应用概率统计　第二十八卷
第六期　2012年12月

Chinese Journal of Applied Probability
and Statistics Vol.28 No.6 Dec. 2012

部分线性模型的Adaptive LASSO变量选择 *

李　锋
(郑州航空工业管理学院经贸学院, 郑州, 450015)

卢一强
(解放军信息工程大学电子技术学院, 郑州, 450004)

李高荣
(北京工业大学应用数理学院, 北京, 100124)

摘　要

部分线性模型是一类常用的半参数统计模型, 本文对部分线性模型的adaptive LASSO参数估计及变量选择方法进行了研究. 首先结合截面最小二乘思想和adaptive LASSO估计方法, 构造了adaptive LASSO惩罚截面最小二乘估计, 并研究了惩罚参数和窗宽的选择问题. 理论上研究了在一定条件下估计量的相合性和渐近正态性, 证明adaptive LASSO估计具有oracle性质. 该估计方法便于计算. 最后通过模拟研究了估计量的小样本性质, 结果表明变量选择和参数估计效果良好.

关键词: 部分线性模型, 变量选择, 渐近分布, LASSO, adaptive LASSO.

学科分类号: O212.1, O212.7.

§1. 引　　言

部分线性模型是由Engle等(1986)提出的, 它是一类常用的半参数模型. 假定响应变量与部分协变量呈线性关系, 与另一部分协变量呈非参数函数关系, 这样部分线性模型既具有线性模型良好的可解释性, 又具有非参回归模型的灵活性. 对部分线性模型的研究已取得了大量的成果. 例Speckman (1988)提出了截面最小二乘思想, 并在一些正则条件下得到了参数估计的$\sqrt{n}$相合性和渐近正态性. 关于部分线性模型的系统论述可参看专著[3].

随着信息技术及计算机的发展, 人们开始关注高维数据的研究, 其根本的问题是变量选择. 近年来高维线性模型研究取得了大量的研究成果, 涌现了众多的变量选择方法, 如LASSO[4–6], SCAD[7, 8], Dantzig selector[9], MCP[10]等. 然而, 对部分线性模型的变量选择研究的文献相对较少, 正如Fan等(2004)指出: 部分线性模型包含非参数成分, 有更多的参数需要选取, 如窗宽和惩罚参数. Fan等(2004)研究了纵向数据部分线性模型的SCAD惩罚截面最小二乘变量选择方法, Xie等(2009)研究了高维数据部分线性模型的SCAD惩罚变量选择, 其采用多项式回归样条进行模型估计; Liang等(2009)研究了部分线性测量误差模型的SCAD惩罚最小二乘变量选择; Ni等(2009)提出了双惩罚的部分线性变量选择方法.

*国家自然科学基金(11101014)、高等学校博士学科点专项科研基金联合资助课题(20101103120016)、北京市属高等学校人才强教深化计划"中青年骨干人才培养计划"项目(PHR20110822)、北京市优秀人才培养资助项目(2010D005015000002)和北京工业大学基础研究基金项目(X4006013201101)资助.

本文2011年8月26日收到, 2012年7月25日收到修改稿.

Zou (2006)指出LASSO变量选择方法不能同时满足模型选择的相合性和参数估计达到$\sqrt{n}$的收敛速度, 为克服此不足之处, 提出了线性模型adaptive LASSO变量选择方法, 证明adaptive LASSO具有Fan等(2001)提出的oracle性质. 而部分线性模型的adaptive LASSO变量选择问题鲜有研究.

本文将对部分线性模型的adaptive LASSO变量选择方法进行研究. 文章第二部分结合Speckman (1988)提出的截面最小二乘估计思想, 构造部分线性模型adaptive LASSO惩罚截面最小二乘估计, 讨论了惩罚参数和窗宽的选择问题. 第三部分研究估计量的渐近性质, 得到了部分线性模型adaptive LASSO参数估计具有oracle性质. 第四部分给出了定理的证明. 最后一部分通过蒙特卡洛模拟研究了所提方法的小样本性质.

§2. 模型与方法

假设$\{(Y_i, X_i, T_i), i = 1, \ldots, n\}$为来自如下部分线性模型的一个样本,

$$Y = X'\beta + f(T) + \epsilon, \tag{2.1}$$

其中X为p维预测变量, β为未知稀疏参数向量即仅部分系数非零(通常建模之初, 为了不遗漏重要变量, 模型会引入众多的预测变量, 然而在这些预测变量中往往只有少数对响应变量有重要影响), $f(\cdot)$为一元协变量T的光滑未知函数. 此方法也适用于多元协变量T, 但由于多元维数灾难问题, 会变得不太实用. 假设T随机, 密度函数为$p(t)$, 且T在某紧集取值, 方便起见, 假定此集合为区间$[0, 1]$. Y为响应变量, ϵ为随机误差项与(X, T)独立且期望为0, 标准差为σ. 由于模型中存在不重要的预测变量, 会直接影响到模型的预测精度和模型的可解释性. 因此本文的目的是采用变量选择方法, 选出重要变量集合, 并给出未知参数和非参函数估计.

记$\mathcal{A} = \{j : \beta_j \neq 0, j = 1, 2, \ldots, p\}$, 不失一般性, 假定$\mathcal{A} = \{1, \ldots, p_0\}$, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$, 其中$X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, \ldots, n$, $\boldsymbol{T} = (T_1, T_2, \ldots, T_n)'$, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$, $\boldsymbol{f} = (f(T_1), f(T_2), \ldots, f(T_n))'$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$. 类似Rice (1986), 我们假定$X_{ij}$和$T_i$满足回归关系$\mathsf{E}(X_{ij}|T_i) = g_j(T_i)$, 其中$g_j(T)$ $(1 \leq j \leq p)$为未知光滑函数且二阶导函数连续, $X_{ij} = g_j(T_i) + \eta_{ij}$ $(1 \leq i \leq n, 1 \leq j \leq p)$, $\mathsf{E}(\eta_{ij}|T_i) = 0$, 且$\eta_{ij}$与$\epsilon_i$独立. 于是$\boldsymbol{X}$可分解为$\boldsymbol{X} = \boldsymbol{g} + \boldsymbol{\eta}$, 其中$\boldsymbol{g} = (g_{ij})_{n\times p}$, $g_{ij} = g_j(T_i)$, $\boldsymbol{\eta} = (\eta_{ij})_{n\times p}$.

