



一般协方差矩阵下的修正 Akaike 信息准则

赵尚威¹, 喻达磊^{2*}, 张新雨³

1. 中央民族大学理学院, 北京 100081;

2. 云南财经大学统计与数学学院, 昆明 650221;

3. 中国科学院数学与系统科学研究院, 北京 100190

E-mail: zsw2011@126.com, yudalei@126.com, xinyu@amss.ac.cn

收稿日期: 2020-04-13; 接受日期: 2020-06-24; 网络出版日期: 2021-03-16; * 通信作者

国家重点研发计划 (批准号: 2020AAA0105200)、国家自然科学基金 (批准号: 11661079, 71925007 和 11688101)、中国科学院青年创新促进会和北京智源人工智能研究院资助项目

摘要 本文在 Stein 恒等式 (Stein's identity) 的框架下, 给出了一种适用于有限样本场合的全新的修正 Akaike 信息准则 (corrected Akaike information criterion), 所提出的新准则适用于非常一般的协方差结构. 在一定的正则性条件下, 本文建立了所提出准则的渐近有效性. 应用带有自回归误差的空间回归模型进行模拟, 结果表明, 在备选模型与真实的数据生成过程之间的差异较小时, 本文所提出方法的表现是令人满意的. 当这种差异变大时, 本文所提出的方法与其他已有方法相比也非常有竞争力. 所提出的方法也被用于一组实际数据 (社区犯罪数据) 的分析中, 所得到的结果更进一步支持了我们的方法在实际数据分析中的应用.

关键词 渐近有效性 修正 Akaike 信息准则 模型选择 空间自回归误差

MSC (2020) 主题分类 62J05, 62F12, 62H11

1 引言

在大多数应用场合, 计量经济学家和统计学家通常面对多个备选模型, 且需要依托已有的数据来确定使用哪个模型. 事实上, 选择最合适的模型通常是数据分析的第一步^[1]. Akaike 信息准则 (Akaike information criterion, AIC) 作为模型选择准则, 在不同领域的实际数据分析中扮演了重要角色 (参见文献 [2, 第 60 页]), 同时也在不同的统计学文献中得到了大量的研究.

在很多常见的模型中, AIC 本质上是相对熵损失的最核心部分的渐近无偏估计^[3], 并通过大样本论证获得 (渐近理论下 AIC 的正式推导可在文献 [2, 第 7 章] 中找到). 但是, 由于在实际应用中我们只有有限个样本, 因此, 研究模型选择准则的有限样本修正总是一个重要的问题. 事实上, 在有限样本场合下, 由渐近理论获得的 AIC 可能有很大的偏差^[4]. 在自回归时间序列分析模型中, 文献 [4] 通过研究常规 AIC 的小样本偏差, 给出了修正的 AIC. 此外, 文献 [5] 提出了用于选择非参数回归的光滑参

英文引用格式: Zhao S, Yu D, Zhang X. Corrected Akaike information criterion with general covariance matrix (in Chinese). Sci Sin Math, 2022, 52: 555–582, doi: 10.1360/SCM-2020-0308

数的修正版本的 AIC. 此外, 依托汇合型超几何函数 (confluent hypergeometric function), 文献 [6] 提出了针对带有独立正态误差的改进型 AIC. 上述修正或者改进的 AIC 通常在正态二次型的比的数学期望的近似表达式的基础上导出, 同时有很好的理论性质和数值表现. 另一方面, Stein 恒等式^[7]作为一种导出正态分布随机变量函数均值无偏估计的重要方法在不同领域得到了广泛的应用. 由于推导 AIC 本质上是寻找无偏估计, 因此, Stein 恒等式在模型选择领域也得到了非常广泛的应用. 例如, 针对线性混合效应模型的模型选择问题, 文献 [8] 使用 Stein 恒等式导出了修正的条件 AIC (conditional Akaike information criterion, cAIC), 此 cAIC 在误差方差已知的情形下是条件 Akaike 信息量的无偏估计. 文献 [8] 使用数据扰动的方法计算 cAIC, 该方法的计算量较大. 为克服此缺点, 文献 [9] 研究了文献 [8] 的修正 cAIC 的解析表达式, 并证明了文献 [8] 的准则就算在参数位于参数空间的边界上时仍然具有很好的理论性质. 针对带有相关结构的数据, 使用固定效应和随机效应对应的联合似然函数, 同样在 Stein 恒等式的框架下, 文献 [1] 提出了线性混合效应模型中的模型选择工具, 并研究了该方法的渐近性质. 在另一个不同的框架下, 文献 [10] 研究了交叉核实准则的迭代纠偏准则版本, 该准则可以给出高至 $2k$ 阶的风险函数的无偏估计, 该方法也拥有很好的小样本表现 (在文献 [10] 的模拟中, 样本量 n 的通常取值为 10 或者 20).

上述一系列工作聚焦于线性模型或者线性混合效应模型的模型选择问题. 但是, 这些工作中考虑的协方差矩阵的结构相对简单, 这些结果也无法被直接用于带有更一般或者更复杂协方差结构的情形 (例如, 当误差项带有空间自回归结构时). 基于上述讨论, 有必要研究适用于更一般或者更复杂协方差结构的有限样本下的修正版本的 AIC. 我们的工作解决了上述问题, 并且有下面两项主要贡献:

- (1) 在 Stein 恒等式的框架下, 我们提出了线性模型中适用于有限样本场合的全新的修正 AIC, 该修正 AIC 适用于带有一般协方差结构的场合;
- (2) 我们研究了所提出的模型选择准则的渐近性质, 证明了该准则是渐近损失有效的 (asymptotically loss efficient).

本文余下内容的结构如下. 第 2 节给出新的修正 AIC 的推导. 第 3 节研究所提出的修正 AIC 的渐近性质. 第 4 节使用 Monte Carlo 模拟评价所提出的准则的数值表现. 第 5 节使用一组实际数据 (社区犯罪数据, the Neighborhood Crimes Data) 展示了修正 AIC 的具体应用. 第 6 节总结全文. 理论证明、额外的讨论和计算细节在附录中给出.

2 Akaike 信息量及其估计

考虑下列数据生成过程:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}_0, \quad \mathbf{e}_0 \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Omega}_0). \quad (2.1)$$

拟合模型为

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim \text{Normal}(\mathbf{0}_{n \times 1}, \boldsymbol{\Omega}(\boldsymbol{\alpha})), \quad (2.2)$$

其中 $\mathbf{X}$ 是 $n \times p$ 维列满秩协变量矩阵, $\boldsymbol{\beta}$ 是 $p \times 1$ 维回归系数向量, $\boldsymbol{\alpha}$ 是 $q \times 1$ 维方差分量参数. 该拟合模型的设定可以是误设定的. 关于 $\mathbf{e}_0$ 正态性的假定只用于证明定理 2.1 中的无偏性. 在第 3 节中, 在证明渐近有效性的过程中, 此类具体的分布假定可以剔除.

现在通过提出一种 Akaike 信息量的无偏估计给出一种模型选择准则. 设 $\mathbf{y}^*$ 是源于数据生成过程 (2.1) 的另外一组实现值, 且与可观测响应变量 $\mathbf{y}$ 独立. 用 $E_{\mathbf{y}}(\cdot)$ 和 $E_{\mathbf{y}^*}(\cdot)$ 分别表示对 $\mathbf{y}$ 和 $\mathbf{y}^*$ 的概率

密度函数求数学期望. 为简单起见, 记 $E(\cdot) = E_{\mathbf{y}}(\cdot)$. 用 $\hat{\beta}$ 和 $\hat{\alpha}$ 分别表示 β 和 α 在模型 (2.2) 下通过可观测数据 $\mathbf{y}$ 获得的极大似然估计, 则拟合模型 (2.2) 对应的 Akaike 信息量为

$$\begin{aligned} \text{AI} &= E(E_{\mathbf{y}^*}[-2\log \text{likelihood}\{\mathbf{y}^*, \mathbf{X}, \hat{\beta}, \Omega(\hat{\alpha})\}]) \\ &= E[E_{\mathbf{y}^*}(n \log 2\pi + \log |\Omega(\hat{\alpha})| + (\mathbf{y}^* - \mathbf{X}\hat{\beta})^T \Omega^{-1}(\hat{\alpha})(\mathbf{y}^* - \mathbf{X}\hat{\beta}))]. \end{aligned} \quad (2.3)$$

用于模型选择的 Akaike 信息准则是 Akaike 信息量的估计量^[2]. 依托 Stein 恒等式, 我们提出了一种新的 Akaike 信息量的估计, 并将其作为模型选择准则. 现在定义一些常用记号. 对于 $j, j_1 = 1, \dots, q$, 记

$$\begin{aligned} U_j(\alpha) &= \frac{\partial \Omega^{-1}(\alpha)}{\partial \alpha_j}, \quad U_{jj_1}(\alpha) = \frac{\partial^2 \Omega^{-1}(\alpha)}{\partial \alpha_j \partial \alpha_{j_1}}, \\ \Pi_1(\Omega_0, \mathbf{X}, \mathbf{y}) &= 2\text{trace}\left\{\hat{\Omega}(\hat{\alpha})\Omega_0^{-1}\mathbf{X}\frac{\partial \hat{\beta}}{\partial \mathbf{y}^T}\right\} - 2\sum_{j=1}^q \frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}^T} \Omega_0 U_j(\hat{\alpha})(\mathbf{y} - \mathbf{X}\hat{\beta}), \\ \Pi_2(\Omega_0, \mathbf{X}, \mathbf{y}) &= \text{trace}\left\{\sum_{j=1}^q U_j(\hat{\alpha})\Omega_0 \frac{\partial^2 \hat{\alpha}_j}{\partial \mathbf{y} \partial \mathbf{y}^T} \Omega_0\right\} + \sum_{j=1}^q \sum_{j_1=1}^q \frac{\partial \hat{\alpha}_{j_1}}{\partial \mathbf{y}^T} \Omega_0 U_{jj_1}(\hat{\alpha})\Omega_0 \frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}} \end{aligned} \quad (2.4)$$

以及

$$\begin{aligned} \hat{\text{AI}}(\Omega_0) &= n \log 2\pi + \log |\Omega(\hat{\alpha})| + (\mathbf{y} - \mathbf{X}\hat{\beta})^T \Omega^{-1}(\hat{\alpha})(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &\quad + \Pi_1(\Omega_0, \mathbf{X}, \mathbf{y}) + \Pi_2(\Omega_0, \mathbf{X}, \mathbf{y}). \end{aligned} \quad (2.5)$$

为了证明的顺畅性, 我们引入一个引理, 将被用在定理 2.1 的证明中.

引理 2.1^[7] (Stein 恒等式) 设 a 为 $\text{Normal}(0, 1)$ 随机变量, $g(a) : \mathbb{R} \rightarrow \mathbb{R}$ 是 Lebesgue 可测函数 $\dot{g}(a)$ (本质上是 $g(a)$ 的导数) 的不定积分. 如果同时还满足 $E|\dot{g}(a)| < \infty$, 则

$$E(\dot{g}(a)) = E(ag(a)).$$

我们有下列结果.

定理 2.1 如果数据由 (2.1) 产生, 且被 (2.2) 拟合, 则

$$E(\hat{\text{AI}}(\Omega_0)) = \text{AI}.$$

定理 2.1 的证明在附录 A.1.1 给出, 由此知道 $\hat{\text{AI}}(\Omega_0)$ 的数学期望恰好等于 AI. 但是, $\hat{\text{AI}}(\Omega_0)$ 作为模型选择准则是不可用的, 原因在于 Ω_0 是未知的. 接下来将讨论如何通过关联 $\hat{\text{AI}}(\Omega_0)$ 和文献 [4] 的修正 AIC 来估计 Ω_0 .

考虑一类最简化的情形, 其中 $\Omega_0 = \sigma_0^2 \mathbf{I}_n$ 和 $\Omega(\alpha) = \sigma^2 \mathbf{I}_n$. 为简化推导, 在此类情形下记 $\alpha = \sigma^{-2}$. 在此场合下, (2.5) 退化为

$$\hat{\text{AI}}(\Omega_0) = n \log 2\pi + n \log \hat{\sigma}^2 + n + 2(p+2) \frac{\sigma_0^2}{\hat{\sigma}^2} + \frac{8 - 2(n-p)}{n} \frac{\sigma_0^4}{\hat{\sigma}^4}, \quad (2.6)$$

其中

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}{n-p}.$$

(2.6) 的证明在附录 A.1.2 给出. 由于 σ_0^2 未知, (2.6) 中的准则 $\hat{\text{AI}}(\Omega_0)$ 还是无法用于模型选择.

文献 [4] 中修正 AIC 的表达式为

$$\text{AICc} = n \log 2\pi + n \log \hat{\sigma}^2 + n \frac{n+p}{n-p-2}. \quad (2.7)$$

若假定拟合模型 (2.2) 设定是正确的, 即存在 β_0 和 α_0 , 使得 $\mathbf{X}\beta_0 = \mu$ 且 $\Omega(\alpha_0) = \sigma_0^2 \mathbf{I}_n$ (这类条件在推导模型选择准则时是常用的条件, 可参见文献 [4, 5]), 则

$$\mathbb{E}(\hat{\text{AI}}(\Omega_0)) = \mathbb{E}(\text{AICc}). \quad (2.8)$$

(2.8) 的证明在附录 A.1.3 给出. 合并 (2.8) 和定理 2.1 可知, 当模型 (2.2) 设定正确时, 修正 AIC 是 (2.3) 中定义的 Akaike 信息量的无偏估计. $\hat{\text{AI}}(\Omega_0)$ 与文献 [4] 中的 AICc 关系非常密切. 事实上, 如果将 (2.6) 中 $\hat{\text{AI}}(\Omega_0)$ 表达式里第 3 和 4 项中的 σ_0^2 分别替换为

$$\frac{n}{n-p-2}\hat{\sigma}^2 \quad \text{和} \quad \frac{n}{(n-p-2)^{1/2}(n-p-4)^{1/2}}\hat{\sigma}^2, \quad (2.9)$$

我们可以立刻得到 (2.7) 中的 AICc. 受此启发, 我们提出使用

$$\hat{\Omega}_{01} = \frac{n}{n-p-2}\Omega(\hat{\alpha}) \quad \text{和} \quad \hat{\Omega}_{02} = \frac{n}{(n-p-2)^{1/2}(n-p-4)^{1/2}}\Omega(\hat{\alpha}) \quad (2.10)$$

分别估计 (2.5) 中 $\hat{\text{AI}}(\Omega_0)$ 表达式里第 4 和 5 项中的 Ω_0 . 在此框架下, 可以得到下列修正 AIC:

$$\begin{aligned} \text{AICc}^* &= n \log 2\pi + \log |\Omega(\hat{\alpha})| + (\mathbf{y} - \mathbf{X}\hat{\beta})^T \Omega^{-1}(\hat{\alpha})(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &\quad + \Pi_1(\hat{\Omega}_{01}, \mathbf{X}, \mathbf{y}) + \Pi_2(\hat{\Omega}_{02}, \mathbf{X}, \mathbf{y}). \end{aligned} \quad (2.11)$$

显然, 文献 [4] 中的 AICc 是 AICc* 在 $\Omega(\hat{\alpha}) = \hat{\sigma}^2 \mathbf{I}_n$ 下的一个特例.

为计算 AICc*, 需计算 $\Pi_1(\hat{\Omega}_{01}, \mathbf{X}, \mathbf{y})$ 和 $\Pi_2(\hat{\Omega}_{02}, \mathbf{X}, \mathbf{y})$ 中的 $\partial \hat{\beta}^T / \partial \mathbf{y}$ 、 $\partial \hat{\alpha}^T / \partial \mathbf{y}$ 和 $\partial^2 \hat{\alpha}_j / \partial \mathbf{y} \partial \mathbf{y}^T$. 如果使用数值方法计算这些表达式 (如数据扰动), 基于最小化 AICc* 的模型选择过程在计算上将变得非常烦琐, 同时数值上也不稳定, 原因在于需要分别重复拟合扰动后的数据 n 和 n^2 次. 附录 A.3.1 给出上述导数的解析表达式, 这大大减轻了计算负担.

3 模型选择有效性

本节将在模型选择有效性的框架下分析所提出的 AICc* 的理论表现. 本文余下的部分, 我们不需要假定数据生成过程 (2.1) 中误差向量 $\mathbf{e}_0$ 来自正态分布, 只需要假定 $\mathbf{e}_0 \sim (\mathbf{0}_{n \times 1}, \Omega_0)$ 即可.

设我们有 M 个备选模型, 其中第 m 个备选模型为

$$\mathbf{y} = \mathbf{X}_{(m)}\beta_{(m)} + \mathbf{e}_{(m)}, \quad \mathbf{e}_{(m)} \sim \text{Normal}(\mathbf{0}_{n \times 1}, \Omega_{(m)}(\alpha_{(m)})), \quad (3.1)$$

其中 $\mathbf{X}_{(m)}$ 是一个 $n \times p_m$ 列满秩矩阵, $\beta_{(m)}$ 和 $\alpha_{(m)}$ 分别是 p_m 和 q_m 维回归系数向量和方差分量参数向量. 在本研究中, p_m 和 q_m 均为固定值. 这里允许所有的备选模型都存在设定偏误, 即

$$\mathbf{X}_{(m)}\beta_{(m)} \neq \mu,$$

(2.1) 中的 $\mathbf{e}_0$ 也没必要来自于正态总体.

