基于 copula 函数的纵向数据复合分位数回归及变量选择

王康宁¹, 李劭珉^{2*}, 林路^{1,3}

1. 山东工商学院统计学院, 烟台 264005;

2. 北京大学光华管理学院, 北京 100871;

3. 山东大学金融研究院, 济南 250100

E-mail: wkn1986@126.com, lsmjim@163.com, linlu@sdu.edu.cn

收稿日期: 2018-12-19; 接受日期: 2019-07-15; 网络出版日期: 2020-03-11; * 通信作者

国家自然科学基金 (批准号: 11901356 和 11231005) 和山东省自然科学基金 (批准号: ZR2017BA002) 资助项目

摘要 复合分位数回归 (composite quantile regression) 具有稳健性好和估计效率高的优势, 所以其经常被用来替代均值回归. 众所周知, 纵向数据具有组内相关的特点, 如果估计过程中能正确地利用组内相关性, 则可以显著地提高估计效率. 因此, 探讨纵向数据复合分位数回归中如何使用相关性是一个有意义的问题. 本文首先利用 copula 函数方法构建纵向数据复合分位数回归的组内协方差矩阵, 进而基于构建的协方差矩阵, 提出一个无偏且有效的基于 copula 函数的复合分位数回归估计方程; 进一步, 为了进行变量选择, 利用基于 copula 函数的估计方程, 提出一个光滑门限 (smooth-threshold) 的复合分位数回归估计方程方法. 本文提出的方法具有很高的灵活性, 而且提高了估计的效率. 理论结果以及数值模拟和实际数据分析都验证了本文的方法.

关键词 复合分位数回归 纵向数据 copula 函数 有效性 变量选择

MSC (2010) 主题分类 62G05, 62G20

1 引言

纵向数据频繁出现在医学以及公共卫生等领域, 其最大特点是具有组内相关性, 而且如何合适地考虑组内相关性对提高估计效率具有重要意义. 文献 [1] 提出的广义估计方程 (generalized estimating equations, GEE) 在纵向数据分析方法发展过程中具有里程碑意义, 其通过工作相关矩阵 (working correlation matrix) 来利用组内相关性. 文献 [1] 指出, 即使工作相关矩阵假定错误, 基于广义估计方程的估计量依然相合. 关于广义估计方程方法的最近研究可参见文献 [2–8] 等.

然而, 广义估计方程与加权最小二乘非常相似, 这就使得它不稳健, 容易受异常值影响. 而纵向数据具有重复观测的特点, 这就使得一个异常值的产生会带来一系列的异常值. 因此, 在纵向数据分析

英文引用格式: Wang K N, Li S M, Lin L. Copula based composite quantile regression for longitudinal data and variable selection (in Chinese). Sci Sin Math, 2020, 50: 1097–1116, doi: 10.1360/N012018-00298

中稳健性极其重要. 近些年来, 为了提高广义估计方程的稳健性, 一个常用的方法是利用稳健的 Huber 得分函数降低响应变量中异常值的影响. 关于这类方法的文献主要有文献 [9–13].

尽管 Huber 得分函数具有稳健性的特点, 但是它在估计效率方面有局限性. 为了提高估计效率, 文献 [14] 提出了复合分位数回归方法. 近年来, 复合分位数回归方法得到了广泛的研究和应用. 文献 [15, 16] 分别将其推广到了非参数和半参数模型中. 另外, 为了进一步提高估计效率, 文献 [17] 提出了加权的复合分位数回归方法, 文献 [18] 基于二次推断函数探讨了复合分位数回归.

然而, 在纵向数据复合分位数回归中, 如何利用组内相关性提高估计效率是一个未知且值得探讨的问题. 近年来, 由于 copula 函数在刻画相关性方面的优势, 其被广泛地应用在纵向数据分析中, 如文献 [19–23]. 然而, 在纵向数据复合分位数回归中, 如何利用 copula 函数刻画组内相关性还是完全未知的. 因此, 本文提出一个基于 copula 函数的纵向数据复合分位数回归方法. 具体来讲, 首先利用 copula 函数得到纵向数据复合分位数回归的组内协方差矩阵, 进而基于 copula 函数组内协方差矩阵, 得到无偏且有效的复合分位数回归估计函数. 另外, 由于得到的估计函数是非光滑的, 我们利用文献 [24] 提出的光滑诱导方法 (induced smoothing method) 将其进行磨光. 作为一个直接的应用, 我们进一步提出了一个光滑门限的复合分位数回归估计方程变量选择方法; 在一些较弱的正则条件下, 建立了本文所提出的方法的大样本性质.

本文后续内容安排如下: 第 2 节介绍基于 copula 函数的复合分位数回归估计方程以及相关的算法和大样本性质. 第 3 节介绍光滑门限的复合分位数回归估计方程变量选择方法. 数值模拟和实际数据分析在第 4 节给出. 所有正则条件和定理证明都放在第 5 节.

2 基于 copula 函数的复合分位数回归估计方程

令 Y_{ij} 和 $\mathbf{X}_{ij} = (X_{ij,1}, \dots, X_{ij,p})^T$ 分别表示第 i 个个体的第 j 次观测的响应变量和 p 维协变量, $i = 1, \dots, n$, $j = 1, \dots, m_i$. 不失一般性, 考虑平衡设计, 即 $m_i = m$ 为有限的. 假定 $\{\mathbf{Y}_i, \mathbf{X}_i, i = 1, \dots, n\}$ 是 $\{\mathbf{Y} = (Y_1, \dots, Y_m)^T, \mathbf{X}\}$ 的一个随机样本, $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im})^T$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{im})^T$, 本文考虑如下的线性模型:

$$Y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m, \quad (2.1)$$

其中, $\boldsymbol{\beta}$ 是 p 维参数, 其真实值为 $\boldsymbol{\beta}_0$, ϵ_{ij} 是随机误差项.

假定 $0 < \tau_1 < \dots < \tau_K < 1$ 是区间 $[0, 1]$ 的一个分割, 特别地, 本文取 $\tau_k = \frac{k}{1+K}$, $k = 1, \dots, K$, 且 K 与 n 无关, $b(\tau_k)$ 是随机误差 ϵ_{ij} 的 $100\tau_k\%$ 条件分位数. 接下来将介绍基于 copula 函数的复合分位数回归估计方程的构造过程. 首先, 对 $\tau_k \in (0, 1)$,

$$E\{\tau_k - I(Y_{ij} - \mathbf{X}_{ij}^T \boldsymbol{\beta}_0 \leq b(\tau_k)) \mid \mathbf{X}_{ij}\} = 0, \quad (2.2)$$

其中, $I(\cdot)$ 是示性函数. 因此可以通过求解如下关于 $\boldsymbol{\beta}$ 的方程得到相合估计:

$$\sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \mathbf{S}_{i(k)}(\boldsymbol{\beta}) = \mathbf{0}, \quad (2.3)$$

其中, $\mathbf{S}_{i(k)}(\boldsymbol{\beta}) = (S_{i1(k)}(\boldsymbol{\beta}), \dots, S_{im(k)}(\boldsymbol{\beta}))^T$, $S_{ij(k)}(\boldsymbol{\beta}) = \tau_k - I(Y_{ij} - \mathbf{X}_{ij}^T \boldsymbol{\beta} \leq b(\tau_k))$. 但是, $b(\tau_k)$ ($k = 1, \dots, K$) 是未知的, 所以, (2.3) 不能直接应用. 因此, 我们将 $b(\tau_k)$ ($k = 1, \dots, K$) 替换为它们的

估计量. 而得到估计量的一个最简单的方法是最小化如下忽略了相关性的复合分位数回归目标函数:

$$(\hat{b}_1, \dots, \hat{b}_K, \hat{\beta}) = \arg \min_{b_1, \dots, b_K, \beta} \left\{ \sum_{k=1}^K \sum_{i=1}^n \sum_{j=1}^m \rho_{\tau_k}(Y_{ij} - b_k - \mathbf{X}_{ij}^T \beta) \right\}, \quad (2.4)$$

其中, $\rho_{\tau}(u) = u(\tau - I(u < 0))$. 经过简单计算, 有

$$\sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \hat{\mathbf{S}}_{i(k)}(\beta) = \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \mathbf{S}_{i(k)}(\beta) + \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \{\hat{\mathbf{S}}_{i(k)}(\beta) - \mathbf{S}_{i(k)}(\beta)\}, \quad (2.5)$$

其中,

$$\begin{aligned} \hat{\mathbf{S}}_{i(k)}(\beta) &= (\hat{S}_{i1(k)}(\beta), \dots, \hat{S}_{im(k)}(\beta))^T, \\ \hat{S}_{ij(k)}(\beta) &= \tau_k - I(Y_{ij} - \mathbf{X}_{ij}^T \beta \leq \hat{b}_k). \end{aligned}$$

根据文献 [14, 定理 1], 可知 $\hat{b}_k - b(\tau_k) = O_p(\frac{1}{\sqrt{n}})$. 因此, 可以简单地得到

$$\begin{aligned} E\{\hat{S}_{ij(k)}(\beta_0) - S_{ij(k)}(\beta_0) \mid \mathbf{X}_{ij}\} &= P(Y_{ij} - \mathbf{X}_{ij}^T \beta_0 \leq b(\tau_k) \mid \mathbf{X}_{ij}) - P(Y_{ij} - \mathbf{X}_{ij}^T \beta_0 \leq \hat{b}_k \mid \mathbf{X}_{ij}) \\ &= o_p(1). \end{aligned}$$

所以, 求解方程 (2.3) 与求解如下的方程是等价的:

$$\sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \hat{\mathbf{S}}_{i(k)}(\beta) = \mathbf{0}. \quad (2.6)$$

很明显, 方程 (2.6) 并没有考虑纵向数据的组内相关性, 尽管这并不影响估计相合, 但是估计效率往往不令人满意. 所以, 为了考虑组内相关性, 受广义估计方程以及文献 [25] 的思想启发, 提出如下的复合分位数回归估计函数:

$$\mathbf{U}(\beta) = \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \mathbf{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1} \hat{\mathbf{S}}_{i(k)}(\beta), \quad (2.7)$$

其中, $\mathbf{V}_{i(k)} = \text{Cov}\{\mathbf{S}_{i(k)}^0 \mid \mathbf{X}_i\}$, $\mathbf{S}_{i(k)}^0 = (S_{i1(k)}^0, \dots, S_{im(k)}^0)^T = (\tau_k - I(Y_{i1} - \mathbf{X}_{i1}^T \beta_0 - b(\tau_k) \leq 0), \dots, \tau_k - I(Y_{im} - \mathbf{X}_{im}^T \beta_0 - b(\tau_k) \leq 0))^T$, $\mathbf{\Lambda}_{i(k)} = \text{diag}(s_{i1(k)}, \dots, s_{im(k)})$, $s_{ij(k)} = f_j(\mathbf{X}_{ij}^T \beta_0 + b(\tau_k) \mid \mathbf{X}_{ij}, T_{ij})$ 度量了 $Y_{ij} - \mathbf{X}_{ij}^T \beta_0 - b(\tau_k)$ 的离散性, $f_j(\cdot \mid \mathbf{X}_{ij})$ 是 Y_j 的条件密度函数.

