Statistics and Application 统计学与应用, 2023, 12(2), 391-397
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122042

Hans汉斯

基于Geman-McClure损失的稳健变量选择

谢传琪，王延新*

宁波工程学院理学院，浙江 宁波

收稿日期：2023年3月24日；录用日期：2023年4月14日；发布日期：2023年4月26日

摘 要

基于惩罚函数的最小二乘估计或似然估计是变量选择的有效方法。但当数据存在异常值时，罚最小二乘或似然估计的稳健性受到极大挑战。本文提出基于Geman-McClure损失的稳健变量选择方法，该损失函数能够有效抵制数据中异常值的影响。数值模拟和实际数据分析验证了该模型的有效性和稳健性。

关键词

稳健变量选择，高维数据，Adaptive LASSO，Geman-McClure损失

Robust Variable Selection Based on Geman-McClure Loss

Chuanqi Xie, Yanxin Wang*

School of Science, Ningbo University of Technology, Ningbo Zhejiang

Received: Mar. 24th, 2023; accepted: Apr. 14th, 2023; published: Apr. 26th, 2023

Abstract

Least squares or likelihood estimation based on penalty function is an effective method for variable selection. However, the robustness of penalized least squares or likelihood estimation is greatly challenged when there are outliers in the data. In this paper, we propose a robust variable selection method based on the Geman-McClure loss, which is an effective loss function to counteract the influence of outliers in the data. Numerical simulations and analysis of real data validate the validity and robustness of the model.

Keywords

Robust Variable Selection, High-Dimensional Data, Adaptive LASSO, Geman-McClure Loss

*通讯作者。

文章引用: 谢传琪, 王延新. 基于 Geman-McClure 损失的稳健变量选择[J]. 统计学与应用, 2023, 12(2): 391-397.
DOI: 10.12677/sa.2023.122042

1. 引言

关于变量选择问题，很多学者做了大量研究。上世纪 70 年代，人们提出信息准则方法，如 1973 年 Akaike [1]提出 AIC 准则，1978 年 Schwarz [2]在贝叶斯理论的基础下提出 BIC 准则。然而，随着变量维数的增加，基于准则选取变量的方法，计算复杂度会急剧增加，效率低下。近年来，变量选择的稀疏正则化方法已逐渐流行起来。1996 年，Tibshirani [3]提出了 LASSO (Least Absolute Shrinkage and Selection Operator)，该方法通过 L1 范数进行惩罚来压缩回归系数的大小，使绝对值较小的回归系数被自动压缩为 0。尽管 L1 范数是凸函数且易于求解，但 LASSO 为有偏估计，并且变量选择的一致性需要满足一定的不可表示条件[4] (Irrepresentable Condition)和系数 Riesz 条件[5] (Sparse Riesz Condition)。为了解决上述问题，一系列非凸正则化方法被提出。2001 年，Fan 和 Li [6]提出了 SCAD (Smoothly Clipped Absolute Deviation Penalty)罚，是一种近似无偏稀疏估计。2006 年，Zou [7]在 LASSO 的基础上提出了 Adaptive LASSO，该方法是一种 LASSO 的改进。SCAD 和 Adaptive LASSO 在一定条件下都满足 Oracle 性质。2010 年，T. Zhang [8]提出 Capped L1 实现模型的稀疏解。Zhang [9]提出了 MCP (Minimax Concave Penalty)惩罚。很多研究表明，非凸惩罚函数在理论分析以及应用中具有更优的表现[10]。

近些年，诸多高维数据变量选择研究工作的前提假设是误差分布为高斯分布或次高斯分布。然而许多实际数据如气候数据，保险理赔数据，电子商务数据等往往服从重尾分布，对于此类数据，上述方法并不适用。模型的稳健性受到极大挑战。统计学家针对具有重尾误差的情形提出了若干稀疏正则化方法，如基于 Huber 损失的高维 M 估计[11]、基于 LAD [12]或分位数的损失函数的估计、稳健 M 估计中拟似然估计方法[13]、基于梯度下降算法的稳健估计[14]、基于指数平方损失[15]或 t 型损失[16]的稳健回归、基于 Wilcoxon 得分函数的秩 LASSO 估计[17]等。

针对误差分布为重尾分布或数据存在异常值的高维模型，本文提出了一种基于 Geman-McClure 损失的稳健罚估计方法。该方法在 X 空间或 Y 空间存在离群值时，依旧能稳健且有效的进行变量选择。

