高维部分线性模型中的变量选择

杨宜平，薛留根，王学娟
（北京工业大学 应用数理学院，北京 100124）

摘 要：研究了高维部分线性模型中的变量选择，结合样条方法和 Dantzig 或 Lasso 变量选择方法，同时进行变量选择和未知参数估计，证明了估计误差的非渐近界. 模拟结果说明，该方法在参数维数较高时优于已有方法.

关键词：样条；估计理论；线性回归；Monte Carlo 方法
中图分类号：O 212.7 **文献标志码**：A **文章编号**：0254－0037(2011)02－0291－05

1 研究背景

考虑部分线性模型

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + g(T_i) + \varepsilon_i,\ i = 1,\cdots,n \tag{1}$$

其中，Y_i 是响应变量；$\boldsymbol{X}_i$ 是 p 维协变量；$\boldsymbol{\beta}$ 是 p 维未知参数；$g(\cdot)$ 是未知的函数；ε_i 是随机误差且 ε_i 与 $(\boldsymbol{X}_i, T_i)$ 独立，均值为 0，方差为 σ^2. 不失一般性，假定 T_i 在闭区间 $[0,1]$ 上取值. 传统的模型都假定 $\boldsymbol{\beta}$ 是有限维，提出了估计参数和非参数的方法，如核方法[1]、样条方法[2]、局部线性方法[3]等. 但是，当 p 很大，与 n 的大小相当甚至大于 n 时，已有的方法不能处理该问题，因此，作者考虑高维的部分线性模型，假定 p 很大，$\boldsymbol{\beta}$ 稀疏，即 $\boldsymbol{\beta}$ 的一些元素是 0，对非参数部分，采用样条方法逼近非参数函数 $g(t)$，然后用 Dantzig 或 Lasso 变量选择方法进行变量选择[4-5]，同时估计未知参数. 关于变量选择的更多研究可参见文献[6-13].

2 方法与主要结果

令 $\boldsymbol{B}(T_i) = [b_1(T_i), b_2(T_i), \cdots, b_q(T_i)]^{\mathrm{T}}$ 是 q 维的基函数，$g(T_i)$ 由 q 维的样条逼近，允许一些误差，即

$$g(T_i) = \boldsymbol{B}^{\mathrm{T}}(T_i)\boldsymbol{\eta} + e(T_i) \tag{2}$$

其中 $\boldsymbol{\eta}$ 是 q 维参数. 由式(1) 和式(2) 可得

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{B}^{\mathrm{T}}(T_i)\boldsymbol{\eta} + \varepsilon_i^* \tag{3}$$

其中 $\varepsilon_i^* = \varepsilon_i + e(T_i)$.

假设 $\boldsymbol{\beta}$ 已知，为了得到 $\boldsymbol{\eta}$ 的估计，需最小化

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{B}^{\mathrm{T}}(T_i)\boldsymbol{\eta})^2 \tag{4}$$

为了求出式(4) 最小值，对 $\boldsymbol{\eta}$ 求偏导并令其为 0，有

$$\sum_{i=1}^{n}\boldsymbol{B}(T_i)\boldsymbol{B}^{\mathrm{T}}(T_i)\boldsymbol{\eta} = \sum_{i=1}^{n}\boldsymbol{B}(T_i)(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}) \tag{5}$$

收稿日期：2009-03-05.
基金项目：国家自然科学基金资助项目(10871013)；高等学校博士学科点专项科研基金资助项目(20070005003)；北京市自然科学基金资助项目(1072004，1062001).
作者简介：杨宜平(1981—)，女，湖北荆州人，博士生.

令 $\boldsymbol{Y}=(Y_1,\cdots,Y_n)^{\mathrm{T}}$,$\boldsymbol{X}=(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)^{\mathrm{T}}$,$\boldsymbol{B}(T)=(\boldsymbol{B}(T_1),\cdots,\boldsymbol{B}(T_n))^{\mathrm{T}}$,将式(5)代入式(3),有

$$\breve{\boldsymbol{Y}}=\breve{\boldsymbol{X}}\boldsymbol{\beta}+\boldsymbol{\varepsilon}^{*} \tag{6}$$

其中

$$\breve{\boldsymbol{Y}}=(\boldsymbol{I}-\boldsymbol{B}(T)(\boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{B}(T))^{-1}\boldsymbol{B}^{\mathrm{T}}(T))\boldsymbol{Y}$$

$$\breve{\boldsymbol{X}}=(\boldsymbol{I}-\boldsymbol{B}(T)(\boldsymbol{B}^{\mathrm{T}}(T)\boldsymbol{B}(T))^{-1}\boldsymbol{B}^{\mathrm{T}}(T))\boldsymbol{Y}$$

$\boldsymbol{I}$ 是 $n\times n$ 单位矩阵.