接下来, 将通过 copula 函数对条件协方差矩阵 $\mathbf{V}_{i(k)}$ 进行刻画. Copula 函数实际上是一类将联合分布函数与边际分布函数连接在一起的函数, 因此也被称为连接函数. 假定 $F(y_1, \dots, y_m \mid \mathbf{x})$ 是 $(Y_1, \dots, Y_m)^T$ 在给定 $\mathbf{X} = \mathbf{x}$ 的条件分布函数, 有连续的边际条件分布 $F_1(\cdot \mid \mathbf{x}), \dots, F_m(\cdot \mid \mathbf{x})$. 注意到 $F_i(Y_i \mid \mathbf{x})$ 是 $[0, 1]$ 上的均匀分布, Sklar 在 1959 年指出存在 $[0, 1]^m$ 上的唯一的 copula 函数 C 满足

$$\begin{aligned} F(y_1, \dots, y_m \mid \mathbf{x}) &= P(Y_1 \leq y_1, \dots, Y_m \leq y_m \mid \mathbf{x}) \\ &= P\{F_1(Y_1 \mid \mathbf{x}) \leq F_1(y_1 \mid \mathbf{x}), \dots, F_m(Y_m \mid \mathbf{x}) \leq F_m(y_m \mid \mathbf{x}) \mid \mathbf{x}\} \\ &= C(F_1(y_1 \mid \mathbf{x}), \dots, F_m(y_m \mid \mathbf{x}) \mid \mathbf{x}). \end{aligned} \quad (2.8)$$

为了模型简洁, 考虑参数的 copula 函数, 即如下简化的 copula 模型:

$$F(y_1, \dots, y_m \mid \mathbf{x}) = C(F_1(y_1 \mid \mathbf{x}), \dots, F_m(y_m \mid \mathbf{x}); \theta_0). \quad (2.9)$$

很明显, 上面的模型可以另写为

$$C(u_1, \dots, u_m; \theta_0) = F(F_1^{-1}(u_1 | \mathbf{x}), \dots, F_m^{-1}(u_m | \mathbf{x}) | \mathbf{x}), \quad (u_1, \dots, u_m) \in [0, 1]^m.$$

Copula 模型 (2.9) 被广泛地应用在对多元分布进行建模的相关文献中 (如文献 [23, 26, 27]). 文献 [23, 26] 也指出, 即使在参数假定不满足时, 简化的模型 (2.9) 依然可以作为很好的近似.

注意到, 对模型 (2.1) 以及任意的 $\tau \in (0, 1)$, 有

$$F_j^{-1}(\tau | \mathbf{X}_{ij}) = \mathbf{X}_{ij}^T \beta_0 + F_{\epsilon_j}^{-1}(\tau | \mathbf{X}_{ij}),$$

其中, $F_{\epsilon_j}(\cdot | \mathbf{X}_{ij})$ 是 ϵ_{ij} 的条件分布函数. 令 $u_{ij} = F_j(Y_{ij} | \mathbf{X}_{ij})$, 其服从 $[0, 1]$ 上的均匀分布, 然后将上式中的 τ 替换为 u_{ij} 可得到

$$Y_{ij} = \mathbf{X}_{ij}^T \beta_0 + F_{\epsilon_j}^{-1}(u_{ij} | \mathbf{X}_{ij}).$$

对于给定的多元 copula 函数 $C(\cdot; \theta_0)$, 令 $\gamma_{ijl(k)}$ 表示 $S_{ij(k)}^0$ 与 $S_{il(k)}^0$ 的条件协方差, 通过一些计算, 有

$$\begin{aligned} \gamma_{ijl(k)} &= P(Y_{ij} - \mathbf{X}_{ij}^T \beta_0 - b(\tau_k) \leq 0, Y_{il} - \mathbf{X}_{il}^T \beta_0 - b(\tau_k) \leq 0 | \mathbf{X}_{ij}, \mathbf{X}_{il}) - \tau_k^2 \\ &= P(F_{\epsilon_j}^{-1}(u_{ij} | \mathbf{X}_{ij}) \leq b(\tau_k), F_{\epsilon_l}^{-1}(u_{il} | \mathbf{X}_{il}) \leq b(\tau_k) | \mathbf{X}_{ij}, \mathbf{X}_{il}) - \tau_k^2 \\ &= P(u_{ij} \leq \tau_k, u_{il} \leq \tau_k | \mathbf{X}_{ij}, \mathbf{X}_{il}) - \tau_k^2 \\ &= P(F_j(Y_{ij} | \mathbf{X}_{ij}) \leq \tau_k, F_l(Y_{il} | \mathbf{X}_{il}) \leq \tau_k | \mathbf{X}_{ij}, \mathbf{X}_{il}) - \tau_k^2 \\ &= C_{jl}(\tau_k, \tau_k; \theta_0) - \tau_k^2, \end{aligned} \quad (2.10)$$

其中, $C_{jl}(\tau_k, \tau_k; \theta_0)$ 表示 $C(\cdot; \theta_0)$ 的第 j 和 l 个分量取为 τ_k , 其他分量取为 1. 那么, $\mathbf{S}_{i(k)}^0$ 的基于 copula 函数的条件协方差矩阵可表示为 $\mathbf{V}_{i(k)}(\theta_0) = (C_{jl}(\tau_k, \tau_k; \theta_0) - \tau_k^2)_{j,l=1}^m$.

注 2.1 常用的多元 copula 函数包括椭圆 copulas 和 Archimedean copulas. 椭圆 copulas 可以包含很多特殊的相关结构, 如可交换相关结构、自回归相关结构和 Toeplitz 相关结构等. 因此, 其被广泛地用来刻画纵向数据的组内相关性. 而 Archimedean copulas 只能刻画纵向数据个体不同观测间的可交换的相关结构, 因此其不适应于个体观测次数多的情形. 为了解决这一问题, 我们也可以考虑 vine copulas, 其由一系列二元 copulas 构成. 更多关于多元 copulas 或者 vine copulas 的介绍可参见文献 [27, 28].

进一步, 我们需要估计 θ_0 , 其可以通过最大化如下的 copula 对数似然函数得到

$$l(\theta) = \sum_{i=1}^n \log[c(u_{i1}, \dots, u_{im}; \theta)], \quad (2.11)$$

其中, $c(\cdot; \theta)$ 表示与 copula 函数 $C(\cdot; \theta)$ 相应的密度函数, $u_{ij} = \{v \in (0, 1) : F_{\epsilon_j}^{-1}(v | \mathbf{X}_{ij}) + \mathbf{X}_{ij}^T \beta_0 = Y_{ij}\} = P(Y_j \leq Y_{ij} | \mathbf{X}_{ij}) = F_j(Y_{ij} | \mathbf{X}_{ij})$. 然而, u_{ij} ($i = 1, \dots, n, j = 1, \dots, m$) 是未知的, 因此首先需要得到 u_{ij} 的估计量 $\tilde{u}_{ij}$, 然后通过最大化 copula 对数似然函数得到估计量 $\hat{\theta}$. 最后, 将得到的基于 copula 函数的条件协方差矩阵 $\mathbf{V}_{i(k)}(\hat{\theta}) = (C_{jl}(\tau_k, \tau_k; \hat{\theta}) - \tau_k^2)_{j,l=1}^m$ 应用到估计方程 (2.7) 中, 就得到了未知参数 β 的稳健有效复合分位数回归估计量. 估计过程的具体细节如下.

第一步, 假定 $0 < \varsigma_1 < \dots < \varsigma_{\kappa_n} < 1$ 是一系列分位点, 其中, 取 $\varsigma_k = \frac{k}{\kappa_n + 1}$. 对于 $k = 1, \dots, \kappa_n$, 令

$$(\tilde{b}(\varsigma_k), \tilde{\beta}(\varsigma_k)) = \arg \min_{b(\varsigma_k), \beta(\varsigma_k)} \sum_{i=1}^n \sum_{j=1}^m \rho_{\varsigma_k}(Y_{ij} - b(\varsigma_k) - \mathbf{X}_{ij}^T \beta(\varsigma_k)).$$

如果 $Y_{ij} \in [\tilde{b}(\varsigma_k) + \mathbf{X}_{ij}^T \tilde{\boldsymbol{\beta}}(\varsigma_k), \tilde{b}(\varsigma_{k+1}) + \mathbf{X}_{ij}^T \tilde{\boldsymbol{\beta}}(\varsigma_{k+1})]$, 定义 $\tilde{u}_{ij} = (1 - \tilde{\alpha}_{ij})\varsigma_k + \tilde{\alpha}_{ij}\varsigma_{k+1}$,

$$\tilde{\alpha}_{ij} = \frac{Y_{ij} - \tilde{b}(\varsigma_k) - \mathbf{X}_{ij}^T \tilde{\boldsymbol{\beta}}(\varsigma_k)}{(\tilde{b}(\varsigma_{k+1}) - \tilde{b}(\varsigma_k)) + (\mathbf{X}_{ij}^T \tilde{\boldsymbol{\beta}}(\varsigma_{k+1}) - \mathbf{X}_{ij}^T \tilde{\boldsymbol{\beta}}(\varsigma_k))},$$

其中, 参见文献 [23] 的建议, κ_n 选择为 $[4 + 3n^{0.4}]$, 符号 $[a]$ 表示 a 的整数部分.

第二步, 通过最大化如下的伪 copula 对数似然函数得到 $\boldsymbol{\theta}_0$ 的估计量 $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^n \log[c(\tilde{u}_{i1}, \dots, \tilde{u}_{im}; \boldsymbol{\theta})] \right\}.$$

第三步, 在方程 (2.7) 中, 将 $\mathbf{V}_{i(k)}$ 和 $\boldsymbol{\Lambda}_{i(k)}$ 分别替换为它们的估计量 $\mathbf{V}_{i(k)}(\hat{\boldsymbol{\theta}})$ 和 $\hat{\boldsymbol{\Lambda}}_{i(k)}$, 通过求解如下估计方程:

$$\hat{\mathbf{U}}(\boldsymbol{\beta}) = \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \hat{\boldsymbol{\Lambda}}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\hat{\boldsymbol{\theta}}) \hat{\mathbf{S}}_{i(k)}(\boldsymbol{\beta}) = \mathbf{0},$$

可得到估计量 $\hat{\boldsymbol{\beta}}^o$. 在具体应用中, $\hat{\boldsymbol{\Lambda}}_{i(k)} = \text{diag}(\hat{s}_{i1(k)}, \dots, \hat{s}_{im(k)})$ 通过利用文献 [29] 的商 (quotient) 估计方法得到, 即

$$\hat{s}_{ij(k)} = \frac{2h}{(\tilde{b}(\tau_k + h) - \tilde{b}(\tau_k - h)) + (\mathbf{X}_{ij}^T \tilde{\boldsymbol{\beta}}(\tau_k + h) - \mathbf{X}_{ij}^T \tilde{\boldsymbol{\beta}}(\tau_k - h))},$$

其中, h 为正的窗宽且随着 $n \rightarrow \infty$, 有 $h \rightarrow 0$, 参见文献 [30] 的建议, 我们选择

$$h = 1.57n^{-1/3} \left\{ \frac{1.5\phi^2[\Phi^{-1}(\tau_k)]}{2[\Phi^{-1}(\tau_k)]^2 + 1} \right\}^{1/3},$$

$\Phi(\cdot)$ 与 $\phi(\cdot)$ 分别表示标准正态分布的分布函数和密度函数.

