中国科学: 数学 2024 年 第 54 卷 第 ? 期: 1 ∼ 22
SCIENTIA SINICA Mathematica
论 文

超高维参数单指标模型的拟合优度检验

周亭攸[1], 张耀武[2*], 朱利平[3]

1. 浙江财经大学数据科学学院, 杭州 310018;
2. 上海财经大学信息管理与工程学院, 上海 200433;
3. 中国人民大学统计与大数据研究院, 北京 100872

E-mail: zhoutingyou@zufe.edu.cn, zhang.yaowu@mail.shufe.edu.cn, zhu.liping@ruc.edu.cn

收稿日期: 2023-06-22; 接受日期: 2024-04-15; 网络出版日期: 2024-06-21; * 通信作者
国家重点研发计划 (批准号: 2023YFA1008702) 和国家自然科学基金 (批准号: 12225113, 12271331 和 72192832) 资助项目

摘要 本文提出一种基于投影的两阶段检验方法, 用来对超高维参数单指标模型进行拟合优度检验. 将样本数据随机等分成两组, 在第一组数据上进行变量筛选, 并保留一部分有用的协变量和一部分候选的协变量; 在第二组数据上基于筛选后的协变量进行基于投影的拟合优度检验. 该两阶段方法避免了高维检验中可能出现的第 I 类错误膨胀和功效损失的问题. 此外, 在第二阶段的检验中提出了一个全新的、不依赖于调节参数并能避免维数灾难的检验统计量. 该统计量在原假设下是 n 相合的, 在备择假设下则是 $\sqrt{n}$ 相合的. 本文使用野自助法来确定该检验的临界值并在理论上证明该方法的有效性. 最后, 通过统计模拟和一个实际数据分析展示该方法的有限样本性质.

关键词 超高维 参数单指标模型 拟合优度检验 第 I 类错误 功效

MSC (2020) 主题分类 62G10

1 引言

本文研究超高维数据下的参数单指标模型的检验问题, 即当协变量的维数 p 远大于样本量 n 时, 如何检验一维响应变量 Y 与 p 维协变量 $\boldsymbol{x} \overset{\text{def}}{=} (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ 之间是否存在一个参数单指标关系. 这个问题在许多科学领域都有重要的应用, 如生物医药、基因诊断、遗传学、农业、地质、经济学和金融学等. 本文的目标是提出一种有效的检验方法, 该方法能够在高维场景下保持良好的真实水平 (size) 和检验功效 (power).

参数单指标模型是一类常用的统计模型, 它假设响应变量 Y 的条件期望 $\mathrm{E}(Y \mid \boldsymbol{x})$ 是协变量 $\boldsymbol{x}$ 的某个线性组合的函数, 即

$$Y = m(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}) + \varepsilon, \tag{1.1}$$

英文引用格式: Zhou T Y, Zhang Y W, Zhu L P. A lack-of-fit test for parametric index model with ultrahigh-dimensional covariates (in Chinese). Sci Sin Math, 2024, 54: 1–22, doi: 10.1360/SSM-2023-0188

其中 $m(\cdot)$ 是一个已知的平方可积的连续函数, $\boldsymbol{\beta}$ 是未知的 p 维参数向量, ε 是随机误差项且满足 $\mathrm{E}(\varepsilon \mid \boldsymbol{x}) = 0$. 选择不同的 $m(\cdot)$ 可以得到不同类型的参数单指标模型, 例如, 当 $m(\cdot)$ 取恒等函数时, 模型 (1.1) 退化为针对连续响应变量的普通线性模型; 而当 $m(\cdot)$ 为 Logit 函数或者 Probit 函数时, 模型 (1.1) 分别对应处理二值响应变量的 Logistic 模型和 Probit 模型; 当 $m(\cdot)$ 取自然对数函数时, (1.1) 是能够对有序离散的响应变量建模的 Poisson 回归模型. 值得注意的是, 本文所讨论的参数单指标模型包含广义线性模型, 但相较于广义线性模型, 参数单指标模型的假设更为简化, 仅假设响应变量只依赖于协变量的某一个线性组合, 而对响应变量的分布并未作出特殊要求. 而广义线性模型则需要进一步假设响应变量的分布属于指数分布族. 因此, 我们认为参数单指标模型相对于广义线性模型具有更广泛的适用性与一般性.

然而, 在使用参数单指标模型之前, 首先需要验证这个模型是否适合当下的数据集, 即是否能够刻画出响应变量 Y 与协变量 $\boldsymbol{x}$ 之间的真实关系. 这涉及一个假设检验问题:

$$\begin{aligned} &H_0 : \Pr\{\mathrm{E}(Y \mid \boldsymbol{x}) = m(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\} = 1 \ \text{ 对某个 } \boldsymbol{\beta} \in \mathbb{R}^p \text{ 成立},\\ &H_1 : \Pr\{\mathrm{E}(Y \mid \boldsymbol{x}) = m(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\} < 1 \ \text{ 对所有的 } \boldsymbol{\beta} \in \mathbb{R}^p \text{ 成立}. \end{aligned} \tag{1.2}$$

如果原假设成立, 则说明参数单指标模型是合适的; 否则, 说明存在一些其他形式的相关关系, 需要使用更复杂或更灵活的模型来拟合数据. 这个检验问题在低维或中等维数的情形下已经有很多研究, 如文献 [3, 8, 23, 24, 29]. 但是, 在超高维情形下, 即当 $p \gg n$ 时, 这些方法都会失效或不适用. 第 2 节通过例 2.1 来说明当协变量的维数增大时, 模型检验的检验功效将会显著下降.

近几年也有一些其他的文章关注了超高维情形下的假设检验问题, 但主要集中在模型的单个参数或者组参数的显著性检验和区间估计方面, 而对模型检验却极少涉猎. 有兴趣的读者可以参见文献 [2] 中的总结. 此外, Shah 和 Bühlmann [19] 针对超高维的线性模型提出了一种基于残差的模型检验方法. 他们的思路是先用普通最小二乘或 LASSO (least absolute shrinkage and selection operator) 回归 (当维数 p 很大时) 拟合数据, 并得到残差 $\widehat{\varepsilon}$. 如果原假设成立, 即数据真实服从某个线性模型, 则残差中应该只含有很少的信息, 与随机误差项 ε 相似; 如果原假设不成立, 即存在一个更合适的非线性模型, 则残差中应该含有一些非线性信号. 基于这个假设, 他们用某一回归模型 (如随机森林) 来预测残差中的非线性信号, 并根据预测能力的强弱来判断是否拒绝原假设. 随后, Janková 等 [10] 提出了针对高维广义线性模型的拟合优度检验方法. 该方法首先利用带有惩罚项的广义线性模型 (GLM (generalized linear model) LASSO) 来估计参数并获取残差, 若模型设定有误, 则残差中将含有特定信号. 通过数据分组和加权 LASSO 回归的方式对检验统计量进行校正, 在模型拟合不佳的情形下, 检验统计量将呈现较大值, 从而有效地揭示模型的不足, 实现模型检验的目的. 另外, 文献 [27, 28] 在充分降维的框架下提出了一种基于变量筛选的拟合优度检验方法, 但是这种方法的检验功效可能会受到变量筛选阶段漏选重要变量的影响.

为了解决超高维情形下的检验问题, 需要引入一个重要的假设: 稀疏性条件. 这个假设认为, 在所有协变量中, 只有一小部分对响应变量有真正的影响作用, 而其他大部分都是无关紧要或冗余的. 用集合 $\mathcal{A}$ 表示那些有影响作用的协变量所对应的下标, 并用 $\boldsymbol{x}_{\mathcal{A}} = \{X_k : k \in \mathcal{A}\}$ 表示这些协变量本身. 稀疏性条件可以表示为

$$\mathcal{A} \overset{\text{def}}{=} \{k : \mathrm{E}(Y \mid \boldsymbol{x}) \text{ 在函数意义上依赖于 } X_k, \text{其中 } k = 1, \ldots, p\}. \tag{1.3}$$

在这种情形下, 原来的检验问题可以简化为

$$H_0 : \Pr\{\mathrm{E}(Y \mid \boldsymbol{x}) = m(\boldsymbol{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})\} = 1 \text{ 对某个 } \boldsymbol{\beta}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|} \text{ 成立},$$

$$H_1 : \Pr\{\mathrm{E}(Y \mid \boldsymbol{x}) = m(\boldsymbol{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})\} < 1 \text{ 对所有的 } \boldsymbol{\beta}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|} \text{ 成立}, \tag{1.4}$$

其中 $|\mathcal{A}|$ 表示集合 $\mathcal{A}$ 的大小 (即集合中元素的个数). 基于这个简化后的检验问题, 考虑一种朴素的两阶段检验方法: 首先进行变量筛选并得到集合 $\mathcal{A}$ 的估计 $\mathcal{S}_a$, 然后基于第一阶段筛选出来的变量 $\boldsymbol{x}_{\mathcal{S}_a}$ 进行模型检验.

为了实现检验问题的简化, 首先需要估计出那些有影响作用的协变量所对应的下标集合 $\mathcal{A}$. 在这方面, 一类受到广泛关注的方法是独立筛选的方法, 它们的目标是在保留真正有用的协变量的同时, 尽可能多地剔除对响应变量没有影响的协变量. 这类方法自从 Fan 和 Lv[5] 首次提出后就引发了大量的研究, 如文献 [1,6,14,15,17,20,32] 等. 感兴趣的读者可参见文献 [16] 在这方面的综述. 理论上这类方法具有良好的确定筛选性质 (sure screening property), 即随着样本量的增大, 这些方法能以趋于 1 的概率保证 $\mathcal{A} \subset \mathcal{S}_a$. 在超高维数据的应用环境中, 最理想的状态是 $\mathcal{A} = \mathcal{S}_a$, 但这基本上是不可能实现的. 在第一阶段中, 由于样本有限, 所以总有一部分没有用的协变量会保留下来, 即 $\mathcal{A} \subsetneq \mathcal{S}_a$, 将这种称作 “过拟合” 情形 (over-fitting case). 随后会在例 2.2 中展示过拟合情形导致后续检验第 I 类错误膨胀, 并提出数据分组的方法解决这一问题.

另外, 在实际应用中, 由于样本量有限、假设条件不满足、数据的随机性等因素, 所以也有可能在第一阶段的筛选中遗漏一些重要的协变量, 即 $\mathcal{S}_a \cap \mathcal{A} \neq \mathcal{A}$. 将这种称作 “欠拟合” 情形 (under-fitting case), 并在例 2.3 中展示欠拟合情形导致后续检验功效损失. 为了解决这一问题, 考虑在第一阶段的变量筛选中, 先后选出一部分有用的变量 $\boldsymbol{x}_{\mathcal{S}_a}$ 和一部分候选的变量 $\boldsymbol{x}_{\mathcal{S}_c}$, 并基于这些筛选出的变量构造检验统计量, 进而能缓解欠拟合情形下的检验失效问题.

在得到估计的指标集 $\mathcal{S} = \mathcal{S}_a \cup \mathcal{S}_c$ 之后, 只要满足 $\mathcal{A} \subset \mathcal{S}$, 就可以基于筛选出来的 (中等维数的) 协变量 $\boldsymbol{x}_{\mathcal{S}}$ 构造模型检验方法. 文献中已经存在很多模型检验方法, 大致可以分为两大类: 局部光滑方法[9,29] 和全局光滑方法[12,21,22]. 感兴趣的读者可参见文献 [7] 在此方面的综述. 然而, 上述模型检验方法多适用于协变量的维数 p 较小的情形, 若 p 不断增加将会出现典型的维数灾难问题进而导致检验失效. 为了避免这种情形, 文献 [13,25] 提出了基于投影的光滑方法, 文献 [8] 针对参数单指标模型提出了基于充分降维的检验方法. 但这些方法的结果很大程度上依赖于窗宽的选择.

本文提出一种改进的两阶段检验方法. 第一阶段, 利用一部分数据, 采用变量筛选的方法, 从所有协变量中筛选出一部分有用的协变量 $\boldsymbol{x}_{\mathcal{S}_a}$ 和一部分候选的协变量 $\boldsymbol{x}_{\mathcal{S}_c}$; 第二阶段, 利用另一部分数据, 基于筛选出来的协变量 $\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}_a \cup \mathcal{S}_c}$, 构造一个检验统计量, 并进行假设检验. 本文的方法有以下几个优点:

(1) 通过将数据分成两部分, 能够避免第一阶段变量筛选的过拟合现象对第二阶段检验的影响, 从而保证检验的大小不会膨胀.

(2) 通过引入候选协变量 $\boldsymbol{x}_{\mathcal{S}_c}$, 能够缓解第一阶段变量筛选的欠拟合现象对第二阶段检验的影响, 从而提高检验的功效.

(3) 构造的检验统计量是基于投影的, 既能够避免高维情形下的维数灾难问题, 又不需要选择任何调节参数.

本文余下内容的结构如下. 第 2 节通过几个例子展示过拟合和欠拟合对后续检验的影响, 进而引出本文的两阶段检验. 第 3 节详细介绍基于数据分组的两阶段检验方法. 该检验的相关理论性质的研究总结在第 4 节. 第 5 节给出大量的数据模拟和一个实际数据分析, 用来展示本文所提检验方法在有限样本下的表现. 第 6 节是介绍多重数据分组策略的扩展章节. 所有的技术细节被归入附录中.

