分类号：_____ 学校代码：10165

密　级：_____ 学　　号：200910784

遼寧師範大學

硕士学位论文

广义线性模型的惩罚高维经验似然

作 者 姓 名：刘　琦

学科、专业：概率论与数理统计

研 究 方 向：应用概率统计

导 师 姓 名：侯文副教授

2012 年 4 月

分类号：　　学校代码：10165

密　级：　　学　　号：200910784

遼寧師範大學

硕 士 学 位 论 文

广义线性模型的惩罚高维经验似然

作 者 姓 名：刘　琦

学科、专业：概率论与数理统计

研 究 方 向：应用概率统计

导 师 姓 名：侯文副教授

2012 年 4 月

学位论文独创性声明

本人承诺：所呈交的学位论文是本人在导师指导下所取得的研究成果。论文中除特别加以标注和致谢的地方外，不包含他人和其他机构已经撰写或发表过的研究成果，其他同志的研究成果对本人的启示和所提供的帮助，均已在论文中做了明确的声明并表示谢意。

学位论文作者签名：刘琦

学位论文版权的使用授权书

本学位论文作者完全了解辽宁师范大学有关保留、使用学位论文的规定，及学校有权保留并向国家有关部门或机构送交复印件或磁盘，允许论文被查阅和借阅。本文授权辽宁师范大学，可以将学位论文的全部或部分内容编入有关数据库并进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，并且本人电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后使用本授权书。

学位论文作者签名：刘琦　　指导教师签名：侯文

签名日期：2012 年 6 月 1 日

摘　　要

经验似然方法是由 Owen[2-3]提出的，它是统计推断中非常重要的方法之一，且有许多的优点. 本文讨论惩罚经验似然方法在广义线性模型下的参数估计和变量选择问题. 在广义线性模型下选取适当的惩罚函数结合经验似然方法对回归系数作统计推断，证明了惩罚经验似然方法具有 Oracle 性质.

本文由以下四部分组成：

第一部分介绍了广义线性模型的基本特点，经验似然方法的定义和它的两个基本的定理，以及一些常见的惩罚函数的表达形式，指出了对于高维数据，经验似然方法也是适用的. 同时，对变量选择加以了简单的介绍，列举了几种常用的变量选择方法.

第二部分通过构造 β 的经验似然函数

$$L(\beta)=\sup\{\prod_{i=1}^{n}\omega_i:\omega_i\geq 0,\sum_{i=1}^{n}\omega_i=1\text{和}\sum_{i=1}^{n}\omega_i U_i(\beta)=0\}. \qquad (2.2)$$

在一定条件下，针对带有发散参数的参数估计和变量选择问题提出了惩罚经验似然方法. 通过选取适当的惩罚函数，我们发现惩罚经验似然方法具有 Oracle 性质. 也就是说，如果模型的稀疏性已知，依概率 1，惩罚经验似然确定的真实模型和估计非零系数一样有效，在假设检验和构造置信区间时，惩罚经验似然方法的优点充分体现出来.

第三部分给出 SCAD 的算法并选择了优化的参数，并给出一个数值模拟来证实本文的主要结果.

第四部分对本文的定理进行了证明. 首先，提出了引理并对其加以证明，通过运用 Lagrange 乘子法，中心极限定理得到本文中部分结果，也显示出其在统计推断中的重要作用.

关键词：广义线性模型；惩罚经验似然； SCAD； 变量选择

Penalized High Dimensional Empirical Likelihood for Generalized Linear models

Abstract

The empirical likelihood method, introduced by Owen[2-3], is one of the most important statistical inference methods. And it has many advantages. This paper considers the problems of variable selection and estimation in generalized linear models via the penalized empirical likelihood (PEL) method. In the generalized linear models, we select an appropriate penalty function combined with the empirical likelihood method for statistical inference on the regression coefficients. Furthermore, we show that PEL has the oracle property.

The paper consists of the following four chapters:

Chapter 1 is mainly focused on introducing the form of generalized linear model, the definition and two theorem of empirical likelihood method, the form of kinds of penalty functions, for high dimensionality data, the empirical likelihood method is still applied, and introduced the variable selection method.

