中国科学: 数学 2018 年 第 48 卷 第 8 期: 1019～1052
SCIENTIA SINICA Mathematica

《中国科学》杂志社
SCIENCE CHINA PRESS

论 文

半参数模型平均估计的渐近理论

朱容[1,2], 邹国华[3*]

1. 中国科学院大学数学科学学院, 北京 100049;
2. 中国科学院数学与系统科学研究院, 北京 100190;
3. 首都师范大学数学科学学院, 北京 100048

E-mail: zhu_rong@amss.ac.cn, ghzou@amss.ac.cn

收稿日期: 2017-01-15; 接受日期: 2017-09-26; 网络出版日期: 2018-04-19; * 通信作者
国家自然科学基金 (批准号: 11331011 和 11529101) 和科技部国家重点研发计划 (批准号: 2016YFB0502301) 资助项目

摘要 本文主要研究半参数模型下的模型平均估计问题, 旨在将 Hjort 和 Claeskens 在 2003 年的工作从参数模型扩展到半参数模型. 尽管 Claeskens 和 Carroll 于 2007 年在完全一样的模型下考虑了相同的问题, 但二者的方法不相同. 本文推导了模型平均估计的渐近分布, 并构造了一个覆盖真实参数的概率趋于给定水平的置信区间. 模拟研究和实际数据分析均表明本文的方法是有效的.

关键词 渐近分布 广义轮廓似然估计 模型平均 模型选择 半参数模型

MSC (2010) 主题分类 62E20, 62F12

1 引言

模型选择一直是统计学的热门研究课题. 在对数据进行统计分析的初始阶段, 为了不遗失重要的变量, 一般会尽可能多地包含潜在的相关变量, 但是其中有些变量其实并不是显著的. 为了获得真实的或近似真实的模型, 更准确地描述变量间的相关关系, 提高预测的精度, 有必要把不显著的变量排除在模型之外, 从而得到一个最佳的模型. 近几十年来, 统计学家已提出了许多的模型选择方法, 如经典的逐步回归、基于信息准则的 AIC (Akaike information criterion) [1]、基于 Bayes 思想的 BIC (Bayesian information criterion) [2]、基于预测误差的 Mallows' C_p [3] 和基于删一思想的 CV (cross validation) [4] 等. 模型选择的目标是为了找到真实或最佳的模型, 然后把这个模型当作真实的数据产生过程, 其后的统计推断完全基于这个模型. 这个过程忽略了模型选择阶段中的不确定性, 从而低估实际的变异, 导致过于乐观的置信区间; 对假设检验而言, 所声明的犯第一类错误的概率比实际的更高. 克服模型选择方法这个缺陷的一个有效方式是采用模型平均方法.

所谓模型平均, 就是将来自不同模型的估计量或预测进行加权平均, 它并没有把某个特定的模型当作真实的数据产生过程, 也就不存在由于进行模型选择而产生的随机性. 另外, 由模型平均方法得

英文引用格式: Zhu R, Zou G H. The asymptotic theory for model averaging in general semiparametric models (in Chinese). Sci Sin Math, 2018, 48: 1019–1052, doi: 10.1360/N012017-00010

到的加权的估计量或预测充分地利用了各个模型的信息, 因而常常具有更高的精度, 也更加稳健. 从纯粹进行估计或预测的角度看, 模型平均方法可视为模型选择方法的推广. 与统计学有两大学派 (Bayes 学派和频率学派) 一样, 模型平均也有这两大方向: Bayes 模型平均和频率模型平均. 关于 Bayes 模型平均的研究, 可参见文献 [5] 和文献 [6, 第 11 章]. 本文考虑频率模型平均方法, 它主要关注两个问题: 一个是最优权重的选取, 目前文献中已在多种模型下提出了权重选择准则, 包括光滑化的 AIC 和 BIC 方法 [7]、基于 Mallows 准则的 MMA (Mallows model averaging) [8,9]、基于 Jackknife 准则的 JMA (jackknife model averaging) [10,11]、基于均方误差无偏估计的 OPT (optimal model averaging) [12] 和基于删组交叉验证的 LsoMA (leave-subject-out model averaging) [13] 等; 另一个是模型平均估计量的渐近分布, 文献 [14] 首次对此进行了研究, 在局部误设定的框架下, 对一般的参数模型建立了模型平均估计的渐近分布理论. 自从文献 [14] 的工作发表以来, 在模型平均估计的渐近分布理论方面已取得了丰富的成果. 例如, 文献 [15] 在 Cox 风险模型下获得了模型平均估计的渐近分布; 文献 [16] 将模型扩展到了一般的半参数模型, 也导出了模型平均估计的渐近分布, 为了简单起见, 文献 [16] 只考虑了全模型和零模型这两个候选模型的情形; 文献 [17, 18] 分别研究了可加部分线性模型和广义可加部分线性模型的情形; 文献 [19] 研究了部分线性单指标模型, 均获得了模型平均估计的渐近分布; 文献 [20] 结合广义估计方程发展了纵向数据下的模型平均方法, 也获得了估计量的渐近分布. 对测量误差模型平均方法的研究, 文献 [21] 首先考虑了线性测量误差模型的情形; 之后, 文献 [22, 23] 发展了变系数部分线性测量误差模型下的模型平均方法, 导出了模型平均估计的渐近分布, 这两篇文献的差别在于前者只考虑了线性部分的协变量带有测量误差的情形, 而后者考虑了模型的线性部分和变系数部分的协变量都有测量误差的情形; 文献 [24] 则研究了响应变量删失情况下线性模型的模型平均估计. 关于混合效应模型平均估计和分位回归模型平均估计的渐近分布的推导, 可参见文献 [25–28].

文献 [14] 研究的是一般的参数模型, 本文将把其工作从参数模型推广到半参数模型, 对感兴趣的模型参数的函数进行模型平均估计. 我们首先对半参数模型的非参数部分进行局部线性估计, 然后在给定该局部线性估计的情况下, 最大化广义轮廓对数似然函数, 得到模型参数的广义轮廓似然估计, 由此获得感兴趣的参数函数的估计, 进而得到其模型平均估计. 我们的重点是推导模型平均估计量的渐近分布, 虽然早在 2007 年文献 [16] 就考虑了完全相同的模型, 但我们的方法与他们的不一样: 文献 [16] 主要利用 Le Cam 的邻近理论, 以真实参数为基准, 给出了全模型和零模型对应的参数估计与真实参数之差的极限分布, 也给出了相应的模型平均估计量的极限分布; 我们则是基于文献 [14] 的思路, 对真实的密度函数 $f_{\rm true}(y, z, \beta_{\rm true}, \theta_{\rm true}(x))$ 在零模型处进行 Taylor 展开, 以零模型的参数为基准, 给出了所有候选模型对应的参数估计和零模型的参数之差的极限分布, 而不仅仅是文献 [16] 中的全模型和零模型, 也给出了相应的模型平均估计量的极限分布. 值得指出的是, 文献 [16] 的渐近分布依赖于真实模型参数, 与样本量有关, 而我们的依赖于零模型参数, 与样本量无关, 尽管两者在形式上是相当的.

本文结构如下: 第 2 节给出模型框架和模型的参数部分与非参数部分的估计; 第 3 节研究候选模型下参数及其函数的估计量的渐近分布; 第 4 节主要推导模型平均估计量的渐近分布, 并构造一个覆盖真实参数的概率趋于给定水平的置信区间; 第 5 节是模拟研究; 第 6 节通过对波士顿郊区房价数据的分析来说明模型平均方法在实际问题中的应用; 第 7 节是总结; 假设条件和定理的证明放在附录中.

2 模型框架和估计

由于我们采用似然的思想研究估计量的渐近分布理论, 故只需要考虑半参数模型的对数似然函数

即可. 记

$$\sum_{i=1}^{n} l(y_i, z_i, \beta, \theta(x_i)) = \sum_{i=1}^{n} \log f(y_i, z_i, \beta, \theta(x_i)), \tag{2.1}$$

其中 $\{y_i, z_i, x_i\}_{i=1}^n$ 表示一组独立同分布的样本; y_i 表示与协变量 z_i 和 x_i 相关的响应变量; β 表示模型的参数; z_i 表示参数协变量; $\theta(x_i)$ 是关于协变量 x_i 的未知函数, 表示模型的非参数部分. 记参数

$$\beta = \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_{(p+q)\times 1},$$

其中 α 是 p 维参数, 表示确定存在于模型中的变量对应的参数; γ 是 q 维参数, 表示不确定是否存在于模型中的变量对应的参数. 又记

$$\beta_0 = \begin{pmatrix} \alpha_0 \\ \gamma_0 \end{pmatrix}_{(p+q)\times 1}$$

表示零模型的参数, 其中 γ_0 是已知的 q 维向量;

$$\beta_{\rm true} = \begin{pmatrix} \alpha_{\rm true} \\ \gamma_{\rm true} \end{pmatrix}_{(p+q)\times 1} = \begin{pmatrix} \alpha_0 \\ \gamma_0 + \frac{\delta}{\sqrt{n}} \end{pmatrix}_{(p+q)\times 1}$$

表示真实模型的参数, 其中 $\gamma_{\rm true}$ 在 γ_0 附近有一个 $O(1/\sqrt{n})$ 的扰动; $\theta_{\rm true}(x_i)$ 表示真实模型的非参数部分. 我们的目的是估计参数向量的函数 $\mu(\beta_{\rm true})$, 记为 $\mu_{\rm true}$.

我们用下标 S 表示候选模型, 其中 $S \subset \{1, 2, \ldots, q\}$. 记 π_S 为 $|S| \times q$ 维的选择矩阵, 将 v 向量映射到它的子向量 v_S 上, 它只保留了 v_j $(j \in S)$ 的那些分量, 即 $\pi_S v = v_S$, 其中 $|S|$ 表示候选模型 S 中 γ_S 对应的参数个数. 再记

$$\beta_{0,S} = \begin{pmatrix} \alpha_0 \\ \gamma_{0,S} \end{pmatrix}_{(p+|S|)\times 1} = \begin{pmatrix} I_p & \\ & \pi_S \end{pmatrix} \beta_0,$$

$$\beta_S = \begin{pmatrix} \alpha_0 \\ \gamma_S \end{pmatrix}_{(p+|S|)\times 1} = \begin{pmatrix} \alpha_0 \\ \gamma_{0,S} + \frac{\delta_S}{\sqrt{n}} \end{pmatrix}_{(p+|S|)\times 1} = \begin{pmatrix} I_p & \\ & \pi_S \end{pmatrix} \beta,$$

后者是候选模型 S 对应的参数.

为了估计 $\mu(\beta_{\rm true})$, 我们先对模型的非参数部分进行估计, 然后在给定非参数估计的情形下, 最大化轮廓对数似然函数, 得到参数的广义轮廓似然估计, 再对目标量进行模型平均. 对非参数部分的估计, 方法很多, 本文类似于文献 [16], 采用局部线性估计:

$$[\hat{\psi}(x, \beta), \hat{\psi}_1(x, \beta)] = \arg\max_{\psi, \psi_1} \frac{1}{n} \sum_{i=1}^{n} l[y_i, z_i, \beta, \psi(x) + \psi_1(x)(x_i - x)] K_h(x_i - x), \tag{2.2}$$

即 $\hat{\theta}(x, \beta) = \hat{\psi}(x, \beta)$ 表示在给定 β 的情形下, $\theta(\cdot)$ 在 $X = x$ 处的局部线性估计, 其中 $K(\cdot)$ 表示定义在有界闭区间上的对称且连续可微的核函数, h 表示窗宽, $K_h(\cdot) = K(\cdot/h)/h$. 我们也假定边际密度函数 $f_X(\cdot)$ 在有界闭区间 $\mathcal{B}$ 上是连续可微的, 且 $\inf_{x\in\mathcal{B}} f_X(x) > 0$. 由文献 [16, 引理 A1] 可知, 在一定的正则条件下, 对于给定的 β, 非参数部分 $\theta(x)$ 的局部线性估计 $\hat{\theta}(x, \beta)$ 总是存在的, 且具有相合性, 即 $\hat{\theta}(x, \beta) - \theta(x, \beta) \xrightarrow{p} 0$, 其中 $\theta(x, \beta) = \arg\max_\theta {\rm E}[l(y, z, \beta, \theta(x, \beta)) \mid x]$, 且在 $\beta_{\rm true}$ 处, 满足 $\theta(x, \beta_{\rm true}) = \theta_{\rm true}(x)$, 而 ${\rm E}(\cdot)$ 表示真实模型下的数学期望.

对非参数部分采用局部线性估计获得的对数似然函数为 $\sum_{i=1}^n l[y_i,z_i,\beta,\hat{\theta}(x_i,\beta)]$, 由此得到 β 的广义轮廓似然估计 $\hat{\beta}$, 即 $\hat{\beta}$ 满足

$$0=\sum_{i=1}^n \frac{d}{d\beta} l[y_i,z_i,\beta,\hat{\theta}(x_i,\beta)]\Big|_{\beta=\hat{\beta}}. \tag{2.3}$$

为简单起见, 在候选模型下非参数部分的估计仍记为 $\hat{\theta}(x,\beta_S)$, 于是, 候选模型 S 下的估计 $\hat{\beta}_S$ 满足

$$0=\sum_{i=1}^n \frac{d}{d\beta_S} l[y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S)]\Big|_{\beta_S=\hat{\beta}_S}. \tag{2.4}$$

3 候选模型下估计量的渐近分布

3.1 记号

为方便起见, 采用文献 [14] 中的符号, 但是含义有所区别. 记 $f_0(y)=f(y,z,\beta_0,\theta_{\text{true}}(x))$ 表示零模型下的密度函数. 由于既有参数部分, 也有非参数部分, 所以得分函数是对数似然函数对参数 β 的全导数, 记为

$$\begin{aligned} u(y)&=\frac{d\log f(y,z,\beta,\hat{\theta}(x,\beta))}{d\alpha}\Big|_{\beta=\beta_0}, \quad v(y)=\frac{d\log f(y,z,\beta,\hat{\theta}(x,\beta))}{d\gamma}\Big|_{\beta=\beta_0},\\ u_0(y)&=\frac{d\log f(y,z,\beta,\theta_0(x,\beta))}{d\alpha}\Big|_{\beta=\beta_0}, \quad v_0(y)=\frac{d\log f(y,z,\beta,\theta_0(x,\beta))}{d\gamma}\Big|_{\beta=\beta_0}, \end{aligned}$$

其中 $\theta_0(x,\beta)=\arg\max_\theta \mathrm{E}_0[l(y,z,\beta,\theta(x,\beta))\mid x]$, 且在 β_0 处, 满足 $\theta_0(x,\beta_0)=\theta_{\text{true}}(x)$, 而 $\mathrm{E}_0(\cdot)$ 表示零模型下的数学期望.

类似于文献 [16, (2) 和 (3)], 假定

$$\mathrm{E}_0\left[\frac{\partial l(y,z,\beta_0,\theta_{\text{true}}(x))}{\partial\beta}\;\Big|\; x\right]=0, \quad \mathrm{E}_0\left[\frac{\partial l(y,z,\beta_0,\theta_{\text{true}}(x))}{\partial\theta}\;\Big|\; x\right]=0. \tag{3.1}$$

由此可得

$$\begin{aligned} &\mathrm{E}_0\left[\frac{dl(y,z,\beta,\theta_0(x,\beta))}{d\beta}\Big|_{\beta=\beta_0}\;\Big|\; x\right]\\ &=\mathrm{E}_0\left[\left(\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\beta}+\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}\frac{\partial\theta_0(x,\beta)}{\partial\beta}\right)\Big|_{\beta=\beta_0}\;\Big|\; x\right]\\ &=\mathrm{E}_0\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\beta}\Big|_{\beta=\beta_0}\;\Big|\; x\right]+\mathrm{E}_0\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}\Big|_{\beta=\beta_0}\;\Big|\; x\right]\frac{\partial\theta_0(x,\beta)}{\partial\beta}\Big|_{\beta=\beta_0}\\ &=\mathrm{E}_0\left[\frac{\partial l(y,z,\beta_0,\theta_{\text{true}}(x))}{\partial\beta}\;\Big|\; x\right]+\mathrm{E}_0\left[\frac{\partial l(y,z,\beta_0,\theta_{\text{true}}(x))}{\partial\theta}\;\Big|\; x\right]\frac{\partial\theta_0(x,\beta)}{\partial\beta}\Big|_{\beta=\beta_0}\\ &=0, \end{aligned}$$

即

$$\mathrm{E}_0(u_0(y))=0, \quad \mathrm{E}_0(v_0(y))=0. \tag{3.2}$$

进一步地, 记

$$\bar{u}_n = \frac{1}{n}\sum_{i=1}^n u(y_i), \quad \bar{v}_n = \frac{1}{n}\sum_{i=1}^n v(y_i), \quad \bar{u}_0 = \frac{1}{n}\sum_{i=1}^n u_0(y_i), \quad \bar{v}_0 = \frac{1}{n}\sum_{i=1}^n v_0(y_i),$$

$u_0(y)$ 和 $v_0(y)$ 在 $f_0(y)$ 下的协方差矩阵为

$$J = \mathrm{Var}_0\begin{pmatrix} u_0(y) \\ v_0(y) \end{pmatrix} = \mathrm{Var}_0\left(\frac{d\log f(y,z,\beta,\theta_0(x,\beta))}{d\beta}\bigg|_{\beta=\beta_0}\right) = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}_{(p+q)\times(p+q)}, \tag{3.3}$$

有逆

$$J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}.$$

对候选模型 S, 有类似的定义. 特别地,

$$\begin{aligned} & u_S(y) = u(y), \quad v_S(y) = \pi_S v(y), \quad u_{0,S}(y) = u_0(y), \quad v_{0,S}(y) = \pi_S v_0(y), \\ & \bar{u}_{n,S} = \frac{1}{n}\sum_{i=1}^n u_S(y_i), \quad \bar{v}_{n,S} = \frac{1}{n}\sum_{i=1}^n v_S(y_i), \quad \bar{u}_{0,S} = \frac{1}{n}\sum_{i=1}^n u_{0,S}(y_i), \quad \bar{v}_{0,S} = \frac{1}{n}\sum_{i=1}^n v_{0,S}(y_i), \\ & J_S = \mathrm{Var}_0\begin{pmatrix} u_0(y) \\ v_{0,S}(y) \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01}\pi_S^{\mathrm{T}} \\ \pi_S J_{10} & \pi_S J_{11}\pi_S^{\mathrm{T}} \end{pmatrix}_{(p+|S|)\times(p+|S|)}, \end{aligned} \tag{3.4}$$

其中 π_S^{T} 表示矩阵 π_S 的转置, 有逆

$$J_S^{-1} = \begin{pmatrix} J^{00,S} & J^{01,S} \\ J^{10,S} & J^{11,S} \end{pmatrix}.$$

令 $K = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$. 经过简单的矩阵计算, 可以得到

$$\begin{aligned} & (\pi_S K^{-1}\pi_S^{\mathrm{T}})^{-1} = K_S, \quad J^{01,S} = -J_{00}^{-1}J_{01}\pi_S^{\mathrm{T}}K_S, \\ & J^{00,S} = J_{00}^{-1} + J_{00}^{-1}J_{01}\pi_S^{\mathrm{T}}K_S\pi_S J_{10}J_{00}^{-1}. \end{aligned}$$

3.2 候选模型下广义轮廓似然估计的渐近分布

为了导出候选模型下广义轮廓似然估计的渐近分布, 我们需要如下引理.

