Statistics and Application 统计学与应用, 2021, 10(1), 145-150
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.101014

Hans汉斯

基于SICA罚的变量选择及应用

吕鹏飞，项　超，王延新*

宁波工程学院，浙江　宁波
Email: 493012141@qq.com, *yxwang@nbut.edu.cn

收稿日期：2021年1月25日；录用日期：2021年2月19日；发布日期：2021年2月26日

摘　要

高维数据的变量选择一直是统计学领域的热门研究方向。本文研究SICA罚估计在线性模型变量选择中的应用，结合LLA (Local linear approximation)和坐标下降算法给出一种有效的迭代算法，并提出BIC准则选择正则化参数。实际数据的分析表明，与其他变量选择方法相比较，SICA方法在参数估计精度和变量选择方面具有较好的表现。

关键词

SICA罚，变量选择，参数估计，线性模型，BIC准则

Variable Selection and Application Based on SICA Penalty

Pengfei Lv, Chao Xiang, Yanxin Wang*

Ningbo University of Technology, Ningbo Zhejiang
Email: 493012141@qq.com, *yxwang@nbut.edu.cn

Received: Jan. 25th, 2021; accepted: Feb. 19th, 2021; published: Feb. 26th, 2021

Abstract

Variable selection of high-dimensional data has always been a hot research direction in the field of statistics. In this paper, we study the application of SICA penalty estimation in variable selection of linear model, give an effective iterative algorithm combined with LLA (local linear approximation) and coordinate descent algorithm, and propose BIC criterion to select regularization parameters. The analysis of actual data shows that SICA method has better performance in parameter estimation accuracy and variable selection compared with other variable selection methods.

*通讯作者。

文章引用: 吕鹏飞, 项超, 王延新. 基于 SICA 罚的变量选择及应用[J]. 统计学与应用, 2021, 10(1): 145-150.
DOI: 10.12677/sa.2021.101014

Keywords

SICA Penalty, Variable Selection, Parameter Estimation, Linear Model, BIC Criteria

1. 引言

在统计建模的过程中，我们总是希望筛选出对响应变量影响较强的变量，剔除没有影响或影响较弱的变量。但面对一些高维数据，往往存在维数灾难(curse of dimension)的现象，即数据的维数要远远多于样本量的大小。这对传统的模型选择方法提出了巨大的挑战。

早在 1970 年，Hoerl 和 Kennard 提出岭回归[1]，是对最小二乘估计的改进。岭回归通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。但岭回归不具有变量选择的功能，近年来，统计学家门提出过一系列变量选择的方法[2] [3] [4]。1996 年，受 NG (Nonnegative Garrot) [5] [6]方法的启发，Robert Tibshirani 首次提出 LASSO 方法[7]。该方法是岭回归的一种特殊形式。它通过构造一个惩罚函数来压缩一些回归系数，即强制系数绝对值之和小于某个固定值；同时设定一些回归系数为零，以此达到变量选择的目的。但 LASSO 估计是有偏估计，2001 年，Fan 和 Li [8]提出 SCAD 方法，并理论上证明了 SCAD 罚估计具有 Oracle 性质。此后，2010 年，Zhang 提出 MCP 估计[9]，也是一种近似无偏估计。

2009 年，Lv and Fan 提出了一种新的罚函数估计方法——SICA 罚[10]用于变量选择和参数估计。本文针对线性模型，研究基于 SICA 罚的变量选择，结合 LLA 和坐标下降提出一种迭代算法，并提出 BIC 准则选择正则化参数。最后通过对实际数据分析，与其他方法进行比较。

2. SICA 罚估计原理

考虑线性模型

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

其中，y 为 $n\times 1$ 的响应变量；X 为 $n\times p$ 设计阵；$\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)$，为 $p\times 1$ 维未知参数向量；ε 为 $n\times 1$ 独立同分布随机误差向量，均值为 0，方差为 δ^2。

基于模型(1)的 SICA 惩罚最小二乘定义如下：

$$Q_{(n)}(\beta) = \frac{1}{2n}\|y - X\beta\|^2 + \sum_{j=1}^{p} p_{\lambda,\tau}\left(|\beta_j|\right) \tag{2}$$