上面第三步中求解方程 $\hat{\mathbf{U}}(\boldsymbol{\beta}) = \mathbf{0}$ 的主要困难在于目标函数 $\hat{\mathbf{U}}(\boldsymbol{\beta})$ 是非光滑的. 我们将利用文献 [24] 提出的光滑诱导方法来解决这一问题. 具体来讲, 假定 $\boldsymbol{\Gamma} = N(\mathbf{0}, \mathbf{I}_p)$ 且 $\boldsymbol{\Omega}$ 为 $p \times p$ 正定矩阵, 则 $\tilde{\mathbf{U}}(\boldsymbol{\beta})$ 的光滑诱导估计函数可以自然地定义为

$$\begin{aligned} \tilde{\mathbf{U}}(\boldsymbol{\beta}) &= \mathbf{E}_{\boldsymbol{\Gamma}}[\mathbf{U}(\boldsymbol{\beta} + \boldsymbol{\Omega}^{1/2}\boldsymbol{\Gamma})] \\ &= \mathbf{E}_{\boldsymbol{\Gamma}} \left[\sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \hat{\boldsymbol{\Lambda}}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\hat{\boldsymbol{\theta}}) \hat{\mathbf{S}}_{i(k)}(\boldsymbol{\beta} + \boldsymbol{\Omega}^{1/2}\boldsymbol{\Gamma}) \right] \\ &= \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \hat{\boldsymbol{\Lambda}}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\hat{\boldsymbol{\theta}}) \tilde{\mathbf{S}}_{i(k)}(\boldsymbol{\beta}), \end{aligned} \quad (2.12)$$

$\tilde{\mathbf{S}}_{i(k)}(\boldsymbol{\beta}) = (\tilde{S}_{i1(k)}(\boldsymbol{\beta}), \dots, \tilde{S}_{im(k)}(\boldsymbol{\beta}))^T$, $\tilde{S}_{ij(k)}(\boldsymbol{\beta}) = \tau_k + \Phi\left(\frac{Y_{ij} - \mathbf{X}_{ij}^T \boldsymbol{\beta} - \tilde{b}_k}{r_{ij}}\right) - 1$, $r_{ij} = [\mathbf{X}_{ij}^T \boldsymbol{\Omega} \mathbf{X}_{ij}]^{1/2}$. 通过求解方程 $\tilde{\mathbf{U}}(\boldsymbol{\zeta}) = \mathbf{0}$, 可以得到估计量 $\hat{\boldsymbol{\beta}}^s$. 因为 $\tilde{\mathbf{U}}(\boldsymbol{\beta})$ 关于 $\boldsymbol{\beta}$ 是光滑的, 为了得到方程 $\tilde{\mathbf{U}}(\boldsymbol{\zeta}) = \mathbf{0}$ 的解, 我们将采用 Newton-Raphson 迭代算法, 具体如下:

步骤 2.1 给定初始值 $\boldsymbol{\beta}^{(0)}$ 和 $\boldsymbol{\Omega}^{(0)} = \frac{1}{n}\mathbf{I}_p$, 其中, $\boldsymbol{\beta}^{(0)}$ 可以通过 (2.4) 得到.

步骤 2.2 对于给定的 $\boldsymbol{\beta}^{(t)}$ 和 $\boldsymbol{\Omega}^{(t)}$, 更新 $\boldsymbol{\beta}^{(t+1)}$ 和 $\boldsymbol{\Omega}^{(t+1)}$,

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} - \tilde{\mathbf{D}}^{-1}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Omega}^{(t)}) \tilde{\mathbf{U}}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Omega}^{(t)}), \\ \boldsymbol{\Omega}^{(t+1)} &= [\tilde{\mathbf{D}}^{-1}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Omega}^{(t)})] \text{cov}[\tilde{\mathbf{U}}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Omega}^{(t)})][\tilde{\mathbf{D}}^{-1}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Omega}^{(t)})]^T, \end{aligned}$$

其中,

$$\tilde{D}(\beta^{(t)}, \Omega^{(t)}) = - \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \hat{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\hat{\theta}) \tilde{\Lambda}_{i(k)} \mathbf{X}_i,$$

$\tilde{\Lambda}_{i(k)}$ 是第 j 个对角元素为 $\phi(\frac{Y_{ij} - \mathbf{X}_{ij}^T \beta^{(t)} - \hat{b}_k}{\tau_{ij}^{(t)}})$ 的对角矩阵, $r_{ij}^{(t)} = [\mathbf{X}_{ij}^T \Omega^{(t)} \mathbf{X}_{ij}]^{1/2}$,

$$\text{cov}[\tilde{U}(\beta^{(t)}, \Omega^{(t)})] = \sum_{k,k'=1}^K \sum_{i=1}^n \mathbf{X}_i^T \hat{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\hat{\theta}) \tilde{S}_{i(k)}(\beta^{(t)}) \tilde{S}_{i(k')}^T(\beta^{(t)}) \mathbf{V}_{i(k')}^{-1}(\hat{\theta}) \hat{\Lambda}_{i(k')} \mathbf{X}_i |_{\beta^{(t)}, \Omega^{(t)}}.$$

步骤 2.3 重复迭代步骤 2.2 直到收敛. 在实际操作中, 当 $\|\beta^{(t+1)} - \beta^{(t)}\| < 10^{-6}$ 时, 停止迭代.

接下来将探索估计量 $\hat{\beta}^o$ 和 $\hat{\beta}^s$ 的渐近性质. 首先, 定义如下符号:

$$\begin{aligned} \Sigma &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \Lambda_{i(k)} \mathbf{V}_{i(k)}^{-1}(\theta_0) \Lambda_{i(k)} \mathbf{X}_i, \\ \Xi &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \sum_{k'=1}^K \mathbf{X}_i^T \Lambda_{i(k)} \mathbf{V}_{i(k)}^{-1}(\theta_0) \mathbf{W}^{kk'} \mathbf{V}_{i(k')}^{-1}(\theta_0) \Lambda_{i(k')} \mathbf{X}_i, \end{aligned}$$

其中,

$$\begin{aligned} \mathbf{W}^{kk'} &= (W_{jj'}^{kk'} \sqrt{\tau_k(1-\tau_k)} \sqrt{\tau_{k'}(1-\tau_{k'})})_{j,j'=1}^m, \\ W_{jj'}^{kk'} &= \text{corr}(S_{ij(k)}^0, S_{ij'(k')}^0). \end{aligned}$$

定理 2.1 假定第 5 节中的正则条件 (C1)–(C7) 成立, 而且 $\kappa_n^{2a+1}/n^{1-2q} \rightarrow 0$, $\kappa_n \rightarrow \infty$ (随着 $n \rightarrow \infty$), 那么, 有如下结果成立:

- (1) $\sqrt{n}(\hat{\beta}^o - \beta_0) \rightarrow_d N(\mathbf{0}, \Sigma^{-1} \Xi \Sigma^{-1})$;
- (2) $\sqrt{n}(\hat{\beta}^s - \beta_0) \rightarrow_d N(\mathbf{0}, \Sigma^{-1} \Xi \Sigma^{-1})$.

定理 2.1 建立了估计量的渐近正态分布性, 另外, 通过定理 2.1, 我们发现光滑后的估计函数 $\tilde{U}(\zeta)$ 与原估计函数 $\hat{U}(\zeta)$ 是等价的.

3 光滑门限的复合分位数回归估计方程变量选择

众所周知, 当数据维数较高时, 变量选择非常重要. 近年来, 惩罚损失函数型变量选择方法得到了广泛关注. 文献 [31] 提出了 LASSO (least absolute shrinkage and selection operator) 惩罚, 文献 [32] 比较了桥回归 (bridge regression) 和 LASSO 惩罚, 文献 [33] 提出了 SCAD (smoothly clipped absolute deviation) 惩罚, 文献 [34] 进一步给出了自适应的 LASSO. 然而这类方法往往要涉及凸优化, 具有一定的计算复杂度. 为了解决这一问题, 文献 [35] 提出了一个新的变量选择方法, 称为光滑门限 (smooth-threshold) 估计方程, 其能够自动地选择变量并且避免凸优化问题.

基于第 2 节提出的估计函数 $\tilde{U}(\zeta)$ 和文献 [35] 中的方法, 我们提出如下的光滑门限的复合分位数回归估计方程:

$$(\mathbf{I}_p - \Lambda) \tilde{U}(\zeta) + \Lambda \beta = \mathbf{0}, \quad (3.1)$$

其中, $\mathbf{I}_p$ 是 p 维单位矩阵, $\Lambda = \text{diag}(\delta_1, \dots, \delta_p)$ 是对角矩阵. 注意到, 如果 $X_{ij,k}$ 是不相关变量, 则在方程 (3.1) 中令 $\delta_k = 1$, 将使最终的解 $\beta_k = 0$. 因此, 通过构造合适的矩阵 Λ , 方程 (3.1) 的解可以用作

变量选择. 另外, 由于估计函数 $\tilde{U}(\zeta)$ 具有稳健性以及估计效的优势, 所以, 最终的估计量和选择结果也具有这方面的优势.

然而, 方程 (3.1) 并不能直接地应用, 因为其涉及未知参数 δ_j ($j = 1, \dots, p$), 这就需要基于数据的准则来选取. 相似于文献 [35], 在得到基于 copula 函数的复合分位数回归估计量 $\hat{\beta}^s$ 后, δ_j ($j = 1, \dots, p$) 被选择为

$$\hat{\delta}_j = \min \left\{ 1, \frac{\lambda}{|\hat{\beta}_j^s|^{1+\tau}} \right\}, \quad j = 1, \dots, p.$$

在某种意义上, 这种选择方法与文献 [34] 中的自适应性 LASSO 有相同的思想. 注意到, 这种选择将会涉及压缩参数 (λ, τ) , 下面将提出基于数据的准则来选择这些参数. 在方程 (3.1) 中将 $\mathbf{\Lambda}$ 替换为对角元素为 $\hat{\delta}_j$ ($j = 1, \dots, p$) 的 $\hat{\mathbf{\Lambda}}$, 方程 (3.1) 变为

$$(\mathbf{I}_p - \hat{\mathbf{\Lambda}})\tilde{U}(\zeta) + \hat{\mathbf{\Lambda}}\beta = \mathbf{0}. \quad (3.2)$$

令 $\tilde{\beta}_{(\lambda, \tau)} = (\tilde{\beta}_{1(\lambda, \tau)}, \dots, \tilde{\beta}_{p(\lambda, \tau)})^T$ 是基于压缩参数 (λ, τ) 得到的方程 (3.2) 的解, 则估计的相关变量指标集合为 $\hat{\mathcal{A}} = \{k : \tilde{\beta}_{k(\lambda, \tau)} \neq 0, k = 1, \dots, p\}$. 不失一般性, 假定 $\beta_{0k} \neq 0$ ($k = 1, \dots, p_1$) 和 $\beta_{0k} = 0$ ($k = p_1 + 1, \dots, p$), 令 $\beta^* = (\beta_1, \dots, \beta_{p_1})^T$, 且其真值为 β_0^* . 下面的定理建立了光滑门限的复合分位数回归估计方程估计量 $\tilde{\beta}_{(\lambda, \tau)}$ 的相合性.

定理 3.1 假定定理 2.1 中的正则条件成立, 另外假设 λ 和 τ 满足 $\sqrt{n}\lambda \rightarrow 0$ 且 $n^{(1+\tau)/2}\lambda \rightarrow \infty$ (随着 $n \rightarrow \infty$), 那么,

$$\|\tilde{\beta}_{(\lambda, \tau)} - \beta_0\| = O_p(n^{-1/2}).$$

定理 3.1 表明 $\tilde{\beta}_{(\lambda, \tau)}$ 可以 $\sqrt{n}$ 相合地估计未知参数. 另外, 下面的定理 3.2 也表明定理 3.1 中的相合估计量具有稀疏性且非零系数的估计量的渐近分布和基于真实模型得到的估计量的渐近分布相同.

定理 3.2 假定定理 3.1 中的条件成立, 那么,

- (1) $\lim_{n \rightarrow \infty} P\{\tilde{\beta}_{k(\lambda, \tau)} = 0, k = p_1 + 1, \dots, p\} = 1$;
- (2) $\sqrt{n}(\tilde{\beta}^* - \beta_0^*) \rightarrow_d N(\mathbf{0}, \Sigma_*^{-1} \Xi_* \Sigma_*^{-1})$,

其中, Σ_* 和 Ξ_* 分别是对应于相关变量的 Σ 和 Ξ 的子矩阵.