2 高维模型检验中存在的问题

本节通过不同的例子说明高维模型检验中可能存在的几个问题, 并提出一种改进的两阶段检验方法来解决这些问题. 具体而言, 例 2.1 说明传统的检验方法的检验功效会随着维数的增加而显著降低; 例 2.2 和 2.3 分别说明当第一阶段的变量筛选时, 多选了没用的变量 (即 $\mathcal{A} \subsetneq \mathcal{S}_a$) 或者漏选了有用的变量 (即 $\mathcal{S}_a \cap \mathcal{A} \neq \mathcal{A}$), 会导致第二阶段中出现怎么样的检验结果. 本文将这两种情形分别称作 "过拟合" 和 "欠拟合". 在本节所有的例子中, 固定样本量 $n = 100$, 并将每个实验重复 1,000 次. 在例 2.2 和 2.3 中, 固定协变量维数 $p = 2{,}000$. 同时使用基于累积差异 (参见文献 [30,31]) 的筛选方法进行第一阶段的变量筛选, 并基于保留的 (中等维数的) 协变量进行第二阶段的拟合优度检验, 检验统计量 (3.6) 在第 3 节中将会详细介绍.

例 2.1 设 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 来自于均值向量为 $\mathbf{0}$、协方差矩阵为 $\boldsymbol{\Sigma} = (\sigma_{kl})_{p\times p}$ 的多元正态分布, 其中 $\sigma_{kl} = 0.5^{|k-l|}$, $k, l = 1, \ldots, p$. 令 $Y = c_0 X_1 + (1 - c_0)\varepsilon$, 其中 ε 与 $\boldsymbol{x}$ 独立且服从标准正态分布. 对于假设检验 $H_0 : \mathrm{E}(Y \mid \boldsymbol{x}) = \mathrm{E}Y$, 考察 $c_0 = 0$ 和 $c_0 = 1$ 两种情形. 当 $c_0 = 0$ 时, 原假设成立, 考察的是检验的真实水平 (size); 而当 $c_0 = 1$ 时, 备择假设成立, 考察的是检验的功效 (power). 考察不同的协变量维数 $p = 5, 10, 20, 100, 2{,}000$ 和不同的检验水平 $\alpha = 0.05, 0.10$, 并将每一个实验重复 1,000 次. 检验的结果列在表 1 中.

可以发现, 当 $c_0 = 0$ 时, 检验的真实水平并没有随着维数 p 的变化而产生较大的波动, 整体而言, 检验的真实水平很接近于预设水平 α. 但是当 $c_0 = 1$ 时, 即原假设不成立时, 检验的功效随着维数 p 的增大而急剧下降; 当 $p = 2{,}000$ 时, 检验已经完全失效. 这种由于超高维数所带来模型检验的失效问题很常见, 这是因为我们的目标备择假设是任意的相关关系, 导致检验方法没有针对性, 而在高维场景下备择假设类别又过于庞大 [26]. 所以在超高维的情形下, 上述的模型检验的方法都不再适用.

例 2.2 本例展示过拟合下的第 I 类错误膨胀, 并提出数据分组的方法解决这一问题. 记 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 是来自于均值为 $\mathbf{0}$、协方差矩阵为 $\boldsymbol{\Sigma} = (\sigma_{kl})_{p\times p}$ 的多元正态, 其中 $\sigma_{kl} = 0.5^{|k-l|}$, $k, l = 1, \ldots, p$. 独立地产生 $Y \sim \mathcal{N}(0, 1)$. 我们检验 $H_0 : \mathrm{E}(Y \mid \boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$ 对某些 $\boldsymbol{\beta} \in \mathbb{R}^p$ 成立. 考虑如下两种检验方式, 这两种检验的结果列在表 2 中.

(1) 不使用数据分组的两阶段检验: 首先基于全部的数据集 $\mathcal{D}$ 进行变量筛选并保留一个有用变量; 随后在同一组数据集 $\mathcal{D}$ 上, 基于这个选中的变量进行拟合优度检验.

(2) 使用数据分组的两阶段检验: 先将整个数据集 $\mathcal{D}$ 随机等分成两组, 分别记作 $\mathcal{D}_1$ 和 $\mathcal{D}_2$. 首先在数据集 $\mathcal{D}_1$ 上进行变量筛选并保留一个有用变量; 随后在数据集 $\mathcal{D}_2$ 上, 基于这个选中的变量进行拟合优度检验.

在此例中, Y 与 $\boldsymbol{x}$ 独立, 所以在第一阶段中保留的那一个变量就属于 "多选出来的无用的变量", 即过拟合. 从表 2 中可以看到, 不使用数据分组的检验会导致显著的第 I 类错误膨胀, 即当一些 (在总

表 1 例 2.1 中基于 1,000 次的重复实验所得到的检验的真实水平 (当 $c_0 = 0$ 时) 和功效 (当 $c_0 = 1$ 时). 考察 $p = 5, 10, 20, 100, 2{,}000$, 检验的水平取 $\alpha = 0.05, 0.10$

c_0	α	$p = 5$	$p = 10$	$p = 20$	$p = 100$	$p = 2{,}000$
0	0.05	0.045	0.044	0.044	0.043	0.042
	0.10	0.108	0.104	0.104	0.102	0.103
1	0.05	1.000	0.999	0.445	0.067	0.046
	0.10	1.000	1.000	0.895	0.148	0.096

表 2 例 2.2 中检验的第 I 类错误. 考虑两种检验方式: 不使用数据分组的两阶段检验和使用数据分组的两阶段检验. 考察不同的显著性水平 $\alpha = 0.05, 0.10$

α	无数据分组	数据分组
0.05	0.194	0.061
0.10	0.286	0.101

体水平上) 无用的协变量 (在样本水平上) "表现地" 好像对响应变量有影响时, 直接基于这些带有虚假相关性的协变量的检验会导致第 I 类错误膨胀 [4]. 若采用随机数据分组的策略, 则在数据集 $\mathcal{D}_1$ 上具有虚假相关性的协变量在数据 $\mathcal{D}_2$ 上对响应变量的影响会减小, 进而能够避免两阶段检验中第 I 类错误膨胀的问题.

例 2.3 本例展示欠拟合下的功效损失, 并通过将候选协变量 $\boldsymbol{x}_{\mathcal{S}_c}$ 引入到后续检验中来有效缓解这一问题. 响应变量 Y 来自于回归模型

$$Y = X_1 + X_2 + c_0(X_2^2 - 1) + \varepsilon,$$

其中 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 的分布与例 2.2 中相同, 误差项 $\varepsilon \sim \mathcal{N}(0,1)$ 并与 $\boldsymbol{x}$ 独立. 考虑检验 $H_0 : \mathrm{E}(Y \mid \boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$ 对某些 $\boldsymbol{\beta} \in \mathbb{R}^p$ 成立, 考察 $c_0 = 0, 1$. 当 $c_0 = 0$ 时, 原假设成立; 当 $c_0 = 1$ 时, 备择假设成立.

为了排除第一阶段中可能筛选出的虚假相关变量后续检验的影响, 选择固定的变量作为第二阶段中的检验变量. 分别选择 $\{X_1, X_2\}$、$\{X_3, X_4\}$ 和 $\{X_{300}, X_{400}\}$ 作为有用的协变量, 在第二阶段中, 根据这些预先指定的协变量进行模型拟合优度检验. 表 3 的左半部分 ("无候选变量") 总结了检验的第 I 类错误的概率 (当 $c_0 = 0$ 时) 和功效 (当 $c_0 = 1$ 时). 结果显示, 当 $c_0 = 0$ 时, 即使在测试过程中未选择 X_1 或 X_2, 也会接受原假设. 这是因为, 当 Y 和 $\boldsymbol{x}$ 服从联合正态分布时, 无论在第一阶段筛选出哪些协变量作为有用变量, Y 与所选的协变量依然服从联合正态分布, 因此线性关系始终成立. 但是, 由于 X_{300} 和 X_{400} 几乎与 Y 独立, 它们不足以刻画 Y 关于 $\boldsymbol{x}$ 的条件均值函数. 当 $c_0 = 1$ 时, 可以清楚地看到, 当缺少重要的协变量 X_1 或 X_2 时, 检验的功效将显著损失.

为了避免上述现象, 我们建议在检验 Y 是否线性依赖于当前筛选出的协变量的同时, 检验当前筛选出的协变量是否足够刻画 Y 的条件均值函数. 具体而言, 在第一阶段, 首先选择一组与 Y 最均值相关的 (中等维数的) 协变量 $\boldsymbol{x}_{\mathcal{S}_a}$, 然后基于 $\boldsymbol{x}_{\mathcal{S}_a}$ 构造 (参数单指标) 回归模型并得到相应的残差, 随后选出一组与残差最均值相关的协变量作为候选变量 $\boldsymbol{x}_{\mathcal{S}_c}$. 这里 $\mathcal{S}_c$ 满足 $\mathcal{S}_c \subset \mathcal{S}_a^c$, 其中 $\mathcal{S}_a^c$ 是集合 $\mathcal{S}_a$ 关于集合 $\{1, 2, \ldots, p\}$ 的补集. 通过上述步骤, 我们既能系统地选择出一组与 Y 直接相关的协变量 $\boldsymbol{x}_{\mathcal{S}_a}$, 又能选出与一组与 Y 中未被 $\boldsymbol{x}_{\mathcal{S}_a}$ 所解释的部分紧密相关的协变量 $\mathcal{S}_c$ 作为候选变量. 这样的选择过程有助于我们更全面地理解 Y 与协变量之间的关系, 并为后续的模型构建和检验提供有力的支持. 在

表 3 例 2.3 中检验的第 I 类错误 (当 $c_0 = 0$ 时) 和功效 (当 $c_0 = 1$ 时). 考察两种检验过程: 一种是没有候选变量, 另一种则是有 X_2 作为候选变量. 对于每个过程, 分别基于选定的协变量 $\{X_1, X_2\}$、$\{X_3, X_4\}$ 和 $\{X_{300}, X_{400}\}$ 来评价检验的结果. 研究不同的显着性水平 $\alpha = 0.05, 0.10$

		无候选变量			候选变量为 X_2	
c_0	α	X_1, X_2	X_3, X_4	X_{300}, X_{400}	X_3, X_4	X_{300}, X_{400}
0	0.05	0.049	0.054	0.054	1.000	1.000
	0.10	0.098	0.108	0.108	1.000	1.000
1	0.05	1.000	0.152	0.067	1.000	1.000
	0.10	1.000	0.245	0.129	1.000	1.000

第二阶段, 进行拟合优度检验, 即考察 $\mathrm{E}[\{Y - m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\} \mid \boldsymbol{x}_{\mathcal{S}}] = 0$ 是否成立, 其中 $\mathcal{S} = \mathcal{S}_a \cup \mathcal{S}_c$ 且 $\boldsymbol{x}_{\mathcal{S}} \overset{\text{def}}{=} (\boldsymbol{x}_{\mathcal{S}_a}^{\mathrm{T}}, \boldsymbol{x}_{\mathcal{S}_c}^{\mathrm{T}})^{\mathrm{T}}$.

下面使用上述检验方法来考察例 2.3 中欠拟合情形下的检验结果. 为简单起见, 固定 X_2 为候选变量. 从表 3 的右侧部分 ("候选变量为 X_2") 可以看到, 带有候选变量的检验可以有效地拒绝基于 $\{X_3, X_4\}$ 或 $\{X_{300}, X_{400}\}$ 的检验, 即检验的结果认为 $\{X_3, X_4\}$ 或 $\{X_{300}, X_{400}\}$ 并不足以刻画原始协变量与响应变量之间的条件均值关系. 此外, 当 $c_0 = 1$ 时, 模型的非线性部分是 X_2 的函数, 当将 X_2 作为候选变量引入后, 能够更直接地考虑到 X_2 对模型的影响, 从而显著提高检验的功效.

结合例 2.2 和 2.3, 整理出改进的两阶段检验方法如下:

(S0) 数据分组: 将样本观测 $\mathcal{D} \overset{\text{def}}{=} \{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$ 随机等分成两组, 分别记作 $\mathcal{D}_1 \overset{\text{def}}{=} \{\boldsymbol{x}_i, Y_i\}_{i=1}^{n_1}$ 和 $\mathcal{D}_2 \overset{\text{def}}{=} \{\boldsymbol{x}_i, Y_i\}_{i=n_1+1}^{n_1+n_2}$, 其中 $n_1 + n_2 = n$.

(S1) 第一阶段: 在数据集 $\mathcal{D}_1$ 上进行 (迭代的) 变量筛选, 并保留一部分有用的协变量 $\boldsymbol{x}_{\mathcal{S}_a}$ 和一部分候选的协变量 $\boldsymbol{x}_{\mathcal{S}_c}$.

(S2) 第二阶段: 在数据集 $\mathcal{D}_2$ 上进行拟合优度检验, 考察 $\mathrm{E}[\{Y - m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\} \mid \boldsymbol{x}_{\mathcal{S}}] = 0$ 是否成立, 其中 $\boldsymbol{x}_{\mathcal{S}} \overset{\text{def}}{=} (\boldsymbol{x}_{\mathcal{S}_a}^{\mathrm{T}}, \boldsymbol{x}_{\mathcal{S}_c}^{\mathrm{T}})^{\mathrm{T}}$.

此外, 漏选了重要的变量 (欠拟合) 也可能导致后续检验中潜在第一类错误膨胀, 第 6 节扩展部分说明此现象并使用多重数据分组策略来缓解这一问题.

3 改进的两阶段检验方法

第 2 节针对超高维情形下模型 (1.2) 的检验问题提出一种改进的两阶段方法. 本节介绍这种两阶段方法的操作细节.

