基于最小二乘近似的模型平均方法

张新雨

中国科学院数学与系统科学研究院, 北京 100190

E-mail: xinyu@amss.ac.cn

收稿日期: 2019-05-18; 接受日期: 2019-06-12; 网络出版日期: 2020-03-13

国家自然科学基金 (批准号: 71522004, 11471324 和 71631008) 资助项目

摘要 本文提出基于最小二乘近似的模型平均方法. 该方法可用于线性模型、广义线性模型和分位数回归等各种常用模型. 特别地, 经典的 Mallows 模型平均方法是该方法的特例. 现存的模型平均文献中, 渐近分布的证明一般需要局部误设定假设, 所得的极限分布的形式也比较复杂. 本文将在不使用局部误设定假设的情形下证明该方法的渐近正态性. 另外, 本文将该方法发展到维数发散的情形. 数值结果显示了该方法的优势.

关键词 渐近正态 相合性 最小二乘近似 模型平均 权重选择

MSC (2010) 主题分类 62F10, 62F99, 62H12

1 引言

作为变量选择或者模型选择的替代, 模型平均方法近年来受到了越来越多的关注. 模型平均通常可以降低估计风险, 避免忽略有用的信息, 类似的讨论参见文献 [1, 2]. 特别地, 当潜在的模型噪声大时, 模型平均在提高估计精度方面的优势将更加明显, 参见文献 [3, 4].

在过去三十年, Bayes 模型平均已经取得长足的发展, 有大量的文献关注这方面的研究, 参见综述性文献 [5]. 本文关注频率模型平均方法, 该方法不需要任何先验信息. 频率模型平均方法已经发展到各类模型下, 如线性模型 [6–10]、结构变化模型 [11]、缺失数据模型 [12, 13]、混合效应模型 [14]、因子模型 [15]、分位数回归模型 [16]、广义线性模型 [17]、部分线性模型 [18] 等. 然而在这些工作中, 预测是主要的关注点, 参数推断被忽略了. 研究基于模型平均的参数推断是一个很困难的问题, 因为权重是随机的且没有显式的表达式. 本文将通过推导权重的极限属性来解决这个难题.

本文将提出一种基于最小二乘近似 (least squares approximation, LSA) 的模型平均方法. LSA 由 Wang 和 Leng [19] 提出, 他们使用最小二乘型函数近似原始的损失函数, 从而统一了不同模型框架下的罚估计方法, 如线性模型、广义线性模型和分位数回归模型等, 具有广泛的适用性. 通过组合模型平均和 LSA 技术, LSA 的广泛适用性将被延续下来, 即新的模型平均也可以用在各种模型框架下. 特别

英文引用格式: Zhang X Y. Model averaging by least squares approximation (in Chinese). Sci Sin Math, 2021, 51: 535–548, doi: 10.1360/N012019-00137

地, 经典的 Mallows 模型平均方法将是新方法的一个特例. 另外, 我们提出使用惩罚的 LSA 的解路径来准备模型平均的候选模型, 这使得计算负担大为降低.

本文的另一个贡献在于我们推出所得模型平均估计的渐近正态性, 这使得统计推断比很多文献中所推出的复杂极限分布下的统计推断更加简单, 而且这些文献中还使用了局部误设定假设 (参见文献 [20–27]). 正如文献 [28] 中评论, 尽管利用这个假设能够得到风险的近似, 但是它本身有固有的缺陷, 例如, 需要候选模型必须介于一些特殊的模型之间. 文献 [28] 在没有使用该假设的前提下提出了一种在参数模型与非参数模型之间进行选择的方法. 最近, 在没有使用该假设的前提下, 文献 [29] 建立了最小二乘模型平均估计的分布理论, 而文献 [30] 提出了一个平均极大似然估计的框架.

本文的结构安排如下. 第 2 节提出模型平均估计方法. 具体地, 先介绍估计步骤, 然后建立我们的方法与 Mallows 模型平均方法之间的联系; 接下来研究权重的收敛性, 并在此基础上建立渐近分布理论. 第 3 节展示模拟和实际例子的结果. 第 4 节将该方法发展到高维情形. 第 5 节是总结. 定理的证明均在附录中.

2 模型平均估计

2.1 记号和估计步骤

记 $\beta = (\beta_1, \dots, \beta_p)^T$, 它是所关注的参数向量. 记 $\mathcal{A} = \{j : \beta_{0j} \neq 0\} = \{\mathcal{A}_1, \dots, \mathcal{A}_q\}$, 它是非零参数的指标集合. 这里暂时假设 p 是固定的, 后文将考虑 p 发散的情形. 令 $\beta_0 = (\beta_{0,1}, \dots, \beta_{0,p})^T$ 为 β 的真实值. 令 n 为样本量. 通过一种估计方法, 如极大似然估计或者最小绝对距离估计, 我们可以得到 β_0 的估计, 记为 $\tilde{\beta}$. 假设 $\tilde{\beta}$ 是 $\sqrt{n}$ 相合的并且有如下的渐近正态性:

$$\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} \text{Normal}(0, \Sigma), \quad (2.1)$$

其中 $\xrightarrow{d}$ 表示依分布收敛, Σ 是渐近协方差矩阵. 本文中的收敛性均是在 n 趋于无穷意义下的. 这个假设对许多常用的估计方法均成立. 令 $\hat{\Sigma}$ 为 Σ 的估计. 我们需要 $\hat{\Sigma}$ 收敛到一个正定矩阵, 但收敛性 $\hat{\Sigma} \xrightarrow{p} \Sigma$ 不是必需的, 这里 $\xrightarrow{p}$ 表示依概率收敛. 我们提出以下 5 步的模型平均方法.

步骤 1 通过极小化以下 LSA 版本的自适应 LASSO (least absolute shrinkage and selection operator) [31] 得到解的路径:

$$Q(\beta) = (\beta - \tilde{\beta})^T \hat{\Sigma}^{-1} (\beta - \tilde{\beta}) + \rho \sum_{j=1}^p |\tilde{\beta}_j|^{-\gamma} |\beta_j|, \quad (2.2)$$

其中 $\tilde{\beta}_j$ 是 $\tilde{\beta}$ 的第 j 个元素, ρ 是关键参数, γ 也是预先设定参数. 在本文的数值部分, 我们选择 $\gamma = 1$. 这个最小化可以由文献 [32] 中的 LARS (least angle regression) 算法实现. 文献 [19] 指出, 当不考虑一个与 β 无关的项时, 公式 (2.2) 是以下罚损失函数 $2\mathcal{L}_n(\beta)/n + \rho \sum_{j=1}^p |\tilde{\beta}_j|^{-\gamma} |\beta_j|$ 的一个合理的近似, 其中 $\mathcal{L}_n(\beta)$ 是损失函数.

步骤 2 利用 BIC (Bayesian information criterion) 准则在解的路径中选择一个估计. 令 $\hat{\mathcal{A}}$ 为由 BIC 选择的估计的参数指标集. 在评论 2 中, 我们将讨论在 BIC 选择出一个估计之后我们仍然组合其他估计的原因.

