

一种有效且稳健的变量选择方法

胡毓榆, 郭子君, 陈梦醒, 樊亚莉

(上海理工大学 理学院, 上海 200093)

摘要: 当数据中存在异常值时, 一些基于最小二乘估计的统计模型会产生较大的偏差, 最小一乘估计对异常值具有比较强的抵抗能力。考虑到数据中可能存在异常值的情况, 用绝对值损失代替平方损失, 针对同时具有变量稀疏性和相邻系数差分稀疏性这种结构的线性模型, 提出了最小一乘融合熔断自适应岭估计模型(LAD-Fused-BAR)。该模型将上一步估计的回归系数倒数的平方作为下一步惩罚权重, 自适应地给予不同变量不同的惩罚, 通过不断迭代得到最终解。运用交替方向乘子法(ADMM)求解 LAD-Fused-BAR 模型, 并证明了 ADMM 算法的收敛性。数值模拟和实证分析也验证了该模型的有效性和稳健性。

关键词: LAD-Fused-BAR 模型; 稳健回归; 交替方向乘子法

中图分类号: O 212.1 文献标志码: A

An effective and robust variable selection method

HU Yuyu, GUO Zijun, CHEN Mengxing, FAN Yali

(College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Some statistical models based on least squares estimation will produce large bias when there are outliers in the data. The least absolute deviation has strong resistance to outliers. Considering the influence of the outliers in the data, the square loss was replaced with the absolute loss. Aiming at the linear model of a structure that has both variable sparsity and sparsity of adjacent coefficient differences, the least absolute deviation fused broken adaptive ridge estimation model (LAD-Fused-BAR) was proposed. The square of the reciprocal of the regression coefficient estimated in the previous step was taken as the penalty weight for the next step, different penalties were adaptively given to different variables, and the final solution was obtained through continuous iteration. The alternating direction multiplier method (ADMM) was adopted to solve the LAD-Fused-BAR model and prove the convergence of the ADMM algorithm. Additionally, numerical simulation and empirical analysis confirm the efficacy and robustness of the proposed methodology.

Keywords: LAD-Fused-BAR; robust regression; ADMM

收稿日期: 2021-11-24

基金项目: 国家自然科学基金青年基金资助项目(11401383)

第一作者: 胡毓榆(1998-), 男, 硕士研究生。研究方向: 概率论与数理统计。E-mail: 2423794725@qq.com

通信作者: 樊亚莉(1978-), 女, 副教授。研究方向: 概率论与数理统计。E-mail: yalifan@usst.edu.cn

大数据时代的生活中到处充满着海量的数据^[1], 通过分析这些数据, 研究者可以探究出潜在的商业信息及其应用价值。然而, 随着科技的发展, 数据中变量的维度呈指数级增长, 并且数据中经常夹杂着噪声和冗余信息。另外, 这些变量中往往存在一些不相关变量和冗余变量, 给数据分析带来了困难, 影响估计的效率以及精度。为了解决这个问题, 通常利用变量选择将重要的变量筛选出来, 从而达到降维的目的。

1 文献综述

关于变量选择问题, 很多研究者从统计学角度作了大量研究。1973年, Akaike等^[2]提出了AIC准则。1978年, Schwarz^[3]在贝叶斯的框架下提出了BIC准则。然而, 当变量的维度增加时, 这种基于AIC和BIC准则选取最优变量子集的传统方法计算效率较低。近十几年来, 通过在损失函数后面加入一个惩罚函数的方法使得变量选择有了较大的发展, 并且这种加入惩罚函数的方法可以同时实现参数估计以及变量选择的目的。1995年, Natarajan^[4]提出了在损失函数后面加入 l_0 范数, 通过惩罚非零元素个数从而达到变量选择的目的。这个惩罚函数是一种很直观的形式, 但是 l_0 范数是非凸且不连续的函数, 这就导致了最小化 l_0 问题是一个NP-Hard问题。1996年, Tibshirani^[5]提出了LASSO(least absolute shrinkage and selection operator)惩罚, 通过加入 l_1 范数从而得到一个稀疏解。 l_1 范数是 l_0 范数的一个凸松弛。尽管 l_1 范数是一个凸函数且易于求解, 但是LASSO估计的参数是有偏的^[6]。为了解决这个问题, 近些年有许多学者提出了非凸的惩罚函数来解决估计的有偏问题。2001年, Fan等^[6]提出了SCAD(smoothly clipped absolute deviation)惩罚。2006年, Zou^[7]在LASSO的基础上提出了Adaptive LASSO惩罚, 该方法是对LASSO的一种改进。SCAD和Adaptive LASSO在一定条件下都满足Oracle性质。2008年, Candès等^[8]提出了对数惩罚。2010年, Zhang^[9]提出了MCP(minimax concave penalty)惩罚。很多研究结果表明非凸惩罚函数在理论分析以及实际应用中具有更优秀的表现^[10]。

然而, 对于一些具有一定结构的复杂数据, 例如信号处理、基因表达等分段常数函数类型的

数据, 上述方法并不适用。2005年, Tibshirani等^[11]在LASSO方法的基础上提出了Fused-LASSO方法, 在考虑变量稀疏性的同时还考虑了变量差分的稀疏性。这种方法能够同时解决变量稀疏性和获取噪声信号或者基因序列中非零片段的情况。2011年, Tibshirani等^[12]提出了(2D)Fused-LASSO方法, 该方法能够处理图像去噪的问题。

2017年, Dai等^[13]提出了一种熔断自适应岭估计(BAR)方法, 该方法本质上是一种迭代重加权岭估计方法, 它能够很有效地同时进行变量选择和参数估计。相比于之前文献提到的方法, BAR方法的优点是参数估计通过迭代的方式进行, 在每一步迭代过程中上一步参数的估计值将作为下一步估计的权重。因此, BAR方法中的权重是自适应更新的, 并且加入的惩罚是 l_2 惩罚, 目标函数是严格凸以及可微的。因此, 每一步迭代过程都可以求得显式解, 最后通过不断迭代得到最终的估计。Dai等^[13]证明了他们提出的估计具有Oracle性质和Group Effect, 并且Dai在最后数值实验中表明了BAR方法比上文所述几种方法更有效。

