

Cox 比例风险模型的弹性 SCAD 变量选择方法及应用*

孙志华¹, 谢向东², 焦东方¹

(1. 中国海洋大学数学科学学院, 山东 青岛 266100; 2. 东北师范大学数学与统计学院, 吉林 长春 130024)

摘要: 本文对基于 Cox 比例风险模型的弹性 SCAD 变量的选择方法进行了研究, 证明了弹性 SCAD 方法在进行变量选择时具有 Oracle 性质、渐近正态性及组效应。通过数值模拟, 将弹性 SCAD 与几种常见变量选择方法作了比较, 得出几种不同参数设定下的变量选择结果。通过肺癌实际数据比较了弹性 SCAD 方法和其他常见变量选择方法。研究表明, 弹性 SCAD 方法选出的变量与临床结论一致并发现了需要进一步深入研究的变量。

关键词: 变量选择; 弹性 SCAD; Cox 模型; 组效应; Oracle 性质

中图分类号: O212.1

文献标志码: A

文章编号: 1672-5174(2020)11-129-10

DOI: 10.16441/j.cnki.hdxh.20180263

引用格式: 孙志华, 谢向东, 焦东方. Cox 比例风险模型的弹性 SCAD 变量选择方法及应用[J]. 中国海洋大学学报(自然科学版), 2020, 50(11): 129-138.

SUN Zhi-Hua, XIE Xiang-Dong, JIAO Dong-Fang. Elastic SCAD variable selection and its application in Cox model[J]. Periodical of Ocean University of China, 2020, 50(11): 129-138.

变量选择问题是统计领域中的研究方向之一。基于惩罚思想的变量选择方法(也称正则化回归方法), 在选出变量的同时也对变量参数作出了估计, 其计算量相对较小, 相比其他变量选择方法呈现出诸多优越性。这使得, 以 Lasso 为代表的基于惩罚的变量选择方法得到广泛研究, 出现了一系列基于惩罚的变量选择方法: Bridge, Lasso, SCAD, Elastic Net, Adaptive Lasso, Dantzig Selector, MCP 等。虽然 Lasso 具有较好的预测结果, 但一般情况下 Lasso 结果是有偏的, 在严格的条件假设下才具有相合性, 并且 Lasso 不具有 oracle 性质^[1]。而 SCAD, Adaptive Lasso, Dantzig Selector 以及 MCP 具有 oracle 性质。West 等^[2]提出当“large p, small n”时, 要特别重视成组变量(Grouped variables)的问题。针对此类问题, Hastie 等^[3]、Díaz-Urriarte^[4]提出了主成分分析法, Hastie 等^[5]提出了监督 tree harvesting 方法, Dettling 和 Bühlmann^[6]将聚类和有监督学习结合在了一起, Segal 等^[7]论证了正则化回归方法处理成组变量的优势, Zou 和 Hastie^[8]提出了变量选择方法的组效应(Grouping effect)。Lasso 处理成组变量的效果非常差, 而 SCAD, Elastic Net, Adaptive Lasso, Dantzig Selector, MCP 中, 仅 Elastic Net 方法具有组效应。

由于生存数据的删失性, 完全数据的变量选择方

法不能直接应用于生存数据, 因而一些学者研究了 Cox 比例风险模型(简称之为 Cox 模型)的变量选择问题: Tibshirani^[9]、Fan 和 Li^[10]分别将 Lasso、SCAD 应用于 Cox 模型, Li 和 Luan^[11]提出了 Cox 核转换方法, Zhang 和 Lu^[12], Zou^[13]将 Adaptive Lasso 应用到 Cox 模型, 闫丽娜^[14]将 Elastic Net 与 Cox 模型结合, 侯文^[15]研究了 Cox 模型的桥估计(Bridge), 邓秋玲^[16]研究了 SCAD 和 Adaptive Dantzig Selector 在 Cox 模型中的应用, 刘丹等^[17]研究了 Adaptive Elastic Net 在 Cox 模型中的应用。

为了使高维生存数据的 Cox 模型的变量选择方法既有 oracle 性质, 又具有组效应, 本文提出了 Cox 模型的弹性 SCAD 方法, 并证明了弹性 SCAD 方法的统计性质, 通过数值模拟, 比较了在 Cox 模型下, 弹性 SCAD 与 Lasso, Adaptive Lasso, SCAD, Elastic Net, Adaptive Elastic Net 等方法的变量选择结果, 得到了弹性 SCAD 在某些情况下的优越性, 最后再结合实例, 探讨了弹性 SCAD 及其他变量选择方法应用于 Cox 模型处理生存数据的表现优劣。

1 基于 Cox 模型的弹性 SCAD 方法及性质

首先, 我们给出 Cox 模型的弹性 SCAD 方法的估计:

* 基金项目: 山东省自然科学基金项目(ZR2016DQ09)资助

Supported by the Natural Science Foundation of Shandong Province of China(ZR2016DQ09)

收稿日期: 2019-04-06; 修订日期: 2019-06-25

作者简介: 孙志华(1977-), 女, 博士生, 讲师, 主要从事生物统计研究。E-mail: zhihuasun@ouc.edu.cn

$$\hat{\beta} = \operatorname{argmin}_{\beta} \{-\ln(\beta) + \sum_{j=1}^p P_{\lambda_1, \lambda_2}(\beta_j)\}. \quad (1)$$

式中: $\ln(\beta)$ 表示 Cox 模型的对数偏似然函数^[5], $\ln(\beta) = \sum_{i=1}^n \delta_i [x_i \beta - \ln(\sum_{j \in R(t_i)} e^{x_j \beta})]$;

$$P_{\lambda_1, \lambda_2}(\beta) = f_{\lambda_1}(\beta) + \lambda_2 \|\beta\|^2 = \sum_{i=1}^p f_{\lambda_1}(\beta) + \lambda_2 \sum_{i=1}^p \beta_i^2.$$

$R(t_i)$ 表示 t_i 时刻的危险集, δ_i 是表示是否删失的示性函数。 $\gamma > 2, \lambda_1 > 0, \lambda_2 > 0$ 为调整参数

$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, 表示矩阵 X 的第 i 行。

故

$$p'_{\lambda_1, \lambda_2}(\beta_j) = \begin{cases} \lambda_1 \operatorname{sgn}(\beta_j) + 2\lambda_2 \beta_j, & |\beta_j| < \lambda_1 \\ -\frac{\beta_j - \gamma \lambda_1 \operatorname{sgn}(\beta_j)}{\gamma - 1} + 2\lambda_2 \beta_j, & \lambda_1 \leq |\beta_j| < \gamma \lambda_1 \\ 2\lambda_2 \beta_j, & |\beta_j| \geq \gamma \lambda_1 \end{cases} \quad (2)$$

$$p''_{\lambda_1, \lambda_2}(\beta_j) = \begin{cases} 2\lambda_2, & |\beta_j| < \lambda_1 \\ -\frac{1}{\gamma - 1} + 2\lambda_2, & \lambda_1 \leq |\beta_j| < \gamma \lambda_1 \\ 2\lambda_2, & |\beta_j| \geq \gamma \lambda_1 \end{cases} \quad (3)$$

