**Statistics and Application** 统计学与应用, 2019, 8(6), 852-858
Published Online December 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.86096

Hans汉斯

# MCP Regularization Parameter Selection in High Dimensional Data Variable Selection

**Xiaoping Zhang, Weiming Wu, Yanxin Wang***

Ningbo University of Technology, Ningbo Zhejiang
Email: *wyxinbj@163.com, *1156957967@qq.com

Received: Nov. 12th, 2019; accepted: Nov. 25th, 2019; published: Dec. 2nd, 2019

## Abstract

**In the era of big data, variable selection of high-dimensional data is one of the hot topics in modern statistics. The MCP regularization method is a commonly used variable selection method, but the merits of the MCP regularization method depend on whether the optimal regularization parameter can be selected. Based on the BIC criterion of regularization parameter selection, an MBIC criterion is proposed for MCP regularization parameter selection. Through data simulation and practical application, the MCP method with MBIC criterion can select the correct model with higher probability, which is obviously superior to other regularization parameter selection methods.**

## Keywords

**Variable Selection, MBIC, MCP, Regularization Parameter, High-Dimensional Data**

# 高维数据变量选择中MCP正则化参数选择研究

**张肖萍，吴炜明，王延新***

宁波工程学院，浙江 宁波
Email: *wyxinbj@163.com, *1156957967@qq.com

收稿日期：2019年11月12日；录用日期：2019年11月25日；发布日期：2019年12月2日

## 摘　要

**大数据时代，高维数据的变量选择是现代统计的研究热点问题之一。MCP正则化方法是常用的变量选取方法，但MCP正则化方法的优劣取决于能否选取出最优的正则化参数。本文在BIC准则的基础上，提出适用于MCP正则化参数选择的MBIC准则。通过数据模拟及实际应用表明，MCP方法在MBIC准则下能够**

*通讯作者。

文章引用: 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. DOI: 10.12677/sa.2019.86096

**以更高的概率选择正确的模型，MBIC准则明显优于其它参数选择方法。**

## 关键词

**变量选择，MBIC，MCP，正则化参数，高维数据**



## 1. 引言

随着大数据时代的来临，庞大的数据资源吸引了越来越多领域的关注。各行各业都希望通过数据处理与挖掘发现数据隐含的信息，为相关决策提供现实依据。特别在分析建模中，为了全面而准确地反应信息的特征及其内在规律，常常引入多个指标，进而形成高维数据。然而并不是高维数据中的所有信息都是有效的，过多的变量反而会导致模型复杂度提升，以至于模型拟合效果和预测精度的降低。因此，如何从海量的高维数据中提取有用特征是一个亟待解决的问题。变量选择就是这样一种从大量信息中提取相关变量从而建立稀疏稳健模型的技术。

传统的变量选择方法如最佳子集选择或逐步向前向后回归，需要结合 AIC [1]，BIC [2]等准则。但在高维数据下，容易出现难以克服的 NP-Hard 问题。为了克服传统方法的缺陷，统计学家们提出了众多基于惩罚函数的变量选择方法[3]，如：Lasso 估计(Least Absolute Shrinkage and Selection Operator) [4]，SCAD 估计(Smoothly Clipped Absolute Deviation) [5]，MCP 估计(Minimax Concave Penalty) [6]等。MCP 由 Zhang 等提出用于高维数据的变量选择，MCP 估计满足变量选择的 Oracle 性质，即一致地选择出正确的模型，且参数的估计满足渐进正态性。在实际应用中，MCP 方法优劣取决于能否选择合适的正则化参数 $\lambda$，正则化参数越小，模型复杂程度越高；正则化参数越大，模型精确程度越低。因此，如何选择合适的正则化参数 $\lambda$ 是一个至关重要的问题。常见的正则化参数选择方法有交叉验证(CV)，广义交叉验证(GCV) [4]以及 AIC，BIC [7]等信息准则。Wang *et al.* [8]考虑到 GCV 方法的过拟合性，提出了参数选择的 BIC 准则，并从理论上证明了模型选择的一致性。SCAD 等常借助 BIC 准则选择正则化参数[9]，但该准则对于 MCP 估计未必能选出最优的模型。