步骤 3 令 d 为集合 $\hat{\mathcal{A}}$ 中元素个数. 按照这些元素进入 LARS 算法解路径的时间, 对这些元素进行排序. 记 $\hat{\mathcal{A}}^* = \{\hat{\mathcal{A}}_1^*, \dots, \hat{\mathcal{A}}_d^*\}$ 为这些元素最终的排序, 由此构造 d 个嵌套模型, 第 s 个模型使用 $\beta_s^* = (\beta_{\hat{\mathcal{A}}_1^*}, \dots, \beta_{\hat{\mathcal{A}}_s^*})^T$ 作为未知参数向量. 令 Π_s 为投影矩阵使得 $\beta_s^* = \Pi_s \beta$. 令 $\hat{B}_s = \hat{\Sigma}^{-1/2} \Pi_s^T$, 它

是 $\widehat{\Sigma}^{-1/2}$ 的子矩阵, 令 $\mathbf{z} = \widehat{\Sigma}^{-1/2} \widetilde{\beta}$. 在第 s 个候选模型中 ($s \in \{1, \dots, d-1\}$), 我们通过如下公式估计 β_s^* :

$$\widehat{\beta}_s^* = (\widehat{B}_s^T \widehat{B}_s)^{-1} \widehat{B}_s^T \mathbf{z}. \quad (2.3)$$

那么, 在第 s 个候选模型下, β_0 的估计为

$$\widehat{\beta}_s = \Pi_s^T \widehat{\beta}_s^*. \quad (2.4)$$

步骤 4 除了步骤 3 中的嵌套候选模型, 我们也使用解路径中的包含 $\widehat{\mathcal{A}}$ 中所有元素的解构造候选模型. 对第 s 个解, 我们使用 $\widehat{\mathcal{A}}_s = \{\widehat{\mathcal{A}}_{s,1}, \dots, \widehat{\mathcal{A}}_{s,p_s}\}$ 作为非零估计的指标集的集合. 第 s 个解对应于一个候选模型, 在这个候选模型中未知参数向量为 $(\beta_{\widehat{\mathcal{A}}_{s,1}}, \dots, \beta_{\widehat{\mathcal{A}}_{s,p_s}})$. 我们使用计算 $\widetilde{\beta}$ 的方式来估计这些模型中的参数向量. 记 $\widehat{\beta}_d, \dots, \widehat{\beta}_S$ 为来自这些模型中 β 的估计.

步骤 5 令 $\mathbf{w} = (w_1, \dots, w_S)^T$ 为权重向量, 它属于集合

$$\mathcal{W} = \left\{ \mathbf{w} \in [0, 1]^S : \sum_{s=1}^S w_s = 1 \right\}.$$

那么, β_0 的模型平均估计的形式为

$$\widehat{\beta}(\mathbf{w}) = \sum_{s=1}^S w_s \widehat{\beta}_s.$$

令 v_s 为第 s 个模型中未知参数的个数. 记 $\mathbf{v} = (v_1, \dots, v_S)^T$. 我们通过极小化以下准则来选择最终的权重向量:

$$C_n(\mathbf{w}) = \{\widehat{\beta}(\mathbf{w}) - \widetilde{\beta}\}^T \widehat{\Sigma}^{-1} \{\widehat{\beta}(\mathbf{w}) - \widetilde{\beta}\} + \phi_n \frac{\mathbf{w}^T \mathbf{v}}{n}, \quad (2.5)$$

即

$$\widehat{\mathbf{w}} = (\widehat{w}_1, \dots, \widehat{w}_S)^T = \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmin}} C_n(\mathbf{w}). \quad (2.6)$$

关键参数 ϕ_n 的选择将在第 2.5 小节讨论. 因此, β_0 的最终模型平均估计为 $\widehat{\beta}(\widehat{\mathbf{w}}) = \sum_{s=1}^S \widehat{w}_s \widehat{\beta}_s$.

为了使以上步骤中的候选模型更加清晰, 我们举一个 $p=5$ 的简单例子. 假设通过使用步骤 1, 我们得到解的路径 $\{3\}$ 、 $\{3, 1\}$ 、 $\{3, 1, 5\}$ 、 $\{3, 1, 5, 2\}$ 和 $\{3, 1, 5, 2, 4\}$, 这里整数表示非零参数的指标. 假设通过步骤 2, 我们选择了解 $\{3, 1, 5\}$. 因此, $d=3$ 且 $\widehat{\mathcal{A}}^* = \{3, 1, 5\}$. 那么由步骤 3 确定的候选模型包括模型 $\{3\}$ 和 $\{3, 1\}$. 在步骤 4 中得到候选模型 $\{3, 1, 5\}$ 、 $\{3, 1, 5, 2\}$ 和 $\{3, 1, 5, 2, 4\}$, 因为它们都包括步骤 2 中选择的第 3、1 和 5 个变量.

评论 1 在步骤 4 选择的候选模型中, 我们使用计算 $\widetilde{\beta}$ 的方式来估计未知参数, 这种估计方法可以是极大似然估计, 也可以是最小绝对距离估计等. 然后, 在步骤 3 选择的候选模型中, 我们使用 (2.3) 和 (2.4) 来估计 β , 因为用这种方式使得最终的模型平均估计的渐近分布可以推出.

评论 2 除了步骤 2 中选择的候选模型, 我们也使用步骤 3 和 4 中选择的更大的和更小的模型. 这里有两个原因. 第一, 在有限样本下, $\widehat{\mathcal{A}}$ 或许与 $\mathcal{A}$ 不同, 那么使用步骤 3 和 4 中选择出的更多的模型可以保留更多的变量信息. 第二, 使用更多的模型可以提供一种保险机制, 防止选择出很差的模型.

在准则 $C_n(\mathbf{w})$ 中, 第一项衡量了平均估计 $\widehat{\beta}(\mathbf{w})$ 与原始估计 $\widetilde{\beta}$ 之间的不同. 第二项是对加权的模型大小的惩罚. 最后计算 $\widehat{\mathbf{w}}$ 是非常简单的. 令 $\widehat{\mathbf{g}}_s = \widehat{\beta}_s - \widetilde{\beta}$, $\widehat{\mathbf{G}} = (\widehat{\mathbf{g}}_1, \dots, \widehat{\mathbf{g}}_{q_n})$, 那么 $C_n(\mathbf{w}) = \mathbf{w}^T \widehat{\mathbf{G}}^T \widehat{\mathbf{G}} \mathbf{w} + \phi_n \mathbf{w}^T \mathbf{v}$. 因而, 最小化 $C_n(\mathbf{w})$ 是一个二次规划问题, 许多软件包可以快速地解这个问题.

2.2 与 Mallows 模型平均的联系

本小节将展示经典的 Mallows 模型平均 (Mallows model averaging, MMA) 方法是我们的方法的一个特例. 考虑线性回归模型

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \quad (2.7)$$

其中 $\mathbf{y}$ 是响应变量, $\mathbf{X}$ 是 $n \times p$ 的解释变量矩阵, $\boldsymbol{\epsilon}$ 是均值为 0 方差为 σ^2 的独立同分布的随机误差构成的向量. 对第 s 个候选模型, 令 $\mathbf{X}_s$ 为 $n \times v_s$ 的解释变量矩阵, 它是 $\mathbf{X}$ 的子矩阵, 那么 Mallows 准则为

$$\mathcal{C}_{\text{Mallows}}(\mathbf{w}) = \{\mathbf{P}(\mathbf{w})\mathbf{y} - \mathbf{y}\}^T \{\mathbf{P}(\mathbf{w})\mathbf{y} - \mathbf{y}\} + 2\hat{\sigma}^2 \mathbf{w}^T \mathbf{v}, \quad (2.8)$$

其中 $\mathbf{P}(\mathbf{w}) = \sum_s w_s \mathbf{P}_s$, $\mathbf{P}_s = \mathbf{X}_s (\mathbf{X}_s^T \mathbf{X}_s)^{-1} \mathbf{X}_s^T$, $\hat{\sigma}^2 = (\mathbf{P}\mathbf{y} - \mathbf{y})^T (\mathbf{P}\mathbf{y} - \mathbf{y}) / (n - p)$, $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. 在线性模型中, $\boldsymbol{\beta}$ 的常用估计为 $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, (2.1) 中的 $\boldsymbol{\Sigma}$ 可以由 $\hat{\boldsymbol{\Sigma}} = n\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}$ 来估计.