根据弹性 SCAD 罚函数的二阶导函数易知, 当调整参数满足条件: $\lambda_1 > 0, \gamma > 1, \lambda_2 > \frac{1}{2(\gamma - 1)}$ 时, 弹性 SCAD 罚函数为关于 β 的严格凸函数。

记 $Q(\beta) = -\ln(\beta) + P_{\lambda_1, \lambda_2}(\beta)$ 。 (4)

我们可以得出基于 Cox 比例风险模型的弹性 SCAD 方法有以下性质(证明见附录):

性质 1 若 $x_i = x_j, i \neq j, i, j \in \{1, 2, \dots, n\}, \lambda_1 > 0, \gamma > 1, \lambda_2 > \frac{1}{2(\gamma - 1)}$, 则 $\hat{\beta}_i = \hat{\beta}_j$ 。

性质 2(变量选择的组效应) 记 β^* 为真实参数向量, 若 $\beta_i^* \cdot \beta_j^* > 0, \lambda_2 > \frac{1}{2(\gamma - 1)}, i, j \in \{1, 2, \dots, n\}, D(i, j) = C|\beta_i^* - \beta_j^*|$, 则

$$D(i, j) < ma + \sqrt{2(1 - \rho)},$$

式中 $\rho = x_i^T \cdot x_j; C = 2\lambda_2 - \frac{1}{\lambda - 1}$;

$a = \max\{|x_{ij} - x_{it}|, i \in \{1, 2, \dots, n\}, j, t \in \{1, 2, \dots, p\}\}$ 。

性质 3(估计一致性) 若 $n \rightarrow +\infty, \lambda_1(n) \rightarrow 0, \sqrt{n} \lambda_2(n) \rightarrow 0$, 则存在 $Q(\beta)$ 的一个局部最优解 $\hat{\beta}(n)$, 满足 $\|\hat{\beta}(n) - \beta^*\| = o_p(b_n)$, 其中 $a = \max\{p'_{\lambda_1, \lambda_2}(|\beta_j^*| \neq 0)\}, b_n = n^{-\frac{1}{2}} + a$ 。

性质 4(Oracle 性质) 若 $n \rightarrow \infty, \lambda_1(n) \rightarrow 0$,

$\sqrt{n} \lambda_1(n) \rightarrow +\infty, \sqrt{n} \lambda_2(n) \rightarrow 0$, 则 $\hat{\beta}(n) = \begin{pmatrix} \hat{\beta}_N(n) \\ \hat{\beta}_Z(n) \end{pmatrix}$,

满足

(1)稀疏性: $P\{\hat{\beta}_Z(n) = 0\} = 1$;

(2)渐近正态性: $\sqrt{n}(\hat{\beta}(n) - \beta^*) \xrightarrow{D} N_p(0, \Sigma)$ 。

式中 $\hat{\beta}_N(n)$ 表示非零系数矩阵的估计; $\hat{\beta}_Z(n)$ 表示零

系数矩阵的估计; $\beta^* = \begin{pmatrix} \beta_N^* \\ \beta_Z^* \end{pmatrix}$ 表示真实参数矩阵; β_N^*

表示 d 维非零系数矩阵, β_Z^* 表示 $(p-d)$ 维零系数矩阵, Σ 表示偏似然估计方程的方差。

性质 5(方差估计的渐近性)

$$\Sigma_1(\hat{\beta}(n)) \xrightarrow{p} \Sigma_1(\beta^*), n \rightarrow \infty.$$

式中: $\Sigma_1(\hat{\beta}(n))$ 表示弹性 SCAD 惩罚偏似然估计方

程的方差, 即 $\frac{Q(\beta)}{\beta_s} = -\frac{\ln(\beta)}{\beta_s} \Big|_{\beta=\hat{\beta}(n)} + n p'_{\lambda_1, \lambda_2}(\beta_s) = 0$ 的方差。

性质 1~5 的证明见附录。

2 数值模拟

弹性 SCAD 方法可以看作是 SCAD 方法和岭回归的结合, 具有 Oracle 性质, 其克服了 Elastic Net 有偏估计的缺点, 属于渐近无偏估计。与 SCAD 相比, 具有组效应, 且与同样具有组效应的 Elastic Net (ENet)、Adaptive elastic net(AENet)相比, 其在处理小样本高维数据、变量间存在高度共线性问题与小样本低维、变量间存在弱相关关系两种情况下表现更优。

本部分数据模拟选用十折交叉验证, 研究的目的是通过运用基于 Cox 模型的 Lasso、Adaptive lasso (Alasso)、ENet、AENet、SCAD、弹性 SCAD(Escad) 6 种变量选择方法, 对模拟生成的高维、并具有不同共线性强弱的生存数据进行变量选择, 比较其变量筛选效果以及模型误差等指标, 进而评价各种方法的优劣。

数值模拟的参数设置情况如下: 设样本量为 n , 协变量个数为 p , 生成 $n \times p$ 的数据矩阵 X , 除前 2 列外, 其余各列服从标准的多元正态分布, 第一列数据服从一元标准正态分布, 第二列数据与第一列数据之间的相关系数分别设为 $r(v_1, v_2) = 0.8, 0.5$ 和 0.2 , 分别代表变量和间存在强共线性、中等共线性和弱共线性的情况。参数设置为:

$v = (v_1, v_2, v_3, \dots, v_p) = (0.8, 0.7, 1, -0.6, 0, 0, \dots, 0)$ 。

当变量 v_1 与 v_2 间存在共线性且为重要变量时, 若所用变量选择方法同时将 v_1, v_2 选进模型, 则说明该方法具有变量选择的组效应, 否则, 说明其没有变量选择的组效应。生存时间删失率设为 0.3 。

利用 R 软件进行蒙特卡洛模拟抽样 100 次(主要

用到的包为 `ncvreg`、`Coxnet`、`survial`、`Matrix`), 结果见表 1。其中真阳性表示选对非零系数的个数, 假阳性表示错选非零系数的个数; `noise` 表示除变量前 4 个系数外, 其他非零系数绝对值之和; `main` 表示变量中前 4 个变量绝对值之和; $\text{bias} = |\text{main} - m|$, 表示前 4 个估计系数与真实系数之间偏差的绝对值, 用来衡量系数估计的优劣, 其中 $m = \sum_{i=1}^4 |\hat{\beta}_i|$; 模型误差 = `noise` × `bias`, 表示 `noise` 与 `bias` 表示 `noise` 与 `bias` 的乘积, 用来衡量模型误差, 之所以用误差而不是 `cve` (交叉验证模型误差) 或 `cvse` (交叉验证标准误差) 来衡量, 主要是因为 `ncvreg` 包中关于交叉验证的定义并不明确, 无法计算 `cvse`, 只能得出 `cve`, 而 `Coxnet` 包中仅能计算 `cvse`, 且 `cvse` 与 `cve` 之间的换算并不清楚, 这导致二者对模型误差的衡量标准不统一, 故采用新的衡量方法, 易知 `noise` 和 `bias` 越小, 模型误差越小, 故用二者的乘机来衡量模型的误差较为合理。

