DOI:10.13718/j.cnki.xsxb.2023.08.005

变量选择的稳健贝叶斯 LASSO 方法①

梁韵婷, 张辉国, 胡锡健

新疆大学 数学与系统科学学院,乌鲁木齐 830046

摘要:针对数据中广泛存在的异常值会扭曲贝叶斯 LASSO 方法的参数估计和变量选择结果的问题,通过引入异方差扰动的先验设定,借此提升贝叶斯 LASSO 方法的稳健性,并推导出各参数的后验分布,利用 Gibbs 抽样得到其估计值与置信区间.该方法在数值模拟中表现出较低的拟合误差与较高的变量识别准确率,对糖尿病数据集和血浆 β-胡萝卜素水平数据集的分析表明该方法能达到简化模型与减少预测误差的平衡,实现稳健的变量选择与系数估计,并对数据中可能包含的异常值与异方差扰动有良好的抑制作用.

关 键 词:变量选择;贝叶斯 LASSO;稳健性;异常值;异方差

中图分类号:O212.8 **文献标志码**:A **文章编号**:1000-5471(2023)08-0033-08

Robust Bayesian LASSO for Variable Selection

LIANG Yunting, ZHANG Huiguo, HU Xijian

College of Mathematics and System Science, Xinjiang University, Urumqi 830046, China

Abstract: Given that the ubiquitous outliers in the data can distort the parameter estimation and variable selection results of Bayesian LASSO, the prior information of heteroscedastic disturbances is introduced to improve the robustness of Bayesian LASSO. The posterior distribution of each parameter is derived, and the estimation and confidence interval of each parameter are obtained by Gibbs sampling. The method exhibits low fitting error and high variable identification accuracy in numerical simulation, and the analyses of diabetes dataset and Plasma Beta-Carotene Level Dataset show that the proposed method achieves the balance between simplifying model and reducing prediction error. The proposed method can realize robust variable selection and coefficient estimation and has a good inhibitory effect to outliers and heteroscedastic disturbances that may be included in the data.

Key words: variable selection; Bayesian LASSO; robustness; outlier; heteroscedasticity

随着信息化时代的到来,大数据的应用越来越广泛,同时也不可避免地出现了异质性问题,表现出异方差特性.而当数据中存在异方差误差或异常点时,变量选择的结果将不再稳定.目前变量选择方法主要分为非贝叶斯方法和贝叶斯方法.基于惩罚函数的变量选择是非贝叶斯方法的主流[1-9],最常见的包括 LASSO(Least Absolute Shrinkage and Selection Operator)及其改进方法,如:EN(Elastic Net)、自适应 LAS-

① 收稿日期:2022-10-23

基金项目:国家自然科学基金项目(11961065);教育部人文社会科学研究规划基金项目(19YJA910007);新疆自然科学基金项目(2019D01C045).

作者简介:梁韵婷,硕士研究生,主要从事贝叶斯空间计量模型的研究.

SO(ALASSO)、组LASSO、SCAD(Smoothly Clipped Absolute Deviation)、MCP(Minimax Convex Penalty)、最小绝对偏差LASSO[7]等. 尽管非贝叶斯方法已经取得了不错的成果，但这类方法都不能提供令人满意的标准差估计.

文献[1]表明当回归参数具有独立且相同的拉普拉斯先验时，LASSO估计可以解释为后验众数估计. 因此，基于该联系和贝叶斯思想，文献[10]提出了贝叶斯LASSO(BLASSO)并构造了全贝叶斯分层模型和相应的采样器. 文献[11]证明在预测均方误差方面，贝叶斯LASSO的表现与频率派LASSO相似甚至在某些情况下更好. 基于文献[10-13]的研究，本文将贝叶斯LASSO与异方差误差先验相结合，以实现稳健的变量选择与系数估计，同时该法能自动产生各参数的置信区间.

1 分层模型

1.1 Gibbs采样器

考虑以下线性回归模型

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{V}) \tag{1}$$

其中：$\boldsymbol{Y}$为$n\times 1$维的因变量，$\boldsymbol{X}$为$n\times p$维的解释变量，误差$\boldsymbol{\varepsilon}$服从异方差的多元正态分布，$\boldsymbol{V}=\mathrm{diag}(V_1,\cdots,V_n)$，则该模型的似然函数如式(2)所示

$$L(\boldsymbol{Y}\mid\boldsymbol{\beta},\ \sigma^2,\ \boldsymbol{V})=(2\pi\sigma^2)^{-\frac{n}{2}}\mid\boldsymbol{V}\mid^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta})\right] \tag{2}$$