其中

$$p_{\lambda,\tau}\left(|\beta_j|\right) = \lambda\frac{(\tau+1)|\beta_j|}{|\beta_j| + \tau} \tag{3}$$

为 SICA 惩罚函数，λ, τ 为正则化参数。λ 控制罚函数的惩罚力度，τ 控制罚函数的凹凸度。当取 $\lambda = 1$，τ 分别取 0.01，0.05，0.1，绘制 SICA 罚函数图像如图 1 所示。由图可见 τ 的取值越小，SICA 罚越趋近于 L0 罚。

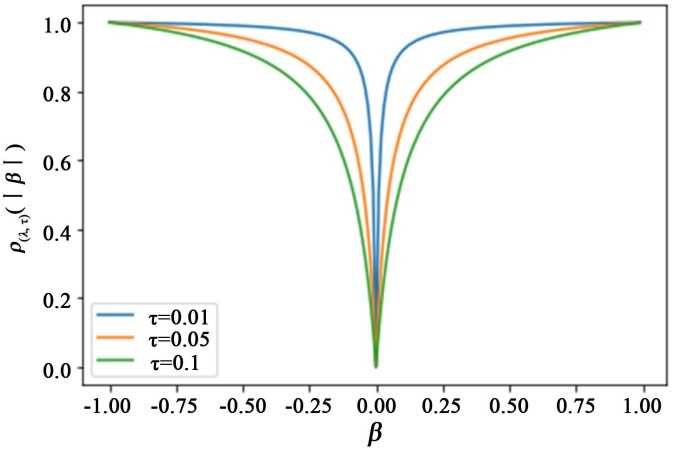

Figure 1. SICA penalty function
图 1. SICA 罚函数

对上述(2)式求极小值，得到 SICA 惩罚最小二乘估计 $\hat{\beta}=\arg\min_{\beta}Q_n(\beta)$。记 $\hat{\beta}=\{\hat{\beta}_j;j=1,\cdots,p\}$。极小化 $Q_n(\beta)$ 的结果使一些系数为 0，从而实现了保留对响应变量影响较大的变量，去除对响应变量影响较小的变量。所以我们可以通过对 $Q_n(\beta)$ 实施极小化同时达到系数估计和变量选择的效果。

3. SICA 罚估计的算法及参数选择

3.1. 算法

在本文中使用局部一次逼近(LLA)方法来求解模型。对于固定正则化参数 λ，LLA 方法采用局部线性近似的方法估计 $\sum_{j=1}^{p}P_{\alpha}(|\beta_j|)$。对于给定的初值 $\beta^0=(\beta_1^0,\cdots,\beta_p^0)^{\mathrm{T}}$，LLA 将 $\sum_{j=1}^{p}p_{\alpha}(|\beta_j|)$ 进行一阶泰勒展开：

$$\sum_{j=1}^{p}\left[p_{\lambda,\tau}(|\beta_j^0|)+p'_{\lambda,\tau}(|\beta_j^0|(|\beta_j|-|\beta_j^0|))\right] \tag{4}$$

将上式带入 SICA 罚函数得到：

$$\frac{1}{2}\|y-X\beta\|^2+\sum_{j=1}^{p}\left[p_{\lambda,\tau}(|\beta_j^{(0)}|)+p'_{\lambda,\tau}(|\beta_j^{(0)}|(|\beta_j|-|\beta_j^0|))\right] \tag{5}$$

其中 $p_{\lambda,\tau}(|\beta_j|)=\lambda\frac{(\tau+1)|\beta_j|}{|\beta_j|+\tau},j=1,2,\cdots,p$。上式其实是一种加权 LASSO 模型，然后采用坐标下降算法求最小值即得到各参数的值，模型得以求解。