3.1 第一阶段: 在观测集 $\mathcal{D}_1$ 上基于累积差异 (cumulative divergence, CD) 的边际筛选

在稀疏性原理的假定下, 检验 (1.2) 与 (1.4) 等价. 但是由于 $\mathcal{A}$ 往往是未知的, 于是需要事先估计. 本小节在数据集 $\mathcal{D}_1$ 上采用基于累积差异 [30,31] 的边际筛选来选择变量 $\boldsymbol{x}_{\mathcal{S}}$. 之所以采用累积差异作为单个协变量的边际效用, 主要原因是: 第一, 累积差异其定义本身就是度量条件均值 (而非条件分布, 参见文献 [15,32]) 的独立性问题, 因此与本文研究的主题相吻合; 第二, 累积差异具有对异常值稳健的性质 (文献 [20] 提出的鞅差相关系数 (martingale difference correlation, MDC) 虽然也是用来度量条件均值的独立性问题, 但是却容易受到极端值的影响), 这种稳健性在超高维的数据分析中格外具有优势, 因为高维数会导致样本观测中出现极端值的概率大大增加; 第三, 基于累积差异的边际筛选不需要对 $\mathrm{E}(Y \mid X_k)$ 的参数或者半参数结构有假定限制, 因此能广泛地使用到各类数据集上.

给定第一部分的观测数据 $\mathcal{D}_1$, 定义

$$\mathcal{S}_a \overset{\text{def}}{=} \{k : \widehat{\mathrm{CD}}(Y \mid X_k) \text{ 是 } \{\widehat{\mathrm{CD}}(Y \mid X_1), \ldots, \widehat{\mathrm{CD}}(Y \mid X_p)\} \text{ 中前 } s_a \text{ 个最大的值}\}, \tag{3.1}$$

其中 s_a 是人为指定的中等维数的模型大小, 并且 $\widehat{\mathrm{CD}}(Y \mid X_k)$ 的定义可参见文献 [31, (2.5)].

在一定的正则条件下, 基于 CD 的筛选方法具有确定筛选性质, 即当样本量趋于无穷时, $\Pr(\mathcal{A} \subset \mathcal{S}_a) \to 1$ 成立. 相关技术细节与文献 [32] 中的相似, 因此在此省略.

首先在数据集 $\mathcal{D}_1$ 上, 基于 (迭代的) 筛选方法选出一部分有用的协变量 $\boldsymbol{x}_{\mathcal{S}_a}$ 和一部分候选协变量 $\boldsymbol{x}_{\mathcal{S}_c}$; 随后在数据集 $\mathcal{D}_2$ 上检验 $\mathrm{E}[\{Y - m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\} \mid \boldsymbol{x}_{\mathcal{S}}] = 0$ 是否成立, 其中 $\boldsymbol{x}_{\mathcal{S}} \overset{\text{def}}{=} (\boldsymbol{x}_{\mathcal{S}_a}^{\mathrm{T}}, \boldsymbol{x}_{\mathcal{S}_c}^{\mathrm{T}})^{\mathrm{T}}$. 在

第二阶段, 我们是基于协变量 $\boldsymbol{x}_{\mathcal{S}}$, 而非协变量 $\boldsymbol{x}_{\mathcal{A}}$, 进行的检验. 我们真正检验的是

$$\begin{aligned} &H_0 : \Pr\{\mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}}) = m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}} \boldsymbol{x}_{\mathcal{S}_a})\} = 1 \ \text{ 对某个 } \boldsymbol{\beta}_{\mathcal{S}_a} \in \mathbb{R}^{|\mathcal{S}_a|} \text{ 成立}, \\ &H_1 : \Pr\{\mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}}) = m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}} \boldsymbol{x}_{\mathcal{S}_a})\} < 1 \ \text{ 对所有的 } \boldsymbol{\beta}_{\mathcal{S}_a} \in \mathbb{R}^{|\mathcal{S}_a|} \text{ 成立}. \end{aligned} \tag{3.2}$$

于是一个很重要的问题是, 检验 (1.4) 和 (3.2) 是否等价. 下面的定理回答了这个问题.

定理 3.1 对 (1.3) 中定义的集合 $\mathcal{A}$ 和任意下标集合 $\mathcal{S}_a$, 只要 $\mathcal{A} \subseteq \mathcal{S}_a$ 成立, 检验 (1.4) 和 (3.2) 就等价.

3.2 第二阶段: 在观测集 $\mathcal{D}_2$ 上的拟合优度检验

本小节在数据集 $\mathcal{D}_2$ 上对模型 (3.2) 提出一种基于投影的拟合优度检验.

定义 $\varepsilon \overset{\text{def}}{=} Y - m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}} \boldsymbol{x}_{\mathcal{S}_a})$, 当 (3.2) 中 H_0 成立时, 有

$$\mathrm{E}\{\varepsilon \mid \boldsymbol{x}_{\mathcal{S}}\} = 0. \tag{3.3}$$

(3.3) 等价于

$$\mathrm{E}(\varepsilon \mid \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{\mathcal{S}}) = 0 \text{ 对所有的 } \boldsymbol{\alpha} \text{ 满足 } \|\boldsymbol{\alpha}\| = 1, \tag{3.4}$$

其中 $\|\cdot\|$ 表示 Euclid 范数. 注意到 (3.4) 进一步地等价于

$$\mathrm{E}\{\varepsilon I(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{\mathcal{S}} \leqslant t)\} = 0 \ \text{ 对所有的 } \boldsymbol{\alpha} \text{ 满足 } \|\boldsymbol{\alpha}\| = 1 \ \text{ 和 } t \in \mathbb{R},$$

或者等价地写作

$$\int_{t \in \mathbb{R}, \|\boldsymbol{\alpha}\|=1} \mathrm{E}\{\varepsilon_1 \varepsilon_2 I(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{1,\mathcal{S}} \leqslant t) I(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{2,\mathcal{S}} \leqslant t)\} dF_{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{\mathcal{S}}}(t) d\boldsymbol{\alpha} = 0.$$

上式又可以等价地表示为

$$\int_{\|\boldsymbol{\alpha}\|=1} \mathrm{E}\{\varepsilon_1 \varepsilon_2 I(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{1,\mathcal{S}} \leqslant \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{3,\mathcal{S}}) I(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{2,\mathcal{S}} \leqslant \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{3,\mathcal{S}})\} d\boldsymbol{\alpha} = 0. \tag{3.5}$$

所以在检验 (3.2) 时, 只需要检验 (3.5) 是否成立. 而根据文献 [3], 有

$$\begin{aligned} &\int_{\|\boldsymbol{\alpha}\|=1} I(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{1,\mathcal{S}} \leqslant \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{3,\mathcal{S}}) I(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{2,\mathcal{S}} \leqslant \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_{3,\mathcal{S}}) d\boldsymbol{\alpha} \\ &\quad = \frac{\pi^{|\mathcal{S}|/2-1}}{\Gamma(|\mathcal{S}|/2+1)} \left| \pi - \arccos\left\{ \frac{(\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}})^{\mathrm{T}} (\boldsymbol{x}_{2,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}})}{\|\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}}\| \|\boldsymbol{x}_{2,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}}\|} \right\} \right|, \end{aligned}$$

其中 $\Gamma(\cdot)$ 是伽马函数. 于是定义

$$T \overset{\text{def}}{=} \mathrm{E}\left[\varepsilon_1 \varepsilon_2 \left| \pi - \arccos\left\{ \frac{(\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}})^{\mathrm{T}} (\boldsymbol{x}_{2,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}})}{\|\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}}\| \|\boldsymbol{x}_{2,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}}\|} \right\} \right| \right].$$

特别地, 当 $\boldsymbol{x}_{1,\mathcal{S}} = \boldsymbol{x}_{2,\mathcal{S}} \neq \boldsymbol{x}_{3,\mathcal{S}}$ 或者 $\boldsymbol{x}_{1,\mathcal{S}} = \boldsymbol{x}_{3,\mathcal{S}} \neq \boldsymbol{x}_{2,\mathcal{S}}$ 或者 $\boldsymbol{x}_{2,\mathcal{S}} = \boldsymbol{x}_{3,\mathcal{S}} \neq \boldsymbol{x}_{1,\mathcal{S}}$ 时, 定义

$$\left| \pi - \arccos\left\{ \frac{(\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}})^{\mathrm{T}} (\boldsymbol{x}_{2,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}})}{\|\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}}\| \|\boldsymbol{x}_{2,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}}\|} \right\} \right| = \pi,$$

当 $\boldsymbol{x}_{1,\mathcal{S}}=\boldsymbol{x}_{2,\mathcal{S}}=\boldsymbol{x}_{3,\mathcal{S}}$ 时, 定义上式的取值为 2π.

定义 T 的一个自然的估计量为

$$T_{n_2}\overset{\text{def}}{=}n_2^{-3}\sum_{i,j,k=n_1+1}^{n}\widehat{\varepsilon}_i\widehat{\varepsilon}_j\left|\pi-\arccos\left\{\frac{(\boldsymbol{x}_{i,\mathcal{S}}-\boldsymbol{x}_{k,\mathcal{S}})^{\mathrm{T}}(\boldsymbol{x}_{j,\mathcal{S}}-\boldsymbol{x}_{k,\mathcal{S}})}{\|\boldsymbol{x}_{i,\mathcal{S}}-\boldsymbol{x}_{k,\mathcal{S}}\|\|\boldsymbol{x}_{j,\mathcal{S}}-\boldsymbol{x}_{k,\mathcal{S}}\|}\right\}\right|, \tag{3.6}$$

其中 $\widehat{\varepsilon}\overset{\text{def}}{=}Y-m(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})$, 而 $\widehat{\boldsymbol{\theta}}_{\mathcal{S}}$ 是通过求解估计方程 (A.3) 得到的 $\boldsymbol{\theta}_{\mathcal{S}}$ 的普通最小二乘估计.

为了简化表达, 定义

$$D(\boldsymbol{x}_{1,\mathcal{S}},\boldsymbol{x}_{2,\mathcal{S}},\boldsymbol{x}_{3,\mathcal{S}})\overset{\text{def}}{=}\pi-\arccos\left\{\frac{(\boldsymbol{x}_{1,\mathcal{S}}-\boldsymbol{x}_{3,\mathcal{S}})^{\mathrm{T}}(\boldsymbol{x}_{2,\mathcal{S}}-\boldsymbol{x}_{3,\mathcal{S}})}{\|\boldsymbol{x}_{1,\mathcal{S}}-\boldsymbol{x}_{3,\mathcal{S}}\|\|\boldsymbol{x}_{2,\mathcal{S}}-\boldsymbol{x}_{3,\mathcal{S}}\|}\right\},$$

于是有 $T=\mathrm{E}[\varepsilon_1\varepsilon_2 D(\boldsymbol{x}_{1,\mathcal{S}},\boldsymbol{x}_{2,\mathcal{S}},\boldsymbol{x}_{3,\mathcal{S}})]$. 由于 $T\geqslant 0$ 恒成立并且当且仅当 (3.2) 中的 H_0 成立时 $T=0$, 所以在实际应用中, 我们倾向于在 T_{n_2} 较大时拒绝 (3.2) 中的原假设. 第 4 节会使用自助法决定该检验的临界值并证明此方法的相合性.

4 理论性质

本节考察上述检验方法的理论性质. 本文研究检验统计量 T_{n_2} 在原假设下的渐近分布, 提出使用野自助法去决定该检验的临界值并证明这种做法的合理性. 此外, 还考察检验统计量 T_{n_2} 在全局备择假设和局部备择假设下的渐近分布, 并探讨检验的功效问题.

首先讨论 T_{n_2} 在原假设下的渐近分布. 给定显著性水平 α, 确定检验 (3.2) 的临界值意味着需要研究清楚检验统计量 T_n 的渐近分布. 在此之前, 假定一些正则条件 (C1)–(C4) 如下:

(C1) $\sup_{l\in\mathcal{S}_a}\mathrm{E}(X_l^2)<\infty$ 对 $l=1,\ldots,|\mathcal{S}_a|$ 和 $\mathrm{E}(\varepsilon^2\mid\boldsymbol{x}_{\mathcal{S}})<\infty$ 成立.

(C2) $m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})$ 是在 $\mathbb{R}^{|\mathcal{S}_a|}$ 上 Borel 可测、在 $\mathbb{R}^{|\mathcal{S}_a|}$ 的紧子集上两阶连续可导的实函数.

(C3) 存在一个可积函数 $L(\boldsymbol{x}_{\mathcal{S}_a})$ 满足 $|g_l(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})|\leqslant L(\boldsymbol{x}_{\mathcal{S}_a})$ 对所有的 $\boldsymbol{\beta}_{\mathcal{S}_a}$ 和 $l=1,\ldots,|\mathcal{S}_a|$ 都成立, 其中, $g_l(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})$ 是 $\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})$ 的第 l 个元素, $\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\overset{\text{def}}{=}\partial m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})/\partial\boldsymbol{\beta}_{\mathcal{S}_a}$.

(C4) 定义 $\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}\overset{\text{def}}{=}\arg\min_{\boldsymbol{\beta}_{\mathcal{S}_a}\in\mathbb{R}^{|\mathcal{S}_a|}}\mathrm{E}[\{Y-m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^2]$. $\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}$ 是上述优化问题的唯一解并且是一个内点. 此外, $\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}]$ 在参数 $\boldsymbol{\beta}_{\mathcal{S}_a}$ 的真值处是正定的.

