中国科学 : 数学　2010 年　第 40 卷　第 4 期 : 393 ∼ 402
www.springerlink.com　math.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

含发散维数自变量的单指标模型中方向向量的稳健估计

朱利平[①*], 朱力行[②]

① 华东师范大学金融与统计学院, 上海 200241;
② 香港浸会大学数学系, 香港
E-mail: lpzhu@stat.ecnu.edu.cn, lzhu@hkbu.edu.hk

收稿日期: 2008-10-21; 接受日期: 2009-08-07; * 通信作者
国家自然科学基金 (批准号: 10701035) 和上海市教育发展基金会晨光计划 (批准号: 2007CG33) 资助项目

摘要　在回归分析中, 常常引入大量的自变量来减少模型拟合的误差. 本文考虑如下非常一般的单指标模型: 在给定自变量 X 的线性组合 $\beta_0^\tau X$ 的条件下, 响应变量 Y 和维数发散的自变量 X 相互独立, 其中 β_0 是 p_n 维向量. 本文在这样的单指标模型假设下讨论当 $p_n \to \infty$ 时单指标模型中方向向量的稳健估计问题. 我们发现, 当 $p_n = o(\sqrt{n})$ 时, 最小二乘估计 β_{n0} 能够相合地估计 β_0 的方向. 为了剔除不相关的自变量, 从而提高回归模型的可解释能力, 我们提出基于 ℓ_1-正则化的算法, 通过加二次限制得到稀疏的最小二乘估计. ℓ_1-正则化的解 β_n 不仅可以相合地估计 β_0 的方向, 而且可以产生稀疏的估计. 因此, 我们可以选择一些重要的自变量, 在保持预测准确度的同时使模型解释变得容易. 模拟分析和汽车价格数据应用分析表明, 我们所提出的估计方法在有限样本场合具有良好的表现.

关键词　ℓ_1-正则化　发散维数　逆回归　限制正交性　稀疏性　充分降维
MSC (2000) 主题分类　62H15, 62G20

1　引言

回归分析被广泛地应用于研究响应变量 Y 和 p_n 维自变量 $X = (X_1, \ldots, X_{p_n})^\tau$ 之间的关系, 这里 "τ" 指代转置运算. 伴随着科技的进步, 以及对诸如网页文档数据, 基因表达数据和金融历史数据等各种形式数据采集的巨额投入, 这样的数据往往有两个特征: 大样本容量和大维数自变量. 典型的例子是 Framingham 心脏病研究 [1], 这个研究中样本容量约为 $n = 25000$, 维数 $p_n = 100$. Donoho[1] 建议考虑自变量的维数满足 $p_n = O(n^{1/3})$ 甚至 $p_n = O(n^{1/2})$, 也就是说, 当 $n \to \infty$ 时, 自变量的维数 $p_n \to \infty$. 显然, 这种情形下数据可视化将会很困难.

回归分析中, 有两种方法可以减少自变量的维数. 一是选择部分与因变量高度相关的自变量子集. 这是因为, 自变量中往往包含许多不相关的变量, 当自变量的维数很高时 [2], 参数估计的精度和相应模型预测的准确度将会大大降低 [3]. 于是, 在建立模型之前将不相关的变量排除是很关键的. 另一个方法是, 在不损失回归信息和不要求预先假定的参数模型的条件下, 将原有的高维变量 X 替换为少量的线性组合, 这种方法在充分降维中已经被广泛研究过. 详细的回顾可参见 [4].

通过假设条件均值 $E(Y|X)$ 具有参数结构, Fan 和 Li[5] 作了许多奠基性的工作. 与逐步回归方法

英文版见: Zhu L P, Zhu L X. Stable direction recovery in single-index models with a diverging number of predictors. Sci China Math, 2010, 53, doi: 10.1007/s11425-010-0028-4

等文献中其他计算密集的方法不一样的是, 当自变量维数固定时, 他们通过非凹惩罚似然提出了一个统一的方法, 可以自动选择变量且同时估计系数. Fan 和 Peng[6] 将这个创新的工作进一步发展成一般的方法, 考虑了自变量维数发散的情形. 其他的最小二乘方法, 如 Tibishirani[7] 提出的 Lasso 也可以在这个框架下讨论.

通过假定条件均值 $E(Y|X)$ 具有半参数的结构, Naik 和 Tsai[8] 以及 Kong 和 Xia[9] 仔细考虑了单指标模型

$$Y \perp\!\!\!\perp E(Y|X)|\beta^\tau X \tag{1}$$

的变量选择问题, 其中 "$\perp\!\!\!\perp$" 是指条件独立性, β 是 $p_n \times 1$ 向量. 本质上, 这个模型等价于 $E(Y|X)$ 是 $\beta^\tau X$ 的函数 [10]. 通过极小化真实模型和候选模型之间的 Kullback-Leibler 距离, Naik 和 Tsai[8] 推导出了一个新的用于单指标模型的 AIC_C 准则. Kong 和 Xia[9] 提出了可分交叉核实法, 它是一个迭代算法, 因此计算量比较大.

