中国科学: 数学 2024 年 第 54 卷 第 2 期: 181 ~ 210
SCIENTIA SINICA Mathematica
论 文

复杂高维异质性数据的加权分位回归方法

熊巍[1*], 潘晗[2], 虞克明[3], 田茂再[4]

1. 对外经济贸易大学统计学院, 北京 100029;
2. 北京大学数学科学学院, 北京 100871;
3. Department of Mathematics, Brunel University London, Kingston Lane, Uxbridge UB9 3PH, UK;
4. 中国人民大学统计学院, 北京 100872
E-mail: xiongwei@uibe.edu.cn, scott_pan@163.com, Keming.Yu@brunel.ac.uk, mztian@ruc.edu.cn

收稿日期: 2022-05-05; 接受日期: 2023-05-15; 网络出版日期: 2023-12-04; * 通信作者
国家自然科学基金 (批准号: 12001101)、对外经济贸易大学中央高校基本科研业务费专项资金 (批准号: CXTD14-05) 和对外经济贸易大学优秀青年学者 (批准号: 20YQ18) 资助项目

摘要 随着数字化智能技术的发展, 信息泛滥、算力膨胀、数据异构性及混杂性等问题频现, 给数据建模的理论方法带来极大挑战. 本文从众数角度出发, 提出最优分位水平概念和基于众数的加权分位回归 (mode-based weighted quantile regression, MWQR) 方法, 以求最大程度利用样本信息. 与已有估计方法相比, MWQR 方法具有如下优势: (1) 适用于复杂高维异质性数据, 在误差分布厚尾和偏态时仍能保证稳健性; (2) 解决了分位回归建模中分位水平主观选择的问题; (3) 通过赋予不同分位水平不同权重, 极大提升估计效率, 减少运算时间; (4) 有效探测响应变量的条件分布. 鉴于 MWQR 方法的优势, 本文进一步将其应用于部分线性可加模型, 提出两种算法进行变量选择和系数估计, 并探究理论性质. 数值模拟及城投债 "隐性担保" 和血浆 β- 胡萝卜素浓度两组实际数据分析, 表明该方法能很好地挖掘数据内蕴结构, 显著提高运算效率, 具有广泛的应用价值.

关键词 众数 最优分位水平 加权分位回归 部分线性可加模型 变量选择

MSC (2020) 主题分类 62G05, 62P10, 62P20

1 引言

随着全球新一轮科技革命和产业变革的深入推进, 以 5G 和人工智能等为代表的信息技术日益成为经济社会数字化转型发展的关键驱动力量, 全球正在迈入数字化智能技术的新时代. 数字化如无形的触手渗透到生活、工作、学习和娱乐中, 深刻改变着人类的生产生活方式. 然而数字化也带来了信息泛滥、算力膨胀、数据异构及数据混杂等多种问题, 给传统的数据分析方法和数据建模理论带来了极大的挑战. 目前, 国内外学者已通过数据分布式计算、随机抽样等方式提升计算机运行速度, 以缓解数据的海量化问题, 如 Li 等 [13] 和 Guo 等 [4]. 然而大数据本身的复杂化、高维化和局部异质性等特

英文引用格式: Xiong W, Pan H, Yu K M, et al. A weighted quantile regression approach for complex high-dimensional heterogeneous data (in Chinese). Sci Sin Math, 2024, 54: 181–210, doi: 10.1360/SSM-2022-0080

征以及其潜在的错综复杂的变量关系仍是困扰研究者的重点与难点. 据此, 本文从统计模型角度出发, 提出一种全新的高效稳健建模方法以适应数字化时代数据的多样性和信息的多元化需求.

数字化信息时代的经济活动产生了更多高维及复杂经验数据, 决策行为将日益基于数据分析而非经验和直觉. 统计模型理论的不断创新发展及其在科学前沿领域的深入应用为新范式下的社会经济研究提供了可能, 研究方式从传统的线性、低维、有限样本、简单参数模型得以向当前的非线性、高维、大样本、复杂非参数混合模型转化. 在众多统计模型中, 部分线性可加模型 (partially linear additive model, PLAM) 由于综合了参数模型和非参数模型的多种优势, 解释力强, 极具灵活性, 被广泛应用于经济金融、医药卫生和环境生态等领域. Sun 等 [26] 选取现金占净比和消费者价格指数等指标, 基于部分线性可加模型对余额宝收益率影响因素进行了探究; Song 等 [24] 将部分线性可加模型拓展到面板数据, 选取外商直接投资依存度等 4 个衡量经济开放程度的指标分析经济开放程度对企业创新的提升与促进作用; Ibarra-Espinosa 等 [9] 利用部分线性可加模型研究了新型冠状病毒肺炎的每日确诊人数和死亡人数与人口流动性和空气污染之间的关系. 部分线性可加模型表示如下:

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{p} g_j(Z_j) + \varepsilon, \tag{1.1}$$

其中, Y 为响应变量, $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ 为 d 维协变量, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$ 为参数向量, $Z_j \in [0,1]$ 为非参数部分的协变量, $g_j(\cdot)$ $(j = 1, \ldots, p)$ 为一维光滑函数, ε 为随机误差项. 如 (1.1) 所示, 部分线性可加模型既包含参数部分又包含非参数可加部分, 不仅能够刻画复杂数据结构, 揭示数据模式, 还适用于高维数据, 能有效避免维数灾难. 例如, Kazemi 等 [10] 基于部分线性可加模型对基因数据进行分析, 识别出了影响 G 蛋白偶联受体超量表达的 12 个重要基因.

由于 (1.1) 中包含未知的参数部分和非参数部分, 因此现有部分线性可加模型的估计方法常采用迭代, 算法可以分为 3 类: 反向拟合 (backfitting) 算法 [21, 25]、核回归算法 [1, 15] 和分位回归算法 [7, 11]. 但上述算法存在一定局限性, 例如, 无法同时估计模型中的参数系数及非参数函数, 需要迭代; 误差项通常假定服从正态分布或方差有限, 不够稳健. 另外, 随着数据收集能力的提高, 信息来源和获取方式呈现多样性, 高维异质性数据频现. 尽管部分线性可加模型在一定程度上能够处理高维数据, 但不够有效. 于是近年来, 部分线性可加模型的变量选择方法逐渐成为研究热点. Liu 等 [16] 基于多项式样条展开法实现了非参数部分的变量选择. Lian [14] 结合自适应最小绝对收缩和选择算子 (least absolute shrinkage and selection operator, LASSO) 实现了部分线性可加模型的变量选择和参数估计. 然而上述算法过度依赖最小二乘 (least square, LS) 方法, 其估计精度和效率对于极端值和厚尾分布异常敏感. 据此, Guo 等 [3] 采用复合分位回归 (composite quantile regression, CQR) 方法并结合自适应 LASSO, 提出了稳健的 B 样条逼近复合分位数回归 (B-spline approximation composite quantile regression, BSA-ACQR) 算法以实现部分线性可加模型的参数估计及变量选择, 但 BSA-ACQR 算法仅考虑了参数部分的变量选择, 有一定局限性. Lv 等 [17] 基于众数回归并利用平滑剪裁绝对偏差 (smoothly clipped absolute deviation, SCAD) 惩罚函数实现了参数及非参数部分的同时变量选择. 近来, Nguelifack 和 Kemajou-Brown [20] 也提出了一种稳健的符号秩估计方法, 并有效实现了部分线性可加模型的变量选择.

计算的可行性、算法的稳健性和估计的有效性一直是应用统计关注的核心问题. 据此, 本文提出一种全新的估计方法—基于众数的加权分位回归 (mode-based weighted quantile regression, MWQR) 方法, 在保证算法稳健性的同时, 还能有效提升运算速度, 且估计效率优于已有方法. 该方法建立在分位回归方法和众数理论的框架下, 能够有效利用数据分布中的位置信息. 之所以选取众数是因为众数作为 “最有可能出现的值”, 既是描述数据中心位置的重要参数, 也是对均值和分位数的重要补充; 其

作为连续分布密度函数的最高点, 不仅具有丰富的信息量, 而且还对离群值和厚尾分布具有较强的耐抗性. 本文所构建的 MWQR 方法试图通过众数选取令每个样本最具有信息量的分位水平, 也即最优分位水平, 并通过对不同的最优分位水平进行加权汇总, 有效综合样本信息, 极大降低估计误差. 因此, MWQR 方法不仅具有双重稳健性, 而且更具有广泛适用性, 能够应用于多种模型, 助力实际问题解决. 鉴于 MWQR 方法的优势, 本文以部分线性可加模型为例, 提出适用的 MWQR-PLAM 算法, 并结合自适应 LASSO 提出惩罚 MWQR-PLAM (penalized MWQR-PLAM, PMWQR-PLAM) 算法, 实现模型中参数及非参数部分的同时变量选择和估计. 在非参数函数的估计中, 本文使用 B 样条基函数逼近非参数部分. 与核方法相比, 使用 B 样条基函数具有计算更快、结果更准确等特点, 因此在非参数和半参数领域得到了广泛应用. He 和 Shi [5] 探究了基于 B 样条基函数的非参数条件分位数函数的估计量的收敛速率; 为有效分析高维异质性数据, He 等 [6] 提出了一种分位数适应的非线性特征筛选方法, 并通过 B 样条基模拟感兴趣分位水平的边际效应; Sherwood 和 Wang [23] 研究了超高维下的部分线性可加分位数回归模型, 并利用 B 样条基估计模型的非参数部分. Wang 等 [29] 使用 B 样条基逼近未知时间趋势, 实现了非平稳时间序列的建模. 与 LS 相比, MWQR-PLAM 算法更加准确有效, 能够规避极端值带来的影响, 在误差服从厚尾分布下依然有效; 与加权分位回归 (weighted quantile regression, WQR) 方法和 CQR 方法相比, MWQR-PLAM 算法能够避免分位水平主观选取所带来的偏误, 在误差服从偏态分布下依然表现良好; 与分位回归方法 (quantile regression, QR) 及众数回归方法相比, 该算法使用多个分位水平探测响应变量的整个条件分布, 更加全面准确. 理论上, 本文探究了所提出算法的优良性质; 数值模拟和实证研究验证了所提出方法在估计精度、算法效率、计算稳健性和数据应用中的明显优势.

本文余下内容的安排如下. 第 2 节提出基于众数的加权最优分位方法, 并探究该方法的渐近相对效率. 第 3 节给出适用于部分线性可加模型的 MWQR-PLAM 算法, 并结合自适应 LASSO 惩罚给出用于变量选择的 PMWQR-PLAM 算法. 第 4 节为模拟研究. 第 5 节进行实证分析, 分析经济学及卫生领域的两组实际数据. 第 6 节给出本文结论. 附录 A 给出定理的证明.

2 基于众数的加权分位回归方法

2.1 分位回归与加权分位回归

本小节从简单线性模型 (2.1) 出发, 给出 MWQR 方法的思想和原理, 再逐步将其拓展到部分可加线性模型 (1.1) 中. 首先考虑如下线性模型:

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon, \tag{2.1}$$

其中 $\boldsymbol{X}$、Y、$\boldsymbol{\beta}$ 和 ε 如 (1.1) 中定义. 假定一组容量为 n 的样本 $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ 来自模型 (2.1), 分位回归方法旨在求解 $(\hat{c}_\tau, \hat{\boldsymbol{\beta}}_\tau) = \arg\min_{c,\boldsymbol{\beta}} \sum_{i=1}^n \rho_\tau(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - c)$, 其中 $\rho_\tau(u) = u(\tau - I(u<0))$ 为分位损失函数. 且有分位回归估计量 $\hat{\boldsymbol{\beta}}_\tau$ 渐近服从正态分布:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}) \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau))}\boldsymbol{D}^{-1}\right), \tag{2.2}$$

其中, $f(\cdot)$ 和 $F(\cdot)$ 分别为误差项 ε 的密度函数和分布函数, $\boldsymbol{D} = \lim_{n\to\infty} n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 为正定的对称矩阵, $\xrightarrow{d}$ 表示依分布收敛. 若 $\varepsilon \sim N(0, \sigma^2)$, 则参数 $\boldsymbol{\beta}$ 的最小二乘估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ 同样渐近服从正态分布, 且渐近方差为 $\sigma^2\boldsymbol{D}^{-1}$. Xiong 和 Tian [30] 指出, (2.2) 中的 $\tau(1-\tau)/f^2(F^{-1}(\tau))$ 与 $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ 渐近方差中的

σ^2 扮演着相同的角色, 可将 $\sqrt{\tau(1-\tau)}/f(F^{-1}(\tau))$ 视为分位回归框架下误差标准差的一种度量. 此外, $1/f(F^{-1}(\tau))$ 刻画了分位水平 τ 附近的数据分布情形, 取值越小, 数据越稠密; 取值越大, 数据越稀疏. 因此 $1/f(F^{-1}(\tau))$ 也称为稀疏函数或分位密度函数[28], 本文用 $s(\tau)$ 加以表示. Xiong 和 Tian[30] 基于稀疏函数 $s(\tau)$ 的特点, 通过提出一种适用于不同分位水平的权重 $w_k = 1/s(\tau)\sqrt{\tau(1-\tau)}$, 构造参数 $\boldsymbol{\beta}$ 的 WQR 估计量. 具体地, 针对选定的 K 个分位水平 $\{\tau_k, k = 1, \ldots, K\}$, 分别计算不同分位水平的参数估计 $\hat{\boldsymbol{\beta}}_{\tau_k}$, 并对各 $\hat{\boldsymbol{\beta}}_{\tau_k}$ 赋以权重 w_k, 再将权重标准化 $w_k^* = w_k/\sum_k w_k$, 最终得到模型 (2.1) 中参数 $\boldsymbol{\beta}$ 的 WQR 估计量

$$\hat{\boldsymbol{\beta}}_{\mathrm{WQR}} = \sum_{k=1}^{K} w_k^* \hat{\boldsymbol{\beta}}_{\tau_k}. \tag{2.3}$$

2.2 基于众数的加权分位回归

上述 WQR 方法中, 分位水平的选择具有较强主观性. 为充分利用样本信息并规避分位水平人为选择的主观问题, 本文提出一种基于众数的 MWQR 方法. 首先给出最优分位水平的定义.

定义 2.1 给定一组来自总体 $(\boldsymbol{X}, Y)$ 的样本 $(\boldsymbol{X}_i, Y_i)_{i=1}^n$, $\boldsymbol{X}$ 为 d 维随机向量, Y 为一维随机变量. 令 $f(y \mid \boldsymbol{x})$ 为给定 $\boldsymbol{X} = \boldsymbol{x}$ 下 Y 的条件密度, 并假定 $f(y \mid \boldsymbol{x})$ 关于 $y \in \mathcal{R}$ 连续, 并且对于每一个给定 $\boldsymbol{x}$, 存在唯一的众数 $m(\boldsymbol{x})$ 使得 $f(m(\boldsymbol{x})|\boldsymbol{x}) = \max_{y\in\mathcal{R}} f(y \mid \boldsymbol{x})$, 则第 i 个样本 $\boldsymbol{x}_i$ 的最优分位水平 $\tau_{\boldsymbol{x}_i}$ 定义为 $\tau_{\boldsymbol{x}_i} = \arg\max_{\tau\in(0,1)} f(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x}_i)$, 其中 $Q_{\boldsymbol{x}}(\tau) := Q_Y(\tau \mid \boldsymbol{X} = \boldsymbol{x})$ 为给定 $\boldsymbol{X} = \boldsymbol{x}$ 下 Y 的 $\tau\%$ 条件分位数.

注 2.1 (1) 由定义 2.1 可知, 最优分位水平 $\tau_{\boldsymbol{x}_i}$ 的获取依赖于条件众数的估计. 由于 $f(y \mid \boldsymbol{x})$ 关于 y 连续性和众数唯一性的假设, 因此存在唯一 $\tau_{\boldsymbol{x}_i}$ 使得 $f(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x}_i) = \max_{y\in\mathcal{R}} f(y \mid \boldsymbol{x}_i)$, 即对于每个样本 $\boldsymbol{x}_i$, 其最优分位水平存在且唯一. (2) 如上最优分位水平的定义可以推广到条件分布具有多个众数的情形. 定义局部众数集合为 $M(\boldsymbol{x}) = \{y : \frac{\partial}{\partial y} f(y \mid \boldsymbol{x}) = 0, \frac{\partial^2}{\partial y^2} f(y \mid \boldsymbol{x}) < 0\}$. 于是对于给定的 $\boldsymbol{x}$, $M(\boldsymbol{x})$ 可能包含多个点, 也即 $M(\boldsymbol{x})$ 是一个多值函数. 按照如上最优分位水平的定义, 每个局部众数对应于一个局部最优分位水平. 假定 $f(y \mid \boldsymbol{x})$ 存在 t 个局部众数, 则对应有 t 个局部最优分位水平. 最优分位水平则为使得条件密度达到全局最大值所对应的局部最优分位水平. 若存在多个局部分位水平使得条件密度达到全局最大值, 则最优分位水平不唯一, 理论上任意满足条件的局部最优分位水平均为最优分位水平. 若存在多个局部分位水平使得条件密度达到全局最大值, 此时最优分位水平不唯一, 理论上任意满足条件的局部最优分位水平均为最优分位水平. 当条件密度存在多个局部众数时, 本文所提出的 MWQR 估计方法依然有效, 详见第 4 节模拟例 4.2. (3) 最优分位水平依赖于样本容量, 一个样本容量为 n 的样本对应有 n 个最优分位水平, 但不同样本可能享有相同的最优分位水平. 为计算简便, 本文建议在预先给定的一组分位水平 $\{\tau_k\}_{k=1}^K$ 中确定样本 $\boldsymbol{x}_i$ 的最优分位水平, 例如, 考虑一组均匀分位水平: $\tau_k = k/(K+1)$, $k = 1, \ldots, K$.

