

线性回归模型中基于GMD算法的两阶段组Lasso多变点估计

安子祯, 董翠玲*

(新疆师范大学 数学科学学院, 新疆 乌鲁木齐 830017)

摘要: 利用变量选择方法估计和检测变点是目前流行且有效的方法。文章提出了一种基于GMD算法的两阶段组Lasso多变点估计方法,该方法可以同时估计出线性回归模型中多变点的位置和数量。数值模拟结果显示,与基于GMD算法未分段的组Lasso、未分段的自适应Lasso和未分段的Lasso三种变量选择算法的多变点估计方法相比,基于GMD算法的两阶段组Lasso多变点估计方法在估计精度和计算速度两方面均有显著优势。

关键词: 变量选择;组Lasso;GMD算法;线性回归模型;多变点

中图分类号: O212.1

文献标识码: A

文章编号: 1008-9659(2025)04-0001-09

变点不仅在质量控制方面有广泛应用^[1],还被应用于生物医疗^[2]、金融学^[3-4]和地震学^[5]等领域,变量选择技术是目前检测和估计多变点的流行且有效的方法。Jin等人通过引入一个特殊的设计矩阵将多变点检测与高维线性回归模型的变量选择问题联系起来,提出了利用极小极大凹惩罚(Minimax Concave Penalty, MCP)和带平滑削边绝对偏离法(Smoothly Clipped Absolute Deviation, SCAD)检测并估计了分段平稳自回归过程(Piecewise Stationary Autoregressive Process, PSAR)中的多变点^[6]。之后Jin等人又提出了一种快速且精准的两阶段变点检测方法,应用MCP、SCAD、最小绝对值收缩(Least absolute shrinkage and selection operator, Lasso)和自适应最小绝对值收缩(Adaptive Least absolute shrinkage and selection operator, Adaptive Lasso)四种变量选择方法,检测线性回归模型中的多变点^[7]。周佳琪等人将两阶段变点估计方法拓展到面板数据模型中,建立了空间网络自回归模型^[8]。Sun等人针对广义线性回归模型提出了一种两阶段多变点估计方法,并证明了估计量的相合性^[9]。吕丽等人基于两阶段多变点估计方法,应用随机加权自助法结合高斯混合模型得到线性回归模型中多变点的置信区间^[10]。Wang等人针对多阈值比例风险模型(Multi-threshold Proportional Hazards Model)提出了一种两阶段多变点估计方法,通过数值模拟和实证分析验证了该方法的有效性和可行性^[11]。李美琪等人通过分割数据建立相依观测数据的高维线性回归模型,将多变点估计问题转化为变量选择问题,得到了多变点的相合估计量^[5]。

组最小绝对值收缩(Group Least absolute shrinkage and selection operator, Group Lasso)最早是由Yuan等人在Lasso的基础上提出的,用于组变量的选择^[12]。Yang等人提出了分组主要下降(Groupwise Majorization Descent, GMD)算法,是一种快速求解组Lasso的算法,该算法在求解过程中不要求组间变量正交^[13]。李傲梅运用基于GMD算法的自适应组Lasso来求解多元线性回归模型的变点问题,并通过数值模拟和实证分析验证了该方法的有效性^[14]。文章将组Lasso应用于两阶段多变点检测方法中变量选择的阶段,并利用GMD算法求解组Lasso。数值模拟结果表明,基于GMD算法的两阶段组Lasso多变点估计方法相较于基于GMD算法

[收稿日期] 2024-07-05

[修回日期] 2024-09-29

[基金项目] 新疆维吾尔自治区自然科学基金项目(2023D01A37)。

[作者简介] 安子祯(1999-),女,硕士研究生,主要从事多变点的统计推断方面研究,E-mail:1311175702@qq.com。

*[通讯作者] 董翠玲(1978-),女,副教授,主要从事多变点的统计推断、面板数据分析方面研究,E-mail:75894804@qq.com。

的未分段的组 Lasso、未分段的自适应 Lasso 和未分段的 Lasso 对线性回归模型中多变点的估计在精确度和计算速度上均有显著优势。

1 基于 GMD 算法求解组 Lasso 的两阶段多变点估计

应用两阶段多变点估计方法,第一阶段首先对数据进行分割,引入一个特殊的设计矩阵,将线性回归模型的变点检测问题转化为高维线性回归模型的变量选择问题;其次利用组 Lasso 对高维线性回归模型进行变量选择,并应用 GMD 算法求解。第二阶段基于拟似然比检验,确定分割数据中准确的变点位置。

1.1 含有多变点的线性回归模型

Jin 等人考虑一个含有 s 个变点的线性回归模型,其中 $s \geq 2, 1 < a_1 < \dots < a_s < n$, 模型如下^[7]

$$\begin{aligned} y_i &= \sum_{j=1}^q x_{i,j} \beta_{j,0} + \sum_{l=1}^s \sum_{j=1}^q x_{i,j} \delta_{j,l} I(a_l < i \leq n) + \varepsilon_i \\ &= \mathbf{x}_i^T \left[\boldsymbol{\beta}_0 + \sum_{l=1}^s \boldsymbol{\delta}_l I(a_l < i \leq n) \right] + \varepsilon_i, \quad i = 1, \dots, n \\ &= \begin{cases} \mathbf{x}_i^T \boldsymbol{\beta}_0 + \varepsilon_i, & 1 \leq i \leq a_1 \\ \mathbf{x}_i^T (\boldsymbol{\beta}_0 + \boldsymbol{\delta}_1) + \varepsilon_i, & a_1 < i \leq a_2 \\ \dots & \dots \\ \mathbf{x}_i^T \left(\boldsymbol{\beta}_0 + \sum_{l=1}^s \boldsymbol{\delta}_l \right) + \varepsilon_i, & a_s < i \leq n \end{cases} \end{aligned} \quad (1)$$

其中, $\{\mathbf{y} = (y_1, \dots, y_n)^T\}$ 是 n 维观测值, $\{\mathbf{x}_i = (x_{i,1}, \dots, x_{i,q})^T\}$ 是 q 维解释变量序列, $\boldsymbol{\beta}_0 = (\beta_{1,0}, \dots, \beta_{q,0})^T \neq \mathbf{0}$ 是 q 维回归系数, s 是变点数量, $a_1, \dots, a_s$ 是变点位置, $\boldsymbol{\delta}_l = (\delta_{1,l}, \dots, \delta_{q,l})^T, (l = 1, \dots, s)$ 是变点处回归系数的增量, 且 $\boldsymbol{\beta}_0, s, a_j, \boldsymbol{\delta}_l$ 都是未知的, $\varepsilon_1, \dots, \varepsilon_n$ 是随机误差。