本文讨论 β 的方向估计问题. 我们不假定 $Y|X$ 的条件分布具有某种参数形式, 只是简单地假设 Y 关于 X 的回归信息可以全部由自变量的线性组合 $\beta^\tau X$ 来描述. 也就是说,

$$Y \perp\!\!\!\perp X|\beta^\tau X. \tag{2}$$

显然, 模型 (2) 并不限于条件均值 $E(Y|X)$. 譬如, 它也可应用于具有常数均值和单指标方差函数的模型, $\mathrm{Var}(Y|X) = \mathrm{Var}(Y|\beta^\tau X)$. 我们研究的半参数模型 (2) 介于参数模型和非参数模型中间. 一方面, 可以减轻因错误确定连接函数带来的风险, 而且比参数模型假设更加灵活; 另一方面, 可以避免非参数模型拟合中的维数祸根问题. 鉴于这些优点, 许多统计学和计量经济学的文献都讨论了单指标模型 (2) 中 β 的方向的估计问题. 例如, 当 p_n 固定时, Ni, Cook 和 Tsai[11], Li 和 Nachtshein[12], Li 和 Yin[13], 以及 Li[14] 提供了一些算法, 可以用来同时估计 β 的方向, 并剔除不相关的自变量.

本文考查单指标模型 (2) 在 $p_n \to \infty$ 时 β 的方向的稳健估计问题, 这里仅仅假设在给定 $\beta^\tau X$ 时, 响应变量 Y 和预测变量 X 独立, 其中 β 是 $p_n \times 1$ 向量.

在第二节, 我们将会指出, 在线性条件下,

$$E(X|\beta^\tau X) = \Sigma_{XX}\beta[\beta^\tau \Sigma_{XX}\beta]^{-1}\beta^\tau X, \tag{3}$$

有

$$\beta_0 := \Sigma_{XX}^{-1}\Sigma_{XY} = \kappa \cdot \beta, \tag{4}$$

其中, κ 是某个未知常数, $\Sigma_{XX} = E[X - E(X)]X^\tau$, $\Sigma_{XY} = E[(X - E(X))F(Y)]$, $F(y)$ 是连续响应变量 Y 的边际分布函数. 如果存在样本, $\{(x_i, y_i), i = 1, \ldots, n\}$, 我们能够用样本形式估计 β_0,

$$\beta_{n0} := \widehat{\Sigma}_{XX}^{-1}\widehat{\Sigma}_{XY}, \tag{5}$$

其中, $\widehat{\Sigma}_{XX} = \frac{1}{n}\boldsymbol{X}^\tau\boldsymbol{X}$, $\widehat{\Sigma}_{XY} = \frac{1}{n}\boldsymbol{X}^\tau F_n(\boldsymbol{Y})$, $\boldsymbol{X} = (x_1 - \overline{x}, \ldots, x_n - \overline{x})^\tau$, $\overline{x} = n^{-1}\sum_{i=1}^n x_i$, $F_n(\boldsymbol{Y}) = [F_n(y_1), \ldots, F_n(y_n)]^\tau$, $F_n(y) = n^{-1}\sum_{i=1}^n 1_{\{y_i \leqslant y\}}$. 在很弱的充分条件下, 我们依概率有 $\|\beta_{n0} - \beta_0\|_{\ell_2} = O(p_n/\sqrt{n})$.

在第三节, 为了寻找 β_0 的方向的稀疏估计, 我们建议使用 ℓ_1 正则化算法. 为了记号方便, 定义 $\|\alpha\|_{\ell_1} = \sum_{i=1}^{p_n}|\alpha_i|$, $\|\alpha\|_{\ell_2} = [\sum_{i=1}^{p_n}\alpha_i^2]^{1/2}$, $\alpha = (\alpha_1, \ldots, \alpha_{p_n})^\tau$. 我们记 $\tau_n = C \cdot \|F_n(\boldsymbol{Y}) - \boldsymbol{X}\beta_{n0}\|_{\ell_2}$, C 是

某个大的常数. 考虑估计 β_0 的凸规划搜索,

$$\begin{aligned} &\min_{\beta\in\Omega} \quad \|\beta\|_{\ell_1}, \\ &\text{s.t.} \quad \|F_n(\boldsymbol{Y}) - \boldsymbol{X}\beta\|_{\ell_2} \leqslant \tau_n, \end{aligned} \tag{6}$$

其中, Ω 是 $\mathbb{R}^{p_n}$ 的紧子集. 在这一节, 我们指出 (6) 的解 β_n 识别 β_0, 并且稀疏误差至多是 τ_n 的倍数. 在 ℓ_1 正则限制下, β_n 不仅能够相合地识别 β_0, 而且能够得到充分稀疏的估计, 从而可以选择重要的变量, 在保持预测精度的同时便于模型解释.

2 基于分布的最小二乘估计

为了便于说明, 我们不失一般性地假设自变量都已经被中心化了, 即 $E(X) = 0$, 并且 $\mathrm{Cov}(X) = \Sigma_{XX}$. 我们定义最小二乘估计的总体形式为

$$\beta_0 := \arg\min_{\beta\in\Omega} E[F(Y) - X^\tau\beta]^2 = \Sigma_{XX}^{-1}\Sigma_{XY}. \tag{7}$$

Zhu 和 Zhu[15] 指出, 在最小二乘准则 (7) 下, 最小值 β_0 能够识别 β 的方向, 下面的引理陈述了这个结论.

引理 2.1 假设 β 满足条件独立性 (2). 如果 Y 是连续变量, 那么由线性条件 (3) 可得到 (4).

条件独立性模型 (2) 涵盖了许多回归模型, 如 $Y \perp\!\!\!\perp E(Y|X)|\beta^\tau X$[10], 其中单指标 $\beta^\tau X$ 包含在条件均值中, $[Y - E(Y|X)] \perp\!\!\!\perp \mathrm{Var}(Y|X)|\beta^\tau X$[16], 这里单指标包含在条件方差内. 在充分降维领域常常假定线性条件成立, 见 [17] 和 [4, 命题 4.2, 57 页]. Hall 和 Li[18] 指出, 如果 $p_n \to \infty$, 线性条件在模型 (2) 中总是近似成立的. 关于这个条件的讨论也可参见 [19]. 关于 Y 的连续性假设在这不是本质的. 如果 Y 是离散的, 我们可以把 $F(y)$ 替换成 $F(y-) = \mathrm{Prob}(Y < y)$, 依然会有相同的结论成立.