鉴于上述原因，本文通过对 BIC 准则进行改进，提出一种更适合于 MCP 正则化参数选择的修正 BIC 准则(MBIC)。通过数据模拟，比较 MBIC 准则与 BIC 准则在 MCP 方法中的效果。最后，讨论不同方法在实际数据中的应用，分析了 1986 和 1987 年赛季美国职业棒球大联盟的棒球运动员收入数据，探究与美国棒球运动员收入相关的影响因素。

## 2. 罚估计方法

考虑线性模型

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i,\ i = 0,1,\cdots,n$$

其中 $y_i$ 是第$i$个响应变量，$x_i$ 是 $p\times1$阶的协变量，$\varepsilon_i$ 是均值为 0，方差为 $\sigma^2$ 的 *i.i.d* 的随机误差项。为了同时进行变量选择和参数估计，常采用很多基于罚函数的稀疏正则化方法，其一般框架为

$$L(\beta;\lambda) = \|y - X\beta_\lambda\|_2^2 + n\sum_{j=1}^{p} p\left(|\beta_j|;\lambda\right) \quad (1)$$

其中 $p\left(\left|\beta_j\right|;\lambda\right)$ 表示惩罚函数。

### 2.1. MCP 估计

2010 年，CUN-Hui Zhang 提出 MCP [6]，MCP 是一种非凸罚函数，在 $[0,\infty)$ 的定义为

$$p\left(\beta_j;\lambda\right)=\lambda\int_0^{\beta_j}\left(1-\frac{x}{\gamma\lambda}\right)_+\mathrm{d}x$$

其一阶导数为

$$p_{\lambda,\gamma}\left(\beta_j\right)=\begin{cases}\lambda-\dfrac{\beta_j}{\gamma}, & \text{if } 0<\beta_j<\gamma\lambda\\ 0, & \text{if } \beta_j>\gamma\lambda\end{cases} \tag{2}$$

其中 $\lambda\geq 0$ 和 $\gamma>1$ 为正则化参数。

结合 MCP 罚函数的一阶导数的形式，可以看出 MCP 从 0 到 $\gamma\lambda$ 惩罚力度呈线性下降趋势，当 $\beta_j>\gamma\lambda$ 是惩罚力度为 0，即不惩罚。MCP 罚函数满足近似连续性，稀疏性和无偏性。

### 2.2. 其他罚函数

Lasso 方法对参数的 L1 范数进行惩罚，Lasso 的惩罚项为 $p_\lambda\left(\left|\beta_j\right|\right)=\lambda\left|\beta_j\right|$，估计形式为 $\min_\beta\left\|y-X\beta_\lambda\right\|^2$, s.t. $\|\beta\|_1\leq t$ 其中 $\|\beta\|_1=\sum_{i=1}^p|\beta_i|$，上式也等价于

$$\hat{\beta}=\arg\min_\beta\left\{\frac{1}{n}\left\|y-X\beta_\lambda\right\|_2^2+\lambda\|\beta\|_1\right\}$$

但是，Lasso 对较大系数的估计是有偏估计，并且 Lasso 估计也不满足变量选择的 Oracle 性质。2001 年 Fan and Li 提出 SCAD 方法[5]，同时证明了其满足变量选择的 Oracle 性质。与 Lasso 相比，SCAD 是无偏估计，因而受到广泛关注。SCAD 罚函数的惩罚项为

$$p_\lambda\left(\left|\beta_j\right|\right)=\begin{cases}\lambda\left|\beta_j\right|, & \left|\beta_j\right|\leq\lambda\\ -\dfrac{\left|\beta_j\right|^2-2\alpha\lambda\left|\beta_j\right|+\lambda^2}{2(\alpha-1)}, & \lambda\leq\left|\beta_j\right|\leq\alpha\gamma\\ \dfrac{(\alpha+1)^2\lambda}{2}, & \left|\beta_j\right|>\alpha\gamma\end{cases} \tag{3}$$

其中 $\lambda\geq 0$ 和 $\alpha>1$ 为正则化参数，在实际应用中常取 $\alpha=3.7$。

## 3. 正则化参数选择方法

在实际应用中，正则化模型(1)的优劣与正则化参数 $\lambda$ 取值密切相关，不同的参数 $\lambda$ 会导致不同的惩罚力度，进而影响最终的模型。因此，参数 $\lambda$ 的选择至关重要。常见的选择参数 $\lambda$ 的方法有 CV，GCV 和各种信息准则，如 AIC 及 BIC 等。