记 $\alpha_{(m)} = (\alpha_{(m),1}, \dots, \alpha_{(m),q_m})^T$ 并使用 $\hat{\alpha}_{(m)} = (\hat{\alpha}_{(m),1}, \dots, \hat{\alpha}_{(m),q_m})^T$ 表示第 m 个备选模型下 $\alpha_{(m)}$ 的极大似然估计. 记 $\hat{\Omega}_{(m)} = \Omega_{(m)}(\hat{\alpha}_{(m)})$ 且 $\hat{\beta}_{(m)} = (\mathbf{X}_{(m)}^T \hat{\Omega}_{(m)}^{-1} \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \hat{\Omega}_{(m)}^{-1} \mathbf{y}$. 定义

$$\text{AICc}_{(m)}^* = n \log(2\pi) + \log |\hat{\Omega}_{(m)}| + (\mathbf{y} - \mathbf{X}_{(m)} \hat{\beta}_{(m)})^T \hat{\Omega}_{(m)}^{-1} (\mathbf{y} - \mathbf{X}_{(m)} \hat{\beta}_{(m)}) + \hat{\Pi}_{(m),1} + \hat{\Pi}_{(m),2}, \quad (3.2)$$

其中 $\hat{\Pi}_{(m),1}$ 和 $\hat{\Pi}_{(m),2}$ 将在等式 (3.3) 和 (3.4) 中定义. 最小化 $\text{AICc}_{(m)}^*$ 选择模型, 被选出的模型标为

$$\hat{m} = \underset{1 \leq m \leq M}{\operatorname{argmin}} \text{AICc}_{(m)}^*.$$

现在列出一些本文余下部分常用的记号. 将一个一般矩阵 $\mathbf{A}$ 的最大特征根和最小特征根分别记为 $\lambda_{\max}(\mathbf{A})$ 和 $\lambda_{\min}(\mathbf{A})$, $\|\mathbf{A}\|$ 为谱范数. 对于 $m \in \{1, \dots, M\}$ 和 $j, l, s \in \{1, \dots, q_m\}$, 记

$$\begin{aligned} \hat{\Omega}_{1,(m)} &= \frac{n \Omega_{(m)}(\hat{\alpha}_{(m)})}{n - p_m - 2}, \quad \hat{\Omega}_{2,(m)} = \frac{n \Omega_{(m)}(\hat{\alpha}_{(m)})}{(n - p_m - 2)^{1/2} (n - p_m - 4)^{1/2}}, \\ U_{(m),j}(\alpha_{(m)}) &= \frac{\partial \Omega_{(m)}^{-1}(\alpha_{(m)})}{\partial \alpha_{(m),j}}, \quad U_{(m),jl}(\alpha_{(m)}) = \frac{\partial^2 \Omega_{(m)}^{-1}(\alpha_{(m)})}{\partial \alpha_{(m),j} \partial \alpha_{(m),l}}, \\ U_{(m),jls}(\alpha_{(m)}) &= \frac{\partial^3 \Omega_{(m)}^{-1}(\alpha_{(m)})}{\partial \alpha_{(m),j} \partial \alpha_{(m),l} \partial \alpha_{(m),s}}, \quad \hat{U}_{(m),j} = U_{(m),j}(\hat{\alpha}_{(m)}), \\ \hat{U}_{(m),jl} &= U_{(m),jl}(\hat{\alpha}_{(m)}), \quad \hat{U}_{(m),jls} = U_{(m),jls}(\hat{\alpha}_{(m)}). \end{aligned}$$

进一步地, 对于所有 $m \in \{1, \dots, M\}$, 定义

$$\begin{aligned} \hat{\Pi}_{(m),1} &= \Pi_1(\hat{\Omega}_{1,(m)}, \mathbf{X}_{(m)}, \mathbf{y}) \\ &= \frac{2(n - p_m - 2)}{n} \operatorname{trace} \left\{ \mathbf{X}_{(m)} \frac{\partial \hat{\beta}_{(m)}}{\partial \mathbf{y}^T} \right\} \\ &\quad - \frac{2n}{n - p_m - 2} \sum_{j=1}^{q_m} \frac{\partial \hat{\alpha}_{(m),j}}{\partial \mathbf{y}^T} \hat{\Omega}_{(m)} \hat{U}_{(m),j} (\mathbf{y} - \mathbf{X}_{(m)} \hat{\beta}_{(m)}) \end{aligned} \quad (3.3)$$

以及

$$\begin{aligned} \hat{\Pi}_{(m),2} &= \Pi_2(\hat{\Omega}_{2,(m)}, \mathbf{X}_{(m)}, \mathbf{y}) \\ &= \frac{n^2}{(n - p_m - 2)(n - p_m - 4)} \left\{ \operatorname{trace} \left\{ \sum_{j=1}^{q_m} \hat{U}_{(m),j} \hat{\Omega}_{(m)} \frac{\partial^2 \hat{\alpha}_{(m),j}}{\partial \mathbf{y} \partial \mathbf{y}^T} \hat{\Omega}_{(m)} \right\} \right. \\ &\quad \left. + \sum_{j=1}^{q_m} \sum_{l=1}^{q_m} \frac{\partial \hat{\alpha}_{(m),l}}{\partial \mathbf{y}^T} \hat{\Omega}_{(m)} \hat{U}_{(m),l} \hat{\Omega}_{(m)} \frac{\partial \hat{\alpha}_{(m),j}}{\partial \mathbf{y}} \right\}. \end{aligned} \quad (3.4)$$

对于 $m \in \{1, \dots, M\}$ 和 $j, l, s \in \{1, \dots, q_m\}$, 记

$$\begin{aligned} \mathbf{H}_{(m)}(\alpha_{(m)}) &= \mathbf{X}_{(m)} (\mathbf{X}_{(m)}^T \Omega_{(m)}^{-1}(\alpha_{(m)}) \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T, \\ \mathbf{M}_{(m),j}(\alpha_{(m)}) &= (\mathbf{I}_n - \Omega_{(m)}^{-1}(\alpha_{(m)}) \mathbf{H}_{(m)}(\alpha_{(m)})) \mathbf{U}_{(m),j}(\alpha_{(m)}) (\mathbf{I}_n - \mathbf{H}_{(m)}(\alpha_{(m)}) \Omega_{(m)}^{-1}(\alpha_{(m)})), \\ \mathbf{M}_{(m),jl}(\alpha_{(m)}) &= \frac{\partial \mathbf{M}_{(m),j}(\alpha_{(m)})}{\partial \alpha_{(m),l}}, \\ \mathbf{M}_{(m),jls}(\alpha_{(m)}) &= \frac{\partial \mathbf{M}_{(m),jl}(\alpha_{(m)})}{\partial \alpha_{(m),s}}. \end{aligned}$$

此外, 记

$$\begin{aligned}\hat{\mathbf{H}}_{(m)} &= \mathbf{H}_{(m)}(\hat{\boldsymbol{\alpha}}_{(m)}), \quad \hat{\mathbf{M}}_{(m),j} = \mathbf{M}_{(m),j}(\hat{\boldsymbol{\alpha}}_{(m)}), \\ \hat{\mathbf{M}}_{(m),jl} &= \mathbf{M}_{(m),jl}(\boldsymbol{\alpha}_{(m)}), \quad \hat{\mathbf{M}}_{(m),jls} = \mathbf{M}_{(m),jls}(\hat{\boldsymbol{\alpha}}_{(m)}).\end{aligned}$$

用 $\mathbf{B}_{(m)}(\boldsymbol{\alpha}_{(m)})$ 表示一个 $q_m \times q_m$ 维矩阵, 其 (j, l) 元为

$$b_{(m),jl}(\boldsymbol{\alpha}_{(m)}) = \text{trace} \left\{ \mathbf{U}_{(m),j}(\boldsymbol{\alpha}_{(m)}) \frac{\partial \boldsymbol{\Omega}_{(m)}(\boldsymbol{\alpha}_{(m)})}{\partial \alpha_{(m),l}} \right\} + \text{trace} \{ \mathbf{U}_{(m),jl}(\boldsymbol{\alpha}_{(m)}) \boldsymbol{\Omega}_{(m)}(\boldsymbol{\alpha}_{(m)}) \}.$$

用 $\mathbf{D}_{(m)}(\boldsymbol{\alpha}_{(m)}, \mathbf{y})$ 表示一个 $q_m \times q_m$ 维矩阵, 其 (j, l) 元为 $\mathbf{y}^T \mathbf{M}_{(m),jl}(\boldsymbol{\alpha}_{(m)}) \mathbf{y}$,

$$b_{(m),jls}(\boldsymbol{\alpha}_{(m)}) = \frac{\partial b_{(m),jl}(\boldsymbol{\alpha}_{(m)})}{\partial \alpha_{(m),s}}.$$

$\hat{b}_{(m),jls} = b_{(m),jls}(\hat{\boldsymbol{\alpha}}_{(m)})$, $\hat{\mathbf{B}}_{(m)} = \mathbf{B}(\hat{\boldsymbol{\alpha}}_{(m)})$, $\hat{\mathbf{D}}_{(m)} = \mathbf{D}(\hat{\boldsymbol{\alpha}}_{(m)}, \mathbf{y})$, 且 $\hat{\boldsymbol{\Delta}}_{(m)} = (\hat{\mathbf{B}}_{(m)} - \hat{\mathbf{D}}_{(m)})/n$. 最后, 使用 $\boldsymbol{\theta}_{(m)}^* = (\boldsymbol{\beta}_{(m)}^{*\text{T}}, \boldsymbol{\alpha}_{(m)}^{*\text{T}})^T$ 表示最小化第 m 个备选模型与数据生成过程间 Kullback-Leibler 离差的拟真 (pseudo-true) 值.

本文余下的部分假设下列所有正则性条件对于所有 m ($1 \leq m \leq M$) 均成立.

条件 3.1 存在两个常数 $c_0 > 0$ 和 $c_1 > 0$, 使得

(i) 对于充分大的 n , 有

$$c_0 \leq \lambda_{\min} \left(\frac{\mathbf{X}_{(m)}^T \mathbf{X}_{(m)}}{n} \right) \leq \lambda_{\max} \left(\frac{\mathbf{X}_{(m)}^T \mathbf{X}_{(m)}}{n} \right) \leq c_1;$$

(ii) 对于参数空间里的所有 $\boldsymbol{\alpha}_{(m)}$, 有

$$\begin{aligned}& \lambda_{\min}^{-1}(\boldsymbol{\Omega}_{(m)}(\boldsymbol{\alpha}_{(m)})), \quad \lambda_{\max}(\boldsymbol{\Omega}_{(m)}(\boldsymbol{\alpha}_{(m)})), \\ & \max_{j \in \{1, \dots, q_m\}} \|\mathbf{U}_{(m),j}(\boldsymbol{\alpha}_{(m)})\|, \quad \max_{j, l \in \{1, \dots, q_m\}} \|\mathbf{U}_{(m),jl}(\boldsymbol{\alpha}_{(m)})\|, \\ & \max_{j, l, s \in \{1, \dots, q_m\}} \|\mathbf{U}_{(m),jls}(\boldsymbol{\alpha}_{(m)})\|\end{aligned}$$

以及

$$\max_{j, l, s \in \{1, \dots, q_m\}} \left\| \frac{1}{n} b_{(m),jls}(\boldsymbol{\alpha}_{(m)}) \right\|$$

一致地小于或等于 c_1 .

条件 3.2 存在两个常数 $c_1 > 0$ 和 $\delta > 0$, 使得

(i) $\sup_{n \geq 1} \max_{i \in \{1, \dots, n\}} \mathbb{E}|e_{0i}|^{4+\delta} \leq c_1$, 其中 e_{0i} 是 $\mathbf{e}_0$ 的第 i 个分量;

(ii) $\sup_{n \geq 1} \|\boldsymbol{\mu}\|^2/n \leq c_1$;

(iii) $\lambda_{\min}^{-1}(\boldsymbol{\Omega}_0)$ 和 $\lambda_{\max}(\boldsymbol{\Omega}_0)$ 均小于或等于 c_1 .

条件 3.3 当 $n \rightarrow \infty$ 时, 存在一个 $q_m \times q_m$ 维的可逆常数矩阵 $\bar{\boldsymbol{\Delta}}_{(m)}$, 使得

$$\|\hat{\boldsymbol{\Delta}}_{(m)} - \bar{\boldsymbol{\Delta}}_{(m)}\| = o_p(1).$$

条件 3.4 存在固定的 $\delta > 0$, 使得

$$\mathbb{E}\|n^{1/2}(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{(m)}^*)\|^{3+\delta} = O(1).$$

在条件 3.1 中, (i) 的第一个不等式本质上确保了拟合模型是渐近可识别的, 而第二个不等式确保了 $\hat{\beta}_{(m)}$ 拥有有界的渐近方差. 这类条件是模型选择文献中的常见条件 (如文献 [11] 的假定 (A1') 和文献 [12] 的条件 1). 条件 3.1(ii) 对随机误差的协方差矩阵施加了一定的限制, 并且在特定的模型下是可以验证的. 条件 3.2 给出了一些基本的矩条件假设. 条件 3.3 规范化了矩阵 $\hat{B}_{(m)}$ 的 $\hat{D}_{(m)}$ 行为. 该条件保证了在极限意义下, $\hat{D}_{(m)}$ 的列不会落在 $\hat{B}_{(m)}$ 的列空间中. 条件 3.4 是严格证明模型选择准则的渐近有效性的标准条件, 并常常被用于模型选择的文献中 (如文献 [11] 的条件 R6 和文献 [13] 的条件 6). 在本文的背景下, 条件 3.4 可以进一步化为更本质的条件. 记

$$\psi_{(m)}(\theta) = \begin{pmatrix} \mathbf{X}^T \Omega^{-1}(\alpha)(\mathbf{y} - \mathbf{X}\beta) \\ (\mathbf{y} - \mathbf{X}\hat{\beta})^T U_1(\hat{\alpha})(\mathbf{y} - \mathbf{X}\hat{\beta}) - \text{trace}\{U_1(\hat{\alpha})\Omega(\hat{\alpha})\} \\ \vdots \\ (\mathbf{y} - \mathbf{X}\hat{\beta})^T U_{q_m}(\hat{\alpha})(\mathbf{y} - \mathbf{X}\hat{\beta}) - \text{trace}\{U_{q_m}(\hat{\alpha})\Omega(\hat{\alpha})\} \end{pmatrix}.$$

由 $\hat{\theta}_{(m)}$ 的定义可知, $\psi_{(m)}(\hat{\theta}_{(m)}) = \mathbf{0}_{(p_m+q_m) \times 1}$, 因此,

$$n^{1/2}(\hat{\theta}_{(m)} - \theta_{(m)}^*) = \left(\frac{1}{n} \frac{\partial \psi_{(m)}(\theta_{(m)})}{\partial \theta_{(m)}} \bigg|_{\theta_{(m)} = \tilde{\theta}_{(m)}^*} \right)^{-1} \frac{\psi_{(m)}(\theta_{(m)}^*)}{n^{1/2}},$$

其中 $\tilde{\theta}_{(m)}^*$ 落在 $\hat{\theta}_{(m)}$ 与 $\theta_{(m)}^*$ 之间. 上述论证表明, 如果

$$\mathbb{E} \left\| \left(\frac{1}{n} \frac{\partial \psi_{(m)}(\theta_{(m)})}{\partial \theta_{(m)}^T} \bigg|_{\theta_{(m)} = \tilde{\theta}_{(m)}^*} \right)^{-1} \right\|^{2(3+\delta)} \quad \text{和} \quad \mathbb{E} \| n^{-1/2} \psi_{(m)}(\theta_{(m)}^*) \|^2(3+\delta)$$

均关于 n 一致有界, 则条件 3.4 自然成立.

当所有的备选模型均存在设定偏误时, 模型选择的目的变为如何选择能够提供更为精确预测的模型 [11, 12]. 考虑下列 Kullback-Leibler 型损失函数度量预测误差:

$$\begin{aligned} L_{\text{KL}}(m) &= 2\mathbb{E}_{\mathbf{y}^*} \{ \log f_{\text{true}}(\mathbf{y}^*) - \log \text{likelihood}(\mathbf{y}^*, \mathbf{X}_{(m)}, \hat{\beta}_{(m)}, \hat{\Omega}_{(m)}) \} \\ &= \bar{c}_0 + n \log(2\pi) + \log |\hat{\Omega}_{(m)}| + (\boldsymbol{\mu} - \mathbf{X}_{(m)} \hat{\beta}_{(m)})^T \hat{\Omega}_{(m)}^{-1} (\boldsymbol{\mu} - \mathbf{X}_{(m)} \hat{\beta}_{(m)}) + \text{trace}(\hat{\Omega}_{(m)}^{-1} \Omega_0), \end{aligned}$$

其中 $f_{\text{true}}(\cdot)$ 表示真实数据生成过程的概率密度函数,

$$\bar{c}_0 = \mathbb{E}_{\mathbf{y}^*} \{ -2 \log f_{\text{true}}(\mathbf{y}^*) \}.$$

Kullback-Leibler 型损失函数在模型选择的文献中应用非常广泛 (如文献 [11, 12]). 此外, $L_{\text{KL}}(m)$ 在方差分量参数已知时退化到二次损失函数. 记 $R_{\text{KL}}(m) = \mathbb{E}_{\mathbf{y}} \{ L_{\text{KL}}(m) \}$ 为第 m 个备选模型的风险函数. 为证明 $\hat{m}$ 的渐近有效性, 我们进一步引入下列正则性条件.

条件 3.5 当 $n \rightarrow \infty$ 时, 对于某个正整数 k , 所有备选模型的损失函数满足

$$\sum_{m=1}^M \left(\frac{R_{\text{KL}}(m)}{n^{1/2}} \right)^{-k} \rightarrow 0.$$

条件 3.5 排除了至少存在一个正确设定的备选模型的情形, 类似的条件也在其他研究模型选择和模型平均的文献中使用过 (如文献 [12, 14]). 附录 A.2 提供了关于此条件的进一步讨论, 展示了在有具体分布假定的前提下该条件可以被关于二次损失函数的条件所推出. 现在需要引入一个引理给出 $\hat{\Pi}_{(m),1}$ 和 $\hat{\Pi}_{(m),2}$ 的具体阶数.