2020年, Dai等^[14]将BAR推广为广义的形式, 这种广义的形式能够估计出回归系数的任意结构, 例如稀疏结构、分段常数结构以及线性趋势结构。融合熔断自适应岭估计(Fused-BAR)是广义BAR方法的一种特殊形式, 它能够很有效地进行变量选择以及识别出噪声信号或者基因序列中非零片段的情况。文献^[14]证明了在一定条件下, 在BAR方法迭代过程中, 估计值会收敛到一个不动点。如果事先给出一个正确的初始值, 那么最终BAR方法估计的效果就很好。文献^[14]在数值实验中证实了Fused-BAR方法比Fused-LASSO方法更有效。

然而BAR方法采用的损失是平方损失, 平方损失对于异常值和重尾数据比较敏感。当数据中存在异常值时, 估计的参数会产生较大的偏差和方差, 通常采用稳健的损失函数作为代替。例如Huber损失、t型损失^[15]、稳健的估计方程^[16]、绝对值损失^[17]。这些损失函数对异常值和重尾数据具有一定的抵抗能力。2007年, Wang等^[17]在LASSO方法的基础上, 将平方损失替换为绝对值损失, 提出了LAD-LASSO方法, 与LASSO方法相比, LAD-LASSO方法也可以同时进行变量选择和参数估计, 并且对于重尾分布的数据和带有异

常值的数据有抵抗力。2016年,张环^[18]在 Fused-LASSO 方法的基础上用绝对值损失代替平方损失,提出了 LAD-Fused-LASSO 方法,既能有效地将相邻特征选出来,又能抵抗异常值。

基于以上文献的启发,本文在 Fused-BAR 方法的基础上提出了一种稳健且有效的变量选择模型,称为最小一乘融合熔断自适应岭估计(LAD-Fused-BAR)。它是将 Fused-BAR 估计方法中的平方损失替换成绝对值损失,从而实现稳健且有效的变量选择。Fused-BAR 在迭代过程中每一步可以求出显式解,然而把平方损失替换成绝对值损失之后,由于多了一项非光滑项,要优化的目标函数无法求显示解,并且基于梯度的一些方法也失效了,这给计算带来了很大的挑战。在处理非光滑函数 l_1 上,文献[19]中提出了最小角回归,文献[20]提出了内点法。然而,这些算法的计算复杂度较大,在大规模数据上表现不佳。2010年,Boyd等^[21]首次提出了交替方向乘子法(ADMM)。Glowinski等^[22]受到启发,将 ADMM 方法应用到 LASSO 以及 Group-LASSO 上。因此,本文采取交替方向乘子法(ADMM)进行求解,并证明了 ADMM 算法的收敛性。

2 最小一乘融合熔断自适应岭估计(LAD-Fused-BAR)

考虑一般的线性回归模型

$$Y = X^T \beta + \varepsilon \quad (1)$$

式中: $Y = (y_1 \cdots y_n)$, 是响应变量; $\beta \in \mathbb{R}^p$, 是回归系数; $X = (x_1 \cdots x_n)^T$, 是协变量; $\varepsilon = (\varepsilon_1 \cdots \varepsilon_n)$, ε_i 的均值为 0, 方差为 σ^2 , ε 的各分量相互独立。

Dai 在文献[14]中提出的 Fused-BAR 估计方法如下所示:

$$g(\tilde{\beta}) \equiv \argmin_{\beta} \|Y - X\beta\|_2^2 + \lambda_1 \sum_{i=1}^n \frac{\beta_i^2}{\tilde{\beta}_i^2} + \lambda_2 \sum_{i=2}^n \frac{(\beta_i - \beta_{i-1})^2}{(\tilde{\beta}_i - \tilde{\beta}_{i-1})^2} \quad (2)$$

式中: $g(\tilde{\beta})$ 表示估计的参数; λ_1 和 λ_2 是调节参数。

若在式(2)中只考虑 λ_1 对应的惩罚项, 这就是文献[13]提出的 BAR 方法。式(2)所定义的 $g(\tilde{\beta})$ 是在上一步的估计 $\tilde{\beta}$ 基础上的更新公式。可以看到在更新公式中, 不仅对变量施加惩罚, 也对相邻变量的差分进行惩罚。BAR 估计和文献[7]中提出

的 Adaptive LASSO 方法在思想上类似, 对于估计值比较大的变量给它较小的惩罚, 而对于估计值比较小的变量给它较大的惩罚。但是 Adaptive LASSO 的权重并不是根据上次估计自适应的, 而是事先估计的一个常数。文献[14]认为, 在每次迭代过程中, 将上一次估计的 $\{g_k(\tilde{\beta})\}^{-2}$ 作为权重比用常数作为权重更好。随着迭代的不断进行, 对于真实 β 中为零的变量的权重会趋于无穷, 而对于真实 β 中非零变量的权重会趋于一个常数。由于 BAR 方法使用的惩罚是 l_2 惩罚, 因此, 每一步迭代都能产生显式解。

文献[14]中将岭估计

$$\hat{\beta}_R = (X'X + \xi I)^{-1} X^T Y$$

作为初始值, 式中, $\xi > 0$ 是一个正数, I 是单位矩阵。因此, 上述提出的估计即为融合熔断自适应岭估计(Fused-BAR), 最终估计可以通过迭代算法 $\hat{\beta}_j = g(\hat{\beta}_{j-1})$ 的极限求出来, 即

$$\hat{\beta}_{\text{Fused-BAR}} = \lim_{j \rightarrow \infty} \hat{\beta}_j \quad (3)$$