先将观测数据划分为 $p_n + 1$ 段, 其中 $p_n + 1 > s$, 第一段的长度为 $n - p_n m$, 其他所有段的长度都为 m , $m = \lfloor n / (p_n + 1) \rfloor$ (这里 $\lfloor c \rfloor$ 满足 $0 \leq c - \lfloor c \rfloor < 1$), 记第一段观测点的指标集为 $I_1 = \{1, \dots, n - p_n m\}$, 第 j 段观测点的指标集为 $I_j = \{n - (p_n - j + 2)m + 1, \dots, n - (p_n - j + 1)m\}, (j = 2, \dots, p_n + 1)$ 。

令 $\mathbf{X}$ 是一个矩阵

$$\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(p_n+1)})_{n \times q(p_n+1)} = \begin{pmatrix} \mathbf{X}_{(1)} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{X}_{(2)} & \mathbf{X}_{(2)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{X}_{(3)} & \mathbf{X}_{(3)} & \mathbf{X}_{(3)} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{X}_{(p_n+1)} & \mathbf{X}_{(p_n+1)} & \mathbf{X}_{(p_n+1)} & \dots & \mathbf{X}_{(p_n+1)} \end{pmatrix} \quad (2)$$

其中, $\mathbf{X}^{(1)} = (\mathbf{X}_{(1)}^T, \mathbf{X}_{(2)}^T, \dots, \mathbf{X}_{(p_n+1)}^T)^T$, $\mathbf{X}^{(j)} = (\mathbf{0}_{q \times (n - (p_n - j + 2)m)}, \mathbf{X}_{(j)}^T, \mathbf{X}_{(j+1)}^T, \dots, \mathbf{X}_{(p_n+1)}^T)^T$, $\mathbf{X}_{(1)} = (\mathbf{x}_1, \dots, \mathbf{x}_{n - p_n m})^T$ 为序列 $\{\mathbf{x}_i, i = 1, 2, \dots, n\}$ 第 1 个子段的解释变量, $\mathbf{X}_{(j)} = (\mathbf{x}_{n - (p_n - j + 2)m + 1}, \dots, \mathbf{x}_{n - (p_n - j + 1)m})^T$ 为第 j 个子段的解释变量, $(j = 2, 3, \dots, p_n + 1)$, 这样原来 $n \times q$ 维的自变量矩阵被拓展为 $n \times q(p_n + 1)$ 维。假设每个子段中至多有一个变点 (At Most One Change-point, AMOC), 若变点 a_j 位于第 k_j 个子段中, 即 $a_j \in I_{k_j} = \{n - (p_n - k_j + 2)m + 1, \dots, n - (p_n - k_j + 1)m\}, (j = 1, 2, \dots, s)$ 。记 $A_n = \{1, k_1, k_1 + 1, \dots, k_s, k_s + 1\}$, 即 A_n 表示由第一个子段和变点 a_j 所在的子段以及其后一子段的指标集构成的集合。

基于以上数据序列的分割及下三角矩阵 (2) 的构造, 模型 (1) 的矩阵形式为

$$\mathbf{y} = \mathbf{X} \boldsymbol{\theta}^* + \mathbf{X}_\omega \vec{\omega} + \boldsymbol{\varepsilon} \quad (3)$$

其中, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, $\mathbf{X}$ 是由 (2) 式定义的下三角矩阵, $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \dots, \boldsymbol{\theta}_{p_n+1}^T)^T$, 关于 $\mathbf{X}_\omega, \vec{\omega}, \boldsymbol{\theta}^*$ 的定义以及模型 (3) 的推导详见文献 [7]。这样, 模型 (1) 中多变点的估计问题就转化为高维线性回归模型 (3) 的变量选择问题。下面使用 GMD 算法求解组 Lasso 进行变量选择。

1.2 应用 GMD 算法求解组 Lasso 进行变量选择

Yuan 等人提出的组 Lasso 将高维线性回归模型的系数分为互不重复的 K 组, 在组水平上对系数进行变

量选择,提高了计算速度和组间稀疏性^[12]。结合文章模型,将 $q(p_n + 1)$ 维向量 θ^* 中的元素分成 $p_n + 1$ 组,即 $K = p_n + 1$ 。由文献[12]和[13]知,基于GMD算法的高维线性回归模型(3)中回归系数 θ^* 的组Lasso估计量为

$$\hat{\theta}^* = \arg \min_{\theta} \left\{ \frac{1}{n} \|y - X\theta\|_2^2 + \frac{\lambda_n}{n} \sum_{k=1}^K \omega_k \|\theta^{*(k)}\|_2 \right\} \quad (4)$$

其中, $\lambda_n \geq 0$ 是正则化参数(可利用交叉验证法选择), $\omega_k = \sqrt{q_k}$,这里 q_k 表示第 k 组中元素的个数, $\theta^{*(k)}$ 表示 θ^* 中对应于第 k 组的系数,且 $\|\theta^{*(k)}\|_2 = \sum_{i=1}^q \|\theta_{ki}\|_2$ ($k = 1, 2, \dots, K$)。

1.3 变点的估计

令 $\hat{\theta}^* = (\hat{\theta}_1^T, \hat{\theta}_2^T, \dots, \hat{\theta}_{p_n+1}^T)^T$,其中 $\hat{\theta}_j = (\hat{\theta}_{j1}, \dots, \hat{\theta}_{jq})^T$, ($j = 1, \dots, p_n + 1$)。记 $\hat{A}_n = \{j: \hat{\theta}_j \neq 0, j = 1, \dots, p_n + 1\}$ 为系数不为零的子段指标集构成的集合,则 $\hat{A}_n^* = \{j: j \in \hat{A}_n, j-1 \notin \hat{A}_n, j = 2, \dots, p_n + 1\} = \{\hat{k}_1, \dots, \hat{k}_s\}$ 为变点可能存在的子段指标集构成的集合($\hat{k}_1 < \dots < \hat{k}_s$)。为了得到变点 $\{a_j\}$ 的估计值 $\{\hat{a}_j\}$,应用文献[15]中的定理3.1.1,利用拟似然比检验去检测 $\hat{I}_{(j)} = I_{\hat{k}_{j-1}} \cup I_{\hat{k}_j} = \{n - (p_n - \hat{k}_j + 3)m + 1, \dots, n - (p_n - \hat{k}_j + 1)m\}$, ($j = 1, \dots, \hat{s}$)内是否存在变点,具体如下:

考虑以下模型,其可能的变点位置为 ζ_j , ($j = 1, \dots, \hat{s}$)

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_j I(n_j^{(l)} \leq i \leq \zeta_j) + \mathbf{x}_i^T \boldsymbol{\beta}_{j+1} I(\zeta_j < i \leq n_j^{(r)}) + \varepsilon_i \quad (5)$$

其中, $n_j^{(l)} = n - (p_n - \hat{k}_j + 3)m + 1$, $n_j^{(r)} = n - (p_n - \hat{k}_j + 1)m$, $\boldsymbol{\beta}_j$ 和 $\boldsymbol{\beta}_{j+1}$ 是未知的 q 维回归系数, ($i = n_j + 1, \dots, n_{j+1}$)。区间 $\hat{I}_{(j)} = I_{\hat{k}_{j-1}} \cup I_{\hat{k}_j}$, ($j = 1, \dots, \hat{s}$)中是否存在变点等价于假设检验问题 $H_0: \boldsymbol{\beta}_j = \boldsymbol{\beta}_{j+1} \leftrightarrow H_1: \boldsymbol{\beta}_j \neq \boldsymbol{\beta}_{j+1}$ 。

记 $N_j = n_j^{(r)} - n_j^{(l)}$ 表示区间 $\hat{I}_{(j)} = I_{\hat{k}_{j-1}} \cup I_{\hat{k}_j}$ 的长度, $b_j = (2 \log \log N_j + q(\log \log \log N_j)/2 - \log \Gamma(q/2))^2 / 2 \log \log N_j$, $c_j = (b_j / (2 \log \log N_j))^{1/2}$, $\hat{\sigma}_j^2 = \min_{\boldsymbol{\beta}} \sum_{i=n_j^{(l)}}^{n_j^{(r)}} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$ 为回归模型中最小残差平方和的估计值,其中 $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$ ($x > 0$)是Gamma函数。由文献[15]中的引理3.1.9可知,模型(5)中的回归系数是否存在变点的拟似然比检验统计量为

$$T_j = N_j \left(\hat{\sigma}_j^2 - \min_{\boldsymbol{\beta}} \sum_{i=n_j^{(l)}}^{\zeta_j} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 - \min_{\boldsymbol{\beta}} \sum_{i=\zeta_j+1}^{n_j^{(r)}} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right) / \hat{\sigma}_j^2 \quad (6)$$

且当 $T_j > b_j + 2c_j \log(-2/\log(1-\alpha))$ 拒绝原假设 H_0 ,说明模型(5)中的回归系数存在一个变点 $\zeta_j \in \{n_j^{(l)}, \dots, n_j^{(r)}\}$ 使得 $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}_{j+1}$,其中 $1-\alpha$ 是置信水平。变点 ζ_j 的相合估计量为

$$\hat{\zeta}_j = \arg \min_{n_j^{(l)} + q < k < n_j^{(r)} - q} \left[\min_{\boldsymbol{\beta}} \sum_{i=n_j^{(l)}}^k (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \min_{\boldsymbol{\beta}} \sum_{i=k+1}^{n_j^{(r)}} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right]$$

若接受原假设 H_0 ,则删除第 j 段。记所有能够检测出的变点为 $\tilde{\zeta}_1 < \dots < \tilde{\zeta}_{\tilde{s}}$,由于每个 $\tilde{\zeta}_j$, ($j = 1, \dots, \tilde{s}$)仅用了 $2m$ 个观测值得到,为了避免“过度估计”,提高变点估计的精准度,令 $\tilde{n}_0 = 0$, $\tilde{n}_j = \tilde{\zeta}_j$, ($j = 1, \dots, \tilde{s}$), $\tilde{n}_{\tilde{s}+1} = n$,再次使用拟似然比检验式(6)。检测区间 $\{\tilde{n}_{j-1} + 1, \dots, \tilde{n}_{j+1}\}$, ($j = 1, \dots, \tilde{s}$)中是否存在变点。若存在变点,变点估计量为

$$\hat{a}_j = \arg \min_{\tilde{n}_{j-1} + q < k < \tilde{n}_{j+1} - q} \left[\min_{\boldsymbol{\beta}} \sum_{i=\tilde{n}_{j-1}+1}^k (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \min_{\boldsymbol{\beta}} \sum_{i=k+1}^{\tilde{n}_{j+1}} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right], (j = 1, \dots, \tilde{s}) \quad (7)$$

1.4 变点估计的相合性

假设 1 如果 $s \geq 1$,对于 $1 \leq j \leq s$, $a_j/n \rightarrow \tau_j > 0$,进一步地,当 $s \geq 2$ 时,有 $\min_{1 \leq j \leq s-1} (\tau_{j+1} - \tau_j) > 0$ 。

假设 2 当 $t_2 - t_1 \rightarrow \infty$ 时, $\sum_{i=t_1}^{t_2} \frac{\mathbf{x}_i \mathbf{x}_i^T}{t_2 - t_1} \rightarrow W$,其中 W 是一个严格正定矩阵。

定理 1 假设当 $n \rightarrow \infty$ 时, $m \rightarrow \infty$, $q \rightarrow \infty$, $m/n \rightarrow 0$, $\lambda_n/n \rightarrow 0$,并且 $\lambda_n \omega_k m/n \rightarrow \infty$,如果假设1和假设2成立,模型(1)中变点数量的估计值为 $\hat{s}$,则有

- (1) $\lim_{n \rightarrow \infty} P(\hat{s} = s) = 1;$
 (2) $\lim_{n \rightarrow \infty} P(a_j \in \hat{I}_{(j)}, j = 1, 2, \dots, s | \hat{s} = s) = 1.$

注 假设1和假设2为文献[7]中关于高维线性回归模型(3)的假设,其中假设1说明一个子段至多有一个变点,假设2要求 $\mathbf{x}_i, (i = 1, 2, \dots, n)$ 是独立同分布的随机向量,其协方差矩阵为 W . 定理1中的结论(1)说明变点估计的数量是相合的,结论(2)说明当估计的变点数量相合时,真实变点 a_j 位于 $\hat{I}_{(j)}$ 的概率为1,定理1的证明与文献[7]中定理1的证明类似。