引理 3.1 (模型 (3.1) 的有效自由度) 如果数据产生自 $\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}_0$, 其中 $\mathbf{e}_0 \sim (\mathbf{0}_{n \times 1}, \Omega_0)$, 且条件 3.1–3.3 均得到满足, 则对于所有 $1 \leq m \leq M$, 有

$$\hat{\Pi}_{(m),1} = O_p(p_m + q_m^2), \quad \hat{\Pi}_{(m),2} = O_p(q_m^4).$$

该引理表明, AICc^* 中惩罚项的阶数为

$$O_p(p_m + q_m^2 + q_m^4) = O_p(1).$$

引理 3.1 的证明在附录 A.1.4 给出.

定理 3.1 如果数据由 $\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}_0$ 产生, 其中

$$\mathbf{e}_0 \sim (\mathbf{0}_{n \times 1}, \Omega_0), \quad \sup_{1 \leq m \leq M} \|\boldsymbol{\theta}_{(m)}^*\| = O(1),$$

且条件 3.1–3.5 满足, 则当 $n \rightarrow \infty$ 时,

$$\frac{L_{\text{KL}}(\hat{m})}{\inf_{1 \leq m \leq M} L_{\text{KL}}(m)}$$

依概率收敛到 1.

定理 3.1 的证明在附录 A.1.5 给出. 该结果表明 AICc^* 是渐近有效的, 即当样本量足够大时, 最小化 AICc^* 选出的模型的预测损失同未知的最优模型的预测损失间的差别是可忽略的. 下一节将使用 Monte Carlo 模拟研究 AICc^* 在有限样本下的表现.

4 模拟研究

模拟 4.1 在本模拟研究中, 我们使用带有空间自回归误差的空间回归模型生成数据: $\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}_0$, $\boldsymbol{\mu} = \mathbf{X}_0 \boldsymbol{\beta}_0$, $\mathbf{X}_0 = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, $\mathbf{x}_i = (x_{i1}, \dots, x_{i6})^T$, $x_{i1} = 1$, 对于所有 $j = 2, \dots, 6$, $x_{ij} \sim \text{Normal}(0, 1)$ 且独立, $\boldsymbol{\beta}_0 = (0.3, -0.2, 0.5, -0.8, -0.4, 0.1)^T$, $\mathbf{e}_0 = \rho_0 \mathbf{W}_{n,0} \mathbf{e}_0 + \boldsymbol{\epsilon}_0$ 且 $\boldsymbol{\epsilon}_0 \sim \text{Normal}(\mathbf{0}_{n \times 1}, \mathbf{I}_n)$. 与文献 [15, 第 4.1 小节] 一致, 我们产生随机的组交互矩阵 (random group interaction matrix) $\mathbf{W}_{n,0}$, 并确保其中的平均组规模 (average group size) 近似为 $n^{1/2}$.

使用下列空间自回归误差模型作为拟合模型:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} = \rho \mathbf{W}_n \mathbf{e} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}_{n \times 1}, \sigma^2 \mathbf{I}_n), \quad (4.1)$$

其中 $\mathbf{W}_n$ 是空间权重矩阵, ρ 是未知空间自回归参数. 在拟合模型中, 考虑 3 种空间设定, 引入 3 种空间权重矩阵 $\mathbf{W}_n$:

- (i) $\mathbf{W}_n = \mathbf{0}_{n \times n}$, 即没有任何空间结构;
- (ii) $\mathbf{W}_n = \mathbf{W}_{n,0}$, 即空间结构与真实模型的一致;
- (iii) $\mathbf{W}_n = \mathbf{I}_k \otimes (\mathbf{1}_{r \times 1} \mathbf{1}_{r \times 1}^T - \mathbf{I}_r) / (r - 1)$, 其中 $k = r = n^{1/2}$, 该设定与文献 [16] 中一致.

附录 A.3.2 给出了模型 (4.1) 中 AICc^* 的具体计算方法.

常数项 x_{i1} 总是被包含在所有拟合模型中, 而 x_{i2} 、 x_{i3} 和 x_{i4} 则被考虑为备选协变量, 这些变量有可能被包含也有可能不被包含在拟合模型里. x_{i5} 和 x_{i6} 被故意排除在所有备选模型之外, 以此模拟模型设定存在偏误时的情形. 因此, 考虑所有协变量和空间结构的可能组合, 我们一共有 $2^3 \times 3 = 24$

个备选模型, 我们将考虑协变量和空间权重矩阵的联合选择问题. 设置 $n \in \{16, 36, 64, 100\}$ 且 $\rho_0 \in \{-0.8, -0.7, \dots, 0.8\}$. 在每个参数环境下, 重复 1,000 次模拟. 当样本量很小时, 有时参数估计过程会遭遇不收敛的情形, 遭遇这类情形时我们会将这些不收敛的模拟舍弃.

使用 Kullback-Leibler 型损失函数评价 (3.2) 中提出的 $AICc^*$ 的表现, 同时我们也将其与常规 AIC 和常规 BIC 进行比较. 图 1 给出了在不同的 n 和 ρ_0 组合下的平均 Kullback-Leibler 型损失. 从图 1 中可以发现, 当 $|\rho_0| \leq 0.4$ 且 $n \leq 64$ 时, 我们所提出的方法有最好的表现. 当 n 和 $|\rho_0|$ 逐渐变大时, 我们提出的 $AICc^*$ 与常规 AIC 和常规 BIC 的表现渐渐趋同, 我们还发现常规 BIC 在绝大多数场合的平均 Kullback-Leibler 型损失均较大.

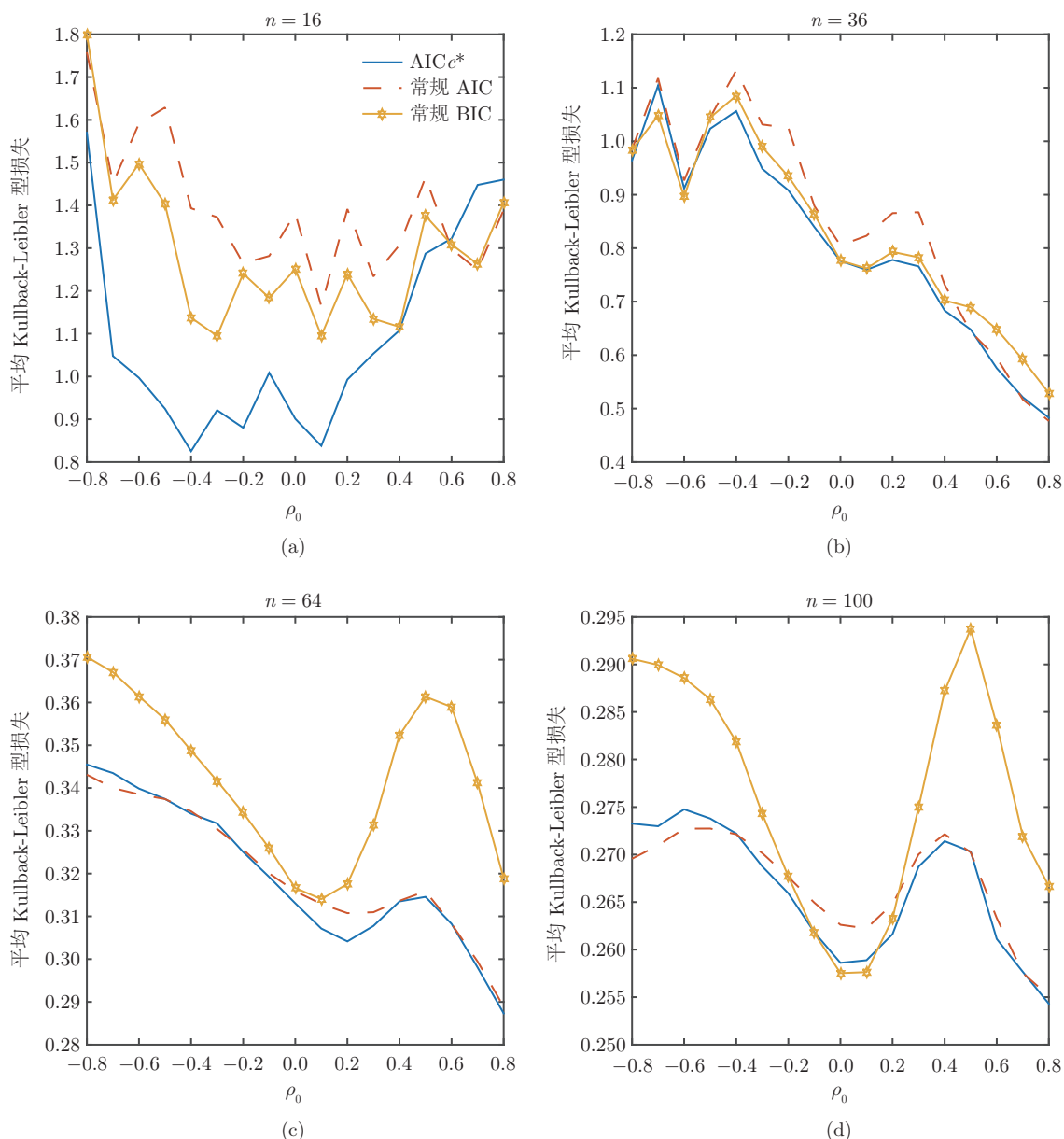


图 1 模拟 4.1 的结果: $AICc^*$ (蓝色实线)、常规 AIC (红色短划线) 以及常规 BIC (黄色带六芒星实线) 的平均 Kullback-Leibler 型损失

为了提供进一步的分析, 我们考虑数据生成过程和备选模型集合之间的离差与 ρ_0 间的相对关系. 使用 $D_{\text{KL}}(\boldsymbol{\theta}_{(m)})$ 表示第 m 个备选模型与真实的数据生成过程之间的 Kullback-Leibler 离差. 考虑下列特征度量数据生成过程和所使用的备选模型集合间的相似度:

$$\tau^* = \inf_{1 \leq m \leq M} D_{\text{KL}}(\boldsymbol{\theta}_{(m)}^*),$$

经过 $1/n$ 单位化的不同的 τ^* 值在表 1 中给出.

综合考虑图 1 和表 1 中的结果, 可以看出当给定 n 时, 我们提出的方法在 τ^* 相对较小时 (即当数据生成过程与备选模型之间差异较小时) 能给出比其竞争者更好的预测效果. 当 τ^* 变大时 (即数据生成过程与备选模型之间的差异变得较为明显时), AICc^* 与常规 AIC 的表现非常类似. 图 1 和表 1 中的结果表明, 当数据生成过程与备选模型之间的相似度较高时, 现有方法在选择合适的模型时会遇到困难, 而我们提出的方法可以有效克服此困难.

模拟 4.2 在本模拟研究中, 我们使用带有空间自回归误差的空间回归模型生成数据, 但是均值是非线性结构, 即 $\boldsymbol{\mu}$ 的第 i 个分量为

$$\mu_i = \beta_1 x_{i1} + \cdots + \beta_5 x_{i5} + \frac{1}{1 + \exp\{\beta_4(\beta_5 x_{i5} + \beta_6 x_{i6})\}},$$

回归系数、协变量及误差的分布设定与模拟 4.1 中的设定保持一致. 备选模型的设定也与模拟 4.1 中的设定保持一致. 在此环境下, 所有备选模型的均值部分的函数形式设定都存在偏误, 且忽略了 x_{i5} 和 x_{i6} . 图 2 给出了本设定下在不同的 n 和 ρ_0 组合下的平均 Kullback-Leibler 型损失. 表 2 则给出了经过 $1/n$ 单位化的不同的 τ^* 值. 图 2 和表 2 中的结果表明, 在 $n = 16$ 时, 本文提出的 AICc^* 依然有较为明显的优势. 特别地, 在 $n = 16$ 且 $\rho_0 \leq 0.4$ 时, AICc^* 的优势非常明显. 而当样本量增大且 τ^* 较大

表 1 模拟 4.1 的结果: τ^* 的值 (被 $1/n$ 标准化)

ρ	$n = 16$	$n = 36$	$n = 64$	$n = 100$
-0.8	0.0998	0.1625	0.1522	0.1582
-0.7	0.1003	0.1606	0.1490	0.1581
-0.6	0.1011	0.1589	0.1460	0.1580
-0.5	0.1023	0.1572	0.1432	0.1581
-0.4	0.1036	0.1557	0.1405	0.1582
-0.3	0.1052	0.1543	0.1380	0.1584
-0.2	0.1070	0.1531	0.1357	0.1587
-0.1	0.1091	0.1506	0.1337	0.1591
0.0	0.1102	0.1509	0.1320	0.1597
0.1	0.1143	0.1510	0.1306	0.1605
0.2	0.1174	0.1510	0.1295	0.1614
0.3	0.1210	0.1513	0.1289	0.1626
0.4	0.1252	0.1521	0.1287	0.1640
0.5	0.1299	0.1534	0.1289	0.1656
0.6	0.1353	0.1553	0.1297	0.1676
0.7	0.1415	0.1577	0.1310	0.1698
0.8	0.1484	0.1608	0.1329	0.1724

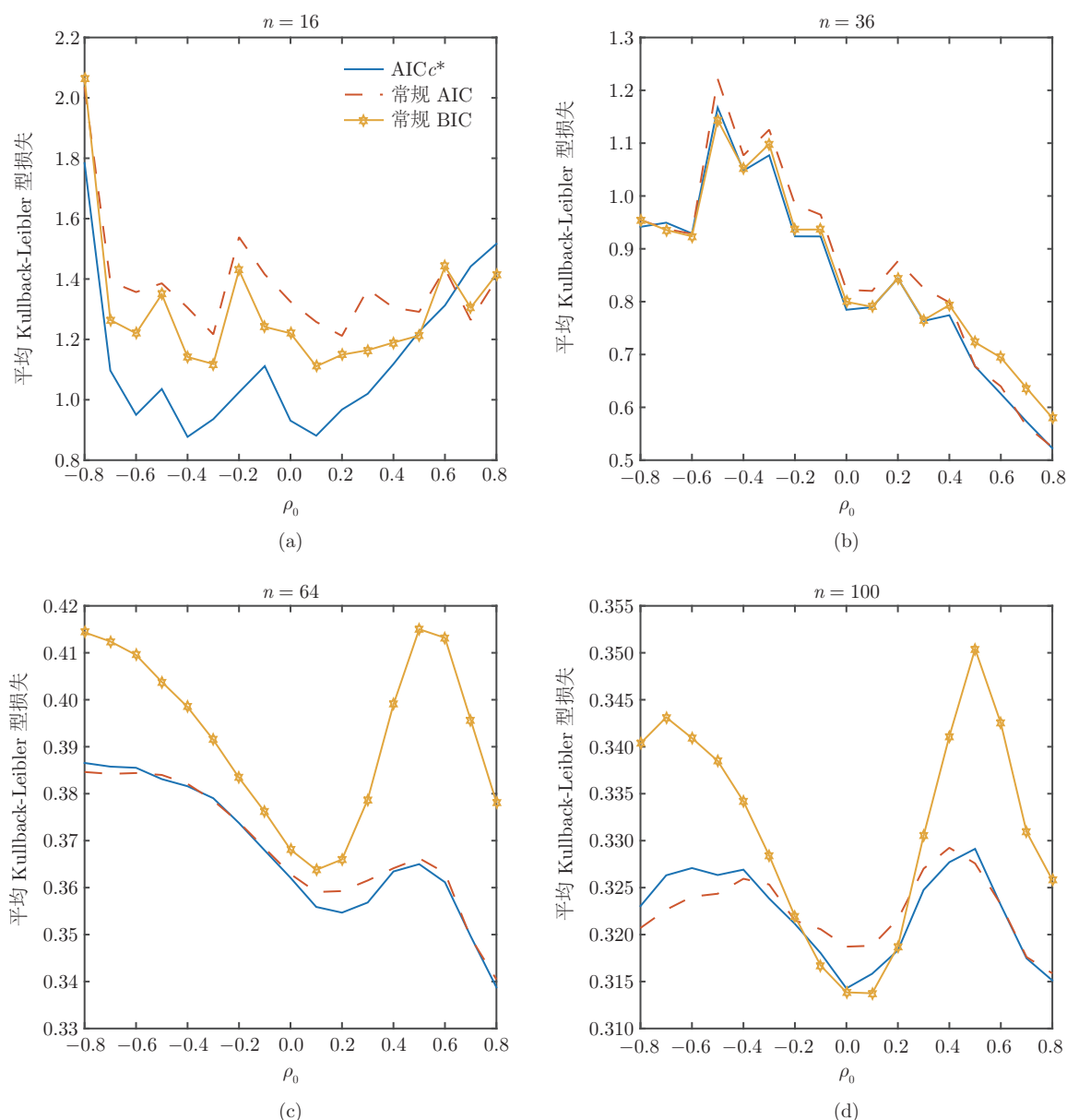


图 2 模拟 4.2 的结果: AICc* (蓝色实线)、常规 AIC (红色短划线) 以及常规 BIC (黄色带六芒星实线) 的平均 Kullback-Leibler 型损失

时, AICc* 与常规 AIC 有非常类似的表现. 而常规 BIC 的表现相对而言波动较大, 在一些场合下其表现好于 AICc*, 但是另外一些场合下常规 BIC 总是给出最大的预测误差.