如果我们有样本 $\{(x_i, y_i), i = 1, \ldots, n\}$, 我们可以将 (7) 中的未知量替换成相应的估计值从而得到估计形式. 具体来说, $R_n(\beta) = n^{-1}\sum_{i=1}^n [F_n(y_i) - x_i^\tau\beta]^2$, 定义

$$\beta_{n0} := \arg\min_{\beta\in\Omega} R_n(\beta) = \widehat{\Sigma}_{XX}^{-1}\widehat{\Sigma}_{XY}. \tag{8}$$

下面的定理说明了 β_{n0} 可以依概率收敛到 β_0.

定理 2.1 假设 $\lambda_{\min}\{\Sigma_{XX}\} \geqslant C_1 > 0$, $\max_{1\leqslant i\leqslant p_n} E|X_i|^4 < C_2 < \infty$, 对 p_n 一致地成立, 并且 $p_n = o(\sqrt{n})$ 关于 n 单调增加. 设 $\Omega \subseteq \mathbb{R}^{p_n}$ 是紧集. 那么依概率有

$$\|\beta_{n0} - \beta_0\|_{\ell_2} = O(p_n/\sqrt{n}). \tag{9}$$

定理指出, 如果 $p_n = o(\sqrt{n})$, 我们可以得到一个相合估计. 这个速度要比 $p_n = o(n^{1/3})$ 快, 后者是在 Fan 和 Peng[6] 讨论非凹惩罚似然估计以及 Huber[20] 讨论 M 估计时的速度.

3 ℓ_1 正则化下的变量选择

我们常常引入大量的自变量以减少模型拟合的偏差. 但是, 在高维场合, 我们很有可能引进了很多不相关的变量 [8], 如果引入了太多的不重要的自变量则会影响到参数估计的精度和模型预测的准确

度 [3]. 为此, 从单指标模型中将无关紧要的自变量剔除往往是很重要的. 在这一节, 我们考虑 $\beta \in \mathbb{R}^{p_n}$ 的方向的稀疏估计. 也就是说, 我们假设方向 β 的支撑 $T_0 = \{i : \beta_i \neq 0\}$ 具有很小的势. 另外, 我们假设观测的自变量矩阵 $\boldsymbol{X}$ 服从一致不确定原则, 即 $\boldsymbol{X}$ 遵循 "限制的等容假设". 为了引进这个概念, 设 $n \times |T|$ 子矩阵 $\boldsymbol{X}_T$ 和 $T \subset \{1, \ldots, n\}$ 是从 $\boldsymbol{X}$ 提取 T 指标集中相应的列得到的. Candès, Romberg 和 Tao[21] 定义 $\boldsymbol{X}$ 的 S-限制等容常量 δ_S 为最小的满足下面条件的值:

$$(1 - \delta_S)\|c\|_{\ell_2}^2 \leqslant \|\boldsymbol{X}_T c\|_{\ell_2}^2 \leqslant (1 + \delta_S)\|c\|_{\ell_2}^2 \tag{10}$$

对所有的子集 $|T| \leqslant S$ 以及系数 $(c_j)_{j \in T}$. 这一性质本质上要求, 对于势小于 S 的 $\boldsymbol{X}$ 的每个列子集, 它们的表现近似于一个标准正交系统.

在一致不确定原则下, 遵循 Candès, Romberg 和 Tao[21] 的思想, 我们考虑凸规划问题 (6). 下面的定理说明, (6) 得到的估计 β_n 的误差最多和 τ_n 成比例.

定理 3.1 除假设定理 2.1 的条件成立外, 假设 S 满足 $\delta_{3S} + 3\delta_{4S} < 2$. 如果 β_0 的支撑 $|T_0| \leqslant S$, 那么 (6) 的解 β_n 满足

$$\|\beta_n - \beta_0\|_{\ell_2} \leqslant C_S \cdot \tau_n, \tag{11}$$

其中常数 C_S 可以依赖于 δ_{4S}.

平稳识别的条件包含限制的等容常量. 我们感兴趣的是服从一致不确定原则的具有好的等容常量的矩阵. 利用随机矩阵理论的工具, Candès, Romberg 和 Tao[21] 给出了一些矩阵, 如元素是独立同分布的正态变量的随机矩阵 $\boldsymbol{X}$, Fourier 和一般正交测度族等等, 这些矩阵都满足条件 (10).

和 [21] 不一样的是, 在 ℓ_1-正则化问题 (6) 中, 我们使用基于数据的界. 值得注意的是, 估计 β_n 的误差项的大小和最小二乘成比例, 而且是我们能够得到的最好的结果.

4 随机模拟

本节考虑模拟. 我们将从两方面评估估计的效果. 第一, 我们关注单指标模型中方向向量的估计 $\hat{\beta}$ (如 β_{n0} 或者其它估计) 的准确度. 为了比较, 我们采用绝对相关系数 $|\mathrm{Corr}(\beta^\tau X, \hat{\beta}^\tau X)|$ 作为准则. 这个准则曾在 Li, Cook 和 Tsai[22] 中使用过. 我们重复试验 1000 次, 然后给出这些绝对相关系数的均值和标准差. 第二, 我们需要比较各个估计量在选择重要变量上的表现. 因为 β_n 能够将某些回归系数估计为零, 我们期望 β_n 能够产生稀疏和准确的 β 的估计, 从而有助于选择正确的变量. 为了度量准确性, 我们运用生物医学中常用的两种准则: 真阳性率 (TPR), 即重要变量中系数被估为非零的比例; 以及假阳性率 (FPR), 即非重要变量中系数被估为非零的比例. 我们希望 TPR 接近 1, 同时 FPR 接近于 0.