模型(1)转化为线性回归模型(6),由于假定 $\boldsymbol{\beta}$ 稀疏,一系列的变量选择方法都可以估计模型(6)中的 $\boldsymbol{\beta}$. 本文采用 Dantzig 或 Lasso 变量选择方法. James 等[14]指出了 Dantzig 和 Lasso 变量选择方法的一些优点:2 种方法对高维的参数模型有较好的经验结果;2 种方法都有有效的算法,Lars 算法用于计算 Lasso,Dasso 算法用于计算 Dantzig.

注意到 2 种方法都需假定设计阵标准化,即协变量矩阵的列范数等于 1. 因此,首先标准化 $\breve{\boldsymbol{X}}$,模型(6)变为

$$\breve{\boldsymbol{Y}}=\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}+\boldsymbol{\varepsilon}^{*} \tag{7}$$

其中,$\tilde{\boldsymbol{\beta}}=\boldsymbol{D}_{\breve{X}}\boldsymbol{\beta}$ 且 $\boldsymbol{D}_{\breve{X}}$ 是对角阵;对角线元素是 $\breve{\boldsymbol{X}}$ 每一列的范数;$\tilde{\boldsymbol{X}}$ 是 $\breve{\boldsymbol{X}}$ 标准化的矩阵,列范数为 1.

考虑模型(7),Lasso 估计 $\hat{\tilde{\boldsymbol{\beta}}}_L$ 为

$$\hat{\tilde{\boldsymbol{\beta}}}_L=\arg\min_{\tilde{\boldsymbol{\beta}}}\frac{1}{2}\|\breve{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\|_2^2+\lambda\|\tilde{\boldsymbol{\beta}}\|_1 \tag{8}$$

其中,$\|\cdot\|_1$ 和 $\|\cdot\|_2$ 分别表示 L_1 和 L_2 范数;$\lambda\geqslant 0$ 是调整参数.

Dantzig 估计 $\hat{\tilde{\boldsymbol{\beta}}}_{\mathrm{DZ}}$ 为

$$\hat{\tilde{\boldsymbol{\beta}}}_{\mathrm{DZ}}=\arg\min_{\tilde{\boldsymbol{\beta}}}\|\tilde{\boldsymbol{\beta}}\|_1 \text{ 且 } |\tilde{\boldsymbol{X}}_j(\breve{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})|\leqslant\lambda,\ j=1,\cdots,p \tag{9}$$

其中,$\tilde{\boldsymbol{X}}_j$ 是 $\tilde{\boldsymbol{X}}$ 的第 j 列;$\lambda\geqslant 0$ 是调整参数.

James 等[14]给出了 Dantzig 选择和 Lasso 等价的充分条件,即

$$u=(\boldsymbol{D}_\lambda\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{D}_\lambda)^{-1}\mathbf{1}\geqslant 0 \text{ 且 } \|\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{D}_\lambda u\|_\infty\leqslant 1 \tag{10}$$

其中,$\boldsymbol{D}_\lambda$ 是对角阵;$\mathbf{1}$ 是全为 1 的向量. 令 $\hat{\tilde{\boldsymbol{\beta}}}_\lambda$ 是用调整参数 λ 得到的 Lasso 估计,如果 $\hat{\tilde{\boldsymbol{\beta}}}_\lambda$ 的第 j 个分量是正数、负数或 0,相应的 $\boldsymbol{D}_\lambda$ 的第 j 个对角线元素是 -1、1 或 0.

由式(8)或式(9)可以得到 $\tilde{\boldsymbol{\beta}}$ 的估计,令 $\hat{\tilde{\boldsymbol{\beta}}}$ 是采用 Dantzig 或 Lasso 方法得到的估计,则 $\hat{\boldsymbol{\beta}}=\boldsymbol{D}_{\breve{X}}^{-1}\hat{\tilde{\boldsymbol{\beta}}}$. 定理 1 将给出 $\hat{\boldsymbol{\beta}}$ 的非渐近界,在定理 1 中,δ、θ、$N_{n,p}$ 都是已知常数,在定理证明中将给出其定义.