针对 LASSO 估计，Zou H. *et al.* [8]给出了估计的自由度，并提出了适用于 Lasso 估计的 BIC 准则，定义如下

$$\mathrm{BIC}(\lambda)=\frac{\left\|y-X\beta_\lambda\right\|_2^2}{n\sigma^2}+\frac{\log(n)}{n}\widehat{df}_\lambda \tag{4}$$

其中 $\widehat{df}_\lambda=\sum_{i=1}^n cov\left(x_i\beta_\lambda-y_i\right)/\sigma^2$，$\sigma^2=Var(\varepsilon)$。

此外，Wang *et al.* [7]证明了 GCV 方法易出现模型选择的过拟合现象，针对 SCAD 估计提出了 BIC 准则。BIC 准则是在有限的模型集合中的模型选择准则，BIC 准则认为具有最小 BIC 值的模型是模型集合中最优良的模型。BIC 具体定义如下

$$\mathrm{BIC}(\lambda)=\log\frac{\left\|y-X\beta_{\lambda}\right\|_2^2}{n}+\frac{\widehat{df}_{\lambda}}{n}\log(n) \tag{5}$$

其中 $\widehat{df}_{\lambda}$ 为广义自由度，

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

其中 $\sum\lambda=diag\left(p'_{\lambda}\left(\left|\hat{\beta}_{\lambda 1}\right|\right)\Big/\left|\hat{\beta}_{\lambda 1}\right|,\cdots,p'_{\lambda}\left(\left|\hat{\beta}_{\lambda d}\right|\right)\Big/\left|\hat{\beta}_{\lambda d}\right|\right)$。

但在实际应用中，MCP 估计在 BIC 准则下选择了较为复杂的模型，故本文提出 MBIC 准则，定义如下

$$\mathrm{MBIC}(\lambda,\alpha)=\log\left(\frac{\left\|y-X\beta_{\lambda,\alpha}\right\|_2^2}{n-p_0}\right)+\frac{p_0}{n}\log(n) \tag{6}$$

其中，$p_0$ 表示非零变量个数。

## 4. 模型研究和实际数据分析

### 4.1. 模拟研究

小节通过模拟实验比较 LASSO，SCAD，MCP 变量选择方法的性能。

考虑线性模型

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

进行随机模拟，从而产生数据 $x$ 和 $y$。在模拟实验中，$n=200$，$\varepsilon\sim N(0,1)$，$\sigma=2$，变量个数 $p$ 分别取 8，12，20，且 $\beta=(3,1.5,0,0,2,0,0,0,\cdots)_{1\times p}^{\mathrm{T}}$，$x_i,x_j$ 之间的相关系数为 $cor(j_1,j_2)=0.5^{|j_1-j_2|}$。

算法上，Lasso 估计，SCAD 估计和 MCP 估计均采用坐标下降算法[10]。MCP 估计分别利用 BIC 准则(5)，MBIC 准则(6)选择正则化参数，SCAD 估计采用 BIC 准则(5)，而 LASSO 估计采用 BIC 准则(4)选择正则化参数。所有模拟实验重复进行 100 次，模拟结果如表 1 所示。

为比较 Lasso、SCAD、MCP 估计精确性，给出模型误差公式

$$ME(\hat{\mu})=\left(\hat{\beta}-\beta\right)^{\mathrm{T}}E\left(XX^{\mathrm{T}}\right)\left(\hat{\beta}-\beta\right)$$

其中，“MME”表示 100 次重复实验中模型误差 ME 的中位数；“SD”表示 100 次重复实验中模型误差 ME 的标准差；“C”表示 100 次重复实验中非零系数被正确估计为非零个数的均值；“IC”表示 100 次重复实验中零系数被错误估计为非零个数的均值；“Underfit”表示欠拟合，即在 100 次模拟实验中将非零系数错误估计为零的比例；“Correctfit”表示正确拟合，即在 100 次模拟实验中将非零系数正确估计为非零的比例；“Overfit”表示过拟合，即 100 次模拟实验中选择了 3 个重要变量并且包含了非零系数的比例。