由模型(2.1)可知$f(T) = \mathsf{E}(Y|T) - \mathsf{E}(X|T)'\beta$, 代入(2.1)整理可得

$$Y - \mathsf{E}(Y|T) - (X - \mathsf{E}(X|T))'\beta - \epsilon = 0. \tag{2.2}$$

记$m_X(T) = \mathsf{E}(X|T)$, $m_Y(T) = \mathsf{E}(Y|T)$. 令$\widehat{m}_X(T)$和$\widehat{m}_Y(T)$分别为$m_X(T)$和$m_Y(T)$的估

计, 本文采用核估计, 如

$$\widehat{m}_X(T)=\frac{\sum\limits_{i=1}^{n}K\Big(\dfrac{T_i-T}{h}\Big)X_i}{\sum\limits_{i=1}^{n}K\Big(\dfrac{T_i-T}{h}\Big)},\qquad \widehat{m}_Y(T)=\frac{\sum\limits_{i=1}^{n}K\Big(\dfrac{T_i-T}{h}\Big)Y_i}{\sum\limits_{i=1}^{n}K\Big(\dfrac{T_i-T}{h}\Big)},$$

其中$K(\cdot)$为核函数, h为窗宽. 记$\widetilde{Y}_i=Y_i-\widehat{m}_Y(T_i)$, $\widetilde{X}_i=X_i-\widehat{m}_X(T_i)$, $\widetilde{\boldsymbol{X}}=(\widetilde{X}_1,\widetilde{X}_2,\ldots,\widetilde{X}_n)'$, $\widetilde{\boldsymbol{Y}}=(\widetilde{Y}_1,\widetilde{Y}_2,\ldots,\widetilde{Y}_n)'$, $\widetilde{\boldsymbol{f}}=(I-\boldsymbol{K})\boldsymbol{f}$, $\widetilde{\boldsymbol{g}}=(I-\boldsymbol{K})\boldsymbol{g}$, $\widetilde{\boldsymbol{\eta}}=(I-\boldsymbol{K})\boldsymbol{\eta}$, 其中

$$\boldsymbol{K}=(K_{ij})_{n\times n},\qquad K_{ij}=\frac{K\Big(\dfrac{T_i-T_j}{h}\Big)}{\sum\limits_{j=1}^{n}K\Big(\dfrac{T_i-T_j}{h}\Big)}.$$

Speckman (1988)提出的截面最小二乘估计量为

$$\widehat{\beta}_{\mathrm{PLS}}=(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}. \tag{2.3}$$

结合Speckman (1988)的截面最小二乘思想及Zou (2006)提出的adaptive LASSO线性回归模型变量选择方法, 我们定义部分线性模型系数向量的adptive LASSO估计为

$$\widehat{\beta}_{\mathrm{ads}}=\arg\min_{\Phi}\Big\{\|\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{X}}\Phi\|_2^2+\lambda_n\sum_{j=1}^{p}w_j|\phi_j|\Big\}, \tag{2.4}$$

其中λ_n为惩罚参数, $\|\Phi\|_l=\Big(\sum\limits_{i=1}^{p}|\phi_i|^l\Big)^{1/l}$, $w_j\,(j=1,2,\ldots,p)$为权重函数, w_j应为$|\phi_j|$的减函数, 即对较大的$|\phi_j|$应给予较小的惩罚, 以得到较小的偏差甚至无偏的参数估计量, 而对较小的$|\phi_j|$应给予较大的惩罚, 以得到较为精简的模型; 这里可取$w_j=|(\widehat{\beta}_{\mathrm{PLS}})_j|^{-\gamma}$, $\gamma>0$. 由Speckman (1988)可知, 在一些正则条件下, $\widehat{\beta}_{\mathrm{PLS}}$为$\beta$的$\sqrt{n}$相合估计. 记$\mathcal{A}_n=\{j:(\widehat{\beta}_{\mathrm{ads}})_j\neq 0,j=1,\ldots,p\}$, 进而可得非参数成分的估计

$$\widehat{f}(T)=\widehat{m}_Y(T)-(\widehat{m}_X(T))'\widehat{\beta}_{\mathrm{ads}}. \tag{2.5}$$

令$w_j\phi_j=\psi_j$, $\Psi=(\psi_1,\ldots,\psi_p)'$, $\widetilde{\boldsymbol{Z}}=\widetilde{\boldsymbol{X}}\Omega^{-1}$, 其中$\Omega=\mathrm{diag}(w_1,\ldots,w_p)$, 于是部分线性模型adaptive LASSO优化问题可转化为

$$\Psi^*=\arg\min_{\Psi}\Big\{\|\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{Z}}\Psi\|_2^2+\lambda_n\sum_{j=1}^{p}|\psi_j|\Big\},$$

即优化问题(2.4)可变换为线性模型LASSO解, 本质上仍为线性优化问题, 分别用$\widetilde{\boldsymbol{Y}}$和$\widetilde{\boldsymbol{Z}}$替换$\boldsymbol{Y}$和$\boldsymbol{Z}$后可采用LARS算法[6]实现, 进而可得$\widehat{\beta}_{\mathrm{ads}}=\Omega^{-1}\Psi^*$.