对于步骤 4 中的估计, 有

$$\mathbf{X}(\hat{\boldsymbol{\beta}}_s - \tilde{\boldsymbol{\beta}}) = \mathbf{X}\{\boldsymbol{\Pi}_s^T (\mathbf{X}_s^T \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{y} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\} = (\mathbf{P}_s - \mathbf{P})\mathbf{y}. \quad (2.9)$$

对于步骤 3 中的估计, 有

$$\begin{aligned} \mathbf{X}(\hat{\boldsymbol{\beta}}_s - \tilde{\boldsymbol{\beta}}) &= \mathbf{X}\boldsymbol{\Pi}_s^T (\hat{\mathbf{B}}_s^T \hat{\mathbf{B}}_s)^{-1} \hat{\mathbf{B}}_s^T \mathbf{z} - \mathbf{P}\mathbf{y} \\ &= \mathbf{X}\boldsymbol{\Pi}_s^T (\boldsymbol{\Pi}_s \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Pi}_s^T)^{-1} \boldsymbol{\Pi}_s \hat{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\beta}} - \mathbf{P}\mathbf{y} \\ &= \mathbf{P}_s \mathbf{P}\mathbf{y} - \mathbf{P}\mathbf{y} \\ &= (\mathbf{P}_s - \mathbf{P})\mathbf{y}. \end{aligned} \quad (2.10)$$

那么, 由 (2.5)、(2.9) 和 (2.10) 可知, 准则 $\mathcal{C}_n(\mathbf{w})$ 可以写成

$$\begin{aligned} \mathcal{C}_n(\mathbf{w}) &= n^{-1} \hat{\sigma}^{-2} \{\hat{\boldsymbol{\beta}}(\mathbf{w}) - \tilde{\boldsymbol{\beta}}\}^T \mathbf{X}^T \mathbf{X} \{\hat{\boldsymbol{\beta}}(\mathbf{w}) - \tilde{\boldsymbol{\beta}}\} + \phi_n \frac{\mathbf{w}^T \mathbf{v}}{n} \\ &= n^{-1} \hat{\sigma}^{-2} \left\{ \sum_{s=1}^S w_s \hat{\boldsymbol{\beta}}_s - \tilde{\boldsymbol{\beta}} \right\}^T \mathbf{X}^T \mathbf{X} \left\{ \sum_{s=1}^S w_s \hat{\boldsymbol{\beta}}_s - \tilde{\boldsymbol{\beta}} \right\} + \phi_n \frac{\mathbf{w}^T \mathbf{v}}{n} \\ &= n^{-1} \hat{\sigma}^{-2} \left\{ \sum_{s=1}^S w_s (\mathbf{P}_s - \mathbf{P})\mathbf{y} \right\}^T \sum_{s=1}^S w_s (\mathbf{P}_s - \mathbf{P})\mathbf{y} + \phi_n \frac{\mathbf{w}^T \mathbf{v}}{n} \\ &= n^{-1} \hat{\sigma}^{-2} \mathbf{y}^T \{\mathbf{P}(\mathbf{w}) - \mathbf{P}\}^T \{\mathbf{P}(\mathbf{w}) - \mathbf{P}\} \mathbf{y} + \phi_n \frac{\mathbf{w}^T \mathbf{v}}{n} \\ &= n^{-1} \hat{\sigma}^{-2} \mathbf{y}^T \{\mathbf{P}(\mathbf{w}) - \mathbf{I} + \mathbf{I} - \mathbf{P}\}^T \{\mathbf{P}(\mathbf{w}) - \mathbf{I} + \mathbf{I} - \mathbf{P}\} \mathbf{y} + \phi_n \frac{\mathbf{w}^T \mathbf{v}}{n} \\ &= n^{-1} \hat{\sigma}^{-2} \mathbf{y}^T \{\mathbf{P}(\mathbf{w}) - \mathbf{I}\}^T \{\mathbf{P}(\mathbf{w}) - \mathbf{I}\} \mathbf{y} + \phi_n \frac{\mathbf{w}^T \mathbf{v}}{n} \\ &\quad + n^{-1} \hat{\sigma}^{-2} \mathbf{y}^T (\mathbf{I} - \mathbf{P}) \mathbf{y}, \end{aligned} \quad (2.11)$$

其中最后一项与 $\mathbf{w}$ 没有关系. 当 $\phi_n = 2$ 时, 前两项等于 $n^{-1} \hat{\sigma}^{-2} \mathcal{C}_{\text{Mallows}}(\mathbf{w})$. 因此, 在线性模型下, MMA 方法是我们的方法的一个特例.

2.3 权重收敛性

本小节研究权重向量 $\hat{\mathbf{w}}$ 的收敛性, 它依赖于关键参数 ϕ_n . 令 $\mathcal{O}$ 为过拟合模型的指标构成的集合, 即对任意 $s \in \mathcal{O}$, 模型 s 包含所有的非零参数对应的变量, 也包含一个或多个零参数对应的变量. 令 $\mathcal{U}$ 为欠拟合模型的指标构成的集合, 即对任意 $s \in \mathcal{U}$, 存在至少一个非零参数对应的变量没有包含在模型 s 中. 首先建立欠拟合模型的权重的收敛性.

定理 1 如果 $\phi_n = o(n)$, 那么对任何 $s \in \mathcal{U}$, 有

$$\hat{w}_s = O_p\left(\frac{\phi_n}{n}\right). \quad (2.12)$$

定理 1 的证明见附录 A.1. 接下来建立过拟合模型的权重的收敛性质.

定理 2 如果 $\phi_n \rightarrow \infty$ 且 $\phi_n = O(n^{1/2})$, 那么对任意 $s \in \mathcal{O}$, 有

$$\hat{w}_s = O_p(\phi_n^{-1/2}). \quad (2.13)$$

这个定理意味着, 当 $\phi_n \rightarrow \infty$, 准则 $\mathcal{C}_n(\mathbf{w})$ 的第二项, 即 $\mathbf{w}^T \mathbf{v}$, 能够对过拟合模型施加足够的权重, 使得过拟合模型的权重收敛到 0. 定理 2 的证明见附录 A.2.

2.4 渐近分布

模型平均中分布理论的研究总是一个困难的问题. 尽管在局部误设定假设下, 基于信息准则的模型平均估计的渐近分布理论已经得到 (参见文献 [23]), 但是这个分布比较复杂, 使得基于此的统计推断比较困难. 令 $\beta_{0,\mathcal{A}}$ 为真实模型中非零参数向量, 令 $\hat{\beta}_{\mathcal{A}}(\hat{\mathbf{w}})$ 为通过我们的方法得到的 $\beta_{0,\mathcal{A}}$ 的估计. 本节将展示 $\hat{\beta}_{\mathcal{A}}(\hat{\mathbf{w}})$ 的渐近正态性.

假设第 $\tilde{s}$ 个模型是真实模型. 注意 $\tilde{s}$ 是未知的. 令 $\hat{\beta}_{\tilde{s},\mathcal{A}}$ 是来自这个模型的 $\beta_{0,\mathcal{A}}$ 的估计, 显然, 它是 $\hat{\beta}_{\tilde{s}}$ 的子向量. 假设

$$\sqrt{n}(\hat{\beta}_{\tilde{s},\mathcal{A}} - \beta_0) \xrightarrow{d} \text{Normal}(0, \Sigma_{\tilde{s}}), \quad (2.14)$$

这个假设可以由一些常用的正则条件推出. 由 (2.14)、定理 1 和 2 中显示的权重的收敛性可直接得到如下定理.

定理 3 如果 $\phi_n \rightarrow \infty$ 且 $\phi_n = o(n^{1/2})$, 那么

$$\sqrt{n}(\hat{\beta}_{\mathcal{A}}(\hat{\mathbf{w}}) - \beta_{0,\mathcal{A}}) \xrightarrow{d} \text{Normal}(0, \Sigma_{\tilde{s}}).$$

定理 3 意味着在大样本下, 我们的模型平均方法可以取得真实模型下一样的估计效率.

2.5 ϕ_n 的选择

我们推荐 $\phi_n = \log n$. 一个原因是 $\log n$ 满足定理 1-3 中使用的条件 $\phi_n \rightarrow \infty$ 和 $\phi_n = o(n^{1/2})$. 第二个原因是下一节中的数值结果支持这个选择. 第三个原因是这个选择可以把我们的方法与 BIC 建立如下联系.

