Estimation and variable selection for semiparametric additive partial linear models

Key words: BIC; Lasso; penalized likelihood ; regression spline; SCAD
Presenter: Fei Yang
2022/11/03

CONTENT

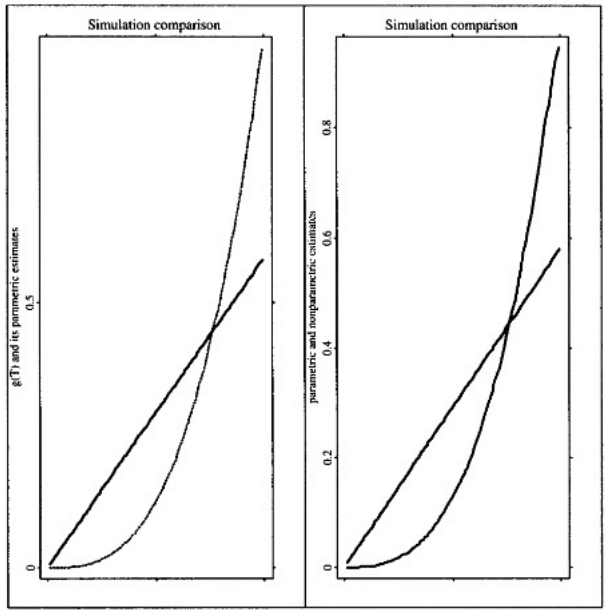
1.研究背景

半参数模型

线性部分　　非线性部分

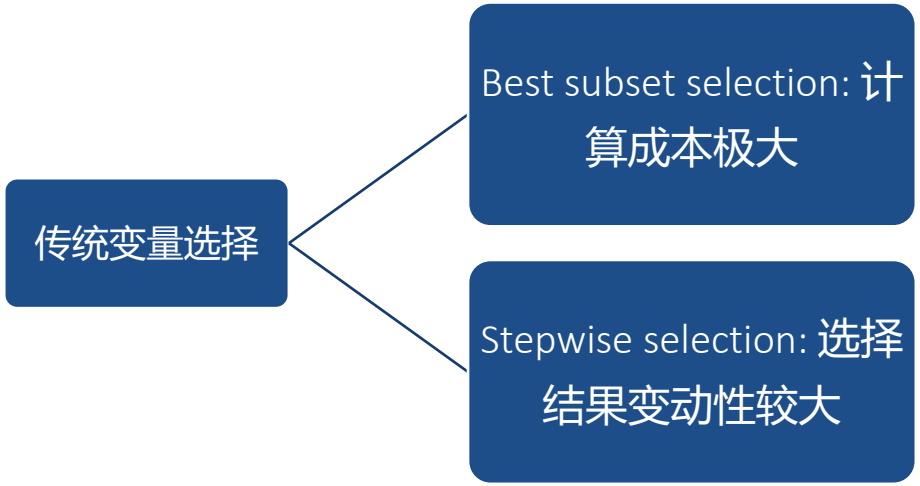
$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=1}^{K} g_k(Z_k) + \varepsilon,$$

- **半参数部分线性可加模型，**
  - 对于n个独立同分布的*随机样本*$\{(X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n)\}$，*其中*
  - $\mathrm{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$是参数部分，$\mathrm{Z} = (Z_1, \ldots, \mathrm{Z_K})^{\mathrm{T}}$是非参数部分，$\mathrm{g}_1, \ldots, g_K$是某未知的光滑函数，$\beta = (\beta_1, \ldots, \beta_{\mathrm{d}})$是线性部分的参数，模型的随机扰动项$\varepsilon$的条件均值为零，方差为$\sigma^2$。此外为了非参数部分的可识别性，需要增加假设$\mathrm{E}\{g_k(Z_k)\} = 0\,, \mathrm{k} = 1, \ldots, K$
- **优点**
  - **降低过拟合风险**
  - **对参数部分的估计更有效**
- **缺点**
  - **上述优点建立在模型被正确指定的基础上**
  - **与广义可加模型相比，多增加一步对于线性部分的识别(structure identification)**



1.研究背景

研究动机



左述方法对于超高维问题不适用

基于SCAD penalty的变量选择

1. Cox models-Fan and Li(2002)
2. Semi-parametric models- Li and Liang(2008)
3. Partially linear models with measurement errors-Li and Liang(2008)-单线性部分

A SCAD-based variable selection procedure for APLM combining the spline approximation

2.估计和变量选择过程

样条近似

- **半参数部分线性可加模型，**
  - 满足特定条件时, 非参数部分可以完全由样条函数逼近，本文中采用三次样条函数cubic splines

$$\mathrm{f}_j(x) \approx \sum_{\mathrm{k}} b_{\mathrm{jk}} B_{\mathrm{jk}}(x)$$

  - 回顾 An Introduction to Statistical Learning： CH7 Moving beyond linearity

Regression Splines 样条回归

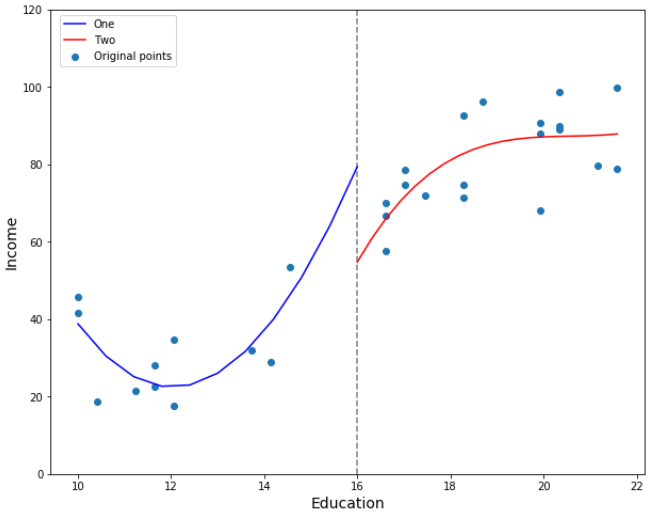
1） Piecewise Polynomials 分段多项式回归

- 基函数,e.g. **结合了多项式和逐段线性回归**的形式，即分段多项式回归

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i,$$

- **动机：**
  - Fitting separate low-degree polynomials over different region of X to avoid fitting a high-degree polynomial over entire range of X 可理解为具有变系数的多项式回归
- **结点：**
  - 改变点称为knots，K个结点, 即K+1个方程
  - 自由度: $(d+1)\times(K+1)$
  - 实际举例：税法改革使相关支出的回归线会在法案生效时点发生变化
- **缺点：**受异常点影响较大，需要加上额外的边界约束。
  - 举例：一个三次多项式分段回归如下，但该回归不满足“平滑特性”,在结点处可能出现断裂或跳跃（如右图）。