我们将把本文提出的估计方法和其他 5 个估计方法作比较. 这些估计是: 本文提出的最小二乘估计 β_{n0}, 切片数为 5 的切片逆回归估计 β_{sir}[17], 改进的 MAVE 估计 β_{rmave}[23], 稀疏 MAVE 估计 β_{smave}[24] 以及可分交叉核实估计 β_{scv}[9]. 因为前两个竞争者不能产生稀疏估计, 我们仅仅给出其绝对相关系数的均值以及标准差.

我们使用 Candès, Romberg 和 Tao[21] 提出的相同的算法, 区别只是在于如何选择 (6) 中 τ_n 的调整参数 C. 关于算法的细节, 可以参考 [21]. 另外, 这个算法的 Matlab 程序可以从 http://www.acm.caltech.edu/l1magic/下载. 显然, 大家所熟知的交叉核实法 (GCV) 和 Ni, Cook 和 Tsai[11], Li 和 Yin[13]

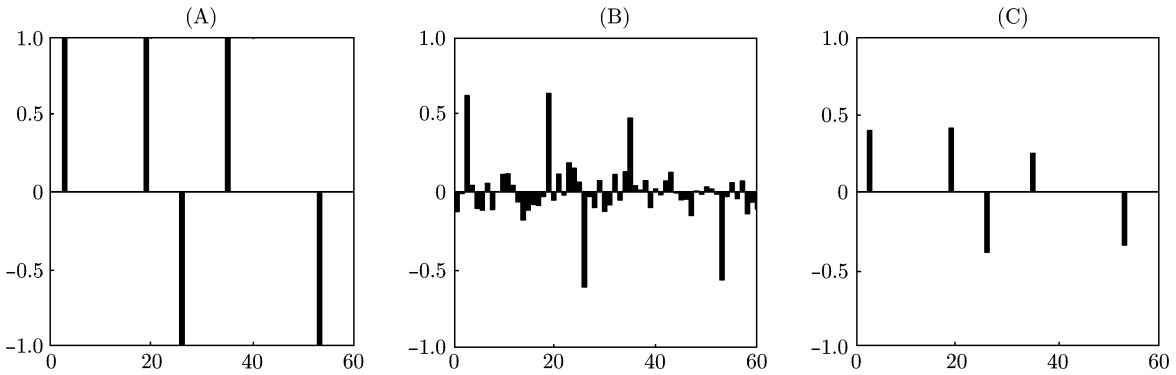

图 1 (A) 原始向量 β; (B) 基于分布的最小二乘估计 β_{n0}; (C) 稀疏估计 β_n

提出的信息准则在这里可以直接使用. 但是, 我们发现提出的估计对 C 取 $1.5 \leqslant C \leqslant 2.5$ 并不敏感. 为此, 简单起见, 在整个模拟中我们取 $C = 1.6$.

例 1 Tobit 模型在计量经济学领域被广泛使用. 模型中部分响应变量可能是随机删失的. 例如, Tobin[25] 最初提出的线性回归模型考虑响应变量是消费耐用品支出, 因为收入为负值时耐用消费品支出可能是不可观测的, 所以响应变量会出现删失. 之后, 线性 Tobit 模型被 Li, Simonoff 和 Tsai[26] 推广至非线性模型. 模拟中, 我们选择下面两个模型:

$$Y = (\beta^\tau X + \varepsilon) \cdot \mathbf{1}_{(\beta^\tau X + \varepsilon > 0)}, \tag{12}$$

$$Y = \exp(\beta^\tau X + \varepsilon) \cdot \mathbf{1}_{(\beta^\tau X + \varepsilon > 0)}, \tag{13}$$

其中 $\mathbf{1}_{(\cdot)}$ 是示性函数. 显然, 模型 (13) 是非线性的. 自变量 $X = (X_1, \ldots, X_{p_n})^\tau \sim N(0, I_{p_n})$, ε 是均值为零, 标准差为 0.05 的正态随机变量. 样本容量 $n = 100$, 维数分别取为 $p_n = 20$, 40 和 60.

在这个例子中, 我们取 $|T_0| = 5$, 即有 5 个预测变量与响应变量相关, 从而其他 $p_n - 5$ 个变量与响应变量无关. 模拟时, 我们从 β 的 p_n 个分量中随机地选取 5 个非零变量. 然后使得非零变量以等概率取值 1 或者 -1. 关于这一点, 我们可以通过下面的图形来说明: 原始向量 β, 基于分布的最小二乘估计 β_{n0} 和稀疏估计 β_n, 可分别见图 1 中 (A), (B) 和 (C).

因为 β_{n0} 和 β_n 只是和 β 成比例, 所以相应估计的数值大小可能不同于真实模型中的 β. 但是, 我们在这里只关心 β 的方向, 从图 1 可以看出 β_{n0} 和 β_n 在估计 β 的方向上都不错, 而且, β_n 是一个稀疏估计.