假设 $\tilde{\beta}$ 是由极大似然方法得到. 记 $\mathcal{L}_n(\beta)$ 为似然函数, 那么第 s 个候选模型的 BIC 为 $\text{BIC}_s = -2\mathcal{L}_n(\hat{\beta}_s) + \log(n)v_s$. 对于模型平均估计 $\hat{\beta}(\mathbf{w})$, 我们可以定义它的 BIC 为 $\text{BIC}(\mathbf{w}) = -2\mathcal{L}\{\hat{\beta}(\mathbf{w})\}$

$+\log(n)\mathbf{w}^T\mathbf{v}$. 通过 Taylor 展开有

$$-2n^{-1}\mathcal{L}\{\hat{\beta}(\mathbf{w})\} \approx -2n^{-1}\mathcal{L}(\tilde{\beta}) + \{\hat{\beta}(\mathbf{w}) - \tilde{\beta}\}^T \left\{ n^{-1} \frac{\partial^2 \mathcal{L}_n(\beta)}{\partial \beta \partial \beta^T} \bigg|_{\beta=\tilde{\beta}} \right\} \{\hat{\beta}(\mathbf{w}) - \tilde{\beta}\}.$$

通常 $n^{-1}\partial^2 \mathcal{L}_n(\beta)/(\partial \beta \partial \beta^T)$ 与渐近方差 Σ 密切相关. 因此,

$$\begin{aligned} n^{-1}\text{BIC}(\mathbf{w}) &= -2n^{-1}\mathcal{L}\{\hat{\beta}(\mathbf{w})\} + n^{-1}\log(n)\mathbf{w}^T\mathbf{v} \\ &\approx \{\hat{\beta}(\mathbf{w}) - \tilde{\beta}\}^T \hat{\Sigma}^{-1} \{\hat{\beta}(\mathbf{w}) - \tilde{\beta}\} + \log(n) \frac{\mathbf{w}^T\mathbf{v}}{n} + \text{常数} \\ &= \mathcal{S}(\mathbf{w}) |_{\phi_n=\log n} + \text{常数}, \end{aligned} \quad (2.15)$$

即当 $\phi_n = \log n$ 时, 近似地说, 准则 $\mathcal{S}(\mathbf{w})$ 是模型平均的 BIC.

3 数值研究

3.1 模拟研究

本节通过模拟来比较我们的模型平均方法 (model averaging by least squares approximation, 记为 MA_{LSA}) 与 LSA 方法, 其中的关键参数采用 AIC (Akaike information criterion) 或者 BIC 进行选择 (分别记为 LSA-AIC 和 LSA-BIC). 文献 [19] 的数值结果已经显示 LSA-BIC 方法有很好的表现.

例 1 (逻辑回归) 在这个例子中我们使用文献 [19, 33] 中的数据产生方式. 具体地,

$$\Pr(y_i | \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \beta_0)}{1 + \exp(\mathbf{x}_i^T \beta_0)},$$

其中 $\beta_0 = (3, 0, 0, 1.5, 0, 0, 2, 0, 0)^T$, $\mathbf{x}_i = (x_{i1}, \dots, x_{id})^T$, 它的分量是零均值方差为 1 的相互独立的正态变量, 各个分量之间的相关系数为 $0.5^{|j_1-j_2|}$ (j_1 和 j_2 表示 1 到 d 的整数). 样本量 n 属于集合 $\{100, 200, 400, 800\}$. 基于 1,000 次重复, 我们计算参数估计的均方误差 (mean squared error, MSE): $E\{(\hat{\beta} - \beta_0)^T(\hat{\beta} - \beta_0)\}$.

另外由定理 2 所显示的权重的收敛性, 我们知道模型平均估计 $\hat{\beta}(\mathbf{w})$ 的许多分量应该接近 0. 因此, 我们人为地用一个非常小的数值 10^{-5} 去截断这些权重, 从而取得稀疏性. 然后用平均的模型大小 (mean model size, MMS) 来评价各种方法在变量使用上的节俭程度: $E\{\sum_{j=1}^p I(|\hat{\beta}_j| > 10^{-5})\}$, 其中 $I(\cdot)$ 是示性函数.

模拟结果显示在表 1 中. 不论何种样本量, MA_{LSA} 方法的 MSE 都比 LSA-BIC 的小. 当 $n \geq 200$ 时, MA_{LSA} 的 MSE 比 LSA-AIC 小; 当 $n = 100$ 时, LSA-AIC 的 MSE 更小. 但是, MA_{LSA} 的 MMS 总是比 LSA-AIC 的小, 这意味着 MA_{LSA} 使用了更少的变量却取得了与 LSA-AIC 相当的 MSE.

例 2 (最小绝对距离回归) 在这个例子中, 我们使用文献 [19, 34] 中的数据产生过程. 具体地, $y_i = \mathbf{x}_i^T \beta_0 + \epsilon_i$, 其中 $\beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\mathbf{x}_i$ 的产生机制与例 1 中一样, ϵ_i 来自一个混合分布, 其中 10% 来自标准的 Cauchy 分布, 90% 来自标准的正态分布. 样本量属于集合 $\{50, 100, 200, 400\}$, 这个集合与例 1 不同, 因为对于更大的样本量, MA_{LSA} 与 LSA-BIC 表现得很接近.

模拟结果显示在表 2 中. 不同于之前的例子, 在这个例子中, 各个样本量下 MA_{LSA} 均产生了比 LSA-AIC 方法更小的 MSE. 当 $n \leq 200$ 时, MA_{LSA} 的 MSE 比 LSA-BIC 更小, 而当 $n = 400$ 时它们的 MSE 接近. 在三种样本下, MA_{LSA} 和 LSA-BIC 的 MMS 比较接近, 它们均小于 LSA-AIC 的 MMS.

表 1 例 1 的模拟结果. SE 表示标准差

	估计方法	MSE	(SE)	MMS	(SE)
$n = 100$	LSA-AIC	1.422	(0.051)	3.302	(0.025)
	LSA-BIC	2.247	(0.106)	2.908	(0.025)
	MA _{LSA}	1.474	(0.056)	3.020	(0.018)
$n = 200$	LSA-AIC	0.888	(0.046)	3.612	(0.035)
	LSA-BIC	0.932	(0.043)	3.152	(0.018)
	MA _{LSA}	0.861	(0.052)	3.058	(0.012)
$n = 400$	LSA-AIC	0.415	(0.019)	3.470	(0.032)
	LSA-BIC	0.366	(0.016)	3.082	(0.012)
	MA _{LSA}	0.352	(0.021)	3.024	(0.007)
$n = 800$	LSA-AIC	0.187	(0.007)	3.528	(0.033)
	LSA-BIC	0.161	(0.007)	3.058	(0.011)
	MA _{LSA}	0.144	(0.007)	3.008	(0.004)

表 2 例 2 的模拟结果. SE 表示标准差

	估计方法	MSE	(SE)	MMS	(SE)
$n = 50$	LSA-AIC	0.347	(0.014)	3.756	(0.036)
	LSA-BIC	0.261	(0.012)	3.352	(0.027)
	MA _{LSA}	0.209	(0.011)	3.206	(0.022)
$n = 100$	LSA-AIC	0.129	(0.005)	3.534	(0.033)
	LSA-BIC	0.086	(0.004)	3.144	(0.019)
	MA _{LSA}	0.083	(0.003)	3.126	(0.017)
$n = 200$	LSA-AIC	0.062	(0.003)	3.464	(0.030)
	LSA-BIC	0.039	(0.002)	3.074	(0.012)
	MA _{LSA}	0.038	(0.002)	3.048	(0.010)
$n = 400$	LSA-AIC	0.029	(0.001)	3.390	(0.028)
	LSA-BIC	0.018	(0.001)	3.036	(0.008)
	MA _{LSA}	0.019	(0.001)	3.018	(0.006)

3.2 实际数据研究

本小节中所使用的实际数据来源于文献 [35], 它是 2005 年 921 家美国公司的标普信用评级. 被解释变量是 0-1 变量, 如果评级在 BB+ 以上取 1 (即为投资级别), 否则取 0. 解释变量包括截距项、公司的营运资金 (它代表公司的短期流动性)、留存收益 (它代表历史盈利能力)、息税前利润 (它代表当前盈利能力)、账面杠杆 (即公司债务与资产的比率) 和销售量 (它代表公司的规模). 前三个非截距项的解释变量均除以了总资产.