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$



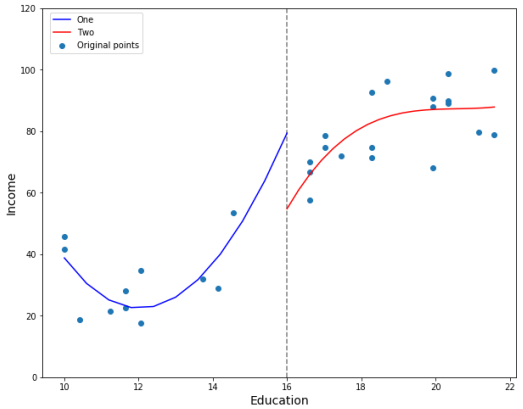
Regression Splines 样条回归

2） Constraints and Splines带约束的分段多项式回归=样条回归

- 样条是一个函数，由多项式构造的分段函数，并且在分段节点处要具有高度平滑的特性，即在分段结点处连续的导数。

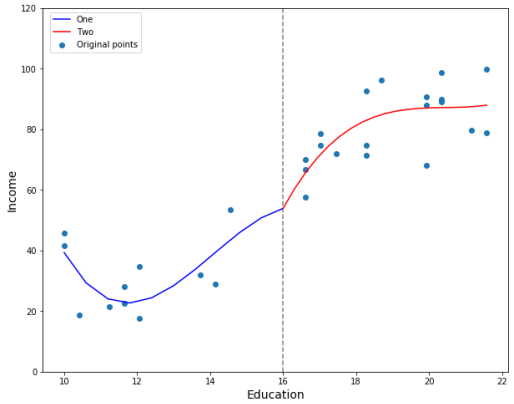
- **性质：**
  - **样条函数具有连续性和光滑性；（**在结点处的函数值、斜率上相同**）**

三次多项式分段回归



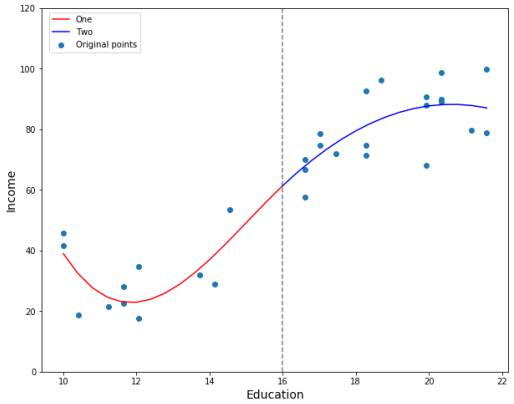
- 无约束

连续的三次多项式分段回归



- 结点处连续
- $f(C_-) = f(C_+)$

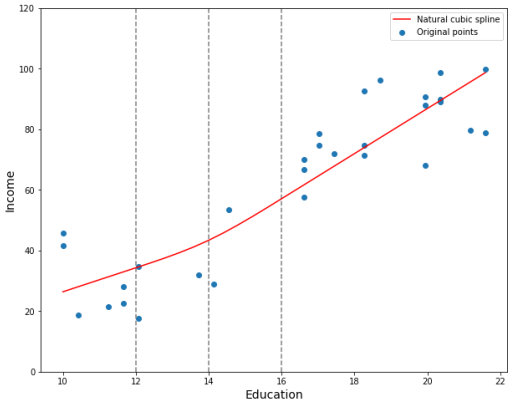
三次样条回归



- 结点处连续且光滑

(一阶导与二阶导在节点处连续)

自然三次样条回归



- 结点处连续且光滑
- 边界线性-超出末端节点之外的拟合是线性的

自由度$(d+1)\times(K+1)$

由于每个节点上有d个约束（从0到$d-1$阶导数相等），最终自由度是总自由度减去总约束度：$(d+1)\times(K+1)-K\times d = K+d+1$

Smoothing Splines 平滑样条

- 避免多项式样条估计的节点选择问题对光滑程度造成过多主观性影响，我们采用正则化手段，在自然三次样条中，引入光滑参数λ对拟合的粗糙度变化进行惩罚。即假设我们样条函数为g(x), 寻找一个光滑函数使得残差平方和最小

- Goal:
  - Find some function g(x) that makes RSS small: $\boldsymbol{RSS} = \sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{g(x_i)})^2$
  - Guarantee g(x) is also smooth.
    - Find the function g that minimizes: (加入光滑因子后的均方误差)

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

how g fit the data well　　penalizes the variability in g

    - Where λ is a nonnegative tuning parameter.(bias-variance tradeoff)
      - λ =0, penalty term no effect, function g will be jumpy,
      - $\lambda \to \infty$ g will be smooth,
      - 超参λ ，用来衡量惩罚项的重要性占比，一般用n折交叉验证或留一交叉验证法来确定。
- ”Effective Degrees of Freedom”
  - The number of free parameters is an inappropriate measure of model complexity due to λ

- 一阶导：该点曲线的斜率；
- 二阶导 ：斜率的变化率，即，惩罚项即表示为函数曲线在该点的平滑性。
- 二阶导的积分为：对区间t内，二阶导数累积的变化情况，因此可以用来衡量该段区间整体的平滑性。