模拟 4.3 在本模拟研究中, 我们依旧使用带有空间自回归误差的空间回归模型生成数据, 但是 ϵ_0 的每个分量用自由度为 8 的 t -分布独立产生. 其他数据生成过程和备选模型的设定与模拟 4.1 中的设定保持一致. 在此设定下, 所有备选模型的均值部分和误差分布部分均存在设定偏误. 表 3 给出了经过 $1/n$ 单位化的不同的 τ^* 值. 图 3 给出了本设定下在不同的 n 和 ρ_0 组合下的平均 Kullback-Leibler 型损失. 观察图 3 和表 3 中的结果可知, 在样本量较小且 τ^* 较小时, 本文提出的 AICc* 有较为明显的优势. 而当样本量增大且 τ^* 较大时, AICc* 与常规 AIC 有非常类似的表现. 而常规 BIC 表现出的

表 2 模拟 4.2 的结果: τ^* 的值 (被 $1/n$ 标准化)

ρ	$n = 16$	$n = 36$	$n = 64$	$n = 100$
-0.8	0.1352	0.2081	0.1869	0.2150
-0.7	0.1372	0.2059	0.1851	0.2152
-0.6	0.1395	0.2037	0.1833	0.2155
-0.5	0.1421	0.2017	0.1817	0.2158
-0.4	0.1449	0.1997	0.1803	0.2162
-0.3	0.1479	0.1976	0.1791	0.2166
-0.2	0.1512	0.1938	0.1781	0.2172
-0.1	0.1547	0.1920	0.1774	0.2179
0.0	0.1559	0.1929	0.1769	0.2187
0.1	0.1628	0.1932	0.1768	0.2197
0.2	0.1675	0.1928	0.1770	0.2210
0.3	0.1728	0.1929	0.1776	0.2224
0.4	0.1787	0.1936	0.1787	0.2242
0.5	0.1855	0.1949	0.1802	0.2263
0.6	0.1931	0.1968	0.1822	0.2288
0.7	0.2017	0.1996	0.1846	0.2316
0.8	0.2114	0.2031	0.1877	0.2350

表 3 模拟 4.3 的结果: τ^* 的值 (被 $1/n$ 标准化)

ρ	$n = 16$	$n = 36$	$n = 64$	$n = 100$
-0.8	0.1064	0.1558	0.1485	0.1518
-0.7	0.1068	0.1541	0.1457	0.1516
-0.6	0.1074	0.1526	0.1431	0.1515
-0.5	0.1083	0.1511	0.1406	0.1514
-0.4	0.1093	0.1498	0.1383	0.1515
-0.3	0.1105	0.1486	0.1362	0.1516
-0.2	0.1119	0.1476	0.1343	0.1519
-0.1	0.1135	0.1461	0.1326	0.1522
0.0	0.1146	0.1459	0.1312	0.1527
0.1	0.1176	0.1459	0.1301	0.1533
0.2	0.1200	0.1459	0.1293	0.1541
0.3	0.1229	0.1462	0.1288	0.1550
0.4	0.1262	0.1469	0.1286	0.1561
0.5	0.1300	0.1480	0.1289	0.1575
0.6	0.1342	0.1495	0.1295	0.1590
0.7	0.1391	0.1515	0.1306	0.1608
0.8	0.1445	0.1539	0.1321	0.1628

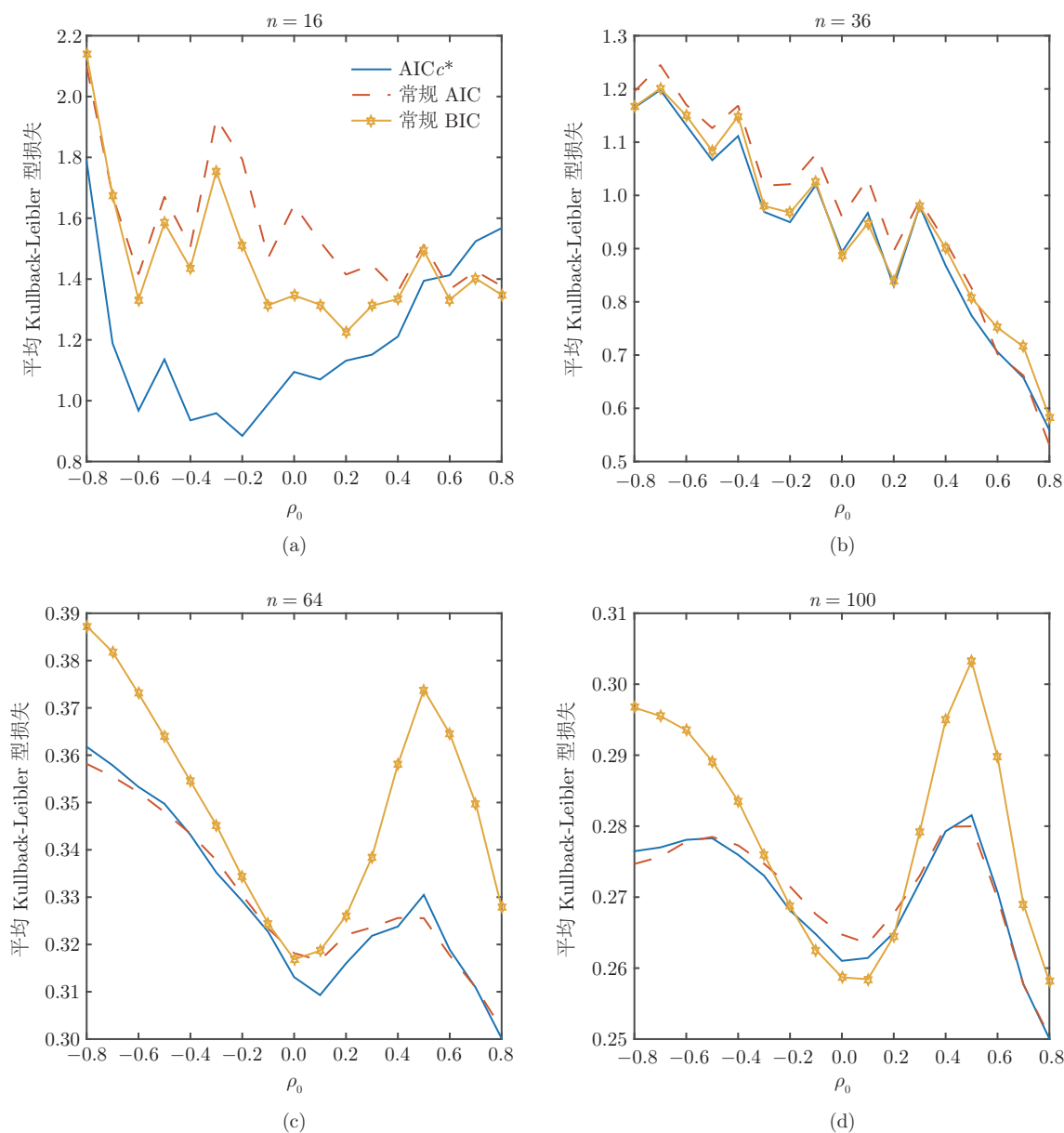


图 3 模拟 4.3 的结果: AIC_c^* (蓝色实线)、常规 AIC (红色短划线) 以及常规 BIC (黄色带六芒星实线) 的平均 Kullback-Leibler 型损失

模型选择模式与其在模拟 4.2 中的表现类似, 其在不同的 ρ_0 水平下预测误差的波动水平较 AIC_c^* 和常规 AIC 更加剧烈.

5 实际数据分析

本节使用一组实际数据展示所提出的方法. 社区犯罪数据包含了 Ohio 州的 Columbus 市 49 个相邻的规划社区的数据 (完整数据集在文献 [17] 中第 189 页的表 12.1 中). 响应变量为社区内每千户中的合计入室盗窃和机动车盗窃量. 该数据集有两个协变量: 收入 (INC) 和住宅价值 (HOUSE),

表 4 社区犯罪数据分析结果: $\hat{R}_{\text{KL}}(i, \hat{m}_i)$ 的样本均值和样本 25%、50% 和 75% 分位数, $\hat{m}_i$ 表示第 i 次重复下由 AICc*, 常规 AIC 或者常规 BIC 选出的模型

	$r_{\text{test}} = 0.7$			$r_{\text{test}} = 0.5$			$r_{\text{test}} = 0.3$		
	AICc*	常规 AIC	常规 BIC	AICc*	常规 AIC	常规 BIC	AICc*	常规 AIC	常规 BIC
均值	1.603	1.699	1.537	0.672	0.680	0.703	0.691	0.692	0.795
25%	0.274	0.279	0.268	0.143	0.134	0.180	0.037	0.035	0.076
50%	0.675	0.713	0.684	0.423	0.423	0.452	0.257	0.243	0.349
75%	1.714	1.783	1.648	0.912	0.922	0.958	0.947	0.973	1.155

我们将这些变量进行了中心化和标准化. 空间相关关系使用社区间的一阶邻接关系 (the first order contiguity), 相关信息可以在文献 [17] 中第 190 页的表 12.2 中找到. 应用模型 (4.1) 来分析数据, 我们使用行正则化权重矩阵 (row normalized weight matrix) 建模不同社区间潜在的空间相关关系. 常数项总是出现在不同备选模型中. INC 和 HOUSE 为各模型中的候选协变量. 我们还考虑一种特殊情形 $\rho \equiv 0$, 即没有空间相关关系. 因此, 一共有 $2 \times 2^2 = 8$ 种备选模型. 为了评价不同方法的表现, 原始数据集被随机分为训练集和测试集, 其中测试数据所占比例为 r_{test} . 基于训练集, 使用极大似然估计给出未知参数, 并使用我们提出的 AICc* 及其竞争者: 常规 AIC 和常规 BIC 选择模型. 上述过程被重复 1,000 次. 在第 i ($i = 1, \dots, 1000$) 次重复中, 第 m 个备选模型的预测误差由

$$\hat{L}_{\text{KL}}(i, m) = n_{\text{test}} \log(2\pi) + \log |\hat{\Omega}_{(m)}^{(i)}| + (\mathbf{y}_{\text{test}}^{(i)} - \mathbf{X}_{\text{test},(m)}^{(i)} \hat{\beta}_{(m)}^{(i)})^T (\hat{\Omega}_{(m)}^{(i)})^{-1} (\mathbf{y}_{\text{test}}^{(i)} - \mathbf{X}_{\text{test},(m)}^{(i)} \hat{\beta}_{(m)}^{(i)})$$

度量, 其中 $\mathbf{y}_{\text{test}}^{(i)}$ (或 $\mathbf{X}_{\text{test},(m)}^{(i)}$) 是测试集中对应的响应变量 (或协变量矩阵), $n_{\text{test}} = \text{floor}(r_{\text{test}} \times n)$ 是测试样本量, $\hat{\beta}_{(m)}^{(i)}$ 和 $\hat{\Omega}_{(m)}^{(i)}$ 基于训练数据获得. 结合附录 A.2 中的讨论可知, $\hat{L}_{\text{KL}}(i, m)$ 本质上度量了第 m 个模型预测第 i 个测试样本的损失, 这种损失的下界控制了常用的二次损失. 由于在实际数据分析中, 真实的损失函数是未知的, 我们通过下列标准化后的相对损失度量第 i 次重复下的预测误差:

$$\hat{R}_{\text{KL}}(i, \hat{m}_i) = \frac{\hat{L}_{\text{KL}}(i, \hat{m}_i) - \min_{1 \leq m \leq M} \hat{L}_{\text{KL}}(i, m)}{n_{\text{test}}}, \quad i = 1, \dots, 1000,$$

其中 $\hat{m}_i$ 表示第 i 次重复下由 AICc*、常规 AIC 或者常规 BIC 选出的模型. 在不同的 r_{test} 下, $\hat{R}_{\text{KL}}(i, \hat{m}_i)$ 的样本均值和样本 25%、50% 和 75% 分位数在表 4 中给出.

表 4 中的结果表明没有一种方法一致地胜过其他方法. 如果考虑平均相对损失, AICc* 在 $r_{\text{test}} = 0.5$ 和 $r_{\text{test}} = 0.3$ 时表现最好, 并有最小的 75% 分位数. AICc* 给出的相对损失在 $r_{\text{test}} = 0.7$ 和 $r_{\text{test}} = 0.5$ 时有最小的 50% 分位数. 此外, 常规 AIC 给出的相对损失在 $r_{\text{test}} = 0.5$ 和 $r_{\text{test}} = 0.3$ 时有最小的中位数, 但是在 $r_{\text{test}} = 0.7$ 时无论以何种标准, 常规 AIC 的预测表现均是最差的. 常规 BIC 在 $r_{\text{test}} = 0.7$ 时有最小的平均预测损失, 但是在 $r_{\text{test}} = 0.5$ 和 $r_{\text{test}} = 0.3$ 时预测精度最差. 这些现象表明, 在本组数据中, AICc* 在不同设定下的表现是最稳定的, 这进一步支持了我们的方法在实际数据分析中的应用.

6 总结

在本文中, 虽然 AICc* 是在极大似然估计的框架下推导出的, 但是我们的理论框架可以用于任意合理的估计方法 (如限制极大似然估计或者调整极大似然估计及其高阶纠偏版本^[18]). 本文的研究聚

焦于线性模型, 且 $AICc^*$ 在 Stein 恒等式的框架下给出. 我们严格地研究了 $AICc^*$ 的渐近性质. 事实上, 在计数数据模型中, 作为 Stein 恒等式在 Poisson 分布下的推广, Stein-Chen 恒等式^[19] 也被用于导出模型选择准则^[20, 21]. 如何研究这些方法的渐近性质还是一个公开的问题, 我们考虑该问题作为未来的研究方向. 此外, 模型选择后推断 (post-selection inference) 并未在本文中得到研究. 最近, 文献 [22] 给出了常规 AIC 框架下的选择后推断的理论结果. 我们也将 $AICc^*$ 下的选择后推断作为我们未来的研究方向.

致谢 作者感谢两位匿名审稿人提出的非常有洞察力和建设性的意见, 这些意见极大地提升了本文初稿的质量.

参考文献

- 1 Zhang X, Liang H, Liu A, et al. Selection strategy for covariance structure of random effects in linear mixed-effects models. *Scand J Statist*, 2016, 43: 275–291
- 2 Konishi S, Kitagawa G. *Information Criteria and Statistical Modeling*. New York: Springer, 2008
- 3 Yang Y, Barron A R. An asymptotic property of model selection criteria. *IEEE Trans Inform Theory*, 1998, 44: 95–116
- 4 Hurvich C M, Tsai C L. Regression and time series model selection in small samples. *Biometrika*, 1989, 76: 297–307
- 5 Hurvich C M, Simonoff J S, Tsai C L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J R Stat Soc Ser B Stat Methodol*, 1998, 60: 271–293
- 6 Reschenhofer E. Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification. *Econom Theory*, 1999, 15: 377–387
- 7 Stein C M. Estimation of the mean of a multivariate normal distribution. *Ann Statist*, 1981, 9: 1135–1151
- 8 Liang H, Wu H L, Zou G. A note on conditional AIC for linear mixed-effects models. *Biometrika*, 2008, 95: 773–778
- 9 Greven S, Kneib T. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, 2010, 97: 773–789
- 10 Yanagihara H, Fujisawa H. Iterative bias correction of the cross-validation criterion. *Scand J Statist*, 2012, 39: 116–130
- 11 Flynn C J, Hurvich C M, Simonoff J S. Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *J Amer Statist Assoc*, 2013, 108: 1031–1043
- 12 Yu D, Zhang X, Yau K K W. Asymptotic properties and information criteria for misspecified generalized linear mixed models. *J R Stat Soc Ser B Stat Methodol*, 2018, 80: 817–836
- 13 Lv J, Liu J S. Model selection principles in misspecified models. *J R Stat Soc Ser B Stat Methodol*, 2014, 76: 141–167
- 14 Ando T, Li K-C. A model-averaging approach for high-dimensional regression. *J Amer Statist Assoc*, 2014, 109: 254–265
- 15 Yang Z. A robust LM test for spatial error components. *Reg Sci Urban Econ*, 2010, 40: 299–310
- 16 Lee L F. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 2004, 72: 1899–1925
- 17 Anselin L. *Spatial Econometrics: Methods and Models*. Dordrecht: Springer, 1988
- 18 Yu D, Bai P, Ding C. Adjusted quasi-maximum likelihood estimator for mixed regressive, spatial autoregressive model and its small sample bias. *Comput Statist Data Anal*, 2015, 87: 116–135
- 19 Chen L H Y. Poisson approximation for dependent trials. *Ann Probab*, 1975, 3: 534–545
- 20 Lian H. A note on conditional Akaike information for Poisson regression with random effects. *Electron J Stat*, 2012, 6: 1–9
- 21 Yu D. Conditional Akaike information criteria for a class of Poisson mixture models with random effects. *Scand J Statist*, 2016, 43: 1214–1235
- 22 Charkhi A, Claeskens G. Asymptotic post-selection inference for the Akaike information criterion. *Biometrika*, 2018, 105: 645–664
- 23 Kelejian H H, Prucha I R. On the asymptotic distribution of the Moran I test statistic with applications. *J Econometrics*, 2001, 104: 219–257
- 24 Yang Y. Combining forecasting procedures: Some theoretical results. *Econometric Theory*, 2004, 20: 176–222

附录 A

附录 A.1 理论证明

附录 A.1.1 定理 2.1 的证明

首先, 注意到

$$\begin{aligned} \text{AI} &= \text{E}[\text{E}_{\mathbf{y}^*}\{n \log 2\pi + \log |\boldsymbol{\Omega}(\hat{\boldsymbol{\alpha}})| + (\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}})\}] \\ &= \text{E}[n \log 2\pi + \log |\boldsymbol{\Omega}(\hat{\boldsymbol{\alpha}})| + (\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \text{trace}\{\boldsymbol{\Omega}_0\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})\}], \end{aligned} \quad (\text{A.1})$$

其中

$$\begin{aligned} &\text{E}\{(\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}})\} \\ &= \text{E}\{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\mu} - \mathbf{y})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\mu} - \mathbf{y})\} \\ &= \text{E}\{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\} + \text{E}\{(\mathbf{y} - \boldsymbol{\mu})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\mathbf{y} - \boldsymbol{\mu})\} \\ &\quad - 2\text{E}\{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\mathbf{y} - \boldsymbol{\mu})\}. \end{aligned} \quad (\text{A.2})$$