定义

$$M = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

是一个 $(p-1) \times p$ 的矩阵, 则式(2)可以写成

$$g(\tilde{\beta}) \equiv \argmin_{\beta} \|Y - X\beta\|^2 + \lambda_1 \beta^T H_1(\tilde{\beta}) \beta + \lambda_2 \beta^T M^T H_2(\tilde{\beta}) M \beta \quad (4)$$

其中

$$H_1(\tilde{\beta}) = \text{diag}(\tilde{\beta}_i^{-2}), H_2(\tilde{\beta}) = \text{diag}((M\tilde{\beta})_i^{-2}) \quad (5)$$

如果要解决的是信号处理、基因检测以及图像去噪等问题, 只要令 $X = I$ 即可。式(4)中目标函数用的是平方损失, 当响应变量中存在异常值时, 该损失函数会放大异常值的影响。受到文献[17-18]的启发, 本文用绝对值损失代替平方损失, 得到最小一乘融合熔断自适应岭估计模型(LAD-Fused-BAR)为

$$g(\tilde{\beta}) \equiv \argmin_{\beta} \|Y - X\beta\|_1 + \lambda_1 \beta^T H_1(\tilde{\beta}) \beta + \lambda_2 \beta^T M^T H_2(\tilde{\beta}) M \beta \quad (6)$$

3 算法求解

由于本文的方法是将文献[14]中的平方损失

替换成了绝对值损失, 在目标函数中多了一项非光滑项, 使得在每一次迭代过程中无法求出显式解。当数据规模很大时, 最小角回归以及内点法计算复杂度高。Boyd 等^[21]在 2010 年提出了 ADMM 算法, 通过选取一个光滑函数来逼近模型的非光滑项, 并且 ADMM 方法在大规模数据上计算效率高。除此之外, ADMM 算法在处理该问题时, 使得每一步迭代都有显式解, 这充分地保障了 ADMM 求解 LAD-Fused-BAR 问题的有效性。

为了应用交替方向乘子方法, 引入辅助变量 τ , 则式 (6) 等价于

$$\begin{cases} \min_{\beta, \tau} \|\tau\|_1 + \lambda_1 \beta^T H_1(\tilde{\beta})\beta + \lambda_2 \beta^T M^T H_2(\tilde{\beta})M\beta \\ \text{s.t. } Y - X\beta = \tau \end{cases} \quad (7)$$

式 (7) 的增广拉格朗日形式为

$$L_u(\beta, \tau, \alpha) = \|\tau\|_1 + \lambda_1 \beta^T H_1(\tilde{\beta})\beta + \lambda_2 \beta^T M^T H_2(\tilde{\beta})M\beta + \frac{u}{2} \left\| Y - X\beta - \tau - \frac{\alpha}{u} \right\|^2 - \frac{1}{2u} \|\alpha\|_2^2 \quad (8)$$

β 的更新为

$$\begin{aligned} \beta^{k+1} = \arg \min_{\beta} \lambda_1 \beta^T H_1(\tilde{\beta})\beta + \lambda_2 \beta^T M^T H_2(\tilde{\beta})M\beta + \\ \frac{u}{2} \left\| Y - X\beta - \tau^k - \frac{\alpha^k}{u} \right\|^2 \end{aligned} \quad (9)$$

令式 (9) 右端微分为 0 得到

$$\begin{aligned} \beta^{k+1} = (2\lambda_1 H_1(\tilde{\beta}) + 2\lambda_2 M^T H_2(\tilde{\beta})M + uX^T X)^{-1} \cdot \\ X^T (uY - u\tau^k - \alpha^k) \end{aligned} \quad (10)$$

τ 的更新为

$$\tau^{k+1} = \arg \min_{\tau} \frac{u}{2} \left\| Y - X\beta^{k+1} - \tau - \frac{\alpha^k}{u} \right\|^2 + \|\tau\|_1 \quad (11)$$

该优化问题可以用软阈值算法进行求解得到, 即

$$\begin{aligned} \tau^{k+1} = \text{sign} \left(Y - X\beta^{k+1} - \frac{\alpha^k}{u} \right) \cdot \\ \max \left\{ 0, \left| Y - X\beta^{k+1} - \frac{\alpha^k}{u} \right| - \frac{1}{u} \right\} \end{aligned} \quad (12)$$

α 的更新为

$$\alpha^{k+1} = \alpha^k - u(Y - X\beta^{k+1} - \tau^{k+1}) \quad (13)$$

运用算法 ADMM 求解本文 LAD-Fused-BAR 模型的具体流程如下:

输入 $\{x_i, y_i\}_{i=1}^n, \lambda_1, \lambda_2, \mu, \tilde{\beta}^0, (\alpha^0, \tau^0) = (0, 0)$

$k = 0, \tilde{\beta} = \tilde{\beta}^0$

while $\|\tilde{\beta} - \beta^k\| \geq \delta$

$\tilde{\beta} = \hat{\beta}^k$

$H_1(\tilde{\beta}) = \text{diag}(\tilde{\beta}_i^2)$

$H_2(\tilde{\beta}) = \text{diag}((M\tilde{\beta})_i^2)$

while $\|\beta^{j+1} - \beta^j\| \geq \delta$

$\beta^{j+1} = \arg \min_{\beta} \lambda_1 \beta^T H_1(\tilde{\beta})\beta + \lambda_2 \beta^T M^T H_2(\tilde{\beta})M\beta +$

$\frac{u}{2} \left\| Y - X\beta - \tau^j - \frac{\alpha^j}{u} \right\|^2$

$\tau^{j+1} = \text{sign} \left(Y - X\beta^{j+1} - \frac{\alpha^j}{u} \right) \max \left\{ 0, \left| Y - X\beta^{j+1} - \frac{\alpha^j}{u} \right| - \frac{1}{u} \right\}$