定理 3.2 表明, 通过选择合适的压缩参数 λ 和 τ , 并通过估计方程 (3.2) 得到的估计量可以相合地选择模型的相关变量, 即

$$\lim_{n \rightarrow \infty} P\{\hat{\mathcal{A}} = \{1, \dots, p_1\}\} = 1.$$

另外, $\tilde{\beta}^*$ 的渐近分布与在模型相关变量已知的情形下得到的估计量一样.

方程 (3.2) 的数值解也可以通过 Newton-Raphson 迭代方法得到. 我们选择 β 的初始值为 $\hat{\beta}^s$, 对于给定的 λ 和 τ , 相比较于方程 $\tilde{U}(\zeta) = \mathbf{0}$ 的求解过程, 求方程 (3.2) 的数值解的过程的唯一不同是将步骤 2.2 替换为下式:

$$\begin{aligned} \beta^{(t+1)} &= \beta^{(t)} - [\tilde{D}(\beta^{(t)}, \Omega^{(t)}) + \hat{G}]^{-1}[\tilde{U}(\beta^{(t)}, \Omega^{(t)}) + \hat{G}\beta^{(t)}], \\ \Omega^{(t+1)} &= [\tilde{D}^{-1}(\beta^{(t)}, \Omega^{(t)}) + \hat{G}]\text{cov}[\tilde{U}(\beta^{(t)}, \Omega^{(t)})][\tilde{D}^{-1}(\beta^{(t)}, \Omega^{(t)}) + \hat{G}]^T, \end{aligned}$$

其中, $\hat{G} = (\mathbf{I}_p - \hat{\mathbf{\Lambda}})^{-1}\hat{\mathbf{\Lambda}}$. 另外, 为了选择最优的 (λ, τ) , 我们最小化如下的信息准则:

$$\text{SIC}(\lambda, \tau) = \log \left\{ \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n \mathbf{s}_{i(k)}^T(\tilde{\beta}_{\lambda, \tau}) \mathbf{V}_{i(k)}^{-1}(\hat{\theta}) \mathbf{s}_{i(k)}(\tilde{\beta}_{\lambda, \tau}) \right\} + \frac{\log n}{n} df_{\lambda, \tau}, \quad (3.3)$$

其中, $df_{\lambda, \tau}$ 表示估计量 $\tilde{\beta}_{(\lambda, \tau)}$ 中非零分量的个数. 模拟结果显示, 通过最小化 $SIC(\lambda, \tau)$ 选择的参数在有限样本时具有良好的表现, 而如何从理论上证明所选参数的大样本性质是一个有趣的问题, 也具有一定的挑战, 我们将在以后的研究中讨论.

4 数值模拟和实际应用

4.1 数值模拟

例 4.1 考虑模型

$$Y_{ij} = X_{ij,1}\beta_1 + X_{ij,2}\beta_2 + X_{ij,3}\beta_3 + \epsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, 5, \quad (4.1)$$

其中, $\beta = (1.5, 2.5, 3)^T$, $X_{ij,1}, \dots, X_{ij,3}$ 是零均值协方差结构为 $\text{Cov}(X_{ij,k_1}, X_{ij,k_2}) = 4 \exp(-|k_1 - k_2|)$ 的 Gauss 随机过程. 针对随机误差 $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{i5})^T$, 我们考虑三种情形, 情形 1 (正态误差): $\epsilon_i \sim N(\mathbf{0}, \Sigma)$; 情形 2 (t 误差): ϵ_i 服从多元 t_3 分布, 且协方差结构为 Σ ; 情形 3 (异常点数据): 随机误差相同于情形 1, 但是随机选择两组数据并将响应变量 Y_{ij} 变为 $Y_{ij} + 3$. 情形 1 中的协方差矩阵 Σ 具有 $\text{ar}(1)$ 结构, 情形 2 为可交换的结构, 且方差都为 1, 相关系数为 $\delta = 0.5$.

为了更清楚地比较, 我们考虑三种类型的估计量: 文献 [1] 的广义估计方程估计量 (表示为 GEE), 假定工作独立的复合分位数回归估计量 (表示为 WICQR), 文献 [10] 的稳健广义估计方程估计量 (表示为 RGEE). 对于我们的基于 copula 函数的复合分位数回归方法, 考虑两种 copula 函数, 分别为 Gaussian copula 和 t-copula, 相应的结果分别用 GCQR 和 TCQR 表示. 另外, 在 GEE、RGEE、GCQR 和 TCQR 中, 分别考虑了可交换和一阶自回归 $\text{ar}(1)$ 相关结构.

对于 $n = 100$ 和 $K = 9$, 表 1-3 报告了基于 500 次重复计算的均方误差、偏差和标准差. 可以看出, 对于每一种情形, 相比较于忽略相关性的复合分位数回归方法 (WICQR), 新方法在估计效率方面具有明显的优势. 其次, 模拟结果也显示新方法对 copula 函数以及相关结构的选择是稳健的, 即使在 copula 函数以及相关结构指定错误的时候, 新方法依然表现良好. 而且由于充分地利用了纵向数据的相关性以及继承了复合分位数回归的优势, 新方法的表现明显比 RGEE 方法好. 即使在随机误差服从正态时, 新方法也与 GEE 方法表现相当, 而在情形 2 和 3 时, 新方法明显比 GEE 方法要好. 最后, 针对情形 1 (情形 2 和 3 的结果类似, 为了省略篇幅, 在此不再报告), 在样本容量 $n = 100$ 时, 我们将 $K = 9$ 增大到 $K = 12$, 表 4 报告了模拟结果, 我们发现 $K = 9$ 与 $K = 12$ 的模拟结果差别不大, 另外我们也在 $K = 9$ 时, 将样本容量由 $n = 100$ 增大到 $n = 200$, 通过表 5 的模拟结果, 我们发现增加样本容量可以进一步提高估计准确性.

例 4.2 本例探索光滑门限的复合分位数回归估计方程变量选择方法的有限样本性质. 基于采用的 Gauss copula 和 t-copula, 我们将本文的变量选择方法分别表示为 SGCQR 和 STCQR. 另外, 我们也将本文的方法与文献 [9] 提出的惩罚的稳健广义估计方程方法 (表示为 PRGEE) 和文献 [6] 提出的光滑门限的广义估计方程方法 (表示为 SGEE) 进行了比较. 为了更清楚地比较, 利用 “C” 表示被正确地估计为零的回归系数的平均个数, “IC” 表示被错误地估计为零的回归系数的平均个数, “CF” 表示相关变量被选择正确的频率.

数据集通过如下的模型产生:

$$Y_{ij} = \sum_{k=1}^{10} X_{ij,k}\beta_k + \epsilon_{ij}, \quad i = 1, \dots, n = 100, \quad j = 1, \dots, 5, \quad (4.2)$$

表 1 例 4.1 中情形 1 的模拟结果: 均方误差 $\times 10^2$ (MSE)、偏差 $\times 10^2$ (Bias) 和标准差 (SD)

方法	相关性	β_1			β_2			β_3		
		MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD
GCQR	ex	3.795	-0.092	0.198	3.044	0.099	0.183	3.006	-0.095	0.182
	ar1	3.886	-0.089	0.183	3.176	-0.092	0.192	2.569	0.088	0.179
TCQR	ex	3.901	0.102	0.187	3.236	-0.112	0.173	3.121	-0.089	0.188
	ar1	3.879	-0.097	0.186	3.209	0.106	0.169	3.008	-0.091	0.179
WICQR	—	8.989	-0.112	0.266	7.197	0.121	0.273	7.327	0.129	0.287
RGEE	ex	8.235	-0.128	0.271	7.519	-0.129	0.283	6.535	0.131	0.273
	ar1	8.342	0.133	0.268	7.237	-0.138	0.272	6.718	0.126	0.266
GEE	ex	4.387	0.111	0.214	3.813	-0.105	0.201	3.096	-0.097	0.219
	ar1	4.518	0.103	0.216	3.932	-0.106	0.215	3.223	-0.099	0.205

表 2 例 4.1 中情形 2 的模拟结果: 均方误差 $\times 10^2$ (MSE)、偏差 $\times 10^2$ (Bias) 和标准差 (SD)

方法	相关性	β_1			β_2			β_3		
		MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD
GCQR	ex	4.076	-0.113	0.199	3.978	0.105	0.187	3.711	0.103	0.194
	ar1	3.992	-0.115	0.205	4.056	0.109	0.191	3.818	-0.108	0.198
TCQR	ex	4.027	0.109	0.190	4.019	-0.108	0.187	4.012	-0.098	0.191
	ar1	3.915	0.112	0.201	3.975	-0.114	0.189	3.878	-0.110	0.193
WICQR	—	10.121	-0.209	0.331	9.985	0.225	0.313	8.765	0.217	0.377
RGEE	ex	9.876	-0.221	0.313	8.839	0.219	0.279	8.075	-0.233	0.301
	ar1	9.767	-0.215	0.325	8.919	0.225	0.306	7.815	0.228	0.317
GEE	ex	24.331	-1.357	0.589	26.929	2.768	0.512	28.989	1.967	0.672
	ar1	24.467	-1.603	0.612	27.216	-2.879	0.509	30.817	-2.135	0.658

表 3 例 4.1 中情形 3 的模拟结果: 均方误差 $\times 10^2$ (MSE)、偏差 $\times 10^2$ (Bias) 和标准差 (SD)

方法	相关性	β_1			β_2			β_3		
		MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD
GCQR	ex	4.102	-0.121	0.201	4.176	-0.109	0.198	3.976	0.109	0.203
	ar1	4.133	-0.118	0.205	4.201	-0.113	0.197	4.032	0.116	0.201
TCQR	ex	4.187	0.114	0.197	4.155	0.109	0.199	4.098	-0.106	0.198
	ar1	4.225	-0.117	0.204	4.222	0.115	0.193	4.118	-0.117	0.202
WICQR	—	11.076	-0.311	0.379	10.088	0.327	0.401	9.232	-0.364	0.392
RGEE	ex	10.113	-0.327	0.376	9.632	-0.316	0.389	9.326	0.355	0.403
	ar1	9.989	-0.356	0.379	9.732	0.338	0.382	9.755	-0.372	0.411
GEE	ex	45.313	-3.275	0.798	63.254	-3.876	0.812	50.236	-4.198	0.786
	ar1	50.545	-3.718	0.812	65.312	3.675	0.823	51.332	4.232	0.779

其中, $\beta = (1.5, 2.5, 3, 0, 0, \dots, 0)^T$, $X_{ij,1}$ 、 $X_{ij,2}$ 和 $X_{ij,3}$ 是相关变量, 其产生方式与例 4.1 相同, $X_{ij,4}, \dots, X_{ij,10}$ 是不相关变量且服从独立同分布的标准正态分布. 为了考察稳健性和有效性, 例 4.1 中的情

表 4 当 $n = 100$ 且 $K = 12$ 时, 例 4.1 中情形 1 的模拟结果: 均方误差 $\times 10^2$ (MSE)、偏差 $\times 10^2$ (Bias) 和标准差 (SD)

方法	相关性	β_1			β_2			β_3		
		MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD
GCQR	ex	3.817	-0.087	0.183	3.021	0.091	0.189	2.808	0.088	0.183
	ar1	3.877	0.083	0.189	3.087	0.088	0.187	2.681	0.079	0.175
TCQR	ex	3.889	-0.092	0.191	3.178	-0.103	0.171	2.878	-0.091	0.181
	ar1	3.871	-0.095	0.185	3.185	-0.098	0.166	2.837	0.087	0.172

表 5 当 $n = 200$ 且 $K = 9$ 时, 例 4.1 中情形 1 的模拟结果: 均方误差 $\times 10^2$ (MSE)、偏差 $\times 10^2$ (Bias) 和标准差 (SD)

方法	相关性	β_1			β_2			β_3		
		MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD
GCQR	ex	2.901	0.081	0.169	2.818	0.091	0.168	2.673	-0.082	0.161
	ar1	2.897	0.077	0.172	2.767	0.085	0.163	2.476	-0.076	0.157
TCQR	ex	3.013	-0.093	0.171	2.838	-0.089	0.167	2.787	0.081	0.170
	ar1	2.976	-0.088	0.177	2.903	-0.097	0.166	2.687	-0.077	0.163

形 1-3 依然在本例中被考虑.