从表 1 可以看出，在 BIC 正则化参数选择方法下，MCP 估计和 SCAD 估计方法在变量选择和模型误差方面优于 LASSO 方法。在模型误差中，所有变量选择的方法均能减小模型误差，而 MCP 方法在

MBIC 准则下具有最小的模型误差，而且能够以更高概率选择真实模型。综上，在 MBIC 正则化参数选择方法下，MCP 估计在变量选择能力和模型误差方面均最优。

**Table 1.** Simulation results
**表 1.** 模拟结果

| p | 准则 | MME | SD | C | IC | Underfit | Correctfit | Overfit |
|---|---|---|---|---|---|---|---|---|
| 8 | Lasso_BIC | 0.0979 | 0.0692 | 3 | 1.6200 | 0 | 0.0800 | 0.9200 |
| | SCAD_BIC | 0.0584 | 0.0645 | 3 | 0.3300 | 0 | 0.7500 | 0.2500 |
| | MCP_BIC | 0.0482 | **0.0489** | 3 | 0.3700 | 0 | 0.6900 | 0.3100 |
| | MCP_MBIC | **0.0429** | 0.0546 | 3 | **0.0400** | 0 | **0.9600** | **0.0400** |
| 12 | Lasso_BIC | 0.1276 | 0.0911 | 3 | 2.3600 | 0 | 0.0650 | 0.9350 |
| | SCAD_BIC | 0.0664 | 0.0581 | 3 | 0.9300 | 0 | 0.5000 | 0.5000 |
| | MCP_BIC | 0.0726 | **0.0570** | 3 | 0.7050 | 0 | 0.5450 | 0.4550 |
| | MCP_MBIC | **0.0530** | 0.0594 | 3 | **0.0650** | 0 | **0.9400** | **0.0600** |
| 20 | Lasso_BIC | 0.1394 | 0.0885 | 3 | 3.5300 | 0 | 0.0000 | 1.0000 |
| | SCAD_BIC | 0.0753 | **0.0688** | 3 | 1.9500 | 0 | 0.3300 | 0.6700 |
| | MCP_BIC | 0.0840 | 0.7210 | 3 | 1.4100 | 0 | 0.3900 | 0.6100 |
| | MCP_MBIC | **0.0446** | 0.0690 | 3 | **0.1000** | 0 | **0.9300** | **0.0700** |

## 4.2. 实际数据分析

本文利用来自 R 语言的 ISLR 包中的数据集 Hitters，该数据集包含 20 个变量，322 次样本，描述关于 1986 和 1987 赛季的棒球大联盟中的棒球运动员收入的相关信息。数据集各个变量描述如下：

X1：1986 年击球的次数；X2：1986 年的点击次数；X3：1986 年的本垒打数量；X4：1986 年的运行次数；X5：1986 年击败的次数；X6：1986 年的散步次数；X7：联赛的年份；X：职业生涯中击球的次数；X9：职业生涯中的点击次数；X10：职业生涯中的本垒打数量；X11：职业生涯中的跑步次数；X12：职业生涯中击球的次数；X13：职业生涯中的散步次数；X14：表示 1986 年底的球员联赛 A 级和 N 级的因素；X15：表示 1986 年底的分裂 E 和 W 等级的因素；X16：1986 年的罢工数量；X17：1986 年的助攻数量；X18：1986 年的错误数量；X19：表示 1987 年初的球员联赛 A 级和 N 级的因素；Y：1987 年开业日的年薪数(千美元)。

假设棒球运动员在各个赛季的表现与棒球运动员收入呈线性关系，即有如下线性模型

$$y_i = \sum_{j=1}^{20} x_{ij}\beta_j + \varepsilon_i,\ i = 0,\cdots,n$$

其中 $y_i$ 表示第 $i$ 个运动员的收入，$x_{ij}$ 是他的第 $j$ 个变量，$\varepsilon_i$ 是均值为 0，方差为 $\sigma^2$ 的 $i.i.d$ 的随机误差项。