结合文献[10，12]的工作，则全模型的分层表示为

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{V})$$

$$\mathrm{p}(\boldsymbol{\beta}\mid\tau_1^2,\ \tau_2^2,\ \cdots,\ \tau_p^2)\sim \mathrm{N}(\boldsymbol{0},\ \sigma^2\boldsymbol{D}_\tau)$$

$$\boldsymbol{D}_\tau=\mathrm{diag}(\tau_1^2,\ \tau_2^2,\ \cdots,\ \tau_p^2)$$

$$\mathrm{p}(\tau_1^2,\ \tau_2^2,\ \cdots,\ \tau_p^2)\sim\prod_{j=1}^{p}\frac{\lambda^2}{2}\mathrm{e}^{-\frac{\lambda^2\tau_j^2}{2}}$$

$$\mathrm{p}(\sigma^2)\sim\frac{\gamma^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\mathrm{e}^{-\frac{\gamma}{\sigma^2}}(\alpha>0,\ \gamma>0)$$

$$\mathrm{p}\left(\frac{r}{V_i}\right)\sim \text{i. i. d. }\chi^2(r),\ i=1,\ \cdots,\ n$$

将该模型的似然函数与各参数的先验分布相乘，可得联合后验分布为

$$\begin{aligned}&\mathrm{p}(\boldsymbol{\beta},\ \sigma^2,\ \boldsymbol{V},\ \tau_1^2,\ \cdots,\ \tau_p^2\mid\boldsymbol{Y},\ \boldsymbol{X})\\ \propto{}&\mid\boldsymbol{V}\mid^{-\frac{1}{2}}(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta})\right]\frac{\gamma^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\mathrm{e}^{-\frac{\gamma}{\sigma^2}}\times\\ &\prod_{j=1}^{p}\frac{1}{(2\pi\sigma^2\tau_j^2)^{\frac{1}{2}}}\mathrm{e}^{-\frac{\beta_j^2}{2\sigma^2\tau_j^2}}\frac{\lambda^2}{2}\mathrm{e}^{-\frac{\lambda^2\tau_j^2}{2}}\times\\ &\left(\frac{r}{2}\right)^{\frac{nr}{2}}\left[\Gamma\left(\frac{r}{2}\right)\right]^{-n}\prod_{i=1}^{n}V_i^{-\frac{r+2}{2}}\mathrm{e}^{-\frac{r}{2V_i}}\end{aligned} \tag{3}$$

基于式(3)，可得$\boldsymbol{\beta}$的全条件后验分布服从均值为$\boldsymbol{B}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}$，方差为$\sigma^2\boldsymbol{B}^{-1}$的多元正态分布，其中：$\boldsymbol{B}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{D}_\tau^{-1}$；$\sigma^2$的全条件后验分布服从形状参数为$\frac{n}{2}+\frac{p}{2}+\alpha$，尺度参数为$\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta})}{2}+\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}}{2}+\gamma$的逆伽马分布；$\frac{1}{\tau_j^2}$的全条件后验分布服从形状参数为$\lambda'=\lambda^2$，均值参数为$\mu'=\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}$的逆高斯分布；文献[12]得出$\boldsymbol{V}$的全条件后验分布服从以下形式的卡方分布

$$\mathrm{p}\left(\frac{e_i^2\sigma^{-2}+r}{V_i}\,\middle|\,\boldsymbol{\beta},\ \sigma^2,\ \boldsymbol{V}_{-i},\ \tau_1^2,\ \cdots,\ \tau_p^2\right)\propto\chi^2(r+1)$$

式中e_i项为向量$\boldsymbol{e}=\boldsymbol{Y}-\boldsymbol{X\beta}$的第$i$个元素，$\boldsymbol{V}_{-i}=(V_1,\ \cdots,\ V_{i-1},\ V_{i+1},\ \cdots,\ V_n)$，$i=1,\cdots,n$. 根据各参数后

验分布可构造出稳健贝叶斯 LASSO 的 Gibbs 采样算法：

算法 1：稳健贝叶斯 LASSO 的 Gibbs 采样器

输入：$\boldsymbol{Y}$，$\boldsymbol{X}$，迭代次数 T_{draw}，预热次数 T_{omit}，初值 $\boldsymbol{\beta}_{(0)}$，$\sigma^2_{(0)}$，$\boldsymbol{\tau}^2_{(0)}$，$\boldsymbol{V}_{(0)}$

输出：$\hat{\boldsymbol{\beta}}$，$\hat{\sigma}^2$，$\hat{\boldsymbol{\tau}}^2$，$\hat{\boldsymbol{V}}$

1：$k \leftarrow 1$

2：当 $k \leqslant T_{\text{draw}}$

3：　从后验分布 $\mathrm{p}(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{X}, \sigma^2_{(k-1)}, \boldsymbol{V}_{(k-1)}, \boldsymbol{\tau}^2_{(k-1)})$ 中抽样并记为 $\boldsymbol{\beta}_{(k)}$