引理 1 在条件 1 和 2 下, 有

$$\begin{pmatrix} \sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_n \end{pmatrix} - \begin{pmatrix} \sqrt{n}\bar{u}_0 \\ \sqrt{n}\bar{v}_0 \end{pmatrix} = o_p(1). \tag{3.5}$$

由此利用 $\bar{u}_n$ 和 $\bar{v}_n$ 与 $\bar{u}_{n,S}$ 和 $\bar{v}_{n,S}$, 以及 $\bar{u}_0$ 和 $\bar{v}_0$ 与 $\bar{u}_{0,S}$ 和 $\bar{v}_{0,S}$ 之间的关系, 可得到

$$\begin{pmatrix} \sqrt{n}\bar{u}_{n,S} \\ \sqrt{n}\bar{v}_{n,S} \end{pmatrix} - \begin{pmatrix} \sqrt{n}\bar{u}_{0,S} \\ \sqrt{n}\bar{v}_{0,S} \end{pmatrix} = o_p(1). \tag{3.6}$$

该引理表明广义轮廓对数似然函数的全导数和对数似然函数的全导数在 β_0 处实际上很接近, 是渐近等价的.

引理 2 在条件 1–3 下, 有

$$\begin{pmatrix} \sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} J_{01}\delta \\ J_{11}\delta \end{pmatrix} + \begin{pmatrix} M \\ N \end{pmatrix}, \tag{3.7}$$

其中

$$\begin{pmatrix} M \\ N \end{pmatrix} \sim N_{p+q}(0, J).$$

由此知, 对基于候选模型 S 的估计, 有

$$\begin{pmatrix} \sqrt{n}\bar{u}_{n,S} \\ \sqrt{n}\bar{v}_{n,S} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} J_{01}\delta \\ \pi_S J_{11}\delta \end{pmatrix} + \begin{pmatrix} M \\ N_S \end{pmatrix}, \tag{3.8}$$

其中

$$\begin{pmatrix} M \\ N_S \end{pmatrix} \sim N_{p+|S|}(0, J_S),$$

而 $|S|$ 表示 S 中元素的个数.

该引理得到了广义轮廓对数似然函数关于 β 的全导数的渐近分布.

引理 3 在条件 1 和 2 下, 可得

$$\sup_{\beta\in\mathcal{N}(\beta_0)} \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\beta d\beta^{\mathrm{T}}} l(y_i, z_i, \beta, \hat{\theta}(x_i, \beta)) - \frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\beta d\beta^{\mathrm{T}}} l(y_i, z_i, \beta, \theta_0(x_i, \beta)) \right\| = o_p(1), \tag{3.9}$$

其中 $\mathcal{N}(\beta_0)$ 表示 β_0 的邻域, $\|A\| = \max_{i,j}|a_{ij}|$, a_{ij} 是矩阵 A 的 (i,j) 元素.

该引理说明广义轮廓对数似然函数的二阶全导数和对数似然函数的二阶全导数之间的差值在 β_0 的邻域是一致为 $o_p(1)$ 的.

引理 4 在条件 1 和 2 下, $\beta_{0,S}$ 的相合估计一定存在; 进一步, 如果不存在处处收敛于 $\beta_{0,S}$ 的估计, 则广义轮廓对数似然方程以趋于 1 的概率存在相合解.

该引理说明了在一定的正则条件下 $\hat{\beta}_S$ 具有相合性, 而下述引理则建立了其渐近正态性.

引理 5 在条件 1–3 下, 有

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\hat{\alpha}_S - \alpha_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} &= \sqrt{n}(\hat{\beta}_S - \beta_{0,S}) \doteq J_S^{-1} \begin{pmatrix} \sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} C_S \\ D_S \end{pmatrix} \\ &\sim N\left(J_S^{-1} \begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix} \delta,\ J_S^{-1} \right), \end{aligned} \tag{3.10}$$

其中 $\gamma_{0,S} = \pi_S\gamma_0$, 而 $x_n \doteq y_n$ 表示两边变量之差依概率趋于 0, 即 $x_n - y_n \xrightarrow{p} 0$, 故有相同的极限分布.

类似文献 [16] 中的记号, 令 $\hat{\beta}_{\text{full}}$ 和 $\hat{\alpha}_{\text{red}}$ 分别表示全模型和零模型下相应参数的估计. 当 S 取全集时可得

$$\sqrt{n}(\hat{\beta}_{\text{full}} - \beta_0) \xrightarrow{d} N\left(\begin{pmatrix} 0_{p\times q} \\ I_q \end{pmatrix} \delta, J^{-1} \right),$$

由此知,

$$\sqrt{n}(\hat{\beta}_{\text{full}} - \beta_{\text{true}}) = \sqrt{n}(\hat{\beta}_{\text{full}} - \beta_0) - \sqrt{n}(\beta_{\text{true}} - \beta_0) = \sqrt{n}(\hat{\beta}_{\text{full}} - \beta_0) - \begin{pmatrix} 0_{p\times q} \\ I_q \end{pmatrix} \delta \xrightarrow{d} N(0, J^{-1}).$$

类比文献 [16, 定理 1] 可以看出, 两者有相同的渐近分布结构, 只不过本文采用的是基于 β_0 的协方差矩阵, 而文献 [16] 采用的是基于 β_{true} 的协方差矩阵. 当 S 取空集时情形类似: $\sqrt{n}(\hat{\alpha}_{\text{red}} - \alpha_{\text{true}}) \xrightarrow{d} N(J_{00}^{-1}J_{01}\delta, J_{00}^{-1})$, 这类似于文献 [16, 定理 3].

3.3 候选模型下参数函数估计的渐近分布

为了方便叙述, 我们先引入一些记号. 令

$$W = J^{10}M + J^{11}N = K(N - J_{10}J_{00}^{-1}M).$$

简单计算知 $\mathrm{Cov}(W, M) = 0$, 且 W 和 M 是联合正态的, 故 W 和 M 相互独立. 由引理 5 可得

$$\hat{\delta}_S \hat{=} \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \xrightarrow{d} D_S = K_S\pi_S K^{-1}(\delta + W),$$

特别地,

$$D_n \hat{=} \hat{\delta}_{\text{full}} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0) \xrightarrow{d} D = \delta + W \sim N_q(\delta, K).$$

记 $H_S = K^{-1/2}\pi_S^{\mathrm{T}}K_S\pi_S K^{-1/2}$, 它是 $q \times q$ 的对称幂等阵, $H_{\emptyset}$ 是 $q \times q$ 的零矩阵; $\omega = J_{10}J_{00}^{-1}\frac{\partial\mu}{\partial\alpha} - \frac{\partial\mu}{\partial\gamma}$, 其中偏导数是在 (α_0, γ_0) 处取值的; $\mu(\alpha, \gamma)$ 在候选模型 S 下的估计记为 $\hat{\mu}_S = \mu(\hat{\alpha}_S, \hat{\gamma}_S, \gamma_{0,S^c})$.

引理 6 在条件 1–3 下, 假设 $\mu(\alpha, \gamma)$ 在 (α_0, γ_0) 的邻域内有二阶连续偏导数, 且偏导数有限, 则有

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S = \left(\frac{\partial\mu}{\partial\alpha}\right)^{\mathrm{T}} J_{00}^{-1}M + \omega^{\mathrm{T}}(\delta - K^{\frac{1}{2}}H_S K^{-\frac{1}{2}}D). \tag{3.11}$$

易见 Λ_S 服从均值为 $\omega^{\mathrm{T}}(I - K^{1/2}H_S K^{-1/2})\delta$, 方差为 $\tau_0^2 + \omega^{\mathrm{T}}K^{1/2}H_S K^{1/2}\omega$ 的正态分布, 其中

$$\tau_0^2 = \left(\frac{\partial\mu}{\partial\alpha}\right)^{\mathrm{T}} J_{00}^{-1}\left(\frac{\partial\mu}{\partial\alpha}\right).$$

当 S 取全集时,

$$\sqrt{n}(\mu(\hat{\beta}_{\text{full}}) - \mu(\beta_{\text{true}})) \xrightarrow{d} \Lambda_{\text{full}} = \left(\frac{\partial\mu}{\partial\beta}\right)^{\mathrm{T}} N(0, J^{-1});$$

而当 S 取空集时,

$$\sqrt{n}(\mu(\hat{\beta}_{\text{red}}) - \mu(\beta_{\text{true}})) \xrightarrow{d} \Lambda_{\text{red}} = \left(\frac{\partial\mu}{\partial\alpha}\right)^{\mathrm{T}} N(J_{00}^{-1}J_{01}\delta, J_{00}^{-1}) - \left(\frac{\partial\mu}{\partial\gamma}\right)^{\mathrm{T}}\delta.$$

与引理 5 类似, 这两种情形下的渐近分布与文献 [16, 定理 5] 中的渐近分布有相同的结构.

4 模型平均估计及推断

4.1 0-1 权重估计量

模型选择的结果最终导致的其实是 0-1 权重的估计量 (从纯粹估计的角度, 即是选择后估计量, 通常的预检验估计量是其一种形式), 也就是对所选中的候选模型赋予权重 1, 而其余候选模型的权重全为 0, 它可看作是模型平均方法的特例. 类似于文献 [16], 我们对模型选择准则是 AIC 的情形讨论权重的形式. 在本文的模型下, AIC 准则定义为

$$\mathrm{AIC}_{n,S} = 2\sum_{i=1}^{n} l(y_i, z_i, \hat{\alpha}_S, \hat{\gamma}_S, \hat{\theta}(x_i, \hat{\alpha}_S, \hat{\gamma}_S)) - 2|S|, \tag{4.1}$$

此处为记号简单起见, 我们省略了 γ_{0,S^C} 这个包含在第 S 个候选模型中的已知向量, 我们的目标是选出使 AIC 值达到最大的候选模型. 以 $\hat{S}_{\mathrm{AIC}}$ 表示被选中的模型, $\hat{\mu}(S)$ 表示 μ 在候选模型 S 下的估计量, 则通过 AIC 模型选择方法得到的估计量为

$$\hat{\mu} = \hat{\mu}(\hat{S}_{\mathrm{AIC}}) = \sum_{S} I(\mathrm{AIC}\text{ 选中模型 } S)\hat{\mu}(S).$$

考虑似然比统计量

$$\begin{aligned}
G_{n,S} &= 2\sum_{i=1}^{n}\log\left(\frac{f(y_i, z_i, \hat{\alpha}_S, \hat{\gamma}_S, \hat{\theta}(x_i, \hat{\alpha}_S, \hat{\gamma}_S))}{f(y_i, z_i, \alpha_0, \gamma_0, \hat{\theta}(x_i, \alpha_0, \gamma_0))}\right) \\
&= 2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{d}{d\beta_S}l(y_i, z_i, \beta_S, \hat{\theta}(x_i, \beta_S))\,\Big|_{\beta_S=\beta_{0,S}}\right]^{\mathrm{T}}\sqrt{n}(\hat{\beta}_S - \beta_{0,S}) \\
&\quad + \sqrt{n}(\hat{\beta}_S - \beta_{0,S})^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}}l(y_i, z_i, \beta_S, \hat{\theta}(x_i, \beta_S))\,\Big|_{\beta_S=\beta_{0,S}^*}\right]\sqrt{n}(\hat{\beta}_S - \beta_{0,S}),
\end{aligned}$$

其中 $\beta_{0,S}^*$ 介于 $\hat{\beta}_S$ 与 $\beta_{0,S}$ 之间. 由引理 1、3、5、大数定律和 Slutsky 定理可得

$$\begin{aligned}
G_{n,S} &\doteq 2\begin{pmatrix}\sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix}^{\mathrm{T}} J_S^{-1}\begin{pmatrix}\sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix} - \left(J_S^{-1}\begin{pmatrix}\sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix}\right)^{\mathrm{T}} J_S\left(J_S^{-1}\begin{pmatrix}\sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix}\right) \\
&= \begin{pmatrix}\sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix}^{\mathrm{T}} J_S^{-1}\begin{pmatrix}\sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix} \\
&\xrightarrow{d} \begin{pmatrix}J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S\end{pmatrix}^{\mathrm{T}} J_S^{-1}\begin{pmatrix}J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S\end{pmatrix},
\end{aligned}$$

它服从非中心的自由度为 $p+|S|$ 的卡方分布. 令 $G_{n,\emptyset}$ 表示零模型下的似然比统计量, 结合第 3.1 小节的记号, 进一步可得

$$\begin{aligned}
G_{n,S} - G_{n,\emptyset} &\doteq \begin{pmatrix}\sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix}^{\mathrm{T}} J_S^{-1}\begin{pmatrix}\sqrt{n}\bar{u}_n \\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix} - n\bar{u}_n^{\mathrm{T}} J_{00}^{-1}\bar{u}_n \\
&= n(\bar{v}_{n,S} - J_{10,S}J_{00}^{-1}\bar{u}_n)^{\mathrm{T}} J_{11,S}(\bar{v}_{n,S} - J_{10,S}J_{00}^{-1}\bar{u}_n) \\
&\xrightarrow{d} (K_S^{-1}\delta + N_S - J_{10,S}J_{00}^{-1}M)^{\mathrm{T}} K_S(K_S^{-1}\delta + N_S - J_{10,S}J_{00}^{-1}M),
\end{aligned}$$

它服从非中心的自由度为 $|S|$ 的卡方分布. 由于 $\hat{\delta}_S = \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S})$ 和 $\sqrt{n}K_S\pi_S(\bar{v}_n - J_{10}J_{00}^{-1}\bar{u}_n)$ 至多相差 $o_p(1)$, 即 $\hat{\delta}_S \doteq \sqrt{n}K_S\pi_S(\bar{v}_n - J_{10}J_{00}^{-1}\bar{u}_n)$, 而当 S 取全集时, $D_n = \hat{\delta}_{\mathrm{full}} \doteq \sqrt{n}K(\bar{v}_n - J_{10}J_{00}^{-1}\bar{u}_n)$, 故可得 $\hat{\delta}_S \doteq K_S\pi_S K^{-1}D_n$. 这样 AIC 可以由 D_n 来表示:

$$\mathrm{AIC}_{n,S} = G_{n,S} - G_{n,\emptyset} - 2|S| = D_n^{\mathrm{T}} K^{-\frac{1}{2}} H_S K^{-\frac{1}{2}} D_n - 2|S| + o_p(1). \tag{4.2}$$

我们以只有全模型和零模型这两个候选模型的情形为例. 令 $\mathrm{AIC}_{\mathrm{full}}$ 和 $\mathrm{AIC}_{\mathrm{red}}$ 分别表示全模型和零模型对应的 AIC 值: $\mathrm{AIC}_{\mathrm{full}} = D_n^{\mathrm{T}} K^{-1} D_n - 2q$, $\mathrm{AIC}_{\mathrm{red}} = 0$. 当 $\mathrm{AIC}_{\mathrm{full}} \geqslant \mathrm{AIC}_{\mathrm{red}}$, 即 $D_n^{\mathrm{T}} K^{-1} D_n \geqslant 2q$ 时, 将选全模型; 反之, 则选零模型. 因此, AIC 模型选择方法的权重为 $c(D_n) = I(D_n^{\mathrm{T}} K^{-1} D_n \geqslant 2q)$, 因而 AIC 方法的估计量为 $\hat{\mu} = c(D_n)\mu(\hat{\beta}_{\mathrm{full}}) + (1 - c(D_n))\mu(\hat{\beta}_{\mathrm{red}})$.

4.2 模型平均估计的渐近分布

对于感兴趣的参数函数, 考虑如下形式的模型平均估计:

$$\hat{\mu} = \sum_S c(S \mid D_n)\hat{\mu}_S, \tag{4.3}$$

其中 $c(S \mid D_n)$ 是权重函数, 其和为 1.

记 $G(d) = K^{-\frac{1}{2}}(\sum_S c(S \mid d)H_S)K^{\frac{1}{2}}$ 是关于 $d = (d_1, d_2, \ldots, d_q)^{\mathrm{T}}$ 的 $q \times q$ 矩阵函数,

$$\hat{\delta}(D) = G^{\mathrm{T}}(D)D = K^{\frac{1}{2}}\left(\sum_S c(S \mid D)H_S\right)K^{-\frac{1}{2}}D$$

是 δ 的基于 D 的估计, 则对模型平均估计 $\hat{\mu}$, 有如下的渐近结果:

定理 1 假设权重函数 $c(S \mid d)$ 关于 d 连续, 则有

$$\sqrt{n}(\hat{\mu} - \mu_{\mathrm{true}}) \xrightarrow{d} \Lambda = \sum_S c(S \mid D)\Lambda_S = \Lambda_0 + \omega^{\mathrm{T}}(\delta - \hat{\delta}(D)), \tag{4.4}$$

其中 $\Lambda_0 = (\frac{\partial\mu}{\partial\alpha})^{\mathrm{T}} J_{00}^{-1} M$. Λ 的均值和方差分别为 $\omega^{\mathrm{T}}[\delta - \mathrm{E}\hat{\delta}(D)]$ 和 $\tau_0^2 + \omega^{\mathrm{T}}\mathrm{Var}(\hat{\delta}(D))\omega$. 因而, Λ 的均方误差为 $\mathrm{E}\Lambda^2 = \tau_0^2 + R(\delta)$, 其中 $R(\delta) = \mathrm{E}(\omega^{\mathrm{T}}\hat{\delta}(D) - \omega^{\mathrm{T}}\delta)^2 = \omega^{\mathrm{T}}\mathrm{E}[(\hat{\delta}(D) - \delta)(\hat{\delta}(D) - \delta)^{\mathrm{T}}]\omega$.