条件 (C1)–(C3) 用来保证参数 $\boldsymbol{\beta}_{\mathcal{S}_a}$ 的最小二乘估计的 $\sqrt{n}$ 相合性, 条件 (C4) 保证了相关估计量的渐近正态性.

定理 4.1 假定正则条件 (C1)–(C4) 成立, 当 (3.2) 中的 H_0 成立时, 有 $T=0$ 且

$$n_2T_{n_2}\xrightarrow{d}c^{-1}(|\mathcal{S}|)\int\zeta^2(\boldsymbol{\alpha},t)dF_{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha},\quad\text{当 } n_2\to\infty\text{ 时,}$$

其中 $c(|\mathcal{S}|)=\frac{\pi^{|\mathcal{S}|/2-1}}{\Gamma(|\mathcal{S}|/2+1)}$; $\zeta(\boldsymbol{\alpha},t)$ 是一个均值为 0、协方差函数为 $\mathrm{cov}\{\zeta(\boldsymbol{\alpha},t),\zeta(\boldsymbol{\alpha}_0,t_0)\}$ 的 Gauss 过程, 这里 $\mathrm{cov}\{\zeta(\boldsymbol{\alpha},t),\zeta(\boldsymbol{\alpha}_0,t_0)\}$ 的定义参见 (B.1); "$\xrightarrow{d}$" 表示 "依分布收敛".

考虑到检验统计量 T_{n_2} 在 H_0 下的渐近分布的特殊形式, 在给定显著性水平 α 的条件下, 本文使用自助法去决定检验 (3.2) 的临界值. 具体步骤如下所述:

步骤 1 在数据集 $\mathcal{D}_2$ 上, 估计参数 $\boldsymbol{\beta}_{\mathcal{S}_a}$ 并得到估计量 $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}$. 计算残差 $\widehat{\varepsilon}_i=Y_i-m(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})$, $i=n_1+1,\ldots,n$, 并根据 (3.6) 计算检验统计量 T_{n_2}.

步骤 2 定义 $Y_i^*=m(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})+\delta_i|\widehat{\varepsilon}_i|$, $i=n_1+1,\ldots,n$, 其中 δ_i $(i=n_1+1,\ldots,n)$ 是独立同分布的随机权重并且满足 $\Pr(\delta_i=1)=\Pr(\delta_i=-1)=1/2$.

步骤 3 基于新的数据集 $(\boldsymbol{x}_{i,\mathcal{S}}, Y_i^*)$, $i = n_1+1, \ldots, n$, 重复步骤 1 并得到新的检验统计量 $\widetilde{T}_{n_2}$.

步骤 4 将步骤 2 和 3 重复 B 次, 进而得到 B 个检验统计量 $\widetilde{T}_{n_2}^{(1)}, \widetilde{T}_{n_2}^{(2)}, \ldots, \widetilde{T}_{n_2}^{(B)}$.

定义检验的 p 值为

$$B^{-1}\sum_{b=1}^{B} I(\widetilde{T}_{n_2}^{(b)} \geqslant T_{n_2}),$$

其中 $I(\cdot)$ 是示性函数. 给定显著性水平 α, 在 p 值 $< \alpha$ 时拒绝 H_0. 下面的定理证明了该自助法的相合性.

定理 4.2 当定理 4.1 中的条件都成立时, $\widetilde{T}_{n_2}$ 是 n 相合的, 且其渐近分布为

$$n_2\widetilde{T}_{n_2} \xrightarrow{d} c^{-1}(|\mathcal{S}|)\int \zeta^2(\boldsymbol{\alpha}, t)dF_{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha}, \quad \text{当 } n_2 \to \infty \text{ 时}.$$

定理 4.2 意味着, 尽管 T_{n_2} 在原假设下的渐近分布形式棘手, 但可以使用自助法来近似该极限分布, 进而确定检验的 p 值.

接下来, 研究检验的功效问题. 考虑一列备择假设如下:

$$H_{1,n_2}: Y = m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a}) + C_{n_2}G(\boldsymbol{x}_{\mathcal{S}}) + \varepsilon, \tag{4.1}$$

其中, $G(\cdot)$ 是一个未知的光滑函数且满足 $\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})\} = 0$ 和 $\mathrm{E}\{G^2(\boldsymbol{x}_{\mathcal{S}})\} < \infty$, 残差项满足 $\mathrm{E}(\varepsilon \mid \boldsymbol{x}_{\mathcal{S}}) = 0$. 当 C_{n_2} 为一个非零的固定常数时, H_{1,n_2} 成为一个全局备择假设; 当 $C_{n_2} \to 0$ 时, H_{1,n_2} 是一个局部备择假设.

定理 4.3 假定正则条件 (C1)–(C4) 成立, 则

(1) 在检验 (3.2) 的全局备择假设下, ${n_2}^{1/2}(T_{n_2} - T)$ 依分布收敛到均值为 0、方差为 σ_0^2 的正态分布, 其中 $\sigma_0^2 \overset{\mathrm{def}}{=} \mathrm{var}(Z_1 + Z_2 + Z_3)$, 且 Z_1、Z_2 和 Z_3 的定义参见 (B.2)–(B.4).

(2) 在检验 (3.2) 的局部备择假设下, 若 $C_{n_2} = n_2^{-1/2}$, 则有

$$n_2T_{n_2} \xrightarrow{d} c^{-1}(|\mathcal{S}|)\int \zeta_0^2(\boldsymbol{\alpha}, t)dF_{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha}, \quad \text{当 } n_2 \to \infty \text{ 时},$$

其中 $\zeta_0(\boldsymbol{\alpha}, t)$ 是均值函数为 (B.5)、协方差函数为 $\mathrm{cov}\{\zeta_0(\boldsymbol{\alpha}, t), \zeta_0(\boldsymbol{\alpha}_0, t_0)\}$ (定义参见 (B.1)) 的 Gauss 过程.

从定理 4.3 中可以看出, 本文的方法不仅对于全局备择假设有功效, 同时对于收敛速度为 $n_2^{-1/2}$ 的局部备择假设也有功效.

5 数值分析

第 5.1 小节利用大量的 Monte Carlo 模拟详细展示两阶段检验方法的有限样本性质. 接着, 第 5.2 小节将这一检验方法应用到实际数据中, 以进一步验证其可行性.

5.1 Monte Carlo 模拟

固定样本量 $n = 200$, 协变量的维数 $p = 2{,}000$, 自助法次数 $B = 1{,}000$, 并且对每一个实验重复 1,000 次. 考虑以下 3 个实验:

$$\text{实验 (I)}: Y = X_1 + X_2 + c_0(X_1^2 + X_2^2) + \varepsilon,$$

$$\text{实验 (II)}: Y = 5X_1 + 5X_2 + 5X_3 + 2(X_4 + c_0X_4^2) + \varepsilon,$$
$$\text{实验 (III)}: Y = (2 + X_1 + X_2 + X_3)^2 + c_0(X_1^2 + X_2^2) + \varepsilon.$$

在这 3 个实验中, $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 来自于多元正态分布 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, 其中 $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{p\times p}$. 独立误差项 ε 来自于标准正态分布. 分别考察 c_0 取 0、0.2、0.4、0.6、0.8 和 1.0 的不同情形, 其中当且仅当 $c_0 = 0$ 时原假设 (1.2) 成立.

变量筛选结果 由定理 3.1 可知, 只要在第一阶段中保留了所有真正有用的变量, 检验 (1.4) 和 (3.2) 就等价. 因此, 首先需要考察有限样本水平上的一致筛选性是否成立.

本文在第一部分的数据集 $\mathcal{D}_1$ 上进行基于累积差异的边际筛选, 并采用以下两个标准评价筛选结果的好坏:

$\mathcal{M}$ 表示当所有有用的协变量都被选中时最小的模型大小. 我们给出的是基于 1,000 次重复实验所得到的 $\mathcal{M}$ 的 0%、25%、50%、75% 和 100% 的分位数.

$\mathcal{P}$ 表示给定最终的模型大小 $|\mathcal{S}|$, 所有有用的协变量同时被选中的经验概率, 其中 $\mathcal{S} = \mathcal{S}_a \cup \mathcal{S}_c$, $|\mathcal{S}|$ 是集合 $\mathcal{S}$ 中元素的个数. 分别考察选择的有用协变量的个数 $|\mathcal{S}_a| = 3, 6$ 和候选协变量的个数 $|\mathcal{S}_c| = 0, 1$ 共 4 种情形. 相应的结果列在表 4 中.

由于实验 (I)、(II) 和 (III) 中真正有用的协变量的个数分别为 2、4 和 3, 从表 4 中可看出, 在所有的情形下, 最小的模型大小 $\mathcal{M}$ 都非常接近真实的模型大小. 此外, 除了实验 (II) 中取 $|\mathcal{S}_a| = 3$ 和

表 4 基于累积差异的边际筛选的筛选结果. 考察 $\mathcal{M}$ 的 0%、25%、50%、75% 和 100% 的分位数, 以及 $|\mathcal{S}_a| = 3, 6$ 和 $|\mathcal{S}_c| = 0, 1$ 下的 $\mathcal{P}$

实验	c	$\mathcal{M}$					$\mathcal{P}$			
							$\lvert\mathcal{S}_c\rvert = 1$		$\lvert\mathcal{S}_c\rvert = 0$	
		0%	25%	50%	75%	100%	$\lvert\mathcal{S}_a\rvert = 3$	$\lvert\mathcal{S}_a\rvert = 6$	$\lvert\mathcal{S}_a\rvert = 3$	$\lvert\mathcal{S}_a\rvert = 6$
(I)	0.0	2	2	2	2	2	1.000	1.000	1.000	1.000
	0.2	2	2	2	2	2	1.000	1.000	1.000	1.000
	0.4	2	2	2	2	2	1.000	1.000	1.000	1.000
	0.6	2	2	2	2	2	1.000	1.000	1.000	1.000
	0.8	2	2	2	2	10	0.999	0.999	0.999	0.999
	1.0	2	2	2	2	37	0.986	0.997	0.986	0.997
(II)	0.0	4	4	4	4	50	1.000	1.000	0.000	0.966
	0.2	4	4	4	4	55	1.000	1.000	0.000	0.968
	0.4	4	4	4	4	71	1.000	1.000	0.000	0.970
	0.6	4	4	4	4	101	1.000	1.000	0.000	0.974
	0.8	4	4	4	4	130	0.999	1.000	0.000	0.973
	1.0	4	4	4	4	163	0.997	1.000	0.000	0.970
(III)	0.0	3	3	3	3	5	1.000	1.000	0.998	1.000
	0.2	3	3	3	3	6	1.000	1.000	0.997	1.000
	0.4	3	3	3	3	9	1.000	1.000	0.996	0.999
	0.6	3	3	3	3	10	1.000	1.000	0.993	0.999
	0.8	3	3	3	3	19	1.000	0.999	0.990	0.998
	1.0	3	3	3	3	44	0.999	0.999	0.988	0.998

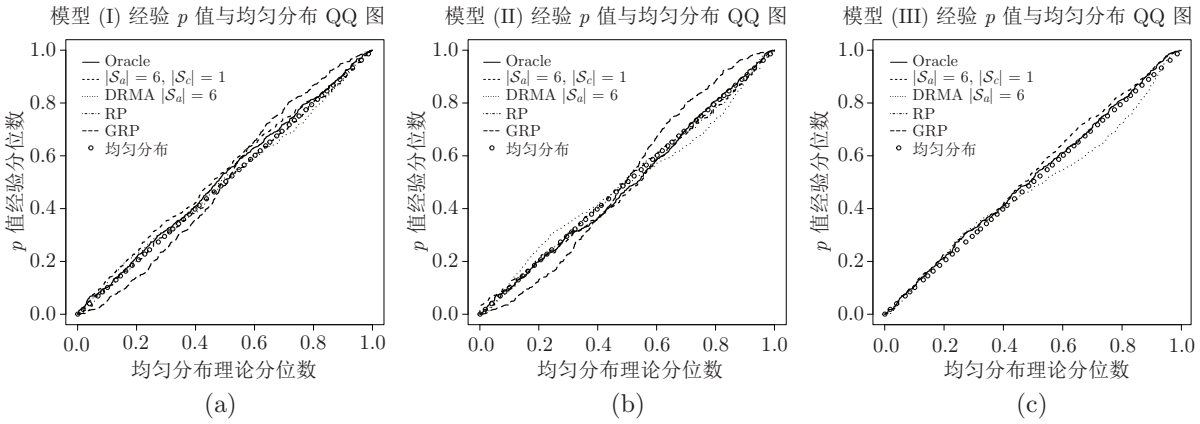

图 1 关于经验 p 值与均匀分布的 QQ 图: (a)、(b) 和 (c) 分别对应实验 (I)、(II) 和 (III). 每张图中展示 4 种不同的检验的结果: Oracle 检验 (实线), 改进后的取 $\mathcal{S}_a = 6$ 和 $\mathcal{S}_c = 1$ 的两阶段检验 (虚线), 取 $\mathcal{S}_a = 6$ 的 DRMA 检验 (点线)、RP 检验 (点虚线) 和 GPR 检验 (长虚线)

$|\mathcal{S}_c| = 0$ 的情形外, 第一阶段的筛选结果将所有有用的变量都保留下来的概率趋近于 1.