表 1 是基于 1000 次重复将 β 的稀疏估计制图而成的. 可以看出, 稀疏 MAVE 线性 Tobit 模型 (12) 中表现最好. 但是, 在非线性模型 Tobit 模型 (13) 中, β_{n0} 是最好的. 在 p_n 相对小的时候, 改进的

表 1 模型 (12) 和 (13) 中绝对相关系数的均值和标准差

		绝对相关系数				
	$n = 100$	β_{n0}	β_n	β_{sir}	β_{rmave}	β_{smave}
(12)	$p_n = 20$	0.9657±0.0116	0.9538±0.0366	0.9296±0.0236	0.9747±0.0093	0.9972±0.0018
	$p_n = 40$	0.9341±0.0154	0.9696±0.0241	0.8614±0.0285	0.9354±0.0213	0.9974±0.0015
	$p_n = 60$	0.9045±0.0177	0.9810±0.0138	0.8034±0.0355	0.8413±0.0541	0.9964±0.0029
(13)	$p_n = 20$	0.9660±0.0109	0.9517±0.0437	0.9283±0.0234	0.9472±0.0163	0.9416±0.1501
	$p_n = 40$	0.9333±0.0151	0.9695±0.0243	0.8627±0.0269	0.8800±0.0314	0.8711±0.2294
	$p_n = 60$	0.9043±0.0173	0.9814±0.0128	0.7995±0.0308	0.7949±0.0529	0.7907±0.2559

表 2　模型 (12) 和 (13) 中基于 1000 次重复的平均 TPR 和 FPR 值

		β_n		β_{smave}		β_{scv}	
	$n=100$	TPR	FPR	TPR	FPR	TPR	FPR
(12)	$p_n=20$	0.6336	0.0000	1.0000	0.1907	0.9980	0.0000
	$p_n=40$	0.9804	0.0017	1.0000	0.1274	1.0000	0.0100
	$p_n=60$	1.0000	0.0107	1.0000	0.0742	0.9980	0.0000
(13)	$p_n=20$	0.6436	0.0000	0.9880	0.6733	1.0000	0.0047
	$p_n=40$	0.9806	0.0028	0.9760	0.6366	1.0000	0.0034
	$p_n=60$	1.0000	0.0117	0.9600	0.5524	1.0000	0.3536

MAVE 方法和 SIR 方法 p_n 表现得不错, 但是, 在 p_n 很大时表现欠佳.

表 2 总结了估计 β_n, β_{smave} 和 β_{scv} 的基于 1000 次重复的平均 TPR 和 FPR 值. 可以看出, 如果维数 p_n 小, 那么基于 TPR, β_{smave} 和 β_{scv} 的表现要比其他估计好得多. 但是, 它们的 FPR 值很大, 尤其是对 $p_n=60$ 的非线性 Tobit 模型更为如此. 另一方面, 当 $|T_0|/p_n$ 越来越小时, β_n 的 TPR 值接近于 1, 同时 FPR 值接近于 0, 这是与我们的期望相符合的.

例 2　我们进一步考虑如下的 4 个模型:

$$Y=\sin(\beta^\tau X)+\varepsilon, \tag{14}$$

$$Y=\sin(\beta^\tau X+\varepsilon), \tag{15}$$

$$Y=\exp(\beta^\tau X)+\varepsilon, \tag{16}$$

$$Y=\exp(\beta^\tau X+\varepsilon), \tag{17}$$

其中 X, ε 和 β 的产生机制和上个例子相同. 样本容量和维数分别是 $n=100$, $p_n=60$. 显然, 模型 (14) 和 (15) 中的连接不是单调的, 模型 (16) 和 (17) 中方向 β 并不在条件均值里面. 模拟结果放在表 3 和表 4.

表 3　模型 (14) 至 (17) 中绝对相关系数的均值和标准差

	绝对相关系数				
	β_{n0}	β_n	β_{sir}	β_{rmave}	β_{smave}
model (14)	0.9700±0.0079	0.9966±0.0026	0.8960±0.0223	0.9683±0.0063	0.2638±0.1666
model (15)	0.9698±0.0082	0.9966±0.0026	0.8933±0.0272	0.9697±0.0060	0.2733±0.1488
model (16)	0.9694±0.0081	0.9965±0.0026	0.8948±0.0229	0.9695±0.0062	0.8725±0.2156
model (17)	0.9697±0.0082	0.9965±0.0027	0.8988±0.0214	0.9686±0.0064	0.8726±0.2259

表 4　模型 (14) 至 (17) 中基于 1000 次重复的平均 TPR 和 FPR 值

	β_n		β_{smave}		β_{scv}	
	TPR	FPR	TPR	FPR	TPR	FPR
model (14)	1.0000	0.0123	0.4400	0.3084	1.0000	0.0465
model (15)	1.0000	0.0113	0.4480	0.3167	1.0000	0.0678
model (16)	1.0000	0.0121	0.9840	0.4698	1.0000	0.0169
model (17)	1.0000	0.0121	0.9840	0.5102	1.0000	0.0264

通过这些模拟, 我们可以看出, 在有限样本容量下, 我们提出的估计单指标模型 (2) 的估计量表现良好. 在绝对相关系数准则下, β_{n0} 和 β_n 都能很好的识别真实的方向 β. 另外, β_{n0} 总是能够选到真正起作用的变量 (TPR), 并且无关紧要的变量只以很小的概率被选入 (FPR). 最后我们指出, 即使方向含在 β 条件方差项, 我们的估计方法依然很有效. 相反, 当连接函数如模型 (14) 和 (15) 时, 稀疏 MAVE 方法表现欠佳.