2.估计和变量选择过程 基于惩罚函数的变量选择

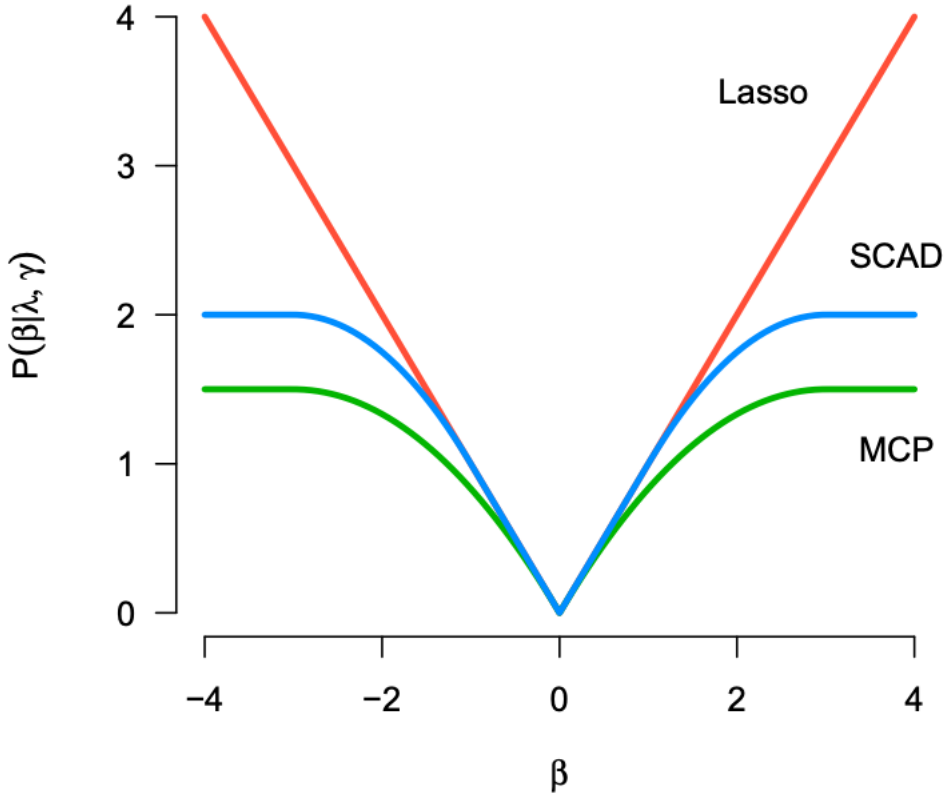
- 对模型复杂度和估计准确度的trade-off

方法	2012penGAM	2011APLMs
应用模型	非参数-广义可加模型	半参数-部分线性可加模型
模型表达式	$\mathrm{Y}_i = c + \sum_{i=1}^{p} f_j\left(x_i^{(j)}\right) + \varepsilon_i$	$Y = \mathrm{X}^{\mathrm{T}}\beta + \sum_{k=1}^{K} g_k(Z_k) + \varepsilon,$
惩罚项	Group LASSO：组间选择 基于对各组内特征对应系数的L2范数，达到组级别上的稀疏性	SCAD：单变量选择-对称非凹函数 优点：可以处理奇异矩阵；连续进行压缩,无偏估计 缺点：速度慢; 高噪下表现不佳
惩罚项表达式	$\mathrm{J}(f_j) = \lambda_1 \sqrt{\|f_j\|_n^2 + \lambda_2 \int (f_j''(x))^2 dx}$	$\mathrm{P}(\beta) = \begin{cases} \lambda\|\beta\|, if\|\beta\| \le \lambda \\ \dfrac{2a\lambda\|\beta\| - \beta^2 - \lambda^2}{2(a-1)}, if\lambda < \|\beta\| \le a\lambda \\ \dfrac{\lambda^2(a+1)}{2}, otherwise \end{cases}$
求解目标函数	$\mathcal{L}(\lambda_1, \lambda_2) = \left\| Y - \sum_{\mathrm{j}=1}^{p} f_j \right\|_n^2 + \sum_{j=1}^{p} J(f_j)$	$\mathcal{L}(\beta, \gamma) = \frac{1}{2}\sum_{i}^{n} [Y_{\mathrm{i}} - \{\gamma^T \mathrm{b}(Z_{\mathrm{i}}) + X_i^{\mathrm{T}}\beta\}]^2 + n\sum_{j=1}^{d} p_{\lambda_j}(\|\beta_j\|)$

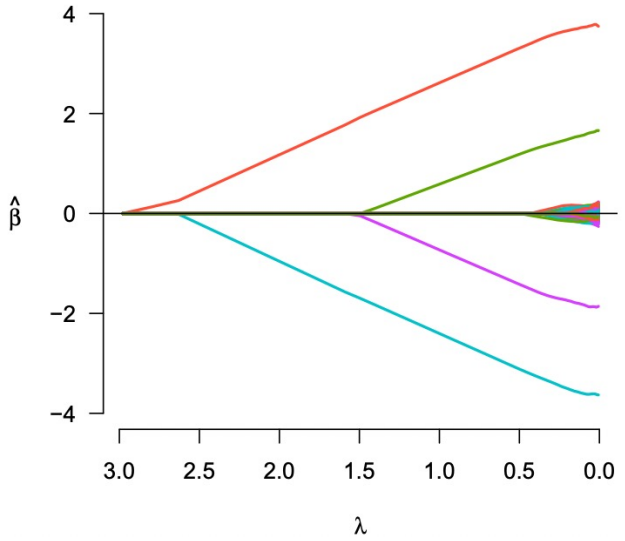
2.估计和变量选择过程

基于惩罚函数的变量选择

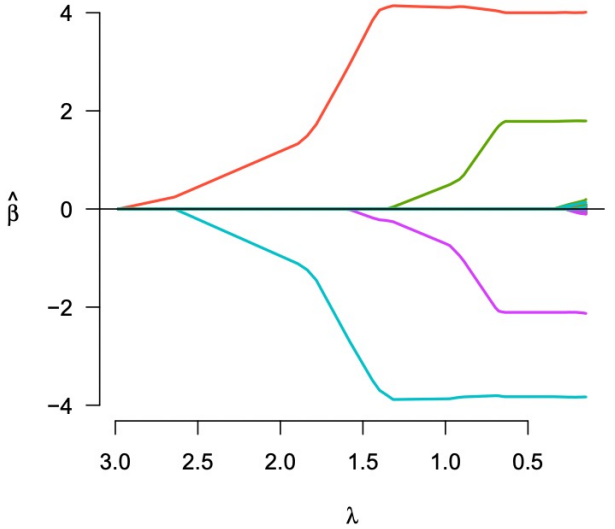
- 对模型复杂度和估计准确度的trade-off



Solution path-LASSO



Solution path-SCAD($\lambda = 4$)



2.估计和变量选择过程

基于惩罚函数的变量选择

- SCAD-一类软阈值算子
- Fan & Li, 2001 JASA, Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties

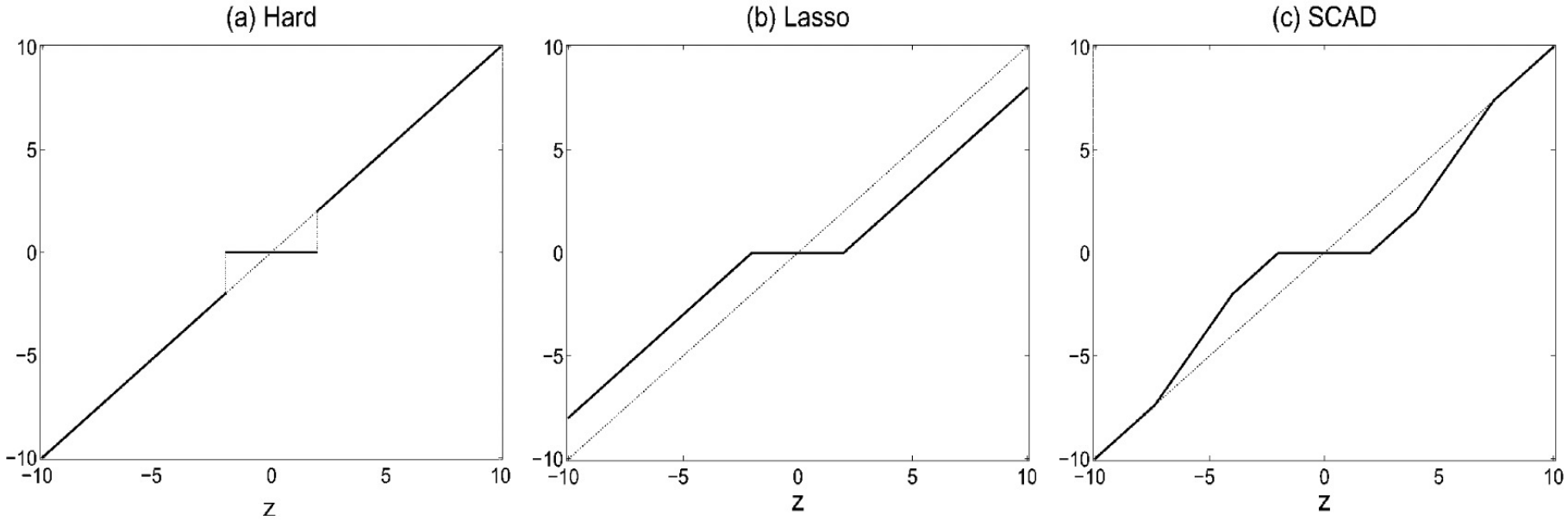


Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.