拟合优度检验结果 在检验 (3.2) 时, 本文比较检验统计量 T_{n_2} 和以下 3 种检验统计量的表现: 文献 [8] 中提出的基于降维的模型自适应的 (dimension-reduction model-adaptive, DRMA) 检验统计量 (考虑基于切片逆回归 (sliced inverse regression, SIR) 的 DEE 方法所得到的检验统计量), 文献 [19] 提出的基于残差投影的检验统计量 (residual prediction, RP), 文献 [10] 提出的针对高维广义线性模型基于残差投影的检验统计量 (generalized residual prediction, GRP). 由于不使用数据分组会导致两阶段检验出现第 I 类错误膨胀的问题, 所以这里只展示基于数据分组的检验结果.

此外, 为了缓解欠拟合所导致的功效损失, 我们建议在第一阶段的筛选中保留一个候选变量, 即 $|\mathcal{S}_c| = 1$. 同时也比较了无候选变量的情形, 即 $|\mathcal{S}_c| = 0$. 定义 Oracle 检验: 假定事先知道了真正有用的变量, 并给予这些变量进行后续检验. 这是在最理想的情形下得到的检验, 并将作为一个比较的基准. 考察不同的显著性水平 $\alpha = 0.01, 0.02, 0.05, 0.10$, 并使用第 4 节中介绍的自助法确定检验的 p 值. 基于 1,000 次重复实验的检验的第 I 类错误总结在表 5 中.

可以看到, 只要第一阶段的筛选过程将所有真正有用的协变量都保留下来, 本文的方法的第 I 类错误的概率就非常接近真实的水平 α. 对于实验 (II), 由于真正有用的变量个数为 4, 若在筛选阶段仅选择 3 个最重要的变量, 则至少会遗漏一个重要变量. 如果在第二阶段中, 直接基于这 3 个筛选出的变量进行后续检验, 即 $|\mathcal{S}_c| = 0$, 我们也会接受 Y 是关于 $\boldsymbol{x}_{\mathcal{S}_a}$ 线性的假设, 但仅用 3 个变量去预测 Y 会导致效率损失. 这时, 可以通过再添加一个候选变量, 即 $|\mathcal{S}_c| = 1$, 来避免这种情形. 根据表 5, 当实验 (II) 在第一阶段仅保留 3 个变量时, 我们拒绝零假设.

至于文献 [8] 提出的 DRMA 方法, 有限的经验表明它对所选窗宽非常敏感. 另外, 从图 1 中也能看出, 本文的方法的经验 p 值更接近于均匀分布. 基于残差投影的两类方法表现比较类似. 在模型 (I) 和 (II) 中, RP 检验对第 I 类错误的控制较好, 其概率非常接近预设的显著性水平 α. 相较之下, GRP 检验犯第一类错误的概率却略低于给定的显著性水平, 意味着检验的功效会受到一定程度的影响. 另外, 值得注意的是, 无论是 RP 检验还是 GRP 检验, 在非线性模型 (III) 中均不适用, 第一类错误概率均显著超出了给定的显著性水平 α, 这进一步显示了这两种检验方法在应用上的局限性.

下面固定显著性水平 $\alpha = 0.05$, 考察 $c_0 = 0.2, 0.4, 0.6, 0.8, 1.0$ 时检验的功效. 不同的检验功效总结的在表 6 中. 可以看到本文的检验方法随着 c_0 的增大, 功效也在随之上升, 尤其是在模型大小较小

时. 在实验 (II) 中, 当 $|\mathcal{S}_a| = 3$ 且 $|\mathcal{S}_c| = 0$ 时出现了功效损失, 这是因为在第一阶段中漏选了某些重要变量 ($|\mathcal{S}_a| = 3$), 并基于单次筛选出来的、并不完整的 ($|\mathcal{S}_c| = 0$) 的协变量进行了后续的检验. 这个结果与引言中声明的一致. 此外, 对于线性模型 (即实验 (I) 和 (II)), RP 检验的功效增长相对缓慢, 且整体上处于较低水平. 相比之下, 更一般的 GRP 检验方法显示出显著的功效增长趋势. 然而, 即便 GRP

表 5　给定显著性水平 α, 不同的检验方法基于 1,000 次的重复实验所得到的犯第 I 类错误 (当 $c_0 = 0$ 时) 的经验概率. 考虑 4 种不同的显著性水平 $\alpha = 0.01, 0.02, 0.05, 0.10$

实验	α	$\lvert\mathcal{S}_c\rvert = 1$		$\lvert\mathcal{S}_c\rvert = 0$		Oracle	DRMA	RP	GRP
		$\lvert\mathcal{S}_a\rvert = 3$	$\lvert\mathcal{S}_a\rvert = 6$	$\lvert\mathcal{S}_a\rvert = 3$	$\lvert\mathcal{S}_a\rvert = 6$		$\lvert\mathcal{S}_a\rvert = 6$		
(I)	0.01	0.010	0.013	0.015	0.013	0.018	0.006	0.005	0.003
	0.02	0.020	0.021	0.029	0.029	0.028	0.013	0.014	0.006
	0.05	0.057	0.065	0.061	0.060	0.060	0.049	0.045	0.017
	0.10	0.111	0.125	0.124	0.125	0.100	0.111	0.100	0.047
(II)	0.01	1.000	0.039	0.019	0.011	0.011	0.004	0.009	0.002
	0.02	1.000	0.044	0.026	0.017	0.019	0.010	0.015	0.007
	0.05	1.000	0.065	0.060	0.038	0.053	0.038	0.042	0.020
	0.10	1.000	0.105	0.118	0.083	0.094	0.112	0.092	0.037
(III)	0.01	0.010	0.007	0.006	0.008	0.010	0.001	0.395	0.843
	0.02	0.016	0.015	0.019	0.016	0.019	0.006	0.410	0.855
	0.05	0.045	0.049	0.046	0.049	0.051	0.040	0.438	0.870
	0.10	0.111	0.106	0.106	0.107	0.107	0.101	0.465	0.885

表 6　给定显著性水平 $\alpha = 0.05$, 不同的检验方法基于 1,000 次的重复实验所得到的检验的功效. 考察 $c_0 = 0.2, 0.4, 0.6, 0.8, 1.0$

实验	c	$\lvert\mathcal{S}_c\rvert = 1$		$\lvert\mathcal{S}_c\rvert = 0$		Oracle	DRMA	RP	GRP
		$\lvert\mathcal{S}_a\rvert = 3$	$\lvert\mathcal{S}_a\rvert = 6$	$\lvert\mathcal{S}_a\rvert = 3$	$\lvert\mathcal{S}_a\rvert = 6$		$\lvert\mathcal{S}_a\rvert = 6$		
(I)	0.2	0.294	0.213	0.560	0.424	0.641	0.076	0.284	0.054
	0.4	0.820	0.682	0.965	0.899	0.981	0.198	0.356	0.218
	0.6	0.978	0.909	0.998	0.996	1.000	0.317	0.363	0.464
	0.8	0.998	0.975	1.000	1.000	1.000	0.425	0.349	0.635
	1.0	0.992	0.994	1.000	1.000	1.000	0.545	0.385	0.759
(II)	0.2	1.000	0.252	0.078	0.425	0.565	0.071	0.193	0.071
	0.4	1.000	0.673	0.107	0.887	0.963	0.140	0.192	0.307
	0.6	1.000	0.892	0.145	0.971	0.998	0.212	0.208	0.569
	0.8	1.000	0.956	0.181	0.978	1.000	0.336	0.218	0.712
	1.0	0.999	0.977	0.219	0.979	1.000	0.507	0.223	0.777
(III)	0.2	0.628	0.636	0.654	0.646	0.630	0.072	–	–
	0.4	0.993	0.990	0.969	0.977	0.980	0.104	–	–
	0.6	1.000	1.000	0.996	1.000	1.000	0.245	–	–
	0.8	1.000	1.000	0.997	1.000	1.000	0.386	–	–
	1.0	1.000	0.999	0.998	1.000	1.000	0.472	–	–

方法有了这样的改进, 本文提出的检验方法仍然在这两种模型上展现出更高的功效.

5.2 实际数据分析

下面将两阶段检验方法应用到心肌病微阵列的数据集上. 这个数据集曾经被分别在文献 [15] 中使用 DC-SIS 方法和文献 [20] 中使用 MDC-SIS 方法分析过. 响应变量 Y 是基因 *Ro1* 的表达水平, 协变量 $\boldsymbol{x}=(X_1,\ldots,X_p)^{\mathrm{T}}$ 是其他的基因表达水平. 样本量 $n=30$, 协变量维数 $p=6{,}319$. 本文的目标是, 检验给定协变量 $\boldsymbol{x}$ 的条件下, Y 的条件均值是否是一个关于 $\boldsymbol{x}$ 的线性模型.

首先对单个协变量 X_k, $k=1,\ldots,p$ 进行边际标准化, 并对响应变量 Y 进行中心化. 接下来, 将样本数据随机地等分为两组, 并在第一组数据集上基于累积差异进行变量筛选. 本文的筛选结果显示基因 *Msa.2877.0* 和基因 *Msa.2134.0* 排在最前面, 这与 DC-SIS 和 MDC-SIS 的结果一致 (尽管基于 DC-SIS 的筛选结果的顺序略有不同). 在 Li 等 [15] 以及 Shao 和 Zhang [20] 的研究中, 他们推荐基于筛选出来的变量拟合一个可加模型. 然而考虑到线性模型的简易性和可解释性, 我们自然而然地想要探究: 最简单的线性模型是否足够刻画响应变量 Y 和筛选出来的基因 *Msa.2877.0* 与基因 *Msa.2134.0* 之间的回归关系? 即检验下述模型是否成立:

$$Y=\beta_1X_1+\beta_2X_2+\varepsilon, \tag{5.1}$$

其中 X_1 和 X_2 分别是基因 *Msa.2877.0* 和基因 *Msa.2134.0* 的表达水平.

对上述模型 (5.1) 进行基于普通最小二乘的参数估计, 考察参数检验与模型检验, 并将相关结果总结在表 7 中. 从表 7 中可以看出, 基因 *Msa.2877.0* 和基因 *Msa.2134.0* 的回归系数都是显著非零的. 此外, 调整后的 R 平方为 0.777, 而相应的 F 检验的 p 值为 0. 这些检验结果意味着线性模型 (5.1) 足够刻画出筛选出的基因 $(X_1,X_2)^{\mathrm{T}}$ 与响应变量 Y 之间的回归关系.

上述结果显然与 Li 等 [15] 以及 Shao 和 Zhang [20] 所声称的相互矛盾, 于是在第二组数据集上使用 (3.6) 中的检验统计量 T_{n_2}, 探究到底哪个结论更加合理. 本文选出候选基因 *Msa.15450.0*, 并将自助法的重复次数设定为 $B=1{,}000$, 然后计算出检验的 p 值为 0.007. 这意味着, 在给定显著性水平 $\alpha=0.05$ 的条件下, 拒绝线性模型 (5.1) 的假定. 这一结论与 Li 等 [15] 以及 Shao 和 Zhang [20] 的结论一致, 即

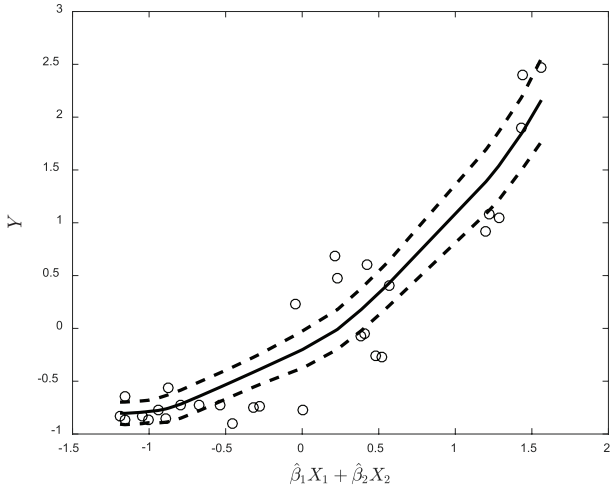

图 2 响应变量 $\boldsymbol{Y}$ 与拟合值 $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\boldsymbol{X}_{\mathbf{1}}+\widehat{\boldsymbol{\beta}}_{\mathbf{2}}\boldsymbol{X}_{\mathbf{2}}$ 之间的散点图. 实线是局部线性回归拟合, 虚线是 95% 的逐点置信带

表 7 模型 (5.1) 的参数估计及相应的检验的 p 值

	估计值	标准误	t 值	p 值
Msa.2877.0	0.65	0.13	5.02	2.9×10^{-5}
Msa.2134.0	0.29	0.13	2.21	3.5×10^{-2}

认为一个非参数的可加模型更适用于该数据集. 此外, 从图 2 中也可以看出, Y 与 $\widehat{\beta}_1X_1+\widehat{\beta}_2X_2$ 之间的确存在非线性关系.