2 数值模拟

运用数值模拟的方法,在100次独立重复试验下,将基于GMD算法的两阶段组Lasso多变点估计方法与基于GMD算法的未分段的组Lasso、未分段的自适应Lasso和未分段的Lasso分别从计算时间和估计精度两个方面进行比较。其中,利用GMD算法求解组Lasso用R中的gglasso包实现,求解自适应Lasso和Lasso用R中glmnet包自带的循环坐标下降算法实现,种子设置为123.所有的模拟都是在华硕 Vivobook 15 X1502ZA 计算机上实现,其配置为12核16线程 Intel(R)Core(TM)i5-12500H(2.50GHz)和512G内存。

2.1 小样本情况下的变点估计

考虑一个含有2个变点的线性回归模型,变量个数为4,样本量为100,真实变点位置为 $1 < a_1 = 30 < a_2 = 70 < n$,即线性回归模型形式如下

$$y_i = \begin{cases} \mathbf{x}_i^T \boldsymbol{\beta}_0 + \varepsilon_i, & 1 \leq i < 30 \\ \mathbf{x}_i^T (\boldsymbol{\beta}_0 + \boldsymbol{\delta}_1) + \varepsilon_i, & 30 \leq i < 70 \\ \mathbf{x}_i^T (\boldsymbol{\beta}_0 + \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2) + \varepsilon_i, & 70 \leq i \leq 100 \end{cases} \quad (8)$$

其中, $\boldsymbol{\beta}_0 = (\beta_1, \beta_2, \beta_3, \beta_4) = (-2, 3, 0, 8)$, $\boldsymbol{\delta}_1 = (\delta_{1,1}, \delta_{1,2}, \delta_{1,3}, \delta_{1,4}) = (7, -3, 0, -8)$, $\boldsymbol{\delta}_2 = (\delta_{2,1}, \delta_{2,2}, \delta_{2,3}, \delta_{2,4}) = (-5, 5, -2, 0)$. 设 $\{\mathbf{x}_i^T\}$ 和 $\{\varepsilon_i\}$ 的分布为以下三种情况:(1) $\{\mathbf{x}_i^T\}$ 和 $\{\varepsilon_i\}$ 独立同分布(IID),都服从标准正态分布 $N(0, 1)$; (2) $\{\mathbf{x}_i^T\}$ 独立同分布,服从自由度为5的 t 分布 $t(5)$, $\{\varepsilon_i\}$ 独立同分布,服从正态分布 $N(0, 1.5^2)$; (3) $\{\mathbf{x}_i^T\}$ 和 $\{\varepsilon_i\}$ 独立同分布,且都服从自由度为6的 t 分布 $t(6)$. 数值模拟过程中每个子段的长度 $m_l = \lfloor \kappa_l \sqrt{n} \rfloor, (l = 1, 2, \dots, L)$, 其中 κ_l 取区间内 L 个网格点的值,例如 $[0.1, 5]$, 这里 κ_l 取值为1.

表1给出了模型(8)在以上三种情况下变点数量的估计结果,其中变点数量估计值的均值表示100次独立重复试验得到的变点数量估计值的均值,变点数量估计的准确度表示100次独立重复试验得到的变点数量的估计值和真实值相等的比例。

表1 变点数量估计结果

$\{\mathbf{x}_i^T\}$ 和 $\{\varepsilon_i\}$ 的 分布情况	$\{\mathbf{x}_i^T\} \sim \text{IID } N(0, 1)$ $\{\varepsilon_i\} \sim \text{IID } N(0, 1)$	$\{\mathbf{x}_i^T\} \sim \text{IID } t(5)$ $\{\varepsilon_i\} \sim \text{IID } N(0, 1.5^2)$	$\{\mathbf{x}_i^T\} \sim \text{IID } t(6)$ $\{\varepsilon_i\} \sim \text{IID } t(6)$
基于GMD算法的两阶段组Lasso(用R中gglasso包实现)			
变点数量估计值的均值	2.0200	2.0100	2.0000
变点数量估计的准确度	0.9800	0.9700	0.9600
基于GMD算法的未分段的组Lasso(用R中gglasso包实现)			
变点数量估计值的均值	2.3400	2.5200	2.4600
变点数量估计的准确度	0.6900	0.5800	0.5700
未分段的自适应Lasso(用R中glmnet包实现)			
变点数量估计值的均值	2.9200	2.9300	2.9000
变点数量估计的准确度	0.1300	0.1800	0.1800
未分段的Lasso(用R中glmnet包实现)			
变点数量估计值的均值	1.7600	1.9300	1.9100
变点数量估计的准确度	0.3500	0.2600	0.3400

表2给出了100次独立重复试验变点位置的估计结果,其中, $\{\tau_i = \frac{a_i}{n}, i = 1, 2\}$ 表示真实变点位置, $\{\hat{\tau}_i, i = 1, 2\}$ 表示在100次试验中,当变点数量的估计值和真实值一致时,变点位置估计结果的平均值,括号里是对应的均方误差。