从图 1 可以看出, SCAD 相比于 Lasso、Alasso, 其模型误差较小, 模型稳定性强; 弹性 SCAD 相比于 Enet、AENet, 由于其具有 Oracle 性质, 故模型误差较小, 模型稳定性强, 虽然 AENet 也具有 Oracle 性质, 但模型表现不如 Escad 稳定。

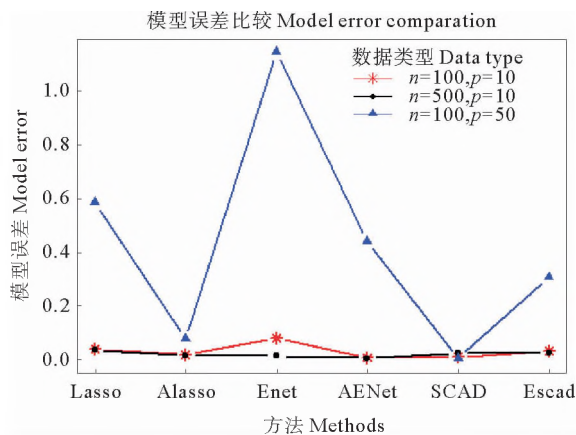
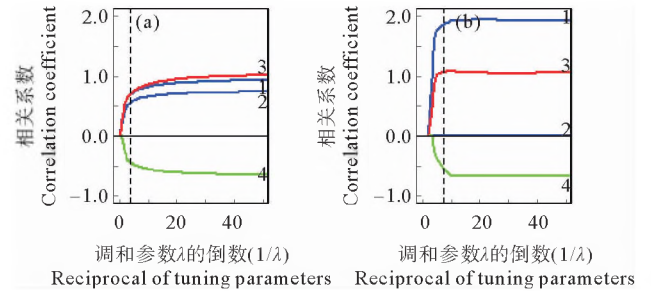


图 1 6 种方法在模型误差方面的表现 (相关系数为 0.8)

Fig.1 The performance of six methods in model error (correlation coefficient 0.8)

图 2 给出了当 $n=100$, $p=10$ 且相关系数为 0.8 时, 弹性 SCAD 与 SCAD 的求解路径的比较, 其中图 2(a) 为弹性 SCAD, 图 2(b) 为 SCAD。

结合图 1 和 2 可以看出, 弹性 SCAD 在保留了 SCAD 变量选择优点的同时, 克服了 SCAD 方法在进行变量选择时不具有组效应的缺点, 在变量间存在高度相关性时, 能够把相关的变量同时选进模型, 具有变量选择的组效应。



($n=100$, $p=10$, $r(v_1, v_2)=0.8$). 曲线 1~4 分别为 $v_1 \sim v_4$ 的估计值。
 $n=100$, $p=10$, $r(v_1, v_2)=0.8$. line 1~4 are the estimators of variables $v_1 \sim v_4$.)

图 2 弹性 SCAD(a) 与 SCAD(b) 的求解路径

Fig.2 Solution path comparison of elastic SCAD(a) and SCAD(b)

从数据模拟结果可以看出: 不具有变量选择组效应的方法是基于 Cox 模型的 Lasso、Alasso、SCAD; 具有变量选择组效应的方法是 ENet、AENet、弹性 SCAD。其中三者对于有组效应的变量选择方法, 在 $n=100$, $p=50$, 变量间存在强相关关系与 $n=100$, $p=10$, 变量间存在弱相关关系两种情况下, 弹性 SCAD 模型误差最小, 系数估计方面也表现最佳; 在 $n=500$, $p=10$ 情况、变量间存在非强 (中等强度及较弱强度) 相关关系时, 弹性 SCAD 在系数估计方面表现最佳; 在 $n=100$, $p=50$ 情况, 变量间存在弱相关关系时, 弹性 SCAD 的模型误差最小。基于 Cox 模型的 Lasso 和 Alasso 均倾向于多选变量, 而 SCAD 与二者相比, 虽然也存在假阳性, 但除小样本低维、变量间存在弱相关性的情况外, 其假阳性的个数均小于二者, 故变量选择的一致性方面, SCAD 优势明显; 基于 Cox 模型的 ENet、AENet 和弹性 SCAD 均倾向于多选变量, 而弹性 SCAD 与二者相比, 虽然也存在假阳性, 但在 $n=100$, $p=50$ 情况下假阳性的个数均小于二者, 故变量选择的一致性方面, 此情况下基于 Cox 模型的弹性 SCAD 方法最优。当 $n=100$, $p=200$ 时, 不具有变量选择组效应的方法是基于 Cox 模型的 Lasso、Alasso、SCAD; 具有变量选择组效应的方法仍是基于 Cox 模型的 ENet、AENet、弹性 SCAD; 在变量间 3 种不同程度的相关关系情况下, 基于 Cox 模型的弹性 SCAD 相比于 ENet 和 AENet 方法, 均具有较少的噪声系数、较低的模型误差。

3 实例分析

本实例数据来源于 Kalbfleish 和 Prentice 的一组肺癌治疗方案的临床试验数据^[18], 这是一个标准的生存分析数据集。我们对这组数据利用前文提到的 6 种方法进行变量选择, 结果见表 2。