2. 基于 Geman-McClure 损失的稳健模型

考虑线性回归模型

$$y = X^{\mathrm{T}}\beta + \varepsilon \tag{1}$$

其中，$y \in R^N$ 为响应变量，$X \in R^{N\times p}$ 为设计矩阵，$\beta \in R^p$ 为回归系数向量，$\varepsilon \in R^N$ 为误差向量且服从独立同分布 $\varepsilon_n \in N\left(0,\sigma^2\right)$，$n \in \{1,2,\cdots,N\}$。

高维数据变量选择的稀疏正则化方法的一般框架为：

$$\hat{\beta} = \arg\min_{\beta\in R^p}\left\{\sum_{i=1}^{n}\phi\left(y_i - x_i^{\mathrm{T}}\beta\right) + n\sum_{j=1}^{p}p_\lambda\left(\left|\beta_j\right|\right)\right\} \tag{2}$$

其中，$\phi(\cdot)$ 为损失函数，$p_\lambda\left(\left|\beta_j\right|\right)$ 为罚函数，$\lambda \geq 0$。常见的罚函数如 LASSO、SCAD、Adaptive LASSO 和 MCP 等。鉴于 Adaptive LASSO 的优良统计性质[7]，本文选取 Adaptive LASSO 作为罚函数。Adaptive LASSO 形式如下：

$$\hat{\beta} = \arg\min_{\beta\in R^p}\frac{1}{2}\left\|y - X\beta\right\|_2^2 + \lambda\sum_{p=1}^{P}\hat{w}_p\left|\beta_p\right| \tag{3}$$

其中，$\|\cdot\|_2$ 表示 L_2 范数，$\lambda \geq 0$，$\hat{w}_p \geq 1/\left|\hat{\beta}_p^{OLS}\right|^{\gamma}$ 为第 p 个回归系数的权值，$\gamma > 0$，$\hat{\beta}_p^{OLS}$ 为普通最小二乘估计的解。可以看出，Adaptive LASSO 的实现需通过两步进行：1) 先进行最小二乘估计，将系数估计值的 γ 次方的倒数作为第 p 个变量的权值；2) 对每个变量赋予“量身定做”的权值后，将权值代入(3)式进行求解。

Adaptive LASSO 采取的损失函数为平方损失，该损失注定了 Adaptive LASSO 不适合用于数据存在异常值的情形。本文在 Adaptive LASSO 以及 Geman-McClure 损失的基础上，提出一种稳健且有效的变量选择模型，形式如下：

$$\hat{\beta} = \arg\min_{\beta \in R^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi\left(y_i - x_i^{\mathrm{T}} \beta\right) + \lambda \sum_{p=1}^{P} \hat{w}_p \left|\beta_p\right| \right\} \tag{4}$$

其中，Geman-McClure 损失函数为 $\phi(t) = t^2 / \left(\gamma + t^2\right)$，其中 $t = y_i - x_i^{\mathrm{T}} \beta$。该损失函数保证正常样本情况下，其自身灵敏度的同时，又降低了对异常样本的敏感程度，提高对变量选择的稳健性。

3. 模拟研究

在本例中，模拟数据由线性回归模型生成，

$$y = X^{\mathrm{T}} \beta + \varepsilon$$

其中变量个数为 40，非零系数分别为 $\beta_1 = 3$，$\beta_2 = 1.5$，$\beta_5 = 2$，其余系数均为 0，样本大小 $n = 200$，对于每种情况，重复模拟 100 次。

情景 1： X_i 中存在异常值。每个预测值 X_i 均为 40 维正态分布的混合样本，$(1-c\%)N(0,1) + c\%N\left(3,3^2\right)$，其中 c 为异常样本比例，误差项服从标准正态分布。

情景 2： y_i 中存在异常值。每个预测值 X_i 均服从标准正态分布，误差项服从 $(1-c\%)N(0,1) + c\%N\left(10,6^2\right)$，其中 c 为异常样本比例。

情景 3： 误差项为 t 分布，每个预测值 X_i 均服从标准正态分布，误差项服从 $t(k)$，其中 $k = 2,3,5$。

为了与本文提出的方法比较，本模拟也利用 Adaptive LASSO 估计，用 10 折 CV 进行参数调节，通过使用 R 软件包“glmnet”来实现。而基于 Geman-McClure 损失和自适应 LASSO 的变量选择方法(以下简记为 GM-ALASSO)采用 BCGD 算法[18]对模型进行求解。