记 $\mathbf{A}(\hat{\boldsymbol{\alpha}}) = \boldsymbol{\Omega}_0^{1/2}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})\boldsymbol{\Omega}_0^{1/2}$, $\mathbf{z} = \boldsymbol{\Omega}_0^{-1/2}\mathbf{y}$, $\boldsymbol{\mu}_z = \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\mu}$. 然后, 由引理 2.1, 有

$$\begin{aligned} &\text{E}\{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\mathbf{y} - \boldsymbol{\mu})\} \\ &= \text{E}\{(\mathbf{z} - \boldsymbol{\Omega}_0^{-1/2}\mathbf{X}\hat{\boldsymbol{\beta}})^{\text{T}}\mathbf{A}(\hat{\boldsymbol{\alpha}})(\mathbf{z} - \boldsymbol{\mu}_z)\} \\ &= \text{E}\left\{\text{trace}\left\{\frac{\partial(\mathbf{A}(\hat{\boldsymbol{\alpha}})(\mathbf{z} - \boldsymbol{\Omega}_0^{-1/2}\mathbf{X}\hat{\boldsymbol{\beta}}))}{\partial \mathbf{z}^{\text{T}}}\right\}\right\} \quad (\text{使用引理 2.1}) \\ &= \text{E}\left[\text{trace}\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\} - \text{trace}\left\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\boldsymbol{\Omega}_0^{-1/2}\mathbf{X}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{z}^{\text{T}}}\right\}\right] \\ &\quad + \text{E}\left[\text{trace}\left\{\sum_{j=1}^q \left(\frac{\partial \mathbf{A}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j}(\mathbf{z} - \boldsymbol{\Omega}_0^{-1/2}\mathbf{X}\hat{\boldsymbol{\beta}})\frac{\partial \hat{\alpha}_j}{\partial \mathbf{z}^{\text{T}}}\right)\right\}\right] \\ &= \text{E}\left[\text{trace}\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\} - \text{trace}\left\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\boldsymbol{\Omega}_0^{-1/2}\mathbf{X}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}^{\text{T}}}\boldsymbol{\Omega}_0^{1/2}\right\}\right] \\ &\quad + \text{E}\left[\text{trace}\left\{\sum_{j=1}^q \left(\frac{\partial \mathbf{A}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j}(\boldsymbol{\Omega}_0^{-1/2}\mathbf{y} - \boldsymbol{\Omega}_0^{-1/2}\mathbf{X}\hat{\boldsymbol{\beta}})\frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}^{\text{T}}}\boldsymbol{\Omega}_0^{1/2}\right)\right\}\right] \\ &= \text{E}\left[\text{trace}\left\{\mathbf{A}(\hat{\boldsymbol{\alpha}}) - \hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\alpha}})\boldsymbol{\Omega}_0^{-1}\mathbf{X}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}^{\text{T}}}\right\} + \sum_{j=1}^q \left[\frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}^{\text{T}}}\boldsymbol{\Omega}_0\left(\frac{\partial \boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right)\right]\right] \\ &= \text{E}\left[\text{trace}\left\{\mathbf{A}(\hat{\boldsymbol{\alpha}}) - \hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\alpha}})\boldsymbol{\Omega}_0^{-1}\mathbf{X}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}^{\text{T}}}\right\} + \sum_{j=1}^q \left[\frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}^{\text{T}}}\boldsymbol{\Omega}_0(U_j(\hat{\boldsymbol{\alpha}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}))\right]\right] \end{aligned} \quad (\text{A.3})$$

以及

$$\begin{aligned} &\text{E}\{(\mathbf{y} - \boldsymbol{\mu})^{\text{T}}\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})(\mathbf{y} - \boldsymbol{\mu})\} \\ &= \text{E}\{(\mathbf{z} - \boldsymbol{\mu}_z)^{\text{T}}\mathbf{A}(\hat{\boldsymbol{\alpha}})(\mathbf{z} - \boldsymbol{\mu}_z)\} \\ &= \text{E}\left\{\text{trace}\left\{\frac{\partial(\mathbf{A}(\hat{\boldsymbol{\alpha}})(\mathbf{z} - \boldsymbol{\mu}_z))}{\partial \mathbf{z}^{\text{T}}}\right\}\right\} \quad (\text{使用引理 2.1}) \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E} \left[\text{trace}\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\} + \text{trace} \left\{ \sum_{j=1}^q \frac{\partial \mathbf{A}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j} (\mathbf{z} - \boldsymbol{\mu}_z) \frac{\partial \hat{\alpha}_j}{\partial \mathbf{z}^T} \right\} \right] \\
&= \mathbb{E} \left[\text{trace}\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\} + \sum_{j=1}^q \frac{\partial \hat{\alpha}_j}{\partial \mathbf{z}^T} \frac{\partial \mathbf{A}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j} (\mathbf{z} - \boldsymbol{\mu}_z) \right] \\
&= \mathbb{E} \left[\text{trace}\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\} + \sum_{j=1}^q \frac{\partial \frac{\partial \mathbf{A}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j} \frac{\partial \hat{\alpha}_j}{\partial \mathbf{z}^T}}{\partial \mathbf{z}^T} \right] \quad (\text{使用引理 2.1}) \\
&= \mathbb{E} \left[\text{trace}\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\} + \text{trace} \left\{ \sum_{j=1}^q \frac{\partial \mathbf{A}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j} \frac{\partial^2 \alpha_j}{\partial \mathbf{z} \partial \mathbf{z}^T} \right\} + \sum_{j=1}^q \sum_{j_1=1}^q \frac{\partial \hat{\alpha}_{j_1}}{\partial \mathbf{z}^T} \frac{\partial \mathbf{A}^2(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j \partial \hat{\alpha}_{j_1}} \frac{\partial \hat{\alpha}_j}{\partial \mathbf{z}} \right] \\
&= \mathbb{E} \left[\text{trace}\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\} + \text{trace} \left\{ \sum_{j=1}^q \boldsymbol{\Omega}_0^{1/2} \frac{\partial \boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j} \boldsymbol{\Omega}_0 \frac{\partial^2 \hat{\alpha}_j}{\partial \mathbf{y} \partial \mathbf{y}^T} \boldsymbol{\Omega}_0^{1/2} \right\} \right. \\
&\quad \left. + \sum_{j=1}^q \sum_{j_1=1}^q \frac{\partial \hat{\alpha}_{j_1}}{\partial \mathbf{y}^T} \boldsymbol{\Omega}_0 \frac{\partial^2 \boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}})}{\partial \hat{\alpha}_j \partial \hat{\alpha}_{j_1}} \boldsymbol{\Omega}_0 \frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}} \right] \\
&= \mathbb{E} \left[\text{trace}\{\mathbf{A}(\hat{\boldsymbol{\alpha}})\} + \text{trace} \left\{ \sum_{j=1}^q \mathbf{U}_j(\hat{\boldsymbol{\alpha}}) \boldsymbol{\Omega}_0 \frac{\partial^2 \hat{\alpha}_j}{\partial \mathbf{y} \partial \mathbf{y}^T} \boldsymbol{\Omega}_0 \right\} + \sum_{j=1}^q \sum_{j_1=1}^q \frac{\partial \hat{\alpha}_{j_1}}{\partial \mathbf{y}^T} \boldsymbol{\Omega}_0 \mathbf{U}_{jj_1}(\hat{\boldsymbol{\alpha}}) \boldsymbol{\Omega}_0 \frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}} \right].
\end{aligned}$$

基于上述结果, 我们证明了定理 2.1.

附录 A.1.2 (2.6) 的证明

由于在 (2.6) 的设定下, $\boldsymbol{\Omega}(\boldsymbol{\alpha}) = \sigma^2 \mathbf{I}_n$, 因此有

$$\mathbf{U}_j(\hat{\boldsymbol{\alpha}}) = \mathbf{I}_n, \quad \mathbf{U}_{jj_1}(\hat{\boldsymbol{\alpha}}) = 0. \quad (\text{A.4})$$

由极大似然估计的定义可知, $\hat{\sigma}^2 = \mathbf{y}\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T\} \mathbf{y}/n$ 且 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, 因此,

$$\frac{\partial \hat{\boldsymbol{\beta}}^T}{\partial \mathbf{y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1}, \quad (\text{A.5})$$

$$\frac{\partial \hat{\sigma}^{-2}}{\partial \mathbf{y}} = -\frac{2\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T\} \mathbf{y}}{n \hat{\sigma}^4} \quad (\text{A.6})$$

以及

$$\begin{aligned}
\frac{\partial^2 \hat{\sigma}^{-2}}{\partial \mathbf{y} \partial \mathbf{y}^T} &= -\frac{2\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T\}}{n \hat{\sigma}^4} + \frac{4\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T\} \mathbf{y}}{n \hat{\sigma}^6} \frac{\partial \hat{\sigma}^2}{\partial \mathbf{y}^T} \\
&= -\frac{2\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T\}}{n \hat{\sigma}^4} \\
&\quad + \frac{8\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T\} \mathbf{y} \mathbf{y}^T \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T\}}{n^2 \hat{\sigma}^6}.
\end{aligned} \quad (\text{A.7})$$

直接将 (A.4)–(A.7) 代入 (2.5), 可得

$$\begin{aligned}
\hat{\mathbf{A}}\mathbf{I}(\boldsymbol{\Omega}_0) &= n \log 2\pi + \log |\boldsymbol{\Omega}(\hat{\boldsymbol{\alpha}})| + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}}) (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&\quad + 2 \text{trace} \left\{ \hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\alpha}}) \boldsymbol{\Omega}_0^{-1} \mathbf{X} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}^T} \right\} - 2 \sum_{j=1}^q \frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}^T} \boldsymbol{\Omega}_0 \mathbf{U}_j(\hat{\boldsymbol{\alpha}}) (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})
\end{aligned}$$

$$\begin{aligned}
& + \text{trace} \left\{ \sum_{j=1}^q U_j(\hat{\alpha}) \Omega_0 \frac{\partial^2 \hat{\alpha}_j}{\partial \mathbf{y} \partial \mathbf{y}^T} \Omega_0 \right\} + \sum_{j=1}^q \sum_{j_1=1}^q \frac{\partial \hat{\alpha}_{j_1}}{\partial \mathbf{y}^T} \Omega_0 U_{jj_1}(\hat{\alpha}) \Omega_0 \frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}} \\
& = n \log 2\pi + n \log \hat{\sigma}^2 + \hat{\sigma}^{-2} (\mathbf{y} - \mathbf{X}\hat{\beta})^T (\mathbf{y} - \mathbf{X}\hat{\beta}) \\
& \quad + 2\sigma_0^2 \hat{\sigma}^{-2} \text{trace} \left\{ \mathbf{X} \frac{\partial \hat{\beta}}{\partial \mathbf{y}^T} \right\} - 2\sigma_0^2 \frac{\partial \hat{\sigma}^{-2}}{\partial \mathbf{y}^T} (\mathbf{y} - \mathbf{X}\hat{\beta}) + \sigma_0^4 \text{trace} \left\{ \frac{\partial^2 \hat{\sigma}^{-2}}{\partial \mathbf{y} \partial \mathbf{y}^T} \right\} \\
& = n \log 2\pi + n \log \hat{\sigma}^2 + n + 2p\sigma_0^2 \hat{\sigma}^{-2} + 4\sigma_0^2 \frac{\mathbf{y}^T (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)}{n\hat{\sigma}^4} (\mathbf{y} - \mathbf{X}\hat{\beta}) \\
& \quad - 2(n-p)n^{-1}\sigma_0^4 \hat{\sigma}^{-4} + 8n^{-1}\sigma_0^4 \hat{\sigma}^{-4} \\
& = n \log 2\pi + n \log \hat{\sigma}^2 + n + 2(p+2) \frac{\sigma_0^2}{\hat{\sigma}^2} + \frac{8-2(n-p)}{n} \frac{\sigma_0^4}{\hat{\sigma}^4}.
\end{aligned}$$

附录 A.1.3 (2.8) 的证明

由于我们已经假定此时 (2.2) 是设定正确的, 可知 $\sigma_0^2 n^{-1} \hat{\sigma}^{-2}$ 是均值为 $(n-p-2)^{-1}$ 、方差为 $2(n-p-2)^{-2}(n-p-4)^{-1}$ 的 χ^2 -分布随机变量的倒数, 这表明

$$\begin{aligned}
\mathbb{E} \left(\frac{\sigma_0^2}{n\hat{\sigma}^2} \right)^2 &= \frac{1}{(n-p-2)^2} + \frac{2}{(n-p-2)^2(n-p-4)} \\
&= \frac{n-p-4+2}{(n-p-2)^2(n-p-4)} \\
&= \frac{1}{(n-p-2)(n-p-4)}. \tag{A.8}
\end{aligned}$$

由 (2.6) 和 (A.8) 可知,

$$\begin{aligned}
\mathbb{E}\{\hat{\text{AI}}(\Omega_0)\} &= \mathbb{E} \left(n \log 2\pi + n \log \hat{\sigma}^2 + n + 2(p+2) \frac{\sigma_0^2}{\hat{\sigma}^2} + (8-2(n-p))n^{-1} \frac{\sigma_0^4}{\hat{\sigma}^4} \right) \\
&= \mathbb{E} \left(n \log 2\pi + n \log \hat{\sigma}^2 + n + \frac{(2p+4)n}{n-p-2} + \frac{n^2(8-2(n-p))}{n(n-p-2)(n-p-4)} \right) \\
&= \mathbb{E} \left(n \log 2\pi + n \log \hat{\sigma}^2 + n + \frac{(2p+4)n}{n-p-2} - \frac{2n}{n-p-2} \right) \\
&= \mathbb{E} \left(n \log 2\pi + n \log \hat{\sigma}^2 + n + \frac{2pn+2n}{n-p-2} \right) \\
&= \mathbb{E} \left(n \log 2\pi + n \log \hat{\sigma}^2 + n \frac{n+p}{n-p-2} \right) \\
&= \mathbb{E}(\text{AICc}).
\end{aligned}$$

至此证明了 (2.8).

附录 A.1.4 引理 3.1 的证明

注意到

$$\begin{aligned}
& \text{trace} \left\{ \mathbf{X}_{(m)} \frac{\partial \hat{\beta}_{(m)}}{\partial \mathbf{y}^T} \right\} \\
& = p_m + 2 \text{trace} \{ (\mathbf{X}_{(m)}^T \hat{\Omega}_{(m)}^{-1} \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \hat{\mathbf{K}}_{(m)}^T (\hat{\mathbf{B}}_{(m)} - \hat{\mathbf{D}}_{(m)})^{-1} \hat{\mathbf{S}}_{(m)} \mathbf{X}_{(m)} \}, \tag{A.9}
\end{aligned}$$

其中 $\hat{\mathbf{S}}_{(m)}$ 为一个 $q_m \times n$ 阶矩阵, 其第 j 行为

$$(\mathbf{y} - \mathbf{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)})^T \hat{\mathbf{U}}_{(m),j} = \mathbf{y}^T (\mathbf{I}_n - \hat{\mathbf{H}}_{(m)}\hat{\boldsymbol{\Omega}}_{(m)}^{-1})^T \hat{\mathbf{U}}_{(m),j},$$

且 $\hat{\mathbf{K}}_{(m)}$ 是一个 $q_m \times n$ 阶矩阵, 其第 j 行为 $\mathbf{y}^T \hat{\mathbf{M}}_{(m),j}$. 由条件 3.1–3.3 可得

$$\begin{aligned} & |\text{trace}\{(\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \hat{\mathbf{K}}_{(m)}^T (\hat{\mathbf{B}}_{(m)} - \hat{\mathbf{D}}_{(m)})^{-1} \hat{\mathbf{S}}_{(m)} \mathbf{X}_{(m)}\}| \\ &= \left| \frac{1}{n} \text{trace}\{(\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \hat{\mathbf{K}}_{(m)}^T \hat{\boldsymbol{\Delta}}_{(m)}^{-1} \hat{\mathbf{S}}_{(m)} \mathbf{X}_{(m)}\} \right| \\ &= \left| \frac{1}{n} \text{trace} \left\{ (\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T (\hat{\mathbf{M}}_{(m),1} \mathbf{y}, \dots, \hat{\mathbf{M}}_{(m),q_m} \mathbf{y}) \hat{\boldsymbol{\Delta}}_{(m)}^{-1} \right. \right. \\ &\quad \times \left. \begin{pmatrix} \mathbf{y}^T (\mathbf{I}_n - \hat{\mathbf{H}}_{(m)} \hat{\boldsymbol{\Omega}}_{(m)}^{-1})^T \hat{\mathbf{U}}_{(m),1} \\ \vdots \\ \mathbf{y}^T (\mathbf{I}_n - \hat{\mathbf{H}}_{(m)} \hat{\boldsymbol{\Omega}}_{(m)}^{-1})^T \hat{\mathbf{U}}_{(m),q_m} \end{pmatrix} \mathbf{X}_{(m)} \right\} \Big| \\ &\leq \frac{1}{n} \sum_{j=1}^{q_m} \sum_{l=1}^{q_m} |\hat{\delta}_{(m)}^{jl}| \text{trace}\{(\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \hat{\mathbf{M}}_{(m),j} \mathbf{y} \mathbf{y}^T (\mathbf{I}_n - \hat{\mathbf{H}}_{(m)} \hat{\boldsymbol{\Omega}}_{(m)}^{-1})^T \hat{\mathbf{U}}_{(m),l} \mathbf{X}_{(m)}\} \\ &\leq \frac{1}{n^2} \sum_{j=1}^{q_m} \sum_{l=1}^{q_m} |\hat{\delta}_{(m)}^{jl}| \left\| \mathbf{y}^T (\mathbf{I}_n - \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \hat{\mathbf{H}}_{(m)}) \hat{\mathbf{U}}_{(m),l} \mathbf{X}_{(m)} \left(\frac{\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)} \mathbf{X}_{(m)}}{n} \right)^{-1} \mathbf{X}_{(m)}^T \hat{\mathbf{M}}_{(m),j} \mathbf{y} \right\| \\ &\leq \sum_{j=1}^{q_m} \sum_{l=1}^{q_m} |\hat{\delta}_{(m)}^{jl}| \frac{\|\mathbf{y}\|^2}{n} \frac{\|\mathbf{X}_{(m)}\|^2}{n} \lambda_{\min}^{-1} \left(\frac{\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)} \mathbf{X}_{(m)}}{n} \right) \\ &\quad \times \|\hat{\mathbf{U}}_{(m),l}\| \|\hat{\mathbf{M}}_{(m),j}\| (1 + \|\hat{\boldsymbol{\Omega}}_{(m)}^{-1}\| \|\hat{\mathbf{H}}_{(m)}\|), \end{aligned} \quad (\text{A.10})$$