2.估计和变量选择过程

基于惩罚函数的变量选择

- 基于惩罚的目标函数的统一框架

$$\mathcal{L}_P(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{b}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\}\right]^2 + n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \qquad (2.5)$$

惩罚项名称	惩罚项表达式
L-0 penalty	$p_{\lambda_j}(\lvert\beta_j\rvert) = 0.5\lambda_j^2 I\{\lvert\beta_j\rvert \neq 0\}$
Extend to a general case	
	$p_{\lambda_j}(\lvert\beta_j\rvert) = \frac{n}{2}\sum_{j=1}^{d}\lambda_j^2 I\{\lvert\beta_j\rvert \neq 0\}$
AIC	$\lambda_j = \sqrt{2/n}\sigma$
BIC	$\lambda_j = \sqrt{\log(n)/n}\sigma$
RIC	$\lambda_j = \sqrt{\log(d)/n}\sigma$
L_q $penalty$Bridge regression	$p_{\lambda_j}(\lvert\beta_j\rvert) = q^{-1}\lambda\lvert\beta_j\rvert^q$

3.理论分析

关于估计量的性质

- 线性部分的估计量 $\hat{\beta}$是$\beta_0\sqrt{n}$一致估计，$\hat{\beta}$- β_0=$O_p\left(\frac{1}{\sqrt{n}}\right)$
- 非线性部分的估计量g趋近于真实值g_0的收敛速率小于$\sqrt{n}$（引理A.4）

Theorem 1. *Under the conditions* (C1)−(C5) *given in the Appendix,* $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ *converges to* $N(\mathbf{0}, \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1})$ *in distribution, where* $\mathbf{D} = E(\widetilde{\boldsymbol{X}}^{\otimes 2})$ *and* $\boldsymbol{\Sigma} = E(\varepsilon^2\widetilde{\boldsymbol{X}}^{\otimes 2})$. *Furthermore, if* ε *and* $(\boldsymbol{X}, \boldsymbol{Z})$ *are independent,* $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \to N\left(0, \sigma^2\mathbf{D}^{-1}\right)$, *where* $\sigma^2 = E\left(\varepsilon^2\right)$.

3.理论分析

关于变量筛选方法的性质

- 基于SCAD惩罚的变量筛选步骤可以有效识别真实重要变量

Theorem 2. *Suppose that* $a_n = O(n^{-1/2})$, $b_n \to 0$, *and* (C1)−(C5) *in the Appendix hold. Then* (I) *With probability approaching one, there exists a local minimizer* $\widehat{\boldsymbol{\beta}}$ *of* $\mathcal{L}_P(\boldsymbol{\beta}, \boldsymbol{\gamma})$ *such that* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$. (II) *If* $\lambda_j \to 0$, $n^{1/2}\lambda_j \to \infty$, *and*

$$\liminf_{n\to\infty} \liminf_{u\to 0^+} \frac{p'_{\lambda_j}(u)}{\lambda_j} > 0, \tag{2.7}$$

then, with probability approaching one, the root-n consistent estimator $\widehat{\boldsymbol{\beta}}$ *in* (I) *satisfies* (a) $\widehat{\boldsymbol{\beta}}_2 = 0$, *and* (b) $\widehat{\boldsymbol{\beta}}_1$ *has an asymptotic normal distribution*

$$\sqrt{n}\{E(\widetilde{\boldsymbol{X}}_1^{\otimes 2}) + \Sigma_\lambda\}[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \{E(\widetilde{\boldsymbol{X}}_1^{\otimes 2}) + \boldsymbol{\Sigma}_\lambda\}^{-1}\kappa_n] \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_s),$$

where $\boldsymbol{\Sigma}_s = \mathrm{Var}\left(\varepsilon\widetilde{\boldsymbol{X}}_1\right)$.

4.数值模拟

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

模型设定

- $\beta = (3,1.5,0,0,2,0,0,0)^T$ ，$\sigma = 1,3,5$ ，X与ε独立,且X具有自相关结构X_i与X_j的自相关系数为$\rho^{|i-j|}$, $\rho = 0.5$
- $g_1(z) = 5\sin(4\pi z)$, $g_2(z) = 100\{\exp(-3.25z) - 4\exp(-6.5z) + 3\exp(-9.75z)\}$,
- (i) $g(Z) = g_1(Z_1); (ii)g(Z) = g_2(Z_2); (iii)g(Z) = g_1(Z_1) + g_2(Z_2)$ 这里Z_1, Z_2 相互独立并满足[0,1]内的均匀分布
- 100次重复实验, n=60, 100, 200

对比方法

- SCAD，LASSO，BIC指标进行变量筛选
  - 其中SCAD和LASSO方法局部二次近似(Fan and Li(2001))和通过GCV进行协调参数的选取

比较指标

- C：线性部分中，5个真实零系数被压缩至0的个数（平均值）
- I：线性部分中，3个真实非零系数被错误压缩至0的个数（平均值）
- MRME： 模型误差的相对中位数 (筛选后模型v.s. 全模型)

4.数值模拟

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

Table 1. Simulation Results for Case (i)

n	method	$\sigma = 1$			$\sigma = 3$			$\sigma = 5$		
		C	I	MRME	C	I	MRME	C	I	MRME
60	scad	4.49	0.00	0.852	4.39	0.12	0.899	4.29	0.61	0.903
	lasso	3.38	0.00	0.882	3.49	0.02	0.750	3.41	0.31	0.723
	bic	4.66	0.00	0.869	4.76	0.14	0.948	4.57	0.86	0.969
	oracle	5.00	0.00	0.662	5.00	0.00	0.680	5.00	0.00	0.635
100	scad	4.45	0.00	0.838	4.44	0.03	0.870	4.35	0.32	0.947
	lasso	3.31	0.00	0.906	3.53	0.00	0.775	3.40	0.09	0.768
	bic	4.84	0.00	0.876	4.80	0.03	0.880	4.77	0.60	1.036
	oracle	5.00	0.00	0.717	5.00	0.00	0.704	5.00	0.00	0.706
200	scad	4.40	0.00	0.798	4.37	0.00	0.818	4.38	0.03	0.788
	lasso	3.27	0.00	0.884	3.37	0.00	0.829	3.37	0.00	0.797
	bic	4.91	0.00	0.803	4.88	0.00	0.916	4.90	0.06	0.772
	oracle	5.00	0.00	0.723	5.00	0.00	0.668	5.00	0.00	0.693