表 6-8 报告了在 $K = 9$ 时基于 500 次重复的 3 种情形的模拟结果. 通过模拟, 我们有如下几点发现. 当数据服从正态分布时 (情形 1), 新方法 with 光滑门限广义估计方程方法表现相当, 而在重尾分布和异常值数据的情形 (情形 2 和 3), 新方法在变量选择和估计效率方面明显好于光滑门限的广义估计方程方法. 另外, 由于利用了纵向数据的相关性以及具备了复合分位数回归的优势, 新方法在各种情形下的表现都明显优于惩罚的稳健估计方程方法. 最后, 新提出的变量选择方法对 copula 函数以及相关结构的选择也是稳健的.

4.2 实际应用

本节将本文提出的方法应用到纵向 AIDS 数据分析中. 这个数据集记录了 283 名感染 HIV 病毒的男性同性恋患者在 1984 年到 1991 年间的随访记录. 令 x_1 表示被调查者的吸烟状况 (1 代表吸烟,

表 6 例 4.2 中情形 1 的变量选择结果: C、IC、CF 和均方误差 $\times 10^2$ (MSE)

方法	相关性	C	IC	CF	MSE		
					β_1	β_2	β_3
SGCQR	ex	6.986	0	98.20	4.152	5.023	3.464
	ar1	6.978	0	97.60	4.232	5.176	3.512
STCQR	ex	6.980	0	97.80	4.202	5.108	3.507
	ar1	6.972	0	97.40	4.256	4.989	3.612
PRGEE	—	6.946	0	95.20	9.378	8.902	9.713
SGEE	ex	6.974	0	97.60	5.636	6.592	4.673
	ar1	6.978	0	97.00	6.032	6.873	4.858

表 7 例 4.2 中情形 2 的变量选择结果: C、IC、CF 和均方误差 $\times 10^2$ (MSE)

方法	相关性	C	IC	CF	MSE		
					β_1	β_2	β_3
SGCQR	ex	6.972	0	96.20	6.127	5.067	6.023
	ar1	6.964	0	96.00	6.078	4.985	6.115
STCQR	ex	6.968	0	96.20	5.989	5.132	5.976
	ar1	6.970	0	97.00	6.109	5.067	6.067
PRGEE	—	6.826	0	87.80	11.207	11.547	10.878
SGEE	ex	6.167	0	61.40	24.122	27.897	29.613
	ar1	6.176	0	62.20	24.231	28.012	29.776

表 8 例 4.2 中情形 3 的变量选择结果: C、IC、CF 以及均方误差 $\times 10^2$ (MSE)

方法	相关性	C	IC	CF	MSE		
					β_1	β_2	β_3
SGCQR	ex	6.962	0	94.80	6.936	7.131	7.165
	ar1	6.958	0	94.20	7.066	6.912	6.979
STCQR	ex	6.954	0	94.60	7.133	7.208	7.211
	ar1	6.946	0	94.40	7.208	7.132	7.099
PRGEE	—	6.794	0	87.20	13.135	12.689	12.708
SGEE	ex	5.932	0	59.20	46.237	62.115	52.236
	ar1	5.916	0	58.40	47.176	61.276	53.178

2 代表不吸烟), x_2 表示被调查者的年龄, 且已经被标准化, x_3 表示标准化的 PreCD4 水平. 另外, 我们对变量的交叉效应也很感兴趣, 因此, 3 个变量的交叉效应也被考虑在如下的纵向数据线性模型中:

$$y_{ij} = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \epsilon_{ij}, \quad (4.3)$$

其中, y_{ij} 表示第 i 个被调查者在时刻 t_{ij} 的 CD4 水平.

我们应用新提出的基于 copula 函数的复合分位数回归方法估计未知参数, 而且将它与广义估计方程方法进行比较, 其中, 我们采用多元 Gaussian copula 以及可交换的相关结构, 而且选择 $K = 9$. 另外, 为了考察稳健性, 相同于文献 [10] 的方法, 我们随机地选取 5% 的数据将响应变量 y_{ij} 变为 $y_{ij} + 8$. 表 9 和 10 报告了估计的结果, 尽管我们不知道参数的真实值, 但是通过比较有异常值和无异常值两种情形下的估计结果, 我们发现本文新提出的基于 copula 的复合分位数回归方法受异常值的影响很低, 而广义估计方程方法受异常值的影响非常严重. 另外, 为了展示新方法的拟合和预测效果, 我们依次将数据集中一个个体的数据作为验证集, 剩下的数据作为训练集, 得到了交叉验证预测误差和拟合误差. 表 9 和 10 也报告了预测误差和拟合误差的绝对值的中位数, 分别表示为 MPE 和 MFE, 可以进一步看出新方法优于广义估计方程方法且基本不受异常值的影响.

另外, 由于模型 (4.3) 中变量个数较少, 因此不太适合于变量选择. 为了展示本文方法在变量选择方面的表现, 采用文献 [4] 的方法, 在模型 (4.3) 中, 我们增加了 5 个服从均匀分布 $U(0, 1)$ 的不相关变量 $r_1, \dots, r_5$. 我们进一步应用本文提出的光滑门限的复合分位数回归估计方程变量选择方法而且将之与光滑门限的广义估计方程方法比较. 另外, 为了考察稳健性, 我们依然随机选择 5% 的数据, 将响应变量 y_{ij} 变为 $y_{ij} + 8$, 进而重新分析数据. 表 11 报告了这 11 个变量基于 200 次重复的被选中频率,

表 9 无异常值情形下 $(\beta_1, \dots, \beta_8)$ 的估计结果以及交叉验证预测误差的绝对值的中位数 (MPE) 和拟合误差的绝对值的中位数 (MFE)

	本文的方法			GEE		
	估计值	标准差	95% 置信区间	估计值	标准差	95% 置信区间
β_1	0.3759	0.1336	(0.114, 0.637)	1.3254	0.2976	(0.742, 1.909)
β_2	0.8741	0.1879	(0.506, 1.242)	-0.7687	0.3012	(-1.359, -0.178)
β_3	0.2176	0.1820	(-0.139, 0.574)	0.3769	0.2577	(-0.128, 0.882)
β_4	-0.1387	0.1758	(-0.483, 0.206)	0.9857	0.2712	(0.454, 1.517)
β_5	0.5752	0.1034	(0.373, 0.778)	-0.1318	0.2157	(-0.554, 0.291)
β_6	-0.9821	0.1894	(-1.352, -0.609)	-0.3769	0.3217	(-1.367, -0.106)
(MPE, MFE)	(5.371, 4.895)			(7.804, 6.178)		

表 10 5% 异常值情形下 $(\beta_1, \dots, \beta_8)$ 的估计结果以及交叉验证预测误差的绝对值的中位数 (MPE) 和拟合误差的绝对值的中位数 (MFE)

	本文的方法			GEE		
	估计值	标准差	95% 置信区间	估计值	标准差	95% 置信区间
β_1	0.4107	0.1391	(0.138, 0.683)	5.3272	0.6712	(4.011, 6.643)
β_2	1.0137	0.2032	(0.615, 1.412)	1.3103	0.7138	(-0.089, 2.709)
β_3	0.1832	0.1912	(-0.191, 0.558)	-0.9709	0.6819	(-2.307, 0.366)
β_4	-0.2039	0.1951	(-0.586, 0.178)	4.1083	0.7856	(2.568, 5.648)
β_5	0.7127	0.1153	(0.486, 0.939)	-2.0117	0.8139	(-3.606, -0.416)
β_6	-0.8325	0.1994	(-1.223, -0.442)	2.2709	0.9102	(-3.795, -0.228)
(MPE, MFE)	(5.989, 5.017)			(12.378, 10.673)		

表 11 变量选择结果

	无异常值		5% 异常值	
	本文的方法	SGEE	本文的方法	SGEE
x_1	1.000	1.000	1.000	0.880
x_2	1.000	1.000	1.000	0.765
x_3	1.000	0.980	1.000	0.840
x_1x_2	0.990	1.000	0.980	0.630
x_1x_3	1.000	1.000	1.000	0.490
x_2x_3	1.000	1.000	1.000	0.685
r_1	0.030	0.070	0.035	0.325
r_2	0.020	0.120	0.040	0.305
r_3	0.025	0.130	0.030	0.380
r_4	0.030	0.090	0.060	0.345
r_5	0.010	0.110	0.025	0.325

可以看出新提出的变量选择方法能够以很高的频率选中真实变量, 而且不相关变量被选中的频率很低. 另一方面, 相比较于光滑门限广义估计方程方法, 本文的变量选择方法对异常值非常稳健.

从上面的数据分析结果可以看出, 本文新提出的新方法适应于纵向数据的统计推断过程, 即使数据存在异常值, 新方法依然有很好的表现.

5 正则条件和主要证明

为了证明渐近性质, 我们需要假定如下的正则条件:

(C1) 方程 $E_{\mathbf{u}}[\frac{\partial \log[c(u_1, \dots, u_m; \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}] = \mathbf{0}$ 存在唯一的解 $\boldsymbol{\theta}_0 \in \Theta$, 其中, $\mathbf{u} = (u_1, \dots, u_m)$, Θ 是紧的凸集;

(C2) 积分 $\int_{\mathbf{u}} \sup_{\boldsymbol{\theta} \in \Theta} \|\frac{\partial \log[c(u_1, \dots, u_m; \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}\| d\mathbf{u}$ 存在;

(C3) 偏导数 $\sup_{1/(\kappa_n+1) \leq \mathbf{u} \leq \kappa_n/(\kappa_n+1)} \|\frac{\partial \log[c(u_1, \dots, u_m; \boldsymbol{\theta}_0)]}{\partial \mathbf{u}}\| \leq C\kappa_n^a$, 其中, C 和 a 是正常数;

(C4) 分布函数 $P(Y_{ij} - b(\tau_k) - \mathbf{X}_{ij}^T \boldsymbol{\beta} \leq y \mid \mathbf{X}_{ij})$ 是绝对连续的, 其密度函数 $f_{ij(k)}(\cdot)$ 是连续且一致有界的, 一阶导数 $f'_{ij(k)}(\cdot)$ 是一致有界的且 $0 < f_{ij(k)}(b(\tau_k)) < \infty$;

(C5) 协变量 $\mathbf{X}_{ij}$ 满足 $\sum_{i=1}^n \sum_{j=1}^m \|\mathbf{X}_{ij}\|^2 = O_p(n)$, 且对某个正的常数 $q < \frac{1}{6}$, 有

$$\max_{1 \leq i \leq n, 1 \leq j \leq m} \|\mathbf{X}_{ij}\|^2 = O_p(n^q);$$

(C6) 对任意的正定矩阵 $\boldsymbol{\Upsilon}_{i(k)}$, $\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_{i(k)}^T \boldsymbol{\Lambda}_{i(k)} \boldsymbol{\Upsilon}_{i(k)} \boldsymbol{\Lambda}_{i(k)} \mathbf{X}_{i(k)}$ 收敛到正定矩阵;

(C7) (2.12) 中的 $p \times p$ 矩阵 $\boldsymbol{\Omega}$ 是正定矩阵, 且 $\|\boldsymbol{\Omega}\| = O(\frac{1}{n})$, $\|\boldsymbol{\Omega}\|$ 表示 $\boldsymbol{\Omega}$ 的最大奇异值.