4.3 置信区间

用传统方法构造的 μ 的置信水平为 $1 - \alpha$ 的置信区间是

$$\left[\hat{\mu}_{\hat{S}} - u_{\alpha/2}\frac{\hat{\tau}_{\hat{S}}}{\sqrt{n}}, \hat{\mu}_{\hat{S}} + u_{\alpha/2}\frac{\hat{\tau}_{\hat{S}}}{\sqrt{n}}\right], \tag{4.5}$$

其中 $\hat{S}$ 表示选中的模型, $\frac{\hat{\tau}_{\hat{S}}}{\sqrt{n}}$ 表示 $\hat{\mu}_{\hat{S}}$ 的标准差的估计, $u_{\alpha/2}$ 表示标准正态分布的上 $\alpha/2$ 分位数. 由于没有考虑模型选择的不确定性, 导致估计的结果过于乐观, 方差偏小, 使得在平均意义下, 置信区间偏小, 实际上达不到之前设定的置信水平. 我们参照文献 [14] 提出的方法构造一个覆盖真实参数的概率趋于预定水平的置信区间. 考虑统计量

$$T_n = \frac{\sqrt{n}(\hat{\mu} - \mu_{\mathrm{true}}) - \hat{\omega}^{\mathrm{T}}(D_n - \hat{\delta}(D_n))}{\hat{\kappa}}, \tag{4.6}$$

其中 $\hat{\omega}$ 和 $\hat{\kappa}$ 分别是 ω 和 $\kappa = (\tau_0^2 + \omega^{\mathrm{T}}K\omega)^{\frac{1}{2}} = [(\frac{\partial\mu}{\partial\alpha})^{\mathrm{T}} J_{00}^{-1}(\frac{\partial\mu}{\partial\alpha}) + \omega^{\mathrm{T}}K\omega]^{\frac{1}{2}}$ 的相合估计. 定义 μ_{true} 的置信区间的上下限分别为

$$\begin{aligned} \mathrm{low}_n &= \hat{\mu} - \hat{\omega}^{\mathrm{T}}\frac{(D_n - \hat{\delta}(D_n))}{\sqrt{n}} - \frac{u_{\alpha/2}\hat{\kappa}}{\sqrt{n}}, \\ \mathrm{up}_n &= \hat{\mu} - \hat{\omega}^{\mathrm{T}}\frac{(D_n - \hat{\delta}(D_n))}{\sqrt{n}} + \frac{u_{\alpha/2}\hat{\kappa}}{\sqrt{n}}. \end{aligned} \tag{4.7}$$

定理 2 统计量 T_n 依分布收敛于标准正态分布, 而置信区间 $[\mathrm{low}_n, \mathrm{up}_n]$ 覆盖真实参数的概率趋于设定的置信水平 $1 - \alpha$, 即

$$\Pr(\mu_{\mathrm{true}} \in [\mathrm{low}_n, \mathrm{up}_n]) \to 2\Phi(u_{\alpha/2}) - 1 = 1 - \alpha, \tag{4.8}$$

其中 Φ 表示标准正态分布函数.

5 模拟研究

本节通过随机模拟研究模型平均方法在有限样本量下的表现. 考虑如下的半参数模型:

$$y_i = z_i^{\mathrm{T}}\beta + \theta(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{5.1}$$

其中

$$z_i = (z_{i1}, \ldots, z_{i6})^{\mathrm{T}} \sim N_6(0, \Sigma_z), \quad \beta = \left(1, 2, 3, \frac{c}{\sqrt{n}}, \frac{c}{\sqrt{n}}, \frac{c}{\sqrt{n}}\right)^{\mathrm{T}}, \quad x_i \sim U(0,1),$$

协方差阵 Σ_z 的 (j,k) 元素为 $\Sigma_{zjk} = (1/2)^{|j-k|}$, $j,k = 1,2,\ldots,6$, c 取值为 0 或 1, 随机误差项服从正态分布, 即 $\epsilon_i \sim N(0,\sigma^2)$, 而 σ 取值为 0.5 或 2. 我们考虑 4 种不同的非参数函数, 分别为 $\theta(x_i) = \sin(8x_i - 2)$、$\theta(x_i) = x_i^2 - 1$、$\theta(x_i) = \exp(x_i)$ 和 $\theta(x_i) = \log(x_i + 1)$. 为节省空间, 我们只给出 $\theta(x_i) = \sin(8x_i - 2)$ 的结果. 对于模型非参数部分的估计, 涉及非参估计的窗宽的选取问题, 由于该问题不是本文的重点, 因此, 我们直接采用 R 语言 ‘np’ 软件包里的 ‘npreg’ 函数, 设置了 Epanechnikov 核函数对非参部分进行局部线性估计. 设定 $\{z_{i1}, z_{i2}, z_{i3}\}$ 为每个子模型的必选变量, $\{z_{i4}, z_{i5}, z_{i6}\}$ 为子模型的候选变量, 这样共有 $S = 2^3 = 8$ 个候选模型. 我们感兴趣的参数为 $\mu_1 = \beta_2$ 和 $\mu_2 = \beta_1 + \beta_2 + \beta_3$. 样本量取为 $n = 100$ 和 $n = 200$, 并进行 $D = 500$ 次循环.

对于模型平均估计量, 我们采用 SAIC (smoothed-AIC) 和 SBIC (smoothed-BIC) 权重 (参见文献 [7]), 即

$$\mathrm{SAIC}_s = \frac{\exp(\mathrm{AIC}_s/2)}{\sum_S \exp(\mathrm{AIC}_S/2)} \quad 和 \quad \mathrm{SBIC}_s = \frac{\exp(\mathrm{BIC}_s/2)}{\sum_S \exp(\mathrm{BIC}_S/2)}. \tag{5.2}$$

除了这两种模型平均方法, 我们还考虑两种相应的模型选择方法, 即 AIC 和 BIC, 以及基于全模型的估计方法.

通过下式计算在不同设置下感兴趣参数的估计量的模拟均方误差 (mean squared error, MSE), 即

$$\mathrm{MSE} = \frac{1}{D}\sum_{d=1}^{D}(\hat{\mu}_i^{(d)} - \mu_i)^2, \quad i = 1, 2, \tag{5.3}$$

其中 $\hat{\mu}_i^{(d)}$ 表示在第 d 次循环中感兴趣的参数 μ_i 的估计值. 计算结果见表 1.

进一步, 我们通过 (4.5) 计算使用模型选择方法时感兴趣参数的置信区间, 通过 (4.7) 计算使用模型平均方法时感兴趣参数的置信区间, 置信水平分别取为 90% 和 95%. 表 2 和 3 给出了不同设置下置信区间的覆盖率和平均长度的计算结果. 在表 2 和 3 中, CP 表示置信区间的覆盖率, Length 表示置信区间的平均长度. SAIC 和 SBIC 的计算结果在保留四位小数之后是一样的, 因此合并到一起.

从表 1–3 我们可以得到如下结论:

(1) 表 1 表明大部分情况下, SAIC 和 SBIC 这两种模型平均方法的 MSE 比相应的模型选择方法 AIC 和 BIC 的 MSE 要小, 这说明模型平均方法比相应的模型选择方法更有优势, 而且基于全模型下的估计方法也是不如模型平均方法表现好.

(2) 表 1 中, 当 $c = 0$ 时, 模型平均和模型选择方法的 MSE 都小于全模型的 MSE; 当 $c = 1$ 时, 几乎在所有的情况下模型平均方法都优于全模型估计方法, 但是有小部分情况下的模型选择方法不如全模型方法表现好, 注意此时全模型是真实模型. 直观上看, 当全模型是真实模型时, 全模型方法的表现应是最好的, 但此处候选变量的系数很小 (是由一个常数 c 除以 $\sqrt{n}$ 组成的), 使得候选变量相对于必

表 1 不同设置下 5 种方法的 MSE ($\times 10^2$)

					$n=100$					$n=200$		
σ	c	μ	SAIC	SBIC	AIC	BIC	Full	SAIC	SBIC	AIC	BIC	Full
0.5	0	μ_1	2.077	2.065	2.092	2.069	2.099	0.418	0.414	0.418	0.411	0.426
		μ_2	9.708	9.683	9.737	9.668	9.832	1.080	1.067	1.085	1.065	1.116
	1	μ_1	1.251	1.263	1.266	1.293	1.251	1.310	1.310	1.320	1.321	1.313
		μ_2	2.326	2.319	2.421	2.539	2.463	8.085	8.049	8.170	8.129	8.170
2.0	0	μ_1	9.027	9.035	9.035	9.086	9.036	3.886	3.865	3.894	3.867	3.903
		μ_2	18.362	18.222	18.527	18.382	18.546	4.975	4.868	5.021	4.906	5.076
	1	μ_1	8.210	8.209	8.244	8.225	8.226	3.442	3.440	3.441	3.433	3.444
		μ_2	12.324	12.221	12.474	12.408	12.463	5.012	4.930	5.073	4.954	5.118

表 2 5 种方法置信区间的覆盖率和平均长度, $\sigma = 0.5$

					CP				Length		
置信水平	c	n	μ	SAIC (SBIC)	AIC	BIC	Full	SAIC (SBIC)	AIC	BIC	Full
90%	0	100	μ_1	0.888	0.880	0.880	0.870	0.3294	0.3169	0.3177	0.3194
			μ_2	0.866	0.852	0.846	0.854	0.3650	0.3262	0.3235	0.3539
		200	μ_1	0.890	0.880	0.881	0.882	0.2083	0.2045	0.2048	0.2051
			μ_2	0.874	0.858	0.872	0.872	0.2608	0.2530	0.2520	0.2711
	1	100	μ_1	0.882	0.874	0.876	0.876	0.3316	0.3216	0.3267	0.3217
			μ_2	0.850	**0.816**	0.810	0.836	0.3662	0.3395	0.3371	0.3545
		200	μ_1	0.910	0.904	0.905	0.906	0.2081	0.2055	0.2061	0.2061
			μ_2	**0.842**	0.822	**0.796**	**0.826**	0.2277	0.2134	0.2102	0.2242
95%	0	100	μ_1	0.952	0.948	0.946	0.946	0.3979	0.3842	0.3840	0.3858
			μ_2	**0.900**	0.888	0.892	**0.896**	0.4391	0.3923	0.3976	0.4258
		200	μ_1	0.948	0.948	0.948	0.944	0.2447	0.2404	0.2406	0.2410
			μ_2	0.934	0.912	0.908	0.926	0.2698	0.2443	0.2426	0.2657
	1	100	μ_1	0.936	0.930	0.932	0.930	0.4007	0.3887	0.3907	0.3885
			μ_2	0.924	0.912	0.880	0.918	0.4404	0.4072	0.4051	0.4270
		200	μ_1	0.944	0.936	0.942	0.936	0.2497	0.2462	0.2495	0.2466
			μ_2	0.906	**0.880**	**0.868**	0.902	0.2736	0.2560	0.2528	0.2691

选变量而言, 变得不那么重要了, 而估计系数会带来额外的方差, 因此全模型的优势并没有体现出来. 当我们增大 c 的取值 (如取 $c=5$ 和 $c=10$) 时, 结果确实表明全模型方法是最优的 (限于篇幅, 具体数值结果没有列出). 整体来说, 模型平均方法是最优的, 模型选择方法次之, 而当全模型是真实模型时, 全模型方法也会表现得好一些, 甚至是最好的.

(3) 对于两种模型平均方法之间以及两种模型选择方法之间的比较, 从表 1 可以看出, 大部分情况下, SBIC 比 SAIC 的 MSE 要小, 而 BIC 比 AIC 的 MSE 要小.

(4) 由于 SAIC 和 SBIC 方法的覆盖率和平均长度的计算结果在保留四位小数之后是一样的, 所以表 2 和 3 将两者进行了合并, 而表中黑体数字为在给定的置信水平下各个方法的最小覆盖率. 由表 2 知, 对于 90% 的置信水平, AIC 和 BIC 方法的最小覆盖率分别为 0.816 和 0.796, 与给定的置信水平差

表 3　5 种方法置信区间的覆盖率和平均长度, $\sigma = 2$

				CP				Length			
置信水平	c	n	μ	SAIC (SBIC)	AIC	BIC	Full	SAIC (SBIC)	AIC	BIC	Full
90%	0	100	μ_1	0.892	0.882	0.882	0.886	0.9068	0.8748	0.8736	0.8797
			μ_2	0.844	0.826	0.814	0.828	1.0064	0.9458	0.9289	0.9760
		200	μ_1	0.892	0.876	0.874	0.878	0.6249	0.6141	0.6136	0.6157
			μ_2	0.864	0.848	0.842	0.854	0.6881	0.6531	0.6420	0.6779
	1	100	μ_1	0.890	0.880	0.884	0.882	0.9080	0.8749	0.8749	0.8809
			μ_2	**0.836**	**0.810**	**0.798**	**0.828**	1.0086	0.9455	0.9334	0.9781
		200	μ_1	0.902	0.898	0.898	0.900	0.6221	0.6114	0.6112	0.6130
			μ_2	0.878	0.858	0.852	0.874	0.6860	0.6543	0.6449	0.6758
95%	0	100	μ_1	0.952	0.946	0.944	0.948	1.0875	1.0495	1.0473	1.0549
			μ_2	0.934	0.926	0.924	0.932	1.1914	1.1147	1.0962	1.1554
		200	μ_1	0.918	0.918	0.916	0.918	0.7396	0.7268	0.7261	0.7288
			μ_2	0.944	0.932	0.926	0.936	0.8151	0.7783	0.7636	0.8029
	1	100	μ_1	0.946	0.944	0.944	0.946	1.0926	1.0558	1.0545	1.0611
			μ_2	**0.914**	**0.898**	**0.892**	**0.904**	1.2040	1.1316	1.1141	1.1688
		200	μ_1	0.946	0.944	0.940	0.944	0.7472	0.7346	0.7343	0.7363
			μ_2	0.934	0.918	0.918	0.930	0.8205	0.7852	0.7713	0.8083

距较大, 而 SAIC 方法和全模型方法的最小覆盖率分别为 0.842 和 0.826, 比 AIC 和 BIC 的覆盖率明显高; 对于给定的 95% 的置信水平, AIC 和 BIC 方法的最小覆盖率分别为 0.88 和 0.868, 而 SAIC 方法和全模型方法的最小覆盖率分别为 0.9 和 0.896, 也比 AIC 和 BIC 的覆盖率高. 可以看出, 在所有情况下, SAIC 比 AIC 和 BIC 的覆盖率都更接近于给定的置信水平. 从表 3 中可以得到同样的结论.

(5) 随着样本量 n 从 100 增加到 200, 估计效果都变得更好. 另外, 表 1–3 是非参数函数

$$\theta(x) = \sin(8x - 2)$$

的结果, 对其他三种非参数函数, 结果相似, 因此没有罗列出.

6　实际数据分析

本节用模型平均方法研究波士顿郊区房价数据集, 该数据集来自于机器学习资源库 [29]. 数据集包含了 506 个自有房屋的价格 (P) 和 13 个可能的影响因素, 分别为每个房子包含的房间数量 (RM)、距离五个波士顿商务区的加权距离 (DIS)、附近的向外辐射的道路指数 (RAD)、不动产税 (TAX)、地位较低的人口指标 (LSTAT)、人均犯罪率 (CRIM)、查尔斯河畔 (CHAS)、一氧化氮浓度 (NOX)、1940 年之前所建的房屋比例 (AGE)、小学生和老师的比例 (PTRATIO)、超过 25,000 平方英尺的住宅用地的比例 (ZN)、每英亩非零售业务用地比例 (INDUS) 和住所周围的黑种人比例 (B). 文献 [30] 对该数据集进行了研究, 基于他们的研究, 我们先对原始数据集进行处理: 对 RM 取平方, 对 P、DIS、RAD 和 LSTAT 取对数, 分别记为 RM2、$\log P$、logDIS、logRAD 和 logLSTAT, 并代替原来相应的变量. 同时他们的研究表明 B 对 $\log P$ 的影响是有非线性趋势的, 而其他变量对 $\log P$ 的影响是线性的, 所以,

我们以 B 作为半参数模型中的非参数协变量, 这样, 我们考虑的模型为

$$\begin{aligned}\log P = {}& \mathrm{RM2}\beta_1 + \mathrm{logDIS}\beta_2 + \mathrm{logRAD}\beta_3 + \mathrm{TAX}\beta_4 + \mathrm{logLSTAT}\beta_5 + \mathrm{CRIM}\beta_6 + \mathrm{CHAS}\beta_7 + \mathrm{NOX}\beta_8 \\ & + \mathrm{AGE}\beta_9 + \mathrm{PTRATIO}\beta_{10} + \mathrm{ZN}\beta_{11} + \mathrm{INDUS}\beta_{12} + \theta(B) + \epsilon. \end{aligned} \tag{6.1}$$

当 $\theta(B) = \mathrm{constant}_1 + \mathrm{constant}_2(B - 0.63)^2$ 时, 该模型与文献 [30, (A.1)] 一致.