6 扩展: 多重数据分组策略

在之前的分析中, 都是针对 (3.2) 进行假设检验, 而不是直接针对 (1.4), 这主要是因为确定筛选性保证了这两个检验在渐近的角度等价. 但正如在引言中提到的那样, 由于样本量有限或某些假设不满足, 在特征筛选的过程中, 可能会漏选一些重要的变量, 从而导致第 I 类错误膨胀. 为了解决上述问题, 可以在变量筛选的过程中多保留一些变量或使用迭代筛选的方法, 这样可以尽可能减小漏选重要变量的风险. 但是, 保留变量的增加, 可能会导致第二步模型检验的功效降低, 如表 1 所示. 类似文献 [18] 的想法, 本节讨论另外一种有效的解决方案. 具体而言, 对样本数据集进行多次 (如 M 次) 随机分组, 对于每次分组后的样本, 使用之前讨论的检验方法计算得到一个 p 值. 这样, 能够获得 M 个 p 值, 将它们分别记为 $p_1,\ldots,p_M$. 对于每个固定的 γ, 对 p 值作如下修正:

$$Q(\gamma)=\min\left[1,q_\gamma\left(\frac{p_i}{\gamma}\right)\right],$$

其中 $\gamma\in(\gamma_{\min},1)$, 且 $q_\gamma(\{p_i/\gamma\})$ 是 $\{p_i/\gamma\}$, $i=1,\ldots,M$ 的 γ 分位数. 这里的 $\gamma_{\min}\in(0,1)$ 是 γ 可能取到的下界, 在实际中, 一般将其设为 0.05 或者 $1/M$. 但是, 由于这个修正后的 p 值, $Q(\gamma)$ 取决于 γ 的选取, 而在实践中可能很难正确地选择 γ. 因此, 进一步定义自适应 p 值如下:

$$Q^*=\min\left\{1,(1-\log\gamma_{\min})\inf_{\gamma\in(\gamma_{\min},1)}Q(\gamma)\right\}.$$

接下来的定理表明, 基于修正后的 p 值或自适应的 p 值进行检验, 可以将第 I 类错误渐近地控制在给定的显著性水平.

定理 6.1 假设 $\lim_{n\to\infty}\Pr(\mathcal{A}\subseteq\mathcal{S}_{a,i})=1$, 其中 $\mathcal{S}_{a,i}$ 是根据第 i 次对样本进行分割时, 在特征筛选阶段挑选出来的变量的下标, $i=1,\ldots,M$, 则

$$\lim_{n\to\infty}\sup\Pr\{Q(\gamma)\leqslant\alpha\}\leqslant\alpha,\quad \lim_{n\to\infty}\sup\Pr\{Q^*(\gamma)\leqslant\alpha\}\leqslant\alpha.$$

下面通过一个具体的例子来说明, 当某些重要的变量以不可忽略的概率被漏选时, 本文提出的多重数据分组的方法可以将第 I 类错误控制在给定的 α 水平. 令

$$Y=0.3X_1+2X_2+\varepsilon,$$

其中 $\boldsymbol{x}=(X_2,\ldots,X_p)^{\mathrm{T}}$ 来自于均值为 $\mathbf{0}$、协方差矩阵为 $\boldsymbol{\Sigma}=(\sigma_{kl})_{(p-1)\times(p-1)}$ 的多元正态, 这里 $\sigma_{kl}=0.5^{|k-l|}$, $k,l=2,\ldots,p$. 定义 $X_1=(1-X_2)^2+\varepsilon_1$, 独立误差项 ε 服从标准正态分布, ε_1 与 ε 独

表 8 基于单次数据分组的检验和多重数据分组的检验得到的第 I 类错误的比较

	单次数据分组				多重数据分组			
α	0.01	0.02	0.05	0.10	0.01	0.02	0.05	0.10
第 I 类错误	0.050	0.072	0.149	0.246	0.000	0.010	0.050	0.081

立同分布. 固定样本量 $n = 200$, 协变量维数 $p = 2{,}000$, 每次筛选 $|\mathcal{S}_a| = 3$ 个有用变量, 1 个候选变量, 每次检验中自助法次数为 200 次, 并重复数据分组过程 $M = 50$ 次.

若只基于单次数据分组进行检验, 则在 1,000 次的重复实验中, 重要变量 X_1 和 X_2 分别会被漏选 262 和 0 次, 这说明在变量筛选的过程中, X_1 会以不可忽略的比例被漏选. 从表 8 中可以看到, 如果只基于单次数据分组进行检验, X_1 的漏选会导致第 I 类错误显著膨胀. 若采用多重数据分组的策略, 可以很好地将第 I 类错误维持在名义水平 (尤其当水平 $\alpha = 0.05$ 和 $\alpha = 0.10$ 时), 这说明多重数据分组的策略可以显著地改善因重要变量漏选所导致的第 I 类错误膨胀的问题.

致谢 感谢审稿人对本文提出的许多优秀的修改意见.

参考文献

1 Chang J, Tang C Y, Wu Y. Marginal empirical likelihood and sure independence feature screening. Ann Statist, 2013, 41: 2123–2148
2 Dezeure R, Bühlmann P, Meier L, et al. High-dimensional inference: Confidence intervals, p-values and R-software hdi. Statist Sci, 2015, 30: 533–558
3 Escanciano J C. A consistent diagnostic test for regression models using projections. Econom Theory, 2006, 22: 1030–1051
4 Fan J, Guo S, Hao N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. J R Stat Soc Ser B Stat Methodol, 2012, 74: 37–65
5 Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol, 2008, 70: 849–911
6 Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Statist, 2010, 38: 3567–3604
7 González-Manteiga W, Crujeiras R M. An updated review of goodness-of-fit tests for regression models. TEST, 2013, 22: 361–411
8 Guo X, Wang T, Zhu L X. Model checking for parametric single-index models: A dimension reduction model-adaptive approach. J R Stat Soc Ser B Stat Methodol, 2016, 78: 1013–1035
9 Hardle W, Mammen E. Comparing nonparametric versus parametric regression fits. Ann Statist, 1993, 21: 1926–1947
10 Janková J, Shah R D, Bühlmann P, et al. Goodness-of-fit testing in high dimensional generalized linear models. J R Stat Soc Ser B Stat Methodol, 2020, 82: 773–795
11 Jennrich R I. Asymptotic properties of non-linear least squares estimators. Ann of Math Stud, 1969, 40: 633–643
12 Khmaladze E V, Koul H L. Goodness-of-fit problem for errors in nonparametric regression: Distribution free approach. Ann Statist, 2009, 37: 3165–3185
13 Lavergne P, Patilea V. Breaking the curse of dimensionality in nonparametric testing. J Econometrics, 2008, 143: 103–122
14 Li G, Peng H, Zhang J, et al. Robust rank correlation based screening. Ann Statist, 2012, 40: 1846–1877
15 Li R Z, Zhong W, Zhu L P. Feature screening via distance correlation learning. J Amer Statist Assoc, 2012, 107: 1129–1139
16 Liu J Y, Zhong W, Li R Z. A selective overview of feature screening for ultrahigh-dimensional data. Sci China Math, 2015, 58: 1–22
17 Mai Q, Zou H. The Kolmogorov filter for variable screening in high-dimensional binary classification. Biometrika, 2012, 100: 229–234
18 Meinshausen N, Meier L, Bühlmann P. p-values for high-dimensional regression. J Amer Statist Assoc, 2009, 104: 1671–1681
19 Shah R D, Bühlmann P. Goodness-of-fit tests for high dimensional linear models. J R Stat Soc Ser B Stat Methodol, 2018, 80: 113–135

20 Shao X, Zhang J. Martingale difference correlation and its use in high-dimensional variable screening. J Amer Statist Assoc, 2014, 109: 1302–1318
21 Stute W. Nonparametric model checks for regression. Ann Statist, 1997, 25: 613–641
22 Stute W, Thies S, Zhu L X. Model checks for regression: An innovation process approach. Ann Statist, 1998, 26: 1916–1934
23 Stute W, Zhu L X. Model checks for generalized linear models. Scand J Stat, 2002, 29: 535–545
24 Stute W, Zhu L X. Nonparametric checks for single-index models. Ann Statist, 2005, 33: 1048–1083
25 Xia Y. Model checking in regression via dimension reduction. Biometrika, 2009, 96: 133–148
26 Zhang X, Yao S, Shao X. Conditional mean and quantile dependence testing in high dimension. Ann Statist, 2018, 46: 219–246
27 Zhang Y, Zhong W, Zhu L P. A lack-of-fit test with screening in sufficient dimension reduction. Statist Sinica, 2020, 30: 1971–1993
28 Zhang Y, Zhou Y, Zhu L P. A post-screening diagnostic study for ultrahigh dimensional data. J Econometrics, 2024, 239: 105354
29 Zheng J X. A consistent test of functional form via nonparametric estimation techniques. J Econometrics, 1996, 75: 263–289
30 Zhou T Y, Zhu L P. Cumulative divergence revisited (in Chinese). Sci Sin Math, 2021, 51: 2049–2064 [周亭攸, 朱利平. 关于累积差异的再次讨论. 中国科学: 数学, 2021, 51: 2049–2064]
31 Zhou T Y, Zhu L P, Xu C, et al. Model-free forward screening via cumulative divergence. J Amer Statist Assoc, 2020, 115: 1393–1405
32 Zhu L P, Li L, Li R, et al. Model-free feature screening for ultrahigh-dimensional data. J Amer Statist Assoc, 2011, 106: 1464–1475

附录 A 引理

引理 A.1 在局部备择假设 (4.1) 下, 若正则条件 (C1)–(C4) 成立, 则有

$$\begin{aligned} n_2^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}-\boldsymbol{\beta}_{\mathcal{S}_a}) &= (\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}n_2^{-1/2}\sum_{i=n_1+1}^{n}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\varepsilon_i \\ &\quad+\left(n_2^{-1}\sum_{i=n_1+1}^{n}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\}^{\mathrm{T}}]\right)^{-1}C_{n_2}n_2^{1/2}\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\} \\ &\quad+o_p(1), \end{aligned} \tag{A.1}$$

其中 $\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\overset{\mathrm{def}}{=}\partial m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})/\partial\boldsymbol{\beta}_{\mathcal{S}_a}$.

特别地, 当 $C_{n_2}=0$ 时, 局部备择假设 (4.1) 退化成检验 (3.2) 中的原假设, 则有

$$\begin{aligned} n_2^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}-\boldsymbol{\beta}_{\mathcal{S}_a}) &= (\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}n_2^{-1/2}\sum_{i=n_1+1}^{n}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\varepsilon_i \\ &\quad+o_p(1). \end{aligned} \tag{A.2}$$

证明 首先验证 $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}$ 是参数 $\boldsymbol{\beta}_{\mathcal{S}_a}$ 的一个强相合估计. 证明的细节与文献 [11, 定理 6 和 7] 类似, 因此省略. $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}$ 一定满足如下方程:

$$n_2^{-1}\sum_{i=n_1+1}^{n}[\{Y_i-m(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\}\mathbf{g}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})]=\mathbf{0}. \tag{A.3}$$

等号左边应用 Lagrange 定理可得

$$n_2^{-1}\sum_{i=n_1+1}^{n}[\{Y_i-m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})]+A_{n_2}(\boldsymbol{\beta}_{\mathcal{S}_a}^{*})(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}-\boldsymbol{\beta}_{\mathcal{S}_a})=\mathbf{0}.$$

其中 $\boldsymbol{\beta}^*_{\mathcal{S}_a}$ 落于 $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}$ 和 $\boldsymbol{\beta}_{\mathcal{S}_a}$ 之间, 且 $A_{n_2}(\boldsymbol{\beta}^*_{\mathcal{S}_a}) \overset{\text{def}}{=} n_2^{-1}\sum_{i=n_1+1}^{n}[\{Y_i - m(\boldsymbol{\beta}^{*\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})\}\partial\mathbf{g}(\boldsymbol{\beta}^{*\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})/\partial\boldsymbol{\beta}^{*\mathrm{T}}_{\mathcal{S}_a} - \mathbf{g}(\boldsymbol{\beta}^{*\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}^{*\mathrm{T}}_{\mathcal{S}_a}{}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\}^{\mathrm{T}}]$. 于是, 有

$$(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a} - \boldsymbol{\beta}_{\mathcal{S}_a}) = -\{A_{n_2}(\boldsymbol{\beta}^*_{\mathcal{S}_a})\}^{-1}\cdot n_2^{-1}\sum_{i=n_1+1}^{n}[\{Y_i - m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})\}\mathbf{g}(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})].$$

在 H_{1,n_2} 下有 $Y_i = m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a}) + C_{n_2}G(\boldsymbol{x}_{\mathcal{S}}) + \varepsilon_i$, 进而得到结论 (A.1). 剩下的证明细节与文献 [8, 引理 3] 类似, 因此省略. □

附录 B 技术证明

定理 3.1 的证明 由于 $\mathcal{S}_a \subset \mathcal{S}$, 所以在稀疏性原则下有 $\mathrm{E}(Y \mid \boldsymbol{x}) = \mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{A}})$ 成立, 再利用重期望定理, 有

$$\mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}}) = \mathrm{E}\{\mathrm{E}(Y \mid \boldsymbol{x}) \mid \boldsymbol{x}_{\mathcal{S}}\} = \mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{A}})$$

对所有满足 $\mathcal{A} \subseteq \mathcal{S}_a$ 的集合 $\mathcal{S}_a$ 都成立. 一方面, 如果 $\Pr\{\mathrm{E}(Y \mid \boldsymbol{x}) = m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}})\} = 1$ 成立, 则通过选择 $\boldsymbol{\beta}_{\mathcal{S}_a} = (\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{A}}, \mathbf{0}_{1\times(|\mathcal{S}_a|-|\mathcal{A}|)})^{\mathrm{T}}$, 有 $\mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}}) = m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}}) = m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{\mathcal{S}_a})$ 成立.