首先利用所有的 6 个变量来拟合数据得到参数估计

$$(-0.430, -7.203, -0.945, 1.235, 4.142, -4.281)^T.$$

为了模仿稀疏性, 我们把后三个系数设置为 0, 即 $\beta_0 = (-0.430, -7.203, -0.945, 0, 0, 0)^T$. 然后我们对

表 3 实际数据分析结果. SE 表示标准差

估计方法	MSE	(SE)	MMS	(SE)
LSA-AIC	1.733	(0.065)	2.886	(0.043)
LSA-BIC	1.670	(0.054)	2.084	(0.014)
MA _{LSA}	1.571	(0.062)	1.788	(0.042)

921 个样本进行重抽样, 并依据概率 $\exp(\mathbf{x}_i^T \boldsymbol{\beta}_0) / \{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}_0)\}$ 生成 0-1 变量. 利用以上过程产生 $n = 921$ 个新的数据. 我们重复以上过程 500 次来计算参数估计的 MSE 和 MMS. 表 3 显示了估计结果. 从中可以看出 MA_{LSA} 的 MSE 和 MMS 均是最小的.

4 维数发散情形下的推广

本节把权重的收敛性推广到维数发散的情形, 即我们允许, 当 $n \rightarrow \infty$ 时, 未知参数的维度 $p = p_n \rightarrow \infty$. 在这种情形下, 假设 $\tilde{\boldsymbol{\beta}}$ 有如下的渐近正态性:

$$\sqrt{n} \boldsymbol{\alpha}^T (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \text{Normal}(0, \boldsymbol{\alpha}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}), \quad (4.1)$$

其中 $\boldsymbol{\alpha}$ 是 p 维的范数为 1 的向量. 这个渐近正态性是一个标准的结果, 参见文献 [36, 37]. 针对这种情形, 在步骤 2 中我们使用拓展的 BIC^[38] 来选择集合 $\hat{\mathcal{A}}$. 通过使用定理 1 和 2 中类似的证明技术可以得到如下的权重收敛性.

定理 4 如果 $\phi_n = o(n)$, 那么对任意 $s \in \mathcal{U}$, 有

$$\hat{w}_s = O_p\left(\frac{p_n}{n} + \frac{\phi_n}{n}\right). \quad (4.2)$$

如果 $\phi_n \rightarrow \infty$, $\phi_n = O(n^{1/2})$ 且 $p_n = o(\phi_n)$, 那么对任意 $s \in \mathcal{O}$, 有

$$\hat{w}_s = O_p\left(\frac{\phi_n^{1/2}}{n^{1/4}} + \frac{p_n^{1/2}}{\phi_n^{1/2}}\right). \quad (4.3)$$

通过比较 (4.2) 与 (2.12) 我们发现, 当 $p_n \rightarrow \infty$ 时, 欠拟合模型的权重的收敛速度依赖于 p_n . 通过比较 (4.3) 与 (2.13) 我们发现, 当 $p_n \rightarrow \infty$ 时, 过拟合模型的权重的收敛结果不是令人很满意的, 因为条件 $p_n = o(\phi_n)$ 限制了 p_n 的发散速率. 定理 4 的证明在附录 A.3 中.

5 结论

本文通过使用最小二乘近似 (LSA), 提出了一种全新的模型平均估计方法. 该方法可用于线性模型、广义线性模型和分位数回归等各种常用模型. 通过使用惩罚 LSA 的解路径及二次规划使得该方法在计算上也很方便. 我们还证明了最终的模型平均估计的渐近正态性质. 数值结果显示了该方法相对以往方法具有优势.

需要注意的是, 尽管我们把该方法拓展到发散的 p_n 的情形, 但是我们仍然需要 $p_n = o(n^{1/2})$. 因此维度不能太高, 如何将我们的方法发展到高维甚至超高维值得将来去研究. 另外, 如何提出一种数据驱动的 ϕ_n 的选择方法也是未来的研究方向.

致谢 感谢编委和审稿人提出的宝贵意见.

参考文献

- 1 Leung G, Barron A R. Information theory and mixing least-squares regressions. *IEEE Trans Inform Theory*, 2006, 52: 3396–3410
- 2 Bates J M, Granger C W J. The combination of forecasts. *J Oper Res Soc*, 1969, 20: 451–468
- 3 Yuan Z, Yang Y H. Combining linear regression models: When and how? *J Amer Statist Assoc*, 2005, 100: 1202–1214
- 4 Zhang X Y, Wan A T K, Zhou S Z. Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. *J Bus Econom Statist*, 2012, 30: 132–142
- 5 Hoeting J A, Madigan D, Raftery A E, et al. Bayesian model averaging: A tutorial. *Statist Sci*, 1999, 14: 382–417
- 6 Hansen B E. Least squares model averaging. *Econometrica*, 2007, 75: 1175–1189
- 7 Wan A T K, Zhang X Y, Zou G. Least squares model averaging by Mallows criterion. *J Econometrics*, 2010, 156: 277–283
- 8 Liu Q, Okui R. Heteroskedasticity-robust C_p model averaging. *Econom J*, 2013, 16: 463–472
- 9 Ando T, Li K C. A model-averaging approach for high-dimensional regression. *J Amer Statist Assoc*, 2014, 109: 254–265
- 10 Zhang X Y, Zou G, Carroll R. Model averaging based on Kullback-Leibler distance. *Statist Sinica*, 2015, 25: 1583–1598
- 11 Hansen B E. Averaging estimators for regressions with a possible structural break. *Econometric Theory*, 2009, 25: 1498–1514
- 12 Zhang X Y. Model averaging with covariates that are missing completely at random. *Econom Lett*, 2013, 121: 360–363
- 13 Fang F, Lan W, Tong J, et al. Model averaging for prediction with fragmentary data. *J Bus Econom Statist*, 2019, 37: 517–527
- 14 Zhang X Y, Zou G, Liang H. Model averaging and weight choice in linear mixed-effects models. *Biometrika*, 2014, 101: 205–218
- 15 Cheng X, Hansen B E. Forecasting with factor-augmented regression: A frequentist model averaging approach. *J Econometrics*, 2015, 186: 280–293
- 16 Lu X, Su L. Jackknife model averaging for quantile regressions. *J Econometrics*, 2015, 188: 40–58
- 17 Zhang X Y, Yu D, Zou G, et al. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *J Amer Statist Assoc*, 2016, 111: 1775–1790
- 18 Zhang X Y, Wang W. Optimal model averaging estimation for partially linear models. *Statist Sinica*, 2019, 29: 693–718
- 19 Wang H, Leng C. Unified LASSO estimation by least squares approximation. *J Amer Statist Assoc*, 2007, 102: 1039–1048
- 20 Behl P, Claeskens G, Dette H. Focused model selection in quantile regression. *Statist Sinica*, 2014, 24: 601–624
- 21 Charkhi A, Claeskens G, Hansen B E. Minimum mean squared error model averaging in likelihood models. *Statist Sinica*, 2016, 26: 809–840
- 22 Hansen B E. Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*, 2014, 5: 495–530
- 23 Hjort N L, Claeskens G. Frequentist model average estimators. *J Amer Statist Assoc*, 2003, 98: 879–899
- 24 Hjort N L, Claeskens G. Focused information criteria and model averaging for the Cox hazard regression model. *J Amer Statist Assoc*, 2016, 101: 1449–1464
- 25 Liu C A. Distribution theory of the least squares averaging estimator. *J Econometrics*, 2015, 186: 142–159
- 26 Xu G, Wang S, Huang J Z. Focused information criterion and model averaging based on weighted composite quantile regression. *Scand J Statist*, 2014, 41: 365–381
- 27 Zhang X Y, Liang H. Focused information criterion and model averaging for generalized additive partial linear models. *Ann Statist*, 2011, 39: 174–200
- 28 Jullum M, Hjort N L. Parametric or nonparametric: The FIC approach. *Statist Sinica*, 2017, 27: 951–981
- 29 Zhang X Y, Liu C A. Inference after model averaging in linear regression models. *Econometric Theory*, 2019, 35: 816–841
- 30 Mitra P, Lian H, Mitra R, et al. A general framework for frequentist model averaging. *Sci China Math*, 2019, 62: 205–226
- 31 Zou H. The adaptive lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429
- 32 Efron B, Hastie T, Johnstone I, et al. Least angle regression. *Ann Statist*, 2004, 32: 407–499
- 33 Hunter D R, Li R. Variable selection using MM algorithms. *Ann Statist*, 2005, 33: 1617–1642
- 34 Wang H, Li R, Tsai C L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*,