引理 5.1 假定正则条件 (C1)–(C5) 成立, 如果随着 $n \rightarrow \infty$, $\kappa_n^{2a+1}/n^{1-2q} \rightarrow 0$ 且 $\kappa_n \rightarrow \infty$, 那么, $\hat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_0$, 其中, $\rightarrow_p$ 表示依概率收敛.

引理 5.1 的证明 证明方法与文献 [23, 引理 1] 的方法相似, 为了节省空间, 在此省略. $\square$

定理 2.1 的证明 令 $\bar{U}(\boldsymbol{\beta}) = \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \mathbf{P}_{i(k)}(\boldsymbol{\beta})$, 其中,

$$\mathbf{P}_{i(k)}(\boldsymbol{\beta}) = (\tau_k - P(Y_{i1} - \mathbf{X}_{i1}\boldsymbol{\beta} < \hat{b}_k), \dots, \tau_k - P(Y_{im} - \mathbf{X}_{im}\boldsymbol{\beta} < \hat{b}_k))^T.$$

由引理 5.1 和文献 [31] 可知, $\mathbf{V}_{i(k)}(\hat{\boldsymbol{\theta}}) \rightarrow_p \mathbf{V}_{i(k)}(\boldsymbol{\theta}_0)$ 且 $\hat{\boldsymbol{\Lambda}}_{i(k)} \rightarrow_p \boldsymbol{\Lambda}_{i(k)}$. 因此可以直接考虑方程

$$\hat{U}(\boldsymbol{\beta}) = \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \hat{\mathbf{S}}_{i(k)}(\boldsymbol{\beta}) = \mathbf{0}.$$

通过直接计算, 可得到

$$\begin{aligned} \frac{1}{n} [\hat{U}(\boldsymbol{\beta}) - \bar{U}(\boldsymbol{\beta})] &= \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) [\hat{\mathbf{S}}_{i(k)}(\boldsymbol{\beta}) - \mathbf{P}_{i(k)}(\boldsymbol{\beta})] \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \begin{bmatrix} P(Y_{i1} - \mathbf{X}_{i1}\boldsymbol{\beta} < \hat{b}_k) - I(Y_{i1} - \mathbf{X}_{i1}\boldsymbol{\beta} < \hat{b}_k) \\ \vdots \\ P(Y_{im} - \mathbf{X}_{im}\boldsymbol{\beta} < \hat{b}_k) - I(Y_{im} - \mathbf{X}_{im}\boldsymbol{\beta} < \hat{b}_k) \end{bmatrix} \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^K \mathbf{a}_{ij(k)} [P(Y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta} < \hat{b}_k) - I(Y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta} < \hat{b}_k)], \end{aligned}$$

其中, $\mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) = (\mathbf{a}_{i1(k)}, \dots, \mathbf{a}_{im(k)})$. 由条件 (C5) 和文献 [36] 可得到

$$\sup_{\boldsymbol{\beta}} \left| \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^K \mathbf{a}_{ij(k)} [P(Y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta} < \hat{b}_k) - I(Y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta} < \hat{b}_k)] \right| = o\left(\frac{1}{\sqrt{n}}\right).$$

因此,

$$\sup_{\beta} \left\| \frac{1}{n} [\hat{U}(\beta) - \bar{U}(\beta)] \right\| = o\left(\frac{1}{\sqrt{n}}\right).$$

因为 $P(Y_{ij} - \mathbf{X}_{ij}\beta_0 < \hat{b}_k | \mathbf{X}_{ij}) \rightarrow_p \tau_k$, 而且 $\hat{\beta}^o$ 是 $\hat{U}(\beta) = \mathbf{0}$ 的解, 因此, $\hat{\beta}^o \rightarrow_p \beta_0$ 随着 $n \rightarrow \infty$. 对任意满足 $\|\beta - \beta_0\| = O(n^{-1/3})$ 的 β , 有

$$\begin{aligned} \hat{U}(\beta) - \hat{U}(\beta_0) &= \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) [\hat{\mathbf{S}}_{i(k)}(\beta) - \hat{\mathbf{S}}_{i(k)}(\beta_0)] \\ &= \left\{ \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \mathbf{P}_{i(k)}(\beta) \right. \\ &\quad \left. + \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) [\hat{\mathbf{S}}_{i(k)}(\beta) - \hat{\mathbf{S}}_{i(k)}(\beta_0) - \mathbf{P}_{i(k)}(\beta)] \right\} \\ &= \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \mathbf{P}_{i(k)}(\beta) \\ &\quad + \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^K \mathbf{a}_{ij(k)} [P(Y_{ij} - \mathbf{X}_{ij}\boldsymbol{\zeta} < \hat{b}_k) - I(Y_{ij} - \mathbf{X}_{ij}\beta < \hat{b}_k) \\ &\quad + I(Y_{ij} - \mathbf{X}_{ij}\beta_0 < \hat{b}_k) - \tau_k]. \end{aligned}$$

根据文献 [25, 引理 3], 可以得到

$$\sup \left| \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^K \mathbf{a}_{ij(k)} [P(Y_{ij} - \mathbf{X}_{ij}\beta < \hat{b}_k) - I(Y_{ij} - \mathbf{X}_{ij}\beta < \hat{b}_k) + I(Y_{ij} - \mathbf{X}_{ij}\beta_0 < \hat{b}_k) - \tau_k] \right| = o_p(\sqrt{n}).$$

因此, $\hat{U}(\beta) - \hat{U}(\beta_0) = \bar{U}(\beta) + o_p(\sqrt{n})$. 因为 $\bar{U}(\beta_0) = \mathbf{0}$, 利用 Taylor 近似可以得到

$$\hat{U}(\beta) - \hat{U}(\beta_0) = \frac{\partial \bar{U}(\beta)}{\partial \beta} \bigg|_{\beta=\beta_0} (\beta - \beta_0) + o_p(\sqrt{n}).$$

因为 $\hat{U}(\hat{\beta}^o) = \mathbf{0}$ 且 $\hat{\beta}^o$ 在 β_0 的 $n^{-1/3}$ 邻域内, 所以有

$$\sqrt{n}(\hat{\beta}^o - \beta_0) = \left(\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \boldsymbol{\Lambda}_{i(k)} \mathbf{X}_i \right)^{-1} \frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \hat{\mathbf{S}}_{i(k)}(\beta_0) + o_p(1).$$

令 $\frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \hat{\mathbf{S}}_{i(k)}(\beta_0) = \sum_{k=1}^K \mathbf{A}_k$, 其中,

$$\mathbf{A}_k = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \hat{\mathbf{S}}_{i(k)}(\beta_0).$$

注意到, $\hat{\mathbf{S}}_{ij(k)}(\beta_0) = I(\epsilon_{ij} \leq b(\tau_k)) - I(\epsilon_{ij} \leq \hat{b}_k) + \tau_k - I(\epsilon_{ij} \leq b(\tau_k))$, $\hat{b}_k - b(\tau_k) = O_p(\frac{1}{\sqrt{n}})$, 所以, $E(\hat{\mathbf{S}}_{ij(k)}(\beta_0)) = 0$. 因此, $E(\mathbf{A}_k) = \mathbf{0}$,

$$\text{Var} \left(\sum_{k=1}^K \mathbf{A}_k \right) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \sum_{k'=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \mathbf{W}^{kk'} \mathbf{V}_{i(k')}^{-1}(\boldsymbol{\theta}_0) \boldsymbol{\Lambda}_{i(k')} \mathbf{X}_i.$$

所以, 利用多元中心极限定理, 可得

$$\frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \hat{\mathbf{S}}_{i(k)}(\boldsymbol{\beta}_0) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Xi}).$$

另外, 基于大数定律可知,

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \boldsymbol{\Lambda}_{i(k)} \mathbf{X}_i \rightarrow_p \boldsymbol{\Sigma}.$$

所以, 利用 Slutsky 定理, 有 $\sqrt{n}(\hat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}_0) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma}^{-1} \boldsymbol{\Xi} \boldsymbol{\Sigma}^{-1})$. 这就完成了定理 2.1(1) 的证明.

首先证明 $\frac{1}{\sqrt{n}}[\tilde{\mathbf{U}}(\boldsymbol{\beta}) - \hat{\mathbf{U}}(\boldsymbol{\beta})] = o_p(1)$. 注意到 $\tilde{S}_{ij(k)}(\boldsymbol{\beta}) - \hat{S}_{ij(k)}(\boldsymbol{\beta}) = I(-\Delta_{ij(k)})\Phi(-|\Delta_{ij(k)}|)$, 其中,

$$\Delta_{ij(k)} = \frac{Y_{ij} - \mathbf{X}_{ij}^T \boldsymbol{\beta} - \hat{b}_k}{r_{ij}}.$$

经计算, 可得到

$$\begin{aligned} \frac{1}{\sqrt{n}}[\tilde{\mathbf{U}}(\boldsymbol{\beta}) - \hat{\mathbf{U}}(\boldsymbol{\beta})] &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{k=1}^K \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \begin{bmatrix} I(-\Delta_{i1(k)})\Phi(-|\Delta_{i1(k)}|) \\ \vdots \\ I(-\Delta_{im(k)})\Phi(-|\Delta_{im(k)}|) \end{bmatrix} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^K \mathbf{a}_{ij(k)} I(-\Delta_{ij(k)})\Phi(-|\Delta_{ij(k)}|). \end{aligned}$$

因为

$$\begin{aligned} \mathbb{E}(\tilde{S}_{ij(k)}(\boldsymbol{\beta}) - S_{ij(k)}(\boldsymbol{\beta})) &= \mathbb{E}[I(-\Delta_{ij(k)})\Phi(-|\Delta_{ij(k)}|)] \\ &= \int_{-\infty}^{\infty} \Phi\left(-\frac{|\epsilon - \hat{b}_k|}{r_{ij}}\right) \{2I(\epsilon - \hat{b}_k \leq 0) - 1\} f_{ij(k)}(\epsilon) d\epsilon \\ &= r_{ij} \int_{-\infty}^{\infty} \Phi(-|t|) \{2I(t \leq 0) - 1\} [f_{ij(k)}(\hat{b}_k) + f'_{ij(k)}(\eta(t)) r_{ij} t] dt, \end{aligned} \quad (5.1)$$

其中, $\eta(t)$ 位于 $\hat{b}_k$ 与 $\hat{b}_k + r_{ij}t$ 之间. 因为 $\int_{-\infty}^{\infty} \Phi(-|t|) \{2I(t \leq 0) - 1\} dt = 0$, 所以有

$$\int_{-\infty}^{\infty} \Phi(-|t|) \{2I(t \leq 0) - 1\} f_{ij(k)}(\hat{b}_k) dt = 0.$$

因为 $\int_{-\infty}^{\infty} \Phi(-|t|) |t| dt = 1/2$, 而且由条件 (C4), 存在常数 M , 满足 $\sup_{ij} |f'_{ij(k)}(\eta(t))| \leq M$. 因此,

$$|\mathbb{E}(\tilde{S}_{ij(k)}(\boldsymbol{\beta}) - S_{ij(k)}(\boldsymbol{\beta}))| \leq r_{ij}^2 \int_{-\infty}^{\infty} \Phi(-|t|) |t| |f'_{ij(k)}(\eta(t))| dt \leq \frac{M r_{ij}^2}{2}.$$