光滑参数窗宽h和惩罚参数λ_n的选择是一个关键的问题, 与线性模型变量选择方法不同, 这里需要选择两个参数. 一种常用的方法就是数据驱动的二维网格搜索法, 如交

叉核实法(CV)或广义交叉核实法(GCV)[17], 然而这样做将极大地增加计算量, 运算效率低. 这里我们也采用类似Wang等(2007)的方法, 分别选择窗宽h和惩罚参数λ. 首先通过Fan等(2004)提出的差分法给出全模型β的估计$\widehat{\beta}_{\rm DF}$, 由Yatchew (1997)和Wang等(2011)可知, $\widehat{\beta}_{\rm DF}$是β的一个$\sqrt{n}$相合估计; 接下来用$\widehat{\beta}_{\rm DF}$来替换部分线性模型(2.1)中的β, 模型(2.1)转化为一元非参函数估计问题, 对窗宽h, 可以采用Ruppert等(1995)提出的插入法或者采用CV, GCV选择. 由条件(5)可知, 最优的窗宽阶数为$O(n^{-1/5})$, 在此条件下可得定理3.3成立. 在选定h后, 惩罚参数λ_n的选择可通过一些常用的调整参数选取方法实现, 如CV、GCV、AIC或BIC等.

§3. 渐近性质

本节我们研究部分线性模型兴趣参数β的adaptive LASSO估计量$\widehat{\beta}_{\rm ads}$的性质. 为了方便定理证明, 本文需要如下假设条件:

(1) 核函数$K(\cdot)$的支撑集为$[-1,1]$, 存在常数$0\le M_1<M_2$, 核函数满足

$$M_1\le K(u)_{u\in[-1,1]}\le M_2,\quad \text{且}\int K(u){\rm d}u=1,\ \int uK(u){\rm d}u=0,\ \int u^2K(u){\rm d}u\neq 0.$$

(2) $n^{-1}\boldsymbol{\eta}'\boldsymbol{\eta}\to_p V$, $V=\begin{pmatrix} V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}$, V_{11}为$p_0\times p_0$阵.

(3) ${\rm tr}(\boldsymbol{K}'\boldsymbol{K})=\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}K_{ij}^2=O_p(h^{-1})$, ${\rm tr}(\boldsymbol{K})=O_p(h^{-1})$.

(4) $\|\widetilde{\boldsymbol{f}}\|_2^2=O_p(nh^4)$.

(5) $h=O(n^{-1/5})$.

注记 1　(1)–(5)都是半参数模型常见的假设条件. (1)是对核函数的一般假定, (2)–(5)为Speckman (1988)中的条件, (5)为普通核估计的最优窗宽.

定理 3.1　在条件(1)–(5)下, 若V非奇异, $\lambda_n/n\to\lambda_0$, 则$\widehat{\beta}_{\rm ads}\to_p\arg\min(Z)$, 其中

$$Z(\Phi)=(\Phi-\beta)'V(\Phi-\beta)+\lambda_0\sum_{j=1}^{p}w_j|\phi_j|.$$

定理3.1表明: 若$\lambda_n=o(n)$, 则$\arg\min(Z)=\beta$, 于是$\widehat{\beta}_{\rm ads}$为相合估计.

定理 3.2　假设条件(1)–(5)成立, 若$\lambda_n/\sqrt{n}\to\lambda_0\ge 0$且$V$非奇异, 则

$$\sqrt{n}(\widehat{\beta}_{\rm ads}-\beta)\to_d\arg\min(Q),$$

其中$Q(u)=-2u'W+u'Vu+\lambda_0\sum\limits_{j=1}^{p}w_j[u_j{\rm sgn}(\beta_j)I(\beta_j\neq 0)+|u_j|I(\beta_j=0)]$, $W\sim N(\boldsymbol{0},\sigma^2V)$.

定理3.2表明: 若$\lambda_n = o(\sqrt{n})$, 则$\sqrt{n}(\widehat{\beta}_{\text{ads}} - \beta) \to_d V^{-1}W \sim N(\mathbf{0}, \sigma^2 V^{-1})$; 若$\lambda_n = O(\sqrt{n})$, 则$\beta$非零系数$\beta_{\mathcal{A}}$的估计$\widehat{\beta}_{\text{ads}\mathcal{A}}$不是$\sqrt{n}$-相合的.

定理 3.3 假设条件(1)–(5)成立, $w_j = |(\widehat{\beta}_{\text{PLS}})_j|^{-\gamma}$, $j = 1, 2, \ldots, p$, 若$\lambda_n/\sqrt{n} \to 0$且$\lambda_n n^{(\gamma-1)/2} \to \infty$, V非奇异, 则

(i) 相合性: $\lim\limits_n \mathsf{P}(\mathcal{A}_n = \mathcal{A}) = 1$,

(ii) 渐近正态性: $\sqrt{n}(\widehat{\beta}_{\text{ads}\mathcal{A}} - \beta_{\mathcal{A}}) \to_d N(0, \sigma^2 V_{11}^{-1})$.

定理3.3说明: 在一些正则条件下, 部分线性adaptive LASSO参数估计量具有Fan等(2001)提出的oracle性质.

注记 2 与线性模型相比, 部分线性模型包含非参数函数部分, 需要更多的假设条件, 如(1), (3)–(5), 技术上构造adaptive LASSO参数估计量也需先消除非参函数的影响, 同时还需考虑窗宽的选择问题. 但定理3.1–定理3.3的结果表明, 部分线性模型的adaptive LASSO参数估计和线性模型的adaptive LASSO参数估计(见Zou (2006))有着相似的渐近结果. 这与部分线性模型截面最小二乘参数估计量和线性模型最小二乘参数估计量具有相似性质相吻合.