3.2. 正则化参数选择

使用 SICA 方法时要确定参数 λ 和 τ。对于参数 τ 可以直接取 0.01，相当于 SCAD 中取 $a=3.7$ [11]。所以我们只需要固定 $\tau=0.01$，选取适当的参数 λ 即可。AIC 准则[12]，GCV 准则[13]是常用的参数选择方法，但其不具有一致性，易发生过拟合的现象。BIC 准则没有这一缺点，所以本文提出基于 BIC 准则的λ的选取：

$$\mathrm{BIC}_{\lambda}=\log\frac{\|y-X\beta_{\lambda}\|_2^2}{n-k}+\frac{df_{\lambda}}{n}\log(n) \tag{6}$$

其中 df_{λ} 为广义自由度：

$$df_\lambda = tr\left\{X\left(X'X + n\sum\lambda\right)^{-1}X\right\}$$

其中 $\sum\lambda = diag\left(p'_\lambda\left(\left|\hat{\beta}_{\lambda 1}\right|\right)\Big/\left|\hat{\beta}_{\lambda 1}\right|, \cdots, p'_\lambda\left(\left|\hat{\beta}_{\lambda k}\right|\right)\Big/\left|\hat{\beta}_{\lambda k}\right|\right)$，$k$ 为模型中非零参数个数，n 为样本数量，取 $\hat{\lambda} = \arg\min_\lambda\left(\mathrm{BIC}_\lambda\right)$。

4. 实际数据应用

实际数据分析

本文使用 Python 的 Keras 库中的数据集 Boston_House，该数据集包含美国人口普查局收集的美国马萨诸塞州波士顿住房价格的有关信息，共 506 个样本，13 个变量。通过多种方法对房价与各自变量之间建立回归模型。对数据的具体描述如下：

X_1 代表犯罪率；X_2 代表住宅用地所占比例；X_3 代表非零售地区所占比例；X_4 代表是否在河边(0 代表不是，1 代表是)；X_5 代表一氧化氮浓度；X_6 代表平均每居民房数；X_7 代表建筑年龄；X_8 代表与市中心的距离；X_9 代表公路可达指数；X_{10} 代表物业税率；X_{11} 代表城镇师生比例；X_{12} 代表黑人比例；X_{13} 代表低收入人口所占比例。

首先将数据以 7:3 的比例划分为训练集与测试集。分别采用经典最小二乘法，岭回归，LASSO 回归，SICA 方法对测试集进行回归建模，将所得模型运用于测试集做预测。从变量选择的效果与预测准确性来比较不同方法的优劣，准确性以 MSE 作为判断标准。

MSE 称为均方误差(Mean Square Error)，是真实值与预测值的差值的平方然后求平均数，计算公式如下：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{7}$$

其中 y_i 表示真实值，$\hat{y}_i$ 表示预测值。

SICA 方法首先要确定参数 λ 的值，事先给定 λ 一个取值范围，查看不同 λ 下 BIC 的值，取使得 BIC 最小的 λ 为参数。BIC 随着参数 λ 变化而变化的结果如图 2 所示：