4：　从后验分布 $\mathrm{p}(\boldsymbol{\tau}^2 \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}_{(k)}, \sigma^2_{(k-1)}, \boldsymbol{V}_{(k-1)})$ 中抽样并记为 $\boldsymbol{\tau}^2_{(k)}$

5：　从后验分布 $\mathrm{p}(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}_{(k)}, \boldsymbol{V}_{(k-1)}, \boldsymbol{\tau}^2_{(k)})$ 中抽样并记为 $\sigma^2_{(k)}$

6：　从后验分布 $\mathrm{p}(\boldsymbol{V} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}_{(k)}, \sigma^2_{(k)}, \boldsymbol{\tau}^2_{(k)})$ 中抽样并记为 $\boldsymbol{V}_{(k)}$

7：　$k \leftarrow k+1$

8：结束

9：删去前 T_{omit} 轮样本，取后 $T_{\text{draw}} - T_{\text{omit}}$ 轮样本计算各参数的后验平均值作为估计值

1.2　超参数选取

关于超参数 λ^2 的选取，借鉴文献[10]提出的基于边际最大似然的经验贝叶斯法，具体算法如下：

1）令 $k=0$ 并设初值为 $\lambda_{(0)} = \dfrac{p\sqrt{\hat{\sigma}^2_{\text{WLS}}}}{\sum_{j=1}^{p} |\hat{\boldsymbol{\beta}}^2_{\text{WLS}}|}$，其中 $\hat{\sigma}^2_{\text{WLS}}$ 和 $\hat{\boldsymbol{\beta}}^2_{\text{WLS}}$ 为以普通线性最小二乘估计残差值的绝对值的倒数为权重的加权最小二乘估计值；

2）令 $\lambda = \lambda_{(k)}$ 并利用上述 Gibbs 采样器从 $\boldsymbol{\beta}$，σ^2，$\boldsymbol{\tau}^2$，$\boldsymbol{V}$ 的后验分布中生成第 k 轮样本；

3）利用第 k 轮样本近似计算更新 $\lambda_{(k+1)} = \sqrt{\dfrac{2p}{\sum_{j=1}^{p} E_{\lambda_{(k)}}[\tau_j^2 \mid \boldsymbol{Y}]}}$ 并令 $k = k+1$；

4）重复步骤 2）－3）直至所需的收敛水平.

由于经验贝叶斯法需要多次 Gibbs 采样，因此该法计算量极大. 文献[14]提出了一种基于随机近似的单步方法作为替代，该方法可以仅使用单次 Gibbs 采样器来获得超参数的极大似然估计，从而极大减少计算量. 该法首先作变换 $\lambda_{(k)} = \mathrm{e}^{s(k)}$，具体算法如下：

1）令 $k=0$ 并设初值为 $s_{(0)}=0$，$\boldsymbol{\theta}_{(0)} = (\boldsymbol{\beta}_{(0)}, \sigma^2_{(0)}, \boldsymbol{\tau}^2_{(0)}, \boldsymbol{V}_{(0)})$；

2）从 $K_{s_{(k)}}(\boldsymbol{\theta}_{(k)}, \cdot)$ 中生成 $\boldsymbol{\theta}_{(k+1)}$，其中 K_s 为联合后验分布 $\mathrm{p}(\cdot \mid \boldsymbol{Y}, s)$ 的 Gibbs 采样器的马尔科夫核；

3）令 $s_{(k+1)} = s_{(k)} + a_k(2p - \mathrm{e}^{2s_{(k)}} \sum_{j=1}^{p} \tau^2_{j,(k+1)})$ 令 $k = k+1$；

4）重复步骤 2）－3）直至所需的迭代次数.

其中 $\{a_k, k \geqslant 0\}$ 为一个非降的正数序列，并满足以下性质

$$\lim_{k\to\infty} a_k = 0,\ \sum a_k = \infty,\ \sum a_k^2 < \infty$$

2　数值模拟

本节将评估异方差误差先验下稳健贝叶斯 LASSO 的实验特性与优点. 根据式(1)生成数据，令 $\boldsymbol{X} = [\boldsymbol{\iota}_n, \boldsymbol{X}']$，$\boldsymbol{\iota}_n$ 为 n 维的单位向量，$\boldsymbol{X}' = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_{p-1}]$ 为多元正态分布 $\mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ 生成，其中 $\Sigma_{ij} = 0.5^{|i-j|}$. 为了考虑系数向量不同的稀释度，所有模拟均设置 $n=100$ 和 $p=50$ 并令非零系数的个数 $q \in \{10, 20\}$. 此外，为了测试收缩的适应性，一半的非零系数从正态分布 $\mathrm{N}(0, 1)$ 中生成，另一半非零系数从正态分布 $\mathrm{N}(0, 5)$ 中抽样，从而使得一半的非零系数接近于 0，另一半的非零系数则表现出更大的变化，剩余系数则设置为 0. 每次模拟均使用 5 000 次迭代并取后 2 500 次抽样计算各参数的后验均值作为估计值，为了避免偶然性，模拟均重复 100 次. 为了考察所提方法对异常值的稳健性，本文考虑了 4 种不同的 $\boldsymbol{\varepsilon}$.