估计不同样本对应的最优分位水平即确定不同样本的条件众数. 令 $s_{\boldsymbol{x}_i} = 1/f(Q_{\boldsymbol{x}}(\tau) \mid \boldsymbol{x}_i)$, 易知 $s_{\boldsymbol{x}_i}(\tau) = \partial Q_{\boldsymbol{x}_i}(\tau)/\partial\tau$. 于是 $s_{\boldsymbol{x}_i}$ 的估计可以通过下式实现:

$$\hat{s}_{\boldsymbol{x}_i}(\tau) = \frac{\hat{Q}_{\boldsymbol{x}_i}(\tau + h) - \hat{Q}_{\boldsymbol{x}_i}(\tau - h)}{2h},$$

其中, $\hat{Q}_{\boldsymbol{x}_i} = \boldsymbol{x}_i^{\mathrm{T}}\hat{\beta}_\tau$, h 为窗宽. 根据最优分位水平的定义, 最大化条件众数即最小化 $s_{\boldsymbol{x}_i}(\tau)$, 于是最优分位水平 $\tau_{\boldsymbol{x}_i}$ 可由 (2.4) 得到,

$$\tau_{\boldsymbol{x}_i} = \arg\min_{\tau\in(0,1)} \hat{s}_{\boldsymbol{x}_i}(\tau), \quad i = 1, \ldots, n. \tag{2.4}$$

本文借鉴 Xiong 和 Tian [30] 提出的权重形式, 赋予第 i 个样本权重 $w_i = 1/s(\tau_{\boldsymbol{x}_i})\sqrt{\tau_{\boldsymbol{x}_i}(1-\tau_{\boldsymbol{x}_i})}$, 其中 $\tau_{\boldsymbol{x}_i}$ 为对应的最优分位水平, 并对权重进行正规化修正 $w_i^* = w_i/\sum_{i=1}^n w_i$. 于是模型 (2.1) 中参数 $\boldsymbol{\beta}$ 的基于条件众数的 MWQR 估计量为

$$\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}} = \sum_{i=1}^{n} w_i^* \hat{\boldsymbol{\beta}}_{\tau_{\boldsymbol{x}_i}}. \tag{2.5}$$

注 2.2 (1) 权重的计算. 本文采用插入法 (plug-in) 计算权重, 即 $\hat{w}_k = \varphi(\Phi^{-1}(\tau_k))/\sqrt{\tau_k(1-\tau_k)}$, 其中 $\varphi(\cdot)$ 与 $\Phi(\cdot)$ 分别为标准正态分布的密度函数与分布函数. 实际应用中当样本量充分大时, 假定随机误差服从正态分布是合理的. (2) 窗宽 h 的选择. Koenker 和 Machado [12] 建议最优窗宽为 $h^{KM}(\tau) = n^{-1/3} z_\alpha^{2/3}\{1.5\varphi^2(\Phi^{-1}(\tau))/(2\Phi^{-1}(\tau)^2+1)\}^{1/3}$, 其中 z_α 满足 $\Phi(z_\alpha) = 1-\alpha$. 为使估计值更加精准并兼具优良统计性质, 本文参考文献 [22], 对窗宽作如下调整: 首先, 选取 $\tau = 0.5$, 计算窗宽 $h_{n,1} = n^{1/6}h^{KM}(0.5)$ 并得到每个样本的初始最优分位水平 $\tau_{\boldsymbol{x}_i,1}$; 然后, 根据初始最优分位水平更新窗宽 $h_n = n^{1/6}h^{KM}(\tau_{\boldsymbol{x}_i,1})$, 并计算得到最终的最优分位水平 $\tau_{\boldsymbol{x}_i}$ $(i = 1,\ldots,n)$. (3) $\hat{s}_{\boldsymbol{x}}(\tau)$ 的求解. 在计算 $\hat{s}_{\boldsymbol{x}}(\tau)$ 的过程中, 若出现 $\tau - h < 0$ 或 $\tau + h > 1$, 则在估计过程作如下修正:

$$\hat{s}_{\boldsymbol{x}}(\tau) = \frac{\hat{Q}_{\boldsymbol{x}}(\tau + \min\{h, \tau_{\max} - \tau\}) - \hat{Q}_{\boldsymbol{x}}(\tau - \min\{h, \tau - \tau_{\min}\})}{\min\{h, \tau_{\max} - \tau\} + \min\{h, \tau - \tau_{\min}\}}, \tag{2.6}$$

其中 $\tau_{\min}$ 和 $\tau_{\max}$ 为预设分位水平的最小值和最大值.

为探究模型 (2.1) 下 $\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}}$ 的渐近性质, 首先给出如下正则化条件.

条件 2.1 误差项 ε 的分布函数为 $F(\cdot)$, 密度函数为 $f(\cdot)$. 对于任意 d 维向量 $\boldsymbol{u}$, 都有

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\int_0^{u_0+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{u}}\sqrt{n}\left[F\left(a+\frac{t}{\sqrt{n}}\right)-F(a)\right]dt = \frac{1}{2}f(a)(u_0,\boldsymbol{u}^{\mathrm{T}})\begin{pmatrix}1 & 0\\ 0 & \boldsymbol{D}\end{pmatrix}(u_0,\boldsymbol{u}^{\mathrm{T}})^{\mathrm{T}}.$$

条件 2.2 设计阵 $\boldsymbol{X}$ 满足 $\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{D}$, 且 $\boldsymbol{D}$ 为 $d\times d$ 维正定矩阵.

条件 2.1 和 2.2 为建立分位回归渐近正态性的常规条件.

定理 2.1 若条件 2.1–2.2 成立, 则有

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\mathrm{MWQR}}),$$

其中

$$\boldsymbol{\Sigma}_{\mathrm{MWQR}} = \frac{\boldsymbol{D}^{-1}\sum_{i=1}^K\sum_{j=1}^K n_i n_j \frac{\tau_i\wedge\tau_j - \tau_i\tau_j}{\sqrt{\tau_i(1-\tau_i)}\sqrt{\tau_j(1-\tau_j)}}}{(\sum_{k=1}^K n_k \frac{f(F^{-1}(\tau_k))}{\sqrt{\tau_k(1-\tau_k)}})^2},$$

$$n_k = \#\{i\in\{1,\ldots,n\} : \tau_{\boldsymbol{x}_i} = \tau_k\}, \quad \tau_i\wedge\tau_j = \min(\tau_i,\tau_j).$$

当 $n_1 = \cdots = n_k$ 时, 渐近分布退化为 $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \Sigma_{\mathrm{WQR}})$, 其中

$$\Sigma_{\mathrm{WQR}} = \frac{\boldsymbol{D}^{-1}\sum_{i=1}^K\sum_{j=1}^K \frac{\tau_i\wedge\tau_j - \tau_i\tau_j}{\sqrt{\tau_i(1-\tau_i)}\sqrt{\tau_j(1-\tau_j)}}}{(\sum_{k=1}^K \frac{f(F^{-1}(\tau_k))}{\sqrt{\tau_k(1-\tau_k)}})^2}.$$

定理 2.1 的证明见附录 A.

2.3 渐近相对效率

为探究 MWQR 方法的有效性, 本小节讨论线性模型 (2.1) 下的 $\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}}$ 与 $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ 和 $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$ 的渐近相对效率 (asymptotic relative efficiency, ARE), 分别记为 ARE_1^* 和 ARE_2^*. 令 MSE 为均方误差, σ^2 为误差方差, 则根据定理 2.1 有

$$\mathrm{ARE}_1^* = \frac{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}})}{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}})} = \frac{\sigma^2\big(\sum_{i=1}^K n_i \frac{f(F^{-1}(\tau_i))}{\sqrt{\tau_i(1-\tau_i)}}\big)^2}{\sum_{i=1}^K\sum_{j=1}^K n_i n_j \frac{\tau_i\wedge\tau_j-\tau_i\tau_j}{\sqrt{\tau_i(1-\tau_i)}\sqrt{\tau_j(1-\tau_j)}}}, \tag{2.7}$$

$$\mathrm{ARE}_2^* = \frac{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}})}{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}})} = \frac{(\sum_{i=1}^K\sum_{j=1}^K \tau_i\wedge\tau_j-\tau_i\tau_j)\big(\sum_{i=1}^K n_i \frac{f(F^{-1}(\tau_i))}{\sqrt{\tau_i(1-\tau_i)}}\big)^2}{\sum_{i=1}^K\sum_{j=1}^K n_i n_j \frac{\tau_i\wedge\tau_j-\tau_i\tau_j}{\sqrt{\tau_i(1-\tau_i)}\sqrt{\tau_j(1-\tau_j)}}\big(\sum_{i=1}^K f(F^{-1}(\tau_i))\big)^2}. \tag{2.8}$$

可见, ARE_1^* 和 ARE_2^* 依赖于误差分布、所选取的分位水平以及不同分位水平所对应的样本容量. 当 $n_1 = n_2 = \cdots = n_K = n/K$ 时, ARE_1^* 和 ARE_2^* 分别退化为文献 [30] 中的 ARE_1 和 ARE_2. 相较于 WQR 方法而言, MWQR 方法更为一般. Xiong 和 Tian [30] 指出, 针对不同的误差分布, 恒有 $\mathrm{ARE}_1 \geqslant 16\sigma^2(\mathrm{E}[f(\varepsilon)])^2/(\pi^2-8)$, $\mathrm{ARE}_2 \geqslant 4/3(\pi^2-8) \approx 0.7132$. 可见 ARE_2 的下界并不依赖于误差分布, 即在极端差的情形下 MWQR 的有效性都会达到 CQR 的 70% 以上, 而一般情形下, ARE 的值都会远大于 1. 本文通过如下 4 种不同的误差分布说明此问题: (1) 标准正态分布 $N(0,1)$; (2) 自由度为 4 的卡方分布 χ_4^2; (3) 自由度为 3 的 t 分布 t_3; (4) 混合正态分布 $MN(0,0,1,10^2,0.9)$, 其中 $MN(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\alpha) := \alpha N(\mu_1,\sigma_1^2) + (1-\alpha)N(\mu_2,\sigma_2^2)$. 考虑两种不同情形. 情形 1: 各分位水平所对应的样本容量相同, 即 $n_1 = n_2 = \cdots = n_K = n/K$. 情形 2: 各分位水平所对应样本容量均不相等, 取 $n_k = 2n[2-(k-1)/(K-1)]/(3K)$, $k = 1,2,\ldots,K$. 该设定下, $n_1 = 2n_K$. 计算过程中, 样本容量取 $n = 100{,}000$, MWQR 和 CQR 采用均匀分位水平 $\tau_k = k/(K+1)$, $k = 1,\ldots,K$, 并设定 $K = 3,5,9,19,99$. ARE_1^* 和 ARE_2^* 值详见表 1.

由表 1 可得如下结论: (1) 对比情形 1 和 2, 当误差分布服从对称分布或近似对称分布时, 若各最优分位水平对应的样本量相同, 则 WQR 方法更有效; 而当误差分布为有偏分布时, 若各最优分位水平对应的样本量不同, 则 MWQR 方法的效率更优. (2) 当误差项服从标准正态分布时, MWQR 方法几乎在所有 K 取值下均优于 CQR 方法, 而与 LS 方法相比, 其渐近相对效率随着 K 的增加逐渐趋向于 1. 当误差项服从偏态分布时 (如 χ_4^2), MWQR 方法表现最优, 这是由于基于众数的 MWQR 方法能

表 1 ARE_1^* 和 ARE_2^* 在不同误差分布及 K 下的理论值

情形	误差分布	ARE_1^*					ARE_2^*				
		$K=3$	$K=5$	$K=9$	$K=19$	$K=99$	$K=3$	$K=5$	$K=9$	$K=19$	$K=99$
情形 1	$N(0,1)$	0.859	0.915	0.952	0.973	0.985	1.003	1.010	1.017	1.025	1.032
	χ_4^2	1.203	1.359	1.496	1.613	1.725	1.009	1.031	1.062	1.101	1.152
	t_3	1.858	1.863	1.844	1.819	1.797	0.995	0.983	0.970	0.957	0.946
	$MN(0,0,1,10^2,0.9)$	7.365	7.396	7.290	6.954	6.804	0.998	0.996	0.983	0.953	0.935
情形 2	$N(0,1)$	0.838	0.898	0.937	0.960	0.973	0.979	0.991	1.001	1.011	1.019
	χ_4^2	1.358	1.550	1.723	1.871	2.017	1.140	1.175	1.221	1.275	1.342
	t_3	1.815	1.830	1.816	1.796	1.776	0.970	0.965	0.955	0.944	0.934
	$MN(0,0,1,10^2,0.9)$	7.218	7.266	7.183	6.868	6.738	0.978	0.977	0.967	0.939	0.925

够更好地识别信息量最大的分位水平, 从而得到最优的估计效率; 而在误差服从厚尾分布下 (如 t 分布及混合正态分布), MWQR 方法与 CQR 方法表现相当, 但显著优于 LS 方法. 综上, MWQR 方法是一种优良的参数估计方法. (3) MWQR 方法对分位水平个数 K 并不敏感. 考虑到计算便捷, 本文均使用 $K=9$.

3 稳健的部分线性可加模型估计方法及变量选择

3.1 基于 MWQR 的稳健部分线性可加模型估计方法

为估计部分线性可加模型 (1.1) 中的非参数函数, 本文使用 3 阶 B 样条基进行近似. 假定不同协变量使用相同的节点个数 k_n, 则共有 $L=k_n+4$ 个样条基函数 $\{B_u, 1\leqslant u\leqslant L\}$, 于是 $g_j(z)$ 可近似为 $g_j(z)\approx\sum_{u=1}^{L}\theta_{ju}B_u(z)$, 其中 θ_{ju} 为对应的系数. 给定分位水平 τ_k, 可通过最小化目标函数 (3.1) 得到模型 (1.1) 的 QR 估计量:

$$\sum_{i=1}^{n}\rho_{\tau_k}\bigg(Y_i-c-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-\sum_{j=1}^{p}g_j(Z_{ij})\bigg). \tag{3.1}$$

为保证参数的可识别性, 通常施加约束条件: $\sum_{i=1}^{n}g_j(Z_{ij})=0$, $j=1,2,\ldots,p$. 于是 (3.1) 转化为

$$\sum_{i=1}^{n}\rho_{\tau_k}\bigg(Y_i-c-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-\sum_{j=1}^{p}\sum_{u=1}^{L}\theta_{ju}B_u(Z_{ij})\bigg),\quad \text{s.t. }\sum_{i=1}^{n}\sum_{u=1}^{L}\theta_{ju}B_u(Z_{ij})=0,\quad j=1,\ldots,p. \tag{3.2}$$

参数估计中, 通过中心化处理上述带约束的优化可以转化为无约束的优化问题, 令

$$\bar{B}_{ju}=n^{-1}\sum_{i=1}^{n}B_u(Z_{ij}),\quad \psi_{ju}=B_u(Z_j)-\bar{B}_{ju},\quad \boldsymbol{\varphi}_{ij}=(\psi_{j1}(Z_{ij}),\ldots,\psi_{jL}(Z_{ij}))^{\mathrm{T}},$$
$$\boldsymbol{\psi}_i=(\boldsymbol{\varphi}_{i1}^{\mathrm{T}},\ldots,\boldsymbol{\varphi}_{ip}^{\mathrm{T}})^{\mathrm{T}},\quad \boldsymbol{\theta}_j=(\theta_{j1},\ldots,\theta_{jL})^{\mathrm{T}},\quad \Theta=(\boldsymbol{\theta}_1^{\mathrm{T}},\ldots,\boldsymbol{\theta}_p^{\mathrm{T}})^{\mathrm{T}},\quad j=1,\ldots,p,\quad u=1,\ldots,L,$$

则最小化 (3.1) 最终转化为最小化目标函数 (3.3):

$$\sum_{i=1}^{n}\rho_{\tau_k}(Y_i-c-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{\psi}_i^{\mathrm{T}}\Theta). \tag{3.3}$$

由于 (3.3) 中各参数估计结果与节点个数 k_n 有关, 本文通过最小化 Schwarz 信息准则 (Schwarz information criterion, SIC) 确定最优节点个数, 即 $\mathrm{SIC}(k_n)=\log\{\sum_{i=1}^{n}\rho_{\tau_k}(Y_i-\hat{c}-\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(k_n)}-\boldsymbol{\psi}_i^{\mathrm{T}}\hat{\Theta}_{(k_n)})\}+p(k_n+4)\cdot\log n/(2n)$, 其中 $\hat{\boldsymbol{\beta}}_{(k_n)}$ 和 $\hat{\boldsymbol{\theta}}_{(k_n)}$ 表示使用 k_n 节点个数下参数的 MWQR 估计值. 实际应用中, 节点个数选取的方式不尽相同. 例如, Ma 和 Song[18] 应用 B 样条基函数时建议选取 $k_n=[2n^{1/(2r+1)}]$, He 等[6] 指出, 当 $k_n=[n^{1/(2r+1)}]$ 时, 基于 B 样条的估计量可以达到最优收敛速率, 其中 r 为样条阶数, $[\cdot]$ 为取整函数; 据此, 给出部分线性可加模型的 MWQR 估计给出部分线性可加模型的 MWQR 估计算法 MWQR-PLAM (算法 1).

为了得到模型 (1.1) 中基于 MWQR 方法下参数及非参数函数的渐近性质, 令 $\mathcal{H}$ 表示支撑为 $[a,b]$ 的所有满足如下条件的函数 g 的集合, 其中 g 的 γ 阶导数存在, 且满足 Hölder 条件, 即对于任意 $s,t\in[a,b]$, $v\in(0,1]$, $r=\gamma+\nu>0.5$, 有 $|g^{(\gamma)}(s)-g^{(\gamma)}(t)|\leqslant\delta|s-t|^{\nu}$, δ 为正常数, $0\leqslant\gamma\leqslant\rho-1$. 为了证明方便, 本文假定 $\boldsymbol{x}_i$ 已进行中心化处理, 具有零均值, 且 $z_{ij}\in[0,1]$. 此外还需要如下几个正则条件.

算法 1 MWQR-PLAM

输入: 样本 $\{\boldsymbol{x}_i, \boldsymbol{z}_i, y_i\}_{i=1}^n$, K 个分位水平 $\{\tau_1, \ldots, \tau_K\}$, 节点个数初值 $k_n^{(0)}$.

输出: 参数估计值 $\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}}$、$\hat{\boldsymbol{\theta}}_{\mathrm{MWQR}}$ 和非参数函数估计值 $\hat{g}_j(z)$ $(j = 1, \ldots, p)$.

for $i = 1, \ldots, n$; **do**

  for $m = 1, \ldots, M$; **do**

    计算: $\tau_{\boldsymbol{x}_i}^{(m)} = \arg\min_\tau \frac{\hat{Q}_{(\boldsymbol{x}_i,\psi_i)}(\tau+\min\{h,\tau_{\max}-\tau\}) - \hat{Q}_{(\boldsymbol{x}_i,\psi_i)}(\tau-\min\{h,\tau-\tau_{\min}\})}{\min\{h,\tau_{\max}-\tau\}+\min\{h,\tau-\tau_{\min}\}}$,

    $(\hat{\boldsymbol{\beta}}_{\tau_{\boldsymbol{x}_i}^{(m)}}, \hat{\boldsymbol{\theta}}_{\tau_{\boldsymbol{x}_i}^{(m)}}, \hat{c}_{\tau_{\boldsymbol{x}_i}^{(m)}}) = \arg\min_{\boldsymbol{\beta},\boldsymbol{\theta},c} \sum_{i=1}^n \rho_{\tau_{\boldsymbol{x}_i}^{(m)}}(Y_i - c - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \psi_i^{\mathrm{T}}\boldsymbol{\theta})$,

    $k_n^{(m)} = \arg\min_k \log\{\sum_{i=1}^n \rho_{\tau_{\boldsymbol{x}_i}^{(m)}}(Y_i - \hat{c}_{\tau_{\boldsymbol{x}_i}^{(m)}} - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\tau_{\boldsymbol{x}_i}^{(m)}} - \psi_i^{\mathrm{T}}\hat{\theta}_{\tau_{\boldsymbol{x}_i}^{(m)}})\} + p(k+4)\cdot\log n/(2n)$.

  end for

  $k_n^{(m)} = k_n^{(m+1)}$ 或 $m = M$.