为了评价模拟效果，我们计算了估计系数与真实系数之间均方误差的中位数(MSE)；平均模型大小(即非零系数的数量)，MS。更好的模型选择应该产生更准确的预测结果(即较小的 MSE 值)、更正确的模型大小(即 MS)。此外，为了评估变量选择表现，我们还考虑了假阳性率(FPR)和假阴性率(FNR)，定义如下

$$\mathrm{FPR} = \frac{\#\text{of selected unimportant vαriables}}{\#\text{of selected variables}},$$

$$\mathrm{FNR} = \frac{\#\text{of removed unimportant vαriables}}{\#\text{of removed variables}}.$$

当未选择变量时，FPR 为 0，当所有变量已选择，FNR 为 0。同时，我们还计算了 Hamming 距离(HD)，其中 HD = FN + FP，其中 FN 表示非零系数被判为零系数的平均次数，FP 表示零系数被判为非零系数的平均次数。在结果上，更好的模型应具备更小的 HD。

表格中，在标记为“Under fit”的列中，我们给出了 200 次重复实验中去除了任何非零系数的比例。同理，“Correct-fit”表示正确选择模型的概率，“Over-fit”列表示选择了一些噪声变量的概率。

从表 1 可以看出，在 X 空间数据未被污染的情况下，Adaptive LASSO 和 GM-ALASSO 估计的准确率均在 0.9 附近，两种方法的估计效果均良好，这证明了 GM-ALASSO 在 X 空间无污染情况下估计的有

效性。然而，在 X 空间数据被污染的情况下，随着污染程度的增加，Adaptive LASSO 变量选择的能力急剧下降，模型过拟合程度急剧增加，而 GM-ALASSO 基本稳定在 0.7 附近，过拟合程度也显著低于 Adaptive LASSO，这证明了 GM-ALASSO 在 X 空间被污染情况下变量选择方法的稳健性。

Table 1. Scenario 1 estimated results
表 1. 情景 1 估计结果

参数	无污染		污染 5%		污染 10%		污染 20%	
	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO
MS	3.25	3.14	4.25	4.07	4.17	3.65	4.23	4.06
Underfit	0.01	0.01	0	0	0	0	0	0
Correctfit	0.86	0.92	0.68	0.75	0.56	0.73	0.53	0.71
Overfit	0.13	0.07	0.32	0.25	0.44	0.27	0.47	0.29
FPR	0.05	0.03	0.11	0.08	0.14	0.08	0.14	0.09
FNR	0.0003	0.0003	0	0	0	0	0	0
MSE (median)	0.37	0.44	0.33	0.37	0.28	0.32	0.19	0.21
HD	0.27	0.16	1.25	1.07	1.17	0.65	1.23	1.06

同理，从表 2 可看出，在 Y 空间数据未被污染的情况下，Adaptive LASSO 和 GM-ALASSO 估计的准确率均较为良好，证实 GM-ALASSO 在 Y 空间未被污染情况下的有效性。在 Y 空间被污染的情况下，随着污染程度的加剧，GM-ALASSO 的估计效果愈加强于 Adaptive LASSO，模型的过拟合程度较 Adaptive LASSO 也较低，证实 GM-ALASSO 在 Y 空间存在污染情况下的有效性。但两者估计效果在 Y 空间被污染情况下的准确率均会高于在 X 空间存在污染的情况，说明 GM-ALASSO 对 Y 空间存在异常值的抵抗力更为强大。

Table 2. Scenario 2 estimated results
表 2. 情景 2 估计结果

参数	无污染		污染 5%		污染 10%		污染 20%	
	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO
MS	3.22	3.18	3.09	3.09	3.23	3.15	3.19	3.1
Underfit	0.01	0.02	0.01	0.01	0.01	0.01	0.05	0.05
Correctfit	0.86	0.85	0.89	0.89	0.78	0.85	0.75	0.83
Overfit	0.13	0.13	0.10	0.10	0.21	0.14	0.20	0.12
FPR	0.04	0.04	0.03	0.03	0.06	0.04	0.06	0.04
FNR	0.0003	0.0005	0.0003	0.0003	0.0003	0.0003	0.001	0.001
MSE (median)	0.39	0.45	0.36	0.41	0.40	0.47	0.54	0.63
HD	0.24	0.22	0.11	0.11	0.25	0.17	0.29	0.2

同理，从表 3 可看出，随着模型重尾程度的加剧，GM-ALASSO 的变量选择能力会显著强于 Adaptive LASSO，证明了 GM-ALASSO 在数据服从重尾分布情况下的稳健性。