例 1(异方差误差)：为了生成异方差误差，对于样本量 n 按照文献[15]生成随机组，其中组的个数由均

匀分布 U(3, 20) 抽样得出. 如果组个数大于 10, 则将该组所有样本的方差设置为等于组个数, 否则将方差设置为组个数倒数的平方, 并令 $\boldsymbol{\varepsilon}$ 的第 i 个元素为

$$\varepsilon_i = \sigma_i \xi_i$$

其中: σ_i 为第 i 个观测样本的标准差, ξ_i 来自独立同分布的标准正态分布 N(0, 1).

例 2(污染分布): $\boldsymbol{\varepsilon}$ 服从污染分布, 其中前 90% 来自标准正态分布, 后 10% 服从标准柯西分布.

例 3(柯西分布): $\boldsymbol{\varepsilon}$ 服从标准柯西分布.

例 4(拉普拉斯分布): $\boldsymbol{\varepsilon}$ 服从标准拉普拉斯分布.

为了衡量系数估计与变量选择的性能, 本文采用均方误差(MSE)与平衡准确率(BAR)作为指标. 平衡准确率能综合衡量变量选择方法正确选择、错选、漏选变量的个数, 其计算公式如下

$$BAR = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

其中 TP, TN, FP, FN 分别表示真阳性、真阴性、假阳性和假阴性的数量.

将本文提出的稳健贝叶斯 LASSO 方法简记为 RBLASSO. 表 1 列出了不施加异方差误差先验下几种常见方法与 RBLASSO 的实验结果, 其中每项指标为基于 100 次模拟的平均值. 值得注意的是, 贝叶斯方法的变量选择结果基于参数的 95% 置信区间. 若 95% 置信区间含 0, 则可认为该参数被识别为 0.

从模拟结果可得, 本文方法在大多数情况下都具有较好的综合表现, 其中当误差分布为异方差时 RBLASSO 的各项性能指标均为最优. 根据对比可得, 当非零系数的个数 q 增大时, 即系数向量越密集时, 每种方法的估计值往往会稍差, 这是因为需要用相同数量的观测值估计更多的非零参数. 当误差分布服从标准柯西分布, 即例子 3 时, 不施加异方差误差先验下的贝叶斯 LASSO 的 MSE($\hat{\boldsymbol{\beta}}$)相比其他误差分布大得多, 而 RBLASSO 依然能保持较好的系数估计与变量选择能力, 甚至在 q 增大时 MSE($\hat{\boldsymbol{\beta}}$)反而减小, 这表明了施加异方差误差先验对抵抗异常值具有重大作用.

表 1 不同模型在 4 种扰动下基于 100 次模拟试验的变量选择结果

	方法	$q=10$		$q=20$	
		MSE($\hat{\boldsymbol{\beta}}$)	BAR	MSE($\hat{\boldsymbol{\beta}}$)	BAR
Example 1	BLASSO	0.078 8	0.726 9	0.105 2	0.729 4
	LASSO	0.056 8	0.720 1	0.087 8	0.688 4
	ALASSO	0.051 0	0.734 1	0.103 8	0.705 7
	RBLASSO	0.014 8	0.837 0	0.048 4	0.793 3
Example 2	BLASSO	0.414 4	0.742 7	0.343 2	0.784 6
	LASSO	0.099 8	0.715 1	0.264 8	0.653 3
	ALASSO	0.100 8	0.764 7	0.240 6	0.769 5
	RBLASSO	0.112 4	0.768 3	0.272 4	0.808 2
Example 3	BLASSO	19.856 6	0.582 9	60.057 4	0.523 5
	LASSO	0.538 4	0.620 0	0.466 6	0.556 1
	ALASSO	0.706 2	0.607 9	0.787 2	0.552 1
	RBLASSO	0.659 4	0.636 9	0.354 2	0.587 9
Example 4	BLASSO	0.030 4	0.935 3	0.047 4	0.827 6
	LASSO	0.019 4	0.785 4	0.045 0	0.669 9
	ALASSO	0.017 8	0.852 4	0.035 2	0.813 4
	RBLASSO	0.030 2	0.924 4	0.055 2	0.813 2

3 案例研究

3.1 糖尿病数据集

将本文提出的稳健贝叶斯 LASSO 方法应用到糖尿病数据集中, 该数据集由文献[16]提供, 共有 442

个样本和 11 个变量，其中 10 个解释变量分别为年龄(age)、性别(sex)、体重指数(bmi)、平均血压(map)及 6 种血清测量 (tc,ldl,hdl,tch,ltg,glu)，因变量为基线点一年后疾病进展的定量测量. 本文所使用的数据集来自 R 包 care，所有变量均已标准化使得均值为 0、方差为 1. 为了研究所提方法的稳健性，随机选取 20%的样本在因变量上加上噪音 c，其中 c 取为 3 倍的因变量标准差，并随机划分 70%的数据集作为训练集，剩余 30%作为测试集. 评估指标采用预测均方误差(MSE)与中值绝对预测误差(MAPE).