其中 $\hat{\delta}_{(m)}^{jl}$ 是 $\hat{\boldsymbol{\Delta}}_{(m)}^{-1}$ 的第 (j, l) 元, 且满足对于所有 $j, l \in \{1, \dots, q_m\}$, 有

$$|\hat{\delta}_{(m)}^{jl}| \leq \|\hat{\boldsymbol{\Delta}}_{(m)}^{-1}\| = O_p(1). \quad (\text{A.11})$$

在条件 3.3、3.1(i) 和 3.1(ii) 下,

$$\|\hat{\mathbf{H}}_{(m)}\| \leq \left\| \frac{\mathbf{X}_{(m)}}{n^{1/2}} \right\|^2 \lambda_{\min}^{-1} \left(\frac{\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)} \mathbf{X}_{(m)}}{n} \right) = O_p(1). \quad (\text{A.12})$$

基于条件 3.1, 可以验证

$$\begin{aligned} \max_{j \in \{1, \dots, q_m\}} \|\hat{\mathbf{M}}_{(m),j}\| &\leq \left(1 + \|\hat{\boldsymbol{\Omega}}_{(m)}\| \left\| \frac{\mathbf{X}_{(m)}}{n^{1/2}} \right\| \lambda_{\min}^{-1} \left(\frac{\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)} \mathbf{X}_{(m)}}{n} \right) \right)^2 \max_{j \in \{1, \dots, q_m\}} \|\hat{\mathbf{U}}_{(m),j}\| \\ &= O_p(1). \end{aligned} \quad (\text{A.13})$$

由 (A.10)–(A.13), 在条件 3.1 和 3.2 下, 有

$$|\text{trace}\{(\mathbf{X}_{(m)}^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \hat{\mathbf{K}}_{(m)}^T (\hat{\mathbf{B}}_{(m)} - \hat{\mathbf{D}}_{(m)})^{-1} \hat{\mathbf{S}}_{(m)} \mathbf{X}_{(m)}\}| = O_p(q_m^2). \quad (\text{A.14})$$

类似地, 由条件 3.3 可以得到

$$|\hat{c}_{(m),js}| \leq \frac{\|(\hat{\mathbf{B}}_{(m)}/n - \hat{\mathbf{D}}_{(m)}/n)^{-1}\|}{n} = \frac{\|\hat{\Delta}_{(m)}^{-1}\|}{n} = O_p(n^{-1}), \quad (\text{A.15})$$

其中 $\hat{c}_{(m),js}$ 是向量 $\hat{\mathbf{c}}_{(m),j}$ 的第 s 个分量, 这里

$$\hat{\mathbf{c}}_{(m),j} = (\hat{\mathbf{B}}_{(m)} - \hat{\mathbf{D}}_{(m)})^{-1} \mathbf{l}_{n,j}, \quad j = 1, \dots, q,$$

$\mathbf{l}_{a,j}$ 表示第 j 个分量为 1、其他分量是 0 的 a 维单位向量. 然后, 由 (A.13) 和 (A.15), 可得

$$\begin{aligned} & \left| \sum_{j=1}^{q_m} \frac{\partial \hat{\alpha}_{(m),j}}{\partial \mathbf{y}^T} \hat{\Omega}_{(m)} \hat{\mathbf{U}}_{(m),j} (\mathbf{y} - \mathbf{X}_{(m)} \hat{\beta}_{(m)}) \right| \\ & \leq \frac{2}{n} \sum_{j=1}^{q_m} \sum_{s=1}^{q_m} n |\hat{c}_{(m),js}| \|\mathbf{y}^T \hat{\mathbf{M}}_{(m),s} \hat{\Omega}_{(m)} \hat{\mathbf{U}}_{(m),j} (\mathbf{I}_n - \hat{\mathbf{H}}_{(m)} \hat{\Omega}_{(m)}^{-1}) \mathbf{y}\| \\ & \leq \frac{2 \|\mathbf{y}\|^2}{n} \sum_{j=1}^{q_m} \sum_{s=1}^{q_m} \|\hat{\Delta}_{(m)}^{-1}\| \|\hat{\mathbf{M}}_{(m),s}\| \|\hat{\Omega}_{(m)}\| \|\hat{\mathbf{U}}_{(m),j}\| \|\mathbf{I}_n - \hat{\mathbf{H}}_{(m)} \hat{\Omega}_{(m)}^{-1}\| \\ & = O_p(q_m^2). \end{aligned} \quad (\text{A.16})$$

合并 (A.9)、(A.14) 和 (A.16), 并结合 $\hat{\Pi}_{(m),1}$ 的定义可以得到 $\hat{\Pi}_{(m),1} = O_p(p_m + q_m^2)$.

接下来证明 $\hat{\Pi}_{(m),2} = O_p(q_m^4)$. 用 $\hat{\mathbf{B}}_{(m)}^{(s)}$ 表示一个 $q_m \times q_m$ 阶矩阵, 其 (l, k) 元为 $\hat{b}_{(m),lks}$, $\hat{\mathbf{D}}_{(m)}^{(s)}$ 表示一个 $q_m \times q_m$ 阶矩阵, 其 (l, k) 元为 $\mathbf{y}^T \hat{\mathbf{M}}_{(m),lks} \mathbf{y}$,

$$\begin{aligned} \hat{\mathbf{G}}_{(m),j} &= (\hat{\mathbf{M}}_{(m),jl} \mathbf{y}, \dots, \hat{\mathbf{M}}_{(m),jq_m} \mathbf{y})^T, \\ \hat{\mathbf{f}}_{(m),js} &= \hat{\mathbf{B}}_{(m)}^{(s)} \hat{\mathbf{c}}_{(m),j}, \quad \hat{\mathbf{v}}_{(m),js} = \hat{\mathbf{D}}_{(m)}^{(s)} \hat{\mathbf{c}}_{(m),j}, \\ \hat{\mathbf{F}}_{(m),j} &= (\hat{\mathbf{f}}_{(m),j1}, \dots, \hat{\mathbf{f}}_{(m),jq_m}) \end{aligned}$$

且 $\hat{\mathbf{V}}_{(m),j} = (\hat{\mathbf{v}}_{(m),j1}, \dots, \hat{\mathbf{v}}_{(m),jq_m})$. 与 (A.13) 的推导类似, 我们可以验证

$$\max_{j,l \in \{1, \dots, q_m\}} \|\hat{\mathbf{M}}_{(m),jl}\| = O_p(1), \quad \max_{j,l,s \in \{1, \dots, q_m\}} \|\hat{\mathbf{M}}_{(m),jls}\| = O_p(1). \quad (\text{A.17})$$

此外,

$$\begin{aligned} & \frac{(n - p_m - 2)(n - p_m - 4) \hat{\Pi}_{(m),2}}{n^2} \\ &= \sum_{j=1}^{q_m} \text{trace} \left\{ \hat{\mathbf{U}}_{(m),j} \hat{\Omega}_{(m)} \left[2 \sum_{s=1}^q \left(\hat{c}_{(m),js} \hat{\mathbf{G}}_{(m),s}^T \frac{\partial \hat{\alpha}_{(m)}}{\partial \mathbf{y}^T} + \hat{c}_{(m),js} \hat{\mathbf{M}}_{(m),s} \right) \right. \right. \\ & \quad \left. \left. - \frac{2}{n} \hat{\mathbf{K}}_m^T \hat{\Delta}_{(m)}^{-1} \left((\hat{\mathbf{F}}_{(m),j} - \hat{\mathbf{V}}_{(m),j}) \frac{\partial \hat{\alpha}_{(m)}}{\partial \mathbf{y}^T} - 2 \sum_{s=1}^q \hat{c}_{(m),js} \hat{\mathbf{G}}_{(m),j} \right) \right] \hat{\Omega}_{(m)} \right\} \\ & \quad + \sum_{j=1}^{q_m} \sum_{l=1}^{q_m} \text{trace} \left\{ \frac{\partial \hat{\alpha}_{(m),l}}{\partial \mathbf{y}^T} \hat{\Omega}_{(m)} \mathbf{U}_{(m),jl} \hat{\Omega}_{(m)} \frac{\partial \hat{\alpha}_{(m),j}}{\partial \mathbf{y}^T} \right\} \\ &= \sum_{j=1}^{q_m} \text{trace} \left\{ \hat{\mathbf{U}}_{(m),j} \hat{\Omega}_{(m)} \left[2 \sum_{s=1}^q \left(2 \hat{c}_{(m),js} \hat{\mathbf{G}}_{(m),s}^T \frac{\hat{\Delta}_{(m)}^{-1}}{n} \hat{\mathbf{K}}_{(m)} + \hat{c}_{(m),js} \hat{\mathbf{M}}_{(m),s} \right) \right. \right. \end{aligned}$$

$$\begin{aligned}
& -\frac{4}{n}\hat{\mathbf{K}}_m^T\hat{\Delta}_{(m)}^{-1}\left((\hat{\mathbf{F}}_{(m),j}-\hat{\mathbf{V}}_{(m),j})\frac{\hat{\Delta}_{(m)}^{-1}}{n}\hat{\mathbf{K}}_{(m)}-\sum_{s=1}^q\hat{c}_{(m),js}\hat{\mathbf{G}}_{(m),s}\right)\hat{\Omega}_{(m)}\Big\} \\
& +\sum_{j=1}^{q_m}\sum_{l=1}^{q_m}\text{trace}\left\{\frac{\partial\hat{\alpha}_{(m),l}}{\partial\mathbf{y}^T}\hat{\Omega}_{(m)}\mathbf{U}_{(m),jl}\hat{\Omega}_{(m)}\frac{\partial\hat{\alpha}_{(m),j}}{\partial\mathbf{y}^T}\right\}.
\end{aligned} \tag{A.18}$$

由 (A.11)、(A.15) 和 (A.17), 结合条件 3.1–3.3, 可得

$$\begin{aligned}
& \left|\sum_{j=1}^{q_m}\text{trace}\left\{\hat{\mathbf{U}}_{(m),j}\hat{\Omega}_{(m)}\left[2\sum_{s=1}^q\left(2\hat{c}_{(m),js}\hat{\mathbf{G}}_{(m),s}^T\frac{\hat{\Delta}_{(m)}^{-1}}{n}\hat{\mathbf{K}}_{(m)}+\hat{c}_{(m),js}\hat{\mathbf{M}}_{(m),s}\right)\right]\hat{\Omega}_{(m)}\right\}\right| \\
& \leq 2\sum_{j=1}^{q_m}\sum_{s=1}^{q_m}|\hat{c}_{(m),js}|\left|\text{trace}\left\{\hat{\mathbf{U}}_{(m),j}\hat{\Omega}_{(m)}\left(2\hat{\mathbf{G}}_{(m),s}^T\frac{\hat{\Delta}_{(m)}^{-1}}{n}\hat{\mathbf{K}}_{(m)}+\hat{\mathbf{M}}_{(m),s}\right)\hat{\Omega}_{(m)}\right\}\right| \\
& = 4\sum_{j=1}^{q_m}\sum_{s=1}^{q_m}\sum_{l=1}^{q_m}\sum_{k=1}^{q_m}|\hat{c}_{(m),js}||\hat{\delta}_{(m)}^{lk}||\mathbf{y}^T\hat{\mathbf{M}}_{(m),k}\hat{\Omega}_{(m)}\hat{\mathbf{U}}_{(m),j}\hat{\Omega}_{(m)}\hat{\mathbf{M}}_{(m),sl}\mathbf{y}| \\
& \quad + 2\sum_{j=1}^{q_m}\sum_{s=1}^{q_m}|\hat{c}_{(m),js}||\text{trace}\{\hat{\mathbf{U}}_{(m),j}\hat{\Omega}_{(m)}\hat{\mathbf{M}}_{(m),s}\hat{\Omega}_{(m)}\}| \\
& \leq \frac{2\|\mathbf{y}\|^2}{n}\sum_{j=1}^{q_m}\sum_{s=1}^{q_m}\sum_{l=1}^{q_m}\sum_{k=1}^{q_m}|\hat{c}_{(m),js}||\hat{\delta}_{(m)}^{lk}||\|\hat{\mathbf{M}}_{(m),k}\hat{\Omega}_{(m)}\hat{\mathbf{U}}_{(m),j}\hat{\Omega}_{(m)}\hat{\mathbf{M}}_{(m),sl}\| \\
& \quad + \frac{2\text{trace}(\hat{\Omega}_{(m)})}{n}\sum_{j=1}^{q_m}\sum_{s=1}^{q_m}|n\hat{c}_{(m),js}||\|\hat{\mathbf{U}}_{(m),j}\| \|\hat{\Omega}_{(m)}\| \|\hat{\mathbf{M}}_{(m),s}\| \\
& = O_p\left(\frac{q_m^4}{n}\right) + O_p(q_m^2).
\end{aligned} \tag{A.19}$$

另外, 记 $\hat{\Delta}_{(m)}^{-1}(\hat{\mathbf{F}}_{(m),j}-\hat{\mathbf{V}}_{(m),j})\hat{\Delta}_{(m)}^{-1}/n=\hat{\Delta}_{(m),2}=(\hat{\delta}_{(m),2jl})_{q_m\times q_m}$, 结合 (A.15)、条件 3.1 和 3.2, 可以直接得到 $\|\hat{\Delta}_{(m),2}\|=O_p(1)$. 综上, 由 (A.15)、(A.17) 和条件 3.1–3.3, 有

$$\begin{aligned}
& \sum_{j=1}^{q_m}\left|\text{trace}\left\{\frac{1}{n}\hat{\mathbf{U}}_{(m),j}\hat{\Omega}_{(m)}\hat{\mathbf{K}}_m^T\hat{\Delta}_{(m)}^{-1}\left((\hat{\mathbf{F}}_{(m),j}-\hat{\mathbf{V}}_{(m),j})\frac{\hat{\Delta}_{(m)}^{-1}}{n}\hat{\mathbf{K}}_{(m)}-2\sum_{s=1}^q\hat{c}_{(m),js}\hat{\mathbf{G}}_{(m),s}\right)\hat{\Omega}_{(m)}\right\}\right| \\
& \leq \sum_{j=1}^{q_m}\frac{\text{trace}(\hat{\Omega}_{(m)})}{n}\left\|\hat{\mathbf{U}}_{(m),j}\hat{\Omega}_{(m)}\hat{\mathbf{K}}_m^T\hat{\Delta}_{(m)}^{-1}\frac{(\hat{\mathbf{F}}_{(m),j}-\hat{\mathbf{V}}_{(m),j})}{n}\hat{\Delta}_{(m)}^{-1}\hat{\mathbf{K}}_{(m)}\right\| \\
& \quad + 2\frac{\text{trace}(\hat{\Omega}_{(m)})}{n}\sum_{j=1}^{q_m}\sum_{s=1}^q|\hat{c}_{(m),js}||\|\hat{\mathbf{U}}_{(m),j}\hat{\Omega}_{(m)}\hat{\mathbf{K}}_m^T\hat{\Delta}_{(m)}^{-1}\hat{\mathbf{G}}_{(m),s}\| \\
& \leq \frac{\text{trace}(\hat{\Omega}_{(m)})}{n}\|\hat{\mathbf{U}}_{(m),j}\| \|\hat{\Omega}_{(m)}\| \sum_{j=1}^{q_m}\sum_{s=1}^{q_m}\sum_{l=1}^{q_m}|\hat{\delta}_{(m),2sl}||\|\hat{\mathbf{M}}_{(m),s}\| \|\hat{\mathbf{M}}_{(m),l}\| \\
& \quad + 2\frac{\text{trace}(\hat{\Omega}_{(m)})}{n}\frac{\|\mathbf{y}\|^2}{n}\sum_{j=1}^{q_m}\sum_{s=1}^{q_m}\sum_{l=1}^{q_m}\sum_{k=1}^{q_m}|n\hat{c}_{(m),js}||\hat{\delta}_{(m)}^{kl}||\|\hat{\mathbf{U}}_{(m),j}\| \|\hat{\Omega}_{(m)}\| \|\hat{\mathbf{M}}_{(m),l}\| \|\hat{\mathbf{M}}_{(m),sk}\| \\
& = O_p(q_m^3) + O_p(q_m^4).
\end{aligned} \tag{A.20}$$

类似地, 由 (A.15)、(A.17) 和条件 3.1–3.3, 可以推导出

$$\left|\sum_{j=1}^{q_m}\sum_{l=1}^{q_m}\text{trace}\left\{\frac{\partial\hat{\alpha}_{(m),l}}{\partial\mathbf{y}^T}\hat{\Omega}_{(m)}\mathbf{U}_{(m),jl}\hat{\Omega}_{(m)}\frac{\partial\hat{\alpha}_{(m),j}}{\partial\mathbf{y}^T}\right\}\right|$$

$$\begin{aligned}
&\leq 4 \sum_{j=1}^{q_m} \sum_{l=1}^{q_m} \sum_{s=1}^{q_m} \sum_{k=1}^{q_m} |\hat{c}_{(m),jk}| |\hat{c}_{(m),ls}| \|\mathbf{y}^T \hat{\mathbf{M}}_{(m),s} \hat{\boldsymbol{\Omega}}_{(m)} \hat{\mathbf{U}}_{(m),jl} \hat{\boldsymbol{\Omega}}_{(m)} \hat{\mathbf{M}}_{(m),k} \mathbf{y}\| \\
&\leq \frac{4 \|\mathbf{y}\|^2}{n} \sum_{j=1}^{q_m} \sum_{l=1}^{q_m} \sum_{s=1}^{q_m} \sum_{k=1}^{q_m} |n \hat{c}_{(m),jk}| |n \hat{c}_{(m),ls}| \frac{\|\hat{\mathbf{M}}_{(m),s}\| \|\hat{\boldsymbol{\Omega}}_{(m)}\|^2 \|\hat{\mathbf{U}}_{(m),jl}\| \|\hat{\mathbf{M}}_{(m),k}\|}{n} \\
&= O_p\left(\frac{q_m^4}{n}\right). \tag{A.21}
\end{aligned}$$

将 (A.18) 与 (A.19)–(A.21) 合并, 我们最终证明了 $\hat{\Pi}_{(m),2} = O_p(q_m^4)$, 至此完成了证明.