结果与SCAD(2001)类似，零元素正确识别方面：BIC > SCAD > LASSO；非零元素正确识别方面：LASSO > SCAD >BIC

4.数值模拟

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

Table 2. Simulation Results for Case (ii)

n	method	$\sigma = 1$			$\sigma = 3$			$\sigma = 5$		
		C	I	MRME	C	I	MRME	C	I	MRME
60	scad	4.44	0.00	0.774	4.48	0.14	0.937	4.32	0.69	1.028
	lasso	3.28	0.00	1.017	3.41	0.02	1.003	3.47	0.35	0.889
	bic	4.60	0.00	0.792	4.74	0.19	0.983	4.58	0.88	1.058
	oracle	5.00	0.00	0.673	5.00	0.00	0.674	5.00	0.00	0.662
100	scad	4.49	0.00	0.784	4.46	0.03	0.874	4.47	0.38	1.017
	lasso	3.58	0.00	1.044	3.50	0.00	0.996	3.58	0.11	0.963
	bic	4.85	0.00	0.784	4.76	0.03	0.907	4.78	0.61	1.041
	oracle	5.00	0.00	0.747	5.00	0.00	0.655	5.00	0.00	0.681
200	scad	4.40	0.00	0.768	4.31	0.00	0.805	4.31	0.03	0.870
	lasso	3.29	0.00	1.006	3.38	0.00	0.983	3.36	0.00	0.910
	bic	4.89	0.00	0.767	4.89	0.01	0.839	4.84	0.08	0.954
	oracle	5.00	0.00	0.716	5.00	0.00	0.644	5.00	0.00	0.677

从模型误差角度，SCAD在样本量较大以及误差项方差较小时的表现较好，LASSO在样本量较小和误差方差较大时表现好

4.数值模拟

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

Table 3. Simulation Results for Case (iii)

n	method	$\sigma = 1$			$\sigma = 3$			$\sigma = 5$		
		C	I	MRME	C	I	MRME	C	I	MRME
60	scad	4.43	0.00	0.924	4.39	0.28	1.045	4.37	0.74	1.010
	lasso	3.52	0.00	1.072	3.68	0.10	0.922	3.67	0.26	0.783
	bic	4.32	0.00	0.939	4.42	0.31	1.077	4.43	0.87	1.012
	oracle	5.00	0.00	0.802	5.00	0.00	0.802	5.00	0.00	0.752
100	scad	4.41	0.00	0.926	4.49	0.02	0.957	4.28	0.35	1.052
	lasso	3.58	0.00	1.028	3.58	0.00	0.964	3.56	0.09	0.883
	bic	4.60	0.00	0.939	4.77	0.05	0.977	4.66	0.65	1.112
	oracle	5.00	0.00	0.800	5.00	0.00	0.782	5.00	0.00	0.784
200	scad	4.44	0.00	0.881	4.45	0.00	0.953	4.45	0.06	0.973
	lasso	3.42	0.00	1.022	3.48	0.00	0.995	3.41	0.01	0.891
	bic	4.84	0.00	0.900	4.79	0.00	0.988	4.86	0.11	1.021
	oracle	5.00	0.00	0.821	5.00	0.00	0.806	5.00	0.00	0.797

总体而言，SCAD和BIC在模拟中的表现接近，SCAD通过多选变量以减小预测误差，同时计算效率更快

5.实证分析 营养流行病学调查

- **研究问题：探究个人特点和饮食因素是如何影响血浆中的beta胡萝卜素的浓度**
- **数据集：**一个营养流行病学数据集（Nierenberg et al.(1989)）
  - http://lib.stat.cmu.edu/datasets/Plasma_Retinol

  - 先前调查表明，膳食胡萝卜素以及女性是和beta胡萝卜素水平正相关，而吸烟、BMI与其负相关。年龄因素并不显著
  - 然而这些研究产生了一些相左的结论，并且仅依赖于普通的方差分析或线性模型来进行。
    - **线性模型：**BMI，FIBER，GENDER，SMOKE3为统计意义上显著的变量。
    - **广义可加模型：**由于beta胡萝卜素水平取log后与AGE，CHOL呈现非线性关系，因此采用GAM拟合，检验确认
      - **线性变量：**BMI，CALORIES，FAT，FIBER，ALCOHOL，BETADIET
      - **非线性变量：**AGE，CHOL
    - **部分线性可加模型**：检验全部变量中哪些应当被选入模型

$$\begin{aligned}\log(\text{beta-carotene}) = {} & \beta_0 + \beta_1 \text{BMI} + \beta_2 \text{CALORIES} + \beta_3 \text{FAT} + \beta_4 \text{FIBER} \\ & + \beta_5 \text{BETADIET} + \beta_6 \text{GENDER} + \beta_7 \text{ALCOHOL} + \beta_8 \text{SMOKE2} \\ & + \beta_9 \text{SMOKE3} + g_1(\text{AGE}) + g_2(\text{CHOL}) + \varepsilon.\end{aligned}$$

变量名称	变量含义
BETA-CAROTENE	Beta胡萝卜素（因变量）
AGE	年龄
GENDER	性别
BMI	身体质量指数
CALORIES	每日摄入卡路里
FAT	每日摄入脂肪
FIBER	每日摄入纤维素
ALCOHOL	每周摄入酒精
CHOL	每日摄入胆固醇
BETADIET	每日生日膳食胡萝卜素
SMOKE2	1-曾经吸烟 0-从未吸烟
SMOKE3	1-当前吸烟 0-从未吸烟

5.实证分析

营养流行病学调查

- **参数选取：**
  - **1.非线性部分确定cubic B-splines的结点个数，**测试结点数量2-9时给出最小相对均方误的结点：
  - **2.SCAD，LASSO内的参数：**广义交叉验证GCV选取
- **结果：**

Best knots number

	SCAD	LASSO
AGE	2	5
CHOL	2	2

Table 4. Results for the nutritional study. Left panel: Estimated values, associated standard error, and P-value by using the ordinary least squares. Right panel: Estimates, associated standard errors of the coefficients using the APLM with the proposed variable selection procedures.

	LS				APLM		
	Est.	s.e	z value	$Pr(>\|z\|)$	SCAD (s.e.)	LASSO (s.e.)	BIC (s.e.)
BMI	-0.976	0.202	-4.829	$< 10^{-4}$	-0.947(0.189)	-0.948(0.173)	-1.001(0.188)
CALORIES	0	0	-0.457	0.648	0(0)	0(0)	0(0)
FAT	-0.002	0.003	-0.711	0.477	0(0)	-0.001(0.001)	0(0)
FIBER	0.027	0.012	2.352	0.019	0.021(0.007)	0.019(0.007)	0.025(0.008)
BETADIET	0.137	0.073	1.889	0.060	0.046(0.027)	0.101(0.051)	0(0)
GENDER	0.277	0.135	2.060	0.040	0.194(0.088)	0.201(0.096)	0(0)
ALCOHOL	0.043	0.048	0.901	0.368	0(0)	0(0)	0(0)
SMOKE2	-0.068	0.091	-0.742	0.458	0(0)	0(0)	0(0)
SMOKE3	-0.286	0.130	-2.191	0.029	-0.245(0.097)	-0.224(0.096)	-0.293(0.117)
AGE	0.005	0.003	1.724	0.086			
CHOL	-0.015	0.114	-0.133	0.894			