表2 变点位置估计结果

$\{\mathbf{x}_i^T\}$ 和 $\{\varepsilon_i\}$ 的 分布情况	$\{\mathbf{x}_i^T\} \sim IID N(0, 1)$ $\{\varepsilon_i\} \sim IID N(0, 1)$	$\{\mathbf{x}_i^T\} \sim IID t(5)$ $\{\varepsilon_i\} \sim IID N(0, 1.5^2)$	$\{\mathbf{x}_i^T\} \sim IID t(6)$ $\{\varepsilon_i\} \sim IID t(6)$
基于GMD算法的两阶段组Lasso(用R中gglasso包实现)			
$\tau_1=0.3$	$\hat{\tau}_1=0.29939(1.2245 \times 10^{-5})$	$\hat{\tau}_1=0.30021(2.6804 \times 10^{-5})$	$\hat{\tau}_1=0.30063(1.2500 \times 10^{-5})$
$\tau_2=0.7$	$\hat{\tau}_2=0.69949(4.5918 \times 10^{-5})$	$\hat{\tau}_2=0.69948(1.2887 \times 10^{-4})$	$\hat{\tau}_2=0.70115(3.4375 \times 10^{-5})$
基于GMD算法的未分段的组Lasso(用R中gglasso包实现)			
$\tau_1=0.3$	$\hat{\tau}_1=0.31957(7.0870 \times 10^{-4})$	$\hat{\tau}_1=0.32172(8.5517 \times 10^{-4})$	$\hat{\tau}_1=0.32474(1.0544 \times 10^{-3})$
$\tau_2=0.7$	$\hat{\tau}_2=0.71986(6.0725 \times 10^{-4})$	$\hat{\tau}_2=0.72138(7.4483 \times 10^{-4})$	$\hat{\tau}_2=0.71544(6.0000 \times 10^{-4})$
未分段的自适应Lasso(用R中glmnet包实现)			
$\tau_1=0.3$	$\hat{\tau}_1=0.32692(1.0077 \times 10^{-3})$	$\hat{\tau}_1=0.33222(1.6222 \times 10^{-3})$	$\hat{\tau}_1=0.32444(9.5556 \times 10^{-4})$
$\tau_2=0.7$	$\hat{\tau}_2=0.73308(1.3000 \times 10^{-3})$	$\hat{\tau}_2=0.74611(2.6611 \times 10^{-3})$	$\hat{\tau}_2=0.72389(7.3889 \times 10^{-4})$
未分段的Lasso(用R中glmnet包实现)			
$\tau_1=0.3$	$\hat{\tau}_1=0.38514(4.3743 \times 10^{-2})$	$\hat{\tau}_1=0.40962(4.0588 \times 10^{-2})$	$\hat{\tau}_1=0.44088(4.7268 \times 10^{-2})$
$\tau_2=0.7$	$\hat{\tau}_2=0.98000(7.8594 \times 10^{-2})$	$\hat{\tau}_2=0.98154(7.9415 \times 10^{-2})$	$\hat{\tau}_2=0.98029(7.8797 \times 10^{-2})$

由表1和表2可以看出,在 $n = 100$ 的情况下,基于GMD算法的两阶段组Lasso对变点数量估计的准确率在96%~98%之间,变点位置的估计误差大多保持在 10^{-5} 的水平,而基于GMD算法的未分段的组Lasso和未分段的自适应Lasso均存在过度估计,即估计变点数量大于真实变点数量。

表3给出了在 $n = 100$ 的情况下四种用于多变点估计的变量选择方法在100次独立重复试验中运行时间的均值。

表3 运行时间对比

单位:秒

模拟方法	$\{\mathbf{x}_i^T\} \sim IID N(0, 1)$ $\{\varepsilon_i\} \sim IID N(0, 1)$	$\{\mathbf{x}_i^T\} \sim IID t(5)$ $\{\varepsilon_i\} \sim IID N(0, 1.5^2)$	$\{\mathbf{x}_i^T\} \sim IID t(6)$ $\{\varepsilon_i\} \sim IID t(6)$
基于GMD算法的两阶段组Lasso	0.0898	0.1202	0.1204
基于GMD算法的未分段的组Lasso	1.2456	2.0039	1.9274
未分段的自适应Lasso	0.1596	0.1588	0.1651
未分段的Lasso	0.1084	0.1124	0.1167

从表3可以看出,在 $n = 100$ 的情况下,相较于其他三种方法,基于GMD算法的两阶段组Lasso方法估计变点的运算速度较快,基于GMD算法的未分段的组Lasso方法估计变点的运算速度最慢。

2.2 大样本情况下的变点估计

考虑一种含有3个变点的线性回归模型,变量个数为10,样本量为1000,真实变点位置为 $1 < a_1 = 200 < a_2 = 500 < a_3 = 700 < n$,即线性回归模型形式如下

$$y_i = \begin{cases} \mathbf{x}_i^T \boldsymbol{\beta}_0 + \varepsilon_i, & 1 \leq i < 200 \\ \mathbf{x}_i^T (\boldsymbol{\beta}_0 + \boldsymbol{\delta}_1) + \varepsilon_i, & 200 \leq i < 500 \\ \mathbf{x}_i^T (\boldsymbol{\beta}_0 + \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2) + \varepsilon_i, & 500 \leq i < 700 \\ \mathbf{x}_i^T (\boldsymbol{\beta}_0 + \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2 + \boldsymbol{\delta}_3) + \varepsilon_i, & 700 \leq i \leq 1000 \end{cases} \quad (9)$$

其中, $\beta_0 = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}) = (-2, 3, 0, 8, 0, 0, -5, 0, 0, 2)$

$$\delta_1 = (\delta_{1,1}, \delta_{1,2}, \delta_{1,3}, \delta_{1,4}, \delta_{1,5}, \delta_{1,6}, \delta_{1,7}, \delta_{1,8}, \delta_{1,9}, \delta_{1,10}) = (7, -3, 0, -8, 1, 3, 5, 0, 7, -5)$$

$$\delta_2 = (\delta_{2,1}, \delta_{2,2}, \delta_{2,3}, \delta_{2,4}, \delta_{2,5}, \delta_{2,6}, \delta_{2,7}, \delta_{2,8}, \delta_{2,9}, \delta_{2,10}) = (-5, 5, -2, 0, -1, -3, 3, -5, -6, 3)$$

$$\delta_3 = (\delta_{3,1}, \delta_{3,2}, \delta_{3,3}, \delta_{3,4}, \delta_{3,5}, \delta_{3,6}, \delta_{3,7}, \delta_{3,8}, \delta_{3,9}, \delta_{3,10}) = (0, -4, 2, 3, -5, 0, -3, 7, -1, 1).$$

$\{x_i^T\}$ 和 $\{\varepsilon_i\}$ 的分布仍然取模型(8)的三种情况, κ_i 取值为2.5. 同表1~表3, 表4~表6给出了 $n = 1000$ 的大样本情况下 100 次独立重复试验中变点数量估计结果、变点位置估计结果和运行时间对比结果。

表4 变点数量估计结果

$\{x_i^T\}$ 和 $\{\varepsilon_i\}$ 的 分布情况	$\{x_i^T\} \sim IID N(0, 1)$ $\{\varepsilon_i\} \sim IID N(0, 1)$	$\{x_i^T\} \sim IID t(5)$ $\{\varepsilon_i\} \sim IID N(0, 1.5^2)$	$\{x_i^T\} \sim IID t(6)$ $\{\varepsilon_i\} \sim IID t(6)$
基于GMD算法的两阶段组 Lasso(用 R 中 gglasso 包实现)			
变点数量估计值的均值	3.0000	3.0000	3.0000
变点数量估计的准确度	1.0000	1.0000	1.0000
基于GMD算法的未分段的组 Lasso(用 R 中 gglasso 包实现)			
变点数量估计值的均值	3.3500	3.3600	3.2800
变点数量估计的准确度	0.6500	0.6400	0.7200
未分段的自适应 Lasso(用 R 中 glmnet 包实现)			
变点数量估计值的均值	4.2400	4.5100	4.2100
变点数量估计的准确度	0.1300	0.0500	0.1500
未分段的 Lasso(用 R 中 glmnet 包实现)			
变点数量估计值的均值	26.8400	27.6100	25.9000
变点数量估计的准确度	0.0000	0.0000	0.0000