In chapter 2, Give the empirical likelihood function of β:

$$L(\beta) = \sup\{\prod_{i=1}^{n} \omega_i : \omega_i \geq 0, \sum_{i=1}^{n} \omega_i = 1 \text{和} \sum_{i=1}^{n} \omega_i U_i(\beta) = 0\}.$$

The penalized empirical likelihood for parameter estimation and variable selection for problems with diverging numbers of parameters is proposed. By using an appropriate penalty function, we show that PEL has the oracle property. Our results are demonstrated regression coefficients in generalized linear models. That is, with probability tending to one, penalized empirical likelihood identifies the model and estimates the nonzero coefficients as efficiently as if the sparsity of the true model were known in advance. The advantage of penalized empirical likelihood is illustrated in testing hypothesis and constructing confidence sets.

In chapter 3, we give the computation of the SCAD as well as the choice of the tuning parameter. and numerical simulations confirm our theoretical findings.

The proof of main results is given in chapter 4. Firstly, to prove main conclusion, we establish several lemmas which are of significance on their own right, The estimation techniques Lagrange multiplier method, central-limit theorem play an important role in deriving our results, which show the important position and role in the statistical inference.

Key Words： Generalized linear models; Penalized empirical likelihood; Smoothly clipped absolute deviation(SCAD); Variable selection

目　　录

1 绪论

1.1 广义线性模型

广义线性模型（GLMs）首先由 Nelder and Wedderburn [1] 提出. 该模型的应用范围非常的广泛，在医学、经济、生物等众多领域都有着非常重要的应用. 它是我们熟悉的经典线性模型的一种推广， 线性回归模型，方差分析模型，用于列联表分析的 Log-Log 模型和 Logistic 模型都是广义线性模型的特例. 广义线性模型的三部分假设如下：

（i）随机成分，即响应变量Y服从自然指数族分布，其密度函数为

$$f(y,\theta,\phi)=\exp\{\frac{[y\theta-b(\theta)]}{a(\phi)}\}+c(y,\phi),$$

其中θ为自然参数，即

$$E(Y)=\mu=b'(\theta).\ \ Var(Y)=a(\phi)b''(\theta).$$

设

$$l(y,\theta,\phi)=\log[f(y,\theta,\phi)]$$

那么，

$$E[\frac{\partial l(y,\theta,\phi)}{\partial\theta}]=0,\qquad -E[\frac{\partial^2 l(y,\theta,\phi)}{\partial\theta^2}]=E[\frac{\partial l(y,\theta,\phi)}{\partial\theta}]^2.$$

因此，

$$\frac{\partial l(y,\theta,\phi)}{\partial\theta}=\frac{y-b'(\theta)}{a(\varphi)}\ ,\ \frac{\partial^2 l(y,\theta,\phi)}{\partial\theta^2}=\frac{-b''(\theta)}{a(\phi)}.$$

（ii）系统成分，即自变量的线性组合η，

$$\eta=\beta_0+\beta_1x_1+\ldots+\beta_px_p.$$

尽管系统成分η是X的线性函数，但是$E(Y|X)$不必是X的线性函数.

（iii）联接函数，联接函数g是单调可微函数，它将线性预测与$E(Y|X)$相联系，即

$$g(\mu)=\eta=\beta_0+\beta_1x_1+\ldots+\beta_px_p.$$

g的单调性保证了映射是一对一的，因此我们以联接函数的反函数来表示$E(Y|X)$，即

$$E(Y|X)=g^{-1}(\beta_0+\beta_1x_1+\ldots+\beta_px_p).$$

常见的联接函数有：

1. 正态分布:$\eta=g(\mu)=\mu$.
2. 泊松分布：$\eta=g(\mu)=\log(\mu)$.

3. 二项分布：$\eta = g(\mu) = \log(\frac{\mu}{1-\mu})$.

4. 伽马分布：$\eta = g(\mu) = \frac{-1}{\mu}$.

5. 逆高斯分布：$\eta = g(\mu) = \frac{-1}{2\mu^2}$.