表 1 部分数值模拟结果
Table 1 Partial numerical simulation results

(n, p)	相关系数	对比方面	方法 Methods					
	Correlation coefficient	Comparation aspect	Lasso	Alasso	Enet	AENet	SCAD	Escad
(100,10)	0.8	真阳性	3.000	3.000	4.000	4.000	3.000	3.830
	0.5	True	3.000	3.000	4.000	4.000	3.000	4.000
	0.2	positive	3.000	3.000	4.000	4.000	3.000	4.000
	0.8	假阳性	2.020	0.630	3.580	0.960	0.560	2.000
	0.5	False	2.000	1.000	4.020	1.000	0.840	2.020
	0.2	positive	2.010	0.870	3.180	1.010	0.910	2.090
	0.8	模型误差	0.040	0.018	0.081	0.008	0.010	0.033
	0.5	Model	0.016	0.127	0.106	0.005	0.082	0.098
	0.2	error	0.214	0.298	0.082	0.115	0.196	0.066
(500,10)	0.8	真阳性	3.000	3.000	4.000	4.000	3.000	4.000
	0.5	True	3.000	3.000	4.000	4.000	3.000	4.000
	0.2	positive	3.000	3.000	4.000	4.000	3.000	4.000
	0.8	假阳性	5.900	1.920	6.000	2.380	1.750	5.040
	0.5	False	5.690	1.580	6.000	2.030	1.480	6.000
	0.2	positive	5.460	1.730	6.000	2.080	1.370	6.000
	0.8	模型误差	0.035	0.015	0.013	0.005	0.024	0.025
	0.5	Model	0.179	0.085	0.044	0.022	0.088	0.037
	0.2	error	0.648	0.371	0.386	0.199	0.332	0.363
(100,50)	0.8	真阳性	3.000	3.000	4.000	4.000	3.000	4.000
	0.5	True	3.000	3.000	4.000	4.000	3.000	4.000
	0.2	positive	3.000	3.000	4.000	4.000	3.000	4.000
	0.8	假阳性	7.790	6.000	15.000	8.700	0.630	7.280
	0.5	False	2.160	8.000	16.000	15.92	0.420	11.070
	0.2	positive	1.510	6.000	18.000	16.000	1.870	11.110
	0.8	模型误差	0.585	0.078	1.144	0.439	0.005	0.306
	0.5	Model	0.209	0.341	1.286	0.532	0.026	0.772
	0.2	error	0.010	1.438	0.861	0.479	0.728	0.394
(100,200)	0.8	真阳性	3.000	3.000	4.000	4.000	3.000	4.000
	0.5	True	3.000	3.000	4.000	4.000	3.000	4.000
	0.2	positive	3.000	3.000	4.000	4.000	3.000	4.000
	0.8	假阳性	7.160	10.690	20.660	15.830	0.000	6.970
	0.5	False	5.810	10.240	19.320	18.410	0.000	10.320
	0.2	positive	4.770	10.790	28.990	31.200	0.300	20.430
	0.8	模型误差	0.394	0.313	1.133	1.123	0.000	0.377
	0.5	Model	0.242	0.076	0.954	1.340	0.000	0.509
	0.2	error	0.060	1.342	1.153	1.465	0.005	0.752

由于 Diagnosis time, age 和 prior 的值均为 0, 可以判定其对研究对象的生存时间没有影响; 结合 Escad、ENet、AENet 的选择结果, 可知 small 为噪声系数, 对结果没有明显影响; 若假设变量间存在相关性, 结合数值模拟结果可知, 在小样本低维数情况下,

AENet 表现最优, 故理论上其选择的非零个数要比 SCAD、Lasso、Alasso 要多, 而结合表 2 中数据可知, 这 4 种方法选出的非零个数并无明显差异, 甚至不具有组效应的 Lasso 要比 AENet 选的还多, 这说明这些变量间是不具有相关关系。

表 2 对肺癌数据进行变量选择的结果
Table 2 The result of variable selection for lung cancer data

方法 Methods	治疗指标 Treatment	卡式评分 Karnofsky score	诊断时间 Diagnosis time	年龄 Age	前期治疗 Priortherapy	鳞状细胞癌 Squamous	小细胞肺癌 Small	腺癌 Adeno	大细胞癌 Large
Escad	0.099	-0.028	0.000	0.000	0.000	-0.681	0.000	0.296	-0.334
Enet	0.057	-0.025	0.000	0.000	0.000	-0.430	0.082	0.368	-0.148
AENet	0.000	-0.031	0.000	0.000	0.000	-0.428	0.000	0.416	0.000
SCAD	0.000	-0.033	0.000	0.000	0.000	-0.310	0.000	0.182	0.000
Lasso	0.069	-0.028	0.000	0.000	0.000	-0.504	0.024	0.350	-0.191
Alasso	0.000	-0.032	0.000	0.000	0.000	-0.434	0.000	0.425	0.000

卡式评分 (Karnofsky score) 表现得分越高, 健康状况越好, 越能忍受治疗给身体带来的副作用, 因而也就有可能接受彻底的治疗, 6 种方法均将其选出, 比较符合实际; 肺小细胞癌是肺癌中最凶恶的, 坏死典型且呈广泛性, 扩散转移快; 而与之对应的, 鳞状细胞癌 (鳞癌)、腺癌、大细胞癌均为非小细胞型肺癌, 与小细胞癌相比, 它们扩散转移相对较晚。故根据实际, 可把鳞状细胞癌、腺癌、大细胞癌看作一组变量, 表 2 中只有 Escad、ENet、Lasso 3 种方法同时选进了这 3 种变量, 比较符合实际。但 Lasso 不具有组效应, 故其虽然选进了这 3 种变量, 应该是作为噪声变量选入的, 系数估计会很差, 故与实际相差比较大; 而 AENet 作为在 ENet 的基础上的一种改进, 却没有将这 3 种因素同时选入模型, 可见 ENet 选入这 3 种变量是因为此方法会多选变量。综上, 可推知影响研究对象生存时间的主要因素有: Treatment、Karnofsky、squamous、adeno 和 large 影响效用系数分别为: 0.099、-0.028、-0.683、0.296 和 -0.334。

即癌症类型是否为非小细胞癌, 是决定生存时间的因素。更进一步, 其中是否为鳞状细胞癌的指标为主要决定因素。Karnofsky 表现得分虽然对生存时间有影响, 但是相对来说其影响是次要的, 不起决定性作用。再对比文献 [12] 中研究结果, 可知本文上述研究结果与之相符, 但治疗指标 trt 是否为有效影响因素, 对比之下难以确定, 有待进一步研究。由于在实际应用中, 变量间的关系十分复杂, 故不会完全与模拟数据中的表现完全一致, 经本例的分析结果, 可推知基于 Cox 模型的弹性 SCAD 在实际处理生存数据时表现优

于文中其他变量选择方法。

4 结论与展望

本文主要研究了基于 Cox 模型的弹性 SCAD 变量选择方法的理论性质, 通过数值模拟比较得出了如下结论: 在 $n=100, p=50$ 情况下, 基于 Cox 模型的 Elastic net 和 Adaptive elastic net 方法在变量选择的一致性上的表现不如弹性 SCAD 方法; 对于有组效应的变量选择方法, 在 $n=100, p=50$ 情况, 变量间存在强相关关系与小样本低维、变量间存在弱相关关系两种情况下, 基于 Cox 模型的弹性 SCAD 模型误差最小, 系数估计方面也表现最佳; 在 $n=500, p=10$ 情况, 变量间存在非强 (中等强度及较弱强度) 相关关系时, 基于 Cox 模型的弹性 SCAD 在系数估计方面表现最佳; 在 $n=100, p=50$ 情况、变量间存在弱相关关系时, 基于 Cox 模型的弹性 SCAD 的模型误差最小。

当 $n=100, p=200$ 时, 在变量间 3 种不同程度的相关情况下, 基于 Cox 模型的弹性 SCAD 相比于 Elastic net 和 Adaptive elastic net 方法, 均具有较少的噪声系数、较低的模型误差。进一步通过实例分析发现, 基于 Cox 模型的弹性 SCAD 的变量选择结果优于文中讨论的其余变量选择方法, 变量选择结果较为合理, 与实际更相符。

通过多种方法的实例比较分析, 可以判断变量间是否存在共线性, 有利于在决策时选择更适合自己的方法, 作出更理性、正确的判断, 进一步可知, 通过比较这几种方法在不同类型参数下的表现, 可判断共线性的强弱, 在实际应用时, 提供一个较为理性的方案。