  令 $\hat{\boldsymbol{\beta}}_{\tau_{\boldsymbol{x}_i}} = \hat{\boldsymbol{\beta}}_{\tau_{\tau_i}^{(m)}}$, $\hat{\boldsymbol{\theta}}_{\tau_{\boldsymbol{x}_i}} = \hat{\theta}_{\tau_{\tau_i}^{(m)}}$.

end for

$i = n$.

令: $\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}} = \sum_{i=1}^n w_i^* \hat{\beta}_{\tau_{\boldsymbol{x}_i}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{MWQR}} = \sum_{i=1}^n w_i^* \hat{\boldsymbol{\theta}}_{\tau_{\boldsymbol{x}_i}}$, $\hat{g}_j(z) = \sum_{u=1}^L \hat{\theta}_{\mathrm{MWQR},ju}\psi_{ju}(z)$, $j = 1, \ldots, p$.

条件 3.1 给定 $\boldsymbol{x}_i$ 和 $\boldsymbol{z}_i$, 误差项 ε_i 的条件分布函数和条件密度函数分别为 F_i 和 f_i. 给定分位水平 τ, f_i 在 $F^{-1}(\tau)$ 的邻域内一致有界于 0 和 ∞, 其一阶导数 f_i' 在 $F^{-1}(\tau)$ 的邻域内具有一致上界, $1 \leqslant i \leqslant n$.

条件 3.2 存在正常数 M_1 满足 $|x_{ij}| \leqslant M_1$, 且存在有限正常数 δ_1 和 δ_2, 使得

$$\delta_1 \leqslant \lambda_{\max}(n^{-1}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}) \leqslant \delta_2,$$

$\lambda_{\max}(\cdot)$ 表示矩阵的最大特征值.

条件 3.3 对于 $r > 0.5$, 有 $g_j \in \mathcal{H}$, $j = 1, \ldots, p$, 且 $Eg_j(Z_j) = 0$.

条件 3.4 样条基的维数满足如下条件: $k_n \approx n^{1/(2r+1)}$.

相较于高维均值回归理论中对误差项的假定条件 (如假定尾部服从 Gauss 或亚 Gauss 分布), 条件 3.1 相对宽松. 条件 3.2 要求了模型中线性部分变量及设计阵的表现行为. 条件 3.3 为 B 样条理论的经典假设. Stone [25] 指出使用 B 样条基函数可以有效近似满足 Hölder 条件的函数. 条件 3.4 给出了确定 $\hat{g}$ 最优收敛速率所满足的 L 或 k_n.

定理 3.1 在正则条件 3.1–3.4 下, 有

(1) $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, m\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})$;

(2) $\sqrt{n}(\hat{\Theta}_{\mathrm{MWQR}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, m\boldsymbol{\Sigma}_{\psi}^{-1})$,

其中,

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathrm{cov}(\boldsymbol{X}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\psi}} = \mathrm{cov}(\boldsymbol{\psi}), \quad m = \sum_{i=1}^K \sum_{j=1}^K \frac{\tau_i \wedge \tau_j - \tau_i\tau_j}{\sqrt{\tau_i(1-\tau_i)}\sqrt{\tau_j(1-\tau_j)}} \Bigg/ \left(\sum_{k=1}^K \frac{f(F^{-1}(\tau_k))}{\sqrt{\tau_k(1-\tau_k)}} \right)^2;$$

(3) $\|\hat{g}_j(z_j) - g_j(z_j)\|_2 = O_p((L/n)^{1/2})$, $n^{-1}\sum_{i=1}^n(\hat{g}_j(z_{ij}) - g_j(z_{ij}))^2 = O_p(n^{-2r/(2r+1)})$, 其中, $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n)^{\mathrm{T}}$, $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ 和 $\boldsymbol{\Sigma}_{\boldsymbol{\psi}}$ 表示对应的协方差矩阵.

由定理 3.1 知, PLAM 中参数及非参数部分的渐近性质彼此不相关. 在半参数模型的估计问题中, 通常对非参数部分施加欠光滑 (undersmoothing) 条件以使线性部分参数估计实现 $\sqrt{n}$ 收敛同时满足渐近正态性, 而本文提出的 MWQR-PLAM 方法并不需要施加这些限制, 就可以实现 $\hat{\boldsymbol{\beta}}$ 的 $\sqrt{n}$ 收敛, 同时满足渐近正态分布. 此外, MWQR-PLAM 算法可以实现参数及非参数部分的同时估计.

3.2 基于 MWQR 的部分线性可加模型的变量选择

为有效处理高维数据, 对模型 (1.1) 中的参数与非参数部分同时进行变量选择及参数估计, 定义如下加惩罚项的目标函数:

$$P_{\tau_k}(\boldsymbol{\beta},\Theta)=\sum_{i=1}^{n}\rho_{\tau_k}(Y_i-c-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{\psi}_i^{\mathrm{T}}\Theta)+\sum_{j=1}^{d}p_{\lambda_{k,1j}}(|\boldsymbol{\beta}_j|)+\sum_{t=1}^{p}p_{\lambda_{k,2l}}(\|\Theta_t\|_{H_l}),\tag{3.4}$$

其中, $p_{\lambda_1}(\cdot)$ 和 $p_{\lambda_2}(\cdot)$ 为惩罚函数, λ_k 为分位水平 τ_k 下的调节参数, $\|\Theta_j\|_{H_j}=(\boldsymbol{\theta}_j^{\mathrm{T}}H_j\boldsymbol{\theta}_j^{\mathrm{T}})^{1/2}$, H_j 是一个 $L\times L$ 阶矩阵, 其第 (l,l') 元素为 $\int_0^1\psi_{jl}(z)\psi_{jl'}(z)dz$. 本文选择自适应 LASSO [32] 作为惩罚函数, 即 $p_\lambda(a_j)=\lambda\zeta_j|a_j|$, ζ_j 为相应权重. 据此, (3.4) 可以表示为

$$P_{\tau_k}(\boldsymbol{\beta},\Theta)=\sum_{i=1}^{n}\rho_{\tau_k}(Y_i-c-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{\psi}_i^{\mathrm{T}}\Theta)+\lambda_k\sum_{j=1}^{d}\frac{|\boldsymbol{\beta}_{j,k}|}{|\hat{\boldsymbol{\beta}}_{j,k}^{(0)}|^{1/2}}+\lambda_k\sum_{l=1}^{p}\frac{\|\boldsymbol{\theta}_{l,k}\|_{H_l}}{\|\hat{\boldsymbol{\theta}}_{j,k}^{(0)}\|_{H_l}^{1/2}},\tag{3.5}$$

其中 $\hat{\boldsymbol{\beta}}_j^{(0)}$ 与 $\hat{\boldsymbol{\theta}}_j^{(0)}=(\hat{\theta}_{j1}^{(0)},\ldots,\hat{\theta}_{jL}^{(0)})^{\mathrm{T}}$ 为分位水平 τ_k 下的 QR 估计量 (3.3). 本文使用类 BIC 准则选择调节参数 λ_k, 即最小化 $\mathrm{BIC}(\lambda_k)=\log\{\sum_{i=1}^{n}\rho_{\tau_k}(Y_i-\hat{c}-\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(\lambda_k)}-\boldsymbol{\psi}_i^{\mathrm{T}}\hat{\Theta}_{(\lambda_k)})\}+(\log n/n)df_{\lambda_k}$, 其中 df_{λ_k} 为 λ_k 下非零参数的个数. 据此, 本文给出部分线性可加模型的稳健变量选择算法 PMWQR-PLAM (见算法 2). 为探究估计量的渐近性质, 令 $\mathcal{A}=\{1\leqslant j\leqslant d:\boldsymbol{\beta}_j\neq 0\}$ 为参数部分中非零系数的指标集, $q_1=|\mathcal{A}|$ 为 $\mathcal{A}$ 的基. 不失一般性, 本文假定 $\boldsymbol{\beta}$ 的前 q_1 个元素非零, 余下的 $d-q_1$ 个元素为 0. 进一步令 $\boldsymbol{X}_{\mathcal{A}}$ 为矩阵 $\boldsymbol{X}$ 的前 q_1 列构成的子矩阵. 同样地, 假设非参数部分中前 q_2 个成分非零, 并假设 q_2 为一个固定的常数, 并施加如下正则条件:

条件 3.5 对于 $\delta_3<1/3$, 有 $q=O(n^{\delta_3})$.

条件 3.6 存在正常数 $\delta_\beta>0$ 和 $\delta_g>0$, 使得 $\min_{1\leqslant j\leqslant q_1}|\beta_1|\geqslant\delta_\beta$, $\min_{1\leqslant j\leqslant q_2}\|g_j\|^2\geqslant\delta_g$.

于是可证实 PMWQR 估计量具有神谕 (oracle) 性质, 见如下定理 3.2, 证明见附录 A.

算法 2 PMWQR-PLAM

输入: 样本 $\{\boldsymbol{x}_i,\boldsymbol{z}_i,y_i,i=1,2,\ldots,n\}$, K 个分位水平 $\{\tau_1,\ldots,\tau_K\}$ 初值 $k_n^{(0)}$ 和 $\lambda_n^{(0)}$.

输出: 参数估计值 $\hat{\boldsymbol{\beta}}_{\mathrm{PMWQR}}$ 和 $\hat{\Theta}_{\mathrm{PMWQR}}$ 及非参数函数估计值: $\hat{g}_j(z)$ $(j=1,\ldots,p)$.

1: **for** $i=1,\ldots,n$; **do**

2: **for** $m=1,\ldots,M$; **do**

3: 计算: $\tau_{x_i}^{(m)}=\arg\min_\tau\frac{\hat{Q}^{\lambda_n^{(m-1)}}_{(\boldsymbol{x}_i,\psi_i)}(\tau+\min\{h,\tau_{\max}-\tau\})-\hat{Q}^{\lambda_n^{(m-1)}}_{(\boldsymbol{x}_i,\psi_i)}(\tau-\min\{h,\tau-\tau_{\min}\})}{\min\{h,\tau_{\max}-\tau\}+\min\{h,\tau-\tau_{\min}\}}$,

4:

$$(\hat{\boldsymbol{\beta}}_{\tau_{\boldsymbol{x}_i}^{(m)}},\widehat{\boldsymbol{\theta}}_{\tau_{\boldsymbol{x}_i}^{(m)}},\hat{c}_{\tau_{\boldsymbol{x}_i}^{(m)}})=\arg\min_{\boldsymbol{\beta},\boldsymbol{\theta},c}\sum_{i=1}^{n}\rho_{\tau_{\boldsymbol{x}_i}^{(m)}}(Y_i-c-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-\psi_i^{\mathrm{T}}\theta)+\lambda_n^{(m-1)}\sum_{j=1}^{d}\frac{|\boldsymbol{\beta}_{j,k}|}{|\hat{\boldsymbol{\beta}}_{j,k}^{(0)}|^{1/2}}+\lambda_n^{(m-1)}\sum_{l=1}^{p}\frac{\|\theta_{l,k}\|_{H_l}}{\|\widehat{\theta}_{j,k}^{(0)}\|_{H_l}^{1/2}},$$

5: $k_n^{(m)}=\arg\min_k\log\{\sum_{i=1}^{n}\rho_{\tau_{\boldsymbol{x}_i}^{(m)}}(Y_i-\hat{c}_{\tau_{\boldsymbol{x}_i}^{(m)},p}-\boldsymbol{x}_i^{\tau}\hat{\boldsymbol{\beta}}_{\tau_{\boldsymbol{x}_i}^{(m)},p}-\psi_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{\tau_{\boldsymbol{x}_i}^{(m)},p})\}+(p\log n\cdot(k+4))/(2n)$,

6: $\lambda_n^{(m)}=\arg\min_\lambda\log\{\sum_{i=1}^{n}\rho_{\tau_{\boldsymbol{x}_i}^{(m)}}(Y_i-\hat{c}_{\tau_{\boldsymbol{x}_i}^{(m)},p}-\boldsymbol{x}_i^{\tau}\hat{\boldsymbol{\beta}}_{\tau_{\boldsymbol{x}_i}^{(m)},p}-\psi_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{\tau_{\boldsymbol{x}_i}^{(m)},p})\}+\log ndf_\lambda/n$.

7: **end for**

8: $k_n^{(m)}=k_n^{(m+1)}$, $\lambda_n^{(m)}=\lambda_n^{(m+1)}$ 或 $m=M$.

9: 令: $\hat{\boldsymbol{\beta}}_{\tau_{\boldsymbol{x}_i},p}=\hat{\boldsymbol{\beta}}^{(m)}_{\tau_{\boldsymbol{x}_i},p}$, $\hat{\boldsymbol{\theta}}_{\tau_{\boldsymbol{x}_i},p}=\hat{\boldsymbol{\theta}}^{(m)}_{\tau_{\boldsymbol{x}_i},p}$.

10: **end for**

11: $i=n$.

12: 令 $\hat{\boldsymbol{\beta}}_{\mathrm{PMWQR}}=\sum_{i=1}^{n}w_i^*\hat{\boldsymbol{\beta}}_{\tau_{x_i},p}$, $\hat{\boldsymbol{\theta}}_{\mathrm{PMWQR}}=\sum_{i=1}^{n}w_i^*\hat{\boldsymbol{\theta}}_{\tau_{x_i},p}$, $\hat{g}_j(z)=\sum_{u=1}^{L}\hat{\boldsymbol{\theta}}_{\mathrm{PMWQR},ju}\psi_{ju}(z)$, $j=1,\ldots,p$.

定理 3.2 若条件 3.1–3.6 成立, 假定 $\lim_{n\to\infty} n^{-1}L\log(L) = 0$. 假定非参数部分重要变量为 $Z_1,\ldots,Z_{q_2}$, 若惩罚参数 λ_n 满足 $\lambda_n \to 0$, 且有 $n^{r/(2r+1)}\lambda_n \to \infty$, 则

(1) (相合性) $\hat{\boldsymbol{\beta}}_{\mathrm{PMWQR},l} \xrightarrow{p} 0$ $(l = q_1+1,\ldots,d)$, $\hat{g}_j(u) \xrightarrow{p} 0$ $(j = q_2+1,\ldots,p)$;

(2) (有效性) $n^{-1}\sum_{i=1}^{n}(\hat{g}_j(u_i) - g_j(u_i))^2 = O_p(n^{-2r/(2r+1)})$.

4 数值模拟

本节通过 4 个模拟例子探究所提出 MWQR 方法在有限样本下的实际表现. 例 4.1 和 4.2 分别考察基于单峰分布和多峰分布的 MWQR-PLAM 算法在部分线性可加模型中的估计效果. 例 4.3 和 4.4 应用 PMWQR-PLAM 算法进一步考察其在变量选择中的有效性, 例 4.3 设定参数部分的协变量为高维情形, 例 4.4 设定参数与非参数部分的协变量均为高维情形. 此外, 本节对比了 LS、CQR、QR 和 WQR 这 4 种方法. 为公平起见, MWQR、CQR 和 WQR 方法中均采用均匀分位水平, 并设定分位水平个数 $K = 9$; QR 方法中考虑 0.05、0.50 和 0.75 这 3 个分位水平以全面探究其在不同分位水平的表现, 分别记为 $\mathrm{QR}_{0.05}$、$\mathrm{QR}_{0.50}$ 和 $\mathrm{QR}_{0.75}$. 4 个例子中分别考虑样本容量为 200 和 400 两种不同情形, 并通过 4 种不同的误差分布 $N(0,1)$、χ^2_4、t_3 和 $MN(0,0,1,10^2,0.9) = 0.9N(0,1)+0.1N(0,10^2)$, 比较各方法的优劣和有效性. 模拟研究均使用 R 软件, 并重复 200 次. 使用如下评价准则: (1) 参数 $\boldsymbol{\beta}$: 均方误差 (mean squared error, MSE), $\mathrm{MSE} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$, 其中 $\|\cdot\|_2$ 表示二范数; (2) 非参数可加函数 $g_j(\cdot)$: 平均平方误差 (average squared error, ASE) 准则, 即 $\mathrm{ASE}_j = \sum_{k=1}^{t}\{\hat{g}_j(u_k) - g_j(u_k)\}^2/t$, 其中 $\{u_k : k = 1,\ldots,t\}$ 为区间 $[0,1]$ 中的 t 个等分点, 并取 $t = 150$; (3) 变量选择: 指标 C、IC 和 CF, 其中 C 表示 200 次模拟下所有重要变量被正确选择的平均个数, IC 表示非重要变量被选择的平均个数, CF 表示 200 次模拟下模型被正确选择的概率. 参数 $\boldsymbol{\beta}$ 的估计精度采用几何均方误差 (geometric mean squared error, GMSE) 指标进行度量, 即 $\mathrm{GMSE} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$; 非参数部分仍用 ASE 准则; (4) 时间: 记录各算法平均运行 1 次的时间, 单位为分钟 (min).

例 4.1 考虑部分线性可加模型 $Y = 3X_1 - 1.5X_2 + g_1(Z_1) + g_2(Z_2) + (2 + X_1 - Z_2)^{-\gamma}\epsilon$, 其中, 参数部分协变量 X_1 和 X_2 来自零均值, 协方差矩阵为 $\boldsymbol{\Sigma}$ 的二维正态分布, 且 $\boldsymbol{\Sigma}$ 的对角元素为 1、非对角元素为 2/3. 可加部分中协变量 Z_1 与 Z_2 独立同分布于均匀分布 $U(0,1)$, 并有 $g_1(Z_1) = 3(2Z_1-1)^2 - E(3(2Z_1-1)^2)$ 和 $g_2(Z_2) = 2\sin(2\pi Z_2) - E(2\sin(2\pi Z_2))$. 本例考虑参数 $\gamma = 0$ 和 $\gamma = 1$ 分别对应于同方差及异方差情形, 模拟结果见表 2.

