Least Squares Method: A Brief Introduction

Accelouch
Glasssix Research
July 26,2019

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 $\boldsymbol{Ax}=\boldsymbol{b}$。根据数据向量 $\boldsymbol{b}\in\mathbb{R}^{m\times 1}$ 和数据矩阵 $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ 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 $\boldsymbol{b}$ 和数据矩阵 $\boldsymbol{A}$ 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 $\boldsymbol{b}$ 已知，数据矩阵 $\boldsymbol{A}$ 未知；
3. 欠定稀疏矩阵方程：$m<n$, 并且数据向量 $\boldsymbol{b}$ 和数据矩阵 $\boldsymbol{A}$ 已知，但未知向量 $\boldsymbol{x}$ 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 $\boldsymbol{Ax}=\boldsymbol{b}$，其中 $\boldsymbol{b}\in\mathbb{R}^{m\times 1}$ 为数据向量，$\boldsymbol{A}\in\mathbb{R}^{m\times n}$ 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 $\boldsymbol{b}=\boldsymbol{b_0}+\boldsymbol{e}$，其中 $\boldsymbol{b_0}$ 和 $\boldsymbol{e}$ 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 $\boldsymbol{\Delta b}$，并用它是“扰动”有误差的数据向量 $\boldsymbol{b}$。我们的目标是，是矫正项 $\boldsymbol{\Delta b}$“尽可能小”。同时令 $\boldsymbol{Ax}=\boldsymbol{b}+\boldsymbol{\Delta b}$ 补偿存在于数据向量中的不确定性，使得 $\boldsymbol{b}+\boldsymbol{\Delta b}=\boldsymbol{b_0}+\boldsymbol{e}+\boldsymbol{\Delta b}\rightarrow\boldsymbol{b_0}$，从而实现：

$$\boldsymbol{Ax}=\boldsymbol{b}+\boldsymbol{\Delta b}\implies\boldsymbol{Ax}=\boldsymbol{b_0}\tag{1}$$

的转换[1]。也即，如果直接选择矫正向量 $\boldsymbol{\Delta b}=\boldsymbol{Ax}-\boldsymbol{b}$，并且使其“尽可能小”，则能实现对方程 $\boldsymbol{Ax}=\boldsymbol{b_0}$ 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：

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

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 $\boldsymbol{Ax}=\boldsymbol{b}$ 的最小二乘解为：

$$\hat{\boldsymbol{x}}_{LS}=\arg\min_{\boldsymbol{x}}\|\boldsymbol{Ax}-\boldsymbol{b}\|_2^2\tag{3}$$

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

$$\frac{d(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Ax}-\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{b}-\boldsymbol{b}^T\boldsymbol{Ax}+\boldsymbol{b}^T\boldsymbol{b})}{d\boldsymbol{x}}=2\boldsymbol{A}^T\boldsymbol{Ax}-2\boldsymbol{A}^T\boldsymbol{b}=0\tag{4}$$

[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。

[3]求导算子：$\frac{d\cdot}{d\boldsymbol{x}}=(\frac{d\cdot}{dx_1},\cdots,\frac{d\cdot}{dx_n})^T$。

Glasssix Tech Report 2019.

则 $\boldsymbol{x}_{LS}$ 分两种情况：

1. 列满秩，即 $Rank(\boldsymbol{A}) = n$：则 $\boldsymbol{A}^T\boldsymbol{A}$ 非奇异，方程唯一解：
$$\boldsymbol{x}_{LS} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} \tag{5}$$
2. 秩亏缺，即 $Rank(\boldsymbol{A}) < n$：方程有解[4]：
$$\boldsymbol{x}_{LS} = (\boldsymbol{A}^T\boldsymbol{A})^{\dagger}\boldsymbol{A}^T\boldsymbol{b} \tag{6}$$

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 $\boldsymbol{b}$ 含有噪声的超定方程 $\boldsymbol{Ax} = \boldsymbol{b} + \boldsymbol{e}$，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：*
$$\boldsymbol{Ax} = \boldsymbol{b} + \boldsymbol{e} \tag{7}$$
其中随机误差向量 $\boldsymbol{e} = (e_1, \cdots, e_m)^T$ *的均值向量和协方差矩阵分别为：*
$$E(\boldsymbol{e}) = \boldsymbol{0}, Cov(\boldsymbol{e}) = E(\boldsymbol{ee}^T) = \sigma^2\boldsymbol{I} \tag{8}$$
则，当且仅当 $Rank(\boldsymbol{A}) = n$ *时，参数向量* $\boldsymbol{x}$ *的最优无偏解存在，且由最小二乘解给定：*
$$\boldsymbol{x}_{LS} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} \tag{9}$$
其方差：
$$Var(\boldsymbol{x}_{LS}) \leq Var(\tilde{\boldsymbol{x}}) \tag{10}$$
上式中 $\tilde{\boldsymbol{x}}$ *为方程7的任一其他解。*