为了实施模型平均方法, 我们选定 AGE、PTRATIO、ZN 和 INDUS 这 4 个线性协变量为候选变量, 其余 8 个线性协变量和非参数协变量为必选变量, 包含在各个候选模型中. 原因是 AIC 和 BIC 这两种模型选择方法都排除了 AGE、PTRATIO、ZN 和 INDUS. 另外, 当我们通过求条件期望消除模型 (6.1) 中的非线性项而作线性回归时, 发现 AGE、PTRATIO、ZN 和 INDUS 这 4 个变量都是不显著的. 从直观上来看, 我们也认为这 4 个变量对房价的影响较弱, 因为 1940 年之前所建的房屋比例 (AGE) 体现的是住房的年限, 如果住房的地段好, 环境好, 可能住房的年限对房价不会产生太大的影响; 小学生和老师的比例 (PTRATIO) 体现的是住房周边的教育资源, 这只对部分家庭 (有需要接受小学教育的孩子) 有影响, 而对一般的家庭影响不大; 超过 25,000 平方英尺的住宅用地的比例 (ZN) 体现的是有较大的住房面积比例, 并不是每个单个住房面积的大小, 因此对房价的影响可能也不大; 每英亩非零售业务用地比例 (INDUS) 主要体现的是住房周边的工厂情况, 有的人会认为工厂附近环境不太好, 但是也会有在工厂工作的人认为工作地点离家近很方便, 所以这个影响因素也是视情况而定. 总之, 由于这些变量是影响 $\log P$ 的潜在因素, 我们把它们设定为候选变量, 这样共有 $2^4 = 16$ 个不同的候选模型.

我们分别用 SAIC、SBIC、AIC、BIC 方法和全模型方法 (Full) 计算每个协变量的回归系数及其相应的置信水平为 90% 和 95% 的置信区间, 结果见表 4, 其中方括号记录的是置信区间. 由于置信水平不影响回归系数的估计, 因此, 置信水平为 95% 时回归系数的估计值没有重复列出.

从表 4 可以看出, 模型平均方法和全模型方法估计的系数的方向是一致的, 其中 RM2、logRAD、CHAS 和 ZN 对 $\log P$ 有正的影响, 而其余 8 个线性协变量对 $\log P$ 有负的影响. AIC 和 BIC 这两种模型选择方法都排除了 AGE、PTRATIO、ZN 和 INDUS 这 4 个协变量 (由于两种方法选中的模型是一样的, 我们对计算结果进行了合并), 但对其他 8 个系数的估计也有与模型平均方法和全模型方法相同的方向. 事实上, 每个房子的房间数量的平方 (RM2) 大, 说明相应的居住面积比较大; logRAD 大, 表明房屋附近的道路比较多, 交通便利; CHAS 是一个哑变量, 值为 1 表示在查尔斯河畔附近, 环境比较好; 超过 25,000 平方英尺的住宅用地的比例 (ZN) 越高, 大房子的比例就越高. 以上这 4 个变量取值大时, 是有利因素, 因而对房价有正面影响. 而距离 5 个波士顿商务区的加权距离 (logDIS) 远, 居民就需花费更多的时间在上下班的路上; 不动产税 (TAX) 高, 人们购买的欲望就会减小; 地位较低的人口指标 (logLSTAT) 大, 说明周围居民普遍的收入水平不够高; 人均犯罪率 (CRIM) 高, 说明居住区周围不够安全; 一氧化氮浓度 (NOX) 高, 会影响居民的身体健康; 1940 年之前所建的房屋比例 (AGE) 高, 说明房屋老旧; 小学生和老师的比例 (RATIO) 高, 说明老师相对于小学生来说比较少, 教育资源比较匮乏; 每英亩非零售业务用地比例 (INDUS) 大, 说明周围工厂较多, 会有潜在的环境污染和噪声污染等. 以上 8 个变量取值大时, 是不利因素, 因而对房价有负面影响.

也可以看到, 对 AGE、PTRATIO、ZN 和 INDUS 这 4 个协变量, 模型选择方法是进行了完全地排除, 全模型方法则给予了较大的重视, 而模型平均方法给出的是一个折衷的结果, 它说明了 AGE、PTRATIO、ZN 和 INDUS 对房价是有影响的, 但影响程度没有全模型方法声明的那样高.

表中的 SAIC、SBIC 和全模型方法 (Full) 下的置信区间在保留三位小数之后就相同了, 这与文献中的结果是相吻合的: 基于文献 [14] 中的方法构建的置信区间 (见 (4.7)) 和全模型下的置信区间是渐近等价的, 参见文献 [22,31].

表 4　波士顿郊区房价数据集的估计及置信区间 ($\times 10^2$)

置信水平	变量名	SAIC	SBIC	AIC (BIC)	Full
90%	RM2	0.505 [−0.883, 1.828]	0.513 [−0.883, 1.828]	0.515 [−0.738, 1.768]	0.472 [−0.883, 1.828]
	logDIS	−13.622 [−82.967, 47.722]	−12.347 [−82.967, 47.722]	−12.104 [−58.194, 33.986]	−17.623 [−82.967, 47.722]
	logRAD	7.300 [−36.180, 51.809]	7.159 [−36.180, 51.809]	7.135 [−34.826, 49.095]	7.815 [−36.180, 51.809]
	TAX	−0.014 [−0.122, 0.096]	−0.014 [−0.122, 0.096]	−0.014 [−0.111, 0.082]	−0.013 [−0.122, 0.096]
	logLSTAT	−32.952 [−88.582, 25.127]	−33.355 [−88.582, 25.127]	−33.433 [−84.281, 17.414]	−31.727 [−88.582, 25.127]
	CRIM	−1.151 [−8.256, 5.821]	−1.132 [−8.256, 5.821]	−1.128 [−8.108, 5.852]	−1.217 [−8.256, 5.821]
	CHAS	3.893 [−25.606, 34.046]	3.808 [−25.606, 34.046]	3.793 [−25.788, 33.373]	4.220 [−25.606, 34.046]
	NOX	−5.953 [−38.144, 30.355]	−6.393 [−38.144, 30.355]	−6.462 [−19.187, 6.263]	−3.894 [−38.144, 30.355]
	AGE	−0.004 [−0.482, 0.452]	−0.001 [−0.482, 0.452]	0(-)	−0.015 [−0.482, 0.452]
	PTRATIO	−0.012 [−0.563, 0.455]	−0.002 [−0.563, 0.455]	0(-)	−0.054 [−0.563, 0.455]
	ZN	0.010 [−0.412, 0.479]	0.002 [−0.412, 0.479]	0(-)	0.034 [−0.412, 0.479]
	INDUS	−0.051 [−2.353, 1.974]	−0.008 [−2.353, 1.974]	0(-)	−0.189 [−2.353, 1.974]
95%	RM2	[−1.143, 2.088]	[−1.143, 2.088]	[−0.978, 2.008]	[−1.143, 2.088]
	logDIS	[−95.485, 60.240]	[−95.485, 60.240]	[−67.024, 42.816]	[−95.485, 60.240]
	logRAD	[−44.608, 60.237]	[−44.608, 60.237]	[−42.865, 57.134]	[−44.608, 60.237]
	TAX	[−0.143, 0.117]	[−0.143, 0.117]	[−0.129, 0.100]	[−0.143, 0.117]
	logLSTAT	[−99.474, 36.019]	[−99.474, 36.019]	[−94.022, 27.155]	[−99.474, 36.019]
	CRIM	[−9.605, 7.170]	[−9.605, 7.170]	[−9.445, 7.189]	[−9.605, 7.170]
	CHAS	[−31.320, 39.760]	[−31.320, 39.760]	[−31.455, 39.040]	[−31.320, 39.760]
	NOX	[−44.705, 36.917]	[−44.705, 36.917]	[−21.625, 8.700]	[−44.705, 36.917]
	AGE	[−0.571, 0.541]	[−0.571, 0.541]	0(-)	[−0.571, 0.541]
	PTRATIO	[−0.661, 0.553]	[−0.661, 0.553]	0(-)	[−0.661, 0.553]
	ZN	[−0.497, 0.565]	[−0.497, 0.565]	0(-)	[−0.497, 0.565]
	INDUS	[−2.768, 2.389]	[−2.768, 2.389]	0(-)	[−2.768, 2.389]

AIC 和 BIC 的计算结果在保留三位小数之后是一样的, 因此合并到一起; 0(-) 表示该变量没有被 AIC (BIC) 选中.

7 总结

本文在半参数模型下研究了模型平均估计的渐近分布理论, 与文献 [16] 对该模型采用的研究方法不同, 我们通过对真实模型在零模型处进行 Taylor 展开而获得了模型平均估计量的渐近分布, 也构造了一个覆盖真实参数的概率趋于给定的置信水平的置信区间. 虽然我们的渐近分布与文献 [16] 中所得到的渐近分布在协方差矩阵上不相同, 但是因为在局部误设定的情况下, 相应的参数只是相差 $O(1/\sqrt{n})$, 所以, 两者的渐近分布的协方差矩阵其实是渐近等价的. 我们也进行了许多模拟计算和实际数据分析, 结果均表明, 模型平均方法较模型选择方法和全模型方法具有较大的优势.

值得注意的是, 本文研究的模型平均估计的渐近理论是基于一般的半参数模型展开的, 这包含了许多常见的半参数模型, 如部分线性模型等, 因此, 本文的渐近理论的适用范围还是比较广泛的. 另一方面, 本文基于文献 [14] 的局部误设定框架推导渐近分布, 该框架要求所有候选模型在样本量较大时都比较接近, 这是使用本文结果需要注意的前提条件. 如何去除局部误设定假设而在一般情形下推导模型平均估计量的渐近分布, 是一个尚未解决的问题, 即使是对相对简单的参数模型.

近年来, 关于模型平均估计的渐近分布已有大量研究, 但基本都是在权重形式已知的情形下进行的. 如果权重是按某种准则最优选取的, 则它一般没有显式表达, 此时关于模型平均估计的渐近理论是我们未来的一个研究方向. 相关的一项研究可参见文献 [32]. 另一方面, 对于更复杂的数据类型 (如函数型数据等) 如何进行模型平均以及模型平均估计渐近分布的推导也是值得进一步研究的课题.

致谢 感谢编委和两位审稿人的细致阅读以及他们中肯的建议.

参考文献

1 Akaike H. Information theory and an extension of the maximum likelihood principle. In: Proceedings of the Second International Symposium on Information Theory. Budapest: Akademiai Kaido, 1973, 267–281
2 Schwarz G. Estimating the dimension of a model. Ann Statist, 1978, 6: 461–464
3 Mallows C L. Some comments on C_p. Technometrics, 1973, 15: 661–675
4 Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B Stat Methodol, 1974, 36: 111–147
5 Hoeting J A, Madigan D, Raftery A E, et al. Bayesian model averaging: A tutorial. Statist Sci, 1999, 14: 382–417
6 Koop G. Bayesian Econometrics. Chichester: John Wiley & Sons, 2003, 265–282
7 Buckland S T, Burnham K P, Augustin N H. Model selection: An integral part of inference. Biometrics, 1997, 53: 603–618
8 Hansen B E. Least squares model averaging. Econometrica, 2007, 75: 1175–1189
9 Wan A T K, Zhang X, Zou G. Least squares model averaging by Mallows criterion. J Econometrics, 2010, 156: 277–283
10 Hansen B E, Racine J S. Jackknife model averaging. J Econometrics, 2012, 167: 38–46
11 Zhang X, Wan A T K, Zou G. Model averaging by jackknife criterion in models with dependent data. J Econometrics, 2013, 174: 82–94
12 Liang H, Zou G, Wan A T K, et al. Optimal weight choice for frequentist model average estimators. J Amer Statist Assoc, 2011, 106: 1053–1066
13 Gao Y, Zhang X, Wang S, et al. Model averaging based on leave-subject-out cross-validation. J Econometrics, 2016, 192: 139–151
14 Hjort N L, Claeskens G. Frequentist model average estimators. J Amer Statist Assoc, 2003, 98: 879–899
15 Hjort N L, Claeskens G. Focused information criteria and model averaging for the Cox hazard regression model. J Amer Statist Assoc, 2006, 101: 1449–1464
16 Claeskens G, Carroll R J. An asymptotic theory for model selection inference in general semiparametric problems. Biometrika, 2007, 94: 249–265
17 Deng G H, Liang H. Model averaging for semiparametric additive partial linear models. Sci China Math, 2010, 53: 1363–1376
18 Zhang X, Liang H. Focused information criterion and model averaging for generalized additive partial linear models. Ann Statist, 2011, 39: 174–200

19 Yu Y, Thurston S W, Hauser R, et al. Model averaging procedure for partially linear single-index models. J Statist Plann Inference, 2013, 143: 2160–2170
20 Yang H, Lin P, Zou G, et al. Variable selection and model averaging for longitudinal data incorporating GEE approach. Statist Sinica, 2017, 27: 389–413
21 王海鹰, 邹国华. 线性测量误差模型的平均估计. 系统科学与数学, 2012, 32: 1–14
22 Wang H, Zou G, Wan A T K. Model averaging for varying-coefficient partially linear measurement error models. Electron J Stat, 2012, 6: 1017–1039
23 Wang H Y, Chen X, Flournoy N. The focused information criterion for varying-coefficient partially linear measurement error models. Statist Papers, 2016, 57: 99–113
24 孙志猛, 马景义, 苏治. 响应变量删失情况下线性模型的 FIC 模型选择和模型平均. 中国科学: 数学, 2013, 43: 647–661
25 Chen X, Zou G, Zhang X. Frequentist model averaging for linear mixed-effects models. Front Math China, 2013, 8: 497–515
26 Du J, Zhang Z, Xie T. Focused information criterion and model averaging in quantile regression. Comm Statist Theory Methods, 2013, 42: 3716–3734
27 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
28 孙志猛. 删失分位数回归模型基于扩展兴趣信息准则的平均估计. 中国科学: 数学, 2014, 44: 857–874
29 Lichman M. UCI machine learning repository. Irvine: University of California, School of Information and Computer Science, 2013, http://archive.ics.uci.edu/ml
30 Harrison D Jr, Rubinfeld D L. Hedonic housing prices and the demand for clean air. J Environ Economics Manage, 1978, 5: 81–102
31 Wang H, Zhou S Z F. Interval estimation by frequentist model averaging. Comm Statist Theory Methods, 2013, 42: 4342–4356
32 Liu C A. Distribution theory of the least squares averaging estimator. J Econometrics, 2015, 186: 142–159
33 Severini T A, Wong W H. Profile likelihood and conditionally parametric models. Ann Statist, 1992, 20: 1768–1802
34 Jain N C, Marcus M B. Central limit theorems for $C(S)$-valued random variables. J Funct Anal, 1975, 19: 216–231
35 Lehmann E L, Casella G. Theory of Point Estimation, 2nd ed. New York: Springer-Verlag, 1998, 447–448

附录 A　条件

条件 1　(1) 对任意的 $r,s=0,1,\ldots,3$, $r+s\leqslant 3$, $\frac{\partial^{r+s}}{\partial\beta^r\partial\theta^s}l(y,z,\beta,\theta(x))$ 是存在的, 且每一个分量都被 $f_{\mathrm{true}}(y)$ 下有有限均值的函数控制, 即存在一个非负的 $M(x,y,z)$, 使得

$$\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left\|\frac{\partial^{r+s}}{\partial\beta^r\partial\theta^s}l(y,z,\beta,\theta(x))\right\|\leqslant M(x,y,z), \tag{A.1}$$

其中 $\mathcal{N}(\beta_0)$ 表示 β_0 的邻域, Θ 是 $\mathbb{R}$ 上的紧子集, $\|A\|=\max_{i,j}|a_{ij}|$, a_{ij} 是矩阵 A 的 (i,j) 元素, $\mathrm{E}(M(x,y,z))<\infty$, $\mathrm{E}(\cdot)$ 表示在真实密度下的期望.

(2) 对任意的 $p,q=0,1,2$, $p+q\leqslant 2$, 有

$$\mathrm{E}_0\left(\left\|\frac{\partial^{p+q}}{\partial\beta^p\partial\theta^q}l(y,z,\beta,\theta_0(x,\beta))\,\bigg|_{\beta=\beta_0}\right\|_2^2\right)<\infty \quad 且 \quad \mathrm{E}_0\left(\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{d^3}{d\gamma^3}l(y,z,\beta,\theta_{\mathrm{true}}(x))\right|\right)<\infty, \tag{A.2}$$

其中 $\mathrm{E}_0(\cdot)$ 表示在零模型的密度下的期望, $\|\cdot\|_2$ 表示矩阵或者向量的 2 范数.

条件 2　对任意的 $r,s=0,1,2,3$, $r+s\leqslant 3$, $\frac{\partial^{r+s}}{\partial x^r\partial\beta^s}\theta_0(x,\beta)$ 是存在的, 且

$$\mathrm{E}\left(\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left\|\frac{\partial^{r+s}}{\partial x^r\partial\beta^s}\hat{\theta}(x,\beta)\right\|\right)<\infty.$$

设

$$\sup_{x\in\mathcal{B}}|(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\,|_{\beta=\beta_0}|=o_p(n^{-\alpha_1}),\quad \sup_{x\in\mathcal{B}}\|(\hat{\theta}'(x,\beta)-\theta_0'(x,\beta))\,|_{\beta=\beta_0}\|=o_p(n^{-\alpha_2}), \tag{A.3}$$

其中 x 的定义域 $\mathcal{B}$ 是有界闭区间, $\alpha_1+\alpha_2\geqslant 1/2$, $\alpha_1\geqslant 1/4$, 且存在 $\delta>0$ 满足 $\delta\leqslant\alpha_1$, $\delta\leqslant\alpha_2$, 使得

$$\begin{aligned}&\sup_{x\in\mathcal{B}}\left|\left(\frac{\partial}{\partial x}\hat{\theta}(x,\beta)-\frac{\partial}{\partial x}\theta_0(x,\beta)\right)\Big|_{\beta=\beta_0}\right|=o_p(n^{-\delta}),\\&\sup_{x\in\mathcal{B}}\left\|\left(\frac{\partial}{\partial x}\hat{\theta}'(x,\beta)-\frac{\partial}{\partial x}\theta_0'(x,\beta)\right)\Big|_{\beta=\beta_0}\right\|=o_p(n^{-\delta}),\end{aligned}\tag{A.4}$$

且

$$\begin{aligned}&\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{x\in\mathcal{B}}|\hat{\theta}(x,\beta)-\theta_0(x,\beta)|=o_p(1),\\&\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{x\in\mathcal{B}}\|\hat{\theta}'(x,\beta)-\theta_0'(x,\beta)\|=o_p(1),\\&\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{x\in\mathcal{B}}\|\hat{\theta}''(x,\beta)-\theta_0''(x,\beta)\|=o_p(1),\end{aligned}\tag{A.5}$$

其中 $\theta'(x,\beta)=\frac{d}{d\beta}\theta(x,\beta)$, $\theta''(x,\beta)=\frac{d^2}{d\beta d\beta^{\mathrm{T}}}\theta(x,\beta)$.