对广义线性模型，我们记$G(X^T\beta) = g^{-1}(X^T\beta)$，因此，对于随机向量$X \in R^p$，有

$$E(Y|X) = G(X^T\beta) \quad 和 \quad Var(Y|X) = \sigma^2 V(G(X^T\beta))，\qquad (1.1)$$

其中，$\beta \in \Re^p$一个真实的参数向量，G是一个已知的光滑联接函数和V是一个已知的方差函数. 对于半参数模型，拟似然（Wedderburn[4]）已经成为了一个受欢迎的方法，它是广义线性模型的基本框架，参见 McCullagh and Nelder[5]. 在广义线性模型（1.1）下，本文将研究β的估计问题.

1.2 经验似然方法简介

经验似然作为一种构造置信域的非参数统计推断方法是由 Owen[2]提出的. 该方法较一些经典的统计方法优越. 它足以和经典方法如正态近似理论以及 Bootstrap 与 Jackknife 相媲美. 此外，经验似然方法和 Bootstrap 与 Jackknife 一样都不用预先给定数据所属的分布族. 而且经验似然置信域的形状可以自动由数据决定而不用预先给定，而 Bootstrap 方法才能做到让数据决定置信域的形状.

假设观测值$y_1, y_2, \ldots, y_n$为独立同分布的样本，分布函数为$F(y)$（未知）. 定义F的对数经验似然函数为,

$$l_n(F) = \sum_{i=1}^{n} \log(p_i)$$

其中，$p_i = F(\{y_i\}) = \Pr(Y_i = y_i)$. F的总体均值为$\theta = (\theta_1, \ldots, \theta_d)$，可以转化为剖面经验似然. θ的剖面经验似然定义为

$$l_{EL}(\theta) = \sup\{l_n(F) : p_i > 0, i = 1, \ldots, n; \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i(y_i - \theta) = 0\}$$

其中$F(y) = \sum_{i=1}^{n} p_i I(y_i \le y)$. 最大化问题就非常简单了. 在这些限制条件下，$\theta, l_n(F)$被最大化，当

$$\hat{p}_i = \frac{1}{n\{1 + \lambda^T(y_i - \theta)\}},$$

对于$i=1,2,\ldots,n$. 由下式可以得到拉格朗日乘子λ.

$$\sum_{i=1}^{n} p_i(y_i-\theta)=0.$$

因此，我们可以记作

$$l_{EL}(\theta)=-n\log(n)-\sum_{i=1}^{n}\log\{1+\lambda^T(y_i-\theta)\}.$$

定义剖面对数经验似然比函数为

$$W(\theta)=-2\sum_{i=1}^{n}\log(n\hat{p}_i)=2\sum_{i=1}^{n}\log\{1+\lambda^T(y_i-\theta)\}.$$

Owen [3]指出，当θ_0是真实的总体均值时，$W(\theta_0)\to\chi_d^2$，$n\to\infty$类似于参数似然的结果. 这个事实对于θ的假设检验和置信区间的构造是非常有利的.

下面简单介绍经验似然的两个基本定理：

定理 1.2.1（Owen[2]）设 $X_1,X_2,\ldots,X_n\in\Re^p$独立同分布，分布函数为$F_0$，均值为$\mu_0$，方差$V_0$有限且秩为$q>0$，则$C_{r,n}$是一个凸集，且$-2\ln R(\mu_0)$依分布收敛于$\chi^2(q)$. 其中

$$R(\mu)=\max\{\prod_{i=1}^{n}n\omega_i\Big|\sum_{i=1}^{n}\omega_iX_i=\mu,\omega_i\geq 0.\sum_{i=1}^{n}\omega_i=1\}.$$

且μ_0的置信域可以写成下面的形式

$$C_{r,n}=\{\sum_{i=1}^{n}\omega_iX_i\Big|\prod_{i=1}^{n}n\omega_i\geq r,\omega_i\geq 0.\sum_{i=1}^{n}\omega_i=1\}.$$

利用上述定理，可以构造μ的置信度为$1-\alpha$的置信区间，如下

$$I_\alpha=\{\mu:-2\log R(\mu)\leq\chi_{1-\alpha}^2(q)\},$$

其中$\chi_{1-\alpha}^2(q)$为自由度为q的χ^2分布的$1-\alpha$的分位数.