5.实证分析

营养流行病学调查

- **结果：**

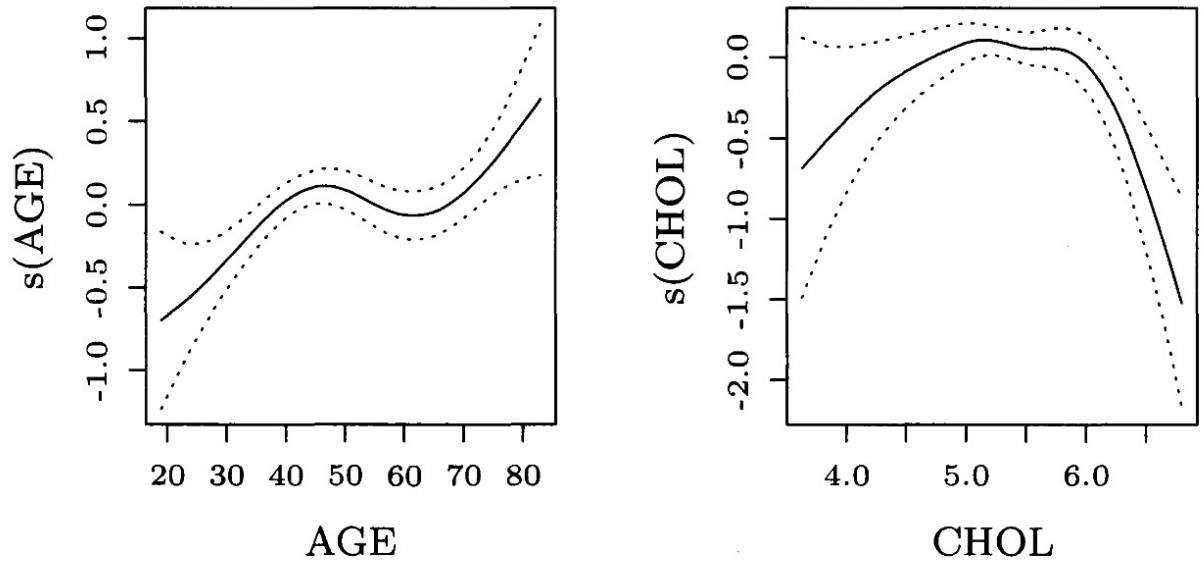


Figure 1. The patterns of AGE and CHOL with $\pm s.e.$ using the R function, gam, for the dataset from a nutritional study.

变量AGE和CHOL的确呈现明显的非线性趋势，验证了部分线性可加模型在此数据集上的适用性。

6.讨论

方法上

- 性能：两阶段惩罚可以实现变量选择、模型估计
- 优势：
  - (1) 避免迭代运算，缩短运算时间
  - (2) 线性部分的估计量在一定条件下是渐近正态的

模拟上

不足：线性部分的表现通过对元素是否真实非零来评估，全模型拟合通过拟合的相对误差来比较
但是在真实模型构造中缺少非线性部分的干扰变量

拓展上

此类方法可以与Sparse additive model(Ravikumar et al. 2009))结合，但需要建立起方法的渐近性质。

Extended Topics: 如何区别非参数/参数部分

- Double penalization based procedure: (Lian, et al. 2013)
- **核心思想：**
  - **惩罚项1: 识别零元素**
  - **惩罚项2: 识别参数元素**
- **来源：**
  - Elastic net(Zou and Hastie(2005)), fused lasso(Tibshirani et al.(2005)), sparse group lass(Peng et al.(2010))**等**
- **亮点：**
  - **构造了一个**regularized oracle estimator, **可以直接对特征的二阶导数 压缩至零**

$$(\hat{\mu}, \hat{b}) = \arg\min_{\mu,b} \frac{1}{2} \sum_i \left\{ Y_i - \mu - \sum_{j=1}^{p} \sum_{k=1}^{K} b_{jk} B_{jk}(X_{ij}) \right\}^2$$

$$+ n\lambda_1 \sum_{j=1}^{p} w_{1j} ||b_j||_{A_j} + n\lambda_2 \sum_{j=1}^{p} w_{2j} ||b_j||_{D_j},$$

ω_{1j}**权重越大,零元素越多;**
ω_{2j}**权重越大,线性部分越多;**

$$\|b_j\|_{A_j} = \|\sum_k b_{jk} B_{jk}\| \qquad \|b_j\|_{D_j} = \|\sum_k b_{jk} B''_{jk}\|$$

对非零元素的惩罚项　　对线性元素的惩罚项

Extended Topics: 如何区别非参数/参数部分

$$Z_j = \begin{pmatrix} B_{j1}(X_{1j}) & B_{j2}(X_{1j}) & \cdots & B_{jK}(X_{1j}) \\ \vdots & \vdots & \ddots & \vdots \\ B_{j1}(X_{nj}) & B_{j2}(X_{nj}) & \cdots & B_{jK}(X_{nj}) \end{pmatrix}_{n \times K},$$

$$\min_b \frac{1}{2}||Y - Zb||^2 + n\lambda_1 \sum_{j=1}^{p} w_{1j}||b_j||_{A_j} + n\lambda_2 \sum_{j=1}^{p} w_{2j}||b_j||_{D_j}.$$

- **参数选取**
  - **样条阶数**spline order: q=4(cubic splines)
  - **基函数个数**number of basis: K =6 (**策略是选择较大的**K, **以降低拟合误差，通过惩罚方式避免常见的过拟合问题**)
  - λ_1, λ_2, **通过**eBIC**选取**(Chen and Chen (Biometrika, 2008))，**可证明该方法选择的**λ_1, λ_2, **可以正确识别非零元素和线性元素**

Extended Topics: 如何区别非参数/参数部分

$$\min_b \frac{1}{2}||Y - Zb||^2 + n\lambda_1 \sum_{j=1}^{p} w_{1j}||b_j||_{A_j} + n\lambda_2 \sum_{j=1}^{p} w_{2j}||b_j||_{D_j}.$$

- **求解过程：**
  - **局部二次近似**LQA ，Fan and Li(2001)：**惩罚项可以由在初始点的泰勒展开来近似（省略无相关项后，目标函数关于**b**是二次的，并且可以得到解析解）**

$$\|b_j\|_{A_j} \approx \|b_j^{(0)}\|_{A_j} + \frac{1}{2}\frac{\|b_j\|_{A_j}^2 - \|b_j^{(0)}\|_{A_j}^2}{\|b_j^{(0)}\|_{A_j}},$$

$$\|b_j\|_{D_j} \approx \|b_j^{(0)}\|_{D_j} + \frac{1}{2}\frac{\|b_j\|_{D_j}^2 - \|b_j^{(0)}\|_{D_j}^2}{\|b_j^{(0)}\|_{D_j}}.$$