表 2 传递了如下信息: (1) 在同方差设定下, 根据 MSE 指标, 随着样本量的增加, MWQR 方法表现更加稳定, 估计误差显著下降, 特别当误差项有偏或存在厚尾分布时, 在参数估计中几乎取得最小偏误. 可见, 基于众数的 MWQR 方法由于综合了多个分位水平的有效信息, 不仅稳健, 而且还能适应于各类数据. MWQR 方法在非参数函数的估计中也有同样的表现. (2) 在异方差情形下, LS 方法及极端分位水平下的 QR 方法估计误差显著增大, 尤其 LS 方法几乎完全失效, 而 MWQR 方法表现依然稳定, 表明该方法对异质性数据下的估计问题具有良好的适应性. 值得指出的是, 虽然个别情形下 CQR 方法及 0.5 分位水平下的 QR 方法在 $g_1(\cdot)$ 的估计中表现略优, 但对 $g_2(\cdot)$ 的估计表现相对较差, 其原因在于, 与 $g_1(\cdot)$ 函数相比, $g_2(\cdot)$ 函数的结构更加复杂, 存在较多拐点, 而 MWQR 方法可以较好恢复非参数函数的高低起伏、崎岖不平的形状. 此外, 由于异方差使得 LS 方法估计的方差增大, 造成预测误差增大, 预测精度降低, 从而出现了在异方差设定下 LS 方法的估计误差随样本量增大反增不减的现象. (3) 从运算效率来看, MWQR、CQR 和 WQR 这 3 种方法由于综合了多个分位水平, 运行

表 2 例 4.1 单峰下的模拟结果

	ε	方法	$\gamma=0$			时间 (min)	$\gamma=1$			时间 (min)
			$\boldsymbol{\beta}$	$g_1(\cdot)$	$g_2(\cdot)$		$\boldsymbol{\beta}$	$g_1(\cdot)$	$g_2(\cdot)$	
			MSE (SD)	ASE_1 (SD)	ASE_2 (SD)		MSE (SD)	ASE_1 (SD)	ASE_2 (SD)	
200	$N(0,1)$	MWQR	0.018 (0.022)	4.045 (1.606)	0.215 (0.233)	0.018	0.014 (0.013)	3.967 (0.840)	0.181 (0.153)	0.017
		LS	0.020 (0.025)	4.097 (1.105)	0.246 (0.316)	0.001	13.213 (49.998)	47.463 (204.974)	26.937 (91.187)	0.001
		CQR	0.021 (0.024)	4.137 (1.121)	0.233 (0.291)	0.039	0.024 (0.031)	3.985 (1.050)	0.196 (0.156)	0.020
		$\text{QR}_{0.05}$	0.086 (0.105)	4.152 (1.084)	0.532 (0.563)	0.002	1.757 (1.746)	9.077 (13.758)	9.376 (45.385)	0.001
		$\text{QR}_{0.50}$	0.025 (0.030)	4.219 (1.329)	0.267 (0.337)	0.002	0.016 (0.017)	3.995 (0.958)	0.194 (0.165)	0.003
		$\text{QR}_{0.75}$	0.034 (0.041)	4.211 (1.466)	0.296 (0.351)	0.003	0.124 (0.087)	4.059 (1.136)	0.309 (0.280)	0.002
		WQR	0.020 (0.025)	4.115 (1.154)	0.231 (0.264)	0.051	0.015 (0.014)	3.983 (0.898)	0.182 (0.154)	0.018
	χ^2_4	MWQR	0.096 (0.109)	4.297 (2.522)	0.409 (0.491)	0.011	0.603 (0.401)	4.148 (2.271)	0.808 (0.988)	0.024
		LS	0.150 (0.189)	4.654 (3.095)	0.778 (1.055)	0.001	151.425 (837.732)	649.846 (4254.523)	288.168 (912.162)	0.001
		CQR	0.995 (1.323)	4.638 (2.705)	0.704 (0.839)	0.014	0.804 (0.714)	4.277 (3.486)	1.302 (1.193)	0.019
		$\text{QR}_{0.05}$	0.076 (0.098)	4.340 (2.532)	0.472 (0.787)	0.001	32.267 (80.756)	121.202 (471.117)	158.912 (447.562)	0.001
		$\text{QR}_{0.50}$	0.189 (0.209)	5.059 (3.891)	0.820 (0.891)	0.001	0.826 (0.599)	4.369 (2.678)	0.877 (0.757)	0.002
		$\text{QR}_{0.75}$	0.498 (0.574)	6.390 (5.772)	1.849 (2.304)	0.001	4.393 (2.101)	4.825 (4.179)	2.888 (2.657)	0.001
		WQR	0.168 (0.176)	4.689 (2.984)	0.643 (0.898)	0.040	1.048 (0.767)	4.418 (3.360)	1.108 (1.187)	0.017
	t_3	MWQR	0.034 (0.036)	3.925 (1.437)	0.286 (0.342)	0.010	0.023 (0.032)	3.900 (1.058)	0.210 (0.197)	0.012
		LS	0.048 (0.061)	4.070 (1.436)	0.351 (0.393)	0.001	12.237 (41.625)	31.338 (67.365)	30.566 (107.664)	0.001
		CQR	0.034 (0.037)	3.941 (1.412)	0.294 (0.375)	0.025	0.031 (0.036)	4.045 (1.526)	0.229 (0.231)	0.015
		$\text{QR}_{0.05}$	0.319 (0.390)	5.886 (4.551)	1.964 (2.997)	0.001	2.645 (2.333)	13.338 (23.447)	19.782 (75.566)	0.001
		$\text{QR}_{0.50}$	0.043 (0.045)	3.918 (1.473)	0.320 (0.400)	0.001	0.024 (0.032)	3.903 (1.120)	0.212 (0.213)	0.002
		$\text{QR}_{0.75}$	0.062 (0.079)	4.024 (1.790)	0.324 (0.342)	0.001	0.157 (0.140)	3.938 (1.531)	0.372 (0.337)	0.002
		WQR	0.034 (0.039)	3.942 (1.476)	0.300 (0.400)	0.044	0.033 (0.037)	3.919 (1.303)	0.231 (0.211)	0.013
	$MN(0,0,1,10^2,0.9)$	MWQR	0.037 (0.049)	4.018 (1.413)	0.283 (0.323)	0.009	0.027 (0.032)	4.035 (1.387)	0.197 (0.178)	0.010
		LS	0.040 (0.050)	4.055 (1.351)	0.303 (0.327)	0.001	378.426 (2999.452)	1234.161 (9309.503)	2460.291 (22026.830)	0.001
		CQR	0.037 (0.048)	4.024 (1.355)	0.294 (0.302)	0.021	0.048 (0.055)	4.092 (1.393)	0.251 (0.243)	0.015
		$\text{QR}_{0.05}$	0.149 (0.164)	4.616 (3.279)	0.919 (1.362)	0.001	2.644 (2.918)	44.314 (281.985)	23.433 (99.323)	0.001
		$\text{QR}_{0.50}$	0.058 (0.068)	4.117 (1.856)	0.366 (0.410)	0.001	0.029 (0.039)	4.093 (1.418)	0.239 (0.232)	0.001
		$\text{QR}_{0.75}$	0.056 (0.074)	3.990 (2.064)	0.365 (0.400)	0.001	0.200 (0.145)	4.077 (1.560)	0.387 (0.355)	0.001
		WQR	0.038 (0.049)	4.052 (1.404)	0.294 (0.323)	0.051	0.031 (0.036)	4.112 (1.585)	0.215 (0.193)	0.010

续表

	ε	方法	$\gamma=0$			时间 (min)	$\gamma=1$			时间 (min)
			$\boldsymbol{\beta}$	$g_1(\cdot)$	$g_2(\cdot)$		$\boldsymbol{\beta}$	$g_1(\cdot)$	$g_2(\cdot)$	
			MSE (SD)	ASE_1 (SD)	ASE_2 (SD)		MSE (SD)	ASE_1 (SD)	ASE_2 (SD)	
400	$N(0,1)$	MWQR	0.012 (0.014)	3.714 (1.158)	0.183 (0.134)	0.026	0.009 (0.012)	3.844 (0.623)	0.178 (0.117)	0.022
		LS	0.014 (0.012)	3.745 (0.836)	0.194 (0.156)	0.001	520.162 (5085.984)	507.218 (4801.315)	1170.925 (11419.720)	0.001
		CQR	0.012 (0.014)	3.757 (0.869)	0.190 (0.146)	0.088	0.020 (0.028)	3.931 (1.003)	0.193 (0.152)	0.078
		$\text{QR}_{0.05}$	0.037 (0.040)	3.864 (0.928)	0.354 (0.418)	0.003	1.277 (0.884)	6.604 (16.043)	2.218 (4.977)	0.001
		$\text{QR}_{0.50}$	0.017 (0.020)	3.750 (1.064)	0.218 (0.188)	0.003	0.011 (0.012)	3.862 (0.675)	0.187 (0.130)	0.002
		$\text{QR}_{0.75}$	0.022 (0.026)	3.904 (1.062)	0.204 (0.158)	0.003	0.095 (0.051)	3.785 (0.712)	0.295 (0.174)	0.002
		WQR	0.012 (0.012)	3.730 (0.856)	0.186 (0.139)	0.055	0.009 (0.009)	3.856 (0.580)	0.181 (0.111)	0.032
	χ^2_4	MWQR	0.031 (0.034)	4.067 (1.891)	0.342 (0.344)	0.029	0.503 (0.284)	4.102 (2.136)	0.671 (0.496)	0.026
		LS	0.087 (0.089)	4.090 (2.356)	0.478 (0.545)	0.001	427.842 (2860.833)	2191.590 (14058.300)	1792.731 (13058.700)	0.002
		CQR	0.392 (0.506)	4.127 (2.102)	0.400 (0.458)	0.046	0.680 (0.544)	3.988 (1.891)	0.893 (0.679)	0.046
		$\text{QR}_{0.05}$	0.040 (0.054)	4.140 (1.412)	0.345 (0.309)	0.002	16.603 (117.474)	29.797 (108.079)	33.459 (57.955)	0.003
		$\text{QR}_{0.50}$	0.090 (0.091)	4.128 (2.514)	0.540 (0.565)	0.003	0.785 (0.367)	4.260 (2.373)	0.857 (0.625)	0.003
		$\text{QR}_{0.75}$	0.213 (0.223)	4.708 (3.859)	1.218 (1.625)	0.003	4.268 (1.522)	4.512 (3.348)	2.581 (1.886)	0.002
		WQR	0.076 (0.066)	4.146 (2.333)	0.442 (0.497)	0.044	0.984 (0.526)	4.395 (2.359)	1.048 (0.817)	0.023
	t_3	MWQR	0.012 (0.015)	3.840 (1.004)	0.171 (0.187)	0.024	0.008 (0.012)	3.880 (0.711)	0.177 (0.111)	0.016
		LS	0.025 (0.020)	3.867 (1.776)	0.209 (0.210)	0.001	124.383 (771.420)	600.047 (3727.758)	66.914 (337.621)	0.001
		CQR	0.013 (0.016)	3.826 (0.987)	0.181 (0.219)	0.055	0.021 (0.022)	3.798 (1.058)	0.188 (0.126)	0.067
		$\text{QR}_{0.05}$	0.195 (0.220)	5.212 (3.515)	1.237 (1.894)	0.002	2.620 (1.893)	8.908 (24.839)	4.140 (9.469)	0.002
		$\text{QR}_{0.50}$	0.016 (0.021)	3.860 (1.044)	0.195 (0.205)	0.001	0.014 (3.905)	3.900 (0.733)	0.184 (0.114)	0.002
		$\text{QR}_{0.75}$	0.021 (0.022)	3.941 (1.198)	0.194 (0.207)	0.002	0.115 (0.063)	3.914 (0.927)	0.323 (0.207)	0.002
		WQR	0.013 (0.016)	3.858 (1.057)	0.176 (0.180)	0.044	0.010 (0.013)	3.911 (0.761)	0.202 (0.134)	0.019
	$MN(0,0,1,10^2,0.9)$	MWQR	0.015 (0.018)	3.822 (1.040)	0.195 (0.195)	0.012	0.014 (0.013)	3.951 (0.856)	0.186 (0.132)	0.012
		LS	0.018 (0.024)	3.857 (1.232)	0.214 (0.213)	0.001	16.168 (61.536)	33.860 (100.788)	39.887 (197.939)	0.001
		CQR	0.016 (0.019)	3.848 (1.164)	0.203 (0.209)	0.049	0.019 (0.020)	3.843 (0.931)	0.220 (0.214)	0.043
		$\text{QR}_{0.05}$	0.085 (0.105)	4.122 (2.233)	0.618 (0.782)	0.002	2.340 (2.276)	6.571 (6.104)	4.423 (13.282)	0.001
		$\text{QR}_{0.50}$	0.027 (0.035)	3.871 (1.416)	0.257 (0.281)	0.002	0.015 (0.013)	4.020 (0.954)	0.197 (0.157)	0.001
		$\text{QR}_{0.75}$	0.033 (0.020)	3.940 (1.582)	0.267 (0.308)	0.001	0.167 (0.089)	4.064 (1.092)	0.383 (0.246)	0.001
		WQR	0.016 (0.021)	3.856 (1.156)	0.203 (0.207)	0.051	0.015 (0.017)	3.984 (0.869)	0.188 (0.144)	0.012

速率相当. 在异方差情形下, MWQR 方法显著优于 CQR 方法. 图 1 分别绘制了同方差标准正态分布和异方差 t_3 分布误差下非参数函数在不同方法下的拟合结果. 不难发现, 在同方差标准正态误差分布下各方法表现相对稳定, 但极端分位水平 0.05 下及中高部分位水平 0.75 下的 QR 方法仍具有较大误差. 此外, 异方差 t_3 分布误差下 LS 方法及 0.05 分位水平下的 QR 方法几乎完全失效. 相较于其他方法, MWQR 方法能更好地恢复非参数函数的真实形状, 特别在拐点和边界尾部处有更好的拟合效果. 以上进一步说明, MWQR 方法有助于探究异质性数据之间的变量关系, 挖掘数据的内蕴结构.

例 4.2 依然考虑例 4.1 中的部分线性可加模型, 误差项设定为两种多峰分布及两种在众数附近峰值平坦的分布: (1) 双峰: $MN(0,4,1,1,0.5) = 0.5N(0,1) + 0.5N(4,1)$; (2) 三峰: $0.3N(4,1) + 0.3N(8,1) + 0.4t_3$, 简记为 $MN + t_3$; (3) 第一种峰值平坦分布: $MN(0,10.2,5,5,0.5) = 0.5N(0,5^2) + 0.5N(10.2,5^2)$; (4) 第二种峰值平坦分布: $0.475N(1.85,1) + 0.525t_3$, 简记为 $N + t_3$. 4 种误差分布图形见图 2. 其余设定同例 4.1, 模拟结果见表 3. 表 3 中的指标 MSE 及 ASE 结果表明, 一方面, 随样本量的增加, MWQR 估计方法的偏差及标准差相应减小, 与定理 3.1 中相合性的结论一致; 另一方面, 当误差分布为多峰分布且峰值明显时, 无论是对参数还是非参数函数的估计, MWQR 方法更有优势,

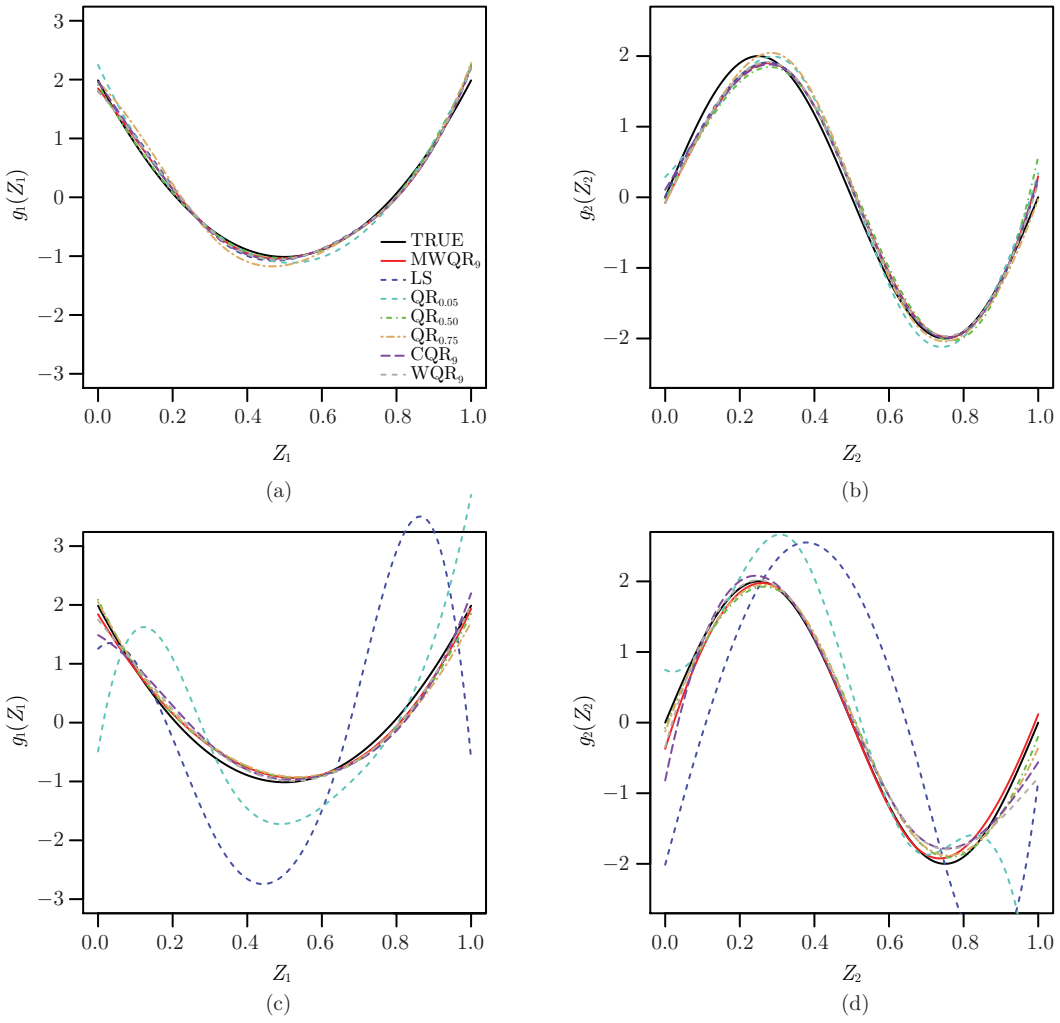

图 1 (网络版彩图) 例 4.1 的拟合曲线. (a) 和 (b) 同方差标准正态分布误差; (c) 和 (d) 异方差 t_3 分布误差

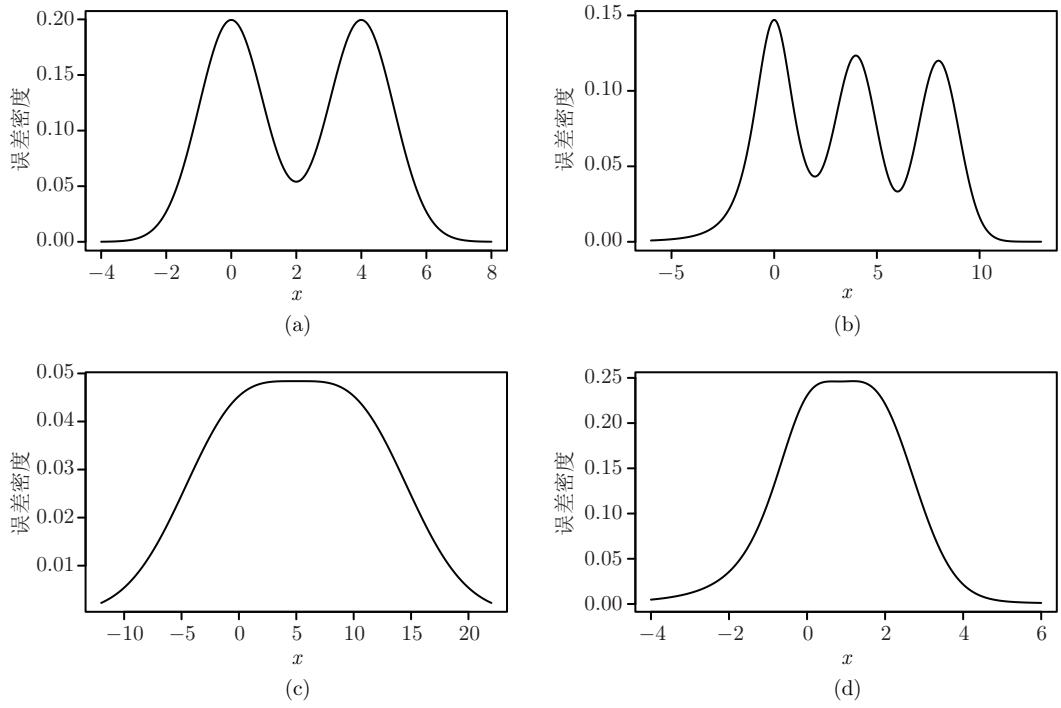

图 2 (a)–(d) 分别对应于例 4.2(1)–4.2(4) 设定的 4 种误差分布情形

显著优于其他估计方法. 这表明当条件分布的众数不唯一时, MWQR 方法能够通过比较不同的局部众数, 识别出全局众数, 进而获取到反映条件分布最大信息量的最优分位水平, 有效提升估计效果. 而当误差分布服从或近似正态分布且在峰值附近比较平坦时, LS 方法最具优势, 这也说明 MWQR 方法更适用于峰值比较明显的情形.