$\alpha^{j+1} = \alpha^j - u(Y - X\beta^{j+1} - \tau^{j+1})$

end while

end while

输出 $\hat{\beta}^k$

4 收敛性证明

文献 [23] 已经给出了关于 ADMM 收敛性的一个框架。为了建立 ADMM 算法的收敛性, 以算法的第 $k+1$ 次迭代值作为一个变分 VI 问题^[23], 可以得到下面引理。其中, k 表示迭代次数。下面 4 个引理在文献 [23] 中已经得到证明, 本文不再赘述。

引理 1 令 $\{w^k\} = \{\beta^k, \tau^k, \alpha^k\}$ 表示由算法产生的序列, $\{v^k\} = \{\tau^k, \alpha^k\}$, 有

$$\begin{aligned} \{w^k\} = \{\beta^k, \tau^k, \alpha^k\} (v^{k+1} - v^*)^T H(v^k - v^{k+1}) \geq \\ (w^{k+1} - w^*)^T \eta(\tau^k, \tau^{k+1}) \end{aligned} \quad (14)$$

其中

$$H = \begin{pmatrix} \mu I_m & 0 \\ 0 & \frac{1}{\mu} I_m \end{pmatrix}, \quad \eta(\tau^k, \tau^{k+1}) = \mu \begin{pmatrix} X^T \\ I_m^T \\ 0 \end{pmatrix} (\tau^k - \tau^{k+1})$$

引理 2 令 $\{w^k\} = \{\beta^k, \tau^k, \alpha^k\}$, 表示由算法产生的序列, 以及 $\{v^k\} = \{\tau^k, \alpha^k\}$, 有

$$\begin{aligned} (w^{k+1} - w^*)^T \eta(\tau^k, \tau^{k+1}) = \\ (\alpha^k - \alpha^{k+1})^T (\tau^k - \tau^{k+1}) \end{aligned} \quad (15)$$

以及

$$(\alpha^k - \alpha^{k+1})^T (\tau^k - \tau^{k+1}) \geq 0 \quad (16)$$

引理 3 令 $\{w^k\} = \{\beta^k, \tau^k, \alpha^k\}$, 表示由算法产生的序列, 以及 $\{v^k\} = \{\tau^k, \alpha^k\}$, 有

$$(v^{k+1} - v^*)^T H(v^k - v^{k+1}) \geq 0, \quad \forall v^* \in V^* \quad (17)$$

引理 4 令 $\{\boldsymbol{w}^k\} = \{\boldsymbol{\beta}^k, \boldsymbol{\tau}^k, \boldsymbol{\alpha}^k\}$, 表示由算法产生的序列, 以及 $\{\boldsymbol{v}^k\} = \{\boldsymbol{\tau}^k, \boldsymbol{\alpha}^k\}$, 有

$$\begin{aligned} \|\boldsymbol{v}^{k+1} - \boldsymbol{v}^*\|_H^2 &\leq \|\boldsymbol{v}^k - \boldsymbol{v}^*\|_H^2 - \\ &\|\boldsymbol{v}^k - \boldsymbol{v}^{k+1}\|_H^2, \forall \boldsymbol{v}^* \in V^* \end{aligned} \tag{18}$$

其中, $\|\boldsymbol{v}^{k+1} - \boldsymbol{v}^*\|_H^2 = (\boldsymbol{v}^{k+1} - \boldsymbol{v}^*)^\top \boldsymbol{H} (\boldsymbol{v}^{k+1} - \boldsymbol{v}^*)$.

定理 1 给定任意起始点 $(\boldsymbol{\beta}^0, \boldsymbol{\tau}^0, \boldsymbol{\alpha}^0) \in T$, 对任意的 u , 由 ADMM 算法产生的序列 $\{\boldsymbol{w}^k\} = \{\boldsymbol{\beta}^k, \boldsymbol{\tau}^k, \boldsymbol{\alpha}^k\}$ 收敛于稳健 LAD-fused-BAR 模型的解 $\boldsymbol{w}^\infty = (\boldsymbol{\beta}^\infty, \boldsymbol{\tau}^\infty, \boldsymbol{\alpha}^\infty)$.

证明 由引理 4 中的不等式

$$\|\boldsymbol{v}^{k+1} - \boldsymbol{v}^*\|_H^2 \leq \|\boldsymbol{v}^k - \boldsymbol{v}^*\|_H^2 - \|\boldsymbol{v}^k - \boldsymbol{v}^{k+1}\|_H^2, \forall \boldsymbol{v}^* \in V^*$$

可以得到 $\{\|\boldsymbol{v}^k - \boldsymbol{v}^*\|_H\}$ 是一个单调递减的序列, 又因为 $\|\boldsymbol{v}^k - \boldsymbol{v}^*\|_H \geq 0$, 所以 $\{\|\boldsymbol{v}^k - \boldsymbol{v}^*\|_H\}$ 是一个有界序列, 从而得到 $\{\boldsymbol{v}^k\}$ 也是有界的。再由不等式 $\|\boldsymbol{v}^{k+1} - \boldsymbol{v}^*\|_H^2 \leq \|\boldsymbol{v}^k - \boldsymbol{v}^*\|_H^2 - \|\boldsymbol{v}^k - \boldsymbol{v}^{k+1}\|_H^2$ 可以得到, $\|\boldsymbol{v}^k - \boldsymbol{v}^{k+1}\|_H \leq \|\boldsymbol{v}^k - \boldsymbol{v}^*\|_H - \|\boldsymbol{v}^{k+1} - \boldsymbol{v}^*\|_H$ 。不等式两边同时 $k \rightarrow \infty$ 且 $\|\boldsymbol{v}^k - \boldsymbol{v}^{k+1}\|_H \geq 0$, 得到 $\lim_{k \rightarrow \infty} \|\boldsymbol{v}^k - \boldsymbol{v}^{k+1}\|_H = 0$, 那么就有

$$\begin{cases} \lim_{k \rightarrow \infty} \|\boldsymbol{\tau}^k - \boldsymbol{\tau}^{k+1}\| = 0 \\ \lim_{k \rightarrow \infty} \|\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k+1}\| = 0 \end{cases}$$