Extended Topics: 如何区别非参数/参数部分

- **数值模拟**

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i,$$

$f_1(x) = 5\sin(2\pi x)$ $f_2(x) = 10x(1-x)$ $f_3(x) = 3x$

$f_4(x) = 2x$ $f_5(x) = -2x$ $f_j(x) = 0, j > 5$

$\mathrm{Cov}(X_{ij_1}, X_{i2}) = 0.5^{|j_1 - j_2|}$

  - N=50,100,200
  - P=50,100,200
  - $\sigma = 0.2, 0.5$

- **对比方法**
  - BIC-**单惩罚项** non-adaptive lasso estimator
  - EBIC-**单惩罚项** non-adaptive lasso estimator
  - BIC/BIC-**双惩罚项**-adaptive group lasso estimator
  - EBIC/EBIC-**双惩罚项**-adaptive group lasso estimator
  - BIC/EBIC-**双惩罚项**-adaptive group lasso estimator

The BIC proposed by Schwarz(1978) selects the model that minimizes, Where $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta(s)$ and $\boldsymbol{\nu}(s)$ is the number of components in s

$$\mathrm{BIC}(s) = -2\log L_n\{\hat{\theta}(s)\} + \nu(s)\log n,$$

$$\mathrm{BIC}_\gamma(s) = -2\log L_n\{\hat{\theta}(s)\} + \nu(s)\log n + 2\gamma\log\tau(\mathcal{S}_j), \quad 0 \leqslant \gamma \leqslant 1,$$

(eBIC)

EBIC-**单惩罚项**

$$\log(\frac{1}{n}\|Y - Z\hat{b}_\lambda\|^2) + d_1\frac{\log(n/K)}{n/K} + \frac{d_1}{n/K}\log p.$$

EBIC-**双惩罚项**

$$\log(\frac{1}{n}\|Y - Z\hat{b}_\lambda\|^2) + d_1\frac{\log(n/K)}{n/K} + d_2\frac{\log n}{n} + \frac{d_1K + d_2}{n}\log p,$$

BIC倾向选择更多非零元素(false positive)，eBIC由于加强惩罚可能会漏选非零元素(false negative)

Extended Topics: 如何区别非参数/参数部分

- **评价指标1-变量筛选结果**
- **说明**
  - #N: **选中的非线性元素个数**
  - #NT: #N **中真实为非线性的个数**
  - #L: **选中的线性元素个数**
  - #LT：#L**中真实为线性的个数**
  - **（小字为标准差）**
- **结果**
  - 单惩罚项-较难识别线性元素
  - BIC-惩罚力度较小，假阳率较高
  - eBIC-惩罚力度较大，假阴率较高

Table 1: 变量筛选结果(n=100)

		#N	#NT	#L	#LT
$n = 100$	BIC	$32.86_{15.1563}$	5_{0}	0_{0}	0_{0}
$p = 50$	EBIC	$6.22_{2.8521}$	$4.36_{1.3667}$	0_{0}	0_{0}
σ=0.2	BIC/BIC	$2.6_{0.8571}$	2_{0}	$2.74_{1.2747}$	$2.4_{0.8571}$
	EBIC/EBIC	$1.94_{0.5115}$	$1.84_{0.3703}$	$2.48_{1.1110}$	$2.42_{1.0515}$
	BIC/EBIC	$2.06_{0.2399}$	2_{0}	$3.06_{0.5500}$	$2.94_{0.2399}$
$n = 100$	BIC	$43.46_{1.5281}$	5_{0}	0_{0}	0_{0}
$p = 50$	EBIC	$3.74_{2.2840}$	$3.3_{1.7871}$	0_{0}	0_{0}
$\sigma = 0.5$	BIC/BIC	$12.64_{12.5727}$	2_{0}	$2.8_{1.7261}$	$1.44_{1.3273}$
	EBIC/EBIC	$1.6_{0.5714}$	$1.56_{0.5014}$	$1.7_{1.3132}$	$1.7_{1.3132}$
	BIC/EBIC	$2.42_{0.5380}$	2_{0}	$3.32_{1.0583}$	$2.64_{0.4849}$
$n = 100$	BIC	$25.7_{19.1644}$	$4.9_{0.5803}$	0_{0}	0_{0}
$p = 100$	EBIC	$4.92_{3.0159}$	$3.78_{1.6817}$	0_{0}	0_{0}
$\sigma = 0.2$	BIC/BIC	$2.98_{1.4497}$	2_{0}	$2.68_{1.7076}$	$2.16_{1.1314}$
	EBIC/EBIC	$1.76_{0.5175}$	$1.72_{0.4536}$	$2_{1.3093}$	$2_{1.3093}$
	BIC/EBIC	$2.04_{0.2828}$	$1.98_{0.1414}$	$3.04_{0.7548}$	$2.86_{0.4953}$
$n = 100$	BIC	$25.2_{24.3788}$	$4.8_{0.8081}$	0_{0}	0_{0}
$p = 100$	EBIC	$3.84_{2.6447}$	$3.16_{1.6826}$	0_{0}	0_{0}
$\sigma = 0.5$	BIC/BIC	$4.2_{6.8512}$	$1.94_{0.2399}$	$2.62_{1.3536}$	$2.2_{1.1429}$
	EBIC/EBIC	$1.68_{0.7677}$	$1.52_{0.5047}$	$1.44_{1.2316}$	$1.42_{1.1968}$
	BIC/EBIC	$2.26_{0.5997}$	$1.96_{0.1979}$	$2.76_{0.8704}$	$2.58_{0.7309}$
$n = 100$	BIC	$13.42_{11.8754}$	$4.6_{1.1066}$	0_{0}	0_{0}
$p = 200$	EBIC	$3.8_{3.0034}$	$2.86_{1.7958}$	0_{0}	0_{0}
$\sigma = 0.2$	BIC/BIC	$2.88_{1.0230}$	2_{0}	$2.36_{1.4107}$	$2.1_{1.0738}$
	EBIC/EBIC	$1.58_{0.6728}$	$1.48_{0.5047}$	$1.3_{1.4178}$	$1.24_{1.3180}$
	BIC/EBIC	$2.18_{0.6289}$	$1.92_{0.2740}$	$2.52_{1.0349}$	$2.4_{0.9258}$
$n = 100$	BIC	$9.48_{9.1724}$	$4.14_{1.4709}$	0_{0}	0_{0}
$p = 200$	EBIC	$2.42_{1.7507}$	$2.1_{1.1995}$	0_{0}	0_{0}

因此，如果第一个估计量使用BIC，第二估计量使用eBIC，便可以修正问题

Extended Topics: 如何区别非参数/参数部分

- **评价指标2-模型拟合结果**
  - RMSE

$$RMSE_j = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\hat{f}_j(t_i) - f_j(t_i))^2},$$

- **说明**
  - Oracle: **真实模型**
  - Sparse Additive: **当$\lambda_2 = 0$时的模型（没有区分线性部分）**
- **结果**
  - **（1）与SA相比，对于非线性部分的估计都较为接近；然而线性部分差异较大，文章提出的估计在RMSE上的表现提升约30%-50%**
  - **（2）估计结果的标准误比SA估计更小**