- 2007, 94: 553–568
- 35 Ashbaugh-Skaife H, Collins D W, LaFond R. The effects of corporate governance on firms' credit ratings. *J Account Econ*, 2006, 42: 203–243
- 36 Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann Statist*, 2004, 32: 928–961
- 37 Zou H, Zhang H H. On the adaptive elastic-net with a diverging number of parameters. *Ann Statist*, 2009, 37: 1733–1751
- 38 Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 2008, 95: 759–771
- 39 Rao C R. *Linear Statistical Inference and Its Applications*, 2nd ed. New York: John Wiley & Sons, 1973
- 40 Leng C, Li B. Least squares approximation with a diverging number of parameters. *Statist Probab Lett*, 2010, 80: 254–261

附录 A 定理的证明

附录 A.1 定理 1 的证明

首先, 由文献 [19] 可得

$$\Pr\{\hat{\mathcal{A}} = \mathcal{A}\} \rightarrow 1. \quad (\text{A.1})$$

因此, 在下面的证明中假设 $\hat{\mathcal{A}} = \mathcal{A}$, 这不影响我们的证明, 因为我们所证明的结论是依概率成立.

令 $\mathbf{C}$ 为 $S \times S$ 矩阵, 其第 s 行第 m 列元素为

$$c_{s,m} = (\hat{\beta}_s - \tilde{\beta})^T \hat{\Sigma}^{-1} (\hat{\beta}_m - \tilde{\beta}) + \frac{\phi_n(v_s + v_m)}{(2n)}, \quad (\text{A.2})$$

那么由 $\sum_{s=1}^S w_s = 1$ 可得

$$\mathbf{C}_n(\mathbf{w}) = \mathbf{w}^T \mathbf{C} \mathbf{w}. \quad (\text{A.3})$$

明显地, $c_{s,m} = c_{m,s}$. 不失一般性, 假设前 $\tilde{s} - 1$ 个模型是欠拟合模型, 第 $\tilde{s}$ 个模型是真实模型, 其他模型是过拟合模型.

由 (2.1)、(2.3) 和假设 $\hat{\Sigma}$ 收敛到一个正定矩阵可得, 对任意 $s \in \mathcal{O} \cup \{\tilde{s}\}$, 有

$$\begin{aligned} \hat{\Sigma}^{-1/2}(\hat{\beta}_s - \tilde{\beta}) &= \hat{\Sigma}^{-1/2}(\hat{\beta}_s - \beta) - \hat{\Sigma}^{-1/2}(\tilde{\beta} - \beta) \\ &= O_p(n^{-1/2}), \end{aligned} \quad (\text{A.4})$$

对任意 $s \in \mathcal{U}$, 有

$$\begin{aligned} \hat{\Sigma}^{-1/2}(\hat{\beta}_s - \tilde{\beta}) &= \hat{\Sigma}^{-1/2}\{\Pi_s^T (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T \mathbf{z} - \tilde{\beta}\} \\ &= \hat{\Sigma}^{-1/2}\{\Pi_s^T (\Pi_s \hat{\Sigma}^{-1} \Pi_s^T)^{-1} \Pi_s \hat{\Sigma}^{-1} \tilde{\beta} - \tilde{\beta}\} \\ &= O_p(1). \end{aligned} \quad (\text{A.5})$$

由 (A.2)–(A.5) 可得, 对任意 $s, m \in \mathcal{O} \cup \{\tilde{s}\}$, 有

$$c_{s,m} = O_p(n^{-1}) + \frac{\phi_n(v_s + v_m)}{(2n)}; \quad (\text{A.6})$$

对任意 $s \in \mathcal{U}$ 和 $m \in \mathcal{O} \cup \{\tilde{s}\}$, 有

$$c_{s,m} = O_p(n^{-1/2}) + \frac{\phi_n(v_s + v_m)}{(2n)}. \quad (\text{A.7})$$

令 $\hat{B}_{s^c}$ 为一个矩阵, 它包含在 $\hat{\Sigma}^{-1/2}$ 中但不在 $\hat{B}_s$ 中的列. 令 Π_{s^c} 为一个投影矩阵使得 $\hat{B}_{s^c} = \hat{\Sigma}^{-1/2} \Pi_{s^c}$. 由文献 [39] 可知, $\hat{B}_{s^c}^T \{I_p - \hat{B}_s (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T\} \hat{B}_{s^c}$ 依概率收敛到一个正定矩阵. 因此, 对任意 $s \in \mathcal{U}$, 由 (2.1) 可知,

$$\begin{aligned} z^T \{I_p - \hat{B}_s (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T\} z &= \tilde{\beta}^T \hat{\Sigma}^{-1/2} \{I_p - \hat{B}_s (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T\} \hat{\Sigma}^{-1/2} \tilde{\beta} \\ &= \tilde{\beta}^T \Pi_{s^c}^T \hat{B}_{s^c}^T \{I_p - \hat{B}_s (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T\} \hat{B}_{s^c} \Pi_{s^c} \tilde{\beta} \\ &\xrightarrow{p} \delta_s > 0, \end{aligned} \quad (\text{A.8})$$

其中 δ_s 是一个正常数. 由 (2.3)、(A.2) 和 (A.8) 可知, 对任意 $s \in \mathcal{U}$, 有

$$\begin{aligned} c_{s,s} &= (\hat{\beta}_s - \tilde{\beta})^T \hat{\Sigma}^{-1} (\hat{\beta}_s - \tilde{\beta}) + \phi_n \frac{v_s}{n} \\ &= \{\hat{\Sigma}^{-1/2} \Pi_s^T (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T z - z\}^T \{\hat{\Sigma}^{-1/2} \Pi_s^T (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T z - z\} + \phi_n \frac{v_s}{n} \\ &= z^T \{I_p - \hat{B}_s (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T\} z + \phi_n \frac{v_s}{n} \\ &= \delta_s + o_p(1) + \phi_n \frac{v_s}{n}. \end{aligned} \quad (\text{A.9})$$