图 1 为该数据集各变量的箱线图，初步可得解释变量和因变量均存在异常值；图 2 为学生化残差与帽子统计量关系图，其中圆圈面积与观测点的 Cook 距离成正比，垂直两条虚线分别为两倍和三倍平均帽子值的参考线，水平两条虚线分别是学生化残差为 0 及 2 的参考线，进一步分析可得该数据集中样本 295 和 305 为离群点，样本 323 和 354 为高杠杆值点，若以$\frac{4}{n-k-1}$为 Cook 距离的阈值则有 35 个强影响点.

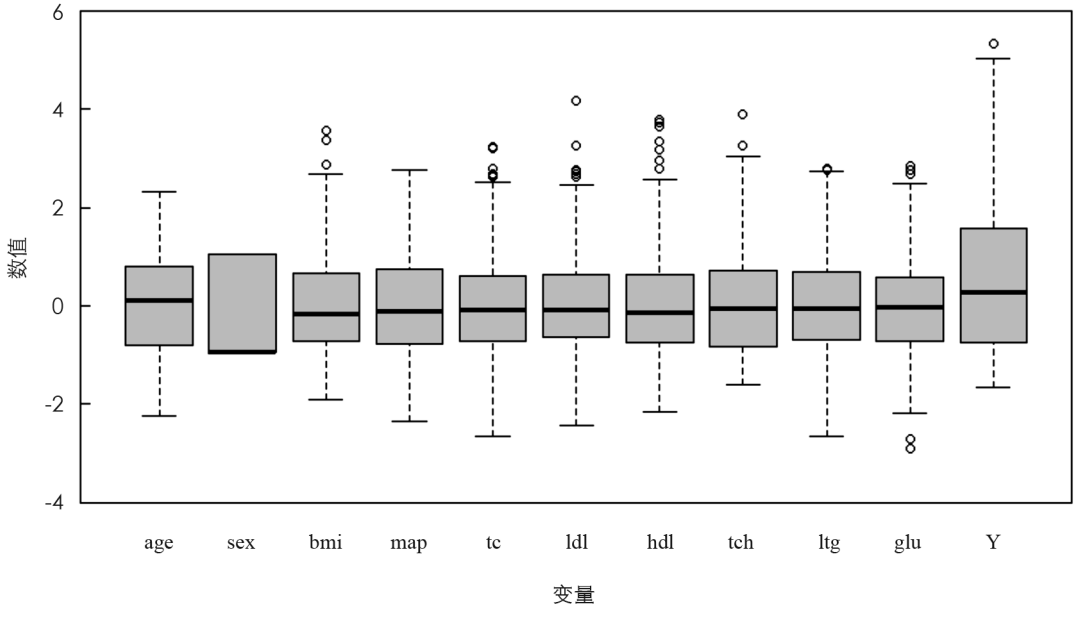

图 1　糖尿病数据集各变量的箱线图

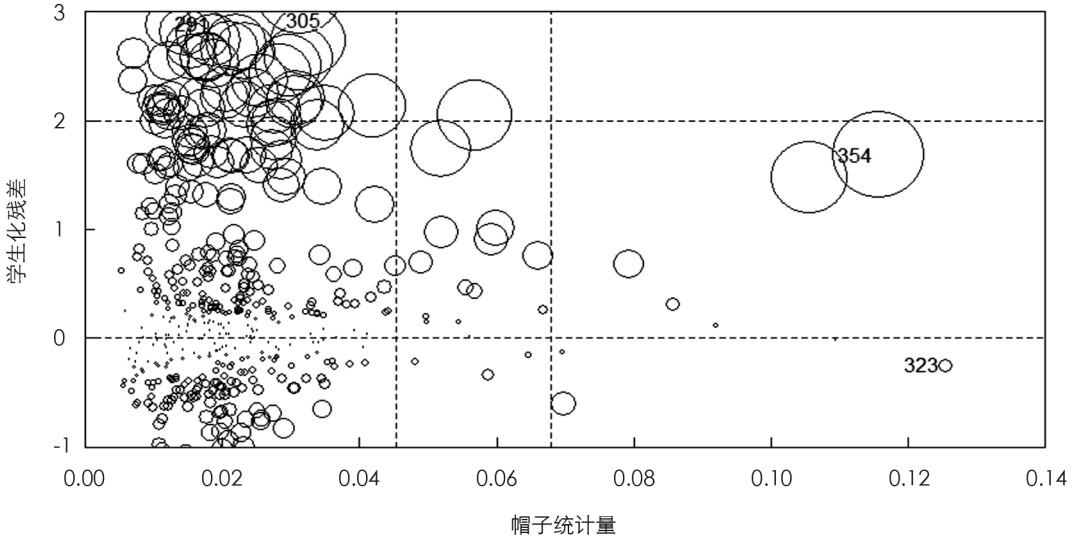

图 2　学生化残差与帽子统计量的气泡图，其中圆圈的面积表示与 Cook 距离成正比的观测值

各模型估计结果如表 2 所示，其中标粗体的系数估计值代表其置信区间含 0. BLASSO 和 RBLASSO 均排除了 7 个相同的非重要变量，而 LASSO 和 ALASSO 仅排除了 4 个非重要变量，且这 4 个非重要变量

均为4个模型所排除的共同变量，分别为sex,ldl,tch,glu. 根据MSE和MAPE，本文所提方法的预测误差最低. 此外，由图3可得相比BLASSO，施加了异方差先验的RBLASSO具有更短的置信区间. 因此，所提方法的结果应具备更高的可靠性.

表2 不同方法下糖尿病数据集的估计结果

	Least Squares	Weighted Least Squares	Bayesian LASSO	Robust Bayesian LASSO	LASSO	Adaptive LASSO
age	−0.002 6	−0.094 9	**−0.066 1**	**−0.049 1**	−0.083 1	−0.111 9
sex	0.012 0	−0.028 2	**0.004 1**	**−0.048 9**	0	0
bmi	0.440 9	0.417 5	0.415 9	0.341 7	0.431 5	0.442 8
map	0.285 0	0.251 3	0.236 8	0.164 3	0.252 5	0.273 5
tc	−1.009 8	−0.851 4	**−0.051 4**	**−0.073 5**	−0.098 7	−0.133 7
ldl	0.750 8	0.582 3	**−0.022 7**	**−0.048 5**	0	0
hdl	0.292 8	0.233 6	**−0.058 2**	**−0.059 1**	−0.037 4	−0.035 9
tch	0.006 7	0.024 3	**0.004 7**	**0.028 8**	0	0
ltg	0.775 4	0.688 1	0.355 8	0.388 2	0.390 3	0.423 6
glu	−0.015 5	−0.002 6	**0.007 7**	**0.018 6**	0	0
MSE	278.734 3	273.482 7	272.943 8	266.531 5	274.390 7	276.019 2
MAPE	0.661 4	0.550 6	0.606 6	0.548 1	0.6122	0.621 1