由于表4中未分段的 Lasso 对变点数量估计的准确率为0%, 故表5中不讨论应用未分段的 Lasso 进行变点位置估计。

表5 变点位置估计结果

$\{x_i^T\}$ 和 $\{\varepsilon_i\}$ 的 分布情况	$\{x_i^T\} \sim IID N(0, 1)$ $\{\varepsilon_i\} \sim IID N(0, 1)$	$\{x_i^T\} \sim IID t(5)$ $\{\varepsilon_i\} \sim IID N(0, 1.5^2)$	$\{x_i^T\} \sim IID t(6)$ $\{\varepsilon_i\} \sim IID t(6)$
基于GMD算法的两阶段组 Lasso(用 R 中 gglasso 包实现)			
$\tau_1=0.2$	$\hat{\tau}_1 = 0.19999 (1.1000 \times 10^{-7})$	$\hat{\tau}_1 = 0.19994 (1.0000 \times 10^{-7})$	$\hat{\tau}_1 = 0.20000 (1.8000 \times 10^{-7})$
$\tau_2=0.5$	$\hat{\tau}_2 = 0.50001 (1.1000 \times 10^{-7})$	$\hat{\tau}_2 = 0.49996 (2.0000 \times 10^{-7})$	$\hat{\tau}_2 = 0.49999 (1.3000 \times 10^{-7})$
$\tau_3=0.7$	$\hat{\tau}_3 = 0.70006 (8.0000 \times 10^{-8})$	$\hat{\tau}_3 = 0.70002 (1.6000 \times 10^{-7})$	$\hat{\tau}_3 = 0.70003 (1.3000 \times 10^{-7})$
基于GMD算法的未分段的组 Lasso(用 R 中 gglasso 包实现)			
$\tau_1=0.2$	$\hat{\tau}_1 = 0.20189 (6.5385 \times 10^{-6})$	$\hat{\tau}_1 = 0.20173 (5.1094 \times 10^{-6})$	$\hat{\tau}_1 = 0.20161 (3.8889 \times 10^{-6})$
$\tau_2=0.5$	$\hat{\tau}_2 = 0.50177 (5.1538 \times 10^{-6})$	$\hat{\tau}_2 = 0.50214 (9.0156 \times 10^{-6})$	$\hat{\tau}_2 = 0.50193 (7.2639 \times 10^{-6})$
$\tau_3=0.7$	$\hat{\tau}_3 = 0.70197 (7.2615 \times 10^{-6})$	$\hat{\tau}_3 = 0.70252 (1.2234 \times 10^{-5})$	$\hat{\tau}_3 = 0.70214 (8.0000 \times 10^{-6})$
未分段的自适应 Lasso(用 R 中 glmnet 包实现)			
$\tau_1=0.2$	$\hat{\tau}_1 = 0.20585 (4.3385 \times 10^{-5})$	$\hat{\tau}_1 = 0.20500 (3.0600 \times 10^{-5})$	$\hat{\tau}_1 = 0.20500 (3.7667 \times 10^{-5})$
$\tau_2=0.5$	$\hat{\tau}_2 = 0.50754 (6.6308 \times 10^{-5})$	$\hat{\tau}_2 = 0.50420 (2.0600 \times 10^{-5})$	$\hat{\tau}_2 = 0.50607 (5.1000 \times 10^{-5})$
$\tau_3=0.7$	$\hat{\tau}_3 = 0.70577 (4.1000 \times 10^{-5})$	$\hat{\tau}_3 = 0.70820 (7.8600 \times 10^{-5})$	$\hat{\tau}_3 = 0.70447 (2.8333 \times 10^{-5})$

由表4和表5可以看出, 在 $n = 1000$ 的大样本情况下, 基于GMD算法的两阶段组 Lasso 方法对变点数量估计的准确率高达100%, 变点位置估计误差大多保持在 10^{-7} 的水平; 基于GMD算法的未分段的组 Lasso、未分段的自适应 Lasso 和未分段的 Lasso 均存在过度估计。

表6 运行时间对比

单位:秒

模拟方法	$\{\mathbf{x}_i^T\} \sim \text{IID } N(0, 1)$	$\{\mathbf{x}_i^T\} \sim \text{IID } t(5)$	$\{\mathbf{x}_i^T\} \sim \text{IID } t(6)$
	$\{\boldsymbol{\varepsilon}_i\} \sim \text{IID } N(0, 1)$	$\{\boldsymbol{\varepsilon}_i\} \sim \text{IID } N(0, 1.5^2)$	$\{\boldsymbol{\varepsilon}_i\} \sim \text{IID } t(6)$
基于GMD算法的两阶段组Lasso	0.8889	1.0151	0.9989
基于GMD算法的未分段的组Lasso	96.9691	150.9882	129.6870
未分段的自适应Lasso	16.5532	16.5312	17.1622
未分段的Lasso	14.5293	14.4340	15.0603

综合以上可以看出,在 $n = 100$ 和 $n = 1000$ 的情况下,基于GMD算法的两阶段组Lasso方法都能够以较高的准确率估计出变点的数量及位置,原因在于该方法在给数据序列分段后加入了拟似然比检验,提高了估计的精度。基于GMD算法的未分段组Lasso的估计效果一般,未分段的自适应Lasso和未分段的Lasso对变点位置及数量的估计效果较差,在大样本情况下体现更加明显,故高维线性回归模型的变量选择更适合使用组Lasso方法。

在运算时间上,相较于其他三种未分段的方法,基于GMD算法的两阶段组Lasso方法估计变点的运算速度具有明显优势,这是由于该方法将 $q(p_n + 1)$ 个变量划分为 $(p_n + 1)$ 组,而未分段的方法是将 nq 个变量划分为 n 组,其中 $(p_n + 1) < n$,所以运算速度得到了较大提高。表3中的小样本未分段的情况下解释变量矩阵维度为 100×400 ,而表6中的大样本情况时解释变量矩阵维度扩展到了 1000×10000 ,故运行时间有相应的增加。