§4. 定理证明

方便起见, 本文定理证明中的期望和方差都是给定T条件下的条件期望和条件方差.

定理3.1的证明: $\widehat{\beta}_{\text{ads}}$的渐近性质由目标函数$Z_n(\Phi)$的渐近性决定, 经简单计算可得,

$$\begin{aligned} Z_n(\Phi) &= \frac{1}{n}\|\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\Phi\|_2^2 + \frac{1}{n}\lambda_n\|\Omega\Phi\|_1 \\ &= \frac{1}{n}(\beta-\Phi)'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}(\beta-\Phi) + \frac{1}{n}\widetilde{\boldsymbol{f}}'\widetilde{\boldsymbol{f}} + \frac{1}{n}\boldsymbol{\epsilon}'(I-\boldsymbol{K})'(I-\boldsymbol{K})\boldsymbol{\epsilon} \\ &\quad + \frac{2}{n}\widetilde{\boldsymbol{f}}'\widetilde{\boldsymbol{X}}(\beta-\Phi) + \frac{2}{n}\widetilde{\boldsymbol{f}}'(I-\boldsymbol{K})\boldsymbol{\epsilon} \\ &\quad + \frac{2}{n}(\beta-\Phi)'\widetilde{\boldsymbol{X}}'(I-\boldsymbol{K})\boldsymbol{\epsilon} + \frac{1}{n}\lambda_n\|\Omega\Phi\|_1, \end{aligned}$$

接下来逐项分析$Z_n(\Phi)$的展式. 对第一项, 由Speckman (1988)可知$n^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} \to_p V$, 于是$n^{-1}(\beta-\Phi)'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}(\beta-\Phi) = (\beta-\Phi)'V(\beta-\Phi) + o_p(1)$. 对第二项, 由条件(4)可得$n^{-1}\widetilde{\boldsymbol{f}}'\widetilde{\boldsymbol{f}} = n^{-1}\|\widetilde{\boldsymbol{f}}\|_2^2 = O_p(h^4)$. 对第三项,

$$\begin{aligned} \frac{1}{n}\boldsymbol{\epsilon}'(I-\boldsymbol{K})'(I-\boldsymbol{K})\boldsymbol{\epsilon} &= \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{K}'\boldsymbol{K}\boldsymbol{\epsilon} - \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{K}'\boldsymbol{\epsilon} - \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{K}\boldsymbol{\epsilon} \\ &= \sigma^2 + \frac{1}{n}\mathrm{tr}([\boldsymbol{K}'\boldsymbol{K} - \boldsymbol{K}' - \boldsymbol{K}]\boldsymbol{\epsilon}\boldsymbol{\epsilon}') + o_p(1) \\ &= \sigma^2 + O_p((nh)^{-1}) + o_p(1), \end{aligned}$$

其中$\mathrm{tr}(\boldsymbol{K}'\boldsymbol{K}-\boldsymbol{K}'-\boldsymbol{K})=O_p(h^{-1})$可由条件(3)得到. 对第四项, 经简单计算可得,

$$\begin{aligned}\frac{2}{n}\widetilde{\boldsymbol{f}}'\widetilde{\boldsymbol{X}}(\beta-\Phi) &= \frac{2}{n}(\widetilde{\boldsymbol{f}}'\widetilde{\boldsymbol{g}}+\widetilde{\boldsymbol{f}}'\boldsymbol{\eta}-\widetilde{\boldsymbol{f}}'\boldsymbol{K}\boldsymbol{\eta})(\beta-\Phi)\\ &= O_p(h^4)+O_p(n^{-1/2}h^2)+O_p((n^{-1}h^3)^{1/2}),\end{aligned}$$

由条件(4)易推出$n^{-1}\widetilde{\boldsymbol{f}}'\widetilde{\boldsymbol{g}}_j=O_p(h^4)$, $j=1,2,\ldots,p$. 又因为$\mathsf{E}(\widetilde{\boldsymbol{f}}'\boldsymbol{\eta}_j)=0$, $\mathsf{Var}\,(\widetilde{\boldsymbol{f}}'\boldsymbol{\eta}_j)=V_{jj}\|\widetilde{\boldsymbol{f}}\|_2^2=O_p(nh^4)$, 从而$n^{-1}\widetilde{\boldsymbol{f}}'\boldsymbol{\eta}_j=O_p(n^{-1/2}h^2)$. 由条件(2), (3)可得, $\|\boldsymbol{K}\boldsymbol{\eta}_j\|_2^2=\mathrm{tr}(\boldsymbol{K}'\boldsymbol{K}\boldsymbol{\eta}_j\boldsymbol{\eta}_j')=O_p(h^{-1})$, $\widetilde{\boldsymbol{f}}'\boldsymbol{K}\boldsymbol{\eta}_j\le\|\widetilde{\boldsymbol{f}}\|_2\cdot\|\boldsymbol{K}\boldsymbol{\eta}_j\|_2=O_p((nh^3)^{1/2})$. 与第四项分析类似, 可得第五项

$$\frac{1}{n}\widetilde{\boldsymbol{f}}'(I-\boldsymbol{K})\boldsymbol{\epsilon}=O_p(n^{-1/2}h^2)+O_p((n^{-1}h^3)^{1/2}).$$

对第六项,

$$\begin{aligned}\frac{1}{n}(\beta-\Phi)'\widetilde{\boldsymbol{X}}(I-\boldsymbol{K})\boldsymbol{\epsilon} &= \frac{1}{n}(\beta-\Phi)'(\widetilde{\boldsymbol{X}}'\boldsymbol{\epsilon}-\widetilde{\boldsymbol{X}}'\boldsymbol{K}\boldsymbol{\epsilon})\\ &= O_p(n^{-1/2})+O_p((nh)^{-1/2}),\end{aligned}$$