例 4.3 考虑部分线性可加模型 $Y_i = X_i^{\mathrm{T}}\boldsymbol{\beta}+g(Z_i)+\epsilon_i$, 其中, 协变量 $\boldsymbol{X}$ 的维数设定为 $d = 200$, 参数向量为 $\boldsymbol{\beta} = (1.5, -2, 3, \mathbf{0}_{d-3})^{\mathrm{T}}$, 可加函数设定如下: $g(Z) = 2\sin(\pi Z) - E(2\sin(\pi Z))$. $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\sigma_{ij})_{d\times d}$, $\sigma_{ij} = 0.5^{|i-j|}$; Z 来自均匀分布 $U(0,1)$. 为进行对比分析, 所有估计方法均采用自适应 LASSO 惩罚进行变量选择, 分别记为 PMWQR、PCQR (penalized CQR)、PQR (penalized QR) 和 PLS (penalized LS). 模拟结果如表 4 所示. 由表 4 可推导出如下结论: (1) 在变量选择方面, 不论误差分布的形式如何, PMWQR 方法均能准确选择重要变量; 而诸如 QR、CQR 和 LS 等方法, 当误差项有偏或服从厚尾分布时, 会遗漏重要变量, 正确选择概率较低. (2) 由 GMSE 和 ASE 指标, 随着样本量的增大, PMWQR 方法在参数估计和非参数函数估计中的准确性不断提高, 且对非参数部分的估计在所有方法中始终保持最优. (3) 在运算时间方面, PMWQR 方法表现与例 4.1 结果相似, 其与 PWQR 方法表现相当, 略优于 PCQR 方法.

例 4.4 考虑部分线性可加模型 $Y_i = X_i^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{p} g_j(Z_{ij}) + \epsilon_i$, 其中, 协变量 $\boldsymbol{X}$ 和 $\boldsymbol{Z}$ 的维数设定为 $p = d = 50$, 参数向量设定为 $\boldsymbol{\beta} = (3, 1.5, 2, 3, \mathbf{0}_{d-4})^{\mathrm{T}}$. 可加函数设定如下: $g_1(Z_1) = (2Z_1+1)^3 - E((2Z_1+1)^3)$, $g_2(Z_2) = 5Z_2 - E(5Z_2)$, $g_2(Z_2) = 5Z_2 - E(5Z_2)$, $g_3(Z_3) = \exp(2Z_3-1) - E(\exp(2Z_3-1))$, 其余 $g_j(Z_j) = 0$. 协变量 $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\sigma_{ij})_{d\times d}$, $\sigma_{ij} = 0.5^{|i-j|}$; 协变量 $Z_1, \ldots, Z_{50}$ 独立同分布于 $U(0,1)$. 为进行对比分析, 所有估计方法均采用自适应 LASSO 惩罚进行变量选择, 分别记为 PMWQR、PCQR、PQR 和 PLS. 模拟结果如表 5 所示. 由表 5 可推导出如下结论:

(1) PMWQR 方法在所有情形下均能准确选择参数部分的所有重要变量, 且不会引入无关变量,

表 3 例 4.2 误差项多峰分布下的估计结果

n	ε	方法	$\boldsymbol{\beta}$	$g_1(\cdot)$	$g_2(\cdot)$	时间 (min)
			MSE (SD)	ASE_1 (SD)	ASE_2 (SD)	
200	$MN(0,4,1,1,0.5)$	MWQR	0.060 (0.078)	4.237 (2.236)	0.450 (0.668)	0.015
		LS	0.078 (0.090)	4.405 (2.616)	0.581 (0.778)	0.001
		CQR	0.095 (0.113)	4.375 (2.486)	0.518 (0.679)	0.019
		$\text{QR}_{0.05}$	0.137 (0.150)	5.039 (3.656)	0.699 (0.925)	0.001
		$\text{QR}_{0.50}$	0.308 (0.293)	5.359 (4.904)	1.720 (1.625)	0.001
		$\text{QR}_{0.75}$	0.092 (0.110)	4.327 (2.500)	0.819 (1.534)	0.001
		WQR	0.086 (0.090)	4.491 (2.733)	0.637 (0.853)	0.010
	$MN+t_3$	MWQR	0.198 (0.236)	4.483 (3.544)	0.986 (1.081)	0.017
		LS	0.271 (0.341)	4.671 (4.114)	1.202 (1.461)	0.001
		CQR	0.248 (0.320)	5.382 (4.252)	1.266 (1.369)	0.019
		$\text{QR}_{0.05}$	0.468 (0.558)	5.952 (6.537)	2.890 (5.413)	0.002
		$\text{QR}_{0.50}$	0.662 (0.769)	6.140 (6.956)	2.777 (3.538)	0.001
		$\text{QR}_{0.75}$	0.582 (0.706)	5.947 (5.268)	3.107 (4.989)	0.001
		WQR	0.292 (0.327)	4.776 (4.263)	1.380 (1.673)	0.014
	$MN(0,10.5,5,5,0.5)$	MWQR	1.084 (1.197)	9.201 (8.660)	5.737 (6.441)	0.017
		LS	0.961 (1.101)	8.781 (8.817)	4.893 (5.565)	0.001
		CQR	1.209 (1.496)	9.806 (9.740)	6.318 (7.591)	0.023
		$\text{QR}_{0.05}$	2.813 (3.195)	22.700 (26.753)	12.615 (17.385)	0.001
		$\text{QR}_{0.50}$	1.974 (2.112)	12.296 (14.217)	10.054 (11.868)	0.001
		$\text{QR}_{0.75}$	1.889 (2.126)	12.076 (13.430)	9.375 (11.281)	0.001
		WQR	1.173 (1.278)	9.566 (9.506)	5.977 (6.839)	0.011
	$N+t_3$	MWQR	0.048 (0.053)	4.060 (1.554)	0.316 (0.446)	0.015
		LS	0.049 (0.052)	4.099 (1.522)	0.308 (0.396)	0.001
		CQR	0.051 (0.059)	4.075 (1.515)	0.332 (0.356)	0.021
		$\text{QR}_{0.05}$	0.210 (0.287)	4.835 (3.565)	1.353 (1.983)	0.001
		$\text{QR}_{0.50}$	0.081 (0.093)	4.125 (1.960)	0.425 (0.641)	0.001
		$\text{QR}_{0.75}$	0.064 (0.072)	4.187 (1.688)	0.383 (0.527)	0.001
		WQR	0.054 (0.063)	4.078 (1.524)	0.330 (0.475)	0.010
400	$MN(0,4,1,1,0.5)$	MWQR	0.026 (0.027)	3.990 (1.286)	0.306 (0.321)	0.022
		LS	0.043 (0.049)	4.059 (1.575)	0.413 (0.471)	0.001
		CQR	0.040 (0.052)	4.368 (1.633)	0.423 (0.462)	0.029
		$\text{QR}_{0.05}$	0.066 (0.082)	4.107 (2.039)	0.413 (0.503)	0.003
		$\text{QR}_{0.50}$	0.225 (0.281)	5.316 (2.875)	1.450 (1.517)	0.001
		$\text{QR}_{0.75}$	0.050 (0.055)	4.071 (1.578)	0.452 (0.707)	0.002
		WQR	0.048 (0.054)	4.151 (1.685)	0.447 (0.541)	0.021
	$MN+t_3$	MWQR	0.084 (0.118)	4.341 (2.295)	0.541 (0.572)	0.023
		LS	0.118 (0.146)	4.415 (2.701)	0.780 (0.803)	0.001
		CQR	0.110 (0.137)	4.472 (2.714)	0.843 (0.985)	0.031
		$\text{QR}_{0.05}$	0.171 (0.195)	5.183 (3.412)	1.071 (1.655)	0.001
		$\text{QR}_{0.50}$	0.276 (0.400)	5.382 (4.403)	1.410 (1.750)	0.002
		$\text{QR}_{0.75}$	0.273 (0.295)	4.740 (4.045)	1.900 (2.066)	0.002
		WQR	0.129 (0.156)	4.513 (2.941)	0.844 (0.911)	0.019
	$MN(0,10.5,5,5,0.5)$	MWQR	0.508 (0.579)	6.370 (7.104)	2.429 (2,913)	0.020
		LS	0.425 (0.495)	6.019 (6.472)	2.103 (2.555)	0.001
		CQR	0.657 (0.665)	6.205 (6.942)	2.764 (3.205)	0.025
		$\text{QR}_{0.05}$	1.430 (1.494)	10.473 (12.803)	7.692 (11.637)	0.002
		$\text{QR}_{0.50}$	0.938 (0.876)	9.628 (11.833)	4.267 (5.433)	0.002
		$\text{QR}_{0.75}$	0.746 (0.804)	7.771 (9.393)	4.946 (5.947)	0.002
		WQR	0.598 (0.631)	6.539 (7.445)	2.759 (3.260)	0.015
	$N+t_3$	MWQR	0.023 (0.026)	3.749 (1.271)	0.292 (0.277)	0.019
		LS	0.028 (0.033)	3.929 (1.340)	0.295 (0.275)	0.001
		CQR	0.029 (0.034)	3.912 (1.188)	0.282 (0.266)	0.025
		$\text{QR}_{0.05}$	0.127 (0.186)	4.436 (2.440)	0.656 (0.764)	0.001
		$\text{QR}_{0.50}$	0.042 (0.043)	3.972 (1.494)	0.320 (0.306)	0.001
		$\text{QR}_{0.75}$	0.033 (0.042)	3.826 (1.446)	0.402 (0.468)	0.001
		WQR	0.027 (0.031)	3.891 (1.193)	0.283 (0.262)	0.013

表 4　例 4.3 估计及变量选择结果

	误差项	方法	参数部分				非参数部分			时间 (min)
			C	IC	CF	GMSE	SD	ASE	SD	
200	$N(0,1)$	PMWQR	3.000	0.000	1.00	0.070	0.048	0.074	0.083	0.352
		PLS	3.000	97.870	0.00	0.285	0.077	0.095	0.172	0.138
		PCQR	3.000	0.030	0.96	0.052	0.036	0.086	0.080	0.617
		$\text{PQR}_{0.05}$	2.565	0.000	0.61	1.780	0.842	1.203	0.678	0.141
		$\text{PQR}_{0.50}$	3.000	0.000	1.00	0.065	0.044	0.077	0.064	0.143
		$\text{PQR}_{0.75}$	3.000	0.000	1.00	0.093	0.063	0.166	0.188	0.135
		PWQR	3.000	0.000	1.00	0.087	0.049	0.092	0.108	0.337
	t_3	PMWQR	3.000	0.000	1.00	0.126	0.078	0.149	0.142	0.320
		PLS	3.000	82.180	0.00	0.774	0.384	0.296	0.463	0.127
		PCQR	3.000	0.000	1.00	0.075	0.062	0.190	0.186	0.580
		$\text{PQR}_{0.05}$	1.835	0.000	0.19	2.720	0.881	1.709	0.574	0.152
		$\text{PQR}_{0.50}$	3.000	0.010	0.99	0.083	0.055	0.208	0.302	0.124
		$\text{PQR}_{0.75}$	3.000	0.010	0.99	0.154	0.107	0.296	0.428	0.135
		PWQR	3.000	0.000	1.00	0.161	0.185	0.183	0.235	0.357
	χ^2_4	PMWQR	2.940	0.000	0.94	0.510	0.451	0.516	0.619	0.344
		PLS	3.000	67.515	0.00	2.116	0.500	1.437	1.238	0.105
		PCQR	3.000	0.515	0.57	0.541	0.454	0.623	0.882	0.680
		$\text{PQR}_{0.05}$	2.130	0.000	0.27	2.339	0.760	1.574	0.705	0.161
		$\text{PQR}_{0.50}$	2.990	0.645	0.52	0.593	0.384	0.527	0.533	0.148
		$\text{PQR}_{0.75}$	2.880	0.775	0.40	1.177	0.728	1.034	0.927	0.128
		PWQR	2.880	0.000	0.88	0.688	0.500	0.590	0.524	0.317
	$MN(0,0,1,10^2,0.9)$	PMWQR	3.000	0.000	1.00	0.160	0.094	0.209	0.257	0.312
		PLS	3.000	91.695	0.00	0.508	0.090	0.282	0.378	0.191
		PCQR	3.000	0.020	0.98	0.121	0.086	0.247	0.334	0.598
		$\text{PQR}_{0.05}$	2.255	0.000	0.39	2.292	0.785	1.644	0.638	0.140
		$\text{PQR}_{0.50}$	3.000	0.000	1.00	0.128	0.082	0.224	0.249	0.126
		$\text{PQR}_{0.75}$	3.000	0.010	0.99	0.163	0.120	0.238	0.320	0.127
		PWQR	3.000	0.000	1.00	0.177	0.099	0.266	0.326	0.348
400	$N(0,1)$	PMWQR	3.000	0.000	1.00	0.046	0.026	0.049	0.065	0.513
		PLS	3.000	57.820	0.00	0.187	0.036	0.077	0.084	0.145
		PCQR	3.000	0.000	1.00	0.048	0.033	0.070	0.090	0.679
		$\text{PQR}_{0.05}$	2.970	0.000	0.97	0.774	0.463	0.706	0.629	0.243
		$\text{PQR}_{0.50}$	3.000	0.000	1.00	0.054	0.031	0.054	0.047	0.197
		$\text{PQR}_{0.75}$	3.000	0.000	1.00	0.056	0.036	0.079	0.100	0.253
		PWQR	3.000	0.000	1.00	0.047	0.030	0.059	0.075	0.611
	t_3	PMWQR	3.000	0.000	1.00	0.085	0.061	0.116	0.144	0.529
		PLS	3.000	57.060	0.00	0.516	0.186	0.212	0.251	0.121
		PCQR	3.000	0.000	1.00	0.067	0.054	0.128	0.210	0.599
		$\text{PQR}_{0.05}$	2.365	0.000	0.51	2.125	0.856	1.275	0.782	0.218
		$\text{PQR}_{0.50}$	3.000	0.000	1.00	0.054	0.041	0.162	0.195	0.202
		$\text{PQR}_{0.75}$	3.000	0.000	1.00	0.091	0.063	0.166	0.274	0.240
		PWQR	3.000	0.000	1.00	0.102	0.055	0.124	0.152	0.617
	χ^2_4	PMWQR	3.000	0.000	1.00	0.315	0.166	0.395	0.482	0.684
		PLS	3.000	38.120	0.00	1.431	0.420	0.751	0.732	0.106
		PCQR	3.000	0.275	0.75	0.255	0.155	0.555	0.646	0.718
		$\text{PQR}_{0.05}$	2.820	0.000	0.84	1.311	0.776	1.082	0.840	0.218
		$\text{PQR}_{0.50}$	3.000	0.305	0.72	0.350	0.301	0.327	0.456	0.176
		$\text{PQR}_{0.75}$	2.910	0.605	0.45	0.867	0.705	0.750	0.641	0.249
		PWQR	2.990	0.000	0.99	0.391	0.249	0.412	0.475	0.549
	$MN(0,0,1,10^2,0.9)$	PMWQR	3.000	0.000	1.00	0.098	0.053	0.085	0.134	0.534
		PLS	3.000	59.425	0.00	0.361	0.086	0.120	0.131	0.133
		PCQR	3.000	0.000	1.00	0.078	0.051	0.106	0.203	0.622
		$\text{PQR}_{0.05}$	2.885	0.000	0.90	1.286	0.703	1.059	0.763	0.295
		$\text{PQR}_{0.50}$	3.000	0.000	1.00	0.105	0.088	0.163	0.159	0.216
		$\text{PQR}_{0.75}$	3.000	0.000	1.00	0.102	0.069	0.138	0.152	0.222
		PWQR	3.000	0.000	1.00	0.106	0.070	0.103	0.146	0.555