对任意 $s, m \in \mathcal{U}$, 当 $s > m$ 时, 由步骤 3 可知 $\hat{B}_m$ 的列包含在 $\hat{B}_s$ 中, 因此, 由 (2.3)、(A.2) 和 (A.8) 可得

$$\begin{aligned} c_{s,m} &= (\hat{\beta}_s - \tilde{\beta})^T \hat{\Sigma}^{-1} (\hat{\beta}_m - \tilde{\beta}) + \frac{\phi_n(v_s + v_m)}{(2n)} \\ &= \{\hat{\Sigma}^{-1/2} \Pi_s^T (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T z - z\}^T \{\hat{\Sigma}^{-1/2} \Pi_m^T (\hat{B}_m^T \hat{B}_m)^{-1} \hat{B}_m^T z - z\} + \frac{\phi_n(v_s + v_m)}{(2n)} \\ &= \{\hat{B}_s (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T z - z\}^T \{\hat{B}_m (\hat{B}_m^T \hat{B}_m)^{-1} \hat{B}_m^T z - z\} + \frac{\phi_n(v_s + v_m)}{(2n)} \\ &= z^T \{I_p - \hat{B}_s (\hat{B}_s^T \hat{B}_s)^{-1} \hat{B}_s^T\} z + \frac{\phi_n(v_s + v_m)}{(2n)} \\ &= \delta_s + o_p(1) + \frac{\phi_n(v_s + v_m)}{(2n)}. \end{aligned} \quad (\text{A.10})$$

对 $k \in \mathcal{U}$, 定义一个新的权重向量 $w^* = (\hat{w}_1, \dots, \hat{w}_{k-1}, 0, \hat{w}_{k+1}, \dots, \hat{w}_{\tilde{s}-1}, \hat{w}_{\tilde{s}} + \hat{w}_k, \hat{w}_{\tilde{s}+1}, \dots, \hat{w}_S)^T$. 由 (A.3)、(A.6)、(A.9) 和 (A.11) 可得

$$\begin{aligned} \mathcal{C}_n(w^*) - \mathcal{C}_n(\hat{w}) &= \hat{w}_k^2 (c_{\tilde{s}, \tilde{s}} + c_{k,k} - 2c_{\tilde{s},k}) + 2\hat{w}_k \sum_{s=1}^S \{\hat{w}_s (c_{\tilde{s},s} - c_{k,s})\} \\ &= 2\hat{w}_k \sum_{s=1}^{k-1} \{\hat{w}_s (c_{\tilde{s},s} - c_{k,s})\} + 2\hat{w}_k \sum_{s=k+1}^{\tilde{s}-1} \{\hat{w}_s (c_{\tilde{s},s} - c_{k,s})\} \\ &\quad + 2\hat{w}_k \sum_{s=\tilde{s}}^S \{\hat{w}_s (c_{\tilde{s},s} - c_{k,s})\} + \hat{w}_k^2 (c_{\tilde{s}, \tilde{s}} - c_{k,k}) \end{aligned}$$

$$\begin{aligned}
 &= 2\hat{w}_k \sum_{s=1}^{k-1} \{o_p(1) - \delta_k\} + 2\hat{w}_k \sum_{s=k+1}^{\tilde{s}-1} \{o_p(1) - \delta_s\} - 2\hat{w}_k \sum_{s=\tilde{s}}^S \hat{w}_s c_{k,s} \\
 &\quad + 2\hat{w}_k \sum_{s=\tilde{s}}^S \hat{w}_s c_{\tilde{s},s} - \hat{w}_k^2 \{\delta_k + o_p(1)\} \\
 &= 2\hat{w}_k \sum_{s=1}^{k-1} \{o_p(1) - \delta_k\} + 2\hat{w}_k \sum_{s=k+1}^{\tilde{s}-1} \{o_p(1) - \delta_s\} - 2\hat{w}_k o_p(1) \\
 &\quad + 2\hat{w}_k \sum_{s=\tilde{s}}^S \hat{w}_s c_{\tilde{s},s} - \hat{w}_k^2 \{\delta_k + o_p(1)\} \\
 &\equiv \Delta + 2\hat{w}_k \sum_{s=\tilde{s}}^S \hat{w}_s c_{\tilde{s},s} - \hat{w}_k^2 \{\delta_k + o_p(1)\} \\
 &= \Delta + 2\hat{w}_k O_p\left(\frac{\phi_n}{n}\right) - \hat{w}_k^2 \{\delta_k + o_p(1)\}, \tag{A.11}
 \end{aligned}$$

其中

$$\Pr\{\Delta \leq 0\} \rightarrow 1. \tag{A.12}$$

进一步, 由 (2.6) 可知, $\mathcal{C}_n(\mathbf{w}^*) - \mathcal{C}_n(\hat{\mathbf{w}}) \geq 0$, 由此及 (A.11) 和 (A.12) 可得

$$\Pr\left\{2\hat{w}_k O_p\left(\frac{\phi_n}{n}\right) - \hat{w}_k^2 \{\delta_k + o_p(1)\} \geq 0\right\} \rightarrow 1,$$

因此, (2.12) 成立.

附录 A.2 定理 2 的证明

令 $\delta = \sum_{s=\tilde{s}+1}^S \tilde{w}_s$, $\tilde{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_{\tilde{s}-1}, \hat{w}_{\tilde{s}} + \delta, 0, \dots, 0)^T$. 划分 $\hat{\mathbf{w}}$ 使得 $\hat{\mathbf{w}} = (\hat{\mathbf{w}}_{(1)}^T, \hat{w}_{\tilde{s}}, \hat{\mathbf{w}}_{(3)}^T)^T$, 据此我们也划分 $\mathbf{C}$ 使得

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \mathbf{C}_{13} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \mathbf{C}_{23} \\ \mathbf{C}_{31} & \mathbf{C}_{32} & \mathbf{C}_{33} \end{pmatrix}, \tag{A.13}$$

其中 $\mathbf{C}_{22} = c_{\tilde{s}, \tilde{s}}$. 令 $\mathbf{t} = (\mathbf{0}_{(\tilde{s}-1) \times 1}^T, -\delta, \hat{\mathbf{w}}_{(3)}^T)^T$, 那么 $\tilde{\mathbf{w}} = \hat{\mathbf{w}} - \mathbf{t}$,

$$\begin{aligned}
 \mathcal{C}_n(\tilde{\mathbf{w}}) - \mathcal{C}_n(\hat{\mathbf{w}}) &= \tilde{\mathbf{w}}^T \mathbf{C} \tilde{\mathbf{w}} - \mathcal{C}_n(\hat{\mathbf{w}}) = (\hat{\mathbf{w}} - \mathbf{t})^T \mathbf{C} (\hat{\mathbf{w}} - \mathbf{t}) - \mathcal{C}_n(\hat{\mathbf{w}}) = (\mathbf{t} - 2\hat{\mathbf{w}})^T \mathbf{C} \mathbf{t} \\
 &= \{(\mathbf{0}_{(\tilde{s}-1) \times 1}^T, -\delta, \hat{\mathbf{w}}_{(3)}^T)^T - 2\hat{\mathbf{w}}\}^T \mathbf{C} (\mathbf{0}_{(\tilde{s}-1) \times 1}^T, -\delta, \hat{\mathbf{w}}_{(3)}^T)^T \\
 &= (-2\hat{\mathbf{w}}_{(1)}^T, -\delta - 2\hat{w}_{\tilde{s}}, -\hat{\mathbf{w}}_{(3)}^T) \mathbf{C} (\mathbf{0}_{(\tilde{s}-1) \times 1}^T, -\delta, \hat{\mathbf{w}}_{(3)}^T)^T \\
 &= 2\hat{\mathbf{w}}_{(1)}^T \mathbf{C}_{12} \delta - 2\hat{\mathbf{w}}_{(1)}^T \mathbf{C}_{13} \hat{\mathbf{w}}_{(3)} + (\delta^2 + 2\hat{w}_{\tilde{s}} \delta) \mathbf{C}_{22} \\
 &\quad - (\delta + 2\hat{w}_{\tilde{s}}) \mathbf{C}_{23} \hat{\mathbf{w}}_{(3)} + \hat{\mathbf{w}}_{(3)}^T \mathbf{C}_{32} \delta - \hat{\mathbf{w}}_{(3)}^T \mathbf{C}_{33} \hat{\mathbf{w}}_{(3)} \\
 &= -2(\hat{\mathbf{w}}_{(1)}^T \mathbf{C}_{13} \hat{\mathbf{w}}_{(3)} - \hat{\mathbf{w}}_{(1)}^T \mathbf{C}_{12} \delta) + 2\hat{w}_{\tilde{s}} (\delta \mathbf{C}_{22} - \mathbf{C}_{23} \hat{\mathbf{w}}_{(3)})
 \end{aligned}$$

$$-(\hat{\mathbf{w}}_{(3)}^T \mathbf{C}_{33} \hat{\mathbf{w}}_{(3)} - \delta^2 \mathbf{C}_{22}). \quad (\text{A.14})$$