记$\widetilde{\boldsymbol{X}}=(\widetilde{\boldsymbol{x}}_1,\widetilde{\boldsymbol{x}}_2,\ldots,\widetilde{\boldsymbol{x}}_p)$, 由$n^{-1}\widetilde{\boldsymbol{x}}_j'\widetilde{\boldsymbol{x}}_j\to_p V_{jj}$及$\epsilon$与$(X,T)$的独立性可得, $\mathsf{E}(\widetilde{\boldsymbol{x}}_j'\boldsymbol{\epsilon})=0$, $\mathsf{Var}\,(\widetilde{\boldsymbol{x}}_j'\boldsymbol{\epsilon})=n\sigma^2V_{jj}+o_p(n)$, 从而$\widetilde{\boldsymbol{x}}_j'\boldsymbol{\epsilon}=O_p(n^{1/2})$. $\widetilde{\boldsymbol{x}}_j\boldsymbol{K}\boldsymbol{\epsilon}\le\|\widetilde{\boldsymbol{x}}_j\|_2\cdot\|\boldsymbol{K}\boldsymbol{\epsilon}\|_2=O_p(n^{1/2}h^{-1/2})$.

在条件(5)下, $Z_n(\Phi)$前六项中除了第一项和第三项外, 其它项均为$o_p(1)$, 即

$$\frac{1}{n}\|\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{X}}\Phi\|_2^2\to_p(\beta-\Phi)'V(\beta-\Phi)+\sigma^2. \tag{4.1}$$

由$Z_n(\Phi)$的凸性及Anderson等(1982)和Pollard (1991)中的结论, 可得$\widehat{\beta}_n=O_p(1)$. 因此, 综合(4.1)和$\widehat{\beta}_n=O_p(1)$可得$\arg\min(Z_n)\to_p\arg\min(Z)$. 定理3.1得证. □

定理3.2的证明: 记$\sqrt{n}(\Phi-\beta)=u$, 由定理3.1对$Z_n(\Phi)$的分解式, $nZ_n(\Phi)$可转换为

$$Q_n(u)=u'\Big(\frac{1}{n}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\Big)u-\frac{2}{\sqrt{n}}u'[\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{f}}+\widetilde{\boldsymbol{X}}'(I-\boldsymbol{K})\boldsymbol{\epsilon}]+\lambda_n\Big\|\Omega\Big(\frac{u}{\sqrt{n}}+\beta\Big)\Big\|_1+S,$$

其中S是与u无关的项. 由定理3.1的证明过程可知:

$$\begin{aligned}&\frac{1}{n}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\to_p V,\qquad \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{f}}=O_p(nh^4)+O_p((nh^4)^{1/2})+O_p((nh^3)^{1/2}),\\ &\widetilde{\boldsymbol{X}}'(I-\boldsymbol{K})\boldsymbol{\epsilon}=O_p(n^{1/2})+O_p(n^{1/2}h^{-1/2}).\end{aligned}$$

在条件(5)下, $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{f}}=O_p(n^{1/5})=o_p(n^{1/2})$, 可以对其不予考虑. 而

$$\begin{aligned}&\mathsf{E}\Big(\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{X}}'(I-\boldsymbol{K})\boldsymbol{\epsilon}\Big)=0,\\ &\mathsf{Var}\,\Big(\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{X}}'(I-\boldsymbol{K})\boldsymbol{\epsilon}\Big)=\frac{1}{n}\sigma^2\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}-\frac{1}{n}\sigma^2\widetilde{\boldsymbol{X}}'(\boldsymbol{K}\boldsymbol{K}'-\boldsymbol{K}-\boldsymbol{K}')\widetilde{\boldsymbol{X}},\end{aligned}$$

由Speckman (1988)式(4.5b)的证明可知$n^{-1}\widetilde{\boldsymbol{X}}'(\boldsymbol{K}\boldsymbol{K}'-\boldsymbol{K}-\boldsymbol{K}')\widetilde{\boldsymbol{X}}=o_p(1)$, 从而可得

$$\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{X}}'(I-\boldsymbol{K})\boldsymbol{\epsilon}\to_d N(0,\sigma^2V).$$

又

$$\lambda_n\Big\|\Omega\Big(\frac{u}{\sqrt{n}}+\beta\Big)\Big\|_1=\lambda_n\sum_{j=1}^p w_j\Big|\frac{u_j}{\sqrt{n}}+\beta_j\Big|,$$

对其在β_j一阶泰勒展开可得

$$\lambda_n\sum_{j=1}^p w_j\Big|\frac{u_j}{\sqrt{n}}+\beta_j\Big|\approx\lambda_n\sum_{j=1}^p w_j\Big[\Big(|\beta_j|+\frac{u_j}{\sqrt{n}}\operatorname{sgn}(\beta_j)\Big)I(\beta_j\neq0)+\frac{|u_j|}{\sqrt{n}}I(\beta_j=0)\Big].$$

记$Q(u)=-2u'W+u'Vu+\lambda_0\sum_{j=1}^p w_j[u_j\operatorname{sgn}(\beta_j)I(\beta_j\neq0)+|u_j|I(\beta_j=0)]$, 其中$W\sim N(\boldsymbol{0},\sigma^2V)$. 由$Q_n(u)$的凸性知$Q_n(u)$有唯一解, 且$\arg\min Q_n(u)\to_d\arg\min Q(u)$; 又因为$\sqrt{n}(\Phi-\beta)=u$, 且$Z_n(\Phi)$在$\widehat{\beta}_{\text{ads}}$有最小值, 所以可得$\sqrt{n}(\widehat{\beta}_{\text{ads}}-\beta)\to_d\arg\min(Q(u))$. 定理3.2得证. □