条件 3 由密度函数 f 关于参数 β 的一阶导数是连续的, 可以得到

$$f_{\mathrm{true}}(y)=f_0(y)\left[1+v_0(y)^{\mathrm{T}}\frac{\delta}{\sqrt{n}}+R_2\left(y,\frac{\delta}{\sqrt{n}}\right)\right],\tag{A.6}$$

其中 $f_{\mathrm{true}}(y)=f(y,z,\beta_{\mathrm{true}},\theta(x,\beta_{\mathrm{true}}))=f(y,z,\beta_{\mathrm{true}},\theta_{\mathrm{true}}(x))$ 表示真实模型的密度函数, $f_0(y)=f(y,z,\beta_0,\theta_0(x,\beta_0))=f(y,z,\beta_0,\theta_{\mathrm{true}}(x))$ 表示零模型的密度函数, 且 $R_2(y,\frac{\delta}{\sqrt{n}})$ 满足

(1)

$$\mathrm{E}_0\left[u_0(y)R_2\left(y,\frac{\delta}{\sqrt{n}}\right)\right]=o\left(\frac{1}{\sqrt{n}}\right),\quad \mathrm{E}_0\left[v_0(y)R_2\left(y,\frac{\delta}{\sqrt{n}}\right)\right]=o\left(\frac{1}{\sqrt{n}}\right);\tag{A.7}$$

(2) 对于任意的第 i、j 和 k 个分量, 有

$$\mathrm{E}_0(u_{0(i)}u_{0(j)}v_{0(k)})=O(1),\quad \mathrm{E}_0(u_{0(i)}v_{0(j)}v_{0(k)})=O(1),\quad \mathrm{E}_0(v_{0(i)}v_{0(j)}v_{0(k)})=O(1);\tag{A.8}$$

(3)

$$\mathrm{E}_0\left(u_{0(i)}u_{0(j)}R_2\left(y,\frac{\delta}{\sqrt{n}}\right)\right)=o(1),\quad \mathrm{E}_0\left(u_{0(i)}v_{0(j)}R_2\left(y,\frac{\delta}{\sqrt{n}}\right)\right)=o(1),\tag{A.9}$$

$$\mathrm{E}_0\left(v_{0(i)}v_{0(j)}R_2\left(y,\frac{\delta}{\sqrt{n}}\right)\right)=o(1).\tag{A.10}$$

条件 1(1) 是对文献 [33] 中的条件 S 的改进, 使得对数似然函数的偏导数在真实模型下, 被具有有限均值的函数一致地界住. 条件 1(2) 类似文献 [33] 中的条件 S, 是对零模型下的对数似然函数的偏导数作了假设. 条件 2 是对文献 [33] 中条件 NP 的改进, 将其从真实的参数 β_{true} 扩展到零模型对应的参数 β_0 处. 由文献 [16, 引理 A1] 可知, $\hat{\theta}(x,\beta)-\theta(x,\beta)\xrightarrow{p}0$. 由于 β_{true} 是与样本量 n 有关的, 因此, 在真实模型下的 $\theta(x,\beta)$ 也是与 n 有关的, 我们假设它的极限存在. 下面说明在一定的正则条件下, 其极限为 $\theta_0(x,\beta)$. 由于 $\theta(x,\beta)=\arg\max_\theta \mathrm{E}[l(y,z,\beta,\theta(x,\beta))\mid x]$, 在一定的正则条件下, $\theta(x,\beta)$ 即是

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}l(y,z,\beta,\theta(x,\beta))\;\Big|\;x\right)=0$$

的解, 也即 $0=\int \frac{\partial}{\partial\theta} l(y,z,\beta,\theta(x,\beta))f(y,z,\beta_{\text{true}},\theta_{\text{true}}(x)\mid x)dydz$. 我们对等式两边取极限得

$$\begin{aligned}0 &= \lim_{n\to\infty}\int \frac{\partial}{\partial\theta} l(y,z,\beta,\theta(x,\beta))f(y,z,\beta_{\text{true}},\theta_{\text{true}}(x)\mid x)dydz\\ &= \int \frac{\partial}{\partial\theta} l\Big(y,z,\beta,\lim_{n\to\infty}\theta(x,\beta)\Big)f\Big(y,z,\lim_{n\to\infty}\beta_{\text{true}},\theta_{\text{true}}(x)\,\Big|\, x\Big)dydz\\ &= \int \frac{\partial}{\partial\theta} l\Big(y,z,\beta,\lim_{n\to\infty}\theta(x,\beta)\Big)f(y,z,\beta_0,\theta_{\text{true}}(x)\mid x)dydz\\ &= \mathrm{E}_0\bigg(\frac{\partial}{\partial\theta} l\Big(y,z,\beta,\lim_{n\to\infty}\theta(x,\beta)\Big)\,\bigg|\, x\bigg).\end{aligned}$$

因此, 在一定的正则条件下, $\lim_{n\to\infty}\theta(x,\beta)$ 将最大化 $\mathrm{E}_0[l(y,z,\beta,\theta(x,\beta))\mid x]$, 即有

$$\lim_{n\to\infty}\theta(x,\beta)=\theta_0(x,\beta).$$

所以, 将文献 [33] 中的条件 NP 从真实模型下的 $\theta(x,\beta)$ 扩展到零模型下的 $\theta_0(x,\beta)$ 也是合理的. 条件 3 是对文献 [14, 条件 (C1)–(C3)] 的改进, 将其从参数模型拓展到半参数模型.

附录 B 引理和定理的证明

引理 1 的证明 由 Taylor 展开可得

$$\begin{aligned}&\begin{pmatrix}\sqrt{n}\bar{u}_n\\ \sqrt{n}\bar{v}_n\end{pmatrix}-\begin{pmatrix}\sqrt{n}\bar{u}_0\\ \sqrt{n}\bar{v}_0\end{pmatrix}\\ &=\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{dl(y_i,z_i,\beta,\hat{\theta}(x_i,\beta))}{d\beta}\,\bigg|_{\beta=\beta_0}-\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{dl(y_i,z_i,\beta,\theta_0(x_i,\beta))}{d\beta}\,\bigg|_{\beta=\beta_0}\\ &=\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{d}{d\beta}[l(y_i,z_i,\beta,\hat{\theta}(x_i,\beta))-l(y_i,z_i,\beta,\theta_0(x_i,\beta))]\,\bigg|_{\beta=\beta_0}\\ &=\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{d}{d\beta}\bigg[\frac{\partial l(y_i,z_i,\beta,\theta_0(x_i,\beta))}{\partial\theta}(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))+r_{n,i}^{(2)}(\beta)\bigg]\,\bigg|_{\beta=\beta_0}\\ &=\frac{1}{\sqrt{n}}\sum_{i=1}^n \bigg[\frac{d}{d\beta}\frac{\partial}{\partial\theta} l(y_i,z_i,\beta,\theta_0(x_i,\beta))\bigg]\,\bigg|_{\beta=\beta_0}[\hat{\theta}(x_i,\beta_0)-\theta_0(x_i,\beta_0)]\\ &\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^n \bigg[\frac{\partial}{\partial\theta} l(y_i,z_i,\beta,\theta_0(x_i,\beta))\,\bigg|_{\beta=\beta_0}\bigg][(\hat{\theta}'(x_i,\beta)-\theta_0'(x_i,\beta))\mid_{\beta=\beta_0}]+\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{dr_{n,i}^{(2)}(\beta)}{d\beta}\,\bigg|_{\beta=\beta_0},\end{aligned}$$

其中

$$r_{n,i}^{(2)}(\beta)=\frac{1}{2}\int_0^1 \frac{\partial^2}{\partial\theta^2} l[y_i,z_i,\beta,\theta_0(x_i,\beta)+t(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))]dt[\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta)]^2.$$

下面分别证明上式中的三项均为 $o_p(1)$.

(i) 首先注意, 第一项中

$$\begin{aligned}&\frac{d}{d\beta}\frac{\partial}{\partial\theta} l(y_i,z_i,\beta,\theta_0(x_i,\beta))\,\bigg|_{\beta=\beta_0}\\ &\quad=\frac{\partial^2}{\partial\beta\partial\theta} l(y_i,z_i,\beta,\theta_0(x_i,\beta))\,\bigg|_{\beta=\beta_0}+\frac{\partial^2}{\partial\theta^2} l(y_i,z_i,\beta,\theta_0(x_i,\beta))\,\bigg|_{\beta=\beta_0}\theta_0'(x_i,\beta)\,\bigg|_{\beta=\beta_0}\end{aligned}$$

$$\doteq B(y,x).$$

类似于文献 [33], 考虑如下的函数空间:

$$\Lambda_0=\left\{h\in C^2_{\mathcal{B}}:\sup_{x\in\mathcal{B}}|h(x)|\leqslant 1,\sup_{x\in\mathcal{B}}\left|\frac{\partial}{\partial x}h(x)\right|\leqslant 1\right\},$$

其中 $C^2_{\mathcal{B}}$ 表示定义在 $\mathcal{B}$ 区间上的二阶连续可导的函数集合; $\mathcal{B}$ 是 x 的定义域, 为有界闭区间. 定义 Λ_0 上的一个度量 $\rho(h_1,h_2)=\sup_{x\in\mathcal{B}}|h_1(x)-h_2(x)|$. 由条件 2 知, 当 $n\to\infty$ 时, $\Pr(n^{\delta}(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\,|_{\beta=\beta_0}\in\Lambda_0)\to 1$.

显然, 在给定 β 的情况下, 对任意的一条曲线 $\theta_{10}(x,\beta)\in\Lambda_0$, 满足 $\theta_{10}(x,\beta_0)=\theta_0(x,\beta_0)$, 有

$$\mathrm{E}_0[l(y,z,\beta,\theta_0(x,\beta))\mid x]\geqslant \mathrm{E}_0[l(y,z,\beta,\theta_{10}(x,\beta))\mid x],$$

且在 β_0 处,

$$\mathrm{E}_0[l(y,z,\beta_0,\theta_0(x,\beta_0))\mid x]=\mathrm{E}_0[l(y,z,\beta_0,\theta_{10}(x,\beta_0))\mid x].$$

因为

$$\mathrm{E}_0\left[\frac{dl(y,z,\beta,\theta_0(x,\beta))}{d\beta}\,\bigg|_{\beta=\beta_0}\,\bigg|\,x\right]=0$$

和

$$\mathrm{E}_0\left[\frac{dl(y,z,\beta,\theta_{10}(x,\beta))}{d\beta}\,\bigg|_{\beta=\beta_0}\,\bigg|\,x\right]=0,$$

故将 $\mathrm{E}_0[l(y,z,\beta,\theta_0(x,\beta))\mid x]$ 和 $\mathrm{E}_0[l(y,z,\beta,\theta_{10}(x,\beta))\mid x]$ 在 β_0 处展开并相减, 可得

$$\frac{1}{2}(\beta-\beta_0)^{\mathrm{T}}\left(\mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_0(x,\beta))}{d\beta d\beta^{\mathrm{T}}}\,\bigg|_{\beta=\beta^*}\,\bigg|\,x\right]-\mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_{10}(x,\beta))}{d\beta d\beta^{\mathrm{T}}}\,\bigg|_{\beta=\beta^*}\,\bigg|\,x\right]\right)(\beta-\beta_0)\geqslant 0,$$

其中 β^* 介于 β_0 与 β 之间. 我们只考虑上式矩阵的对角元素对应的不等式. 记 $\beta_{(i)}$、$\beta_{0(i)}$ 和 $\beta^*_{(i)}$ 分别表示 β、β_0 和 β^* 的第 i 个分量, $\beta^{(i)}$ 是第 i 个元素为 $\beta_{(i)}$、其余 $p+q-1$ 个分量与 β_0 相同的向量, 则对任意的 $i=1,2,\ldots,p+q$, 有

$$\begin{aligned}
&(\beta^{(i)}-\beta_0)^{\mathrm{T}}\left(\mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_0(x,\beta))}{d\beta d\beta^{\mathrm{T}}}\,\bigg|_{\beta=\beta^*}\,\bigg|\,x\right]-\mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_{10}(x,\beta))}{d\beta d\beta^{\mathrm{T}}}\,\bigg|_{\beta=\beta^*}\,\bigg|\,x\right]\right)(\beta^{(i)}-\beta_0)\\
&\quad=\left(\mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_0(x,\beta))}{d\beta^2_{(i)}}\,\bigg|_{\beta=\beta^*}\,\bigg|\,x\right]-\mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_{10}(x,\beta))}{d\beta^2_{(i)}}\,\bigg|_{\beta=\beta^*}\,\bigg|\,x\right]\right)(\beta_{(i)}-\beta_{0(i)})^2\\
&\quad\geqslant 0.
\end{aligned}$$

由于对任意的 β, 都存在 β^* 使得上式成立, 因此, 成立

$$\mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_0(x,\beta))}{d\beta^2_{(i)}}\,\bigg|_{\beta=\beta_0}\,\bigg|\,x\right]\geqslant \mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_{10}(x,\beta))}{d\beta^2_{(i)}}\,\bigg|_{\beta=\beta_0}\,\bigg|\,x\right].$$

又因为

$$-\mathrm{E}_0\left[\frac{d^2l(y,z,\beta,\theta_0(x,\beta))}{d\beta^2_{(i)}}\,\bigg|_{\beta=\beta_0}\,\bigg|\,x\right]=\mathrm{E}_0\left\{\left[\frac{d}{d\beta_{(i)}}l(y,z,\beta,\theta_0(x,\beta))\right]^2\,\bigg|_{\beta=\beta_0}\,\bigg|\,x\right\},$$

所以得到

$$\mathrm{E}_0\left\{\left[\frac{d}{d\beta_{(i)}}l(y,z,\beta,\theta_0(x,\beta))\right]^2\,\Big|_{\beta=\beta_0}\,\Big|\,x\right\}\leqslant \mathrm{E}_0\left\{\left[\frac{d}{d\beta_{(i)}}l(y,z,\beta,\theta_{10}(x,\beta))\right]^2\,\Big|_{\beta=\beta_0}\,\Big|\,x\right\},$$

即有

$$\begin{aligned}&\mathrm{E}_0\left\{\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\beta_{(i)}}+\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}\frac{d\theta_0(x,\beta)}{d\beta_{(i)}}\right]^2\,\Big|_{\beta=\beta_0}\,\Big|\,x\right\}\\&\quad\leqslant \mathrm{E}_0\left\{\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\beta_{(i)}}+\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}\frac{d\theta_{10}(x,\beta)}{d\beta_{(i)}}\right]^2\,\Big|_{\beta=\beta_0}\,\Big|\,x\right\}.\end{aligned}$$

由 $\theta_{10}(x,\beta)$ 的任意性知, 对任意的函数 $h(x)\in\Lambda_0$, 有

$$\mathrm{E}_0\left\{\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\beta_{(i)}}+\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}\frac{d\theta_0(x,\beta)}{d\beta_{(i)}}\right]\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}h(x)\right]\,\Big|_{\beta=\beta_0}\,\Big|\,x\right\}=0.$$

由于上式对任意的 $i=1,2,\ldots,p+q$ 都成立, 故有

$$\mathrm{E}_0\left\{\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\beta}+\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}\frac{d\theta_0(x,\beta)}{d\beta}\right]\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}h(x)\right]\,\Big|_{\beta=\beta_0}\,\Big|\,x\right\}=0.$$

因此, 对任意的 $h(x)\in\Lambda_0$, 有

$$\begin{aligned}&\mathrm{E}_0(B(y,x)h(x))\\&=\mathrm{E}_0\left[\left(\frac{\partial^2}{\partial\beta\partial\theta}l(y,z,\beta,\theta_0(x,\beta))+\frac{\partial^2}{\partial\theta^2}l(y,z,\beta,\theta_0(x,\beta))\theta_0'(x,\beta)\right)\,\Big|_{\beta=\beta_0}h(x)\right]\\&=-\mathrm{E}_x\left(\mathrm{E}_0\left\{\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\beta}+\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}\theta_0'(x,\beta)\right]\left[\frac{\partial l(y,z,\beta,\theta_0(x,\beta))}{\partial\theta}h(x)\right]\,\Big|_{\beta=\beta_0}\,\Big|\,x\right\}\right)\\&=0,\end{aligned}$$

其中 $\mathrm{E}_x(\cdot)$ 表示对 x 的边际密度求期望. 而对任意的函数 $h_1(x),h_2(x)\in\Lambda_0$, 有

$$\|B(y,x)h_1(x)-B(y,x)h_2(x)\|_2=\|B(y,x)(h_1-h_2)(x)\|_2\leqslant\|B(y,x)\|_2\cdot\sup_{x\in\mathcal{B}}|h_1(x)-h_2(x)|,$$

其中 $\|\cdot\|_2$ 表示向量或者矩阵的 2 范数. 又由条件 1(2) 和 2 可以推出

$$\mathrm{E}_0(\|B(y,x)\|_2^2)<\infty.$$

现令 $H(\cdot,\Lambda_0)$ 表示 Λ_0 上的度量熵, 则有 $H(\varepsilon,\Lambda_0)\leqslant A_0\varepsilon^{-1}$, 其中 A_0 为常数 (参见文献 [33]). 文献 [34] 已经证明了, 如果 Γ 是一个度量空间, 其上的度量为 d, $C(\Gamma)$ 是定义在度量空间 Γ 上的连续函数构成的空间, Z 是 $C(\Gamma)$ 上的随机元, 满足 $\mathrm{E}_0(Z)=0$, 且 $|Z(s)-Z(t)|\leqslant Vd(s,t)$, 其中 $\mathrm{E}_0(V^2)<\infty$, 则当 $\int_0^1H^{1/2}(\epsilon,\Gamma)d\epsilon<\infty$ 时, 对于独立且与 Z 同分布的随机元 $Z_1,Z_2,\ldots,Z_n$, 有 $n^{-1/2}(Z_1+Z_2+\ldots+Z_n)$ 满足中心极限定理. 这样, 若视 $Z:h\to B(y,x)h(x)$ 为 $C(\Lambda_0)$ 中的元素, 则由文献 [34, 定理 1] 可知,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{d}{d\beta}\frac{\partial}{\partial\theta}l(y_i,z_i,\beta,\theta_0(x_i,\beta))\,\Big|_{\beta=\beta_0}h(x_i)$$

满足中心极限定理, 因而有

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\beta}\frac{\partial}{\partial\theta}l(y_i,z_i,\beta,\theta_0(x_i,\beta))\Big|_{\beta=\beta_0}n^{\delta}(\hat{\theta}(x_i,\beta_0)-\theta_0(x_i,\beta_0))=O_p(1),$$

于是,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\beta}\frac{\partial}{\partial\theta}l(y_i,z_i,\beta,\theta_0(x_i,\beta))\Big|_{\beta=\beta_0}(\hat{\theta}(x_i,\beta_0)-\theta_0(x_i,\beta_0))=O_p(n^{-\delta})=o_p(1).$$

值得注意的是, 上式是在零模型下证明的, 我们还需要证明在真实模型下该式仍然成立.

