\mu}_e)] \quad (14)$$

其中协方差矩阵 $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$ 。

在定理1的条件下，上式简化为：

$$f(\mathbf{e}) = \frac{1}{(\pi\sigma^2)^m} \exp[-\frac{1}{\sigma^2} \mathbf{e}^T \mathbf{e}] = \frac{1}{(\pi\sigma^2)^m} \exp[-\frac{1}{\sigma^2} \|\mathbf{e}\|_2^2] \quad (15)$$

其似然函数：

$$L(\mathbf{e}) = \ln(f(\mathbf{e})) = -m \ln \pi - 2m \ln \sigma - \frac{\|\mathbf{e}\|_2^2}{\sigma^2} = -m \ln \pi - 2m \ln \sigma - \frac{\|\mathbf{Ax} - \mathbf{b}\|_2^2}{\sigma^2} \quad (16)$$

于是矩阵方程的最大似然解：

$$\hat{\mathbf{x}}_{ML} = \arg \max_{\mathbf{x}} -\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \arg \min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 = \hat{\mathbf{x}}_{LS} \quad (17)$$

也即是在定理1的条件矩阵方程 $\mathbf{Ax} = \mathbf{b}$ 的最大似然解与最小二乘解等价。

References

[1] S. D. Silvey. *Statistical inference*. 1970.

⁷ $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.