由单调有界定理得到, $\{\boldsymbol{v}^k\}$ 收敛至 $\boldsymbol{v}^\infty = (\boldsymbol{\tau}^\infty, \boldsymbol{\alpha}^\infty)$, 根据式 (10) 得到 $\boldsymbol{\beta}^0$ 收敛至 $\boldsymbol{\beta}^\infty$, 证毕。

5 数值模拟

真实的 $\boldsymbol{\beta}^*$ 如下所示生成:

$$\boldsymbol{\beta}^* = (\underbrace{0.6, \dots, 0.6}_5, \underbrace{0, \dots, 0}_{45}, \underbrace{-0.7, \dots, -0.7}_5, \underbrace{0, \dots, 0}_{45})$$

其中, 非零系数的个数为 10, $n = 200, p = 100$, $x_{ij} \sim N(0, 1), 1 \leq i \leq n, 1 \leq j \leq p$. ε_i 服从均值为 0,

标准差为 0.01 的正态分布, y 由式 (1) 生成。为了研究稳健性, 选取 $c\%$ 的数据进行污染, 通过将 $c\%$ 的 ε_i 换成服从均值为 30、标准差为 0.1 的正态分布的噪声值, 下文分别以 $c=0, 5, 10$ 进行模拟。

在 Fused-BAR 程序中, 对 $\xi, \lambda_1, \lambda_2$ 这 3 个参数进行调节。参数选择和文献 [14] 一样, ξ 的选择用 5 折交叉验证的方法。对于 λ_1, λ_2 的选择, 用网格搜索的方法进行筛选, 将 λ_1 和 λ_2 取 $[0.0001, 0.001, 0.01, 0.1, 1, 10]$, 然后通过 5 折交叉验证来选取最优参数。LAD-Fused-BAR 方法里面有 4 个超参数: $\xi, \lambda_1, \lambda_2, u$, 初始值 $\hat{\boldsymbol{\beta}}_R$ 的选取和 Fused-BAR 采取相同的方式。若同时遍历余下的 3 个超参数, 计算时间较慢。因此, 采取先固定 u 然后利用 5 折交叉验证的方法选取最优的 λ_1 和 λ_2 , 然后再固定 λ_1 和 λ_2 来选取最优的 u 。

关于指标选取, 既要考虑估计的偏差又要考虑变量选择的准确率。因此采用平均绝对误差来表示估计的偏差, 即

$$w_{MAE} = \frac{\sum_{i=1}^n |\hat{\beta}_i - \beta_i^*|}{n}$$

式中, w_{MAE} 为平均绝对误差。

假设 θ_{TP} 为正确估计为非零的个数, θ_{FP} 为错误估计为非零的个数, θ_{FT} 为正确估计为零的个数, θ_{FN} 为错误估计为零的个数, 则可以用准确率 $\theta_{ACC} = \frac{\theta_{TP} + \theta_{TN}}{\theta_{TP} + \theta_{TN} + \theta_{FP} + \theta_{FN}}$ 、精确率 $\theta_{PRE} = \frac{\theta_{TP}}{\theta_{TP} + \theta_{FP}}$ 、召回率 $\theta_{REC} = \frac{\theta_{TP}}{\theta_{TP} + \theta_{FN}}$ 、 $F1 = \frac{2\theta_{TP}}{2\theta_{TP} + \theta_{FP} + \theta_{FN}}$ 指标来判别变量选择的能力。将数据划分成 50 份, 其中一份作为测试集, 余下部分作为训练集。运行 50 次后分别求出平均值如表 1 所示。

从表 1 可以看出, 在没有污染的情况下,

表 1 模拟结果
Tab.1 Simulation result

参数	无污染		污染5%		污染10%	
	LAD-Fused-BAR	Fused-BAR	LAD-Fused-BAR	Fused-BAR	LAD-Fused-BAR	Fused-BAR
MAE	9×10^{-4}	5.41×10^{-5}	1.57×10^{-3}	2.56×10^{-2}	2.42×10^{-2}	3.96×10^{-2}
θ_{ACC}	1.00	1.00	1.00	0.98	1.00	0.97
θ_{PRE}	1.00	1.00	1.00	0.91	1.00	0.83
θ_{REC}	1.00	1.00	1.00	1.00	1.00	0.94
F1	1.00	1.00	1.00	0.92	1.00	0.89

LAD-Fused-BAR 和 Fused-BAR 估计的准确率都为 1, 两种方法都能全部准确地估计出真实的 β^* 。虽然 Fused-BAR 估计的偏差会比 LAD-Fused-BAR 更小, 但差距并不是很大, 这证明了 LAD-Fused-BAR 在无污染情况下估计的有效性。然而, 在有污染的情况下, LAD-Fused-BAR 的平均绝对误差比 Fused-BAR 小, 并且其他指标都比 Fused-BAR 要大。这说明 LAD-Fused-BAR 在当数据有异常值的时候表现更稳健, 估计准确率和精度都要比

Fused-BAR 更优秀。进一步, 用不同污染情况下的系数分布来直观地表示两种方法变量选择的能力, 如图 1~3 所示。

图 1 也说明了在无污染的情况下, Fused-BAR 和 LAD-Fused-BAR 都能将 β^* 估计出来, 但是当数据有污染时, Fused-BAR 的估计值出现偏差, 不能将真实的 β^* 准确估计出来, 而 LAD-Fused-BAR 方法依然能将真实的 β^* 估计出来, 说明在数据有异常值的情况下, LAD-Fused-BAR 方法更稳健。