本文只选取了特定删失比例、变量个数、数据个数进行了数值模拟,弹性 SCAD 与 SCAD 的模拟程序也不适于处理带有节点的生存数据,且由于设备的局限性,也未能研究弹性 SCAD 在更高维数据下的表现,结果难免会有一定的片面性。在进一步的研究中,可改进算法、程序、改变删失比例、变量个数及数据个数进行更全面、高效的数值模拟,发现其新的变量选择特性或局限性所在,从而更好地应用于理论研究和实践方面。

参考文献:

- [1] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. *Journal of the American Statistical Association*, 2001, 96(456): 1348-1360.
- [2] West M, Blanchet C, Dressman H, et al. Predicting the clinical status of human breast cancer using gene expression profiles[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2001, 98: 11462-11467.
- [3] Hastie T, Tibshirani R, Eisen M B, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns[J]. *Genome Biology*, 2000, 1(2): 1-21.
- [4] Ramón Díaz-Urriarte, Casado D. Can we Find Molecular Signatures from Gene Expression Data? [C]. Spanish: Spanish Biomformatics Conference, 2004.
- [5] Hastie T, Tibshirani R, Botstein D, et al. Supervised harvesting of expression trees[J]. *Genome Biology*, 2001, 2(1): 1-12.
- [6] Dettling M, Peter Bühlmann. Finding predictive gene groups from microarray data [J]. *Journal of Multivariate Analysis*, 2004, 90(1): 106-131.
- [7] Segal M R, Dahlquist K D, Conklin B R. Regression approaches for microarray data analysis[J]. *Journal of Computational Biology*, 2003, 10(6): 961-980.
- [8] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. *Journal of the Royal Statistical Society*, 2005, 67(2): 301-320.
- [9] Tibshirani R. The lasso method for variable selection in the cox model[J]. *Statistics in Medicine*, 1997, 16(4): 385-395.
- [10] Fan J, Li R. Variable selection for cox's proportional hazards model and frailty model[J]. *The Annals of Statistics*, 2002, 30(1): 74-99.
- [11] Li H, Luan Y. Kernel Cox regression models for linking gene expression profiles to censored survival data[J]. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, 2003, 8(12): 65.
- [12] Zhang H H, Lu W. Adaptive lasso for cox's proportional hazards model[J]. *Biometrika*, 2007, 94(3): 691-703.
- [13] Zou H. A note on path-based variable selection in the penalized proportional hazards model[J]. *Biometrika*, 2008, 95(1): 241-247.
- [14] 闫丽娜. 惩罚 COX 模型和弹性网技术在高维数据生存分析中的应用[D]. 太原: 山西医科大学, 2011.
Yan L N. The Application of Punishment COX Model and Elastic net Technology in Survival Analysis of High-dimensional Data [D]. Taiyuan: Shanxi Medical University, 2011.
- [15] 侯文, 侯向艳, 刘琦. Cox 比例风险模型的桥估计[J]. *辽宁师范大学学报(自然科学版)*, 2012, 35(1): 9-12.
Hou W, Hou X, Liu Q. Bridge estimator for Cox's proportional hazard model[J]. *Journal of Liaoning Normal University (Natural Science Edition)*, 2012, 35(1): 9-12.
- [16] 邓秋玲. SCAD 和 ADS 方法在比例风险模型中的应用[D]. 南宁: 广西大学, 2015.
Deng Q. SCAD and ADS Methods for Proportional Hazards Model [D]. Nanning: Guangxi University, 2015.
- [17] 刘丹, 郑少智. Cox 模型中的自适应 Lasso 变量选择[J]. *统计与决策*, 2016(10): 7-10.
Liu D, Zheng S Z. Adaptive Lasso variable selection method with Cox model[J]. *Statistics and Decision*, 2016(10): 7-10.
- [18] Kalbfleisch D, Prentice R L. *The Statistical Analysis of Failure Time Data*[M]. New York: Wiley, 1980.

Elastic SCAD Variable Selection and Its Application in Cox Model

SUN Zhi-Hua¹, XIE Xiang-Dong², JIAO Dong-Fang¹

(1. School of Mathematical Science, Ocean University of China, Qingdao 266100, China; 2. School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China)

Abstract: In this paper, we studied the application of the elastic SCAD variable selection method based on the Cox proportional hazards model. Our elastic SCAD method has the grouping effect, asymptotic consistency and the Oracle property. We demonstrate the performance of our method via numerical simulations. We show that the elastic SCAD method is better than other variable selections, like Lasso, Adaptive Lasso, SCAD, Elastic Net and Adaptive Elastic Net. Moreover, we apply our method to the lung cancer data. It turns out that the elastic SCAD method performs better than other methods. Elastic SCAD is more reasonable and closer to the practical situation.

Key words: variable selection; elastic SCAD; Cox model; grouping effect; Oracle property

AMS Subject Classification: 62N02

责任编辑 朱宝象

附录

性质 1 证明:用反证法。

对于固定的 $\lambda_1, \lambda_2 > 0$, 若令

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k, k \neq i \text{ 且 } k \neq j \\ \frac{\hat{\beta}_i + \hat{\beta}_j}{2}, k = i \text{ 或 } k = j \end{cases}$$

由 $x_i = x_j$, 知 $\ln(\hat{\beta}^*) = \ln(\hat{\beta})$, 而 $P_{\lambda_1, \lambda_2}(\beta)$ 为一严格凸函数, 故有 $P_{\lambda_1, \lambda_2}(\hat{\beta}^*) < P_{\lambda_1, \lambda_2}(\hat{\beta})$, 这与公式(1)矛盾, 故有 $\hat{\beta}_i = \hat{\beta}_j$ 。

引理 1^[10]: 若 $\beta_i \cdot \beta_j > 0$, $\lambda_2 > \frac{1}{2(\gamma-1)}$, 则 $|p'_{\lambda_1, \lambda_2}(\beta_i) - p'_{\lambda_1, \lambda_2}(\beta_j)| \geq C|\beta_i - \beta_j|, i, j \in \{1, 2, \dots, n\}$ 。

证明 不妨设 $0 < \beta_i \leq \beta_j$, 则可分以下 6 种情况讨论:

$$1. 0 \leq \beta_i \leq \beta_j \leq \lambda_1$$

$$2. \lambda_1 < \beta_i \leq \beta_j \leq \gamma \lambda_1$$

$$3. \gamma \lambda_1 < \beta_i \leq \beta_j$$

$$4. 0 < \beta_i \leq \lambda_1 \leq \beta_j$$

$$5. \lambda_1 < \beta_i \leq \gamma \lambda_1 \leq \beta_j$$

$$6. 0 < \beta_i \leq \lambda_1, \beta_j \geq \gamma \lambda_1$$

当为情况 1 时, 由公式(2)知

$$\begin{aligned} & |p'_{\lambda_1, \lambda_2}(\beta_i) - p'_{\lambda_1, \lambda_2}(\beta_j)| = \\ & |\lambda_1 \operatorname{sgn}(\beta_i) + 2\lambda_2 \beta_i - \lambda_1 \operatorname{sgn}(\beta_j) - 2\lambda_2 \beta_j| = \\ & |2\lambda_2 \beta_i - 2\lambda_2 \beta_j| = 2\lambda_2 |\beta_i - \beta_j| \geq \\ & C|\beta_i - \beta_j|. \end{aligned}$$