所以由正则条件 (C7), 随着 $n \rightarrow \infty$,

$$\left\| \frac{1}{\sqrt{n}} \mathbb{E}[\tilde{\mathbf{U}}(\boldsymbol{\beta}) - \hat{\mathbf{U}}(\boldsymbol{\beta})] \right\| = \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^K \sup_{i,j,k} \|\mathbf{a}_{ij(k)}\| \frac{M r_{ij}^2}{2} = o(1).$$

另外, 将 $\tilde{U}(\beta) - \hat{U}(\beta)$ 写成如下形式:

$$\tilde{U}(\beta) - \hat{U}(\beta) = \sum_{k=1}^K \left\{ \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \begin{bmatrix} I(-\Delta_{i1(k)})\Phi(-|\Delta_{i1(k)}|) \\ \vdots \\ I(-\Delta_{im(k)})\Phi(-|\Delta_{im(k)}|) \end{bmatrix} \right\} =: \sum_{k=1}^K M_k.$$

由 Cauchy-Schwartz 不等式, 可得

$$\begin{aligned} \frac{1}{n} \text{var}[M_k] &= \frac{1}{n} \sum_{i=1}^n \text{var} \left[\sum_{j=1}^m \mathbf{a}_{ij(k)} I(-\Delta_{ij(k)}) \Phi(-|\Delta_{ij(k)}|) \right] \\ &\leq \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \mathbf{a}_{ij(k)}^T \mathbf{a}_{ij(k)} \text{var}[\tilde{S}_{ij(k)}(\beta) - S_{ij(k)}(\beta)] \\ &\quad + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \sum_{j^*=1, j^* \neq j}^m \mathbf{a}_{ij(k)}^T \mathbf{a}_{ij^*(k)} \{ \text{var}[\tilde{S}_{ij(k)}(\beta) - S_{ij(k)}(\beta)] \text{var}[\tilde{S}_{ij^*(k)}(\beta) - S_{ij^*(k)}(\beta)] \}^{1/2}. \end{aligned}$$

对于 $j = 1, \dots, m$, 有

$$\begin{aligned} \text{var}[\tilde{S}_{ij(k)}(\beta) - S_{ij(k)}(\beta)] &\leq \mathbb{E}[\tilde{S}_{ij(k)}(\beta) - S_{ij(k)}(\beta)]^2 \\ &= r_{ij} \int_{-\infty}^{\infty} \Phi^2(-|t|) f_{ij(k)}(r_{ij}t + \hat{b}_k) dt \\ &= r_{ij} \int_{|t|>c} \Phi^2(-|t|) f_{ij(k)}(r_{ij}t + \hat{b}_k) dt \\ &\quad + r_{ij} \int_{|t|\leq c} \Phi^2(-|t|) f_{ij(k)}(r_{ij}t + \hat{b}_k) dt \\ &\leq \Phi^2(-c) + r_{ij} c f_{ij(k)}(\eta), \end{aligned}$$

其中, c 是正值, η 位于 $\hat{b}_k - r_{ij}c$ 与 $\hat{b}_k + r_{ij}c$ 之间. 令 $c = n^{1/3}$, 由条件 (C7) 可得 $r_{ij} = O(n^{-1/2})$, 那么, $r_{ij}c = O(n^{-1/6})$. 注意到, 随着 $n \rightarrow \infty$, $\Phi^2(-c) \rightarrow 0$ 且 $r_{ij}c f_{ij(k)}(\eta) \rightarrow 0$. 所以可得到 $\frac{1}{n} \text{var}[M_k] = o(1)$. 进一步, 利用 Cauchy-Schwarz 不等式可得到

$$\begin{aligned} \frac{1}{n} \text{var}\{\tilde{U}(\beta) - \hat{U}(\beta)\} &= \frac{1}{n} \text{var} \left\{ \sum_{k=1}^K M_k \right\} \\ &\leq \frac{1}{n} \sum_{k=1}^K \text{var}\{M_k\} + \frac{1}{n} \sum_{k=1}^K \sum_{k'=1, k' \neq k}^K \sqrt{\text{var}\{M_k\}} \sqrt{\text{var}\{M_{k'}\}} \\ &= o(1). \end{aligned}$$

所以, 对任意的 β , $\frac{1}{\sqrt{n}}[\tilde{U}(\beta) - \hat{U}(\beta)] \rightarrow 0$, 随着 $n \rightarrow \infty$. 另外, $\sup_{\beta} \|\frac{1}{n}(\tilde{U}(\beta) - \hat{U}(\beta))\| = o(\frac{1}{\sqrt{n}})$, 根据三角不等式可得到

$$\sup_{\beta} \left\| \frac{1}{n}(\tilde{U}(\beta) - \hat{U}(\beta)) \right\| = o\left(\frac{1}{\sqrt{n}}\right).$$

所以, $\hat{\beta}^s \rightarrow_p \beta_0$ 随着 $n \rightarrow \infty$. 为了证明渐近正态性, 需要证明 $\frac{1}{n}\{\tilde{D}(\beta_0) - D(\beta_0)\} \rightarrow_p 0$, 其中

$$\tilde{D}(\beta_0) = \frac{\partial \tilde{U}(\beta)}{\partial \beta} \Big|_{\beta=\beta_0} = - \sum_{k=1}^K \sum_{i=1}^n \mathbf{D}_i^T \boldsymbol{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\boldsymbol{\theta}_0) \boldsymbol{\Phi}_{i(k)} \mathbf{D}_i,$$

$$D(\beta_0) = \frac{\partial \bar{U}(\beta)}{\partial \beta} \Big|_{\beta=\beta_0} = - \sum_{k=1}^K \sum_{i=1}^n D_i^T \Lambda_{i(k)} V_{i(k)}^{-1}(\theta_0) \Lambda_{i(k)} D_i.$$

注意到,

$$\begin{aligned} E[\tilde{D}(\beta_0)] - D(\zeta_0) &= - \sum_{k=1}^K \sum_{i=1}^n D_i^T \Lambda_{i(k)} V_{i(k)}^{-1}(\theta_0) \{E[\Phi_{i(k)}] - \Lambda_{i(k)}\} D_i \\ &= - \sum_{k=1}^K \sum_{i=1}^n \sum_{j=1}^m a_{ij(k)} \left\{ \frac{1}{r_{ij}} E\phi\left(\frac{\epsilon_{ij} - \hat{b}_k}{r_{ij}}\right) - f_{ij(k)}(b(\tau_k)) \right\} D_{ij}, \end{aligned} \quad (5.2)$$

而且,

$$\begin{aligned} \left| \frac{1}{r_{ij}} E\phi\left(\frac{\epsilon_{ij} - \hat{b}_k}{r_{ij}}\right) - f_{ij(k)}(b(\tau_k)) \right| &= \left| f_{ij(k)}(\hat{b}_k) - f_{ij(k)}(b(\tau_k)) + r_{ij} \int_{-\infty}^{+\infty} \phi(t) t f_{ij(k)}(\zeta_t) dt \right| \\ &\leq |f_{ij(k)}(\hat{b}_k) - f_{ij(k)}(b(\tau_k))| + r_{ij} \int_{-\infty}^{+\infty} |\phi(t) t f_{ij(k)}(\zeta_t)| dt. \end{aligned}$$

所以根据条件 (C4) 和 (C7) 可知, 存在常数 M , 使得

$$|f_{ij(k)}(\hat{b}_k) - f_{ij(k)}(b(\tau_k))| + r_{ij} \int_{-\infty}^{+\infty} |\phi(t) t f_{ij(k)}(\zeta_t)| dt \leq \frac{2}{\sqrt{\pi}} r_{ij} M + |f_{ij(k)}(\hat{b}_k) - f_{ij(k)}(b(\tau_k))| \rightarrow 0.$$

因此可以得到 $\frac{1}{n} \{E[\tilde{D}(\beta_0)] - D(\beta_0)\} \rightarrow_p 0$, 进一步, 根据三角不等式和大数定律, 可得到

$$\left| \frac{1}{n} \{\tilde{D}(\beta_0) - D(\zeta_0)\} \right| \leq \left| \frac{1}{n} \{\tilde{D}(\beta_0) - E[\tilde{D}(\beta_0)]\} \right| + \left| \frac{1}{n} \{E[\tilde{D}(\beta_0)] - D(\zeta_0)\} \right| = o(1),$$

这就意味着 $\frac{1}{n} \{\tilde{D}(\beta_0) - D(\beta_0)\} \rightarrow_p 0$. 另外, 关于 $\tilde{U}(\beta)$ 在 β_0 处利用 Taylor 展开得

$$\tilde{U}(\beta) = \tilde{U}(\beta_0) + \tilde{D}(\beta^*)(\beta - \beta_0),$$

其中, β^* 位于 β_0 与 β 之间. 令 $\beta = \hat{\beta}^s$, 则有 $\sqrt{n}(\hat{\beta}^s - \beta_0) = -n\tilde{D}^{-1}(\beta^*)\frac{1}{\sqrt{n}}\tilde{U}(\beta_0)$. 因为 $\hat{\beta}^s \rightarrow_p \beta_0$, 所以, $\beta^* \rightarrow_p \beta_0$ 且 $\tilde{D}^{-1}(\beta^*) \rightarrow_p \tilde{D}^{-1}(\beta_0)$. 由于 $\frac{1}{n} \{\tilde{D}(\beta_0) - D(\beta_0)\} \rightarrow_p 0$, 所以,

$$\sqrt{n}(\hat{\beta}^s - \beta_0) = -nD^{-1}(\beta_0)\frac{1}{\sqrt{n}}\hat{U}(\beta_0) + o_p(1).$$

因此, 利用与定理 2.1(1) 相类似的证明方法可以证明 (2). $\square$

定理 3.1 的证明 令 $\delta_n = n^{-1/2}$, $\beta = \beta_0 + \delta_n T$, $S_n(\beta) = (I_p - \hat{\Lambda})\tilde{U}(\beta) + \hat{\Lambda}\beta$. 我们的目标是证明对任意的 $\varepsilon > 0$, 存在常数 $C > 0$, 使得对充分大的 n , 下式成立:

$$P\left(\sup_{\|T\|=C} \delta_n T^T S_n(\beta_0 + \delta_n T) > 0\right) \geq 1 - \varepsilon.$$

这意味着方程 $S_n(\beta) = 0$ 至少以概率 $1 - \varepsilon$ 存在满足 $\|\tilde{\beta}_{(\lambda, \tau)} - \beta_0\| = O_p(\delta_n)$ 的解. 我们将探究 $\delta_n T^T S_n(\beta_0 + \delta_n T)$ 在 $\{\beta_0 + \delta_n T : \|T\| = C\}$ 上的符号. 利用 Taylor 近似, 可以得到

$$\delta_n T^T S_n(\beta_0 + \delta_n T) = \delta_n T^T S_n(\beta_0) + \delta_n^2 T^T \frac{\partial S_n(\beta)}{\partial \beta} \Big|_{\beta=\beta^*} T = I_{n1} + I_{n2}, \quad (5.3)$$

其中, β^* 位于 β_0 与 $\beta_0 + \delta_n \mathbf{T}$ 之间. 接下来, 分别考虑 I_{n1} 和 I_{n2} . 对于 I_{n1} , 通过一些计算, 有

$$I_{n1} = \delta_n \mathbf{T}^T (\mathbf{I} - \hat{\mathbf{\Lambda}}) \tilde{\mathbf{U}}(\beta_0) + \delta_n \mathbf{T}^T \hat{\mathbf{\Lambda}} \beta_0 = I_{n11} + I_{n12}.$$