5 汽车价格数据

汽车生产商生产的不同汽车品牌具有不同的品质, 诸如每千米行驶的里程数、马力等都不一样. 一方面, 生产商希望获得消费者愿意支付某种汽车的最高价格. 另一方面, 消费者从不同品牌的汽车中寻找, 希望对他们喜欢的一组品质支付最少的价钱. 消费者和生产商博弈的过程, 最后都能导致一种市场平衡, 形成不同属性的各类汽车品牌的市场价格. 市场价格和各种属性之间的关系被称为一类特殊的价格函数 [27]. 经济理论并没有说明这种函数的形状, 因为对不同的市场它可能不一样 (见 [28, 87 页]). 因此, 单指标模型在估计这种价格函数时提供了我们期望的灵活性.

我们收集到的汽车价格数据包含美国的 25 个汽车品牌. 品质差异体现在美国消费者联盟测定的 9 个属性. 具体地说, 自变量分别是每公里行驶的里程数 X_1、马力 X_2、车长 X_3、车宽 X_4、车重 X_5、高度 X_6、满意度 X_7、可靠度 X_8 和总评 X_9. 响应变量 Y 是不可协商的交易价格. 这个数据中自变量的维数相对于样本容量的大小应该可以认为是很高的.

Naik 和 Tsai[8] 首先通过切片逆回归分析了这批数据, 利用 t 检验 [29] 发现马力、车长和满意度对价格有显著的影响. 但是, 对选择重要变量来说, t 检验本身可能并不恰当. 原因有二: 一是切片逆回归估计的标准差并不准确 (见 [29, 219 页]); 二是当模型含有不相干的变量时, 标准差很有可能会不正常地放大 [3]. 因此, 我们应用提出的稀疏估计来选择变量, 得到的稀疏估计是 $\beta_n = (0, 0.0629, 0, 0, 0, 0, 0.1063, 0, 0)^\tau$. 通过变换到预测变量的原始尺度, 然后对方向进行尺度变换使之具有单位长度, 我们得到估计的线性组合 $0.0374X_2 + 0.9993X_7$. 因此, 选择到的变量是马力 (X_2) 和满意度 (X_7), 这与 Naik 和 Tsai[8] 基于 AIC_C 的分析一致.

我们的分析得到, 模型 (2) 恰当地描述了在这个市场的价格函数, 其中

$$\beta_n = (0, 0.0374, 0, 0, 0, 0, 0.9993, 0, 0)^\tau.$$

我们考虑的单指标模型提供了一种线性回归分析方法所不具备的但是又常常在这个领域用到的灵活性 [8].

参考文献

1 Donoho D L. High-dimensional data analysis: the curses and blessings of diemsnionality. Aide-memoire of a lecture at AMS conference on Math Challenges of the 21st Century, 2000

2 Naik P A, Hagerty M R, Tsai C L. A new dimension reduction approach for data-rich marketing environments: sliced inverse regression. J Mark Res, 2000, 37: 113–134

3 Altham P M E. Improving the precision of estimation by fitting a model. J Roy Statistit Soc B, 1984, 46: 118–119

4 Cook R D. Regression Graphics: Ideas for Studying Regressions through Graphics. New York: Wiley & Sons, 1998

5 Fan J Q, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360

6 Fan J Q, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. Ann Statist, 2004, 32: 928–961

7 Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Statist Soc B, 1996, 58: 267–288
8 Naik P A, Tsai C L. Single-index model selections. Biometrika, 2001, 88: 821–832
9 Kong E, Xia Y C. Variable selection for the single-index model. Biometrika, 2007, 94: 217–229
10 Cook R D, Li B. Dimension reduction for conditional mean in regression. Ann Statist, 2002, 30: 455–474
11 Ni L Q, Cook R D, Tsai C L. A note on shrinkage sliced inverse regression. Biometrika, 2005, 92: 242–247
12 Li L X, Nachtsheim C J. Sparse sliced inverse regression. Technometrics, 2006, 48 503–510
13 Li L X, Yin X R. Sliced inverse regression with regularizations. Biometrics, 2007, 64: 124–131
14 Li L X. Sparse sufficient dimension reduction. Biometrika, 2007, 97: 603–613
15 Zhu L P, Zhu L X. On distribution-weighted partial least squares with diverging number of highly correlated predictors. J Roy Statist Soc B, 2009, 71, 525–548
16 Zhu L P, Zhu L X. Dimension reduction for conditional variance in regressions. Statistica Sinica, 2009, 19: 869–883
17 Li K C. Sliced inverse regression for dimension reduction (with discussion). J Amer Statist Assoc, 1991, 86: 316–342
18 Hall P, Li K C. On almost linearity of low dimensional projection from high dimensional data. Ann Statist, 1993, 21: 867–889
19 Cook R D, Ni L. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J Amer Statist Assoc, 2005, 100: 410–428
20 Huber P J. Robust regression: asymptotics, conjectures and Monte Carlo. Ann Statist, 1973, 1: 799–821
21 Candès E J, Romberg J, Tao T. Stable singal recovery from incomplete and inaccurate measurements. Comm Pure Appl Math, 2006, 59: 1207–1223
22 Li L X, Cook R D, Tsai C L. Partial inverse regression. Biometrika, 2007, 94: 615–625
23 Xia Y C, Li W K, Tong H, et al. An adaptive estimation of optimal regression subspace. J Roy Statist Soc B, 2002, 64: 363–410
24 Wang Q, Yin X. A nonlinear multi-dimensional variable selection method for high dimensional data: sparse MAVE. Comp Statist Data Anal, 2008, 52: 4512–4520
25 Tobin J. Estimation of relationsjips for limited dependent variables. Econometrika, 1958, 26: 24–36
26 Li L X, Simonoff J S, Tsai C L. Tobit model estimation and sliced inverse regression. Statistical Modelling, 2007, 7: 107–123
27 Rosen S. Hedonic prices and implicit markets: product differentiation in perfect competition. J Polit Econ, 1974, 83: 34–55
28 Palmquist R B. Hedonic methods. In: Braden J B, Kolstad C D, eds. Measuring the Demand for Environmental Quality. Amsterdam: North-Holland, 1991, 77–120
29 Chen C H, Li K C. Can SIR be as popular as multiple linear regression? Statist Sinica, 1998, 8: 289–316
30 Pollard D. Convergence of Stochastic Processes. New York: Springer-Verlag, 1984