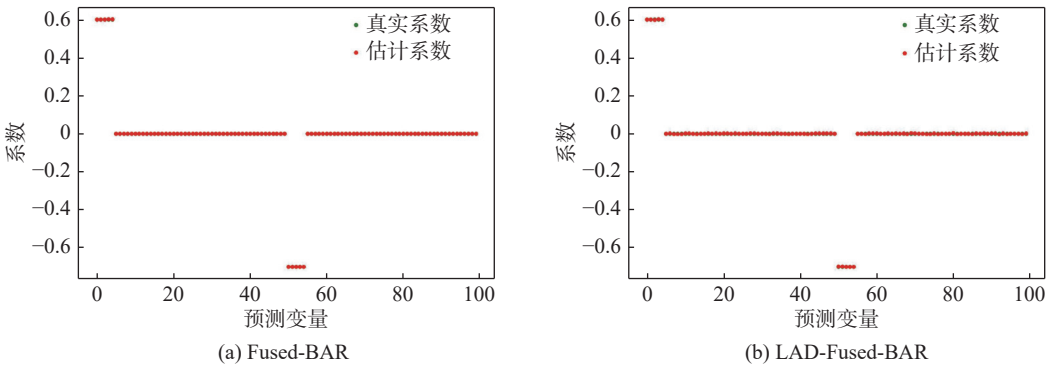


图 1 无污染情况下 Fused-BAR 与 LAD-Fused-BAR 方法的 β 估计
Fig.1 β estimation of Fused-BAR and LAD-Fused-BAR without pollution

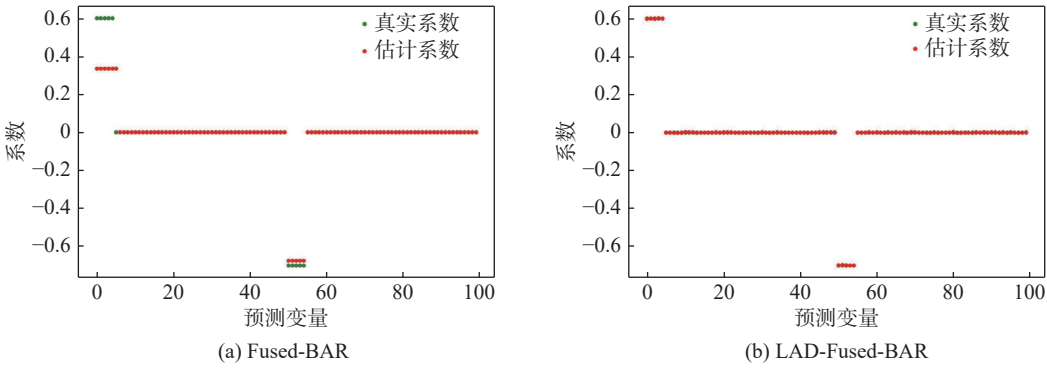


图 2 5% 污染情况下 Fused-BAR 与 LAD-Fused-BAR 方法的 β 估计
Fig.2 β estimation of Fused-BAR and LAD-Fused-BAR under 5% pollution

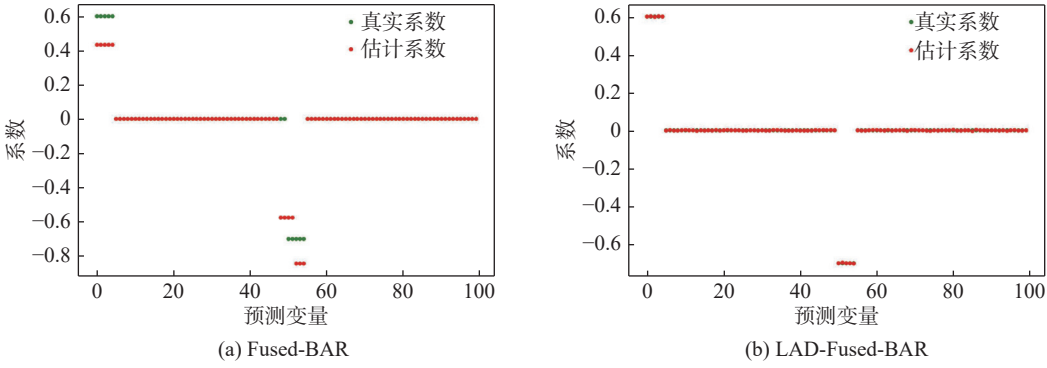


图 3 10% 污染情况下 Fused-BAR 与 LAD-Fused-BAR 方法的 β 估计
Fig.3 β estimated of Fused-BAR and LAD-Fused-BAR under 10% pollution

6 实证分析

在癌症研究中，拷贝数变异数据 (CNV) 是一个很重要的数据集，该数据集具有相邻关系。也就是说，CNV 通常是各种长度的线段的形式^[24]。比较基因组杂交 (CGH) 阵列是扫描基因组中 CNV 的一个很有效的工具。可以通过 CGH 扫描 CNV 来检测基因是否发生改变，即 DNA 拷贝数的缺失和增加。为了方便地检测基因是否改变，通常将 CGH 数据阵列设置为肿瘤细胞中的 DNA 拷贝数与正常或参考细胞中的 DNA 拷贝数的 log 2 比率。因此，当 CGH 为正值时表示 DNA 拷贝数增加，而当其为负值时，表示 DNA 拷贝数缺失。CGH 通常由具有零值分段区域的分段常数序列或函数逼近。

近年来，有许多方法已经对 CGH 数据进行研究。例如 EM 算法^[25]、隐马尔可夫方法^[26]、Fused-LASSO^[27] 以及 Fused-BAR^[14]。这些方法可以用于 CGH 的可视化以及用于 CGH 分段值的推断。本文分别采用 Fused-BAR 和 LAD-Fused-BAR 来分析 CGH 数据，CGH 数据的获得来自于 R 包 cghFlasso。为了验证稳健性，将 $d\%$ 的数据进行污染，通过将 $d\%$

的数据加上 5。在本文中 d 分别取 0, 3, 5 进行实验。

参数选择和模拟时一样采取网格搜索的方式，该实验解决了一个信号去噪的问题。因此，将数据的奇数行作为训练集，偶数行作为验证集，用两折交叉验证选取最优参数。实验结果如表 2 和图 4~6 所示。