由上述定理的证明，我们还可以知道

$$-2\log R(\mu_0)=n(\bar{X}-\mu_0)^T\hat{V}^{-1}(\bar{X}-\mu_0),$$

其中

$$\hat{V}^{-1}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)(X_i-\mu_0)^T.$$

若把上式中的μ_0换成$\bar{X}$，就类似于 Hotelling's T^2,

$$T^2=n(\bar{X}-\mu_0)^TS^{-1}(\bar{X}-\mu_0),$$

其中

$$S=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})^T,$$

经过代数变换得到

$$n(\overline{X}-\mu_0)^T\hat{V}^{-1}(\overline{X}-\mu_0)=T^2(1+\frac{T^2-1}{n})^{-1}=T^2+O_p(\frac{1}{n}).$$

当V_0已知时，

$$n(\overline{X}-\mu_0)^T V_0^{-1}(\overline{X}-\mu_0)\xrightarrow{C}\chi^2(q)$$

当V_0未知时，我们可以用 Hotelling 统计量T^2，即

$$\frac{(n-d)T^2}{(n-1)d}\xrightarrow{C}F_{(d,n-d)}，\ 当 n\to\infty.$$

于是，临界值$\chi^2_{1-\alpha}(q)$换成$\frac{d(n-1)}{n-d}F^{1-\alpha}_{d,n-d}$更为合适. 当$n\to\infty$时，两者是等价的.

定理 1.2.2 设$X_1,X_2,\ldots X_n\in R^d$独立同分布，有共同的分布$F_0$，参数$\theta\in\Theta\subseteq R^p$，$X\in R^d$，$m(X,\theta)\in R^r$. 令$\theta_0\in\Theta$且$Var(m(X_i,\theta_0))$有限，秩为$q>0$, 若$Em(X,\theta_0)=0$, 则

$$-2\log R(\theta_0)\xrightarrow{C}\chi^2(q), n\to\infty$$

其中$R(\theta)=\max\{\prod_{i=1}^{n}n\omega_i\Big|\sum_{i=1}^{n}\omega_i m(X_i,\theta)=0,\omega_i\geq 0,\sum_{i=1}^{n}\omega_i=1\}.$

1.3 变量选择

变量选择对于模型的构建起着重要的作用. 在实践中，一方面为了防止漏掉关键的变量，大量的变量被引入. 另一方面，很多变量并不是具有显著的影响，这就需要选择有用的变量以提高预测的精度和降低模型的复杂度. 随着高维数据的出现，变量选择仍将成为数据分析的重要的内容. 因此，适当的变量选择是必须的，也具有重要的意义.

统计学家的一个目的就是考虑如何从众多的变量中选取出重要的变量，用来提高预测的准确度. 变量选择更多的是从实际角度考虑，从专业或经验的角度看，哪些变量对目标Y有重要影响. 我们可以根据选择的模型做统计分析，这样模型就简化了许多. 根据已经选择的模型来做的统计分析，是可以反映出真实模型的一些性质的. 对于如何进行变量选择，不同的统计学家提出不同的准则，常见的变量选择方法有以下几种：

（1） *AIC* 准则. 这一准则是由 Akaike 提出的，它是基于以下统计量

$$AIC=-2\ln(\sup f(x,\theta))+2(q).$$

来进行判断和选择的. 其中， $\ln(\sup f(x,\theta))$ 表示对数极大似然函数值， $2(q)$ 为惩罚项， q 表示模型最小参数个数. 在模型中参数及参数个数不同时， *AIC* 的值是不同的. 原则是：两个不同模型比较， *AIC* 的值小的入选. 但理论上 *AIC* 不能给出相合估计.

（2） *BIC* 准则. 它是一种具有相合性的变量选择方法，弥补了 *AIC* 准则的缺点，把 *AIC* 准则后面的 $2q$ 换成了 $k\log n$ ，其统计量为：

$$BIC = -2\log(\sup f(x,\theta)) + k\log n.$$

其中， n 是总自变量的个数， k 是每一次选进模型中的自变量的个数.