Table 3. Scenario 3 estimated results
表 3. 情景 3 估计结果

参数	标准正态分布		$t(2)$		$t(3)$		$t(5)$	
	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO
MS	3.38	3.18	3.30	3.16	3.33	3.19	3.30	3.26
Underfit	0	0	0.03	0.05	0.08	0.09	0.01	0
Correctfit	0.80	0.89	0.73	0.79	0.61	0.72	0.73	0.76
Overfit	0.20	0.11	0.24	0.16	0.31	0.19	0.26	0.24
FPR	0.06	0.04	0.08	0.05	0.10	0.07	0.07	0.06
FNR	0	0	0.001	0.001	0.002	0.002	0	0
MSE (median)	0.37	0.43	0.50	0.58	0.46	0.52	0.42	0.46
HD	0.38	0.18	0.36	0.26	0.49	0.37	0.32	0.26

综合而言，GM-ALASSO 在 X 空间或 Y 空间存在异常值时以及样本服从重尾分布的情况下的变量选择能力均强于 Adaptive LASSO，模型整体的稳健性及有效性均较为良好。

4. 实证分析

模拟研究

在本节中，我们将 GM-ALASSO 变量选择方法应用于波士顿房价数据。该数据是统计数据分析类非常著名的一类数据集，其中包括决定房价的结构因素、环境因素和教育因素。同时，该数据集也属于公开研究的数据，本节采用的数据集来自 R 软件自带的波士顿数据集，可通过 data (“Boston”)命令从 R 软件自带数据集中调出。包含 13 个可能影响房价的变量：crim (犯罪率)、zn (高于 25000 平方英尺房屋比率)、indus (非零售商业区比率)、nox (氮氧化物浓度)、rm (住宅平均房间数)、age (1940 年前自住房比率)、dist (与波士顿五大就业中心的加权距离)、rad (高速公路便利指数)、tax (不动产税)、ptratio(学生－老师比例)、black (黑人比例)、lstat (低教育人口比例)、chas (查尔斯河虚拟变量)。响应变量 medv (住房价格中位数)。

近年来，有许多方法对波士顿数据进行研究。例如分位数回归算法[19]、梯度下降算法[20]、SARCH 模型、QPLSIM 模型和 QPLAM 模型[21]、非参数方法和半参数方法探索数据结构[22]。这些方法都可以用来研究波士顿房价数据。波士顿数据集分为两部分：训练集有 354 个样本，测试集有 152 个样本。本文分别采用 Adaptive LASSO 和 GM-ALASSO 两种方法对波士顿房价数据进行变量选择。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 0.05×自身值。在本文中，d 分别取 0，3，5 进行实验。将两种算法变量选择后的结果进行线性建模，结果见表 4。

由表 4 可见，当数据不加污染时，虽然 GM-ALASSO 变量选择方法估计的 RMSE 以及 MAE 比 Adaptive LASSO 方法的要大，但是差别不大，说明在无污染的情况下，GM-ALASSO 变量选择方法是有效的。但当数据被污染时，GM-ALASSO 变量选择方法的 RMSE 以及 MAE 比 Adaptive LASSO 更小，且 R^2 更大，说明 GM-ALASSO 在数据由污染时更稳健。从表 5 可见，在变量选择上，GM-ALASSO 选取的变量较为

稳定，随着污染的加剧，GM-ALASSO 均只在无污染的前提下，新增变量 dis，而 Adaptive LASSO 的波动较为明显，随着污染的加剧，变量选择的改变幅度也在增大，这也说明了 GM-ALASSO 变量选择在污染情况下的稳健性。

Table 4. Experimental results
表 4. 实验结果

参数	无污染		污染 3%		污染 5%	
	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO
R^2	0.786	0.788	0.751	0.764	0.738	0.754
RMSE	4.009	4.005	4.828	4.702	4.972	4.815
MAE	3.047	3.064	3.571	3.421	3.574	3.441

Table 5. Estimated regression coefficients for Boston house price data
表 5. 波士顿房价数据的估计回归系数