同理, 可证情况 2, 3 时亦有

$$|p'_{\lambda_1, \lambda_2}(\beta_i) - p'_{\lambda_1, \lambda_2}(\beta_j)| \geq C|\beta_i - \beta_j| \text{ 成立。}$$

现考虑情况 4, 由公式(2), $\lambda_2 > \frac{1}{2(\gamma-1)}$ 时,

$p'_{\lambda_1, \lambda_2}(\beta_j)$ 为关于的单调递增函数, 从而

$$\begin{aligned} & |p'_{\lambda_1, \lambda_2}(\beta_i) - p'_{\lambda_1, \lambda_2}(\beta_j)| = p'_{\lambda_1, \lambda_2}(\beta_j) - p'_{\lambda_1, \lambda_2}(\beta_i) = \\ & [p'_{\lambda_1, \lambda_2}(\beta_j) - p'_{\lambda_1, \lambda_2}(\lambda_1)] + \\ & [p'_{\lambda_1, \lambda_2}(\lambda_1) - p'_{\lambda_1, \lambda_2}(\beta_i)] \geq \\ & C(\beta_j - \lambda_1) + C(\lambda_1 - \beta_i) = \\ & C(\beta_i - \beta_j) = C|\beta_i - \beta_j|. \end{aligned}$$

同理可证情况 5 和 6 亦有

$$|p'_{\lambda_1, \lambda_2}(\beta_i) - p'_{\lambda_1, \lambda_2}(\beta_j)| \geq C|\beta_i - \beta_j| \text{ 成立。}$$

同样地, 可证 $\beta_i \leq \beta_j < 0$ 时, 仍有

$$|p'_{\lambda_1, \lambda_2}(\beta_i) - p'_{\lambda_1, \lambda_2}(\beta_j)| \geq C|\beta_i - \beta_j| \text{ 成立。}$$

性质 2 证明: 记式子(1)的第 τ 和第 I 个估计值分别为 β_τ, β_I , 则

$$\frac{\partial \ln(\beta)}{\partial \beta_\tau} = \mathbf{x}_\tau \cdot \mathbf{1}_m + \sum_{i=1}^n \delta_i \frac{\sum_{j \in R(t_i)} \mathbf{x}_{j\tau} e^{\beta x_j^\tau}}{\sum_{j \in R(t_i)} e^{\beta x_j^\tau}} =$$

$$\mathbf{x}_\tau \cdot \mathbf{1}_m + M_\tau$$

$$\frac{\partial \ln(\beta)}{\partial \beta_I} = \mathbf{x}_I \cdot \mathbf{1}_m + \sum_{i=1}^n \delta_i \frac{\sum_{j \in R(t_i)} \mathbf{x}_{jI} e^{\beta x_j^\tau}}{\sum_{j \in R(t_i)} e^{\beta x_j^\tau}} =$$

$$\mathbf{x}_I \cdot \mathbf{1}_m + M_I。$$

$\mathbf{x}_\tau$ 表示数据矩 $\mathbf{X}$ 的第 τ 列的非删失部分依次组成的新向量, $\mathbf{x}_I$ 表示数据矩阵 $\mathbf{X}$ 的第 I 列的非删失部分依次组成的新向量。

故 $\Delta M = M_I - M_\tau =$

$$\sum_{i=1}^n \delta_i \frac{\sum_{j \in R(t_i)} (x_{jI} - x_{j\tau}) e^{\beta x_j^\tau}}{\sum_{j \in R(t_i)} e^{\beta x_j^\tau}}$$

易知 $0 \leq |\Delta M| \leq ma$, 从而根据

$$\frac{\partial Q(\beta)}{\partial \beta_\tau} = -\frac{\partial \ln(\beta)}{\partial \beta_\tau} + p'_{\lambda_1, \lambda_2}(\beta_\tau) =$$

$$\mathbf{x}_\tau \cdot \mathbf{1}_m + M_\tau + p'_{\lambda_1, \lambda_2}(\beta_\tau) = 0;$$

$$\frac{\partial Q(\beta)}{\partial \beta_I} = -\frac{\partial \ln(\beta)}{\partial \beta_I} + p'_{\lambda_1, \lambda_2}(\beta_I) =$$

$$\mathbf{x}_I \cdot \mathbf{1}_m + M_I + p'_{\lambda_1, \lambda_2}(\beta_I) = 0。$$

可知

$$p'_{\lambda_1, \lambda_2}(\beta_I) - p'_{\lambda_1, \lambda_2}(\beta_\tau) = \Delta M + \mathbf{x}_I \cdot \mathbf{1}_m - \mathbf{x}_\tau \cdot \mathbf{1}_m$$

再结合引理 1 得

$$\begin{aligned} & C|\beta_I - \beta_\tau| \leq |p'_{\lambda_1, \lambda_2}(\beta_I) - p'_{\lambda_1, \lambda_2}(\beta_\tau)| \leq \\ & \Delta M + |\mathbf{x}_I - \mathbf{x}_\tau| \leq ma + |\mathbf{x}_I - \mathbf{x}_\tau|。 \end{aligned}$$

又 $\rho = \mathbf{x}_\tau^\tau \cdot \mathbf{x}_I$, 此时 $\mathbf{x}_\tau$ 表示 $\mathbf{X}$ 数据矩阵的第 τ 列, $\mathbf{x}_I$ 表示数据矩阵 $\mathbf{X}$ 的第 I 列, 且已经过标准化处理, 故 $|\mathbf{x}_I - \mathbf{x}_\tau|^2 < 2(1 - \rho)$, $|\mathbf{x}_I - \mathbf{x}_\tau| < \sqrt{2(1 - \rho)}$,

即有

$$\begin{aligned} & C|\beta_I - \beta_\tau| \leq |p'_{\lambda_1, \lambda_2}(\beta_I) - p'_{\lambda_1, \lambda_2}(\beta_\tau)| \leq \\ & ma + |\mathbf{x}_I - \mathbf{x}_\tau| < \\ & ma + \sqrt{2(1 - \rho)}。 \end{aligned}$$

从而

$$\begin{aligned} & C|\beta_I - \beta_\tau| \leq |p'_{\lambda_1, \lambda_2}(\beta_I) - p'_{\lambda_1, \lambda_2}(\beta_\tau)| < \\ & ma + \sqrt{2(1 - \rho)}。 \end{aligned}$$