由表 2 可见，当数据不加污染时，虽然 Fused-BAR 方法估计的 MAE 比 LAD-Fused-BAR 方法的小，但差别不大，说明在无污染的情况下，LAD-Fused-BAR 估计是有效的。但是当数据有一部分被污染时，LAD-Fused-BAR 方法估计的 MAE 要比 Fused-BAR 更小，说明 LAD-Fused-BAR 在数据有污染时更稳健。从图 4 可以看出，在无污染的情况下，Fused-BAR 方法能够将 CGH 非零片段识别出来，LAD-Fused-BAR 方法也能将非零片段识别出来。而 LAD-Fused-BAR 识别出来的不像 Fused-BAR 是一条直线，这是因为本文算法求的是近似解，因此，在噪声比较大的情况下估计的信号会有波动。但是，当数据被污染时，LAD-Fused-BAR 估计的 MAE 比 Fused-BAR 方法的小，说明在有污染的情况下，LAD-Fused-BAR 方法更稳健。

表 2 MAE 实验结果
Tab.2 Experimental results of MAE

无污染		3%污染		5%污染	
LAD-Fused-BAR	Fused-BAR	LAD-Fused-BAR	Fused-BAR	LAD-Fused-BAR	Fused-BAR
0.337	0.302	0.409	0.523	0.513	0.614

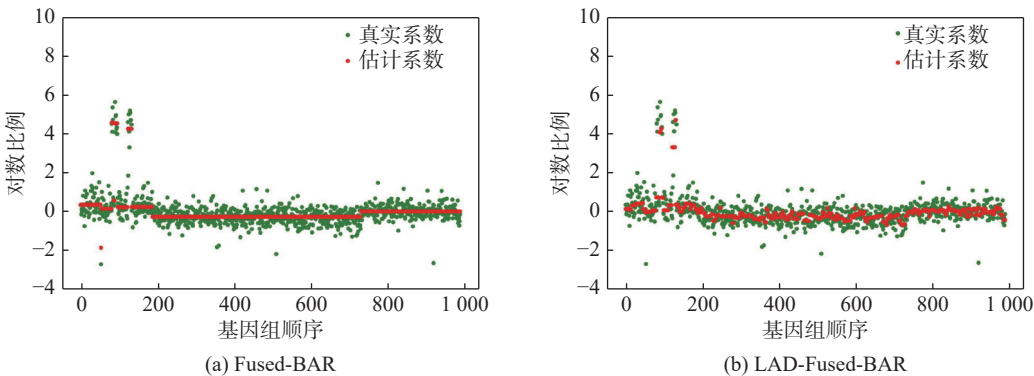


图 4 无污染情况下的 Fused-BAR 与 LAD-Fused-BAR 估计
Fig.4 Estimation of Fused-BAR and LAD-Fused-BAR without pollution

从图 5 和图 6 可以看出，相比于图 4，尽管 BAR 方法能够估计出一条直线，但是在有污染的情况下 Fused-BAR 方法估计的系数绝大部分都被

压缩到 0，体现不出分段常数的形式。这说明 Fused-BAR 估计在有异常值的情况下不稳健。由于本文方法求的是近似解而不是解析解，所以当

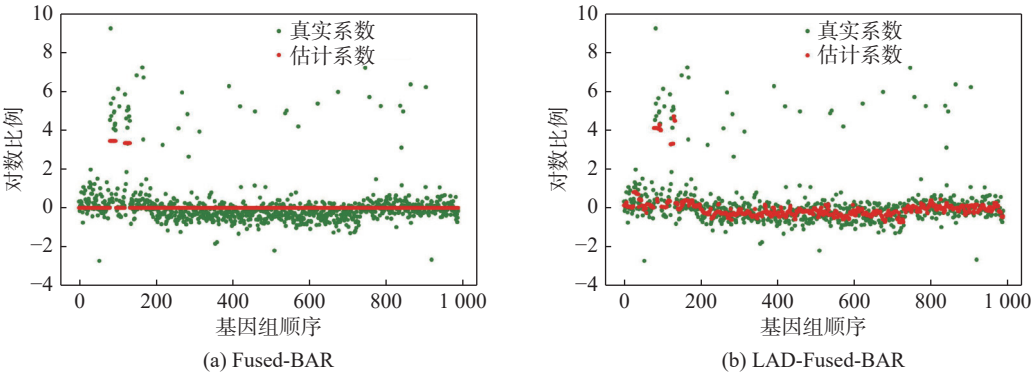


图 5 3% 污染情况下的 Fused-BAR 与 LAD-Fused-BAR 估计

Fig.5 Estimation of Fused-BAR and LAD-Fused-BAR under 3% pollution

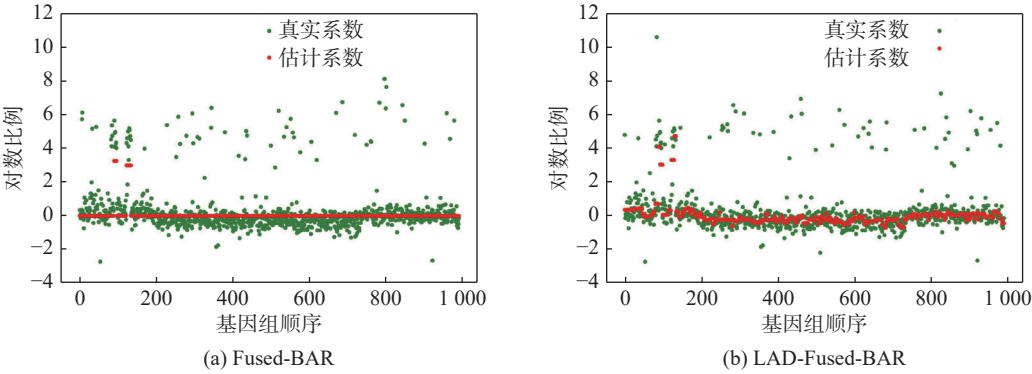


图 6 5% 污染情况下的 Fused-BAR 与 LAD-Fused-BAR 估计

Fig.6 Estimation of Fused-BAR and LAD-Fused-BAR under 5% pollution

噪声比较大时，不能估计出一条分段直线 的形式。但是，LAD-Fused-BAR 方法还是能够判断出 估计的信号是呈分段常数的形式，这说明 LAD-Fused- BAR 在数据中有异常值的情况下更稳健。