利用最小二乘估计(OLS)、Lasso、SCAD 和 MCP 估计分析该数据。变量选择结果如表 2 所示。从表 2 可以看出，无罚的最小二乘估计(OLS)选择了所有的变量，Lasso 选择了 15 个变量，SCAD 选择了 6 个变量，基于 BIC 准则的 MCP 估计选择了 8 个协变量，而基于 MBIC 准则的 MCP 估计选择了 7 个协变量，选择了相对稀疏的模型。从参数估计的结果看，基于 MBIC 的 MCP 估计的结果更接近于最小二乘估计值。

**Table 2.** Parameter estimation under different methods
**表 2.** 不同方法下的参数估计

| 变量 | OLS | LASSO_BIC | SCAD_BIC | MCP_BIC | MCP_MBIC |
|---|---|---|---|---|---|
| $X_1$ | −852.8903 | 33.4617 | 0 | 0 | 0 |
| $X_2$ | 963.6420 | 66.0012 | 101.7722 | 137.8352 | 130.4466 |
| $X_3$ | 72.9293 | 14.9911 | 0 | 0 | 0 |
| $X_4$ | −177.1966 | 56.9189 | 0 | 0 | 0 |
| $X_5$ | −82.1408 | 48.1183 | 0 | 0 | 0 |
| $X_6$ | 337.3167 | 67.8294 | 36.5964 | 124.7608 | 119.1098 |
| $X_7$ | 27.9380 | 2.2523 | 0 | 0 | 0 |
| $X_8$ | −742.3870 | 43.0421 | 0 | 29.8314 | 97.8530 |
| $X_9$ | 183.5095 | 64.3198 | 190.8192 | 151.0980 | 119.9879 |
| $X_{10}$ | −20.0722 | 52.7284 | 20.9577 | 128.7815 | 113.1012 |
| $X_{11}$ | 794.4982 | 66.7017 | 0 | 111.6680 | 0 |
| $X_{12}$ | 416.9650 | 67.5518 | 0 | 0 | 0 |
| $X_{13}$ | −377.9468 | 20.5463 | 0 | 0 | 0 |
| $X_{14}$ | 131.7961 | 77.7381 | 69.4525 | 108.7831 | 107.0552 |
| $X_{15}$ | 70.7693 | 0 | 0 | 0 | 0 |
| $X_{16}$ | −30.3442 | 0 | 0 | 0 | 0 |
| $X_{17}$ | −54.4371 | 0 | 0 | 0 | 0 |
| $X_{18}$ | 100.5867 | 58.8511 | 54.1784 | 73.6822 | 84.1708 |
| $X_{19}$ | 29.6243 | 0 | 0 | 0 | 0 |

## 5. 结论

本文讨论了 MCP 方法在变量选择和参数估计的应用，提出了更适合 MCP 估计的 MBIC 准则。数据模拟以及实际数据分析中都表明在 MBIC 准则下 MCP 估计的结果更优于 BIC 准则估计结果。

## 基金项目

浙江省自然科学基金资助项目(LY18A010026)；全国统计科学研究项目(2019LY06)；宁波市自然科学基金资助项目(2017A610143)；国家级大学生创新创业训练计划项目(201911058025)；王伟明助创基金资助项目(2018020)；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(2018R428027)。

## 参考文献

[1] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. 2*nd Interernationl Symp. on Information Theory*, 1973.

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

[3] 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, 36(4): 678-692.

[4] Tibshirani, R.J. (1996) Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society, Series B*: *Methodological*, **73**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[5] 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

[6] 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

[7] Zou, H., Hastie, T. and Tibshirani, R. (2007) On the "degrees of freedom" of the Lasso. *The Annals of Statistics*, **35**, 2173-2192. https://doi.org/10.1214/009053607000000127

[8] Wang, H., Li, R. and Tsai, C.L. (2007) Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, **94**, 553-568. https://doi.org/10.1093/biomet/asm053

[9] 周荣旺. SCAD 方法的调整参数选择[D]: [硕士学位论文]. 大连: 大连理工大学, 2010.

[10] Breheny, P. and Huang, J. (2011) Coordinate Descent Algorithms for Nonconvex Penalized Regression with Applications to Biological Feature Selection. *Annals of Applied Statistics*, **5**, 232-253. https://doi.org/10.1214/10-AOAS388