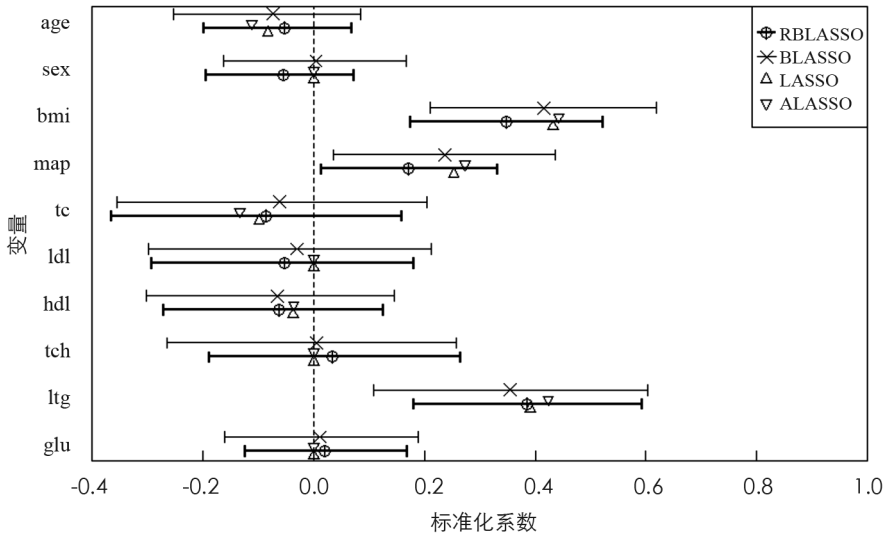

图3 不同方法下糖尿病数据集各变量的系数估计值与对应的95%置信区间

3.2 血浆 β-胡萝卜素水平数据集

文献[17]数据集包含了315名患者，均在3年内进行过活检或切除肺、结肠、乳腺、皮肤、卵巢或子宫的非癌病变，选取其中的273名女性患者作为研究对象. 该数据集共有11个变量，10个解释变量分别为年龄(age)、吸烟状态(smokstat)、Quetelet指数(quetelet)、维生素使用(vituse)、每天摄入的卡路里数(calories)、每天摄入的脂肪克数(fat)、每天摄入的纤维克数(fiber)、每周摄入的酒精饮料数量(alcohol)、胆固醇摄入量(mg/天，chol)、膳食β-胡萝卜素消耗量(mcg/d，betadiet)，因变量为血浆β-胡萝卜素 (ng/ml). 所有变量均已标准化使得均值为0、方差为1，随机划分70%的数据集作为训练集拟合模型，将剩余30%作为测试集并通过计算预测均方误差(MSE)与中值绝对预测误差(MAPE)来评估模型的预测能力.

图4和图5分别为血浆β-胡萝卜素和胆固醇的直方图，由图可得这两个变量均含有异常值. 将各模型应用于该数据，估计结果如表3所示，其中BLASSO和RBLASSO均认为quetelet，vituse和betadiet为重要变量，而LASSO和ALASSO仅排除了calories变量. 尽管RBLASSO的MAPE不是最低，但与MAPE最低的BLASSO差距甚小，且RBLASSO的MSE远低于其他方法，综合来说RBLASSO模型的预测能力

最优. 此外，从图 6 可得 RBLASSO 明显比 BLASSO 具有更短的置信区间，估计精度更高.

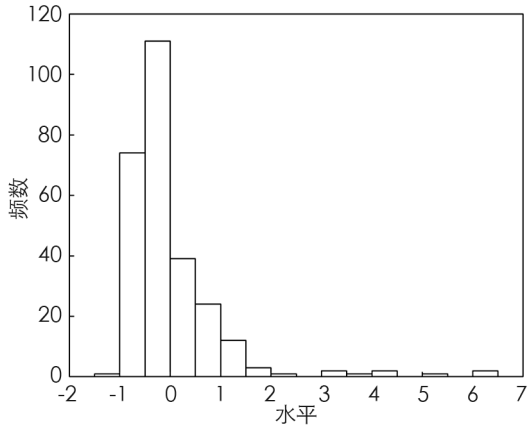

图 4　血浆胡萝卜素的直方图

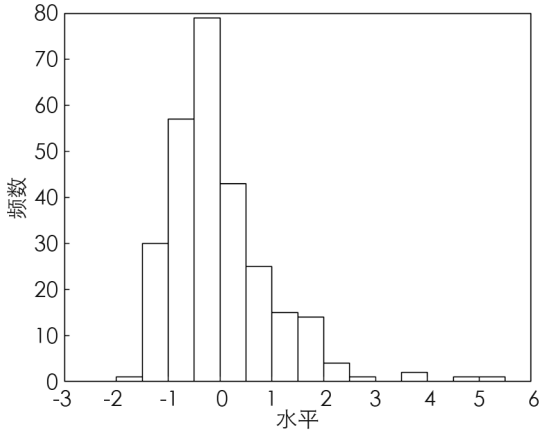

图 5　胆固醇的直方图

表 3　不同方法下血浆胡萝卜素水平数据集的估计结果

	Least Squares	Weighted Least Squares	Bayesian LASSO	Robust Bayesian LASSO	LASSO	Adaptive LASSO
age	0.062 3	0.050 7	**0.048 6**	**0.074 8**	0.054 7	0.064 1
smokstat	−0.046 0	−0.034 6	**−0.033 7**	**−0.020 1**	−0.032 8	−0.042 4
quetelet	−0.205 2	−0.181 8	−0.183 6	−0.138 0	−0.194 6	−0.202 3
vituse	−0.265 5	−0.240 0	−0.228 6	−0.136 7	−0.247 2	−0.256 4
calories	−0.080 4	−0.206 2	**−0.011 7**	**−0.025 7**	0	0
fat	−0.051 4	0.070 9	**−0.059 3**	**−0.006 2**	−0.091 1	−0.102 1
fiber	0.234 1	0.219 7	**0.169 1**	**0.049 5**	0.183 8	0.199 2
alcohol	0.160 0	0.104 4	**0.103 7**	**0.030 4**	0.128 9	0.145 3
chol	−0.046 8	−0.043 0	**−0.038 4**	**−0.016 1**	−0.040 2	−0.047 3
betadiet	0.236 0	0.222 3	0.215 0	0.152 6	0.227 3	0.235 3
MSE	34.914 1	28.592 2	29.465 3	20.370 3	32.385 3	34.267 3
MAPE	0.346 6	0.343 7	0.323 9	0.326 9	0.353 6	0.364 1