7 总结和展望

在 Fused-BAR 变量选择的框架下提出了一种 稳健且有效的变量选择方法，通过把平方损失替 换成绝对值损失从而达到稳健的效果。然而，将 平方损失换成绝对值损失后，导致要优化的目 标函数无法求出显式解。因此，采用 ADMM 进行 求解，并且证明了 ADMM 算法的收敛性。模拟结 果以及实证分析显示，与 Fused-BAR 方法相比， LAD-Fused-BAR 方法在数据有异常值的情况下更 稳健。在面对噪声比信号大的情况时，本文提出 的算法由于得到的是近似解，偏差较大，后续可 以考虑改进算法和其他稳健的损失函数。

参考文献：

[1] CHEN C L P, ZHANG C Y. Data-intensive applications, challenges, techniques and technologies: a survey on Big

Data[J]. *Information Sciences*, 2014, 275: 314–347.
[2] AKAIKE H. Information theory and an extension of the maximum likelihood principle[M]//PETROV B N, CSAKI F. Proceedings of the 2nd International Symposium on Information Theory. Budapest: Akademiai Kiado, 1973: 267–281.
[3] SCHWARZ G. Estimating the dimension of a model[J]. *The Annals of Statistics*, 1978, 6(2): 461–464.
[4] NATARAJAN B K. Sparse approximate solutions to linear systems[J]. *SIAM Journal on Computing*, 1995, 24(2): 227–234.
[5] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, 58(1): 267–288.
[6] FAN J Q, LI R Z. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. *Journal of the American Statistical Association*, 2001, 96(456): 1348–1360.
[7] ZOU H. The adaptive lasso and its oracle properties[J]. *Journal of the American Statistical Association*, 2006, 101(476): 1418–1429.
[8] CANDÈS E J, WAKIN M B, BOYD S P. Enhancing sparsity by reweighted ℓ_1 minimization[J]. *Journal of Fourier Analysis and Applications*, 2008, 14(5): 877–905.

- [9] ZHANG C H. Nearly unbiased variable selection under minimax concave penalty[J]. *The Annals of Statistics*, 2010, 38(2): 894–942.
- [10] XU Z B. Data modeling: visual psychology approach and $L_{1/2}$ regularization theory[C]//*Proceedings of the International Congress of Mathematicians 2010 (ICM 2010)*. Hyderabad: World Scientific, 2010: 3151–3184.
- [11] TIBSHIRANI R, SAUNDERS M, ROSSET S, et al. Sparsity and smoothness via the fused lasso[J]. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005, 67(1): 91–108.
- [12] TIBSHIRANI R J, TAYLOR J. The solution path of the generalized lasso[J]. *The Annals of Statistics*, 2011, 39(3): 1335–1371.
- [13] DAI L L, CHEN K N, SUN Z H, et al. Broken adaptive ridge regression and its asymptotic properties[J]. *Journal of Multivariate Analysis*, 2018, 168: 334–351.
- [14] DAI L L, CHEN K N, LI G. The broken adaptive ridge procedure and its applications[J]. *Statistica Sinica*, 2020, 30(2): 1069–1094.
- [15] 钟先乐, 樊亚莉, 张探探. 基于 t 函数的稳健变量选择方法 [J]. *上海理工大学学报*, 2017, 39(6): 542–548.
- [16] FAN Y L, QIN G Y, ZHU Z Y. Variable selection in robust regression models for longitudinal data[J]. *Journal of Multivariate Analysis*, 2012, 109: 156–167.
- [17] WANG H S, LI G D, JIANG G H. Robust regression shrinkage and consistent variable selection through the LAD-Lasso[J]. *Journal of Business & Economic Statistics*, 2007, 25(3): 347–355.
- [18] 张环. Fused-LASSO 惩罚最小一乘回归的统计分析与优化算法 [D]. 北京: 北京交通大学, 2016.
- [19] EFRON B, HASTIE T, JOHNSTONE I, et al. Least angle regression[J]. *The Annals of Statistics*, 2004, 32(2): 407–499.
- [20] KIM S J, KOH K, LUSTIG M, et al. An interior-point method for large-scale ℓ_1 -regularized least squares[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2007, 1(4): 606–617.
- [21] BOYD S, PARIKH N, CHU E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers[J]. *Foundations and Trends in Machine Learning*, 2011, 3(1): 1–122.
- [22] GLOWINSKI R, MARROCO A. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires[J]. *Revue Française d'automatique, Informatique, et Recherche Opérationnelle. Analyse Numérique*, 1975, 9(R2): 41–76.
- [23] 何炳生. 凸优化和单调变分不等式收缩算法的统一框架 [J]. *中国科学: 数学*, 2018, 48(2): 255–272.
- [24] RIPPE R C A, MEULMAN J J, EILERS P H C. Visualization of genomic changes by segmented smoothing using an L_0 penalty[J]. *PLoS One*, 2012, 7(6): e38230.
- [25] MYERS C L, DUNHAM M J, KUNG S Y, et al. Accurate detection of aneuploidies in array CGH and gene expression microarray data[J]. *Bioinformatics*, 2004, 20(18): 3533–3543.
- [26] FRIDLYAND J, SNIJDERS A M, PINKEL D, et al. Hidden Markov models approach to the analysis of array CGH data[J]. *Journal of Multivariate Analysis*, 2004, 90(1): 132–153.
- [27] TIBSHIRANI R, WANG P. Spatial smoothing and hot spot detection for CGH data using the fused lasso[J]. *Biostatistics*, 2008, 9(1): 18–29.

(编辑: 丁红艺)