附录 A.1.5 定理 3.1 的证明

最小化 $\text{AICc}^*(m)$ 等价于最小化

$$\mathcal{C}(m) = \bar{c}_0 + n \log(2\pi) + \log |\hat{\boldsymbol{\Omega}}_{(m)}| + (\mathbf{y} - \mathbf{X}_{(m)} \hat{\boldsymbol{\beta}}_{(m)})^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} (\mathbf{y} - \mathbf{X}_{(m)} \hat{\boldsymbol{\beta}}_{(m)}) + \hat{\Pi}_{(m),1} + \hat{\Pi}_{(m),2}.$$

要证明定理 3.1, 只需证明当 $n \rightarrow \infty$ 时,

$$\sup_{1 \leq m \leq M} \left| \frac{\mathcal{C}(m) - L_{\text{KL}}(m)}{R_{\text{KL}}(m)} \right| = o_p(1) \tag{A.22}$$

且

$$\sup_{1 \leq m \leq M} \left| \frac{L_{\text{KL}}(m)}{R_{\text{KL}}(m)} - 1 \right| = o_p(1). \tag{A.23}$$

现在一一验证上述等式.

记

$$r_n(m) = \mathbf{e}_0^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{e}_0 - \text{trace}(\hat{\boldsymbol{\Omega}}_{(m)}^{-1} \boldsymbol{\Omega}_0) - 2(\boldsymbol{\mu} - \mathbf{X}_{(m)} \hat{\boldsymbol{\beta}}_{(m)})^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{e}_0. \tag{A.24}$$

由 $\mathcal{C}(m)$ 的定义可知,

$$\mathcal{C}(m) - L_{\text{KL}}(m) = r_n(m) + \hat{\Pi}_{(m),1} + \hat{\Pi}_{(m),2}. \tag{A.25}$$

可以看出, 在条件 3.1 和 3.2 下, 存在常数 $K_1 > 0$, 使得

$$\begin{aligned}
\mathbb{E}|r_n(m)| &\leq \mathbb{E}|\mathbf{e}_0^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{e}_0 - \mathbf{e}_0^T \boldsymbol{\Omega}_{(m)}^{-1}(\boldsymbol{\alpha}_{(m)}^*) \mathbf{e}_0| + \mathbb{E}|\text{trace}\{\boldsymbol{\Omega}_{(m)}^{-1}(\boldsymbol{\alpha}_{(m)}^*) \boldsymbol{\Omega}_0\} - \text{trace}(\hat{\boldsymbol{\Omega}}_{(m)}^{-1} \boldsymbol{\Omega}_0)| \\
&\quad + 2\mathbb{E}|(\boldsymbol{\mu} - \mathbf{X}_{(m)} \hat{\boldsymbol{\beta}}_{(m)})^T \hat{\boldsymbol{\Omega}}_{(m)}^{-1} \mathbf{e}_0 - (\boldsymbol{\mu} - \mathbf{X}_{(m)} \boldsymbol{\beta}_{(m)}^*)^T \boldsymbol{\Omega}_{(m)}^{-1}(\boldsymbol{\alpha}_{(m)}^*) \mathbf{e}_0| \\
&\quad + \mathbb{E}|\mathbf{e}_0^T \boldsymbol{\Omega}_{(m)}^{-1}(\boldsymbol{\alpha}_{(m)}^*) \mathbf{e}_0 - \text{trace}\{\boldsymbol{\Omega}_{(m)}^{-1}(\boldsymbol{\alpha}_{(m)}^*) \boldsymbol{\Omega}_0\}| \\
&\leq K_1 n^{1/2} \mathbb{E}^{1/2} \left(1 + \frac{\|\mathbf{e}_0\|^2}{n} + \left\| \frac{\mathbf{e}_0}{n^{1/2}} \right\|^2 \right) \mathbb{E}^{1/2} \|\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{(m)}^*\|^2 \\
&\quad + n^{1/2} \left[\frac{1}{n} \sum_{i=1}^n (\mu_{4,i} - 3(\sigma_i^2)^2) d_{(m),ii}^2 + \frac{2}{n} \text{trace}\{\boldsymbol{\Omega}_{(m)}^{-1}(\boldsymbol{\alpha}_{(m)}^*) \boldsymbol{\Omega}_0 \boldsymbol{\Omega}_{(m)}^{-1}(\boldsymbol{\alpha}_{(m)}^*) \boldsymbol{\Omega}_0\} \right]^{1/2}, \tag{A.26}
\end{aligned}$$

其中

$$\mu_{4,i} = \mathbb{E}(e_{0i}^4),$$

而 $d_{(m),ii}$ 是 $\Omega_{(m)}^{-1}(\alpha_{(m)}^*)\Omega_0\Omega_{(m)}^{-1}(\alpha_{(m)}^*)\Omega_0$ 的第 i 个主对角元, 这里用到了文献 [16, 23] 中二次型方差的公式. 然后, 由条件 3.1、3.2、3.4 和 (A.26) 可以得到, 对于任意给定的 $\delta > 0$, 存在一个常数 $K_1 > 0$, 使得

$$\begin{aligned} & \mathbb{P}\left(\sup_{1 \leq m \leq M} \left| \frac{r_n(m)}{R_{\text{KL}}(m)} \right| > \delta\right) \\ & \leq \sum_{m=1}^M \frac{\mathbb{E}|\mathcal{C}(m) - L_{\text{KL}}(m)|}{\delta R_{\text{KL}}(m)} \\ & \leq \sum_{m=1}^M \frac{K_1 \mathbb{E}^{1/2}(1 + \frac{\|\mathbf{e}_0\|^2}{n} + \|\frac{\mathbf{e}_0}{n^{1/2}}\|)^2 \mathbb{E}^{1/2}\|n^{1/2}(\boldsymbol{\theta}_{(m)} - \boldsymbol{\theta}_{(m)}^*)\|^2}{\delta R_{\text{KL}}(m)/n^{1/2}} \\ & \quad + \sum_{m=1}^M \frac{[\frac{1}{n} \sum_{i=1}^n (\mu_{4,i} - 3(\sigma_i^2)^2) d_{(m),ii}^2 + \frac{2}{n} \text{trace}\{\Omega_{(m)}^{-1}(\alpha_{(m)}^*)\Omega_0\Omega_{(m)}^{-1}(\alpha_{(m)}^*)\Omega_0\}]^{1/2}}{\delta R_{\text{KL}}(m)/n^{1/2}} \\ & \leq K_1 \sum_{m=1}^M \left(\frac{\delta R_{\text{KL}}(m)}{n^{1/2}} \right)^{-1}. \end{aligned} \quad (\text{A.27})$$

另外, 由引理 3.1, 在条件 3.5 下, 可得

$$\sup_{1 \leq m \leq M} \left| \frac{\hat{\Pi}_{(m),1} + \hat{\Pi}_{(m),2}}{R_{\text{KL}}(m)} \right| = o_p(1). \quad (\text{A.28})$$

在条件 3.5 下, 由 (A.25)–(A.28) 可以得出 (A.22). 类似地, 在条件 3.1 和 3.4 下, 我们还可以证明存在某常数 $K_1 > 0$, 使得

$$\begin{aligned} & \mathbb{E}|L_{\text{KL}}(m) - R_{\text{KL}}(m)| \\ & \leq \mathbb{E}|\log|\hat{\Omega}_{(m)}^{-1}| - \log|\Omega_{(m)}^{-1}(\alpha_{(m)}^*)|| + \mathbb{E}|\log|\hat{\Omega}_{(m)}^{-1}| - \log|\Omega_{(m)}^{-1}(\alpha_{(m)}^*)|| \\ & \quad + \mathbb{E}|(\boldsymbol{\mu} - \mathbf{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)})^T \hat{\Omega}_{(m)}^{-1}(\boldsymbol{\mu} - \mathbf{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)}) - (\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}^*)^T \Omega_{(m)}^{-1}(\alpha_{(m)}^*)(\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}^*)| \\ & \quad + |\mathbb{E}|(\boldsymbol{\mu} - \mathbf{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)})^T \hat{\Omega}_{(m)}^{-1}(\boldsymbol{\mu} - \mathbf{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)}) - (\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}^*)^T \Omega_{(m)}^{-1}(\alpha_{(m)}^*)(\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}^*)|| \\ & \quad + \mathbb{E}|\text{trace}(\hat{\Omega}_{(m)}^{-1}\Omega_0) - \text{trace}\{\Omega_{(m)}^{-1}(\alpha_{(m)}^*)\Omega_0\}| - \mathbb{E}|\text{trace}(\hat{\Omega}_{(m)}^{-1}\Omega_0) - \text{trace}\{\Omega_{(m)}^{-1}(\alpha_{(m)}^*)\Omega_0\}| \\ & \leq n^{1/2} K_1 \mathbb{E}^{1/2}\|n^{1/2}(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{(m)}^*)\|^2. \end{aligned}$$

在条件 3.4 和 3.5 下, 由上述结果可以得出, 对于任意给定的 $\delta > 0$, 当 $n \rightarrow \infty$ 时,

$$\mathbb{P}\left(\sup_{1 \leq m \leq M} \left| \frac{L_{\text{KL}}(m)}{R_{\text{KL}}(m)} - 1 \right| > \delta\right) \leq \sum_{m=1}^M \frac{K_1 \mathbb{E}^{1/2}\|n^{1/2}(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{(m)}^*)\|^2}{R_{\text{KL}}(m)/n^{1/2}} \rightarrow 0. \quad (\text{A.29})$$

由 (A.29) 可以推导出 (A.23). 由 (A.22) 和 (A.23), 使用文献 [12, 定理 2] 的相同证明方法, 我们可以证明定理 3.1.

附录 A.2 关于条件 3.5 的讨论

如果对数据施加确切的分布假定, 则条件 3.5 可以从更为直接的条件导出. 为了展示该事实, 使用 $D_{\text{KL}}(\boldsymbol{\theta}_{(m)})$ 表示第 m 个备选模型与真实数据生成过程之间的 Kullback-Leibler 离差. 如果 $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \Omega_0)$, 则

$$D_{\text{KL}}(\boldsymbol{\theta}_{(m)}) = (\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)})^T \Omega_{(m)}^{-1}(\alpha_{(m)})(\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)})$$

$$+ \text{trace}\{\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)})\mathbf{\Omega}_0\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)})\} - n - \log |\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)})\mathbf{\Omega}_0\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)})|.$$

可以看出 $D_{\text{KL}}(\hat{\boldsymbol{\theta}}_{(m)}) = L_{\text{KL}}(\hat{\boldsymbol{\theta}}_{(m)})$. 使用 $\boldsymbol{\theta}_{(m)}^*$ 的定义, 我们有

$$\left. \frac{\partial D_{\text{KL}}(\boldsymbol{\theta}_{(m)})}{\partial \boldsymbol{\theta}_{(m)}} \right|_{\boldsymbol{\theta}_{(m)} = \boldsymbol{\theta}_{(m)}^*} = \mathbf{0}_{(p_m+q_m) \times 1}.$$

因此, 将 $D_{\text{KL}}(\hat{\boldsymbol{\theta}}_{(m)})$ 在 $\boldsymbol{\theta}_{(m)}^*$ 附近展开至二阶, 可以得到

$$\begin{aligned} R_{\text{KL}}(m) &= E\{D_{\text{KL}}(\hat{\boldsymbol{\theta}}_{(m)})\} \\ &= (\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}^*)^T \mathbf{\Omega}_{(m)}^{-1}(\boldsymbol{\alpha}_{(m)}^*)(\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}^*) \\ &\quad + \text{trace}\{\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\mathbf{\Omega}_0\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\} - n - \log |\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\mathbf{\Omega}_0\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)| \\ &\quad + \frac{E\{(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{(m)}^*)^T \nabla_{(m)}(\tilde{\boldsymbol{\theta}}_{(m)})(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{(m)}^*)\}}{2}, \end{aligned}$$

其中

$$\nabla_{(m)}(\boldsymbol{\theta}_{(m)}) = \frac{\partial^2 D_{\text{KL}}(\boldsymbol{\theta}_{(m)})}{\partial \boldsymbol{\theta}_{(m)} \partial \boldsymbol{\theta}_{(m)}^T},$$

且 $\tilde{\boldsymbol{\theta}}_{(m)}$ 落在 $\hat{\boldsymbol{\theta}}_{(m)}$ 与 $\boldsymbol{\theta}_{(m)}^*$ 之间. 如果进一步假定对任意的 $\boldsymbol{\theta}_{(m)}$, $\nabla_{(m)}(\boldsymbol{\theta}_{(m)})$ 为正定矩阵, 则由条件 3.1(ii) 可知,

$$R_{\text{KL}}(m) \geq c_1^{-1} \|\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}^*\|^2,$$

其中我们使用了下列事实:

$$\begin{aligned} &\text{trace}\{\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\mathbf{\Omega}_0\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\} - n - \log |\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\mathbf{\Omega}_0\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)| \\ &= \sum_{i=1}^n (\lambda_i \{\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\mathbf{\Omega}_0\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\} - 1 - \log[\lambda_i \{\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\mathbf{\Omega}_0\mathbf{\Omega}_{(m)}^{-1/2}(\boldsymbol{\alpha}_{(m)}^*)\}]) \\ &\geq 0, \end{aligned}$$

这里 $\lambda_i(\cdot)$ 表示一个一般矩阵的第 i 个特征值. 上述讨论表明 Kullback-Leibler 型损失永远大于等于其对应的二次损失 (在相差一个正的常数意义下), Kullback-Leibler 型损失与二次损失的进一步关系可以在文献 [24] 中找到. 此外, 上述论证表明, 在本小节的设定下, 条件 3.5 可以被下列条件替代: 当 $n \rightarrow \infty$ 时,

$$\sum_{m=1}^M \left(\frac{\|\boldsymbol{\mu} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}^*\|^2}{n^{1/2}} \right)^{-k} \rightarrow 0. \quad (\text{A.30})$$

应当指出的是, 如果存在 m_0 , 使得 $\boldsymbol{\mu} = \mathbf{X}_{(m_0)}\boldsymbol{\beta}_{(m_0)}^*$ (即第 m_0 个模型是正确设定时), 则 (A.30) 不成立. 另一方面, 如果忽略了重要的协变量, 则 (A.30) 通常是成立的. 为了展示这一点, 假设 $\mathbf{X}_{(M)}$ 是协变量最多的模型对应的协变量矩阵, 此时所有备选模型的协变量矩阵的列向量均落入 $\mathbf{X}_{(M)}$ 的列空间. 另外, 假设真实数据生成过程中,

$$\boldsymbol{\mu} = \mathbf{X}_{(M)}\boldsymbol{\beta}_{(M)}^0 + \mathbf{x}_0\beta_0,$$

其中 $\mathbf{x}_0$ 是所忽略的协变量, β_0 为非零常数. 由拟真值的定义可知,

$$\beta_{(m)}^* = \{\mathbf{X}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1}(\alpha_{(m)}^*) \mathbf{X}_{(m)}\}^{-1} \mathbf{X}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1}(\alpha_{(m)}^*) \boldsymbol{\mu}.$$

另外, 由于矩阵 $\mathbf{I}_n - \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*) \mathbf{X}_{(m)} \{\mathbf{X}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1}(\alpha_{(m)}^*) \mathbf{X}_{(m)}\}^{-1} \mathbf{X}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*)$ 是幂等阵, 其秩为 $n - p_m$, 因此, 其谱分解为

$$\begin{aligned} \mathbf{I}_n - \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*) \mathbf{X}_{(m)} \{\mathbf{X}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1}(\alpha_{(m)}^*) \mathbf{X}_{(m)}\}^{-1} \mathbf{X}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*) \\ = \mathbf{Q}_{(m)} \text{blkdiag}(\mathbf{I}_{n-p_m}, \mathbf{0}_{p_m \times p_m}) \mathbf{Q}_{(m)}^T, \end{aligned}$$

其中 $\mathbf{Q}_{(m)}$ 是正交阵, $\text{blkdiag}(\cdot, \cdot)$ 表示分块对角阵.