定理3.3的证明: 记$\sqrt{n}(\Phi-\beta)=u$, 由定理3.2的证明, 目标函数$nZ_n(\Phi)$可近似转换为

$$\begin{aligned}Q_n'(u)&=u'\Big(\frac{1}{n}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\Big)u-\frac{2}{\sqrt{n}}u'\widetilde{\boldsymbol{X}}'(I-\boldsymbol{K})\boldsymbol{\epsilon}\\&\quad+\lambda_n\sum_{j=1}^p w_j\Big[\frac{u_j}{\sqrt{n}}\operatorname{sgn}(\beta_j)I(\beta_j\neq0)+\frac{|u_j|}{\sqrt{n}}I(\beta_j=0)\Big]+S,\end{aligned}$$

其中S是与u无关的项, 并且如下渐近结果成立,

$$\frac{1}{n}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\to_p V,\qquad\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{X}}'(I-\boldsymbol{K})\boldsymbol{\epsilon}\to_d N(0,\sigma^2V).$$

对$Q_n'(u)$第三项$\lambda_n\sum_{j=1}^p w_j[(u_j/\sqrt{n})\operatorname{sgn}(\beta_j)I(\beta_j\neq0)+(|u_j|/\sqrt{n})I(\beta_j=0)]$, 当$\beta_j\neq0$时, $w_j=|(\widehat{\beta}_{\text{PLS}})_j|^{-\gamma}\to_p|\beta_j|^{-\gamma}$, $\lambda_n/\sqrt{n}\to0$, 可得

$$\frac{\lambda_n}{\sqrt{n}}w_j\operatorname{sgn}(\beta_j)\to_p0;$$

当$\beta_j=0$时, $\sqrt{n}|(\widehat{\beta}_{\text{PLS}})_j|^{-\gamma}=O_p(1)$, 由条件$\lambda_n n^{(\gamma-1)/2}\to\infty$可知,

$$\frac{\lambda_n}{\sqrt{n}}w_j=\frac{\lambda_n}{\sqrt{n}}n^{\gamma/2}(|\sqrt{n}(\widehat{\beta}_{\text{PLS}})_j|)^{-\gamma}\to_p\infty.$$

于是, 由Slutsky定理$Q_n'(u)\to_d Q'(u)$成立, 其中

$$Q'(u)=\begin{cases}u_{\mathcal{A}}'V_{11}u_{\mathcal{A}}-2u_{\mathcal{A}}'W_{\mathcal{A}}, & \text{如果 } u_j=0,\ \forall j\notin\mathcal{A};\\ \infty, & \text{其它.}\end{cases}$$

$Q_n'(u)$为凸函数, $Q'(u)$的唯一最小值解为$(V_{11}^{-1}W_{\mathcal{A}},0)'$, 并且有$\widehat{u}_{n\mathcal{A}} \to_d V_{11}^{-1}W_{\mathcal{A}}$, $\widehat{u}_{n\mathcal{A}^c} \to_d 0$. 又因为$\sqrt{n}(\Phi-\beta)=u$, 且$Z_n(\Phi)$在$\widehat{\beta}_{\text{ads}}$有最小值, 所以可得

$$\sqrt{n}(\widehat{\beta}_{\text{ads}\mathcal{A}}-\beta_{\mathcal{A}}) \to_d N(0,\sigma^2V_{11}^{-1}),$$

渐近正态性得证.

接下来证明模型选择相合性: $\forall j\in\mathcal{A}$, 渐近正态性表明$(\widehat{\beta}_{\text{ads}})_j \to_p \beta_j$, 于是可得$\mathrm{P}(j\in\mathcal{A}_n)\to 1$. 此外只需证明$\forall j'\notin\mathcal{A}$, $\mathrm{P}(j'\in\mathcal{A}_n)\to 0$. 考虑事件$j'\in\mathcal{A}_n$, 由Karush-Kuhn-Tucker最优化条件, 可得$2\widetilde{\boldsymbol{x}}_{j'}'(\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{X}}\widehat{\beta}_{\text{ads}})=\lambda_n w_{j'}\operatorname{sgn}(\beta_{j'})$, 其中$\widetilde{\boldsymbol{x}}_{j'}=(\widetilde{X}_{1j'},\widetilde{X}_{2j'},\ldots,\widetilde{X}_{nj'})'$. 注意到$(\lambda_n/\sqrt{n})w_{j'}=(\lambda_n/\sqrt{n})n^{\gamma/2}(|\sqrt{n}(\widehat{\beta}_{\text{PLS}})_{j'}|)^{-\gamma}\to_p\infty$,

$$\begin{aligned}2\frac{\widetilde{\boldsymbol{x}}_{j'}'(\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{X}}\widehat{\beta}_{\text{ads}})}{\sqrt{n}} &= 2\frac{\widetilde{\boldsymbol{x}}_{j'}'[\widetilde{\boldsymbol{X}}(\beta-\widehat{\beta}_{\text{ads}})+\widetilde{\boldsymbol{f}}+(I-\boldsymbol{K})\boldsymbol{\epsilon}]}{\sqrt{n}}\\ &= 2\frac{\widetilde{\boldsymbol{x}}_{j'}'\widetilde{\boldsymbol{X}}\sqrt{n}(\beta-\widehat{\beta}_{\text{ads}})}{n}+2\frac{\widetilde{\boldsymbol{x}}_{j'}'\widetilde{\boldsymbol{f}}}{\sqrt{n}}+2\frac{\widetilde{\boldsymbol{x}}_{j'}'(I-\boldsymbol{K})\boldsymbol{\epsilon}}{\sqrt{n}},\end{aligned}$$