我们先考虑 $\log(\frac{f_{n,\text{true}}(y)}{f_{n,0}(y)})$ 在零模型下的渐近分布. 易见,

$$\begin{aligned}\log\left(\frac{f_{n,\text{true}}(y)}{f_{n,0}(y)}\right)&=\log\left(\frac{\prod_{i=1}^{n}f_{\text{true}}(y_i)}{\prod_{i=1}^{n}f_0(y_i)}\right)\\&=\sum_{i=1}^{n}[\log f(y_i,z_i,\beta_{\text{true}},\theta_{\text{true}}(x_i))-\log f(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))]\\&=\frac{\delta^{\mathrm{T}}}{\sqrt{n}}\sum_{i=1}^{n}\frac{dl(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma}+\frac{1}{2}\frac{\delta^{\mathrm{T}}}{\sqrt{n}}\sum_{i=1}^{n}\frac{d^2l(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma d\gamma^{\mathrm{T}}}\frac{\delta}{\sqrt{n}}\\&\quad+\frac{1}{6}\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{d^3l(y_i,z_i,\beta,\theta_{\text{true}}(x_i))}{d\gamma_{(j)}d\gamma_{(k)}d\gamma_{(l)}}\Big|_{\beta=\beta^*}\frac{\delta_{(j)}\delta_{(k)}\delta_{(l)}}{n^{3/2}}\\&\hat{=}\ \Delta_1+\Delta_2+\Delta_3,\end{aligned}$$

其中 β^* 介于 β_0 与 β_{true} 之间. 下面依次考虑上式的三项.

在 Δ_1 中, $\frac{dl(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma}$ 是独立同分布的, 其在零模型下的均值和方差分别为

$$\begin{aligned}\mathrm{E}_0\left(\frac{dl(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma}\right)&=\int\left(\frac{dl(y,z,\beta_0,\theta_{\text{true}}(x))}{d\gamma}\right)f_0(y)dydzdx\\&=\int\left(\frac{df_0(y)}{d\gamma}\right)dydzdx\\&=\frac{d}{d\gamma}\int f_0(y)dydzdx\\&=0\end{aligned}$$

和

$$\mathrm{Var}_0\left(\frac{dl(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma}\right)=\mathrm{E}_0\left[\left(\frac{dl(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma}\right)\left(\frac{dl(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma}\right)^{\mathrm{T}}\right]\hat{=}J_0,$$

故由中心极限定理得

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dl(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma}\xrightarrow{d}N(0,J_0),$$

因此, 在零模型下, 有

$$\Delta_1=\frac{\delta^{\mathrm{T}}}{\sqrt{n}}\sum_{i=1}^{n}\frac{dl(y_i,z_i,\beta_0,\theta_{\text{true}}(x_i))}{d\gamma}\xrightarrow{d}N(0,\delta^{\mathrm{T}}J_0\delta).$$

假定 $\int f_0(y)dydzdx$ 关于 γ 的导数在积分下可以二次求导, 则可得

$$J_0 = \mathrm{E}_0\bigg[\bigg(\frac{dl(y_i, z_i, \beta_0, \theta_{\mathrm{true}}(x_i))}{d\gamma}\bigg)\bigg(\frac{dl(y_i, z_i, \beta_0, \theta_{\mathrm{true}}(x_i))}{d\gamma}\bigg)^{\mathrm{T}}\bigg] = -\mathrm{E}_0\bigg(\frac{d^2 l(y_i, z_i, \beta_0, \theta_{\mathrm{true}}(x_i))}{d\gamma d\gamma^{\mathrm{T}}}\bigg),$$

再由大数定律可知, 在零模型下,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d^2 l(y_i, z_i, \beta_0, \theta_{\mathrm{true}}(x_i))}{d\gamma d\gamma^{\mathrm{T}}} \xrightarrow{p} \mathrm{E}_0\bigg(\frac{d^2 l(y_i, z_i, \beta_0, \theta_{\mathrm{true}}(x_i))}{d\gamma d\gamma^{\mathrm{T}}}\bigg) = -J_0,$$

因而, 在零模型下, 有

$$\Delta_2 = \frac{1}{2}\frac{\delta^{\mathrm{T}}}{\sqrt{n}}\sum_{i=1}^{n}\frac{d^2 l(y_i, z_i, \beta_0, \theta_{\mathrm{true}}(x_i))}{d\gamma d\gamma^{\mathrm{T}}}\frac{\delta}{\sqrt{n}} \xrightarrow{p} -\frac{1}{2}\delta^{\mathrm{T}} J_0 \delta.$$

进一步, 根据大数定律和条件 1(2) 知,

$$\frac{1}{n}\sum_{i=1}^{n}\bigg|\frac{d^3 l(y_i, z_i, \beta, \theta_{\mathrm{true}}(x_i))}{d\gamma_{(j)}d\gamma_{(k)}d\gamma_{(l)}}\bigg|_{\beta=\beta^*}\bigg| \leqslant \frac{1}{n}\sum_{i=1}^{n}\sup_{\beta\in\mathcal{N}(\beta_0)}\bigg|\frac{d^3 l(y_i, z_i, \beta, \theta_{\mathrm{true}}(x_i))}{d\gamma_{(j)}d\gamma_{(k)}d\gamma_{(l)}}\bigg| = O_p(1),$$

因此, 在零模型下, Δ_3 为 $o_p(1)$.

这样, 在零模型下有

$$\log\bigg(\frac{f_{n,\mathrm{true}}(y)}{f_{n,0}(y)}\bigg) \xrightarrow{d} V \sim N\bigg(-\frac{1}{2}\delta^{\mathrm{T}} J_0\delta, \delta^{\mathrm{T}} J_0\delta\bigg).$$

注意到 $\mathrm{E}(V) = -\mathrm{Var}(V)/2$, 故零模型下的分布与真实模型下的分布是相互邻近的, 因而, 由 Le Cam 引理知, 在真实模型下, 对第一项成立

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\beta}\frac{\partial}{\partial\theta}l(y_i, z_i, \beta, \theta_0(x_i, \beta))\bigg|_{\beta=\beta_0}(\hat{\theta}(x_i, \beta_0) - \theta_0(x_i, \beta_0)) = o_p(1).$$

(ii) 同理可得第二项

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bigg[\frac{\partial}{\partial\theta}l(y_i, z_i, \beta, \theta_0(x_i, \beta))\bigg|_{\beta=\beta_0}\bigg][\hat{\theta}'(x_i, \beta_0) - \theta_0'(x_i, \beta_0)] = o_p(1).$$

(iii) 下面考虑第三项 $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dr_{n,i}^{(2)}(\beta)}{d\beta}\mid_{\beta=\beta_0}$, 要证其为 $o_p(1)$, 只需证各个分量均为 $o_p(1)$ 即可. 我们考虑第 l ($l = 1, 2, \ldots, p+q$) 个分量.

记 $Q_{n,i}^{(2)}(\beta) = \int_0^1 \frac{\partial^2}{\partial\theta^2}l[y_i, z_i, \beta, \theta_0(x_i, \beta) + t(\hat{\theta}(x_i, \beta) - \theta_0(x_i, \beta))]dt$, 则

$$\begin{aligned}
&\bigg|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dr_{n,i}^{(2)}(\beta)}{d\beta_{(l)}}\bigg|_{\beta=\beta_0}\bigg| \\
&= \bigg|\frac{1}{2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\beta_{(l)}}[Q_{n,i}^{(2)}(\beta)(\hat{\theta}(x_i, \beta) - \theta_0(x_i, \beta))^2]\bigg|_{\beta=\beta_0}\bigg| \\
&\leqslant \bigg|\frac{1}{2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bigg[\frac{d}{d\beta_{(l)}}Q_{n,i}^{(2)}(\beta)\bigg|_{\beta=\beta_0}\bigg][\hat{\theta}(x_i, \beta_0) - \theta_0(x_i, \beta_0)]^2\bigg|
\end{aligned}$$

$$
\begin{aligned}
&+\left|\frac{1}{2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Q_{n,i}^{(2)}(\beta)\,\Big|_{\beta=\beta_0}\left[\frac{d}{d\beta_{(l)}}(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))^2\,\Big|_{\beta=\beta_0}\right]\right|\\
\leqslant&\ \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{d}{d\beta_{(l)}}Q_{n,i}^{(2)}(\beta)\,\Big|_{\beta=\beta_0}\right|\sqrt{n}[\hat{\theta}(x_i,\beta_0)-\theta_0(x_i,\beta_0)]^2\\
&+\frac{1}{n}\sum_{i=1}^{n}|Q_{n,i}^{(2)}(\beta)|_{\beta=\beta_0}|\left|\sqrt{n}(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))\frac{d}{d\beta_{(l)}}(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))\,\Big|_{\beta=\beta_0}\right|.
\end{aligned}
$$

先证明在真实模型下有

$$
\frac{1}{n}\sum_{i=1}^{n}\left|\frac{d^j}{d\beta_{(l)}^j}Q_{n,i}^{(2)}(\beta)\,\Big|_{\beta=\beta_0}\right|=O_p(1),\quad j=0,1.
$$

当 $j=0$ 时, 由条件 1(1) 可知,

$$
|Q_{n,i}^{(2)}(\beta)\,|_{\beta=\beta_0}|\leqslant\sup_{\theta\in\Theta}\left|\frac{\partial^2}{\partial\theta^2}l(y_i,z_i,\beta,\theta(x_i))\,\Big|_{\beta=\beta_0}\right|\leqslant M(x_i,y_i,z_i),
$$

故在真实模型下,

$$
\frac{1}{n}\sum_{i=1}^{n}|Q_{n,i}^{(2)}(\beta)\,|_{\beta=\beta_0}|\leqslant\frac{1}{n}\sum_{i=1}^{n}M(x_i,y_i,z_i)=O_p(1).
$$

当 $j=1$ 时, 由条件 1(1) 知,

$$
\begin{aligned}
&\left|\frac{d}{d\beta_{(l)}}Q_{n,i}^{(2)}(\beta)\,\Big|_{\beta=\beta_0}\right|\\
&=\left|\frac{d}{d\beta_{(l)}}\int_0^1\frac{\partial^2}{\partial\theta^2}l[y_i,z_i,\beta,\theta_0(x_i,\beta)+t(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))]dt\,\Big|_{\beta=\beta_0}\right|\\
&=\left|\int_0^1\frac{d}{d\beta_{(l)}}\frac{\partial^2}{\partial\theta^2}l[y_i,z_i,\beta,\theta_0(x_i,\beta)+t(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))]dt\,\Big|_{\beta=\beta_0}\right|\\
&=\left|\int_0^1\left\{\frac{\partial^3}{\partial\beta_{(l)}\partial\theta^2}l[y_i,z_i,\beta,\theta_0(x_i,\beta)+t(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))]\right.\right.\\
&\quad\left.\left.+\frac{\partial^3}{\partial\theta^3}l[y_i,z_i,\beta,\theta_0(x_i,\beta)+t(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))]\frac{d}{d\beta_{(l)}}(\theta_0(x_i,\beta)+t(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta)))\right\}dt\,\Big|_{\beta=\beta_0}\right|\\
&\leqslant\sup_{\theta\in\Theta}\left|\frac{\partial^3}{\partial\beta_{(l)}\partial\theta^2}l[y_i,z_i,\beta,\theta(x_i)]\,\Big|_{\beta=\beta_0}\right|\\
&\quad+\sup_{\theta\in\Theta}\left|\frac{\partial^3}{\partial\theta^3}l[y_i,z_i,\beta,\theta(x_i)]\,\Big|_{\beta=\beta_0}\right|\left|\sup_{x\in\mathcal{B}}\left|\frac{d}{d\beta_{(l)}}\theta_0(x,\beta)\,\Big|_{\beta=\beta_0}\right|\right.\\
&\quad+\sup_{\theta\in\Theta}\left|\frac{\partial^3}{\partial\theta^3}l[y_i,z_i,\beta,\theta(x_i)]\,\Big|_{\beta=\beta_0}\right|\left|\frac{1}{2}\sup_{x\in\mathcal{B}}\left|\frac{d}{d\beta_{(l)}}(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\,\Big|_{\beta=\beta_0}\right|\right.\\
&\leqslant M(x_i,y_i,z_i)+M(x_i,y_i,z_i)\sup_{x\in\mathcal{B}}\left|\frac{d}{d\beta_{(l)}}\theta_0(x,\beta)\,\Big|_{\beta=\beta_0}\right|\\
&\quad+\frac{1}{2}M(x_i,y_i,z_i)\sup_{x\in\mathcal{B}}\left|\frac{d}{d\beta_{(l)}}(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\,\Big|_{\beta=\beta_0}\right|.
\end{aligned}
$$

故由条件 1(1) 和 2 可得, 在真实模型下,

$$
\frac{1}{n}\sum_{i=1}^{n}\left|\frac{d}{d\beta_{(l)}}Q_{n,i}^{(2)}(\beta)\,\Big|_{\beta=\beta_0}\right|=O_p(1).
$$

又由条件 2 可知, 在真实模型下,

$$\sup_{x\in\mathcal{B}}\sqrt{n}[\hat\theta(x,\beta_0)-\theta_0(x,\beta_0)]^2=\sqrt{n}o_p(n^{-2\alpha_1})=o_p(1),$$

且

$$\sup_{x\in\mathcal{B}}\sqrt{n}(\hat\theta(x,\beta)-\theta_0(x,\beta))\frac{d}{d\beta_{(l)}}(\hat\theta(x,\beta)-\theta_0(x,\beta))\bigg|_{\beta=\beta_0}=\sqrt{n}o_p(n^{-\alpha_1})o_p(n^{-\alpha_2})=o_p(1).$$

故第三项在真实模型下为

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dr_{n,i}^{(2)}(\beta)}{d\beta}\bigg|_{\beta=\beta_0}=o_p(1).$$

结合上面三项的结果可以得到, 在真实模型下有

$$\begin{pmatrix}\sqrt{n}\bar u_n\\ \sqrt{n}\bar v_n\end{pmatrix}-\begin{pmatrix}\sqrt{n}\bar u_0\\ \sqrt{n}\bar v_0\end{pmatrix}=o_p(1).$$

证毕. □

引理 2 的证明 由引理 1 可知, 我们只需证在真实模型下,

$$\begin{pmatrix}\sqrt{n}\bar u_0\\ \sqrt{n}\bar v_0\end{pmatrix}\xrightarrow{d}\begin{pmatrix}J_{01}\delta\\ J_{11}\delta\end{pmatrix}+\begin{pmatrix}M\\ N\end{pmatrix},\quad \begin{pmatrix}M\\ N\end{pmatrix}\sim N(0,J).$$

首先, 由条件 3 可得

$$\begin{aligned}
\mathrm{E}(\bar u_0(y))&=\mathrm{E}(u_0(y_i))\\
&=\int u_0(y)f_{\text{true}}(y)dydzdx\\
&=\int u_0(y)f_0(y)\left[1+v_0^{\mathrm{T}}(y)\frac{\delta}{\sqrt{n}}+R_2\left(y,\frac{\delta}{\sqrt{n}}\right)\right]dydzdx\\
&=\int u_0(y)f_0(y)dydzdx+\int u_0(y)v_0^{\mathrm{T}}(y)f_0(y)dydzdx\frac{\delta}{\sqrt{n}}+\int u_0(y)R_2\left(y,\frac{\delta}{\sqrt{n}}\right)f_0(y)dydzdx\\
&=0+\mathrm{E}_0[u_0(y)v_0^{\mathrm{T}}(y)]\frac{\delta}{\sqrt{n}}+o\left(\frac{1}{\sqrt{n}}\right)\\
&=J_{01}\frac{\delta}{\sqrt{n}}+o\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}\tag{B.1}$$

同理可知,

$$\mathrm{E}(\bar v_0(y))=\mathrm{E}(v_0(y_i))=J_{11}\frac{\delta}{\sqrt{n}}+o\left(\frac{1}{\sqrt{n}}\right).$$

其次, 我们证明

$$\mathrm{Var}\begin{pmatrix}\sqrt{n}\bar u_0\\ \sqrt{n}\bar v_0\end{pmatrix}\to J.$$

考虑

$$\mathrm{Var}(\sqrt{n}\bar u_0)=\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_0(y_i)\right)=\mathrm{Var}(u_0(y_i))=\mathrm{E}[u_0(y_i)u_0^{\mathrm{T}}(y_i)]-\mathrm{E}[u_0(y_i)]\mathrm{E}[u_0^{\mathrm{T}}(y_i)]$$

$$
\begin{aligned}
&= \int u_0(y)u_0^{\mathrm{T}}(y)f_0(y)\left[1+v_0^{\mathrm{T}}(y)\frac{\delta}{\sqrt{n}}+R_2\left(y,\frac{\delta}{\sqrt{n}}\right)\right]dydzdx+O\left(\frac{1}{n}\right)\\
&= \mathrm{E}_0[u_0(y)u_0^{\mathrm{T}}(y)]+\int[u_0(y)u_0^{\mathrm{T}}(y)]\left[v_0^{\mathrm{T}}(y)\frac{\delta}{\sqrt{n}}\right]f_0(y)dydzdx\\
&\quad+\int[u_0(y)u_0^{\mathrm{T}}(y)]R_2\left(y,\frac{\delta}{\sqrt{n}}\right)f_0(y)dydzdx+O\left(\frac{1}{n}\right).
\end{aligned}
$$