性质 3 证明: 此证明中重新定义公式(4)为

$$Q(\beta) = -\ln(\beta) + n P_{\lambda_1, \lambda_2}(\beta),$$

令 $B = \{\beta: \|\beta - \beta^*\| \leq A n^{-\frac{1}{2}} + a\}$, A 为一个常数, 则 B 为一个紧致集。

由 $Q(\beta)$ 为 B 上的一个连续函数知, $Q(\beta)$ 在 B 上存在极小值, 若在 B 的边界上的每一个 β 都有 $Q(\beta^*) < Q(\beta)$ 成立, 则在 B 的内部存在 β^* , 使得 $Q(\beta^*)$ 达到极小值, B 内部这样的 β^* 为 $Q(\beta^*)$ 的局部极小点。

要证明性质 3, 只需证明其一个充分条件:

对给定任意小的 $\epsilon > 0$ 、足够大的 γ_i , 存在一个常数 A (满足 B 包含真实参数), 使得

$$P\{\inf_{\|u\|=A} Q(\beta^* + u b_n) > Q(\beta^*)\} > 1 - \varepsilon,$$

其中 u 是一个 p 维矢量,

$$\text{sgn}(u_j) = \text{sgn}(\beta_j^*), j=1, 2, \dots, p,$$

用 $\hat{\beta}(n)$ 表示 B 内部 $Q(\beta)$ 的局部极小点, 则有

$$\|\beta - \beta^*\| \leq A n^{-\frac{1}{2}} + a.$$

令

$$D(u) = Q(\beta^* + u b_n) - Q(\beta^*) =$$

$$\ln(\beta^*) - \ln(\beta^* + u b_n) + n P_{\lambda_1, \lambda_2}(\beta^* + u b_n) -$$

$$n P_{\lambda_1, \lambda_1}(\beta^*) = [\ln(\beta^*) - \ln(\beta^* + u b_n)] -$$

$$n[P_{\lambda_1, \lambda_1}(\beta^*) - P_{\lambda_1, \lambda_1}(\beta^* + u b_n)].$$

用 $\ln'(\beta^*)$ 代表 $\ln(\beta^*)$ 的梯度向量, 并设 $l'(\beta^*) < \infty$,

$\ln(\beta^*)$ 在 β^* 处的 Fisher 信息阵为:

$$I(\beta^*) = -\frac{1}{n} \ln'(\beta^*) + o_p(1), \text{ 则将 } \ln(\beta^* +$$

$u b_n)$ 在 β^* 处泰勒展开得

$$\ln(\beta^*) - \ln(\beta^* + u b_n) = \ln(\beta^*) -$$

$$\left[\ln(\beta^*) + \ln'(\beta^*)^T u b_n + \frac{1}{2} u^T \ln''(\beta^*) u b_n^2 + O_p(u^T u b_n^2) \right] =$$

$$-\ln'(\beta^*)^T u b_n + \frac{1}{2} u^T I(\beta^*) u b_n^2 - \frac{n+1}{2} O_p(u^T u b_n^2),$$

同理, 可得

$$n\{f_{\lambda_1}(\beta^*) + f_{\lambda_1}(\beta^* + u b_n)\} =$$

$$-n b_n^2 \sum_{j=1}^p f'_{\lambda_1}(\beta^*) u_j - \frac{1}{2} n b_n^2 \sum_{j=1}^p f'_{\lambda_1}(\beta^*) u_j^2 -$$

$$O_p(u^T u b_n^2) n\{\lambda_2 \|\beta^*\|^2 - \lambda_2 \|\beta^* + u b_n\|^2\} =$$

$$-2n b_n \lambda_2(n) \sum_{j=1}^p \beta_j^* u_j - \lambda_2(n) n b_n^2 \sum_{j=1}^p \beta_j^* u_j^2 -$$

$$O_p(u^T u b_n^2).$$

故

$$n[P_{\lambda_1, \lambda_2}(\beta^*) + P_{\lambda_1, \lambda_2}(\beta^* + u b_n)] =$$

$$n[f_{\lambda_1}(\beta^*) - f_{\lambda_1}(\beta^* + u b_n)] +$$

$$n[\lambda_2 \|\beta^*\|^2 - \lambda_2 \|\beta^* + u b_n\|^2] =$$

$$-n b_n^2 \sum_{j=1}^p f'_{\lambda_1}(\beta_j^*) u_j - \frac{1}{2} n b_n^2 \sum_{j=1}^p f'_{\lambda_1}(\beta_j^*) u_j^2 -$$

$$2n b_n \lambda_2(n) \sum_{j=1}^p \beta_j^* u_j - \lambda_2(n) n b_n^2 \sum_{j=1}^p \beta_j^* u_j^2 - 2O_p(u^T u b_n^2),$$

从而可得

$$D_n(u) = -\ln'(\beta^*)^T u b_n + \frac{1}{2} n u^T I(\beta^*) u b_n^2 -$$

$$\frac{n+1}{2} O_p(u^T u b_n^2) + n b_n^2 \sum_{j=1}^p f'_{\lambda_1}(\beta^*) u_j +$$

$$\frac{1}{2} n b_n^2 \sum_{j=1}^p f'_{\lambda_1}(\beta^*) u_j^2 + 2n b_n \lambda_2(n) \sum_{j=1}^p \beta_j^* u_j +$$

$$\lambda_2(n) n b_n^2 \sum_{j=1}^p \beta_j^* u_j^2 + 2O_p(u^T u b_n^2).$$

又有 $n \rightarrow +\infty, \lambda_1(n) \rightarrow 0, \sqrt{n} \lambda_2(n) \rightarrow 0$,

故结合正文中公式(2)、(3)知

$$n b_n^2 \sum_{j=1}^p f'_{\lambda_1}(\beta^*) u_j + \frac{1}{2} n b_n^2 \sum_{j=1}^p f'_{\lambda_1}(\beta^*) u_j^2 =$$

$$n b_n^2 \sum_{j=1}^p 0 \cdot u_j + \frac{1}{2} n b_n^2 \sum_{j=1}^p 2 \lambda_2(n) u_j^2 =$$

$$n b_n^2 \sum_{j=1}^p \lambda_2(n) u_j^2,$$

化简得

$$D_n(u) = -\ln'(\beta^*)^T u b_n + \frac{1}{2} n u^T I(\beta^*) u b_n^2 -$$

$$\frac{n+1}{2} O_p(u^T u b_n^2) + n b_n^2 \sum_{j=1}^p \lambda_2(n) u_j^2 + 2n b_n \lambda_2(n) \sum_{j=1}^p \beta_j^* u_j +$$

$$\lambda_2(n) n b_n^2 \sum_{j=1}^p \beta_j^* u_j^2 + 2O_p(u^T u b_n^2).$$

$$\text{由 } b_n = n^{-\frac{1}{2}} + a = n^{-\frac{1}{2}} + 2b \lambda_2(n),$$

$$b = \max\{|\beta_j^*| : \beta_j^* \neq 0\},$$

$$\text{知 } \lambda_2(n) = o(n^{-\frac{1}{2}}), b_n = n^{-\frac{1}{2}} [1 + 2b \cdot o(1)],$$

$$n b_n^2 = 1 + 4 \cdot o(1) + 4 b^2 \cdot o(1).$$

进一步化简可得

$$D_n(u) = -\ln'(\beta^*)^T u b_n + \frac{1}{2} n u^T I(\beta^*) u b_n^2 -$$

$$\frac{n+1}{2} O_p(u^T u b_n^2) + o(1) = -\ln'(\beta^*)^T u b_n +$$

$$\frac{1}{2} n u^T u b_n^2 (I(\beta^*) - O_p(1)) + o(1) - \frac{1}{2} O_p(1).$$

由 Fisher 信息矩阵 $I(\beta^*) \geq 0$ 可知 $D_n(u) \geq 0$.