定理 1 假设 $\boldsymbol{\beta}$ 是 S 稀疏的参数且 $\delta_{2S}^{\breve{X}}+\theta_{S,2S}^{\breve{X}}<1$,如果式(10)成立且

$$\max\|\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\varepsilon}^{*}\|\leqslant\lambda \tag{11}$$

则

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\leqslant\frac{1}{\sqrt{n}}N_{n,p}\lambda\sqrt{S} \tag{12}$$

注 1 当用 Dantzig 选择计算 $\hat{\boldsymbol{\beta}}$ 时,如果式(10)不成立,仍有定理 1 的结论. 仅用 Lasso 计算 $\hat{\boldsymbol{\beta}}$ 时,才需式(10)成立的条件.

定理 2 假设 $\varepsilon_i\sim N(0,\sigma^2)$,令 $\omega_q=\sup_t|e(t)|$,对任意 $a\geqslant 0$,如果 $\lambda=\sigma\sqrt{2(1+a)\log p}+\omega_q\sqrt{n}$,那么,式(11)成立概率至少 $1-\{p^a\sqrt{4\pi(1+a)\log p}\}^{-1}$,且有

$$\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \| \leqslant \frac{1}{\sqrt{n}} N_{n,p} \sqrt{2S(1+a)\log p} + N_{n,p}\omega_q \sqrt{S} \tag{13}$$

注 2　在适当条件下，随着 n、p、q 的增加，$N_{n,p}$ 收敛到常数，ω_q 退化. 例如，当用分段常数基时，如果 $g'(t)$ 有界，ω_q 以 $1/q$ 的速度收敛到 0；当用分段多项式基时，如果 $g^{d+1}(t)$ 有界，ω_q 以 $1/q^{d+1}$ 的速度收敛到 0 等.

3　模拟研究

为了验证本文的方法，利用 Monte Carlo 方法构造了一个有限样本的模拟. 在数值模拟研究中，考虑如下部分线性模型

$$\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + g(\boldsymbol{T}) + \boldsymbol{\varepsilon}$$

其中，$\boldsymbol{\beta} = (0.25, 0.5, 1.0, 1.5, 0, 0, \cdots, 0)^{\mathrm{T}}$ 是 p 维参数；$\boldsymbol{X}$ 是 p 维正态分布，均值为 1；协方差阵为单位矩阵，$\varepsilon_i \sim N(0, 0.5)$，$i = 1, \cdots, n$；$g(T_i) = \cos(2\pi T_i)$. 在模拟过程中，样本量 $n = 100$，q 取值分别为 10、50、100、120. 本文模拟采用 B 样条函数逼近 $g(T_i)$，样条次数 $m = 3$，节点个数 $k = 3$，节点是 $[0, 1]$ 上 3 等分点. 模拟考虑了 3 种变量选择方法：Dantzig 方法、Lasso 方法和 SCAD 方法，重复计算了 1 000 次，记录下了 3 个指标的值：$\boldsymbol{\beta}$ 非零元素的估计的平均值、4 个非零元素错误设为 0 的平均个数以及 $q-4$ 个非零元素正确设为 0 的平均个数. 模拟结果见表 1.

表 1　变量选择和估计

Table 1　Variable selection and estimators

方法	p	β_1	β_2	β_3	β_4	I	C
SCAD		0.249	0.502	1.003	1.500	0.002	5.281
Dantzig	10	0.248	0.501	1.002	1.500	0.002	5.185
Lasso		0.249	0.502	1.003	1.500	0.002	5.161
SCAD		0.244	0.498	0.999	1.499	0.038	43.515
Dantzig	50	0.243	0.494	0.992	1.496	0.024	39.809
Lasso		0.243	0.494	0.992	1.496	0.030	39.169
SCAD		0.211	0.477	0.974	1.465	0.250	77.348
Dantzig	100	0.214	0.466	0.963	1.467	0.044	87.788
Lasso		0.209	0.459	0.958	1.462	0.056	88.454
SCAD		0.204	0.466	0.933	1.406	0.204	95.531
Dantzig	120	0.211	0.464	0.963	1.464	0.040	101.414
Lasso		0.201	0.455	0.955	1.455	0.043	106.504

注：3 ~ 6 列是 $\boldsymbol{\beta}$ 非零变量的估计值；第 7 列 I 代表 4 个非零变量错误设为 0 的平均个数；第 8 列 C 代表 $q-4$ 个非零变量正确设为 0 的平均个数.

由表 1 的结果可以看出，当 $p < n$ 时，3 种方法都能给出 $\boldsymbol{\beta}$ 的较好估计，SCAD 方法比 Dantzig 和 Lasso 方法更能有效选出系数为 0 对应的协变量，但是，4 个非零元素错误设为 0 的平均个数略高于 Dantzig、Lasso 方法. 当 $p = 100$、120 时，Dantzig 和 Lasso 方法明显优于 SCAD 方法. 因此，在处理 $p \geqslant n$ 的情况时，Dantzig和 Lasso 方法有一定的优势.