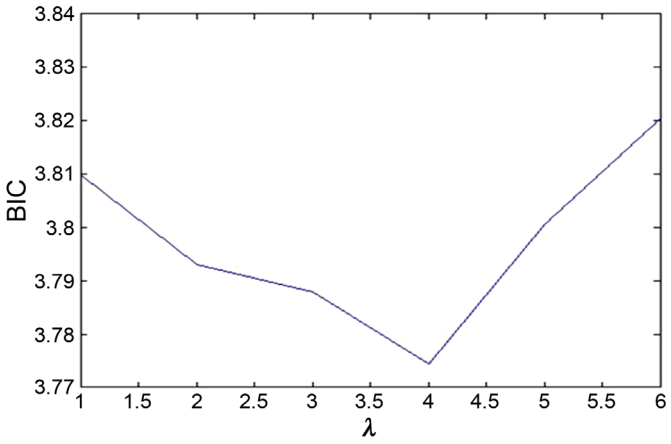

Figure 2. BIC curve
图 2. BIC 变化曲线

可以看出 BIC 曲线呈现 V 型，说明已经找到使得 BIC 最小的 λ 值。模型拟合的系数和 MSE 如表 1，表 2 所示：

Table 1. MSE under different methods
表 1. 不同方法的 MSE

	最小二乘	岭回归	LASSO	逐步回归	SICA
MSE	45.0775	61.1200	96.5622	36.8327	44.9028

Table 2. Coefficient under different methods
表 2. 不同方法的变量系数

	最小二乘	岭回归	LASSO	逐步回归	SICA
X_1	0.4007	0.7135	0	0	0
X_2	2.6494	1.7956	0.8958	0.1033	2.5856
X_3	−2.2908	−1.2861	−0.3943	−0.2543	−2.4047
X_4	1.3066	1.1989	0.1809	3.6142	1.3564
X_5	2.2435	0.2506	0	0	0
X_6	5.7128	1.1778	0	0	6.4960
X_7	2.0795	0.3829	0	0	2.2103
X_8	−10.0440	−6.9809	−3.4737	−1.8660	−9.8784
X_9	−1.2864	0.7831	0	0.0334	0
X_{10}	−0.9701	−1.6955	0	−0.0073	0
X_{11}	0.1385	0.0590	0	0	0
X_{12}	10.2976	1.1828	2.6773	0.0099	9.8597
X_{13}	−8.7597	−4.3576	−9.4379	−0.7102	−8.6776

SICA 方法剔除了 5 个变量，并且所得模型预测结果的 MSE 较小。这说明相对于最小二乘回归和岭回归，SICA 方法准确的剔除了对因变量影响较小的因素，保留了影响较大的因素。LASSO 方法虽然剔除了更多的变量，但是 MSE 却是最大的，说明剔除了应该保留的变量，变量选择的能力上不如 SICA 方法。逐步回归方法剔除了 5 个变量，MSE 也小于 SICA 方法，说明对于这份数据，逐步回归的变量选择与模型预测优于 SICA 方法。综上，SICA 方法无论是变量的选择还是模型的预测精度都优于大部分的传统方法。

5. 结论

本文讨论了 SICA 方法在线性模型的变量选择和参数估计中的应用。通过对实际数据的分析，可以发现相对于大部分的传统方法，本文提出的 SCIA 方法有更强的变量选择能力，对参数的估计具有更高的精度。

基金项目

全国统计科学研究项目(2019LY06)；浙江省统计研究课题(20TJZZ18)；浙江省自然科学基金资助项目(LY18A010026)；浙江省大学生科技创新活动计划暨新苗人才计划资助项目(2020R475013)。

参考文献

[1] Vinod, H.D. (2020) What's the Big Idea? *Ridge Regression and Regularisation*, **17**, 41-41.

https://doi.org/10.1111/1740-9713.01472

[2] 曾津, 周建军. 高维数据变量选择方法综述[J]. 数理统计与管理, 2017(4): 678-692.

[3] 王大荣, 张忠占. 线性回归模型中变量选择方法综述[J]. 数理统计与管理, 2010(4): 615-627.

[4] 李根, 邹国华, 张新雨.高维模型选择方法综述[J]. 数理统计与管理, 2012, 31(4): 640-658.

[5] Breiman, L. (1995) Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, **37**, 373-384. https://doi.org/10.1080/00401706.1995.10484371

[6] Yuan, M. and Lin, Y. (2007) On the Nonnegative Garrote Estimator. *Journal of the Royal Statistical Society* (*Series B*), **69**, 143-161. https://doi.org/10.1111/j.1467-9868.2007.00581.x

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

[8] Fan, J.Q. and Li, R.Z. (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

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

[10] Lv, J.C. and Fan, Y.Y. (2009) A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares. *Annals of Statistics*, **37**, 3498-3528. https://doi.org/10.1214/09-AOS683

[11] 严奇琪, 王延新. 高维部分线性小波模型中的变量选择[J]. 宁波工程学院学报, 2018(2): 13-18.

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

[13] 张肖萍, 吴炜明, 王延新. 高维数据变量选择中 MCP 正则化参数选择研究[J]. 统计学与应用, 2019, 8(6): 852-858.