由定理3.2的证明可知:

$$2\frac{\widetilde{\boldsymbol{x}}_{j'}'\widetilde{\boldsymbol{X}}\sqrt{n}(\beta-\widehat{\beta}_{\text{ads}})}{n}=O_p(1),\qquad 2\frac{\widetilde{\boldsymbol{x}}_{j'}'(I-\boldsymbol{K})\boldsymbol{\epsilon}}{\sqrt{n}}=O_p(1),\qquad 2\frac{\widetilde{\boldsymbol{x}}_{j'}'\widetilde{\boldsymbol{f}}}{\sqrt{n}}=o_p(1),$$

即

$$2\frac{\widetilde{\boldsymbol{x}}_{j'}'(\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{X}}\widehat{\beta}_{\text{ads}})}{\sqrt{n}}=O_p(1).$$

因此可得

$$P(j'\in\mathcal{A}_n)\le P\Big(\Big|2\frac{\widetilde{\boldsymbol{x}}_{j'}'(\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{X}}\widehat{\beta}_{\text{ads}})}{\sqrt{n}}\Big|=\frac{\lambda_n}{\sqrt{n}}w_{j'}\Big)\to 0,$$

模型选择相合性得证. □

§5. 数值模拟

本节我们通过蒙特卡洛模拟实验对比研究部分线性模型adaptive LASSO与LASSO, SCAD, MCP变量选择方法以及oracle估计量的小样本表现. Oracle估计量假定系数的非零分量集合已知, 在实际中并不可用, 但它可以作为衡量变量选择方法优劣的一个重要标准. 为比较以上方法在变量选择及参数估计方面的表现, 程序重复运行500次, 给出以下估计量的近似值: 模型参数β估计的平均二次损失(ML2)、β非零分量估计的偏差(Bias)和标准差(SD)、平均正确选出模型的比率(RSR)、非零分量中被估成零的个数的平均值(F^-)、零分量未被估成零的个数的平均值(F^+).

考虑部分线性模型$Y=X'\beta+f(T)+\epsilon$, $p=15$, $\beta=(3,3,1,1,0.5,0.5,0,\ldots,0)'$, 即$\beta$的15个分量中仅有6个非零, 可看做分别对应相对较大, 适中, 较小的影响变量; $X\sim$

$N_{15}(0,V)$, 其中$V_{ij}=0.5^{|i-j|}$; $f(T)=\sin(2\pi T)$, 其中T服从$[-1,1]$上的均匀分布, $\epsilon\sim N(0,1)$. 分别生成样本容量n为100, 200, 400的样本. 非参估计核函数取Epanechnikov核, 即$K(u)=0.75(1-u^2)_+$. 对权重函数w_j, $j=1,2,\ldots,p$, 取$\gamma=1$, 窗宽依第二节中介绍的方法选出. 对于惩罚参数, Zou等(2007)建议采用BIC选取LASSO调整参数, Wang等(2007)指出对于SCAD, BIC调整参数选择方法优于GCV和AIC. 因此, 这里我们采用BIC选取惩罚参数. MCP惩罚参数的选择采用Breheny和Huang (2011)提出的混合法. 各估计量的估计结果如表1.

表1 蒙特卡洛模拟实验结果表

n	method	ML2	RSR	F^+	F^-	sta.	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$
100	LASSO	0.1609	0.378	1.15	0.006	Bias	-0.0841	-0.0359	-0.0484	-0.0336	-0.0374	-0.0880
						SD	0.1383	0.1535	0.1433	0.1420	0.1408	0.1325
	a.LASSO	0.1800	0.606	0.488	0.094	Bias	-0.0120	0.0138	-0.0330	0.0182	-0.0468	-0.0461
						SD	0.1357	0.1543	0.1528	0.1566	0.1889	0.1668
	SCAD	0.1811	0.714	0.368	0.144	Bias	-0.0070	0.0031	-0.0071	0.0042	0.0163	-0.0373
						SD	0.1344	0.1524	0.1441	0.1445	0.1729	0.2060
	MCP	0.1837	0.618	0.416	0.124	Bias	-0.0048	0.0033	-0.0061	0.0099	-0.0109	-0.0016
						SD	0.1363	0.1516	0.1430	0.1504	0.1882	0.1677
	oracle	0.1138	–	–	–	Bias	-0.0053	0.0044	-0.0071	0.0028	0.0024	0.0023
						SD	0.1337	0.1499	0.1405	0.1386	0.1367	0.1259
200	LASSO	0.0719	0.468	0.848	0	Bias	-0.0640	-0.0230	-0.0291	-0.0363	-0.0294	-0.0639
						SD	0.0864	0.0959	0.0993	0.1029	0.0941	0.0912
	a.LASSO	0.0666	0.818	0.214	0.004	Bias	-0.0087	0.0129	-0.0127	-0.0006	-0.0242	-0.0222
						SD	0.0844	0.0956	0.1036	0.1078	0.1100	0.1005
	SCAD	0.0657	0.802	0.28	0.004	Bias	-0.0047	0.0065	0.0003	-0.0074	-0.0013	-0.0013
						SD	0.0841	0.0947	0.0999	0.1026	0.0967	0.0932
	MCP	0.0628	0.816	0.214	0.006	Bias	-0.0047	0.0068	0.0002	-0.0068	-0.0005	0.0016
						SD	0.0844	0.0949	0.1001	0.1026	0.0971	0.0918
	oracle	0.0523	–	–	–	Bias	-0.0050	0.0067	0.0003	-0.0069	-0.0001	0.0024
						SD	0.0837	0.0946	0.0992	0.1012	0.0941	0.0861
400	LASSO	0.0331	0.564	0.642	0	Bias	-0.0385	-0.0256	-0.0200	-0.0224	-0.0159	-0.0482
						SD	0.0588	0.0688	0.0687	0.0676	0.0634	0.0622
	a.LASSO	0.0273	0.916	0.092	0	Bias	0.0031	-0.0009	-0.0051	0.0023	-0.0053	-0.0139
						SD	0.0580	0.0684	0.0696	0.0692	0.0674	0.0633
	SCAD	0.0267	0.916	0.096	0	Bias	0.0049	-0.0038	0.0017	-0.0006	0.0062	-0.0015
						SD	0.0578	0.0682	0.0685	0.0673	0.0628	0.0591
	MCP	0.0271	0.902	0.112	0	Bias	0.0050	-0.0039	0.0017	-0.0006	0.0062	-0.0013
						SD	0.0578	0.0681	0.0684	0.0673	0.0628	0.0589
	oracle	0.0246	–	–	–	Bias	0.0050	-0.0039	0.0016	-0.0006	0.0061	-0.0013
						SD	0.0578	0.0681	0.0684	0.0673	0.0626	0.0588