4　定理的证明

为了得到本文的结果，首先定义 δ 和 θ，该定义首次被 Candes 等[15] 引入.

定义 1 令 $\boldsymbol{X}$ 是 $n\times p$ 矩阵,将 $\boldsymbol{X}$ 标准化,对给定 $T\subset\{1,\cdots,p\}$,从中提取 T 中元素对应的列向量,构成矩阵 $\boldsymbol{X}_T$,即 $\boldsymbol{X}_T$ 是 $n\times|T|$ 的子矩阵. 那么,对所有子集 T(其中 $|T|\leqslant S$) 和所有长度为 $|T|$ 的向量 $\boldsymbol{c}$,定义 δ_S^X 为使得$(1-\delta_S^X)\ \|\boldsymbol{c}\|_2^2\leqslant\|\boldsymbol{X}_T\boldsymbol{c}\|_2^2\leqslant(1+\delta_S^X)\ \|\boldsymbol{c}\|_2^2$ 成立的最小值.

定义 2 令 T 和 T'是不相关的 2 个集合且 $T,T'\subset\{1,\cdots,p\}$, $|T|\leqslant S,|T'|\leqslant S'$. 那么,如果 $S+S'\leqslant p$,对所有 T 和 T'以及所有对应的向量 $\boldsymbol{c}$ 和 $\boldsymbol{c}'$,定义 $\theta_{S,S'}^X$ 为使得$|(\boldsymbol{X}_T\boldsymbol{c})^{\mathrm{T}}\boldsymbol{X}_{T'}\boldsymbol{c}'|\leqslant\theta_{S,S'}^X\|\boldsymbol{c}\|_2\ \|\boldsymbol{c}'\|_2$ 成立的最小值.

为了证明定理 1,需 James 等[14] 的结果.

引理 1(文献[14]中的定理 4) 假设 $\tilde{\boldsymbol{\beta}}$ 是 S 稀疏的向量且 $\delta_{2S}^{\breve{X}}+\theta_{S,2S}^{\breve{X}}<1$,令 $\hat{\tilde{\boldsymbol{\beta}}}$ 是 Dantzigho 或 Lasso 估计,如果式(10) 和式(11) 成立,则

$$\|\hat{\tilde{\boldsymbol{\beta}}}-\tilde{\boldsymbol{\beta}}\|\leqslant\frac{4\lambda\sqrt{S}}{1-\delta_{2S}^{\breve{X}}-\theta_{S,2S}^{\breve{X}}}$$

定理 1 的证明:

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|=\|\boldsymbol{D}_{\breve{X}}^{-1}(\hat{\tilde{\boldsymbol{\beta}}}-\tilde{\boldsymbol{\beta}})\|\leqslant\|\boldsymbol{D}_{\breve{X}}^{-1}\|\ \|\hat{\tilde{\boldsymbol{\beta}}}-\tilde{\boldsymbol{\beta}}\|=$$

$$\frac{1}{\sqrt{n}}C_{n,p}\|\hat{\tilde{\boldsymbol{\beta}}}-\tilde{\boldsymbol{\beta}}\|\leqslant\frac{1}{\sqrt{n}}\frac{4C_{n,p}\lambda\sqrt{S}}{1-\delta_{2S}^{\breve{X}}-\theta_{S,2S}^{\breve{X}}}=\frac{1}{\sqrt{n}}N_{n,p}\lambda\sqrt{S}$$

其中

$$C_{n,p}=\sqrt{\max_{1\leqslant j\leqslant p}\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\breve{X}_{ij}^2}}$$

$$N_{n,p}=\frac{4C_{n,p}}{1-\delta_{2S}^{\breve{X}}-\theta_{S,2S}^{\breve{X}}}$$

定理 2 的证明:

代入 $\lambda=\sigma\sqrt{2(1+a)\log p}+\omega_q\sqrt{n}$到式(12) ,可得式(13) . 注意到

$$|\hat{\boldsymbol{X}}_j^{\mathrm{T}}\boldsymbol{\varepsilon}^*|=|\hat{\boldsymbol{X}}_j^{\mathrm{T}}\boldsymbol{\varepsilon}+\hat{\boldsymbol{X}}_j^{\mathrm{T}}e(T_j)|\leqslant|\hat{\boldsymbol{X}}_j^{\mathrm{T}}\boldsymbol{\varepsilon}|+|\hat{\boldsymbol{X}}_j^{\mathrm{T}}e(T_j)|\leqslant\sigma|Z_j|+\omega_q\sqrt{n}$$

其中 $Z_j\sim N(0,1)$. 该结果基于事实 $\hat{\boldsymbol{X}}_j$ 范数为 1 且 $\varepsilon_i\sim N(0,\sigma^2)$,则 $\hat{\boldsymbol{X}}_j^{\mathrm{T}}\boldsymbol{\varepsilon}\sim N(0,\sigma^2)$.