另一方面, 如果 $\Pr\{\mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}}) = m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{\mathcal{S}_a})\} = 1$ 成立, 由于已经知道 $\mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}}) = \mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{A}})$ 对所有满足 $\mathcal{A} \subseteq \mathcal{S}_a$ 的集合 $\mathcal{S}_a$ 都成立, 则参数 θ_k, $k \in \mathcal{S}\cap\mathcal{A}^c$ 一定为 0, 即 $\mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}})$ 仅仅只依赖于 $\boldsymbol{x}_{\mathcal{A}}$, 这确保了 $\Pr\{\mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}}) = m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}})\} = 1$. 此外, 由于 $\Pr\{\mathrm{E}(Y \mid \boldsymbol{x}) = \mathrm{E}(Y \mid \boldsymbol{x}_{\mathcal{S}})\} = 1$, 于是有 $\Pr\{\mathrm{E}(Y \mid \boldsymbol{x}) = m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{A}}\boldsymbol{x}_{\mathcal{A}})\} = 1$ 成立.

至此完成了定理 3.1 的证明. □

定理 4.1 的证明 由于 $\widehat{\varepsilon}_i \overset{\text{def}}{=} Y_i - m(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})$, 于是将其改写为

$$\widehat{\varepsilon}_i = \varepsilon_i + \{m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a}) - m(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})\}.$$

定义经验过程

$$\zeta_{n_2}(\boldsymbol{\alpha}, t) \overset{\text{def}}{=} n_2^{-1/2}\sum_{i=n_1+1}^{n}\widehat{\varepsilon}_i I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}} \leqslant t) \overset{\text{def}}{=} V_1(\boldsymbol{\alpha}, t) + V_2(\boldsymbol{\alpha}, t),$$

其中 $V_1(\boldsymbol{\alpha}, t) \overset{\text{def}}{=} n_2^{-1/2}\sum_{i=n_1+1}^{n}\varepsilon_i I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}} \leqslant t)$ 且

$$V_2(\boldsymbol{\alpha}, t) \overset{\text{def}}{=} n_2^{-1/2}\sum_{i=n_1+1}^{n}\{m(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a}) - m(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})\}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}} \leqslant t).$$

注意到 $V_1(\boldsymbol{\alpha}, t)$ 是 n_2 个独立同分布且均值为 0 的随机变量之和, 首先处理 $V_2(\boldsymbol{\alpha}, t)$. 在引理 A.1 中已经证明了

$$n_2^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a} - \boldsymbol{\beta}_{\mathcal{S}_a}) = n_2^{-1/2}(\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}\sum_{i=n_1+1}^{n}\mathbf{g}(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})\varepsilon_i + o_p(1),$$

利用 Taylor 展开, 有

$$V_2(\boldsymbol{\alpha}, t) = n_2^{-1/2}\sum_{i=n_1+1}^{n}\{\mathbf{g}(\boldsymbol{\beta}^{\mathrm{T}}_{\mathcal{S}_a}\boldsymbol{x}_{i,\mathcal{S}_a})\}^{\mathrm{T}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}} \leqslant t)(\boldsymbol{\beta}_{\mathcal{S}_a} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}) + o_p(1)$$

$$= -\mathrm{E}[\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}} I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}} \leqslant t)](\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}$$
$$\cdot\, n_2^{-1/2}\sum_{i=n_1+1}^{n}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\varepsilon_i + o_p(1).$$

由于 $\zeta_{n_2}(\boldsymbol{\alpha},t)=\sum_{k=1}^{2}V_k(\boldsymbol{\alpha},t)$ 且 $V_1(\boldsymbol{\alpha},t)$ 是 n_2 个独立同分布且均值为 0 的随机变量之和, 在原假设下容易得到 $\mathrm{E}\{\zeta_{n_2}(\boldsymbol{\alpha},t)\}=o(1)$ 当 n_2 趋于无穷时成立. 此外,

$$\zeta_{n_2}(\boldsymbol{\alpha},t)\zeta_{n_2}(\boldsymbol{\alpha}_0,t_0) = V_1(\boldsymbol{\alpha},t)V_1(\boldsymbol{\alpha}_0,t_0)+V_2(\boldsymbol{\alpha},t)V_2(\boldsymbol{\alpha}_0,t_0)$$
$$+V_1(\boldsymbol{\alpha},t)V_2(\boldsymbol{\alpha}_0,t_0)+V_2(\boldsymbol{\alpha},t)V_1(\boldsymbol{\alpha}_0,t_0).$$

可计算出上式等号右边第一项的期望为

$$\mathrm{E}\{V_1(\boldsymbol{\alpha},t)V_1(\boldsymbol{\alpha}_0,t_0)\}=\mathrm{E}[\varepsilon^2 I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)I(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t_0)].$$

注意到 $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}$ 是渐近正态的, 即 $n_2^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}-\boldsymbol{\beta}_{\mathcal{S}_a})$ 依分布收敛到一个零均值的正态随机变量. 于是有

$$\mathrm{E}\{V_2(\boldsymbol{\alpha},t)V_2(\boldsymbol{\alpha}_0,t_0)\}=\mathrm{E}[\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}_a})\}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}_a})\cdot I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}}\leqslant t)I(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}}\leqslant t_0)]+o(1).$$

再一次使用引理 A.1 来重新表示 $(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}-\boldsymbol{\beta}_{\mathcal{S}_a})$, 得到

$$\mathrm{E}\{V_1(\boldsymbol{\alpha},t)V_2(\boldsymbol{\alpha}_0,t_0)\} = -\mathrm{E}(\varepsilon_2^2\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}_a})\}^{\mathrm{T}}(\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}$$
$$\cdot\,\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}_a})I(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}}\leqslant t_0,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}}\leqslant t))+o(1),$$
$$\mathrm{E}\{V_2(\boldsymbol{\alpha},t)V_1(\boldsymbol{\alpha}_0,t_0)\} = -\mathrm{E}(\varepsilon_2^2\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}_a})\}^{\mathrm{T}}(\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}$$
$$\cdot\,\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}_a})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}}\leqslant t,\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}}\leqslant t_0))+o(1).$$

记 $\zeta(\boldsymbol{\alpha},t)$ 为一个均值为 0、协方差函数为 $\mathrm{cov}\{\zeta(\boldsymbol{\alpha},t),\zeta(\boldsymbol{\alpha}_0,t_0)\}$ 的 Gauss 过程, 其中 $\mathrm{cov}\{\zeta(\boldsymbol{\alpha},t),\zeta(\boldsymbol{\alpha}_0,t_0)\}$ 的形式如下:

$$\begin{aligned}&\mathrm{E}[\varepsilon^2 I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)I(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t_0)]+\mathrm{E}[\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}_a})\}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}_a})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}}\leqslant t,\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}}\leqslant t_0)]\\&\quad-\mathrm{E}[\varepsilon_2^2\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}_a})\}^{\mathrm{T}}(\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}_a})I(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}}\leqslant t_0,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}}\leqslant t)]\\&\quad-\mathrm{E}[\varepsilon_2^2\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}_a})\}^{\mathrm{T}}(\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}_a})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{1,\mathcal{S}}\leqslant t,\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}}\leqslant t_0)].\end{aligned}\tag{B.1}$$

特别地, 取 $\boldsymbol{\alpha}=\boldsymbol{\alpha}_0$ 和 $t=t_0$ 时, 有

$$\mathrm{E}\{\zeta_{n_2}^2(\boldsymbol{\alpha},t)\}=\mathrm{cov}\{\zeta(\boldsymbol{\alpha},t),\zeta(\boldsymbol{\alpha},t)\}+o(1),$$

于是得到结论: $\zeta_{n_2}(\boldsymbol{\alpha},t)$ 依分布收敛到 $\zeta(\boldsymbol{\alpha},t)$, 进而当 $n_2\to\infty$ 时, 有

$$c(|\mathcal{S}|)n_2T_{n_2}=\int\zeta_{n_2}^2(\boldsymbol{\alpha},t)dF_{n_2,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha}\xrightarrow{d}\int\zeta^2(\boldsymbol{\alpha},t)dF_{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha}.$$

至此完成了定理 4.1 的证明. □

定理 4.2 的证明 由于自助法产生 $Y^*=m(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})+\widetilde{\varepsilon}$, 其中 $\widetilde{\varepsilon}\overset{\mathrm{def}}{=}\delta|\widehat{\varepsilon}|$. 由于 δ 与 $\boldsymbol{x}$ 和 ε 独立, 所以有 $\mathrm{E}(\widetilde{\varepsilon}\mid\boldsymbol{x}_{\mathcal{S}})=0$, 即原假设自动地成立.

与定理 4.1 中的证明类似, 定义经验过程如下:

$$\widetilde{\zeta}_{n_2}(\boldsymbol{\alpha},t) \overset{\text{def}}{=} n_2^{-1/2}\sum_{i=n_1+1}^{n}\widehat{\widetilde{\varepsilon}}_i I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}}\leqslant t).$$

将 $\widetilde{\zeta}_{n_2}(\boldsymbol{\alpha},t)$ 分解成 $\widetilde{V}_1(\boldsymbol{\alpha},t)+\widetilde{V}_2(\boldsymbol{\alpha},t)$ 两部分, 其中

$$\begin{aligned}\widetilde{V}_1(\boldsymbol{\alpha},t) &= n_2^{-1/2}\sum_{i=1}^{n}\widetilde{\varepsilon}_i I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}}\leqslant t),\\ \widetilde{V}_2(\boldsymbol{\alpha},t) &= -\mathrm{E}[\{\mathbf{g}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}} I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)](\mathrm{E}[\mathbf{g}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}\\ &\quad\cdot n_2^{-1/2}\sum_{i=1}^{n}\mathbf{g}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\widetilde{\varepsilon}_i + o_p(1).\end{aligned}$$

与定理 4.1 的证明类似, 可得

$$\begin{aligned}\mathrm{E}\{\widetilde{V}_1(\boldsymbol{\alpha},t)\widetilde{V}_1(\boldsymbol{\alpha}_0,t_0)\} &= \mathrm{E}\{V_1(\boldsymbol{\alpha},t)V_1(\boldsymbol{\alpha}_0,t_0)\}+o(1),\\ \mathrm{E}\{\widetilde{V}_2(\boldsymbol{\alpha},t)\widetilde{V}_2(\boldsymbol{\alpha}_0,t_0)\} &= \mathrm{E}\{V_2(\boldsymbol{\alpha},t)V_2(\boldsymbol{\alpha}_0,t_0)\}+o(1).\end{aligned}$$

进而有

$$\begin{aligned}\mathrm{E}\{\widetilde{V}_2(\boldsymbol{\alpha},t)\widetilde{V}_1(\boldsymbol{\alpha}_0,t_0)\} &= \mathrm{E}\{V_2(\boldsymbol{\alpha},t)V_1(\boldsymbol{\alpha}_0,t_0)\}+o(1),\\ \mathrm{E}\{\widetilde{V}_1(\boldsymbol{\alpha},t)\widetilde{V}_2(\boldsymbol{\alpha}_0,t_0)\} &= \mathrm{E}\{V_1(\boldsymbol{\alpha},t)V_2(\boldsymbol{\alpha}_0,t_0)\}+o(1),\end{aligned}$$

于是得到

$$\mathrm{E}\{\widetilde{\zeta}_{n_2}(\boldsymbol{\alpha},t)\widetilde{\zeta}_{n_2}(\boldsymbol{\alpha}_0,t_0)\} = \mathrm{E}\{\zeta_{n_2}(\boldsymbol{\alpha},t)\zeta_{n_2}(\boldsymbol{\alpha}_0,t_0)\}+o(1).$$

类似于定理 4.1 中的讨论, 最终得到

$$n_2 c(|\mathcal{S}|)\widetilde{T}_{n_2}\xrightarrow{d}\int\zeta^2(\boldsymbol{\alpha},t)dF_{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha}.$$

至此完成了定理 4.2 的证明. □

定理 4.3 的证明 在全局备择假设下, 首先记 $\mathrm{E}(Y\mid\boldsymbol{x}_{\mathcal{S}}) = m(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})+G(\boldsymbol{x}_{\mathcal{S}})$, 其中参数 $\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}$ 与原假设下参数的真值不同. 在全局备择假设下得到的参数 $\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}$ 的估计量 $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}$ 是 $\sqrt{n_2}$ 相合的. 与原假设下的讨论类似, 将经验过程分解如下:

$$\begin{aligned}V_1(\boldsymbol{\alpha},t) &= n_2^{-1/2}\sum_{i=n_1+1}^{n}\{G(\boldsymbol{x}_{i,\mathcal{S}})+\varepsilon_i\}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}}\leqslant t),\\ V_2(\boldsymbol{\alpha},t) &= -\mathrm{E}[\{\mathbf{g}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}} I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)](\mathrm{E}[\mathbf{g}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}\\ &\quad\cdot n_2^{-1/2}\sum_{i=n_1+1}^{n}\mathbf{g}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\{G(\boldsymbol{x}_{i,\mathcal{S}})+\varepsilon_i\}+o_p(1).\end{aligned}$$

由于 $\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})\mid\boldsymbol{x}_{\mathcal{S}}\}=0$ 不成立, 所以有

$$\zeta_{n_2}^2(\boldsymbol{\alpha},t) = 2n_2^{1/2}[\zeta_{n_2}(\boldsymbol{\alpha},t)-n_2^{1/2}\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)\}]\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)\}$$

$$+ n_2\mathrm{E}^2\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}} \leqslant t)\} + O_p(1).$$