表1中的结果表明: (1)adaptive LASSO参数估计量的偏差明显比LASSO估计的偏要小, 这一点与LASSO为有偏估计, adaptive LASSO为渐近无偏估计理论结果一致. (2)adaptive LASSO与SCAD, MCP变量选择方法在正确选出模型的能力和参数估计方面表现相似, 且优于LASSO估计. 特别是随着样本容量的增大, adaptive LASSO, SCAD及MCP的

参数估计结果与oracle估计结果接近. (3)随着样本容量的增大, LASSO, adaptive LASSO, SCAD, MCP的变量选择能力增强, 它们四者以及oracle参数估计的准确度和精确度也明显增加.

参 考 文 献

[1] Engle, R.F., Granger, C.W.J., Rice, J. and Weiss, A., Semiparametric estimates of the relation between weather and electricity sales, *Journal of the American Statistical Association*, **81**(1986), 310–320.

[2] Speckman, P., Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society: Series B*, **50**(1988), 413–436.

[3] Härdle, W., Liang, H. and Gao, J.T., *Partially Linear Models*, Springer Physica, Heidelberg, 2000.

[4] Tibshirani, R., Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B*, **58**(1996), 267–288.

[5] Knight, K. and Fu, W.J., Asymptotics for lasso-type estimators, *The Annals of Statistics*, **28**(2000), 1356–1378.

[6] Efron, B., Hastie, T. and Tibshirani, R., Least angle regression (with discussion), *The Annals of Statistics*, **32**(2004), 407–451.

[7] Fan, J. and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**(2001), 1348–1360.

[8] Fan, J. and Peng, H., Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics*, **32**(2004), 928–961.

[9] Candès, E. and Tao, T., The Dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics*, **35**(2007), 2313–2351.

[10] Zhang, C., Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**(2010), 894–942.

[11] Fan, J. and Li, R., New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *Journal of the American Statistical Association*, **99**(2004), 710–723.

[12] Xie, H.L. and Huang, J., SCAD-penalized regression in high-dimensional partially linear models, *The Annals of Statistics*, **37**(2009), 673–696.

[13] Liang, H. and Li, R.Z., Variable selection for partially linear models with measurement errors, *Journal of the American Statistical Association*, **104**(2009), 234–248.

[14] Ni, X., Zhang, H.H. and Zhang, D., Automatic model selection for partially linear models, *Journal of Multivariate Analysis*, **100**(2009), 2100–2111.

[15] Zou, H., The adaptive lasso and its oracle properties, *Journal of American Statistical Association*, **101**(2006), 1418–1429.

[16] Rice, J., Convergence rates for partially splined models, *Statistics & Probability Letters*, **4**(1986), 203–208.

[17] Craven, P. and Wahba, G., Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerishe Mathematik*, **31**(1979), 377–403.

[18] Wang, H., Li, R. and Tsai, C., Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, **94**(2007), 553–568.

[19] Yatchew, A., An elementary estimator for the partially linear model, *Economics letters*, **57**(1997), 135-143.

[20] Wang, L., Brown, L.D. and Cai, T.T., A difference based approach to the semiparametric partially linear model, *Electronic Journal of Statistics*, **5**(2011), 619-641.

[21] Ruppert, D., Sheather, S.J. and Wand, M.P., An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, **90**(1995), 1257–1270.

[22] Zou, H., Hastie, T. and Tibshirani, R., On the "degrees of freedom" of the lasso, *The Annals of Statistics*, **35**(2007), 2173–2192.

[23] Anderson, P.K. and Gill, R.D., Cox's regression model for counting processes: a large sample study, *The Annals of Statistics*, **10**(1982), 1100–1120.

[24] Pollard, D., Asymptotics for least absolute deviation regression estimators, *Econometric Theory*, **7**(1991), 186–199.

[25] Breheny, P. and Huang, J., Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, *The Annals of Applied Statistics*, **5**(2011), 232–253.

Variable Selection for Partially Linear Models via Adaptive LASSO

Li Feng

(*Resources & Economic Trade Department, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou, 450015*)

Lu Yiqiang

(*Institute of electronic technology, The PLA Information Engineering University, Zhengzhou, 450004*)

Li Gaorong

(*College of Applied Sciences, Beijing University of Technology, Beijing, 100142*)

Partially linear model is a class of commonly used semiparametric models, this paper focus on variable selection and parameter estimation for partially linear models via adaptive LASSO method. Firstly, based on profile least squares and adaptive LASSO method, the adaptive LASSO estimator for partially linear models are constructed, and the selections of penalty parameter and bandwidth are discussed. Under some regular conditions, the consistency and asymptotic normality for the estimator are investigated, and it is proved that the adaptive LASSO estimator has the oracle properties. The proposed method can be easily implemented. Finally a Monte Carlo simulation study is conducted to assess the finite sample performance of the proposed variable selection procedure, results show the adaptive LASSO estimator behaves well.

Keywords: Partially linear models, variable selection, asymptotic distribution, LASSO, adaptive LASSO.

AMS Subject Classification: 62F12, 62G05, 62H12.