附录

定理 2.1 的证明 不失一般性, 我们假设 $E(X)=0$. 从而, $\Sigma_{XX}=E[XX^\tau]$ 且 $\Sigma_{XY}=E[XF(Y)]$. 相应的估计分别为 $\widehat{\Sigma}_{XX}=\frac{1}{n}\boldsymbol{X}^\tau\boldsymbol{X}$, $\widehat{\Sigma}_{XY}=\frac{1}{n}\boldsymbol{X}^\tau F_n(\boldsymbol{Y})$. 将 $\beta_{n0}-\beta_0$ 展为

$$\begin{aligned}\|\beta_{n0}-\beta_0\|_{\ell_2} &= \|\widehat{\Sigma}_{XX}^{-1}\widehat{\Sigma}_{XY}-\Sigma_{XX}^{-1}\Sigma_{XY}\|_{\ell_2}\\ &\leqslant \|\Sigma_{XX}^{-1}[\widehat{\Sigma}_{XY}-\Sigma_{XY}]\|_{\ell_2}+\|\Sigma_{XX}^{-1}[\Sigma_{XX}-\widehat{\Sigma}_{XX}]\widehat{\Sigma}_{XX}^{-1}\widehat{\Sigma}_{XY}\|_{\ell_2}\\ &:= \|T_1\|_{\ell_2}+\|T_2\|_{\ell_2}.\end{aligned}$$

下面分别处理以上两项. 首先, 我们指出

$$\|T_1\|_{\ell_2}=O(\sqrt{p_n/n}),\quad \text{依概率成立.} \tag{18}$$

注意到

$$\widehat{\Sigma}_{XY}-\Sigma_{XY}=\frac{1}{n}\sum_{i=1}^{n}x_iF_n(y_i)-E[XF(Y)]$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i[F_n(y_i) - F(y_i)] + \frac{1}{n}\sum_{i=1}^{n} x_i F(y_i) - E[XF(Y)]$$
$$:= T_{11} + T_{12}.$$

由 [30, 定理 37, 34 页] 不难得到 $\sup_y |F_n(y) - F(y)| = O(n^{-1/2})$ 依概率成立. 因为 T_{11} 可以依类似 T_{12} 的证明处理, 我们只讨论 T_{12}. 也就是说, 我们将会说明

$$\|\Sigma_{XX}^{-1} T_{12}\|_{\ell_2} = O(\sqrt{p_n/n}), \quad \text{依概率成立.} \tag{19}$$

再注意到

$$\begin{aligned} E\|\Sigma_{XX}^{-1} T_{12}\|_{\ell_2}^2 &= E\|\Sigma_{XX}^{-1}[XF(Y) - E(XF(Y))]\|_{\ell_2}^2/n \\ &\leqslant E\|\Sigma_{XX}^{-1} X\|_{\ell_2}^2/n = \operatorname{trace}(\Sigma^{-1})/n \leqslant \lambda_{\max}\{\Sigma^{-1}\}p_n/n \\ &\leqslant \lambda_{\min}^{-1}\{\Sigma\}p_n/n = O(p_n/n). \end{aligned}$$

利用类似 [30, 定理 37] 的证明, 可以得到 (19).

下面处理第二项 T_2. 我们将会指出

$$\|T_2\|_{\ell_2} = O(p_n/n), \quad \text{依概率成立.} \tag{20}$$

便于记号上清楚, 记 $T_{21} = \Sigma_{XX}^{-1}[\widehat{\Sigma}_{XX} - \Sigma_{XX}]\Sigma_{XX}^{-1}\Sigma_{XY}$. 利用类似的证明 (19) 的技术, 我们有

$$\begin{aligned} E\|T_{21}\|_{\ell_2}^2 &= E\|\Sigma_{XX}^{-1}[\widehat{\Sigma}_{XX} - \Sigma_{XX}]\Sigma_{XX}^{-1}\Sigma_{XY}\|_{\ell_2}^2 \\ &= E\|\Sigma_{XX}^{-1}[XX^\tau - \Sigma_{XX}]\Sigma_{XX}^{-1}\Sigma_{XY}\|_{\ell_2}^2/n \\ &= E[\Sigma_{XY}^\tau\Sigma_{XX}^{-1}(XX^\tau - \Sigma_{XX})\Sigma_{XX}^{-2}(XX^\tau - \Sigma_{XX})\Sigma_{XX}^{-1}\Sigma_{XY}]/n \\ &= \operatorname{trace}\{E[\Sigma_{XX}^{-1}(XX^\tau - \Sigma_{XX})\Sigma_{XX}^{-2}(XX^\tau - \Sigma_{XX})\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{XY}^\tau]\}/n. \end{aligned}$$