Table2: 根据前6项拟合结果的RMSE（n=100）

		Oracle	Our Estimator	Sparse Additive
$n = 100$	f_1	$0.3286_{0.01716}$	$0.3301_{0.01622}$	$0.3301_{0.01883}$
$p = 50$	f_2	$0.0761_{0.02542}$	$0.1184_{0.04923}$	$0.0883_{0.03043}$
$\sigma = 0.2$	f_3	$0.0319_{0.02216}$	$0.0361_{0.02643}$	$0.0800_{0.02745}$
	f_4	$0.0366_{0.02271}$	$0.0481_{0.03229}$	$0.0941_{0.04142}$
	f_5	$0.0361_{0.02702}$	$0.0432_{0.03751}$	$0.0929_{0.04113}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
$n = 100$	f_1	$0.3364_{0.01925}$	$0.3468_{0.03131}$	$0.3420_{0.02067}$
$p = 50$	f_2	$0.1186_{0.04531}$	$0.1753_{0.08853}$	$0.1541_{0.06361}$
$\sigma = 0.5$	f_3	$0.0527_{0.03669}$	$0.0645_{0.04389}$	$0.1601_{0.07305}$
	f_4	$0.0494_{0.04048}$	$0.0707_{0.05324}$	$0.1812_{0.06942}$
	f_5	$0.0463_{0.03850}$	$0.0634_{0.05495}$	$0.1733_{0.08265}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$

因此，如果第一个估计量使用BIC，第二估计量使用eBIC，便可以修正问题

Summary

	Single penalization	Double penalization
性能	进行变量选择、模型估计	同时实现变量选择、线性部分的识别、模型估计
参数	通过广义交叉验证选取SCAD内的超参数	结合BIC，eBIC选取两个惩罚项的参数
`优势	（1）避免迭代运算，缩短运算时间 （2）线性部分的估计量在一定条件下是渐近正态的	（1）自主识别线性部分 （2）理论上该方法以概率一可以讲线性/ 非线性部分完美分离
`劣势	线性部分识别依赖于先验信息或预诊断信息	（1）求解过程使用的LQA方法无法获得精确零解
模拟	线性部分的表现通过对元素是否真实非零来评估，全模型拟合通过拟合的相对误差来比较 但是在真实模型构造中缺少非线性部分的干扰变量	（1）n.P关系并未足够模拟超高维情况； （2）对于自相关特征, 随机扰动项不同分布等复杂情况并未探讨
拓展	此类方法可以与Sparse additive model(Ravikumar et al. 2009))结合，但需要建立起方法的渐近性质。	（1）无法得到精确零解，在其他单惩罚项的文献中可以选择坐标下降类型方法代替；双惩罚项是个挑战 （2）拓展到quantile regression等

Reference

[1] Kazemi et al. (2019). A sure independence screening procedure for ultra-high dimensional partially linear additive models

[2] Lian, Heng, Liang, et al. SEPARATION OF COVARIATES INTO NONPARAMETRIC AND PARAMETRIC PARTS IN HIGH-DIMENSIONAL PARTIALLY LINEAR ADDITIVE MODELS.

[3] Lian et al. (2012). Identification of Partially Linear Structure in Additive Models with an Application to Gene Expression Prediction from Sequences

[4]Kazemi et al. (2018). Variable selection and structure identification for ultrahigh-dimensional partially linear additive models with application to cardiomyopathy microarray data

[5]杨晶. 若干半参数模型的稳健推断与模型选择方法[D]. 重庆大学, 2016.

感谢聆听！请大家批评指正!

THANK YOU FOR YOUR CRITICISM

presenter: Fei Yang 2022/11/03

LASSO问题的交替方向乘子法

考虑凸优化问题 $\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\|Ax-b\|_2^{\ 2}+\tau\|x\|_1$ **(1)**

其中$A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m, \tau>0$是正则化参数.

(1)的等价表述

$$\underset{x\in\mathbb{R}^n, y\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\|Ax-b\|_2^{\ 2}+\tau\|y\|_1 \quad \textbf{(2)}$$
$$\text{subject to}\quad x-y=0$$

$$L_\eta(x,y,\lambda)=\frac{1}{2}\|Ax-b\|_2^{\ 2}+\tau\|y\|_1+\lambda^T(x-y)+\frac{\eta}{2}\|x-y\|_2^{\ 2}$$

$$x_{t+1}=(A^TA+\eta I)^{-1}(A^Tb+\eta(y_t-u_t))$$

$y_{t+1}=S_{\tau/\eta}(x_{t+1}+u_t)$ ，其中$S_\rho(\cdot)$是软阈值算子

$u_{t+1}=u_t+(x_{t+1}-y_{t+1})$ ，其中$u=\lambda/\eta$

软阈值算子

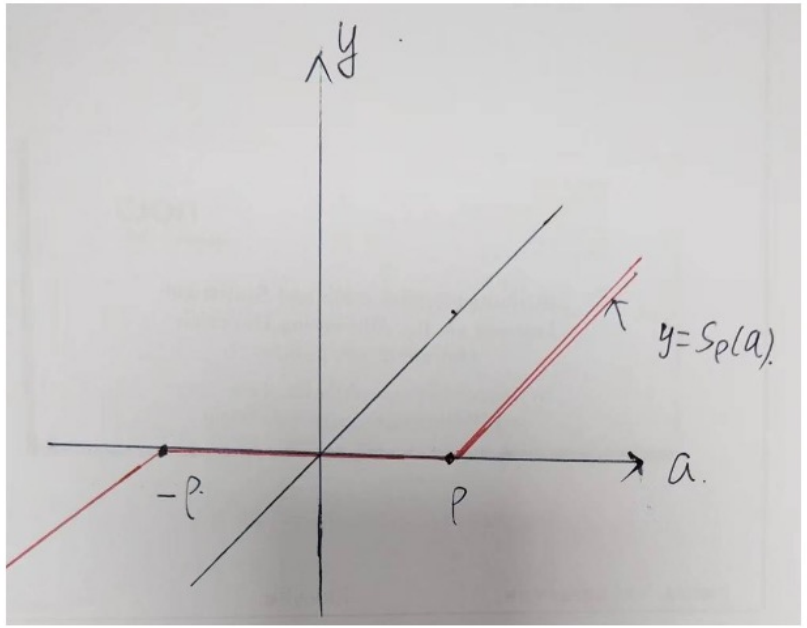
给定正数$\rho > 0$，

- 软阈值算子$S_\rho: \mathbb{R} \to \mathbb{R}$将已知的$a$映射到如下优化问题的解：

$$\min_{t\in\mathbb{R}} \frac{1}{2}(t-a)^2 + \rho|t|$$

- 显式表达式为

$$S_\rho(a) = \begin{cases} a-\rho & a > \rho \\ 0 & |a| \le \rho \\ a+\rho & a < -\rho \end{cases}$$



设$f(t) = |t|$.

t_*是解当且仅当

$$0 \in t_* - a + \rho\partial f(0)$$

$$\Updownarrow$$

$$t_* > 0 \Rightarrow t_* - a + \rho = 0,$$

$$t_* = 0 \Rightarrow t_* - a \in [-\rho, \rho],$$

$$t_* < 0 \Rightarrow t_* - a - \rho = 0.$$