Proof. 显然：
$$E(\boldsymbol{b}) = E(\boldsymbol{Ax}) - E(\boldsymbol{e}) = \boldsymbol{Ax} \tag{11}$$
由于 $Rank(\boldsymbol{A}) = n$，则 $\boldsymbol{A}^T\boldsymbol{A}$ 非奇异[6]，由5式：
$$E(\hat{\boldsymbol{x}}_{LS}) = E((\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}) = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T E(\boldsymbol{b}) = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Ax} = \boldsymbol{x}$$
因此 $\hat{\boldsymbol{x}}_{LS}$ 无偏。下面证明 $\hat{\boldsymbol{x}}_{LS}$ 具有最小方差。

假定 $\boldsymbol{x}$ 还有一候补解 $\tilde{\boldsymbol{x}}$，则可表示为：
$$\tilde{\boldsymbol{x}} = \hat{\boldsymbol{x}}_{LS} + \boldsymbol{Cb} + \boldsymbol{d}$$
其中 $\boldsymbol{C}$ 和 $\boldsymbol{d}$ 分别为常数矩阵和向量。首先要求 $\tilde{\boldsymbol{x}}$ 是无偏的，则下式恒成立：
$$\forall \boldsymbol{x}, E(\tilde{\boldsymbol{x}}) = \hat{\boldsymbol{x}}_{LS} + E(\boldsymbol{Cb}) + \boldsymbol{d} = \boldsymbol{x} + \boldsymbol{CAx} + \boldsymbol{d} = \boldsymbol{x}$$
当且仅当：
$$\boldsymbol{CA} = \boldsymbol{O}, \boldsymbol{d} = \boldsymbol{0} \tag{12}$$
时成立。于是：
$$\begin{aligned} Cov(\tilde{\boldsymbol{x}}) &= Cov(\hat{\boldsymbol{x}}_{LS} + \boldsymbol{Cb}) = E([(\hat{\boldsymbol{x}}_{LS} - \boldsymbol{x}) + \boldsymbol{Cb}][(\hat{\boldsymbol{x}}_{LS} - \boldsymbol{x}) + \boldsymbol{Cb}]^T) \\ &= Cov(\hat{\boldsymbol{x}}_{LS}) + E((\hat{\boldsymbol{x}}_{LS} - \boldsymbol{x})(\boldsymbol{Cb})^T) + E(\boldsymbol{Cb}(\hat{\boldsymbol{x}}_{LS} - \boldsymbol{x})^T) + E(\boldsymbol{Cbb}^T\boldsymbol{C}^T) \end{aligned}$$

[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 $\boldsymbol{Ax} = 0$ 和 $\boldsymbol{A}^T\boldsymbol{Ax} = 0$ 同解，也即 $Rank(\boldsymbol{A}) = Rank(\boldsymbol{A}^T\boldsymbol{A})$。

其中:

$$E(\boldsymbol{b}\boldsymbol{b}^T) = E(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T) + E(\boldsymbol{e}\boldsymbol{e}^T) = \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T + \sigma^2\boldsymbol{I}$$

则有:

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

$$E(\boldsymbol{C}\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{C}^T) = \boldsymbol{C}E(\boldsymbol{b}\boldsymbol{b}^T)\boldsymbol{C}^T = \sigma^2\boldsymbol{C}\boldsymbol{C}^T$$

故:

$$Cov(\tilde{\boldsymbol{x}}) = Cov(\hat{\boldsymbol{x}}_{LS}) + \sigma^2\boldsymbol{C}\boldsymbol{C}^T \tag{13}$$

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

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

即是 $\hat{\boldsymbol{x}}_{LS}$ 为最优无偏解。 □

注意 GM 定理的条件 $Cov(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$ 意味着误差 $\boldsymbol{e}$ 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 $\boldsymbol{e}$ 各个分量均为 iid 的高斯随机变量，则其概率密度函数:

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

其中协方差矩阵 $\boldsymbol{\Sigma}_e = diag(\sigma_1^2, \cdots, \sigma_m^2)$。
在定理1的条件下，上式简化为:

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

其似然函数:

$$L(\boldsymbol{e}) = \ln(f(\boldsymbol{e})) = -m\ln\pi - 2m\ln\sigma - \frac{\|\boldsymbol{e}\|_2^2}{\sigma^2} = -m\ln\pi - 2m\ln\sigma - \frac{\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2}{\sigma^2} \tag{16}$$

于是矩阵方程的最大似然解:

$$\hat{\boldsymbol{x}}_{ML} = \arg\max_{\boldsymbol{x}} -\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \arg\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \hat{\boldsymbol{x}}_{LS} \tag{17}$$

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

References

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

[7] $tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$.