即对任意给定的 $\varepsilon > 0$ 、足够大的 n , 存在一个常数 A (满足 B 包含真实参数), 使得

$$P\{\inf_{\|u\|=A} Q(\beta^* + u b_n) > Q(\beta^*)\} > 1 - \varepsilon,$$

从而性质 3 得证。

性质 4 证明 (1) 证明稀疏性, 只需证明其等价条件即可。

即证明: 当 $n \rightarrow \infty$ 时, 存在 $\varepsilon \geq C n^{-\frac{1}{2}} > 0$, 使得

$$P\left\{\text{sgn}\left(\frac{\partial Q(\beta)}{\partial \beta_j}\right) = \text{sgn}(\beta_j) : |\beta_j| < \varepsilon, j = d+1,$$

$d+2, \dots, p\right\} = 1$ 成立。其中 $\|\beta - \beta^*\| \leq C n^{-\frac{1}{2}}$ 满足,

C 为任意的正常数。

由正文中公式(1)知, 当 $n \rightarrow \infty$ 时, 可取 $\varepsilon <$

$\lambda_1(n)$, 又 $\sqrt{n} \lambda_2(n) \rightarrow 0$, 故

$$n p'_{\lambda_1, \lambda_2}(\beta_s) = n \lambda_1(n) \text{sgn}(\beta_s) + 2n \lambda_2(n) \beta_s =$$

$$\sqrt{n} \{\sqrt{n} \lambda_1(n) \text{sgn}(\beta_s) + 2o_p(1) \beta_s\},$$

又

$$\frac{\ln(\beta)}{\beta_s} = \frac{\ln(\beta)}{\beta_s} \Big|_{\beta=\beta^*} + \sum_{t=1}^p \frac{2 \ln(\beta)}{\beta_s \beta_t} \Big|_{\beta=\beta^*} \cdot$$

$$(\hat{\beta}_t - \beta_t^*) + O_p\left(\sum_{t=1}^p (\hat{\beta}_t - \beta_t^*)\right),$$

结合 $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq Cn^{-\frac{1}{2}}$, 知 $\left. \frac{\ln(\boldsymbol{\beta})}{\beta_s} = \frac{\ln(\boldsymbol{\beta})}{\beta_s} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$

由

$$\frac{\ln(\boldsymbol{\beta})}{\beta_s} = \mathbf{x}_\tau \cdot \mathbf{1}_m + \sum_{i=1}^n \delta_i \frac{\sum_{j \in R(i)} \mathbf{x}_{j\tau} e^{\beta_j \mathbf{x}_j^\top}}{\sum_{j \in R(i)} e^{\beta_j \mathbf{x}_j^\top}} < \infty$$

知

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_s} = & -\left. \frac{\partial \ln(\boldsymbol{\beta})}{\partial \beta_s} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} + \sqrt{n} \{ \sqrt{n} \lambda_1(n) \operatorname{sgn}(\beta_s) + 2o_p(1) \beta_s \} = \\ & \sqrt{n} \{ \sqrt{n} \lambda_1(n) \operatorname{sgn}(\beta_s) + 2o_p(1) \beta_s \} - \text{constant} . \end{aligned}$$

由 $\sqrt{n} \lambda_1(n) \rightarrow \infty$ 知

$$P\left\{ \operatorname{sgn}\left(\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right) = \operatorname{sgn}(\beta_j) : |\beta_j| < \varepsilon, j = d+1, \right.$$

$d+2, \dots, p \} = 1$ 成立, 从而稀疏性得证。

(2) 由性质 3 知若 $n \rightarrow +\infty, \lambda_1(n) \rightarrow 0, \sqrt{n} \lambda_2(n) \rightarrow 0$, 则局部最优解 $\hat{\boldsymbol{\beta}}(n)$ 满足 $\|\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}^*\| = o_p(b_n)$, 从而有

$$\left. \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_s} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(n)} = 0, s = 1, 2, \dots, d.$$

记 $\frac{\partial \ln(\boldsymbol{\beta})}{\partial \beta_s} = 0, s = 1, \dots, p$ 的解为 $\hat{\boldsymbol{\beta}}_{EPL}$ 。

由参考文献[11]中的定理 9, 可知

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EPL} - \boldsymbol{\beta}^*) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}),$$

其中 $\boldsymbol{\Sigma}$ 表示 $\hat{\boldsymbol{\beta}}_{EPL}$ 估计方程的方差。

而 $n \rightarrow +\infty, \gamma \lambda_1(n) \rightarrow 0, |\beta_s| > \gamma \lambda_1(n)$,

故 $np'_{\lambda_1, \lambda_2}(\beta_s) = 2n \lambda_2(n) \beta_s$,

$$\text{从而 } \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_s} = -\left. \frac{\partial \ln(\boldsymbol{\beta})}{\partial \beta_s} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(n)} + np'_{\lambda_1, \lambda_2}(\beta_s) = 0,$$

即为

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_s} = -\left. \frac{\partial \ln(\boldsymbol{\beta})}{\partial \beta_s} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(n)} + 2o_p(1) \beta_s = 0.$$

又 $2o_p(1) \|\hat{\boldsymbol{\beta}}(n)\|^2 \rightarrow o_p(1)$,

故有 $P\{\sqrt{n}\hat{\boldsymbol{\beta}}(n) - \hat{\boldsymbol{\beta}}_{EPL} = \mathbf{0}\} = 1$ 。

从而有 $\sqrt{n}(\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}^*) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, 渐近正态性得证。

性质 5 证明 由文献[11]中的定理 10, 知

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_{EPL}) \xrightarrow{p} \boldsymbol{\Sigma}(\boldsymbol{\beta}^*), n \rightarrow \infty.$$

由性质 4 的(2)中证明, 知

$$P\{\sqrt{n} \|\hat{\boldsymbol{\beta}}(n) - \hat{\boldsymbol{\beta}}_{EPL}\| = 0\} = 1,$$

从而 $\boldsymbol{\Sigma}_1(\hat{\boldsymbol{\beta}}_{EPL}) \xrightarrow{p} \boldsymbol{\Sigma}_1(\boldsymbol{\beta}^*), n \rightarrow \infty$ 。