此外,通过表1至表6的数值模拟结果可以看出随机变量 $\{\mathbf{x}_i^T\}$ 和误差项 $\{\boldsymbol{\varepsilon}_i\}$ 的分布不会对变点估计结果产生显著影响。

3 实证分析

文章选取2022年1月3日至2024年11月21日的黄金期货收盘价日数据(单位:美元/金衡盎司)、美元兑换人民币(USD/CNY)汇率收盘价日数据、WTI原油期货收盘价日数据(单位:美元/桶)(数据均来源于<https://cn.investing.com>),均有729个数据。

由于线性回归模型的解释性强、应用性广,应用上述729个数据建立黄金期货收盘价(y)与美元兑换人民币(USD/CNY)汇率(x_1)和WTI原油期货收盘价(x_2)的线性回归模型为

$$y = 178.254x_1 - 7.263x_2 + 1383.103 \quad (10)$$

应用基于GMD算法的两阶段组Lasso方法对模型(10)中的回归系数进行多变点检测与估计(取 $\kappa_1=2$),由式(7)得到变点位置为194、311、485和616,对应的时间分别为2022年9月30日、2023年3月16日、2023年11月24日和2024年6月14日。上述变点将整个观测数据分割为5个子段(第一子段为1~193,第二子段为194~310,第三子段为311~484,第四子段为485~615,第五子段为616~729)。对每个子段应用R软件分别建立线性回归模型,得到以下分段线性回归模型

$$y = \begin{cases} -282.7797x_1 + 3.3315x_2 + 3364.9449 & 1 \leq t < 194 \\ -475.3796x_1 - 1.4439x_2 + 5234.0312 & 194 \leq t < 311 \\ -162.2254x_1 + 3111.5325 & 311 \leq t < 485 \\ 2608.071x_1 + 6.758x_2 - 17120.710 & 485 \leq t < 616 \\ -566.72x_1 - 14.50x_2 + 7669.87 & 616 \leq t \leq 729 \end{cases} \quad (11)$$

图1给出了2022年1月3日至2024年11月21日标准化美元兑人民币(USD/CNY)汇率、黄金期货收盘价和WTI原油期货收盘价的时间序列图以及变点位置(竖直虚线所示),图中箭头表示每个子段中变量的走势。图中第一个变点(2022年9月30日)出现的原因是WTI原油期货收盘价(x_2)与黄金期货收盘价(y)的关系由正相关转变为负相关;第二个变点(2023年3月16日)出现的原因是WTI原油期货收盘价(x_2)与黄金期货收盘价(y)的关系由负相关转变为没有显著关系;第三个变点(2023年11月24日)出现的原因是WTI原油期货收盘价(x_2)与黄金期货收盘价(y)的关系由不显著转变为正相关;第四个变点(2024年6月14日)出现的

原因是美元兑换人民币(USD/CNY)汇率(x_1)与黄金期货收盘价(y)的关系由正相关转变为负相关,这与模型(11)中回归系数的符号一致。

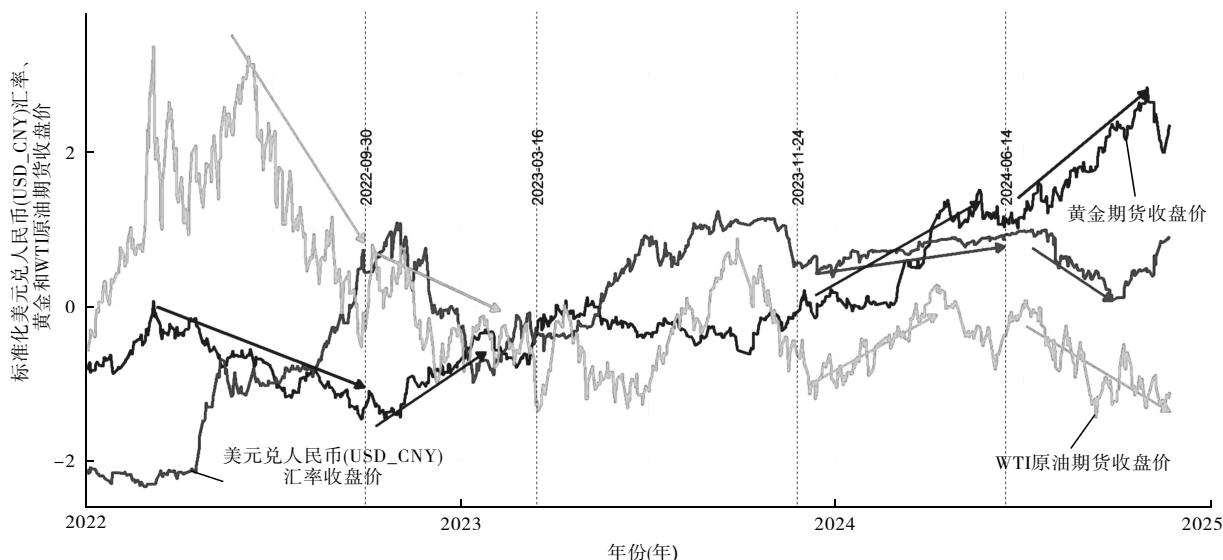


图1 标准化美元兑人民币(USD/CNY)汇率、黄金和WTI原油期货收盘价时间序列图

图2是由全局线性回归模型(10)和分段线性回归模型(11)得到的2022年1月3日至2024年11月21日黄金期货收盘价拟合值与真实值的对比图。从图2可以看出,由分段线性回归模型(11)得到的拟合值曲线与真实值曲线比较接近,而由全局线性回归模型(10)得到的拟合值曲线与真实值曲线相差较远。说明应用线性回归模型刻画黄金期货收盘价格的走势前应先进行变点检测与估计,建立更精准的分段线性回归模型,并且基于GMD算法的两阶段组 Lasso 多变点估计方法可以有效地估计出线性回归模型中回归系数的多变点位置。