变量	无污染		污染 3%		污染 5%	
	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO	Adaptive LASSO	GM-ALASSO
crim	−0.0353	0	−0.0189	0	−0.0193	0
zn	0.0078	0	0	0	0	0
indus	−0.0181	0	−0.0121	0	0	0
chas	2.6218	0	2.2084	0	1.9585	0
nox	−9.3568	0	−0.6280	0	0	0
rm	4.3012	4.3011	3.8472	5.4819	3.4846	5.8161
age	0	−0.0163	0	−0.0096	0	−0.0286
dis	−0.7356	0	−0.0650	−0.0153	0	−0.1318
rad	0	0	0	0	0	0
tax	0	−0.0108	0	−0.0101	0	−0.0080
ptratio	−0.7742	−0.3106	−0.5426	−0.4983	−0.5321	−0.5137
black	0.0054	0.0037	0.0057	0.0089	0.0049	0.0059
lstat	−0.4883	−0.2189	−0.5143	−0.2244	−0.5408	−0.1835

5. 结论

本文在 Adaptive LASSO 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，基于 Geman-McClure 损失提出一种新损失从而达到稳健的效果。模拟结果和实际数据表明，GM-ALASSO 方法能够以比较高的概率选择正确的模型且具有较小的模型误差。与传统方法相比，GM-ALASSO 方法更适合数据存在异常值的情况。此外，相关的理论性质尚未被讨论，这是未来的研究方向之一。

基金项目

宁波工程学院崇本基金项目(2022014)；宁波市自然科学基金项目(2021J143, 2021J144)。

参考文献

[1] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N. and Csaki, F., Eds., *Second International Symposium on Information Theory*, AkademiaiKiado, Budapest, 267-281.

[2] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. https://doi.org/10.1214/aos/1176344136

[3] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[4] Zhao, P. and Yu, B. (2006) On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research*, **7**, 2541-2563.

[5] Zhang, C.-H. and Huang, J. (2008) The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression. *The Annals of Statistics*, **36**, 1567-1594. https://doi.org/10.1214/07-AOS520

[6] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[7] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[8] Zhang, T. (2008) Multi-Stage Convex Relaxation for Learning with Sparse Regularization. *Proceedings of the* 21*st International Conference on Neural Information Processing Systems*, Vancouver, 8-10 December 2008, 1929-1936.

[9] Zhang, C.-H. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-AOS729

[10] Xu, X. (2010) Data Modeling: Visual Psychology Approach and $L_{1/2}$ Regularization Theory. *Proceedings of the International Congress of Mathematicians* 2010 (*ICM* 2010), Hyderabad, 19-27 August 2010.

[11] Fan, J., Li, Q. and Wang, Y. (2017) Estimation of High Dimensional Mean Regression in the Absence of Symmetry and Light Tail Assumptions. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **79**, 247-265. https://doi.org/10.1111/rssb.12166

[12] Wang, H., Li, G. and Jiang, G. (2007) Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, **25**, 347-355. https://doi.org/10.1198/073500106000000251

[13] Avella-Medina, M. and Ronchetti, E. (2018) Robust and Consistent Variable Selection in High-Dimensional Generalized Linear Models. *Biometrika*, **105**, 31-44. https://doi.org/10.1093/biomet/asx070

[14] Prasad, A., Suggala, A.S., Balakrishnan, S. and Ravikumar, P. (2020) Robust Estimation via Robust Gradient Estimation. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **82**, 601-627. https://doi.org/10.1111/rssb.12364

[15] Lozano, A.C., Meinshausen, N. and Yang, E. (2016) Minimum Distance Lasso for Robust High-Dimensional Regression. *Electronic Journal of Statistics*, **10**, 1296-1340. https://doi.org/10.1214/16-EJS1136

[16] 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.

[17] Wang, L., Peng, B., Bradic, J., Li, R. and Wu, Y. (2020) A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression. *Journal of the American Statistical Association*, **115**, 1700-1714. https://doi.org/10.1080/01621459.2020.1840989

[18] Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013) Robust Variable Selection with Exponential Squared Loss. *Journal of the American Statistical Association*, **108**, 632-643. https://doi.org/10.1080/01621459.2013.766613

[19] 陈子亮, 卿清. 影响波士顿不同社区房价水平的因素分析——基于分位数回归方法[J]. 商, 2015(30): 278-279.

[20] 陈泽坤, 程晓荣. 基于梯度下降算法的房价回归分析与预测[J]. 信息技术与信息化, 2020(5): 10-13.

[21] Guo, H., Wu, C.J. and Yu, Y. (2015) Time-Varying Beta and the Value Premium: A Single-Index Varying-Coefficient Model Approach. https://doi.org/10.2139/ssrn.2574080

[22] Qian, J. and Su, L. (2016) Shrinkage Estimation of Regression Models with Multiple Structural Changes. *Econometric Theory*, **32**, 1376-1433. https://doi.org/10.1017/S0266466615000237