表 5 例 4.4 模拟结果

	误差项	方法	参数部分					非参数部分					时间 (min)
			C	IC	CF	GMSE	SD	C	IC	CF	ASE	SD	
200	$N(0,1)$	PMWQR	4.000	0.000	1.00	0.033	0.023	3.000	0.267	0.76	1.833	0.807	0.145
		PLS	4.000	0.000	1.00	0.036	0.025	3.000	0.244	0.81	1.920	0.861	0.222
		PCQR	4.000	0.000	1.00	0.177	0.067	3.000	8.188	0.00	1.962	0.927	0.513
		$\mathrm{PQR}_{0.05}$	4.000	0.000	1.00	0.200	0.168	2.866	0.144	0.77	3.368	1.723	0.021
		$\mathrm{PQR}_{0.50}$	4.000	0.020	0.99	0.034	0.025	3.000	4.044	0.00	2.043	0.914	0.035
		$\mathrm{PQR}_{0.75}$	4.000	0.020	0.99	0.048	0.031	3.000	2.944	0.03	2.063	1.007	0.043
		PWQR	4.000	0.000	1.00	0.035	0.020	3.000	0.411	0.66	1.879	0.652	0.225
	t_3	PMWQR	4.000	0.000	1.00	0.054	0.035	3.000	0.580	0.61	2.072	0.871	0.185
		PLS	4.000	0.010	0.99	0.147	0.165	2.970	0.900	0.39	2.746	1.309	0.267
		PCQR	4.000	0.000	1.00	0.071	0.068	3.000	9.980	0.00	2.103	1.035	0.504
		$\mathrm{PQR}_{0.05}$	3.935	0.065	0.94	0.963	0.700	2.478	1.135	0.15	6.798	3.858	0.031
		$\mathrm{PQR}_{0.50}$	4.000	0.065	0.98	0.051	0.044	3.000	5.200	0.00	2.010	0.988	0.026
		$\mathrm{PQR}_{0.75}$	4.000	0.000	1.00	0.078	0.060	3.000	4.780	0.01	2.637	1.399	0.026
		PWQR	4.000	0.000	1.00	0.057	0.046	3.000	0.700	0.50	2.089	1.083	0.276
	χ^2_4	PMWQR	4.000	0.000	1.00	0.186	0.124	2.790	0.145	0.28	2.936	1.430	0.241
		PLS	3.970	0.030	0.96	0.491	0.367	2.680	0.820	0.30	4.155	1.928	0.360
		PCQR	3.990	0.880	0.72	0.622	0.513	2.940	12.160	0.00	3.305	2.094	0.714
		$\mathrm{PQR}_{0.05}$	3.990	0.000	0.99	0.269	0.283	2.860	0.410	0.56	4.099	2.664	0.024
		$\mathrm{PQR}_{0.50}$	4.000	0.730	0.74	0.246	0.256	2.960	9.520	0.00	4.418	2.904	0.020
		$\mathrm{PQR}_{0.75}$	3.990	1.080	0.64	0.583	0.447	2.920	9.260	0.00	6.831	5.342	0.027
		PWQR	4.000	0.000	1.00	0.230	0.168	2.690	1.555	0.12	3.112	1.516	0.373
	$MN(0,0,1,10^2,0.9)$	PMWQR	4.000	0.000	1.00	0.059	0.048	3.000	0.300	0.71	1.916	0.954	0.170
		PLS	4.000	0.000	1.00	0.072	0.054	3.000	0.395	0.69	2.242	0.930	0.232
		PCQR	4.000	0.010	0.99	0.075	0.069	3.000	9.845	0.00	2.060	1.013	0.586
		$\mathrm{PQR}_{0.05}$	3.995	0.000	0.99	0.395	0.298	2.750	0.395	0.46	4.520	2.218	0.025
		$\mathrm{PQR}_{0.50}$	4.000	0.100	0.95	0.068	0.042	3.000	5.885	0.00	2.438	1.169	0.026
		$\mathrm{PQR}_{0.75}$	4.000	0.060	0.97	0.089	0.063	3.000	4.960	0.00	2.507	1.254	0.037
		PWQR	4.000	0.000	1.00	0.064	0.041	2.995	0.370	0.69	2.176	0.885	0.415
400	$N(0,1)$	PMWQR	4.000	0.000	1.00	0.014	0.008	3.000	0.077	0.92	1.433	0.392	0.308
		PLS	4.000	0.000	1.00	0.015	0.009	3.000	0.244	0.80	1.517	0.412	0.277
		PCQR	4.000	0.000	1.00	0.065	0.029	3.000	5.244	0.00	1.470	0.405	0.758
		$\mathrm{PQR}_{0.05}$	4.000	0.000	1.00	0.085	0.050	3.000	0.311	0.72	2.228	1.015	0.062
		$\mathrm{PQR}_{0.5}$	4.000	0.000	1.00	0.019	0.012	3.000	2.788	0.07	1.624	0.623	0.053
		$\mathrm{PQR}_{0.75}$	4.000	0.000	1.00	0.021	0.013	3.000	2.311	0.03	1.692	0.637	0.062
		PWQR	4.000	0.000	1.00	0.016	0.010	3.000	0.100	0.90	1.512	0.507	0.528
	t_3	PMWQR	4.000	0.000	1.00	0.022	0.015	3.000	0.260	0.80	1.423	0.604	0.302
		PLS	4.000	0.000	1.00	0.055	0.043	2.990	0.260	0.77	2.230	0.872	0.277
		PCQR	4.000	0.000	1.00	0.022	0.012	3.000	7.260	0.00	1.685	0.567	0.649
		$\mathrm{PQR}_{0.05}$	4.000	0.010	0.99	0.331	0.251	2.870	1.550	0.20	4.006	2.020	0.082
		$\mathrm{PQR}_{0.5}$	4.000	0.000	1.00	0.037	0.024	3.000	4.860	0.01	1.605	0.640	0.111
		$\mathrm{PQR}_{0.75}$	4.000	0.000	1.00	0.039	0.030	3.000	4.470	0.01	1.815	0.801	0.072
		PWQR	4.000	0.000	1.00	0.025	0.017	3.000	0.290	0.76	1.591	0.585	0.425
	χ^2_4	PMWQR	4.000	0.000	1.00	0.054	0.042	2.970	0.310	0.74	1.862	0.848	0.325
		PLS	4.000	0.000	1.00	0.129	0.099	3.000	0.350	0.73	2.531	1.202	0.260
		PCQR	4.000	0.010	0.99	0.086	0.063	2.990	10.790	0.00	1.942	1.062	0.718
		$\mathrm{PQR}_{0.05}$	4.000	0.000	1.00	0.078	0.054	2.990	0.330	0.71	2.550	1.093	0.056
		$\mathrm{PQR}_{0.5}$	4.000	0.160	0.95	0.133	0.086	3.000	9.360	0.00	2.917	1.708	0.051
		$\mathrm{PQR}_{0.75}$	3.990	0.260	0.88	0.323	0.240	2.980	9.580	0.00	4.307	2.853	0.052
		PWQR	4.000	0.000	1.00	0.085	0.053	2.930	0.590	0.54	2.105	1.021	0.439
	$MN(0,0,1,10^2,0.9)$	PMWQR	4.000	0.000	1.00	0.028	0.016	3.000	0.080	0.92	1.463	0.598	0.275
		PLS	4.000	0.000	1.00	0.043	0.025	3.000	0.235	0.81	1.726	0.599	0.258
		PCQR	4.000	0.000	1.00	0.026	0.020	3.000	7.720	0.00	1.652	0.649	0.620
		$\mathrm{PQR}_{0.05}$	4.000	0.000	1.00	0.142	0.108	2.995	0.860	0.41	2.850	1.397	0.062
		$\mathrm{PQR}_{0.5}$	4.000	0.015	0.99	0.0366	0.026	3.000	5.630	0.00	1.783	0.776	0.047
		$\mathrm{PQR}_{0.5}$	4.000	0.015	0.99	0.041	0.028	3.000	4.580	0.01	1.838	0.799	0.056
		PWQR	4.000	0.000	1.00	0.027	0.018	3.000	0.105	0.90	1.711	0.663	0.473

而其他方法尤其当样本量较少且误差项存在有偏或厚尾分布时, 易选入非重要变量.

(2) 根据 GMSE 指标, PMWQR 方法有最小估计误差, 且随着样本量增加, 估计误差和标准差进一步减小.

(3) 非参数部分的变量选择中, 当样本量 $n = 400$ 时, PMWQR 方法的模型选择准确率在所有方法中最高, 而 PCQR 方法易引入非重要变量.

(4) 由 ASE 指标, PMWQR 方法对非参数部分函数的估计误差不仅随样本量的增大而减小, 在有偏及厚尾的误差分布下也显著优于其他估计方法.

(5) 相较于 PCQR 方法和 PWQR 方法, PMWQR 方法运算更快. 综上, PMWQR 方法对于高维异质性数据的建模具有广泛的适用性.

5 实证分析

5.1 实例 1: 城投债信用利差及隐性担保

自 1981 年财政部恢复国债发行以来, 中国债券市场上的债券品类从最初的国债发展到包括企业债、中期票据和金融债等各类债券.《关于加强地方政府性债务管理的意见》发布以来, 地方政府债券的发行成为地方政府融资的主要方式, 各地方政府出资成立项目公司以满足当地基础设施建设等需求, 这类城投公司发行的债券称为城投债. 近年来城投债成为诸多学者研究的热点问题 (参见文献 [27,31]).

诸多学者研究发现, 债券评级机构和机构投资者对城投债的 “名义担保” 虽然能够提升债券的评级, 但对债券发行成本的降低并无显著影响, 这一发现也验证了市场上存在着虚假担保的现象. 地方政府的财政收入状况是度量城投债 “隐性担保” 能力的重要指标, 于是本文从 “隐性担保” 对城投债融资成本的影响出发, 对 400 家地级市以下地方融资平台公司发行的企业债数据 (数据来源于 Wind 数据库及《中国城市统计年鉴》) 进行分析和研究. 为度量融资成本, 本文以地方融资平台发行城投债的信用利差 (即债券的票面利率与相同期限的国债利率之差) 为因变量, 以人均公共财政收入作为 “隐性担保” 的衡量指标, 进一步引入债券发行规模及期限作为自变量. 此外, 引入融资平台的主营业务利润率、总资产报酬率和资产负债率及银行授信额度等作为控制变量. 描述统计结果如表 6 所示.

将控制变量作为线性部分, 自变量作为非参数部分, 构建部分线性可加模型分析各变量对融资平

表 6 实例 1 各变量描述统计

变量名称	平均值 (标准差)	中位数 (绝对偏差)	最小值	最大值
信用利差 (Yield)	2.685 (1.179)	2.64 (1.423)	0.099	5
人均公共财政收入 (元, FiscalIncome)	13822.840 (111405.530)	14404.27 (12566.68)	1212.85	43817.58
债券规模 (亿元, Size)	8.545 (5.428)	7.00 (4.447)	1.10	30
债券期限 (年, Age)	4.250 (2.009)	5.00 (2.965)	1.00	10
主营业务利润率 (%, ProRatio)	15.416 (19.279)	11.34 (12.468)	−37.79	87.38
总资产报酬率 (%, ROA)	1.865 (1.648)	1.33 (0.896)	−0.78	12.59
资产负债率 (%, Lev)	57.652 (11.210)	58.553 (10.954)	28.125	79.716
银行授信额度 (亿元, Loan)	666.243 (1354.15)	146.49 (158.76)	3.78	5081.35

台信用利差的影响. 若系数为负, 则该变量对融资成本的降低有一定的促进作用:

$$\text{Yield} = \beta_0 + \beta_1 \text{ProRatio} + \beta_2 \text{ROA} + \beta_3 \text{Lev} + \beta_4 \text{Loan} + g_1(\text{FisicalIncome}) + g_2(\text{Size}) + g_3(\text{Age}) + \epsilon. \quad (5.1)$$

为检验各估计方法的有效性, 对数据进行标准化处理, 并随机取 300 个样本作为训练集, 其余 100 个样本作为测试集. 使用训练集下的中位绝对残差 (median absolute residual, MAR) 和测试集下的中位绝对预测误差 (median absolute prediction error, MAPE) 分别衡量样本内的拟合效果和样本外的预测效果. 为确定 MWQR 方法中的分位水平的个数, 本实例选取各种分位水平及分位水平个数进行试验, 发现当取均匀分位水平且个数为 5 时, 效果最佳. 为进行对比分析, CQR 方法也采用相同的分位水平选择机制. 重复随机抽样 50 次, 分别计算 MAR 与 MAPE 的均值, 结果见表 7.

由表 7 可知, 虽然中位数回归在样本内拟合取得最小拟合残差, 但 MWQR 方法的结果较优, 二者相差仅为 0.02, 且在样本外预测中 MWQR 方法达到最优, 具有最小的预测误差. 这说明利用 MWQR-PLAM 算法对信用利差问题进行研究是合适的. 为此, 本文在表 8 和图 3 中分别汇报了基于 MWQR 方法的模型 (5.1) 的参数估计值以及非参数函数的变化图.

由表 8 和图 3 可知, (1) 参数部分总资产报酬率及银行授信额度系数为负, 表明较高的投资收益及银行信贷额度有利于降低债券的融资成本; 这一点与现有文献的研究结论一致; 而较高的资产负债

表 7 实例 1 各方法估计误差

方法	MAR	MAPE
MWQR_5	0.423	0.463
$\text{QR}_{0.05}$	1.330	1.379
$\text{QR}_{0.5}$	0.421	0.465
$\text{QR}_{0.75}$	0.539	0.574
CQR_5	0.485	0.514
LS	0.461	0.503
WQR_5	0.453	0.496

表 8 实例 1 模型 5.1 的参数估计结果

变量	ProRatio	ROA	Lev	Loan
系数	0.101	−0.147	0.061	−0.155

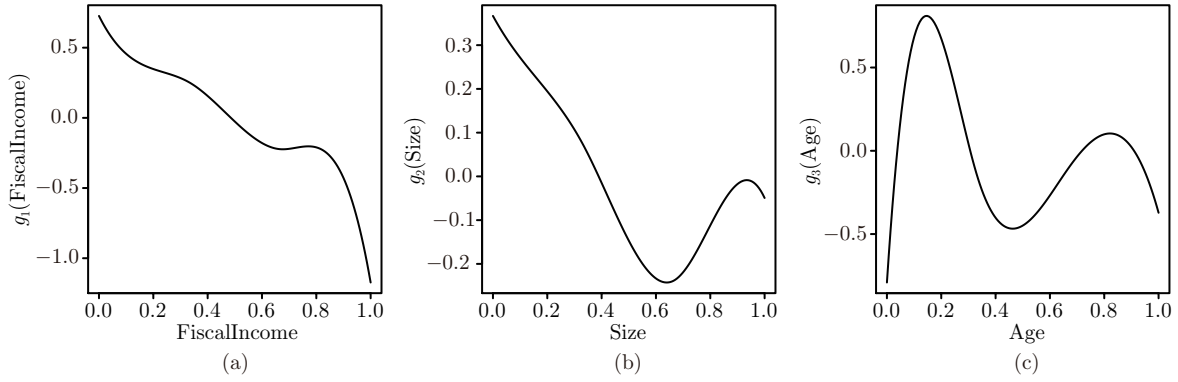

图 3 实例 1 非参数部分函数变化趋势

率表征较大的财务风险, 显然会相应提高信用利差. (2) 非参数部分, 人均公共财政收入曲线呈现明显的下降趋势, 且取值大都为负. 这正验证了城投债的偿还存在着地方政府的 “隐性担保”, 当这些融资机构或融资平台到期不能偿还债务时, 地方政府就会为其 “兜底”, 然而这种 “兜底” 作用会影响评级机构及机构投资者对债券等级质量的评价, 所以这种 “隐性担保” 并不能得到认可. 据此, 应当健全和完善监管机制, 对这些融资平台的运行进行有效监管. 而与债券相关的其他两个指标 (规模和期限), 则呈现较复杂的变化趋势, 现有研究中, 关于这两个变量对信用利差影响的正负仍未有定论. 能否寻找最优的债券发行规模以及债券的偿还期限以降低融资成本, 或将成为相关领域研究的重点问题.

5.2 实例 2: 血浆 β- 胡萝卜素浓度影响因素分析

近年来, 各界对健康问题的关注越来越高, 关于遗传疾病或重大病例的研究与日俱增. 研究表明, 较低的血浆 β- 胡萝卜素浓度会增加某些癌症的致病风险 (参见文献 [2]). 为探究血浆 β- 胡萝卜素浓度与个体特征及饮食习惯间的关系, 本小节使用数据集 (http://lib.stat.cmu.edu/datasets/Plasma_Retinol) 进行分析. 数据共包含 315 个样本, 各变量及其含义如表 9 所示.

建模之前, 除性别等虚拟变量, 本文对各变量均进行标准化处理, 并剔除样本中的一个极端值. 为验证 MWQR 方法的估计效果, 本小节参考文献 [3], 随机选取 214 个样本作为训练集, 其余 100 个样本作为测试集, 对不同方法的预测效果进行对比. 建立 PLAM 如下:

$$\begin{aligned}\text{Beta-Carotene} = {} & \beta_0 + \beta_1 \text{Quetelet} + \beta_2 \text{Calories} + \beta_3 \text{Fat} + \beta_4 \text{Alcohol} + \beta_5 \text{BetaDiet} + \beta_6 \text{Sex} \\ & + \beta_7 \text{Smok1} + \beta_8 \text{Smok2} + g_1(\text{Age}) + g_2(\text{Fiber}) + g_3(\text{Chol}) + \epsilon. \end{aligned} \tag{5.2}$$

参考文献 [3], 重复抽样 20 次, 在 B 样条拟合中选择节点个数为 2. 训练集与测试集评价指标仍分别采用中位绝对残差 MAR 和中位绝对预测误差 MAPE. 表 10 列出了 20 次模拟的两指标的平均值.

表 10 结果表明, 无论是样本内预测还是测试集预测结果, MWQR 方法都显著优于其他算法. 表 11 和图 4 分别给出基于 MWQR 算法的模型参数估计值和 3 个变量对应的非参数函数拟合图.

表 9 实例 2 的变量及其含义

变量名	含义
Beta-Carotene	血浆中 β- 胡萝卜素浓度
Quetelet	Quetelet 指数 = 体重/身高 ^{2}
Calories	每天消耗的热量
Fat	每天摄入的脂肪含量
Alcohol	每周摄入的酒精含量
BetaDiet	每天摄入的 β- 胡萝卜素含量
Sex	性别, 1 = 女性, 2 = 男性
Smok1	1 = 之前吸烟, 0 = 其他
Smok2	1 = 目前吸烟, 0 = 其他
Vit1	1 = 经常使用维生素, 0 = 其他
Vit2	1 = 不经常使用维生素, 0 = 其他
Age	年龄
Chol	每天摄入的胆固醇含量
Fiber	每天摄入的纤维含量

表 10 实例 2 的各方法估计误差

方法	MAR	MAPE
MWQR_9	0.233	0.322
$\mathrm{QR}_{0.05}$	0.422	0.473
$\mathrm{QR}_{0.5}$	0.251	0.352
$\mathrm{QR}_{0.75}$	0.402	0.520
CQR_9	0.312	0.375
LS	0.373	0.409
WQR_9	0.236	0.327

表 11 实例 2 线性部分变量系数

变量	Quetelet	Calories	Fat	Alcohol	BetaDiet	Sex	Smok1	Smok2	Vit1	Vit2
系数	−0.117	0.003	−0.033	0.001	0.104	0.223	−0.071	−0.095	0.087	0.120

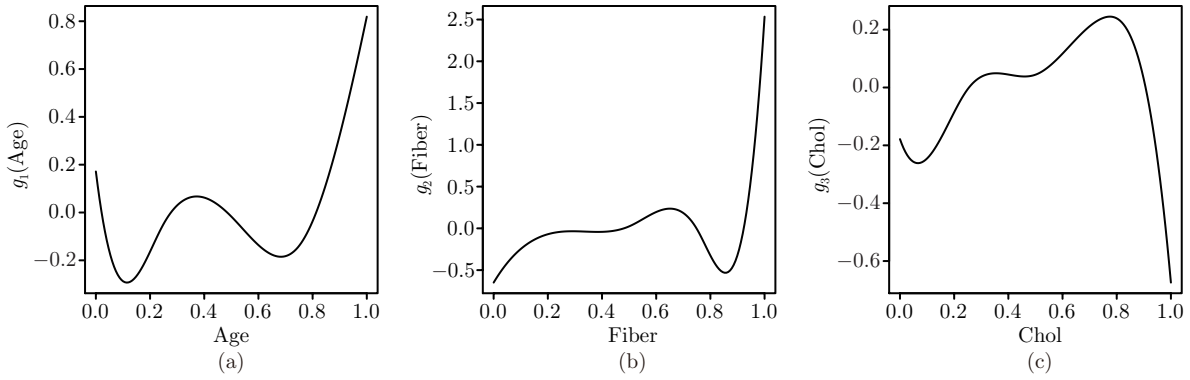

图 4 实例 2 非参数部分函数变化趋势

上述结果表明: (1) 参数部分, 除 Quetelet 指数、脂肪摄入量和吸烟者系数外, 其他系数均为正, 表明除这 3 种因素外, 其他因素都与血浆中 β- 胡萝卜素含量呈正相关关系. 其中每天消耗的热量及每周摄入的酒精含量对 β- 胡萝卜素含量影响较微弱. (2) 非参数部分, 年龄与血浆中 β- 胡萝卜素浓度大致呈现 “U 型” 关系, 而胆固醇含量与胡萝卜素浓度大致呈现 “倒 U 型” 关系. 随着每日纤维摄入量的增加, β- 胡萝卜素浓度会呈现一种骤降进而大幅上升的态势. 为进一步探究 PMWQR 方法在变量选择方面的效果, 首先对线性部分和非参数部分协变量分别设置扰动项: 对线性部分设置 50 个扰动变量, 对非参数部分设置 10 个扰动变量, 各变量独立取自均匀分布 $U(0,1)$, 仍随机抽取 214 个样本, 并重复抽样 20 次, 记录各方法下线性部分及非参数部分选择变量个数的均值以及扰动变量个数的均值, 结果如表 12 所示.