由 (2.12)、(A.7) 和 $\phi_n = o(n)$ 可得

$$\hat{\mathbf{w}}_{(1)}^T \mathbf{C}_{13} \hat{\mathbf{w}}_{(3)} - \hat{\mathbf{w}}_{(1)}^T \mathbf{C}_{12} \delta = O_p\left(\frac{\phi_n^2}{n^2} + \frac{\phi_n}{n^{3/2}}\right). \quad (\text{A.15})$$

由 (A.2)、(A.6) 和 $\phi_n \rightarrow \infty$ 可得

$$\begin{aligned} & \Pr\{\delta \mathbf{C}_{22} - \mathbf{C}_{23} \hat{\mathbf{w}}_{(3)} \leq 0\} \\ &= \Pr\left\{\sum_{s=\tilde{s}+1}^S \hat{w}_s (c_{\tilde{s},s} - c_{\tilde{s},s}) \leq 0\right\} \\ &\rightarrow 1. \end{aligned} \quad (\text{A.16})$$

由 (A.2) 和 (A.6) 可得

$$\begin{aligned} \hat{\mathbf{w}}_{(3)}^T \mathbf{C}_{33} \hat{\mathbf{w}}_{(3)} - \delta^2 \mathbf{C}_{22} &= \sum_{s=\tilde{s}+1}^S \sum_{k=\tilde{s}+1}^S \hat{w}_s \hat{w}_k (c_{s,k} - c_{\tilde{s},s}) \\ &\geq \sum_{s=\tilde{s}+1}^S \sum_{k=\tilde{s}+1}^S \hat{w}_s \hat{w}_k \left(O_p\left(\frac{1}{n}\right) + \frac{\phi_n}{n}\right) \\ &= \delta^2 \frac{\phi_n}{n} + O_p\left(\frac{1}{n}\right). \end{aligned} \quad (\text{A.17})$$

进一步, 由 (2.6) 可知, $\mathcal{C}_n(\mathbf{w}^*) - \mathcal{C}_n(\hat{\mathbf{w}}) \geq 0$, 由此和 (A.14)–(A.17) 可得

$$\Pr\left\{O_p\left(\frac{\phi_n^2}{n^2} + \frac{\phi_n}{n^{3/2}}\right) - \delta^2 \frac{\phi_n}{n} - O_p\left(\frac{1}{n}\right) \geq 0\right\} \rightarrow 1. \quad (\text{A.18})$$

另外, 由 $\phi_n = O(n^{1/2})$ 可得

$$\frac{\phi_n^2}{n^2} + \frac{\phi_n}{n^{3/2}} = O(n^{-1}). \quad (\text{A.19})$$

组合 (A.18)–(A.19) 可得 (2.13).

附录 A.3 定理 4 的证明

首先证明欠拟合模型的权重的收敛性. 由文献 [40], 通过使用拓展的 BIC, 我们仍然有

$$\Pr\{\hat{\mathcal{A}} = \mathcal{A}\} \rightarrow 1. \quad (\text{A.20})$$

利用 (A.6) 和 (A.7) 的推导过程可知, 在一定的正则条件下, 对任意 $s, m \in \mathcal{O} \cup \{\tilde{s}\}$, 有

$$c_{s,m} = O_p(p_n n^{-1}) + \frac{\phi_n(v_s + v_m)}{(2n)}, \quad (\text{A.21})$$

对任意 $s \in \mathcal{U}$ 和 $m \in \mathcal{O} \cup \{\tilde{s}\}$, 有

$$c_{s,m} = O_p(p_n n^{-1/2}) + \frac{\phi_n(v_s + v_m)}{(2n)}. \quad (\text{A.22})$$

那么, 类似于 (A.11) 可得

$$\mathcal{C}_n(\mathbf{w}^*) - \mathcal{C}_n(\hat{\mathbf{w}}) = \Delta + 2\hat{w}_k O_p\left(\frac{\phi_n}{n} + \frac{p_n}{n}\right) - \hat{w}_k^2\{\delta_k + o_p(1)\}. \quad (\text{A.23})$$

因此, 通过使用 $\phi_n = o(n)$ 和定理 1 的证明步骤, 可得 (4.2).

接下来证明过拟合模型的权重的收敛性. 利用 (A.15) 的推导过程和 (4.2) 可知,

$$\hat{\mathbf{w}}_{(1)}^T \mathbf{C}_{13} \hat{\mathbf{w}}_{(3)} - \hat{\mathbf{w}}_{(1)}^T \mathbf{C}_{12} \delta = O_p\left(\phi_n \frac{p_n}{n^{3/2}} + \frac{\phi_n^2}{n^{3/2}} + \phi_n^2 \frac{p_n}{n^2} + \phi_n \frac{p_n^2}{n^2}\right). \quad (\text{A.24})$$

当 $p_n = o(\phi_n)$ 时, 利用 (A.16) 和 (A.17) 的推导过程可得

$$\Pr\{\delta \mathbf{C}_{22} - \mathbf{C}_{23} \hat{\mathbf{w}}_{(3)} \leq 0\} \rightarrow 1, \quad (\text{A.25})$$

$$\hat{\mathbf{w}}_{(3)}^T \mathbf{C}_{33} \hat{\mathbf{w}}_{(3)} - \delta^2 \mathbf{C}_{22} \geq \sum_{s=\tilde{s}+1}^S \sum_{k=\tilde{s}+1}^S \hat{w}_s \hat{w}_k \left(O_p\left(\frac{p_n}{n}\right) + \frac{\phi_n}{n} \right) = \delta^2 \frac{\phi_n}{n} + O_p\left(\frac{p_n}{n}\right). \quad (\text{A.26})$$

那么,

$$\Pr\left\{ O_p\left(\phi_n \frac{p_n}{n^{3/2}} + \frac{\phi_n^2}{n^{3/2}} + \phi_n^2 \frac{p_n}{n^2} + \phi_n \frac{p_n^2}{n^2}\right) - \delta^2 \frac{\phi_n}{n} - O_p\left(\frac{p_n}{n}\right) \geq 0 \right\} \rightarrow 1, \quad (\text{A.27})$$

另外, 由 $\phi_n \rightarrow \infty$, $\phi_n = O(n^{1/2})$ 和 $p_n = o(\phi_n)$ 可得

$$\phi_n \frac{p_n}{n^{3/2}} + \frac{\phi_n^2}{n^{3/2}} + \phi_n^2 \frac{p_n}{n^2} + \phi_n \frac{p_n^2}{n^2} = O\left(\frac{\phi_n^2}{n^{3/2}}\right). \quad (\text{A.28})$$

由 (A.14)–(A.15) 可得 (4.3).

Model averaging by least squares approximation

Xinyu Zhang

Abstract In this paper, we propose a model averaging method by adopting least squares approximation; as a result, the method can be used in linear regression models, generalized linear models, quantile regression, and so on. In particular, the classic Mallows model averaging is a special case of the proposed model averaging method. Unlike the existing literature on the distribution of model averaging where a local misspecification assumption is utilized and the resulting limiting distribution is complicated, we prove the asymptotic normality of the proposed estimator in the current paper without using the local misspecification assumption. In addition, the proposed method is extended to a high-dimensional situation. Numerical results show the promise of the proposed method.

Keywords asymptotic normality, consistency, least squares approximation, model averaging, weight choice

MSC(2010) 62F10, 62F99, 62H12

doi: 10.1360/N012019-00137