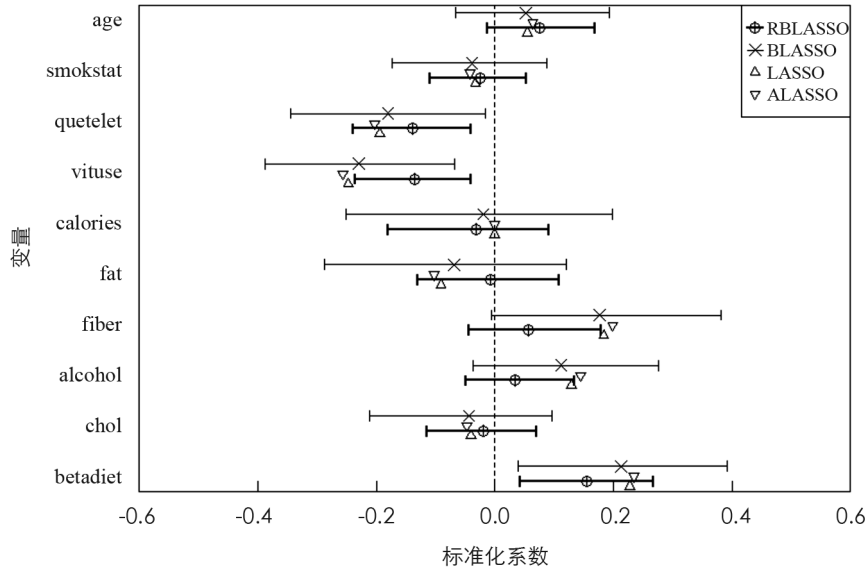

图 6　不同方法下血浆胡萝卜素水平数据集各变量的系数估计值与对应的 95%置信区间

4 结论

本文通过将异方差误差先验引入贝叶斯LASSO，提出了贝叶斯LASSO的稳健模型并建立了相应的贝叶斯分层模型与Gibbs采样器，从而提高了对异常值及异方差误差的稳健性. 数值模拟和实证分析表明当存在异常值或异方差误差时，该方法能实现较简洁的模型与较低的误差，从而实现稳健的变量选择. 此外，该模型立足于贝叶斯思想，能方便地得到估计值的置信区间，从而弥补了LASSO类方法不能给出较好可信度评估的劣势.

参考文献：

[1] TIBSHIRANI R. Regression Shrinkage and Selection via the Lasso [J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996, 58(1): 267-288.

[2] ZOU H, HASTIE T. Regularization and Variable Selection via the Elastic Net [J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005, 67(2): 301-320.

[3] ZOU H. The Adaptive Lasso and Its Oracle Properties [J]. Journal of the American Statistical Association, 2006, 101(476): 1418-1429.

[4] 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(1): 49-67.

[5] 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.

[6] ZHANG C H. Nearly Unbiased Variable Selection under Minimax Concave Penalty [J]. The Annals of Statistics, 2010, 38(2): 894-942.

[7] 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.

[8] WU Y, LIU Y. Variable Selection in Quantile Regression [J]. Statistica Sinica, 2009, 19(2): 801-817.

[9] WANG X Q, JIANG Y L, HUANG M, et al. Robust Variable Selection with Exponential Squared Loss [J]. Journal of the American Statistical Association, 2013, 108(502): 632-643.

[10] PARK T, CASELLA G. The Bayesian Lasso [J]. Journal of the American Statistical Association, 2008, 103(482): 681-686.

[11] KYUNG M, GILL J, GHOSH M, et al. Penalized Regression, Standard Errors, and Bayesian Lassos [J]. Bayesian Analysis, 2010, 5(2): 369-412.

[12] GEWEKE J. Bayesian Treatment of the Independent Student-t Linear Model [J]. Journal of Applied Econometrics, 1993, 8(S1): S19-S40.

[13] LANGE K L, LITTLE R J A, TAYLOR J M G. Robust Statistical Modeling Using the t Distribution [J]. Journal of the American Statistical Association, 1989, 84(408): 881-896.

[14] ATCHADÉ Y F. A Computational Framework for Empirical Bayes Inference [J]. Statistics and Computing, 2011, 21(4): 463-473.

[15] LIN X, LEE L F . GMM Estimation of Spatial Autoregressive Models with Unknown Heteroskedasticity [J]. Journal of Econometrics, 2010, 157(1): 34-52.

[16] EFRON B, HASTIE T, JOHNSTONE I, et al. Least Angle Regression [J]. The Annals of Statistics, 2004, 32(2): 407-499.

[17] NIERENBERG D W, STUKEL T A, BARON J A, et al. Determinants of Plasma Levels of beta-Carotene and Retinol [J]. American Journal of Epidemiology, 1989, 130(3): 511-521.

责任编辑 张枸