利用 Cauchy-Schwarz 不等式, 可以得到

$$|I_{n11}| \leq \delta_n \|\mathbf{T}^T (\mathbf{I} - \hat{\mathbf{\Lambda}})\| \|\tilde{\mathbf{U}}(\beta_0)\| \leq \delta_n (1 - \min\{\hat{\delta}_k : \hat{\delta}_k \neq 1\}) \|\mathbf{T}\| \|\tilde{\mathbf{U}}(\beta_0)\|.$$

因为 $\min\{\hat{\delta}_k : \hat{\delta}_k \neq 1\} \leq \min\{\hat{\delta}_k : k = 1, \dots, q_1\}$, 所以我们只需要得到 $\min\{\hat{\delta}_k : k = 1, \dots, q_1\}$ 的收敛速度. 由定理 2.1 可知, 初始估计 $\hat{\beta}^s$ 满足 $\|\hat{\beta}^s - \beta_0\| = O_p(n^{-1/2})$. 根据条件 $\sqrt{n}\lambda \rightarrow 0$, 对任意的 $\varepsilon > 0$ 和 $k \in \{1, \dots, q_1\}$, 可以得到

$$P(\hat{\delta}_k > n^{-\frac{1}{2}} \varepsilon) = P\left(\frac{\lambda}{|\hat{\beta}_k^s|^{1+\tau}} > n^{-\frac{1}{2}} \varepsilon\right) \leq P\left(\left(\frac{\lambda n^{\frac{1}{2}}}{\varepsilon}\right)^{1/(1+\tau)} > \min_{k \in \{1, \dots, p_1\}} \|\beta_{0k}\| - O_p(n^{-1/2})\right) \rightarrow 0,$$

这意味着对任意的 $k \in \{1, \dots, p_1\}$, $\hat{\delta}_k = o_p(n^{-1/2})$, 而且 $\min_{k=1, \dots, p_1} \hat{\delta}_k = o_p(n^{-1/2})$. 因此,

$$\min_{k \in \{1, \dots, p_1\}} \{\hat{\delta}_k\} = o_p(n^{-1/2}).$$

所以, $|I_{n11}| = O_p(1) \|\mathbf{T}\| - o_p(\delta_n) \|\mathbf{T}\|$. 另外, 对于 I_{n12} , 有 $I_{n12} \leq \delta_n \|\mathbf{T}\| \|\beta_0\| = O_p(\delta_n \|\mathbf{T}\|)$. 因此, $I_{n1} = O_p(\|\mathbf{T}\|)$. 对于 I_{n2} , 可以得到

$$\begin{aligned} I_{n2} &= \delta_n^2 \mathbf{T}^T \frac{\partial \mathbf{S}_n(\beta)}{\partial \beta} \bigg|_{\beta=\beta^*} \mathbf{T} \\ &= n \delta_n^2 \mathbf{T}^T (\mathbf{I} - \hat{\mathbf{\Lambda}}) \left\{ \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n \mathbf{D}_i^T \mathbf{\Lambda}_{i(k)} \mathbf{V}_{i(k)}^{-1}(\theta_0) \mathbf{\Phi}_{i(k)} \mathbf{D}_i \right\} \mathbf{T} \\ &\quad + \delta_n^2 \mathbf{T}^T \hat{\mathbf{\Lambda}} \mathbf{T} + o_p(n \delta_n^2) \\ &=: I_{n21} + I_{n22} + o_p(n \delta_n^2). \end{aligned}$$

利用与上面相同的方法, 可以证明 $I_{n22} = O_p(\delta_n^2 \|\mathbf{T}\|) = o(n \delta_n^2 \|\mathbf{T}\|)$. 因此, 在 $\{\beta_0 + \mathbf{T} : \|\mathbf{T}\| = C\}$ 上, $\delta_n \mathbf{T}^T \mathbf{S}_n(\beta_0 + \mathbf{T})$ 渐近地由 I_{n21} 所控制, 而对于充分大的 C , I_{n21} 是正的. 这就完成了定理 3.1 的证明. $\square$

定理 3.2 的证明 (1) 对于给定的 $k \in \{p_1 + 1, \dots, p\}$, $|\hat{\beta}_k^s - \beta_{0k}| = |\hat{\beta}_k^s| = O_p(n^{-1/2})$, 另外, 基于条件 $n^{(1+\tau)/2} \lambda \rightarrow \infty$, 可以得到

$$P\left(\frac{\lambda}{|\hat{\beta}_k^s|^{1+\tau}} < 0\right) = P(|\hat{\beta}_k^s|^{1+\tau} > \lambda) \leq \lambda^{-1} E(|\hat{\beta}_k^s|^{1+\tau}) = \lambda^{-1} O(n^{-(1+\tau)/2}) \rightarrow 0,$$

这就意味着 $P(\hat{\delta}_k = 1, k \in \{p_1 + 1, \dots, p\}) \rightarrow 1$. 这就完成了定理 3.2(1) 的证明.

(2) 证明方法与文献 [37, 定理 2.2] 的证明方式类似. 为了节省空间, 在此省略. $\square$

致谢 感谢主编和两位审稿人提出的宝贵意见.

参考文献

- 1 Liang K Y, Zeger S L. Longitudinal data analysis using generalized linear models. *Biometrika*, 1986, 73: 13–22

- 2 Lai P, Wang Q, Lian H. Bias-corrected GEE estimation and smooth-threshold GEE variable selection for single-index models with clustered data. *J Multivariate Anal*, 2012, 105: 422–432
- 3 Tian R, Xue L, Hu Y. Smooth-threshold GEE variable selection for varying coefficient partially linear models with longitudinal data. *J Korean Statist Soc*, 2015, 44: 419–431
- 4 Lv J, Yang H, Guo C. An efficient and robust variable selection method for longitudinal generalized linear models. *Comput Statist Data Anal*, 2015, 82: 74–88
- 5 Wang L, Zhou J, Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 2012, 68: 353–360
- 6 Li G, Lian H, Feng S, et al. Automatic variable selection for longitudinal generalized linear models. *Comput Statist Data Anal*, 2013, 61: 174–186
- 7 Lian H, Liang H, Wang L. Generalized additive partial linear models for clustered data with diverging number of covariates using GEE. *Statist Sinica*, 2014, 24: 173–196
- 8 Zhao P, Li G. Modified SEE variable selection for varying coefficient instrumental variable models. *Stat Methodol*, 2013, 12: 60–70
- 9 Fan Y, Qin G, Zhu Z. Variable selection in robust regression models for longitudinal data. *J Multivariate Anal*, 2012, 109: 156–167
- 10 He X, Fung W K, Zhu Z. Robust estimation in generalized partial linear models for clustered data. *J Amer Statist Assoc*, 2005, 100: 1176–1184
- 11 Qin G, Zhu Z, Fung W K. Robust estimation of covariance parameters in partial linear model for longitudinal data. *J Statist Plann Inference*, 2009, 139: 558–570
- 12 Qin G, Bai Y, Zhu Z. Robust empirical likelihood inference for generalized partial linear models with longitudinal data. *J Multivariate Anal*, 2012, 105: 32–44
- 13 Qin G, Zhu Z. Robust estimation in generalized semiparametric mixed models for longitudinal data. *J Multivariate Anal*, 2007, 98: 1658–1683
- 14 Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. *Ann Statist*, 2008, 36: 1108–1126
- 15 Kai B, Li R, Zou H. Local composite quantile regression smoothing: An efficient and safe alternative to local polynomial regression. *J R Stat Soc Ser B Stat Methodol*, 2010, 72: 49–69
- 16 Kai B, Li R, Zou H. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Ann Statist*, 2011, 39: 305–332
- 17 Jiang X J, Jiang J C, Song X Y. Oracle model selection for nonlinear models based on weighted composite quantile regression. *Statist Sinica*, 2012, 22: 1479–1506
- 18 Zhao W H, Lian H, Song X Y. Composite quantile regression for correlated data. *Comput Statist Data Anal*, 2017, 109: 15–33
- 19 Sun J, Frees E W, Rosenberg M A. Heavy-tailed longitudinal data modeling using copulas. *Insurance Math Econom*, 2008, 42: 817–830
- 20 Song P X K. Multivariate dispersion models generated from Gaussian copula. *Scand J Statist*, 2000, 27: 305–320
- 21 Bai Y, Kang J, Song P X K. Efficient pairwise composite likelihood estimation for spatial-clustered data. *Biometrics*, 2014, 70: 661–670
- 22 Fu L, Wang Y G. Efficient parameter estimation via Gaussian copulas for quantile regression with longitudinal data. *J Multivariate Anal*, 2016, 143: 492–502
- 23 Wang H, Feng X, Dong C. Copula-based quantile regression for longitudinal data. *Statist Sinica*, 2019, 29: 245–264
- 24 Brown B M, Wang Y G. Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, 2005, 92: 149–158
- 25 Jung S H. Quasi-likelihood for median regression models. *J Amer Statist Assoc*, 1996, 91: 251–257
- 26 Haff I, Aas K, Frigessi A. On the simplified pair-copula construction—simply useful or too simplistic? *J Multivariate Anal*, 2010, 101: 1296–1310
- 27 Smith M, Min A, Almeida C, et al. Modeling longitudinal data using a pair-copula decomposition of serial dependence. *J Amer Statist Assoc*, 2010, 105: 1467–1479
- 28 Aas K, Czado C, Frigessi A, et al. Pair-copula constructions of multiple dependence. *Insurance Math Econom*, 2009, 44: 182–198
- 29 Hendricks W, Koenker R. Hierarchical spline models for conditional quantiles and the demand for electricity. *J Amer Statist Assoc*, 1992, 87: 58–68
- 30 Hall P, Sheather S. On the distribution of a studentized quantile. *J R Stat Soc Ser B Stat Methodol*, 1988, 50: 381–391
- 31 Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B Stat Methodol*, 1996, 58: 267–288
- 32 Fu W. Penalized regression: The Bridge versus the LASSO. *J Comput Graph Statist*, 1998, 7: 397–416

- 33 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 34 Zou H. The adaptive Lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429
- 35 Ueki M. A note on automatic variable selection using smooth-threshold estimating equations. *Biometrika*, 2009, 96: 1005–1011
- 36 Lv J, Yang H, Guo C. Smoothing combined generalized estimating equations in quantile partially linear additive models with longitudinal data. *Comput Statist*, 2016, 31: 1203–1234
- 37 Wang K, Lin L. Robust and efficient estimator for simultaneous model structure identification and variable selection in generalized partial linear varying coefficient models with longitudinal data. *Statist Papers*, 2019, 60: 1649–1676

Copula based composite quantile regression for longitudinal data and variable selection

Kangning Wang, Shaomin Li & Lu Lin

Abstract Composite quantile regression (CQR) is a good alternative of the mean regression, because of its robustness and efficiency. As is well known, longitudinal data is characterized by within-subject correlation, and if the correlation can be correctly incorporated in the estimation procedure, the efficiency will be improved. Hence, how to specify the correlation structure in CQR with longitudinal data is an interesting issue. We propose a new approach that uses copula to account for intra-subject dependence, and by using the copula based covariance matrix, efficient and unbiased CQR estimating equations are constructed. As a specific application, a smooth-threshold CQR estimation equation is proposed for variable selection. Our proposed new methods are flexible, and can provide efficient estimation. The properties of the proposed methods are established theoretically, and assessed numerically through simulation studies and real data analysis.

Keywords composite quantile regression, longitudinal data, copula function, efficiency, variable selection

MSC(2010) 62G05, 62G20

doi: 10.1360/N012018-00298