由表 12 可以发现, 与其他方法相比, 利用本文提出的 PMWQR 方法得到的模型最精简, 无论是参数部分还是非参数部分, 均选择了最少的变量. 从扰动变量的个数来看, PMWQR 方法倾向选择最少的噪声变量进入模型, 进一步表明本文提出方法的稳健性以及其在变量选择方面的优势. 表 13 还记录了 20 次重复抽样下参数部分 10 个变量及非参数部分 3 个变量被选择的概率. 不难发现, 非参数部分的 3 个变量 (年龄、纤维摄入量及胆固醇摄入量) 均以较高的概率被选入模型, 说明这 3 个变量对血浆中 β- 胡萝卜素浓度影响显著; 而在参数部分中, 2 个变量 (Quetelet 指数及 β- 胡萝卜素摄入

表 12　实例 2 各方法变量选择结果

方法	线性部分		非参数部分	
	变量个数	扰动变量个数	变量个数	扰动变量个数
$PMWQR_9$	1.55	1.45	4.75	2.05
$PQR_{0.05}$	5.50	3.20	10.10	7.30
$PQR_{0.5}$	4.40	2.40	8.95	6.00
$PQR_{0.75}$	4.20	7.70	6.90	7.45
$PCQR_9$	3.95	2.05	8.85	6.10
PLS	14.20	8.35	7.40	5.55
WQR_9	2.35	2.70	6.05	3.25

表 13　实例 2 各变量选择概率

变量	Quetelet	Calories	Fat	Alcohol	BetaDiet	Sex	Smok1	Smok2	Vit1	Vit2	Age	Fiber	Chol
系数	0.40	0.00	0 .00	0.00	0.50	0.05	0.00	0.25	0.35	0.00	1.00	0.95	0.75

量) 选择概率较大, 这与 Guo 等 [3] 的结果一致; 是否吸烟及是否经常使用维生素也会对血浆中 β- 胡萝卜素浓度产生微弱的影响.

6　结论

数据日益的复杂化、高维化与异质性特征给统计分析带来严峻的挑战. 为适应数据多样性与信息多元化的需求, 并兼顾计算可行性、算法稳健性与估计有效性, 本文提出一种基于众数的 MWQR 方法. 该方法通过众数选取每个样本最具信息量的分位水平, 并对不同最优分位水平进行加权综合, 最大程度地利用样本信息. 由于部分线性可加模型的广泛适用性, 本文还提出了针对于部分线性可加模型的 MWQR-PLAM 算法和 PMWQR-PLAM 算法. 与最小二乘方法相比, MWQR-PLAM 算法在误差项厚尾分布下仍能保持稳健性; 与 WQR 和 CQR 方法相比, MWQR-PLAM 算法能避免分位水平主观选取带来的偏误; 与分位回归及众数回归方法相比, 该算法使用多个分位水平有效综合了数据分布中的信息. 此外 PMWQR-PLAM 算法能够实现模型简约并提高解释力. 特别当数据分布存在多个众数时, MWQR 方法显著优于其他方法. 本文不仅从理论上证明了算法的相合性, 推导了其渐近分布, 还通过数值模拟及两组实际数据验证了所提出算法在参数估计和变量选择方面的优势及广泛适用价值.

在未来的研究工作中, 可以考虑算法的进一步优化以及 MWQR 方法中最优权重的求解问题; 为使 MWQR 方法适应于更加复杂的应用问题, 一方面可以探究基于多峰分布及峰值平坦分布的更加合理的最优分位水平的定义方式, 另一方面可将 MWQR 方法推广到更一般的非参数及半参数模型中; 也可以考虑结合分布式运算等方法, 使 MWQR 方法适应于大数据分析.

致谢　感谢审稿人提出的宝贵意见.

参考文献

1　Fan J, Härdle W, Mammen E. Direct estimation of low-dimensional components in additive models. Ann Statist, 1998, 26: 943–971

2 Fiedor J, Przetocki M, Siniarski A, et al. β-carotene-induced alterations in haemoglobin affinity to O_2. Antioxidants, 2021, 10: 451
3 Guo J, Tang M, Tian M, et al. Variable selection in high-dimensional partially linear additive models for composite quantile regression. Comput Statist Data Anal, 2013, 65: 56–67
4 Guo J X, Xu H C, Zhu W Q, et al. Distributed estimation for heterogeneous big data (in Chinese). Statist Res, 2020, 37: 104–114 [郭婧璇, 徐慧超, 祝婉晴, 等. 异质性大数据的分布式估计. 统计研究, 2020, 37: 104–114]
5 He X, Shi P. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. J Nonparametr Stat, 1994, 3: 299–308
6 He X, Wang L, Hong H G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Ann Statist, 2013, 41: 342–369
7 Hoshino T. Quantile regression estimation of partially linear additive models. J Nonparametr Stat, 2014, 26: 509–536
8 Huang J, Horowitz J L, Wei F. Variable selection in nonparametric additive models. Ann Statist, 2010, 38: 2282–2313
9 Ibarra-Espinosa S, Dias de Freitas E, Ropkins K, et al. Negative-Binomial and quasi-poisson regressions between COVID-19, mobility and environment in São Paulo, Brazil. Environ Res, 2022, 204: 112369
10 Kazemi M, Shahsavani D, Arashi M. A sure independence screening procedure for ultra-high dimensional partially linear additive models. J Appl Stat, 2019, 46: 1385–1403
11 Koenker R. Additive models for quantile regression: Model selection and confidence bandaids. Braz J Probab Stat, 2011, 25: 239–262
12 Koenker R, Machado J A F. Goodness of fit and related inference processes for quantile regression. J Amer Statist Assoc, 1999, 94: 1296–1310
13 Li Z D, Lin J H, Wang M J. High-dimensional statistics in big data era: Development and application of sparse modeling (in Chinese). Statist Res, 2015, 32: 3–11 [李仲达, 林建浩, 王美今. 大数据时代的高维统计: 稀疏建模的发展及其应用. 统计研究, 2015, 32: 3–11]
14 Lian H. Variable selection in high-dimensional partly linear additive models. J Nonparametr Stat, 2012, 24: 825–839
15 Liang H, Thurston S W, Ruppert D, et al. Additive partial linear models with measurement errors. Biometrika, 2008, 95: 667–678
16 Liu X, Wang L, Liang H. Estimation and variable selection for semiparametric additive partial linear models. Statist Sinica, 2011, 21: 1225–1248
17 Lv J, Yang H, Guo C. Variable selection in partially linear additive models for modal regression. Commun Stat-Simul Computation, 2017, 46: 5646–5665
18 Ma S, Song P X K. Varying index coefficient models. J Amer Statist Assoc, 2015, 110: 341–356
19 Ma S, Yang L. Spline-backfitted kernel smoothing of partially linear additive model. J Statist Plann Inference, 2011, 141: 204–219
20 Nguelifack B M, Kemajou-Brown I. Robust signed-rank estimation and variable selection for semi-parametric additive partial linear models. J Appl Stat, 2020, 47: 1794–1819
21 Opsomer J D, Ruppert D. Fitting a bivariate additive model by local polynomial regression. Ann Statist, 1997, 25: 186–211
22 Ota H, Kato K, Hara S. Quantile regression approach to conditional mode estimation. Electron J Stat, 2019, 13: 3120–3160
23 Sherwood B, Wang L. Partially linear additive quantile regression in ultra-high dimension. Ann Statist, 2016, 44: 288–317
24 Song Y J, Liu J X, Zhao M L. Has economic openness improved enterprise innovation?—An empirical analysis based on the semi-parametric additive panel model (in Chinese). J Shandong Univ, 2019, 234: 68–80 [宋英杰, 刘俊现, 赵明亮. 经济开放提高了企业创新力吗?—基于半参数可加面板模型的实证分析. 山东大学学报 (哲学社会科学版), 2019, 234: 68–80]
25 Stone C J. Additive regression and other nonparametric models. Ann Statist, 1985, 13: 689–705
26 Sun C Y, Ma X Y, Diao H T, et al. Analysis of influencing factors of Yu E Bao yield based on semi-parametric additive model (in Chinese). J Nanjing Univ Finance Economics, 2018, 212: 62–71 [孙春燕, 马馨悦, 刁海涛, 等. 基于半参数可加模型的余额宝收益率影响因素分析. 南京财经大学学报, 2018, 212: 62–71]
27 Tie Y, He H L. Banking deregulation,financial open source and local government debt regulation (in Chinese). Public Finance Res, 2020, 453: 71–83 [铁瑛, 何欢浪. 银行管制放松、财政开源与地方政府债务治理. 财政研究, 2020, 453: 71–83]
28 Tukey J W. Which part of the sample contains the information? Proc Natl Acad Sci USA, 1965, 53: 127–134
29 Wang S X, You J H, Huang T. Modelling and applications for non-stationary time series in the presence of trend and period. Sci Sin Math, 2022, 52: 177–208 [王守霞, 尤进红, 黄涛. 存在趋势和周期特征的非平稳时间序列的建模及

其应用. 中国科学: 数学, 2022, 52: 177–208]
30 Xiong W, Tian M. Weighted quantile regression theory and its application. J Data Sci, 2019, 17: 145–160
31 Zhu Y, Wang J. Can market discipline affect local government bonds' risk premium?—Evidence from the Chengdu bond market (in Chinese). J Financ Res, 2018, 456: 56–72 [朱莹, 王健. 市场约束能够降低地方债风险溢价吗?—来自城投债市场的证据. 金融研究, 2018, 456: 56–72]
32 Zou H. The adaptive Lasso and its oracle properties. J Amer Statist Assoc, 2006, 101: 1418–1429

附录 A

附录中给出一些引理、定理的证明.

定理 2.1 的证明 由 (2.2) 及 $\sum_{k=1}^{K} w_k^* = 1$, 显然有 $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}}) = \boldsymbol{\beta}$,

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}}) &= \mathrm{Var}\bigg(\sum_{k=1}^{K} w_k^* \hat{\boldsymbol{\beta}}_{\tau_k}\bigg) \\ &= \sum_{i=1}^{K}\sum_{j=1}^{K} w_i^* w_j^* \mathrm{cov}(\hat{\boldsymbol{\beta}}_{\tau_i}, \hat{\boldsymbol{\beta}}_{\tau_j}) \\ &= \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{K} w_i^* w_j^* D^{-1} \frac{\tau_i \wedge \tau_j - \tau_i\tau_j}{f(F^{-1}(\tau_i)) f(F^{-1}(\tau_j))} \\ &= \frac{\frac{1}{n} D^{-1} \sum_{i=1}^{K}\sum_{j=1}^{K} n_i n_j \frac{\tau_i \wedge \tau_j - \tau_i\tau_j}{\sqrt{\tau_i(1-\tau_i)}\sqrt{\tau_j(1-\tau_j)}}}{(\sum_{k=1}^{K} n_k \frac{f(F^{-1}(\tau_k))}{\sqrt{\tau_k(1-\tau_k)}})^2}.\end{aligned}$$

当 $n_1 = n_2 = \cdots = n_k$ 时, 方差退化为

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}}) = \frac{\frac{1}{n} D^{-1} \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\tau_i \wedge \tau_j - \tau_i\tau_j}{\sqrt{\tau_i(1-\tau_i)}\sqrt{\tau_j(1-\tau_j)}}}{(\sum_{k=1}^{K} \frac{f(F^{-1}(\tau_k))}{\sqrt{\tau_k(1-\tau_k)}})^2}.$$

又因为 $\hat{\boldsymbol{\beta}}_{\tau_k}$ 服从正态分布, 所以定理 2.1 得证. □

为证明定理 3.1, 引入如下两个引理.

引理 A.1 若条件 3.1–3.5 成立, 则对于任意 $g_n \in S_n$, 均有

$$\|g_n - g\|_2 = O_p(L^{-\rho} + L^{1/2} n^{-1/2}).$$

特别地, 在给定 $L = O_p(n^{1/(2k_n+1)})$ 的条件下, 有

$$\|g_n - g\|_2 = O_p\{(L/n)^{1/2}\} = O_p(n^{-k_n/(2k_n+1)}).$$

引理 A.1 的证明可参见文献 [8], 这里不再具体说明.

引理 A.2 假定随机变量 V_j 存在有限二阶矩, 且 $\sum_{j=1}^{J} V_j$ 同样存在有限二阶矩, 则有

$$\mathrm{SD}(V_1 + \cdots + V_j) \geqslant \bigg(\frac{1-\delta}{2}\bigg)^{(j-1)/2} (\mathrm{SD}(V_1) + \cdots + \mathrm{SD}(V_j)), \quad 0 < \delta < 1,$$

其中 $\mathrm{SD}(\cdot)$ 为标准差.

引理 A.2 的具体证明过程可参见文献 [25].

定理 3.1 的证明 由定理 3.1 给定的条件, 假定 $L=O(n^{1/(2k_n+1)})$. 此外, 令 g_n 为集合 $\mathcal{H}$ 中对函数 g 拟合最好的函数, 并基于 g_n 定义

$$S_{nj}:=\left\{g_{nj}:g_{nj}=\sum_{u=1}^{L}\theta_{ju}\psi_{ju}(z),(\theta_{j1},\ldots,\theta_{jl})\in R^L\right\}.$$

令 $\mathbf{\Sigma}_X=\mathrm{cov}(X)$ 和 $\mathbf{\Sigma}_\psi=\mathrm{cov}(\psi)$, 由定理 2.1 易得部分线性可加模型中估计量渐近分布满足

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MWQR}}-\boldsymbol{\beta})\xrightarrow{d}N(0,m\mathbf{\Sigma}_X^{-1}),$$
$$\sqrt{n}(\hat{\Theta}_{\mathrm{MWQR}}-\Theta)\xrightarrow{d}N(0,m\mathbf{\Sigma}_\psi^{-1}),$$

其中

$$m=\frac{\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\tau_i\wedge\tau_j-\tau_i\tau_j}{\sqrt{\tau_i(1-\tau_i)}\sqrt{\tau_j(1-\tau_j)}}}{(\sum_{k=1}^{K}\frac{f(F^{-1}(\tau_k))}{\sqrt{\tau_k(1-\tau_k)}})^2}.$$

令 $\sum_{j=1}^{p}g_j(Z_j)=g$, 将 MWQR 估计方法得到的非参数部分函数估计值简记为 $\hat{g}$. 由于 $\mathrm{E}\|\hat{g}-g\|_2\leqslant\mathrm{E}\|\hat{g}-g_n\|_2+\mathrm{E}\|g_n-g\|_2$, 故有

$$\begin{aligned}\mathrm{E}\|\hat{g}-g_n\|_2&=[\mathrm{E}(\hat{g}-g_n)^2]^{1/2}=\{\mathrm{E}[\mathrm{tr}(\hat{\theta}_n-\theta_n)^{\mathrm{T}}\psi_i\psi_i^{\mathrm{T}}(\hat{\theta}_n-\theta_n)]\}^{1/2}\\&=\{\mathrm{tr}[\mathrm{E}(\psi\psi^{\mathrm{T}})\mathrm{E}(\hat{\theta}_n-\theta_n)(\hat{\theta}_n-\theta_n)^{\mathrm{T}}]\}^{1/2}\\&=\{n^{-1}m\cdot\mathrm{tr}[\mathrm{E}(\psi\psi^{\mathrm{T}})\mathrm{Var}(\psi)^{-1}]\}^{1/2}\\&=(mp)^{1/2}\left(\frac{L}{n}\right)^{1/2}\\&=O\left\{\left(\frac{L}{n}\right)^{1/2}\right\}.\end{aligned}$$

根据引理 A.1, 同样有 $\|g_n-g\|_2=O_p\{(L/n)^{1/2}\}$, 故 $\|\hat{g}-g\|_2=O_p\{(L/n)^{1/2}\}$. 由引理 A.2, 有

$$\|\hat{g}_j(Z_j)-g_j(Z_j)\|_2=O_p\left\{\left(\frac{L}{n}\right)^{1/2}\right\},$$

即 $1/n\sum_{i=1}^{n}(\hat{g}_j(Z_{ij})-g_j(Z_{ij}))^2=O_p(n^{-2k_n/(2k_n+1)})$, 定理 3.1(3) 得证. □

推论 A.1 若上述定理中的条件成立, 则基于 Ma 和 Yang [19] 提出的反向拟合算法, 以 $g_1(\cdot)$ 的估计值为例, 可以得到估计值的渐近分布满足

$$\sqrt{nh}\{\hat{g}_{\mathrm{MWQR},1}(z_1)-g_1(z_1)-w\cdot h^2\cdot b_1(z_1)\}\xrightarrow{d}N\left(0,C\sum_i\sum_j\frac{(w_iw_j)^{6/5}}{\sum_i w_i^2}\right),$$

其中 $C=[v_1^2(z_1)]^2+t^2[2v_1^2(z_1)-1]$, 这里,

$$\begin{aligned}t&=g_1(z_1)+w\cdot h^2\cdot b_1(z_1),\\v_1^2(z_1)&=\int K^2(u)du\mathrm{E}[\sigma^2(\mathrm{Z},\mathrm{X})\mid Z_1=z_1]f_1^{-1}(z_1),\\b_1(z_1)&=\int u^2K(u)du\left\{g_1''(z_1)\frac{f_1(z_1)}{2}+g_1'(z_1)f_1'(z_1)\right\}f_1^{-1}(z_1).\end{aligned}$$

为证明推论 A.1, 引入引理 A.3.