注意到 $E[XF(Y)]E[X^\tau F(Y)] \leqslant E[XX^\tau F^2(Y)] \leqslant E(XX^\tau)$. 即 $\Sigma_{XX} - \Sigma_{XY}\Sigma_{XY}^\tau$ 是正定矩阵. 因此, 最后一项小于或等于

$$\begin{aligned} &\operatorname{trace}\{E[\Sigma_{XX}^{-1}(XX^\tau - \Sigma_{XX})\Sigma_{XX}^{-2}(XX^\tau - \Sigma_{XX})]\}/n \\ &\quad = \operatorname{trace}\{E[\Sigma_{XX}^{-1/2}(XX^\tau - \Sigma_{XX})\Sigma_{XX}^{-2}(XX^\tau - \Sigma_{XX})\Sigma_{XX}^{-1/2}]\}/n. \end{aligned}$$

记 $X_0 =: \Sigma_{XX}^{-1/2}X$, 从而上面的项等于

$$\begin{aligned} &\operatorname{trace}\{E[(X_0X_0^\tau - I_{p_n})\Sigma_{XX}^{-1}(X_0X_0^\tau - I_{p_n})]\}/n \\ &\quad \leqslant \operatorname{trace}\{E[X_0X_0^\tau\Sigma_{XX}^{-1}X_0X_0^\tau]\}/n = E[X_0^\tau\Sigma_{XX}^{-1}X_0X_0^\tau X_0]/n \\ &\quad \leqslant \lambda_{\max}\{\Sigma_{XX}^{-1}\}E\|X_0\|_{\ell_2}^4/n = O(p_n^2/n), \end{aligned}$$

因为 T_{21} 是 T_2 的主项, 并且容易得到依概率 $O(\sqrt{p_n/n})$ 是 $\|T_2 - T_{21}\|$ 的上界, 所以 (20) 成立. 因为一致地有 $\max_{1\leqslant i\leqslant p_n} E|X_i|^4 \leqslant C_2 < \infty$, 所以最后一项成立. 从而, 结合 (18) 和 (20), 结论成立.

定理 3.1 的证明 为了便于阅读, 我们将整个证明分成两大步和一些小步.

第 1 步 在这一步, 我们指出

$$\|\beta_n-\beta_{n0}\|_{\ell_2}\leqslant\frac{2(1+\sqrt{\rho})}{C_{|T_0|}}\cdot\tau_n+2\sqrt{\rho}\left[\frac{(1+\sqrt{\rho})\sqrt{1+\delta_{3|T_0|}}}{C_{|T_0|}}+1\right]\cdot\frac{\|\beta_{n0}-\beta_0\|_{\ell_1}}{|T_0|}. \tag{21}$$

步 1.1 由三角不等式可以得到

$$\|\boldsymbol{X}\beta_{n0}-\boldsymbol{X}\beta_0\|_{\ell_2}\leqslant\|F_n(\boldsymbol{Y})-\boldsymbol{X}\beta_n\|_{\ell_2}+\|F_n(\boldsymbol{Y})-\boldsymbol{X}\beta_{n0}\|_{\ell_2}\leqslant 2\times\tau_n. \tag{22}$$

最后的不等式成立是因为 $\|F_n(\boldsymbol{Y})-\boldsymbol{X}\beta_{n0}\|_{\ell_2}\leqslant\|F_n(\boldsymbol{Y})-\boldsymbol{X}\beta_n\|_{\ell_2}$.

步 1.2 因为 β_{n0} 是可行的, 必须有 $\|\beta_n\|_{\ell_1}\leqslant\|\beta_{n0}\|_{\ell_1}$. 将分解 β_n 为 $\beta_n=\beta_{n0}+h$. 令 $\rho=1/\sqrt{3}$, 同样 $C_{|T_0|}=:\sqrt{1-\delta_{4|T_0|}}-\sqrt{\rho}\sqrt{1+\delta_{3|T_0|}}$. 由 Candès, Romberg 和 Tao[21] 的公式 (16) 可得

$$\|h\|_{\ell_2}\leqslant(1+\sqrt{\rho})\frac{\|\boldsymbol{X}h\|_{\ell_2}}{C_{|T_0|}}+2\sqrt{\rho}\left[\frac{(1+\sqrt{\rho})\sqrt{1+\delta_{3|T_0|}}}{C_{|T_0|}}+1\right]\cdot\frac{\|\beta_{n0}-\beta_0\|_{\ell_1}}{|T_0|}, \tag{23}$$

再由 (22), 可以得到 (21).

第 2 步 利用三角不等式可得到

$$\begin{aligned}\|\beta_n-\beta_0\|_{\ell_2}&\leqslant\|\beta_n-\beta_{n0}\|_{\ell_2}+\|\beta_{n0}-\beta_0\|_{\ell_2}\\&\leqslant\frac{2(1+\sqrt{\rho})}{C_{|T_0|}}\cdot\tau_n+\left[\frac{2\sqrt{\rho p_n}[(1+\sqrt{\rho})\sqrt{1+\delta_{3|T_0|}}+C_{|T_0|}]}{|T_0|C_{|T_0|}}+1\right]\cdot\|\beta_{n0}-\beta_0\|_{\ell_2}.\end{aligned} \tag{24}$$

因为 $\|\beta_{n0}-\beta\|_{\ell_1}^2\leqslant p_n\cdot\|\beta_{n0}-\beta\|_{\ell_2}^2$, 所以最后的不等式成立. 定理 2.1 显示第二项的阶为 $O(p_n/\sqrt{n})$, 显然, 当 $n\to\infty$ 时, 它被第一项 $O(\tau_n)$ 所控制. 因此可以得到, 依很大的概率,

$$\|\beta_n-\beta_0\|_{\ell_2}\leqslant\frac{2(1+\sqrt{\rho})}{C_{|T_0|}}\cdot\tau_n.$$