由条件 3 知上式第二项中每一个元素均为 $O(\frac{1}{\sqrt{n}})$, 且第三项中每一个元素均为 $o(1)$, 故

$$
\mathrm{Var}(\sqrt{n}\bar{u}_0)=\mathrm{Var}_0(u_0(y))+O\left(\frac{1}{\sqrt{n}}\right)+o(1)\to\mathrm{Var}_0(u_0(y)).
$$

同理可得

$$
\mathrm{Var}(\sqrt{n}\bar{v}_0)\to\mathrm{Var}_0(v_0(y))=J_{11},
$$

并且

$$
\mathrm{Cov}(\sqrt{n}\bar{u}_0,\sqrt{n}\bar{v}_0)\to\mathrm{Cov}_0(u_0(y),v_0(y))=J_{01}.
$$

又由条件 3 可知 Linderberg 条件成立, 故根据中心极限定理得到, 在真实模型下,

$$
\begin{pmatrix}\sqrt{n}\bar{u}_0\\ \sqrt{n}\bar{v}_0\end{pmatrix}\xrightarrow{d}\begin{pmatrix}J_{01}\delta\\ J_{11}\delta\end{pmatrix}+\begin{pmatrix}M\\ N\end{pmatrix}.
$$

证毕. □

引理 3 的证明　由 Taylor 展开可得

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta d\beta^{\mathrm{T}}}l(y_i,z_i,\beta,\hat{\theta}(x_i,\beta))-\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta d\beta^{\mathrm{T}}}l(y_i,z_i,\beta,\theta_0(x_i,\beta))\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta d\beta^{\mathrm{T}}}[l(y_i,z_i,\beta,\hat{\theta}(x_i,\beta))-l(y_i,z_i,\beta,\theta_0(x_i,\beta))]\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\frac{d^2r_{n,i}(\beta)}{d\beta d\beta^{\mathrm{T}}},
\end{aligned}
$$

其中

$$
r_{n,i}(\beta)=\int_0^1\frac{\partial}{\partial\theta}l[y_i,z_i,\beta,\theta_0(x_i,\beta)+t(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))]dt[\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta)].
$$

记

$$
A^{(1)}(y_i,z_i,x_i,\beta)=\int_0^1\frac{\partial}{\partial\theta}l[y_i,z_i,\beta,\theta_0(x_i,\beta)+t(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))]dt.
$$

考虑 $(p+q)\times(p+q)$ 阶矩阵 $\frac{d^2r_{n,i}(\beta)}{d\beta d\beta^{\mathrm{T}}}$ 的 (j,l) 元素, 则有

$$
\begin{aligned}
&\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{d^2r_{n,i}(\beta)}{d\beta_{(j)}d\beta_{(l)}}\right|\\
&\quad=\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_{(j)}d\beta_{(l)}}[A^{(1)}(y_i,z_i,x_i,\beta)(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))]\right|
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant \sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\left[\frac{d^2}{d\beta_{(j)}d\beta_{(l)}}A^{(1)}(y_i,z_i,x_i,\beta)\right][\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta)]\right| \\
&\quad+\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\left[\frac{d}{d\beta_{(j)}}A^{(1)}(y_i,z_i,x_i,\beta)\right]\left[\frac{d}{d\beta_{(l)}}(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))\right]\right| \\
&\quad+\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\left[\frac{d}{d\beta_{(l)}}A^{(1)}(y_i,z_i,x_i,\beta)\right]\left[\frac{d}{d\beta_{(j)}}(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))\right]\right| \\
&\quad+\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|A^{(1)}(y_i,z_i,x_i,\beta)\left[\frac{d^2}{d\beta_{(j)}d\beta_{(l)}}(\hat{\theta}(x_i,\beta)-\theta_0(x_i,\beta))\right]\right| \\
&\leqslant \sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d^2}{d\beta_{(j)}d\beta_{(l)}}A^{(1)}(y_i,z_i,x_i,\beta)\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}|\hat{\theta}(x,\beta)-\theta_0(x,\beta)| \\
&\quad+\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d}{d\beta_{(j)}}A^{(1)}(y_i,z_i,x_i,\beta)\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{d}{d\beta_{(l)}}(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\right| \\
&\quad+\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d}{d\beta_{(l)}}A^{(1)}(y_i,z_i,x_i,\beta)\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{d}{d\beta_{(j)}}(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\right| \\
&\quad+\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}|A^{(1)}(y_i,z_i,x_i,\beta)|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{d^2}{d\beta_{(j)}d\beta_{(l)}}(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\right|.
\end{aligned}
$$

先证明在真实模型下有

$$
\begin{aligned}
&\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}|A^{(1)}(y_i,z_i,x_i,\beta)| = O_p(1), \\
&\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d}{d\beta_{(j)}}A^{(1)}(y_i,z_i,x_i,\beta)\right| = O_p(1), \quad j=1,2,\dots,p+q, \\
&\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d^2}{d\beta_{(j)}d\beta_{(l)}}A^{(1)}(y_i,z_i,x_i,\beta)\right| = O_p(1), \quad j,l=1,2,\dots,p+q.
\end{aligned}
$$

由条件 1(1) 可得

$$
|A^{(1)}(y,z,x,\beta)| \leqslant \sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta}l(y,z,\beta,\theta(x))\right| \leqslant M(x,y,z),
$$

故

$$
\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}|A^{(1)}(y_i,z_i,x_i,\beta)| \leqslant \frac{1}{n}\sum_{i=1}^{n}M(x_i,y_i,z_i) = O_p(1).
$$

同理, 由条件 1(1) 可得, 在真实模型下,

$$
\begin{aligned}
&\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d}{d\beta_{(j)}}A^{(1)}(y_i,z_i,x_i,\beta)\right| = O_p(1), \quad j=1,2,\dots,p+q, \\
&\sup_{\beta\in\mathcal{N}(\beta_0)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d^2}{d\beta_{(j)}d\beta_{(l)}}A^{(1)}(y_i,z_i,x_i,\beta)\right| = O_p(1), \quad j,l=1,2,\dots,p+q.
\end{aligned}
$$

又由条件 2 知,

$$
\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{x\in\mathcal{B}}|\hat{\theta}(x,\beta)-\theta_0(x,\beta)| = o_p(1),
$$

$$\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{x\in\mathcal{B}}\left|\frac{d}{d\beta_{(j)}}(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\right|=o_p(1),$$
$$\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{x\in\mathcal{B}}\left|\frac{d^2}{d\beta_{(j)}d\beta_{(l)}}(\hat{\theta}(x,\beta)-\theta_0(x,\beta))\right|=o_p(1),\quad j,l=1,2,\ldots,p+q.$$

因此,

$$\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{1}{n}\sum_{i=1}^n\frac{d^2r_{n,i}(\beta)}{d\beta_{(j)}d\beta_{(l)}}\right|=o_p(1).$$

故有

$$\sup_{\beta\in\mathcal{N}(\beta_0)}\left\|\frac{1}{n}\sum_{i=1}^n\frac{d^2r_{n,i}(\beta)}{d\beta d\beta^{\mathrm{T}}}\right\|=o_p(1).$$

证毕. □

引理 4 的证明 证明思路受到文献 [35, 定理 3.7] 证明的启发, 但有本质的区别. 记 $\mathcal{N}(\beta_{0,S},\epsilon)$ 是以 $\beta_{0,S}$ 为中心、ϵ 为半径的球体. 我们先证明, 对于充分小的 $\epsilon>0$, 以趋于 1 的概率成立

$$\sum_{i=1}^n l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))<\sum_{i=1}^n l(y_i,z_i,\beta_{0,S},\hat{\theta}(x_i,\beta_{0,S})),\quad\forall\,\beta_S\in\partial\mathcal{N}(\beta_{0,S},\epsilon).\tag{B.2}$$

事实上, 将广义轮廓对数似然函数在 $\beta_{0,S}$ 处展开, 有

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n[l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))-l(y_i,z_i,\beta_{0,S},\hat{\theta}(x_i,\beta_{0,S}))]\\&=\frac{1}{n}\sum_{i=1}^n\left(\frac{dl(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_S}\Big|_{\beta_S=\beta_{0,S}}\right)^{\mathrm{T}}(\beta_S-\beta_{0,S})\\&\quad+\frac{1}{2n}\sum_{i=1}^n(\beta_S-\beta_{0,S})^{\mathrm{T}}\left(\frac{d^2l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_Sd\beta_S^{\mathrm{T}}}\Big|_{\beta_S=\beta_{0,S}}\right)(\beta_S-\beta_{0,S})\\&\quad+\frac{1}{6}\sum_{j=1}^{p+|S|}\sum_{k=1}^{p+|S|}\sum_{l=1}^{p+|S|}\left(\frac{1}{n}\sum_{i=1}^n\frac{d^3l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_{S,(j)}d\beta_{S,(k)}d\beta_{S,(l)}}\Big|_{\beta_S=\beta_{0,S}^*}\right)\\&\quad\times(\beta_{S,(j)}-\beta_{0,S,(j)})(\beta_{S,(k)}-\beta_{0,S,(k)})(\beta_{S,(l)}-\beta_{0,S,(l)}),\end{aligned}$$

其中 $\beta_{S,(j)}$、$\beta_{S,(k)}$ 和 $\beta_{S,(l)}$ 分别是 β_S 的第 j、k 和 l 个分量, $\beta_{0,S}^*$ 介于 $\beta_{0,S}$ 与 β_S 之间. 下面分别考虑上式中的三项.

(i) 我们先证第一项依概率趋于 0. 由引理 1 知, 在真实模型下,

$$\frac{1}{n}\sum_{i=1}^n\frac{dl(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_S}\Big|_{\beta_S=\beta_{0,S}}=\frac{1}{n}\sum_{i=1}^n\frac{dl(y_i,z_i,\beta_S,\theta_0(x_i,\beta_S))}{d\beta_S}\Big|_{\beta_S=\beta_{0,S}}+o_p\left(\frac{1}{\sqrt{n}}\right),$$

因此只需要证明在真实模型下,

$$\frac{1}{n}\sum_{i=1}^n\frac{dl(y_i,z_i,\beta_S,\theta_0(x_i,\beta_S))}{d\beta_S}\Big|_{\beta_S=\beta_{0,S}}=o_p(1),$$

显然, 在零模型下这是成立的. 而由引理 1 的证明知, 零模型下的分布与真实模型下的分布是相互邻近的, 因此, 由 Le Cam 引理知, 在真实模型下, 该式仍然成立. 这样, 在真实模型下, 第一项以概率趋于 0, 即

$$\frac{1}{n}\sum_{i=1}^{n}\frac{dl(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_S}\Big|_{\beta_S=\beta_{0,S}}=o_p(1).$$

(ii) 其次考虑第二项. 由引理 3 可得

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d^2l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_S d\beta_S^{\mathrm{T}}}\Big|_{\beta_S=\beta_{0,S}}=\frac{1}{n}\sum_{i=1}^{n}\frac{d^2l(y_i,z_i,\beta_S,\theta_0(x_i,\beta_S))}{d\beta_S d\beta_S^{\mathrm{T}}}\Big|_{\beta_S=\beta_{0,S}}+o_p(1).$$

与 (i) 类似地可以证明, 在真实模型下, 有

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d^2l(y_i,z_i,\beta_S,\theta_0(x_i,\beta_S))}{d\beta_S d\beta_S^{\mathrm{T}}}\Big|_{\beta_S=\beta_{0,S}}\xrightarrow{p}-J_S.$$

因此, 在真实模型下, 第二项以趋于 1 的概率有

$$\begin{aligned}&\frac{1}{2n}\sum_{i=1}^{n}(\beta_S-\beta_{0,S})^{\mathrm{T}}\left(\frac{d^2l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_S d\beta_S^{\mathrm{T}}}\Big|_{\beta_S=\beta_{0,S}}\right)(\beta_S-\beta_{0,S})\\&\quad\to-\frac{(\beta_S-\beta_{0,S})^{\mathrm{T}}J_S(\beta_S-\beta_{0,S})}{2}\leqslant-\frac{\lambda_{\min}(J_S)}{2}\epsilon^2,\end{aligned}$$

其中 $\lambda_{\min}(J_S)$ 表示矩阵 J_S 的最小特征值.

(iii) 最后考虑第三项. 假设 $\mathcal{N}(\beta_{0,S},\epsilon)$ 包含在 $\mathcal{N}(\beta_0)$ 中, 则易见

$$\begin{aligned}&\left|\frac{d^3l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_{S,(j)}d\beta_{S,(k)}d\beta_{S,(l)}}\right|\\&\quad\leqslant\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^3l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\beta_{S,(j)}\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\right|\\&\qquad+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^3l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\right|\\&\qquad+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^3l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta\partial\beta_{S,(j)}\partial\beta_{S,(l)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(k)}}\right|\\&\qquad+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^3l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta^2\partial\beta_{S,(l)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(k)}}\right|\\&\qquad+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^2l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta\partial\beta_{S,(l)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial^2\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}\partial\beta_{S,(k)}}\right|\\&\qquad+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^3l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\beta_{S,(j)}\partial\beta_{S,(k)}\partial\theta}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(l)}}\right|\\&\qquad+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^3l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\beta_{S,(k)}\partial\theta^2}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(l)}}\right|\\&\qquad+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^2l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\beta_{S,(k)}\partial\theta}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial^2\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}\partial\beta_{S,(l)}}\right|\end{aligned}$$

$$
\begin{aligned}
&+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^3 l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\beta_{S,(j)}\partial\theta^2}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(k)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(l)}}\right|\\
&+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^3 l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta^3}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(k)}}\right|\\
&\times\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(l)}}\right|\\
&+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^2 l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta^2}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial^2\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(l)}\partial\beta_{S,(k)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(l)}}\right|\\
&+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^2 l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta^2}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(k)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial^2\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}\partial\beta_{S,(l)}}\right|\\
&+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^2 l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\beta_{S,(j)}\partial\theta}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial^2\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\right|\\
&+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial^2 l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta^2}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial^2\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\right|\\
&+\sup_{\beta\in\mathcal{N}(\beta_0)}\sup_{\theta\in\Theta}\left|\frac{\partial l(y_i,z_i,\beta_S,\theta(x_i))}{\partial\theta}\right|\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial^3\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\right|.
\end{aligned}
$$

由条件 1(1) 可知, 上式每一项中的关于对数似然函数偏导数的上确界被 $M(x_i,y_i,z_i)$ 界住; 又由条件 2 可知, 上式每一项中的关于 $\hat{\theta}(x,\beta_S)$ 的偏导数的上确界具有阶 $O_p(1)$, 因此,

$$
\left|\frac{1}{n}\sum_{i=1}^{n}\frac{d^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_{S,(j)}d\beta_{S,(k)}d\beta_{S,(l)}}\Big|_{\beta_S=\beta^*_{0,S}}\right|\leqslant\frac{1}{n}\sum_{i=1}^{n}\left|\frac{d^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_{S,(j)}d\beta_{S,(k)}d\beta_{S,(l)}}\Big|_{\beta_S=\beta^*_{0,S}}\right|=O_p(1).
$$

故存在常数 $c>0$, 使得第三项以趋于 1 的概率有

$$
\begin{aligned}
&\left|\frac{1}{6}\sum_{j=1}^{p+q}\sum_{k=1}^{p+q}\sum_{l=1}^{p+q}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{d\beta_{S,(j)}d\beta_{S,(k)}d\beta_{S,(l)}}\Big|_{\beta_S=\beta^*_{0,S}}\right)\right.\\
&\qquad\left.\times(\beta_{S,(j)}-\beta_{0,S,(j)})(\beta_{S,(k)}-\beta_{0,S,(k)})(\beta_{S,(l)}-\beta_{0,S,(l)})\right|\\
&\qquad\leqslant\frac{1}{6}c\epsilon^3.
\end{aligned}
$$

这样, 当 $\epsilon<3\lambda_{\min}(J_S)/c\doteq c_0$ 时, (B.2) 以趋于 1 的概率成立.

显然, β_S 的函数 $\sum_{i=1}^{n}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))$ 在球体 $\mathcal{N}(\beta_{0,S},\epsilon)$ 中总有最大值点, 记为 $\hat{\beta}_S^{(n)}(\epsilon)$ (此证明中右上角加 n 特意表示与 n 有关). 如最大值点不唯一, 则取离 $\beta_{0,S}$ 最近者 (下同). 这总可以做到, 因为广义轮廓对数似然函数是连续的. 这样, $\|\hat{\beta}_S^{(n)}(\epsilon)-\beta_{0,S}\|_2$ 是 ϵ 的单调递增函数. 因此, 对固定的 n, 记

$$
A_n=\inf_{0<\epsilon<c_0}\|\hat{\beta}_S^{(n)}(\epsilon)-\beta_{0,S}\|_2,
$$

则对于 $m=1,2,\dots$, 存在 $\epsilon_n^{(1)}>\epsilon_n^{(2)}>\cdots>\epsilon_n^{(m)}>\cdots>0$ (此处及以下所涉及的 $\epsilon_n^{(m)}$ 均小于 c_0), 使得

$$
\|\hat{\beta}_S^{(n)}(\epsilon_n^{(m)})-\beta_{0,S}\|_2<A_n+\frac{1}{m}.
$$

对固定的 n, 选定 $\epsilon_n^{(m)}$ $(m=1,2,\ldots)$ 后, 按如下方式选取 $\epsilon_{n+1}^{(m)}$: 首先, 对 $m=1$, 选 $\epsilon_{n+1}^{(1)}<\epsilon_n^{(1)}$, 使得 $\|\hat{\beta}_S^{(n+1)}(\epsilon_{n+1}^{(1)})-\beta_{0,S}\|_2<A_{n+1}+1$; 其次, 对 $m=2$, 选 $\epsilon_{n+1}^{(2)}>0$, 使得 $\epsilon_{n+1}^{(2)}<\min\{\epsilon_n^{(2)},\epsilon_{n+1}^{(1)}\}$, 且 $\|\hat{\beta}_S^{(n+1)}(\epsilon_{n+1}^{(2)})-\beta_{0,S}\|_2<A_{n+1}+\frac{1}{2}$ (由于证明与下面的类似, 故省略); 最后, 在选定 $\epsilon_{n+1}^{(1)},\ldots,\epsilon_{n+1}^{(m-1)}$ 的情况下, 选 $\epsilon_{n+1}^{(m)}>0$, 使得

$$\epsilon_{n+1}^{(m)}<\min\{\epsilon_n^{(m)},\epsilon_{n+1}^{(m-1)}\}\hat{=}\kappa^* \quad 且 \quad \|\hat{\beta}_S^{(n+1)}(\epsilon_{n+1}^{(m)})-\beta_{0,S}\|_2<A_{n+1}+\frac{1}{m}.$$

这样的 $\epsilon_{n+1}^{(m)}$ 是一定存在的. 如若不然, 则 $\forall\,\epsilon'<\kappa^*$, 都有

$$\|\hat{\beta}_S^{(n+1)}(\epsilon')-\beta_{0,S}\|_2\geqslant A_{n+1}+\frac{1}{m}.$$

由此可知, $\forall\,\epsilon''\geqslant\kappa^*$, 也有

$$\|\hat{\beta}_S^{(n+1)}(\epsilon'')-\beta_{0,S}\|_2\geqslant\left\|\hat{\beta}_S^{(n+1)}\left(\frac{1}{2}\kappa^*\right)-\beta_{0,S}\right\|_2\geqslant A_{n+1}+\frac{1}{m}.$$

因而,

$$\inf_{0<\epsilon<c_0}\|\hat{\beta}_S^{(n+1)}(\epsilon)-\beta_{0,S}\|_2\geqslant A_{n+1}+\frac{1}{m},$$

矛盾.