引理 A.3 若文献 [19] 中的条件 (C1)–(C7) 成立, 则以 $\hat{g}_1(z_1)$ 的估计为例, 其渐近分布满足

$$\sqrt{nh}\{\hat{g}_{\mathrm{SBK},1,\tau_k}(z_1) - g_1(z_1) - b_1(z_1)h_k^2\} \xrightarrow{d} N(0, v_1^2(z_1)),$$

其中,

$$v_1^2(z_1) = \int K^2(u)du\mathrm{E}[\sigma^2(Z,X) \mid Z_1 = z_1]f_1^{-1}(z_1),$$
$$b_1(z_1) = \int u^2K(u)du\left\{g_1''(z_1)\frac{f_1(z_1)}{2} + g_1'(z_1)f_1'(z_1)\right\}f_1^{-1}(z_1).$$

推论 A.1 的证明 为得到 $g_\alpha(Z_\alpha)$ 的估计值 $\hat{g}_\alpha(Z_\alpha)$, 首先基于分位回归方法, 在 τ_k 分位点下得到线性部分参数估计值 $\hat{\beta}_{\tau_k}$ 以及非参数部分函数估计值 g_{j,τ_k} $(j = 1, \ldots, p,\ j \neq \alpha)$, 其中非线性部分的拟合仍基于样条方法. 据此, 定义 $\hat{Y}_{i\alpha,\tau_k} = Y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\beta}_{\tau_k} - \sum_{j=1,j\neq\alpha}^{p}\hat{g}_{j,\tau_j}(Z_{ij})$, 不同分位水平 τ_k 下考虑窗宽 $h_k = (\frac{\tau_k(1-\tau_k)}{f^2(F^{-1}(\tau_k))})^{1/5}\cdot h$, 其中 h 为最小二乘估计下的最优窗宽. 进一步可以得到基于样条方法的向后拟合 (spline-backfitted kernel, SBK) 估计量为

$$\hat{g}_{\mathrm{SBK},\alpha,\tau_k}(z_\alpha) = \left\{n^{-1}\sum_{i=1}^{n}K_{h_k}(Z_{i\alpha} - z_\alpha)\hat{Y}_{i\alpha,\tau_k}\right\}\hat{f}_\alpha(z_\alpha),$$

其中, $K_{h_k}(u) = 1/h_k \cdot K(u/h_k)$, K 为核密度函数, $\hat{f}_\alpha(z_\alpha) = 1/n\sum_{i=1}^{n}K_{h_k}(Z_{i\alpha} - z_\alpha)$. 由于仅考虑 $n_1 = n_2 = \cdots = n_k$ 的特殊情形, 基于不同分位点下的估计值 $\hat{g}_{\mathrm{SBK},\alpha,\tau_k}(z_\alpha)$, 可进一步得到 WQR 估计值 $\hat{g}_{\mathrm{MWQR},\alpha}(z_\alpha)$.

根据引理 A.3, 将 $\hat{g}_{\mathrm{SBK},1,\tau_k}(z_1)$ 简记为 $\hat{g}_{k1}(z_1)$, 于是有

$$\begin{aligned}
E(\hat{g}_{\mathrm{MWQR},1}(z_1)) &= E\left(\sum_{k=1}^{K}w_k^*\hat{g}_{k1}(z_1)\right) = \sum_{k=1}^{K}w_k^*E\{\hat{g}_{k1}(z_1)\} \\
&= \sum_{k=1}^{K}w_k^*\{g_1(z_1) + b_1(z_1)h_k^2\} \\
&= g_1(z_1) + b_1(z_1)\left(\sum_{k=1}^{K}w_k^*h_k^2\right).
\end{aligned}$$

而 $w_k^*h_k^2 = \frac{\sum_k w_kh_k^2}{\sum_k w_k} = (\frac{\sum_k w_k^{1/5}}{\sum_k w_k})\cdot h^2 \triangleq w\cdot h^2$, 因此有

$$E(\hat{g}_{\mathrm{MWQR},1}(z_1)) = g_1(z_1) + w\cdot h^2\cdot b_1(z_1),$$
$$\mathrm{Var}(\hat{g}_{\mathrm{MWQR},1}(z_1)) = \mathrm{Var}\left(\sum_{k=1}^{K}w_k^*\hat{g}_{k1}(z_1)\right) = \sum_i\sum_j w_i^*w_j^*\mathrm{cov}(\hat{g}_{i1}(z_1), \hat{g}_{j1}(z_1)).$$

令 $t = E(\hat{g}_{\mathrm{MWQR},1}(z_1)) = g_1(z_1) + w\cdot h^2\cdot b_1(z_1)$, 易得

$$\mathrm{cov}(\hat{g}_{i1}(z_1), \hat{g}_{j1}(z_1)) = [V_1^2(z_1)]^2 + t^2[2v_1^2(z_1) - 1] = C,$$

故有

$$\mathrm{Var}(\hat{g}_{\mathrm{MWQR},1}(z_1)) = \frac{1}{n}\sum_i\sum_j w_i^*w_j^*\frac{1}{\sqrt{h_i\cdot h_j}}\cdot C = \frac{C}{nh}\sum_i\sum_j\frac{(w_iw_j)^{6/5}}{\sum_i w_i^2}.$$

证毕. □

为证明定理 3.2, 同时引入如下两条引理:

引理 A.4 假定存在常数 L, 对于任意 $\epsilon>0$, 当 $n\to\infty$ 时,

$$\mathrm{P}\Big\{\inf_{\|\theta|\geqslant Lk_n^{1/2}}\sum_{i=1}^{n}k_{n,\tau_k}(\varepsilon_i-T_i\Theta-R_i)>\sum_{i=1}^{n}k_{n,\tau_k}(\varepsilon_i-R_i)\Big\}>1-\varepsilon.$$

引理 A.5 若条件 3.1–3.6 成立, 则有

$$(\hat\Theta-\tilde\Theta)^{\mathrm{T}}T'T(\hat\Theta-\tilde\Theta)=O_p(k_n),\quad |\hat\Theta-\tilde\Theta|^2=O_p\Big(\frac{k_n^2}{n}\Big).$$

定理 3.2 的证明 首先假定存在 $\tilde\theta_j\psi_j(u)$ $(j=1,\dots,p)$, 则有 $g(u)=\tilde\theta_j\psi_j(u)-r_j(u)$. 其次, 令 $T_i=(X_i,\psi_i)^{\mathrm{T}}$, $\tilde\Theta=(\tilde\beta_{\tau_k},\tilde\theta_{\tau_k})$, $Q_{\tau_k}(x,z)=T_i^{\mathrm{T}}\tilde\Theta_{\tau_k}=R(x,z)$. 为方便起见, 后文用 θ 和 β 分别代替 θ_{τ_k} 和 β_{τ_k}.

假定参数部分重要变量集合为 $X_1,\dots,X_{s_1}$, 非参数部分重要变量集合为 $Z_1,\dots,Z_{s_2}$, k_{01} 表示 $l>s_1$, k_{02} 表示 $j>s_2$, 令 $\hat\Theta^*$ 表示 $\hat\Theta$ 中所有下标为 k_{01} 和 k_{02} 的系数替换为 0. 基于此, 这里采用反证法证明定理 3.2(1). 具体而言, 假定非重要变量集中存在参数部分变量系数不为 0 或非参数部分的函数取值不为 0, 即 $\|\hat\beta_{k_{01}}\|\neq0$ 或 $\|\hat\theta_{k_{02}}\|\neq0$. 对于任意给定的分位水平 τ_k, 令 $\Theta-\tilde\Theta=\eta_k$, 而 $\eta_k=(\eta_{k1},\eta_{k2})$, 有

$$\begin{aligned}p_k(\Theta)&=\frac1n\sum_{i=1}^{n}\rho_{\tau_k}(y_i-T_i\Theta)+\sum_{l=1}^{d}p_{\lambda_{k1}}(|\beta_l|)+\sum_{j=1}^{p}p_{\lambda_{k2}}(\|\theta_j\|_H)\\&=\frac1n\sum_{i=1}^{n}[\rho_{\tau_k}(\varepsilon_i-T_i\eta_k-R_i)-\rho_{\tau_k}(\varepsilon_i-R_i)]+\sum_{l=1}^{d}\lambda_{k1}w_l(|\tilde\beta_l+\eta_{k1,l}|-|\tilde\beta_i|)\\&\quad+\sum_{j=1}^{p}\lambda_{k2}w_j(\|\tilde\theta_j+\eta_{k2,j}\|_H-\|\tilde\theta_j\|_H).\end{aligned}$$

由于 $\hat\Theta$ 是使得上述表达式最小的参数估计值, 故有 $p_k(\hat\Theta)-p_k(\hat\Theta^*)\leqslant0$, 而

$$\begin{aligned}p_k(\hat\Theta)-p_k(\hat\Theta^*)&=\frac1n\sum_{i=1}^{n}[\rho_{\tau_k}(\varepsilon_i-T_i(\hat\Theta-\tilde\Theta)-R_i)-\rho_{\tau_k}(\varepsilon_i-T_i(\hat\Theta\hat\Theta^*-\tilde\Theta)-R_i)]\\&\quad+\sum_{l=1}^{d}\lambda_{k1}w_l(|\hat\beta_l|-|\hat\beta_l^*|)+\sum_{j=1}^{p}\lambda_{k2}w_j(\|\hat\theta_j\|_H-\|\hat\theta_j\|_H^*)\\&=\frac1n\sum_{i=1}^{n}[\rho_{\tau_k}(\varepsilon_i-T_i(\hat\Theta-\tilde\Theta)-R_i)-\rho_{\tau_k}(\varepsilon_i-T_i(\hat\Theta^*-\tilde\Theta)-R_i)]\\&\quad+\lambda_{k1}w_{k_{01}}\|\hat\beta_{k_{01}}\|+\lambda_{k2}w_{k_{02}}\|\hat\theta_{k_{02}}\|_H\\&\geqslant\frac1n\sum_{i=1}^{n}[\rho_{\tau_k}(\varepsilon_i-R_i)-\rho_{\tau_k}(\varepsilon_i-T_i(\hat\Theta^*-\tilde\Theta)-R_i)]\\&\quad+\lambda_{k1}w_{k_{01}}\|\hat\beta_{k_{01}}\|+\lambda_{k2}w_{k_{02}}\|\hat\theta_{k_{02}}\|_H.\end{aligned}$$

根据引理 A.5, 有 $\|\hat\Theta^*-\tilde\Theta\|=O_p(k_n^{-2r})$. 令

$$I=\frac1n\sum_{i=1}^{n}[\rho_{\tau_k}(\varepsilon_i-R_i)-\rho_{\tau_k}(\varepsilon_i-T_i(\hat\Theta^*-\tilde\Theta)-R_i)],$$

$$u_i = \varepsilon_i - R_i, \quad v_i = T_i(\hat{\Theta}^* - \tilde{\Theta}), \quad \psi_\tau(u) = \tau - I(u < 0),$$

则有

$$\begin{aligned} I &= \frac{1}{n}\sum_{i=1}^{n}[\rho_{\tau_k}(u_i) - \rho_{\tau_k}(u_i - v_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[v_i\psi_{\tau_k}(u_i) - \int_0^{v_i}(I(u_i \leqslant s) - I(u_i \leqslant 0))ds\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}v_i\psi_{\tau_k}(u_i) - \frac{1}{n}\sum_{i=1}^{n}\int_0^{v_i}(I(u_i \leqslant s) - I(u_i \leqslant 0))ds \\ &=: I_1 - I_2, \end{aligned}$$

其中,

$$\begin{aligned} I_1 &\geqslant (\tau_k - 1)(\hat{\Theta}^* - \Theta)\left(\frac{1}{n}\sum_{i=1}^{n}T_i\right) = O_p(n^{-r/(2r+1)}), \\ I_2 &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{v_i}(I(u_i \leqslant s) - I(u_i \leqslant 0))ds \\ &= \mathrm{E}\left(\int_0^{v}(I(u \leqslant s) - I(u \leqslant 0))ds\right) + o(1) \\ &= \int_0^{v}[\mathrm{P}(\varepsilon \leqslant R + s) - \mathrm{P}(\varepsilon \leqslant R)]ds + o(1) \\ &= \int_0^{v}sf(R)ds \\ &= \frac{1}{2}f(R)(\hat{\Theta}^* - \tilde{\Theta})^{\mathrm{T}}T'T(\hat{\Theta}^* - \tilde{\Theta}) + o(1) \\ &= O_p(k_n). \end{aligned}$$

因此,

$$\begin{aligned} p_k(\hat{\Theta}) - p_k(\hat{\Theta}^*) &\geqslant I_1 - I_2 + \lambda_{k1}w_{k_{01}}\|\hat{\beta}_{k_{01}}\| + \lambda_{k2}w_{k_{02}}\|\hat{\theta}_{k_{02}}\|_H \\ &\geqslant (\tau_k - 1)O_p(n^{-r/(2r+1)}) - O_p(k_n) + \lambda_{k1}w_{k_{01}}\|\hat{k}_{k_{01}}\| + \lambda_{k2}w_{k_{02}}\|\hat{\theta}_{k_{02}}\|_H. \end{aligned}$$

由于 $n^{-r/(2r+1)}\lambda_n \to \infty$, 这与 $p_k(\hat{\Theta}) - p_k(\hat{\Theta}^*) \leqslant 0$ 相悖, 所以定理 3.2(1) 得证.

为证明定理 3.2(2), 假定参数 β 已知, 令 $g = ((g^{(1)})^{\mathrm{T}}, (g^{(2)})^{\mathrm{T}})^{\mathrm{T}}$, 其中 $(g^{(1)})^{\mathrm{T}}$ 和 $(g^{(2)})^{\mathrm{T}}$ 分别代表非参数部分的重要成分和非重要成分. 令非参数部分样条基函数的系数 $\theta^{(1)} = (\theta_1^{\mathrm{T}}, \ldots, \theta_{s_1}^{\mathrm{T}})$ 和 $\theta^{(2)} = (\theta_{s+1}^{\mathrm{T}}, \ldots, \theta_p^{\mathrm{T}})$ 分别对应 $(g^{(1)})^{\mathrm{T}}$ 和 $(g^{(2)})^{\mathrm{T}}$, 此外, 记

$$\tilde{\Theta}_{\mathrm{oracle}} = \arg\min_{\theta=(\theta^{(1)\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}}\frac{1}{n}\sum_{i=1}^{n}\rho_{\tau_k}(y_i - T_i\Theta).$$

与之对应, $\tilde{g}_{\mathrm{oracle}}$ 为对应参数 $\tilde{\theta}_{\mathrm{oracle}}$ 下的非参数部分估计量, 据此, 有

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{\tau_k,j}(u_i) - g_j(u_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{\tau_k,j}(u_i) - \tilde{g}_{\mathrm{oracle},j}(u_i) + \tilde{g}_{\mathrm{oracle},j}(u_i) - g_j(u_i))^2$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{\tau_k,j}(u_i)-\tilde{g}_{\text{oracle},j}(u_i))^2+\frac{1}{n}\sum_{i=1}^{n}(\tilde{g}_{\text{oracle},j}(u_i)-g_j(u_i))^2\\
&\quad+\frac{2}{n}\sum_{i=1}^{n}(\hat{g}_{\tau_k,j}(u_i)-\tilde{g}_{\text{oracle},j}(u_i))(\tilde{g}_{\text{oracle},j}(u_i)-g_j(u_i))\\
&=: I+II+III.
\end{aligned}
$$

根据引理 A.5, 有 $II=1/n\sum_{i=1}^{n}(\tilde{g}_{\text{oracle},j}-g_j(u_i))^2=O_p(k_n/n)$. 接下来, 为得到 I 的渐近性质, 令 $\hat{\Theta}_{\tau_k}-\tilde{\Theta}_{\text{oracle}}=\eta_n$, $s_i=y_i-T_i\tilde{\Theta}_{\text{oracle}}$, $t_i=T_i(\hat{\Theta}_{\tau_k}-\tilde{\Theta}_{\text{oracle}})$, 则有

$$
\begin{aligned}
0&\geqslant p_k(\hat{\Theta}_{\tau_k})-p_k(\tilde{\Theta}_{\text{oracle}})\\
&=\frac{1}{n}\sum_{i=1}^{n}[k_{n,\tau_k}(y_i-T_i\hat{\Theta})-k_{n,\tau_k}(y_i-T_i\tilde{\Theta}_{\text{oracle}})]\\
&=-\frac{1}{n}\sum_{i=1}^{n}t_i\psi_{\tau_k}(u_i)+E\bigg[\int_0^t(I(s\leqslant u)-I(s\leqslant 0))du\bigg]+o(1)\\
&\geqslant(1-\tau_k)O_p(\eta_n)+\frac{1}{2}f_{Y|u}(T_i\Theta_{\text{oracle}}\mid u)O_p(\eta_n^2).
\end{aligned}
$$

因此, 由 $p_k(\hat{\Theta}_{\tau_k})-p_k(\tilde{\Theta}_{\text{oracle}})\leqslant 0$, 有 $\eta_n=o_p(1)$, 故

$$
\begin{aligned}
I&=\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{\tau_k,j}(u_i)-\tilde{g}_{\text{oracle},j}(u_i))^2\\
&=\frac{1}{n}(\hat{\Theta}_{\tau_k,j}-\tilde{\Theta}_{\text{oracle},j})^{\mathrm{T}}(\psi_j\psi_j^{\mathrm{T}})(\hat{\Theta}_{\tau_k,j}-\tilde{\Theta}_{\text{oracle},j})\\
&=O_p\bigg(\frac{\eta_n^2}{n}\bigg).
\end{aligned}
$$

进而 $\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{\tau_k,j}(u_i)-g_j(u_i))^2=O_p(n^{-2r/(2r+1)})$, 而

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{\text{PMWQR},j}(u_i)-g_j(u_i))^2&=\frac{1}{n}\sum_{i=1}^{n}\bigg[\sum_{k=1}^{K}w_k^*(\hat{g}_{\tau_k,j}(u_i)-g_j(u_i))\bigg]^2\\
&=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}w_k^{*2}(\hat{g}_{\tau_k,j}(u_i)-g_j(u_i))^2\\
&\quad+\frac{2}{n}\sum_{i=1}^{n}\sum_{k\neq l}w_k^*w_l^*(\hat{g}_{\tau_k,j}(u_i)-g_j(u_i))(\hat{g}_{\tau_l,j}(u_i)-g_j(u_i))\\
&=O_p(n^{-2r/(2r+1)}).
\end{aligned}
$$

定理 3.2(2) 得证. □

A weighted quantile regression approach for complex high-dimensional heterogeneous data

Wei Xiong, Han Pan, Keming Yu & Maozai Tian

Abstract With the development of digital intelligent technology, many problems arise, such as information

flooding, computing power expansion, data heterogeneity, and complexity, which bring great challenges to the theories of data modeling. In this paper, from the perspective of the mode, we propose the concept of the optimal quantile level and mode-based weighted quantile regression (MWQR) to maximize the utilization of sample information. The proposed MWQR method is superior to the existing methods in the following aspects: (1) The proposed method is suitable for complex and high-dimensional heterogeneous data, and the robustness can be ensured even when the error term is thick-tailed and skewed. (2) The MWQR method solves the problem of subjectivity in choosing quantile levels in quantile regression. (3) By assigning different weights to different quantile levels, the estimation efficiency is greatly improved and the computation time is reduced. (4) The entire conditional distribution of response variables can be investigated effectively in the MWQR method. Considering the advantages of the MWQR method, we apply it to partially linear additive models and propose two algorithms for robust coefficient estimation and variable selection, and the consistency and asymptotic distribution of estimators are also demonstrated. The numerical simulation results and empirical study of the "implicit guarantee" of urban investment bonds and plasma β-carotene concentration problems further show that the proposed method can well explore the intrinsic structure of data, significantly improves computational efficiency, and has broad application value.

Keywords **mode, optimal quantile level, weighted quantile regression, partially linear additive model, variable selection**

MSC(2020) **62G05, 62P10, 62P20**

doi: 10.1360/SSM-2022-0080