由 $\zeta_{n_2}(\boldsymbol{\alpha}, t) = V_1(\boldsymbol{\alpha}, t) + V_2(\boldsymbol{\alpha}, t)$, 经过一些计算可得

$$\begin{aligned}&\int 2[V_1(\boldsymbol{\alpha}, t) - n_2^{1/2}\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}} \leqslant t)\}]\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}} \leqslant t)\}dF_{n_2,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha}\\ &= n_2^{-1/2}\sum_{i=n_1+1}^{n} Z_{1,i} + o_p(1),\end{aligned}$$

其中 $Z_{1,i}$, $i = n_1+1, \ldots, n$ 是 n_2 个如下随机变量的独立副本:

$$2c(|\mathcal{S}|)\{G(\boldsymbol{x}_{i,\mathcal{S}}) + \varepsilon_i\}\mathrm{E}\left[G(\boldsymbol{x}_{1,\mathcal{S}})\middle|\pi - \arccos\left\{\frac{(\boldsymbol{x}_{i,\mathcal{S}} - \boldsymbol{x}_{2,\mathcal{S}})^{\mathrm{T}}(\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{2,\mathcal{S}})}{\|\boldsymbol{x}_{i,\mathcal{S}} - \boldsymbol{x}_{2,\mathcal{S}}\|\|\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{2,\mathcal{S}}\|}\right\}\right] - 2T. \tag{B.2}$$

类似地, 可推导出

$$\int 2V_2(\boldsymbol{\alpha}, t)\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}} \leqslant t)\}dF_{n_2,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha} = n_2^{-1/2}\sum_{i=n_1+1}^{n} Z_{2,i} + o_p(1),$$

其中 $Z_{2,i}$, $i = n_1+1, \ldots, n$ 是 n_2 个如下随机变量的独立副本:

$$\begin{aligned}&2c(|\mathcal{S}|)\mathrm{E}\left[G(\boldsymbol{x}_{1,\mathcal{S}})\{\mathbf{g}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{2,\mathcal{S}_a})\}^{\mathrm{T}}\middle|\pi - \arccos\left\{\frac{(\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}})^{\mathrm{T}}(\boldsymbol{x}_{2,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}})}{\|\boldsymbol{x}_{1,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}}\|\|\boldsymbol{x}_{2,\mathcal{S}} - \boldsymbol{x}_{3,\mathcal{S}}\|}\right\}\right]\\ &\cdot (\mathrm{E}[\mathbf{g}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}\mathbf{g}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\{G(\boldsymbol{x}_{i,\mathcal{S}}) + \varepsilon_i\}.\end{aligned} \tag{B.3}$$

此外, 还可以证明

$$\begin{aligned}\int \mathrm{E}^2\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}} \leqslant t)\}dF_{n_2,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha} &= \int \mathrm{E}^2\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}} \leqslant t)\}dF_{n_2,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha}\\ &= n_2^{-1}\sum_{i=n_1+1}^{n}\int \mathrm{E}^2\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}} \leqslant \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}})\}d\boldsymbol{\alpha}\\ &\overset{\mathrm{def}}{=} n_2^{-1}\sum_{i=n_1+1}^{n} Z_{3,i}.\end{aligned} \tag{B.4}$$

将上述结果整理起来, 可将检验统计量 T_{n_2} 简化为

$$T_{n_2} - T = n_2^{-1}\sum_{i=n_1+1}^{n}(Z_{1,i} + Z_{2,i} + Z_{3,i}) + o_p(n_2^{-1/2}),$$

即渐近地是 n_2 个独立同分布的随机变量的均值. 根据中心极限定理, 我们完成了全局备择假设下的证明部分.

在局部备择假设 H_{1,n_2} 下, 有 $Y = m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a}) + C_nG(\boldsymbol{x}_{\mathcal{S}}) + \varepsilon$. 尤其当 $C_{n_2} = n_2^{-1/2}$ 时, 可将 $\zeta_{n_2}(\boldsymbol{\alpha}, t)$ 分解为 $V_1(\boldsymbol{\alpha}, t) + V_2(\boldsymbol{\alpha}, t)$, 其中

$$\begin{aligned}V_1(\boldsymbol{\alpha}, t) &= n_2^{-1/2}\sum_{i=1}^{n}\{\varepsilon_i + n_2^{-1/2}G(\boldsymbol{x}_{i,\mathcal{S}})\}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}} \leqslant t),\\ V_2(\boldsymbol{\alpha}, t) &= n_2^{-1/2}\sum_{i=n_1+1}^{n}\{m(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a}) - m(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}} \leqslant t).\end{aligned}$$

很直观地, 可将 $V_1(\boldsymbol{\alpha},t)$ 化为

$$V_1(\boldsymbol{\alpha},t)=n_2^{-1/2}\sum_{i=1}^{n}\varepsilon_i I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}}\leqslant t)+\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)\}+O_p(n^{-1/2}).$$

如引理 A.1 中展示的,

$$\begin{aligned}n_2^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a}-\boldsymbol{\beta}_{\mathcal{S}_a})=&(\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}n_2^{-1/2}\sum_{i=n_1+1}^{n}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\varepsilon_i\\&+\left(n_2^{-1}\sum_{i=n_1+1}^{n}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\}^{\mathrm{T}}]\right)^{-1}C_{n_2}n_2^{1/2}\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}\\&+o_p(1).\end{aligned}$$

再根据 Taylor 展开, 在局部备择假设下可推导出

$$\begin{aligned}V_2(\boldsymbol{\alpha},t)=&n_2^{-1/2}\sum_{i=n_1+1}^{n}\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\}^{\mathrm{T}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}}\leqslant t)(\boldsymbol{\beta}_{\mathcal{S}_a}-\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a})+o_p(1)\\=&-\mathrm{E}[\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)](\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}\\&\cdot\left[n_2^{-1/2}\sum_{i=1}^{n}\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}_a})\varepsilon_i+\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}\right]+o_p(1).\end{aligned}$$

记 $\zeta_0(\boldsymbol{\alpha},t)$ 是均值函数为 $\mathrm{E}\{\zeta_0(\boldsymbol{\alpha},t)\}$、协方差函数为 $\mathrm{cov}\{\zeta_0(\boldsymbol{\alpha},t),\zeta_0^{\mathrm{T}}(\boldsymbol{\alpha}_0,t_0)\}$ 的 Gauss 过程, 其中

$$\begin{aligned}\mathrm{E}\{\zeta_0(\boldsymbol{\alpha},t)\}=&\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)\}-\mathrm{E}[\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}\leqslant t)]\\&\cdot(\mathrm{E}[\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\{\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\}^{\mathrm{T}}])^{-1}\mathrm{E}\{G(\boldsymbol{x}_{\mathcal{S}})\mathbf{g}(\boldsymbol{\beta}_{\mathcal{S}_a}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}_a})\},\end{aligned}\tag{B.5}$$

$\mathrm{cov}\{\zeta_0(\boldsymbol{\alpha},t),\zeta_0^{\mathrm{T}}(\boldsymbol{\alpha}_0,t_0)\}$ 的形式参见 (B.1).

类似地, 当 $n_2\to\infty$ 时, 有

$$n_2c(|\mathcal{S}|)T_{n_2}\xrightarrow{d}\int\zeta_0^2(\boldsymbol{\alpha},t)dF_{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{S}}}(t)d\boldsymbol{\alpha}.$$

至此完成了定理 4.3 的证明. □

定理 6.1 的证明 与文献 [18] 类似, 为了简化证明表达, 在证明中省略 "inf" 和 "min" 函数部分. 定义 $\pi(u)\overset{\mathrm{def}}{=}M^{-1}\sum_{i=1}^{M}I(p_i\leqslant u)$, 则 $\pi(\alpha\gamma)=M^{-1}\sum_{i=1}^{M}I(p_i\leqslant\alpha\gamma)$, 且 $\{\pi(\alpha\gamma)\geqslant\gamma\}$ 表示 $\{p_i;\,i=1,\ldots,M\}$ 中至少有 γ 的比例不超过 $\alpha\gamma$, 即 $\{p_i;i=1,\ldots,M\}$ 的 γ 分位数不超过 $\alpha\gamma$, 也即 $\{Q(\gamma)\leqslant\alpha\}$, 其中 $Q(\gamma)=q_\gamma(\{p_i/\gamma;i=1,\ldots,M\})$. 也就是 $\{Q(\gamma)\leqslant\alpha\}$ 与 $\{\pi(\alpha\gamma)\geqslant\gamma\}$ 等价. 于是有

$$\Pr\{Q(\gamma)\leqslant\alpha\}=\Pr\{\pi(\alpha\gamma)\geqslant\gamma\}\leqslant\gamma^{-1}\mathrm{E}\{\pi(\alpha\gamma)\}=(\gamma M)^{-1}\sum_{i=1}^{M}P(p_i\leqslant\alpha\gamma),$$

其中第一个不等式成立是由于 Markov 不等式, 最后一个等式成立是由于 $\pi(\cdot)$ 函数的定义.

由于在原假设下, 当 $\mathcal{A}\subset\mathcal{S}_{ai}$ 时, p_i 服从均匀分布, 因此 $\Pr(p_i\leqslant\alpha\gamma\mid\mathcal{A}\subset\mathcal{S}_{ai})=\alpha\gamma$. 又由于 $\Pr(\mathcal{A}\subset\mathcal{S}_{ai})\to1$, 进而推导出 $\Pr(p_i\leqslant\alpha\gamma)\to\alpha\gamma$. 注意到 M 为固定常数, $(\gamma M)^{-1}\sum_{i=1}^{M}P(p_i\leqslant\alpha\gamma)\to\alpha$. 将其代入上述不等式可得定理的第一个结论成立.

接下来验证定理的第二个结论. 由于在原假设下, 当 $\mathcal{A}\subset\mathcal{S}_{ai}$ 时, $p_i\sim U(0,1)$, 于是有

$$\mathrm{E}\Big\{\sup_{\gamma\in(\gamma_{\min},1)}\gamma^{-1}I(p_i\leqslant\alpha\gamma)\Big\}=\int_0^{\alpha\gamma_{\min}}\Big\{\sup_{\gamma\in(\gamma_{\min},1)}\gamma^{-1}I(x\leqslant\alpha\gamma)\Big\}dx+\int_{\alpha\gamma_{\min}}^{\alpha}\Big\{\sup_{\gamma\in(\gamma_{\min},1)}\gamma^{-1}I(x\leqslant\alpha\gamma)\Big\}dx.$$

注意到当 $0\leqslant x\leqslant\alpha\gamma_{\min}$ 时, $\sup_{\gamma\in(\gamma_{\min},1)}\gamma^{-1}I(x\leqslant\alpha\gamma)=\sup_{\gamma\in(\gamma_{\min},1)}\gamma^{-1}=\gamma_{\min}^{-1}$; 当 $\alpha\gamma_{\min}<x\leqslant\alpha$ 时, $\sup_{\gamma\in(\gamma_{\min},1)}\gamma^{-1}I(x\leqslant\alpha\gamma)=\sup_{\gamma\in(\gamma_{\min},1)}\gamma^{-1}I(\gamma\geqslant x/\alpha)=\alpha/x$, 进而得到

$$\mathrm{E}\Big\{\sup_{\gamma\in(\gamma_{\min},1)}\gamma^{-1}I(p_i\leqslant\alpha\gamma)\Big\}=\int_0^{\alpha\gamma_{\min}}\gamma_{\min}^{-1}dx+\int_{\alpha\gamma_{\min}}^{\alpha}\alpha x^{-1}dx=\alpha(1-\log\gamma_{\min}).$$

由 $\{Q(\gamma)\leqslant\alpha\}$ 与 $\{\pi(\alpha\gamma)\geqslant\gamma\}$ 的等价性和 Markov 不等式可得

$$\Pr\Big\{\inf_{\gamma\in(\gamma_{\min},1)}Q(\gamma)\leqslant\alpha\Big\}=\mathrm{E}\Big[\sup_{\gamma\in(\gamma_{\min},1)}I\{\pi(\alpha\gamma)\geqslant\gamma\}\Big]\leqslant\alpha(1-\log\gamma_{\min}).$$

因此, 令 $\theta=\alpha/(1-\log\gamma_{\min})$, 可得

$$\Pr\Big[\inf_{\gamma\in(\gamma_{\min},1)}Q(\gamma)(1-\log\gamma_{\min})\leqslant\theta\Big]\leqslant\theta.$$

至此完成了定理 6.1 的证明. □

A lack-of-fit test for parametric index model with ultrahigh-dimensional covariates

Tingyou Zhou, Yaowu Zhang & Liping Zhu

Abstract In this paper, we propose a modified two-stage model checking for the parametric index model with ultrahigh-dimensional covariates. Specifically, we randomly split the whole data set into two equal halves $\mathcal{D}_1$ and $\mathcal{D}_2$. In the first stage, we carry out a screening procedure to select some active variables and candidate variables based on $\mathcal{D}_1$. Then in the second stage, we propose a lack-of-fit test based on the selected variables using dataset $\mathcal{D}_2$. Our method can avoid potential type-I error inflation and power loss, which widely exist in ultrahigh-dimensional cases. We put forward a novel test statistic with no tuning parameters in the second stage. It can avoid the curse of dimensionality, and is n-consistent under the null hypothesis and root-n-consistent under the alternative hypothesis. A consistent bootstrap procedure is suggested to decide the critical value. Comprehensive simulations as well as an application to a real dataset are conducted to demonstrate the finite sample performances of our proposal.

Keywords **ultrahigh-dimensional, parametric single index model, lack-of-fit test, type-I error, power**

MSC(2020) **62G10**

doi: 10.1360/SSM-2023-0188