由 $\{\epsilon_n^{(m)},m,n=1,2,\ldots\}$ 的选取可知, 序列 $\{\epsilon_n^{(n)}\}$ 是单调下降且有界的, 因而有极限, 记为 $\epsilon^{(0)}$. 如果 $\epsilon^{(0)}=0$, 则 $\|\hat{\beta}_S^{(n)}(\epsilon_n^{(n)})-\beta_{0,S}\|_2\leqslant\epsilon_n^{(n)}\to0$, 于是,

$$\lim_{n\to\infty}\hat{\beta}_S^{(n)}(\epsilon_n^{(n)})=\beta_{0,S},$$

即 $\beta_{0,S}$ 存在处处收敛于它的估计. 如果 $\epsilon^{(0)}>0$, 则对此 $\epsilon^{(0)}$, 由于其小于 c_0, 故 (B.2) 以趋于 1 的概率成立. 这样, $\sum_{i=1}^n l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))$ 以趋于 1 的概率在 $\mathcal{N}(\beta_{0,S},\epsilon^{(0)})$ 内部有最大值点, 记为 $\hat{\beta}_S^{(n)}(\epsilon^{(0)})$. 因此, 广义轮廓对数似然方程以趋于 1 的概率有解 $\hat{\beta}_S^{(n)}(\epsilon^{(0)})$. 下面证明其相合性.

任意固定 $\epsilon^*>0$ (不妨设其小于 c_0), 则同样广义轮廓对数似然方程以趋于 1 的概率有解 $\hat{\beta}_S^{(n)}(\epsilon^*)$, 它是广义轮廓对数似然函数在 $\mathcal{N}(\beta_{0,S},\epsilon^*)$ 内部的最大值点. 因而, $\forall\,\delta>0$, 存在 $N_\delta>0$, 使得当 $n>N_\delta$ 时, 有

$$\Pr(\|\hat{\beta}_S^{(n)}(\epsilon^*)-\beta_{0,S}\|_2<\epsilon^*)>1-\delta.$$

另一方面, $\forall\,m,n=1,2,\ldots$, 都有 $\epsilon^{(0)}\leqslant\epsilon_n^{(m)}$, 故有

$$\|\hat{\beta}_S^{(n)}(\epsilon^{(0)})-\beta_{0,S}\|_2\leqslant\|\hat{\beta}_S^{(n)}(\epsilon_n^{(m)})-\beta_{0,S}\|_2<A_n+\frac{1}{m},\quad\forall\,m,n.$$

由此知,

$$\|\hat{\beta}_S^{(n)}(\epsilon^{(0)})-\beta_{0,S}\|_2=A_n=\inf_{0<\epsilon<c_0}\|\hat{\beta}_S^{(n)}(\epsilon)-\beta_{0,S}\|_2.$$

这样, 当 $n>N_\delta$ 时,

$$\Pr(\|\hat{\beta}_S^{(n)}(\epsilon^{(0)})-\beta_{0,S}\|_2<\epsilon^*)\geqslant\Pr(\|\hat{\beta}_S^{(n)}(\epsilon^*)-\beta_{0,S}\|_2<\epsilon^*)>1-\delta.$$

证毕. □

引理 5 的证明 由 $\hat{\beta}_S$ 是似然方程的解, 可得

$$
\begin{aligned}
0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\beta_S}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\Big|_{\beta_S=\hat{\beta}_S}\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}u_S(y_i)\\ v_S(y_i)\end{pmatrix}+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\Big|_{\beta_S=\beta_{0,S}^*}\right][\sqrt{n}(\hat{\beta}_S-\beta_{0,S})],
\end{aligned}
$$

其中 $\beta_{0,S}^*$ 介于 $\beta_{0,S}$ 与 $\hat{\beta}_S$ 之间.

将上式第二项中的 $\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\,|_{\beta_S=\beta_{0,S}^*}$ 在 $\beta_{0,S}$ 处进行 Taylor 展开得

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\Big|_{\beta_S=\beta_{0,S}^*}\\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\Big|_{\beta_S=\beta_{0,S}}\\
&\qquad+\sum_{j=1}^{p+|S|}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d^3}{d\beta_{S,(j)}d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\Big|_{\beta_S=\beta_{0,S}^{\nabla}}\right][\beta_{S,(j)}^*-\beta_{0,S,(j)}],
\end{aligned}
$$

其中 $\beta_{0,S}^{\nabla}$ 介于 $\beta_{0,S}$ 与 $\beta_{0,S}^*$ 之间.

由引理 3 知,

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\Big|_{\beta_S=\beta_{0,S}}=\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\theta_0(x_i,\beta_S))\Big|_{\beta_S=\beta_{0,S}}+o_p(1),
$$

而在引理 4 的证明中, 我们已经证得在真实模型下, 有

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\theta_0(x_i,\beta_S))\Big|_{\beta_S=\beta_{0,S}}\xrightarrow{p}-J_S,
$$

因此, 下面考虑 $\frac{1}{n}\sum_{i=1}^{n}\frac{d^3}{d\beta_{S,(j)}d\beta_S d\beta_S^{\mathrm{T}}}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\,|_{\beta_S=\beta_{0,S}^{\nabla}}$, 其 (k,l) 元素为

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{d^3}{d\beta_{S,(k)}d\beta_{S,(l)}d\beta_{S,(j)}}l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\Big|_{\beta_S=\beta_{0,S}^{\nabla}},
$$

引理 4 的证明中给出了它的展开式. 我们先考虑展开式的第一项, 即证明

$$
\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{\partial\beta_{S,(j)}\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\Big|_{\beta_S=\beta_{0,S}^{\nabla}}\right|=O_p(1).
$$

事实上, 不难看出,

$$
\begin{aligned}
&\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{\partial\beta_{S,(j)}\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\Big|_{\beta_S=\beta_{0,S}^{\nabla}}\right|>L\right)\\
&\quad=\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{\partial\beta_{S,(j)}\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\Big|_{\beta_S=\beta_{0,S}^{\nabla}}\right|>L,|\beta_{0,S}^{\nabla}-\beta_{0,S}|\leqslant\eta\right)\\
&\qquad+\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{\partial\beta_{S,(j)}\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\Big|_{\beta_S=\beta_{0,S}^{\nabla}}\right|>L,|\beta_{0,S}^{\nabla}-\beta_{0,S}|>\eta\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant \Pr\left(\frac{1}{n}\sum_{i=1}^{n} M(x_i,y_i,z_i) > L\right) + \Pr(|\beta_{0,S}^{\nabla} - \beta_{0,S}| > \eta) \\
&\leqslant \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} M(x_i,y_i,z_i)\right)\frac{1}{L} + \Pr(|\beta_{0,S}^{\nabla} - \beta_{0,S}| > \eta) \\
&= \frac{\mathrm{E}(M(x,y,z))}{L} + \Pr(|\beta_{0,S}^{\nabla} - \beta_{0,S}| > \eta).
\end{aligned}
$$

先令 $n\to\infty$, 由引理 4 可得上式第二项趋于 0; 再令 $L\to\infty$, 由条件 1(1) 可得上式第一项趋于 0, 故展开式的第一项为 $O_p(1)$.

再考虑展开式的第二项, 即证明

$$
\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{\partial\theta\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right|\sup_{x\in\mathcal{B}}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right| = O_p(1).
$$

这只需分别证明

$$
\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3 l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))}{\partial\theta\partial\beta_{S,(k)}\partial\beta_{S,(l)}}\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right| = O_p(1)
$$

和

$$
\sup_{x\in\mathcal{B}}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right| = O_p(1)
$$

即可. 前者类似第一项的证明可以得到; 对于后者, 我们有

$$
\begin{aligned}
&\Pr\left(\sup_{x\in\mathcal{B}}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right| > L\right) \\
&\quad= \Pr\left(\sup_{x\in\mathcal{B}}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right| > L, |\beta_{0,S}^{\nabla}-\beta_{0,S}|\leqslant\eta\right) \\
&\qquad+ \Pr\left(\sup_{x\in\mathcal{B}}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right| > L, |\beta_{0,S}^{\nabla}-\beta_{0,S}| > \eta\right) \\
&\quad\leqslant \Pr\left(\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\right| > L\right) + \Pr(|\beta_{0,S}^{\nabla}-\beta_{0,S}| > \eta) \\
&\quad\leqslant \mathrm{E}\left(\sup_{x\in\mathcal{B}}\sup_{\beta\in\mathcal{N}(\beta_0)}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\right|\right)\frac{1}{L} + \Pr(|\beta_{0,S}^{\nabla}-\beta_{0,S}| > \eta).
\end{aligned}
$$

先令 $n\to\infty$, 由引理 4 可得上式第二项趋于 0; 再令 $L\to\infty$, 由条件 2 可得上式第一项趋于 0, 故

$$
\sup_{x\in\mathcal{B}}\left|\frac{\partial\hat{\theta}(x,\beta_S)}{\partial\beta_{S,(j)}}\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right| = O_p(1),
$$

即展开式的第二项为 $O_p(1)$. 同理可证展开式的其余各项均为 $O_p(1)$.

因此, $\frac{1}{n}\sum_{i=1}^{n}\frac{d^3}{d\beta_{S,(j)}d\beta_S d\beta_S^{\mathrm{T}}} l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\mid_{\beta_S=\beta_{0,S}^{\nabla}}$ 的 (k,l) 元素满足

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{d^3}{d\beta_{S,(k)}d\beta_{S,(l)}d\beta_{S,(j)}} l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right| \\
&\quad\leqslant \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d^3}{d\beta_{S,(k)}d\beta_{S,(l)}d\beta_{S,(j)}} l(y_i,z_i,\beta_S,\hat{\theta}(x_i,\beta_S))\ \Bigg|_{\beta_S=\beta_{0,S}^{\nabla}}\right|
\end{aligned}
$$

$$= O_p(1).$$

由此并利用 $\beta^*_{S,(j)} - \beta_{0,S,(j)} = o_p(1)$ (见引理 4), 我们得到

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\beta_S d\beta_S^{\mathrm{T}}} l(y_i, z_i, \beta_S, \hat{\theta}(x_i, \beta_S))\bigg|_{\beta_S=\beta^*_{0,S}} \xrightarrow{p} -J_S.$$

最后, 根据引理 2, 有

$$\sqrt{n}(\hat{\beta}_S - \beta_{0,S}) \doteq J_S^{-1}\begin{pmatrix}\sqrt{n}\bar{u}_{n,S}\\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix} \xrightarrow{d} J_S^{-1}\begin{pmatrix}J_{01}\delta + M\\ \pi_S J_{11}\delta + N_S\end{pmatrix} \sim N\left(J_S^{-1}\begin{pmatrix}J_{01}\\ \pi_S J_{11}\end{pmatrix}\delta, J_S^{-1}\right).$$

证毕. □

引理 6 的证明　对 $\hat{\mu}_S$ 和 μ_{true} 在 β_0 处 Taylor 展开, 有

$$\hat{\mu}_S = \mu(\hat{\alpha}_S, \hat{\gamma}_S, \gamma_{0,S^c}) = \mu(\alpha_0, \gamma_0) + \begin{pmatrix}\frac{\partial\mu}{\partial\alpha}\\ \frac{\partial\mu}{\partial\gamma_S}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\hat{\alpha}_S - \alpha_0\\ \hat{\gamma}_S - \gamma_{0,S}\end{pmatrix} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

$$\mu_{\mathrm{true}} = \mu\left(\alpha_0, \gamma_0 + \frac{\delta}{\sqrt{n}}\right) = \mu(\alpha_0, \gamma_0) + \left(\frac{\partial\mu}{\partial\gamma}\right)^{\mathrm{T}}\left(\frac{\delta}{\sqrt{n}}\right) + o_p\left(\frac{1}{\sqrt{n}}\right).$$

则

$$\sqrt{n}(\hat{\mu}_S - \mu_{\mathrm{true}}) = \sqrt{n}\begin{pmatrix}\frac{\partial\mu}{\partial\alpha}\\ \frac{\partial\mu}{\partial\gamma_S}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\hat{\alpha}_S - \alpha_0\\ \hat{\gamma}_S - \gamma_{0,S}\end{pmatrix} - \sqrt{n}\left(\frac{\partial\mu}{\partial\gamma}\right)^{\mathrm{T}}\left(\frac{\delta}{\sqrt{n}}\right) + o_p(1).$$

从而, 由引理 5 可得

$$\sqrt{n}(\hat{\mu}_S - \mu_{\mathrm{true}}) \xrightarrow{d} \left(\frac{\partial\mu}{\partial\alpha}\right)^{\mathrm{T}} C_S + \left(\frac{\partial\mu}{\partial\gamma_s}\right)^{\mathrm{T}} D_S - \left(\frac{\partial\mu}{\partial\gamma}\right)^{\mathrm{T}}\delta.$$

经简单计算可以知道上式右边即为

$$\left(\frac{\partial\mu}{\partial\alpha}\right)^{\mathrm{T}} J_{00}^{-1}M + \omega^{\mathrm{T}}(\delta - K^{\frac{1}{2}}H_S K^{-\frac{1}{2}}D) = \Lambda_S.$$

证毕. □

定理 1 的证明　由引理 5 知,

$$\sqrt{n}(\hat{\beta}_S - \beta_{0,S}) \doteq J_S^{-1}\begin{pmatrix}\sqrt{n}\bar{u}_{n,S}\\ \sqrt{n}\bar{v}_{n,S}\end{pmatrix} = J_S^{-1}\begin{pmatrix}I_p & \\ & \pi_S\end{pmatrix}\begin{pmatrix}\sqrt{n}\bar{u}_n\\ \sqrt{n}\bar{v}_n\end{pmatrix},$$

$$D_n = \sqrt{n}(\hat{\gamma}_{\mathrm{full}} - \gamma_0) \doteq (0_{q\times p} I_q)J^{-1}\begin{pmatrix}\sqrt{n}\bar{u}_n\\ \sqrt{n}\bar{v}_n\end{pmatrix} \xrightarrow{d} (0_{q\times p} I_q)J^{-1}\begin{pmatrix}J_{01}\delta + M\\ J_{11}\delta + N\end{pmatrix} = D.$$

可以看出, 对任意的候选模型, 除 $o_p(1)$ 项外, 上两式依赖于相同的部分 $\left(\begin{smallmatrix}\sqrt{n}\bar{u}_n\\ \sqrt{n}\bar{v}_n\end{smallmatrix}\right)$, 故有同时收敛性. 注意到

$$\sqrt{n}(\hat{\mu} - \mu_{\mathrm{true}}) = \sum_S c(S \mid D_n)\sqrt{n}(\hat{\mu}_S - \mu_{\mathrm{true}}),$$

且 $c(S \mid D_n)$ 是 D_n 的连续函数, 故由连续映射定理可得

$$\sqrt{n}(\hat{\mu} - \mu_{\mathrm{true}}) \xrightarrow{d} \sum_S c(S \mid D)\Lambda_S = \Lambda.$$

通过简单的计算知 $\Lambda = \Lambda_0 + \omega^{\mathrm{T}}(\delta - \hat{\delta}(D))$. 证毕. □

定理 2 的证明 由定理 1 的证明可知,

$$(\sqrt{n}(\hat{\mu}-\mu_{\text{true}}), D_n) \xrightarrow{d} (\Lambda_0+\omega^{\mathrm{T}}(\delta-\hat{\delta}(D)), D).$$

故有

$$T_n \xrightarrow{d} \frac{\Lambda_0+\omega^{\mathrm{T}}(\delta-D)}{\kappa} \sim N(0,1),$$

即所构造的统计量 T_n 依分布收敛到标准正态分布. 因此,

$$\Pr\{\mu_{\text{true}} \in [\text{low}_n, \text{up}_n]\} = \Pr\{-u_{\alpha/2} \leqslant T_n \leqslant u_{\alpha/2}\} \to 2\Phi(u_{\alpha/2})-1 = 1-\alpha.$$

证毕. □

The asymptotic theory for model averaging in general semiparametric models

Rong Zhu & Guohua Zou

Abstract This paper focuses on the model averaging method for semiparametric models. We aim to extend the framework of the parameter models given by Hjort and Claeskens in 2003. Although Claeskens and Carroll considered exactly the same problem in 2007, the two methods are not identical. We derive the asymptotic distribution of the model averaging estimator, and also develop a confidence interval with an actual coverage probability that tends toward the nominal level in large samples. Both simulation study and real data analysis show that the proposed model averaging method performs well.

Keywords asymptotic distribution, generalized profile likelihood estimation, model averaging, model selection, semiparametric model

MSC(2010) 62E20, 62F12

doi: 10.1360/N012017-00010