因此

$$P(\max_j|\hat{\boldsymbol{X}}_j^{\mathrm{T}}\boldsymbol{\varepsilon}^*|>\lambda)=P(\max_j|\hat{\boldsymbol{X}}_j^{\mathrm{T}}\boldsymbol{\varepsilon}^*|>\sigma\sqrt{2(1+a)\log p}+\omega_q\sqrt{n})\leqslant$$

$$P(\max_j|Z_j|>\sqrt{2(1+a)\log p})\leqslant p\frac{1}{\sqrt{2\pi}}\exp\{-(1+a)\log p\}/\sqrt{2(1+a)\log p}=$$

$$\{p^a\sqrt{4\pi(1+a)\log p}\}^{-1}$$

$P(\max_j|Z_j|>\sqrt{2(1+a)\log p})\leqslant p\frac{1}{\sqrt{2\pi}}\exp\{-(1+a)\log p\}/\sqrt{2(1+a)\log p}$的证明基于事实 $P(\sup_j|Z_j|>u)\leqslant\frac{p}{u}\frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$. 即可证得定理.

参考文献:

[1] SPECKMAN P. Kernel smoothing in partial linear model[J]. J Roy Statist Soc: Ser B, 1988, 50(3): 413-436.

[2] HECKMAN N E. Spline smoothing in a partly linear model[J]. J Roy Statist Soc: Ser B, 1986, 48(2): 244-248.

[3] FAN J Q, LI R Z. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis [J]. J Amer Statist Assoc, 2004, 99(467): 710-723.

[4] CANDES E, TAO T. The Dantzig selector: statistical estimation when p is much larger than n (with discussion) [J]. Ann

Statist，2007，35(6)：2313-2351.
[5] KNIGHT K，FU W. Asymptotics for Lasso-type estimators [J]. Ann Statist，2000，28(6)：1356-1378.
[6] EFRON B，HASTIE T，JOHNSTON I，et al. Least angle regression (with discusion) [J]. Ann Statist，2004，32(2)：407-451.
[7] FAN J Q，LI R Z. Variable selection via nonconcave penalized likelihood and its oracle properties [J]. J Amer Statist Assoc，2001，96(456)：1348-1360.
[8] TIBSHIRANI R. Regression shrinkage and selection via the LASSO [J]. J Roy Statist Soc：Ser B，1996，58(1)：267-288.
[9] JOHNSON B A. Variable selection in semiparametric linear regression with censored data [J]. J Roy Statist Soc：Ser B，2008，70(2)：351-370.
[10] LI R Z，LIANG H. Variable selection in semiparametric regression modeling [J]. Ann Statist，2008，36(1)：261-286.
[11] ZHOU H. The adaptive Lasso and its oracle properties [J]. J Amer Statist Assoc，2006，101(479)：1418-1429.
[12] NI X，ZHANG H，ZHANG D W. Automatic model selection for partially linear models [J]. J Multi Anal，2009，100(9)：2100-2111.
[13] WANG H S，LI R Z，TSAI C L. Tuning parameter selectors for the smoothly clipped absolute deviation method [J]. Biometrika，2007，94：553-568.
[14] JAMES G，RADCHENKO P，LÜ J. DASSO：connections between the dantzig selector and Lasso [J]. J Roy Statist Soc：Ser B，2009，71(1)：127-142.
[15] CANDES E，TAO T. Decoding by linear programming [J]. IEEE Trans Inform Theory，2005，51(22)：4203-4215.

Variable Selection in High-dimensional Partially Linear Models

YANG Yi-ping，XUE Liu-gen，WANG Xue-juan
(College of Applied Sciences，Beijing University of Technology，Beijing 100124，China)

Abstract：This paper considers the problem of variable selection in high-dimensional partially linear models. By combining spline method and Dantzig selector or Lasso，the authors simultaneously select variables and estimate parameters. The simulation results show that the proposed methods are better than the existing method when the dimension of parameters is much larger than the number of observation.

Key words：splines；estimation theory；linear regression；Monte Carlo method

(责任编辑　梁　洁)