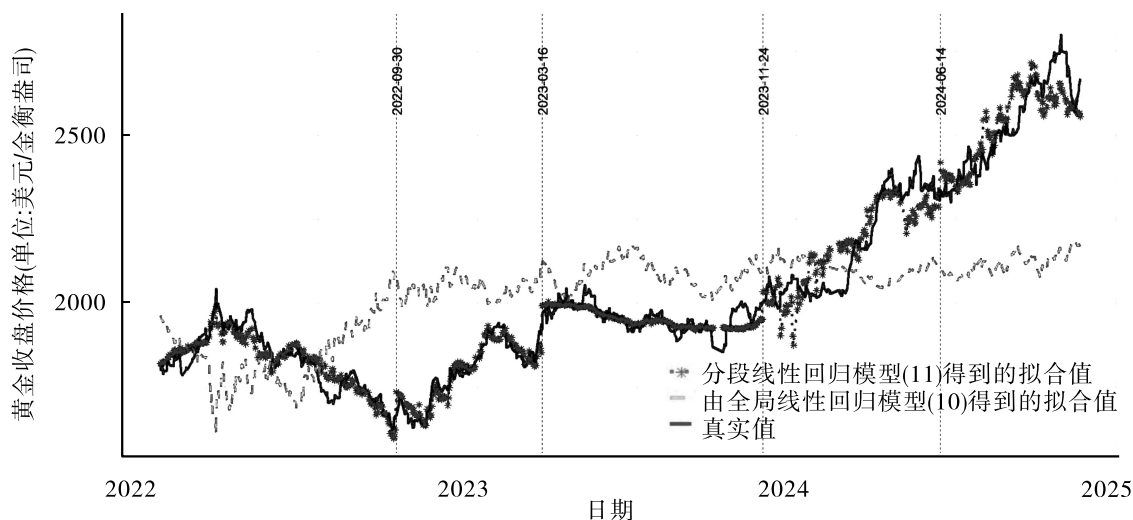


图2 由模型(10)和模型(11)得到的黄金期货收盘价拟合值与真实值的对比图

4 结论

文章将两阶段多变点估计方法与GMD算法求解组 Lasso 方法相结合,提出了一种基于GMD算法的两阶段组 Lasso 多变点估计方法,而且由该方法得到的变点估计量是相合的。模拟结果表明随机变量和误差项的分布对变点估计结果没有显著影响。此外,与基于GMD算法的未分段的组 Lasso、未分段的自适应 Lasso 和未分段的 Lasso 三种变量选择算法的多变点估计方法相比,基于GMD算法的两阶段组 Lasso 方法不仅具有较高的估计精度和运算速度,还避免了其他三种方法过度估计的问题。实证分析表明分段线性回归模型能

更精准地刻画黄金期货收盘价格的走势,说明在建模及统计分析之前检测与估计变点至关重要。

参考文献:

- [1] PAGE E S. Continuous Inspection Schemes[J]. *Biometrika*, 1954, 41(01-02): 100-115.
- [2] GHOSH P, VAIDA F. Random Changepoint Modelling of HIV Immunologic Responses[J]. *Statistics in Medicine*, 2010, 26(09): 2074-2087.
- [3] PEPELYSHEV A, POLUNCHENKO A S. Real-time Financial Surveillance Via Quickest Change-point Detection Methods[J]. *Statistics and its Interface*, 2017, 10(01): 93-106.
- [4] CHEN J, GUPTA A K. Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance[M]. Boston: Birkhäuser, 2014.
- [5] 李美琪, 金百锁, 董翠玲. 线性回归模型中相依数据的多结构变点的估计[J]. *中国科学: 数学*, 2023, 53(07): 1007-1024.
- [6] JIN B, SHI X, WU Y. A Novel and Fast Methodology for Simultaneous Multiple Structural Break Estimation and Variable Selection for Nonstationary Time Series Models[J]. *Statistics and Computing*, 2013, 23(02): 221-231.
- [7] JIN B, WU Y, SHI X. Consistent Two-stage Multiple Change-point Detection in Linear Models[J]. *Canadian Journal of Statistics*, 2016, 44(02): 161-179.
- [8] 周佳琪, 金百锁. 基于空间网络自回归变点模型的合肥市房地产价格影响因素分析[J]. *中国科学院大学学报*, 2020, 37(03): 398-404.
- [9] SUN X, WU Y. Simultaneous Multiple Change Points Estimation in Generalized Linear Models[J]. *Contemporary Experimental Design, Multivariate Analysis and Data Mining: Festschrift in Honour of Professor Kai-tai Fang*, 2020: 341-356.
- [10] 吕丽, 金百锁. 线性模型中多变点的置信区间估计[J]. *系统科学与数学*, 2021, 41(08): 2310-2326.
- [11] WANG B, LI J, WANG X. Multi-threshold Proportional Hazards Model and Subgroup Identification[J]. *Statistics in Medicine*, 2022, 41(29): 5715-5737.
- [12] YUAN M, LIN Y. Model Selection and Estimation in Regression with Grouped Variables[J]. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2006, 68(01): 49-67.
- [13] YANG Y, ZOU H. A Fast Unified Algorithm for Solving Group-Lasso Penalize Learning Problems[J]. *Statistics and Computing*, 2014, 25: 1129-1141.
- [14] 李傲梅. 基于自适应组Lasso算法的线性回归模型多变点估计[D]. 厦门: 厦门大学, 2019.
- [15] CSÖRGÖ M, HORVATH L. Limit Theorems in Change-point Analysis[M]. Chichester: Wiley, 1997.
- [16] 赵军辉, 董翠玲. 面板数据中方差的共同变点估计[J]. *新疆师范大学学报(自然科学版)*, 2024, 43(01): 22-32.

Two-stage Group Lasso Multiple Change-point Estimation based on GMD Algorithm in Linear Regression Model

AN Zi-zhen, DONG Cui-ling*

(School of Mathematical Sciences, Xinjiang Normal University, Urumqi, Xinjiang, 830017, China)

Abstract: Variable selection method is a popular and effective method to estimate and detect multiple change-point. This paper presents a two-stage Group Least absolute shrinkage and selection operator (Group Lasso) multiple change-point estimation method based on Groupwise Majorization Descent (GMD) algorithm, by using this method, the position and number of multiple change-point in the linear regression model can be estimated simultaneously. The numerical simulation results show that the two-stage Group Lasso multiple change-point estimation method based on GMD algorithm has significant advantages in both estimation accuracy and calculation speed, comparing with the multiple change-point estimation method based on GMD algorithm unsegmented group Lasso, unsegmented adaptive Lasso and unsegmented Lasso variable selection algorithms.

Keywords: Variable selection; Group Least absolute shrinkage and selection operator (Group Lasso); Groupwise Majorization Descent (GMD) algorithm; Linear regression model; Multiple change-point