现在将 $\mathbf{X}_{(M)}$ 分割为 $\mathbf{X}_{(M)} = (\mathbf{X}_{(m)}, \mathbf{X}_{(m^c)})$, 与之对应地, 将 $\beta_{(M)}^0$ 分割为

$$\beta_{(M)}^0 = \begin{pmatrix} \beta_{(M,m)}^0 \\ \beta_{(M,m^c)}^0 \end{pmatrix}.$$

此外, 将 $\beta_0 \mathbf{Q}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*) \mathbf{x}_0$ 和 $\mathbf{Q}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*) \mathbf{X}_{(m^c)} \beta_{(M,m^c)}^0$ 的前 $n - p_m$ 个元素所构成的子向量分别记为 $\mathbf{a}_m$ 和 $\mathbf{b}_m$, 容易验证, 对于所有 $1 \leq m \leq M$, 有

$$\begin{aligned} c_1 \|\boldsymbol{\mu} - \mathbf{X}_{(m)} \beta_{(m)}^*\|^2 \\ \geq (\boldsymbol{\mu} - \mathbf{X}_{(m)} \beta_{(m)}^*)^T \boldsymbol{\Omega}_{(m)}^{-1}(\alpha_{(m)}^*) (\boldsymbol{\mu} - \mathbf{X}_{(m)} \beta_{(m)}^*) \\ = (\mathbf{X}_{(m^c)} \beta_{(M,m^c)}^0 + \mathbf{x}_0 \beta_0)^T \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*) \\ \times [\mathbf{I}_n - \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*) \mathbf{X}_{(m)} (\mathbf{X}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1}(\alpha_{(m)}^*) \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*)] \\ \times \boldsymbol{\Omega}_{(m)}^{-1/2}(\alpha_{(m)}^*) (\mathbf{X}_{(m^c)} \beta_{(M,m^c)}^0 + \mathbf{x}_0 \beta_0) \\ = \|\mathbf{a}_m + \mathbf{b}_m\|^2. \end{aligned}$$

因此, 如果对于所有 $1 \leq m \leq M$, $(n - p_m) \times 1$ 向量 $\mathbf{a}_m + \mathbf{b}_m$ 中非零的元素的数量大于 $K_0 n^{1/2+\delta}$, 其中 $K_0 > 0$ 是某常数, 则 (A.30) 自动成立.

附录 A.3 计算细节

附录 A.3.1 $\partial \hat{\beta}^T / \partial \mathbf{y}$ 、 $\partial \hat{\alpha}^T / \partial \mathbf{y}$ 和 $\partial^2 \hat{\alpha}_j / \partial \mathbf{y} \partial \mathbf{y}^T$ ($j = 1, \dots, q$) 的推导

得分函数显然满足

$$\begin{cases} (\mathbf{y} - \mathbf{X} \hat{\beta})^T \boldsymbol{\Omega}^{-1}(\hat{\alpha}) \mathbf{X} = \mathbf{0}_{1 \times p}, \\ (\mathbf{y} - \mathbf{X} \hat{\beta})^T \mathbf{U}_j(\hat{\alpha}) (\mathbf{y} - \mathbf{X} \hat{\beta}) - \text{trace}\{\mathbf{U}_j(\hat{\alpha}) \boldsymbol{\Omega}(\hat{\alpha})\} = 0, \quad j = 1, \dots, q, \end{cases} \quad (\text{A.31})$$

其中 $\hat{\beta} = \tilde{\beta}(\hat{\alpha})$, $\tilde{\beta}(\alpha) = (\mathbf{X}^T \boldsymbol{\Omega}^{-1}(\alpha) \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-1}(\alpha) \mathbf{y}$. 记

$$\mathbf{H}(\hat{\alpha}) = \mathbf{X} (\mathbf{X}^T \boldsymbol{\Omega}^{-1}(\hat{\alpha}) \mathbf{X})^{-1} \mathbf{X}^T$$

以及

$$\mathbf{M}_j(\alpha) = (\mathbf{I}_n - \boldsymbol{\Omega}^{-1}(\alpha) \mathbf{H}(\alpha)) \mathbf{U}_j(\alpha) (\mathbf{I}_n - \mathbf{H}(\alpha) \boldsymbol{\Omega}^{-1}(\alpha)).$$

可以看出 $\hat{\alpha}$ 满足

$$\mathbf{y}^T \mathbf{M}_j(\hat{\alpha}) \mathbf{y} - \text{trace}\{\mathbf{U}_j(\hat{\alpha}) \mathbf{\Omega}(\hat{\alpha})\} = 0, \quad j = 1, \dots, q. \quad (\text{A.32})$$

用 $\mathbf{S}(\beta, \alpha, \mathbf{y})$ 表示一个 $q \times n$ 维矩阵, 其第 j 行为

$$\mathbf{s}_j^T(\beta, \alpha, \mathbf{y}) = (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{U}_j(\alpha).$$

用 $\mathbf{K}(\alpha, \mathbf{y})$ 表示一个 $q \times n$ 维矩阵, 其第 j 行为

$$\mathbf{k}_j^T = \mathbf{y}^T \mathbf{M}_j(\alpha).$$

在 $(\mathbf{y} - \mathbf{X}\hat{\beta})^T \mathbf{\Omega}^{-1}(\hat{\alpha}) = \mathbf{0}_{1 \times p}$ 的两边同时对 $\mathbf{y}$ 求偏导, 得到

$$\frac{\partial \hat{\alpha}^T}{\partial \mathbf{y}} \mathbf{K}(\hat{\beta}, \hat{\alpha}, \mathbf{y}) \mathbf{X} + \mathbf{\Omega}^{-1}(\hat{\alpha}) \mathbf{X} - \frac{\partial \hat{\beta}^T}{\partial \mathbf{y}} \mathbf{X}^T \mathbf{\Omega}^{-1}(\hat{\alpha}) \mathbf{X} = \mathbf{0}_{n \times p}. \quad (\text{A.33})$$

此外, 用 $\mathbf{B}(\alpha)$ 表示一个 $q \times q$ 维矩阵, 其 (j, l) 元为

$$\begin{aligned} b_{jl}(\alpha) &= \frac{\partial \text{trace}\{\mathbf{U}_j(\alpha) \mathbf{\Omega}(\alpha)\}}{\partial \alpha_l} \\ &= \text{trace}\left\{\frac{\partial \mathbf{U}_j(\alpha) \mathbf{\Omega}(\alpha)}{\partial \alpha_l}\right\} \\ &= \text{trace}\left\{\mathbf{U}_j(\alpha) \frac{\partial \mathbf{\Omega}(\alpha)}{\partial \alpha_l}\right\} + \text{trace}\{\mathbf{U}_{jl}(\alpha) \mathbf{\Omega}(\alpha)\} \end{aligned}$$

并用 $\mathbf{D}(\alpha, \mathbf{y})$ 表示一个 $q \times q$ 维矩阵, 其 (j, l) 元为 $\mathbf{y}^T \mathbf{M}_{jl}(\alpha) \mathbf{y}$. 值得指出的是, $\mathbf{B}(\alpha)$ 和 $\mathbf{D}(\alpha, \mathbf{y})$ 是对称阵. 在 (A.32) 的两边同时对 $\mathbf{y}$ 求偏导, 有

$$2\mathbf{K}^T(\hat{\alpha}, \mathbf{y}) + \frac{\partial \hat{\alpha}^T}{\partial \mathbf{y}} (\mathbf{D}(\hat{\alpha}, \mathbf{y}) - \mathbf{B}(\hat{\alpha})) = \mathbf{0}_{n \times q}. \quad (\text{A.34})$$

等式 (A.33) 和 (A.34) 表明, 当涉及的矩阵均可逆时,

$$\frac{\partial \hat{\alpha}^T}{\partial \mathbf{y}} = 2\mathbf{K}^T(\hat{\alpha}, \mathbf{y}) (\mathbf{B}(\hat{\alpha}) - \mathbf{D}(\hat{\alpha}, \mathbf{y}))^{-1}$$

且

$$\frac{\partial \hat{\beta}^T}{\partial \mathbf{y}} = \left(\frac{\partial \hat{\alpha}^T}{\partial \mathbf{y}} \mathbf{S}(\hat{\beta}, \hat{\alpha}, \mathbf{y}) \mathbf{X} + \mathbf{\Omega}^{-1}(\hat{\alpha}) \mathbf{X} \right) (\mathbf{X}^T \mathbf{\Omega}^{-1}(\hat{\alpha}) \mathbf{X})^{-1}.$$

下一步给出 $\partial^2 \hat{\alpha}_j / \partial \mathbf{y} \partial \mathbf{y}^T$ ($j = 1, \dots, q$) 的解析表达式. 用 $\mathbf{l}_{a,j}$ 表示第 i 个分量为 1、其他分量是 0 的 a 维单位向量. 可以得到

$$\frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}} = \frac{\partial \alpha^T}{\partial \mathbf{y}} \mathbf{l}_{n,j}.$$

进一步地, 记 $\mathbf{c}_j(\hat{\alpha}, \mathbf{y}) = (\mathbf{B}(\hat{\alpha}) - \mathbf{D}(\hat{\alpha}, \mathbf{y}))^{-1} \mathbf{l}_{n,j}$ ($j = 1, \dots, q$), 用 $c_{js}(\hat{\alpha}, \mathbf{y})$ 表示 $\mathbf{c}_j(\hat{\alpha}, \mathbf{y})$ 的第 s ($j = 1, \dots, q$) 个分量, 我们立即可以得到

$$\frac{\partial \hat{\alpha}_j}{\partial \mathbf{y}} = 2 \sum_{s=1}^q c_{js}(\hat{\alpha}, \mathbf{y}) \mathbf{M}_s(\hat{\alpha}) \mathbf{y}.$$

因此,

$$\frac{\partial^2 \hat{\alpha}_j}{\partial \mathbf{y} \partial \mathbf{y}^T} = 2 \sum_{s=1}^q c_{js}(\hat{\alpha}, \mathbf{y}) \mathbf{G}_s^T(\hat{\alpha}) \frac{\partial \hat{\alpha}}{\partial \mathbf{y}^T} + 2 \sum_{s=1}^q c_{js}(\hat{\alpha}, \mathbf{y}) \mathbf{M}_s(\hat{\alpha}) + 2 \mathbf{K}^T(\hat{\alpha}, \mathbf{y}) \frac{\partial c_j(\hat{\alpha}, \mathbf{y})}{\partial \mathbf{y}^T},$$

其中

$$\frac{\partial c_j(\hat{\alpha}, \mathbf{y})}{\partial \mathbf{y}^T} = -(\mathbf{B}(\hat{\alpha}) - \mathbf{D}(\hat{\alpha}, \mathbf{y}))^{-1} \left[(\mathbf{F}_j(\hat{\alpha}, \mathbf{y}) - \mathbf{V}_j(\hat{\alpha}, \mathbf{y})) \frac{\partial \hat{\alpha}}{\partial \mathbf{y}^T} - 2 \sum_{s=1}^q c_{js}(\hat{\alpha}, \mathbf{y}) \mathbf{G}_s(\hat{\alpha}, \mathbf{y}) \right],$$

这里 $j, s = 1, \dots, q$,

$$\begin{aligned} \mathbf{G}_j^T(\alpha, \mathbf{y}) &= (g_{j1}(\alpha, \mathbf{y}), \dots, g_{jq}(\alpha, \mathbf{y})), \\ \mathbf{F}_j(\alpha, \mathbf{y}) &= (f_{j1}(\alpha, \mathbf{y}), \dots, f_{jq}(\alpha, \mathbf{y})), \\ \mathbf{V}_j(\alpha, \mathbf{y}) &= (v_{j1}(\alpha, \mathbf{y}), \dots, v_{jq}(\alpha, \mathbf{y})) \end{aligned}$$

且

$$g_{js}(\alpha, \mathbf{y}) = \mathbf{M}_{js}(\alpha) \mathbf{y}, \quad f_{js}(\alpha, \mathbf{y}) = \left(\frac{\partial \mathbf{B}(\alpha)}{\partial \alpha_s} \right) c_j(\alpha, \mathbf{y}), \quad v_{js}(\alpha, \mathbf{y}) = \left(\frac{\partial \mathbf{D}(\alpha)}{\partial \alpha_s} \right) c_j(\alpha, \mathbf{y}).$$

附录 A.3.2 计算空间自回归误差模型下的 AICc^*

记 $\mathbf{S}_n(\rho) = \mathbf{I}_n - \rho \mathbf{W}_n$. 由 $\mathbf{e} = \mathbf{W}_n \mathbf{e} + \boldsymbol{\epsilon}$, 有 $\mathbf{e} = \mathbf{S}_n^{-1}(\rho) \boldsymbol{\epsilon}$. 因此, 在此模型下,

$$\boldsymbol{\alpha} = (\rho, \sigma^2)^T, \quad \boldsymbol{\Omega}(\boldsymbol{\alpha}) = \sigma^2 \mathbf{S}_n^{-1}(\rho) (\mathbf{S}_n^{-1}(\rho))^T$$

且

$$\boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha}) = \sigma^{-2} \mathbf{S}_n^T(\rho) \mathbf{S}_n(\rho).$$

为计算 AICc^* , 我们需要下列结果:

$$\begin{aligned} \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial \rho} &= \mathbf{S}_n^{-1}(\rho) \mathbf{W}_n \boldsymbol{\Omega}(\boldsymbol{\alpha}) + \boldsymbol{\Omega}(\boldsymbol{\alpha}) \mathbf{W}_n^T (\mathbf{S}_n^{-1}(\rho))^T, \\ \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial \sigma^2} &= \mathbf{S}_n^{-1}(\rho) (\mathbf{S}_n^{-1}(\rho))^T, \\ \frac{\partial^2 \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial \rho^2} &= \mathbf{S}_n^{-1}(\rho) \mathbf{W}_n \mathbf{S}_n^{-1}(\rho) \mathbf{W}_n \boldsymbol{\Omega}(\boldsymbol{\alpha}) + \mathbf{S}_n^{-1}(\rho) \mathbf{W}_n \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial \rho} \\ &\quad + \boldsymbol{\Omega}(\boldsymbol{\alpha}) \mathbf{W}_n^T (\mathbf{S}_n^{-1}(\rho))^T \mathbf{W}_n^T (\mathbf{S}_n^{-1}(\rho))^T + \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial \rho} \mathbf{W}_n^T (\mathbf{S}_n^{-1}(\rho))^T, \\ \frac{\partial^2 \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial \rho \partial \sigma^2} &= \mathbf{S}_n^{-1}(\rho) \mathbf{W}_n \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial \sigma^2} + \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial \sigma^2} \mathbf{W}_n^T (\mathbf{S}_n^{-1}(\rho))^T \end{aligned}$$

以及

$$\frac{\partial^2 \boldsymbol{\Omega}(\boldsymbol{\alpha})}{\partial (\sigma^2)^2} = \mathbf{0}_{n \times n}.$$

此外, 我们还需要

$$U_1(\boldsymbol{\alpha}) = \frac{\partial \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial \rho} = -\frac{\mathbf{W}_n^T \mathbf{S}_n(\rho) + \mathbf{S}_n^T(\rho) \mathbf{W}_n}{\sigma^2}, \quad U_2(\boldsymbol{\alpha}) = \frac{\partial \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial \sigma^2} = -\frac{\mathbf{S}_n^T(\rho) \mathbf{S}_n(\rho)}{(\sigma^2)^2},$$

$$\begin{aligned}
U_{11}(\boldsymbol{\alpha}) &= \frac{\partial^2 \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial \rho^2} = \frac{2\mathbf{W}_n^T \mathbf{W}_n}{\sigma^2}, & U_{12}(\boldsymbol{\alpha}) &= \frac{\partial^2 \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial \rho \partial \sigma^2} = -\frac{\mathbf{W}_n^T \mathbf{S}_n(\rho) + \mathbf{S}_n^T(\rho) \mathbf{W}_n}{(\sigma^2)^2}, \\
U_{12}(\boldsymbol{\alpha}) &= U_{12}(\boldsymbol{\alpha}), & U_{22}(\boldsymbol{\alpha}) &= \frac{\partial^2 \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial \rho^2} = \frac{2\mathbf{S}_n^T(\rho) \mathbf{S}_n(\rho)}{(\sigma^2)^3}, \\
U_{111}(\boldsymbol{\alpha}) &= \frac{\partial^3 \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial \rho^3} = \mathbf{0}_{n \times n}, & U_{112}(\boldsymbol{\alpha}) &= \frac{\partial^3 \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial \rho^2 \partial \sigma^2} = -\frac{2\mathbf{W}_n^T \mathbf{W}_n}{(\sigma^2)^2}, \\
U_{121}(\boldsymbol{\alpha}) &= U_{112}(\boldsymbol{\alpha}), & U_{122}(\boldsymbol{\alpha}) &= \frac{\partial^3 \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial \rho \partial (\sigma^2)^2} = -\frac{2(\mathbf{W}_n^T \mathbf{S}_n(\rho) + \mathbf{S}_n^T(\rho) \mathbf{W}_n)}{(\sigma^2)^3}, \\
U_{211}(\boldsymbol{\alpha}) &= U_{112}(\boldsymbol{\alpha}), & U_{212}(\boldsymbol{\alpha}) &= U_{122}(\boldsymbol{\alpha}), \\
U_{221}(\boldsymbol{\alpha}) &= U_{122}(\boldsymbol{\alpha})
\end{aligned}$$

以及

$$U_{222}(\boldsymbol{\alpha}) = \frac{\partial^3 \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha})}{\partial (\sigma^2)^3} = -\frac{6\mathbf{S}_n^T(\rho) \mathbf{S}_n(\rho)}{(\sigma^2)^4}.$$

Corrected Akaike information criterion with general covariance matrix

Shangwei Zhao, Dalei Yu & Xinyu Zhang

Abstract In this paper, within the framework of Stein's identity, we propose a new corrected Akaike information criterion for the finite sample setting. The new criterion applies to the situation where very general covariance structures are involved. Under certain regularity conditions, we establish the asymptotic efficiency of the proposed model selection criterion. Simulations in the spatial regression model with autoregressive errors show that our method is promising when the difference between the candidate models and the true data generating process is small. Our method becomes particularly competitive with its competitors when such difference becomes larger. The proposed model selection criterion is also applied to the analysis of a set of real data (the Neighborhood Crimes Data) and the results further support the use of our method in practical situations.

Keywords asymptotic efficiency, corrected Akaike information criterion, model selection, spatial autoregressive error

MSC(2020) 62J05, 62F12, 62H11

doi: 10.1360/SCM-2020-0308