（3）基于逐步选择的方法. 这种方法包含：向前引入法，向后剔除法和逐步筛选法. 这里只简单介绍逐步筛选法，它是诸多变量选择方法中效果较好的方法，而且应用较为广泛，它是向前引入法和向后剔除法德综合. 其基本思想是在众多的自变量中，根据这些变量各自对回归方程影响的大小，逐一地选入到回归方程中.

（4） *ITC* 准则（也称为 *EDC* 准则）. 具体参见 Zhao,Krishnaiah and Bai[6].

1.4 高维数据惩罚估计

本文介绍以下几种常见的惩罚：

1. SCAD 惩罚（Smoothly Clipped Absolute Deviation penalty）：

$$p'_\tau(t) = \tau\{I(t \le \tau) + \frac{(a\tau - t)_+}{(a-1)\tau} I(t > \tau)\} \qquad (1.4)$$

对于某个 $a > 2$ 和 $t > 0$ （ Fan 和 Li[7]）.

2. 硬门限（hard thresholding）惩罚函数：

$$p_\tau(|t|) = \tau^2 - (|t| - \tau)^2 I(|t| < \tau)$$

3. L_1 （*LASSO*， least absolute shrinkage and selection operator）惩罚函数： $P_\tau(|t|) = \tau|t|$

4. L_2（ridge regression）惩罚函数：

$$p_\tau(|t|) = \tau|t|^2$$

惩罚最大似然估计得到了非常广泛的应用. 在许多实际问题中，针对相关变量的选择稀疏惩罚方法起着至关重要的作用. Tibshirani[8] 的最小二乘估计和缩减估计（LASSO）可称为是显著的例子. LASSO 的相关变式包括由 Frank 和 Friedman[9] 研究的桥估计；Efron，Hastie，Johnstone 和 Tibshirani[10] 的最小角度回归技术；Fan 和 Li[7] 与 Fan 和 Peng[11] 提出了 SCAD 惩罚估计，后来，这个结果被 Kim 等[12] 拓展到高维的情形，针对 LASSO 惩罚，Knight 和 Fu[13] 研究了有限参数的 LASSO 型估计的渐近性质： Zou[14] 进一步将其推广为适应 LASSO, 当缩减参数的权被适当加以控制时，其具有 Oracle 性质.

近年来，高维数据分析的研究已经成为了一个活跃的领域，当数据维数发散时，进行有效的变量选择是可行的. 基于统计模型下的变量选择，尤其对于参数发散的广义线性模型，它可以显著提高拟合模型的性能. 如 Hastie，Tibshirani 和 Friedman[15]，Fan 和 Lv[16] 讨论的. 惩罚经验似然可以恰当的调整偏方差方法使得它的表现形式可以得到改善. 对于惩罚似然方法及研究可以参见文献 Tibshirani[17]；Fan 和 Li[7]；Zou[14]；Wang，Li 和 Tsai[18]；Zhang 和 Lu[19]；Candes 和 Tao[20].

在本文中，我们将主要研究对于不断增加的数据维数下的经验似然方法，Hjort，McKeague 和 Van Keilogom[21] 指出当数据维数增加时，经验似然方法仍然生效. 本文将提出一个统一的框架用于变量选择，参数估计和通过 SCAD 惩罚经验似然进行推断. 当 $p \to n$ 时，对回归系数 $\beta \in \Re^p$ 进行估计，且回归参数是很少的，在某种意义上说 β 的元素为零，这说明对相应的预测因子的影响是较小的. 变量选择的首要目的是辨别 β 中的零元素，其次是在恰当的条件下，说明了非零系数的估计量具有相同的渐近分布，且它可以作为真实模型下的估计量. 所以，惩罚经验似然具有 Oracle 性质(Fan 和 Li[7]).

本文共分四部分，具体组织结构如下：

首先第一章主要介绍了本文课题研究所需的前人的研究结果，以及本论文的主要内容和基本结构. 第二章构造了 β 的惩罚经验似然值，并描述了它的稀疏性和渐近性质. 第三章给出 SCAD 的算法并选择了优化的参数，通过模拟来证实惩罚经验似然的作用，同时给出了一个实例分析. 第四章给出了本文主要结果的证明过程.

2 惩罚经验似然估计

2.1 方法和假设

广义线性模型（1.1）下的经验似然可以构造出标准的方程. 定义 β 的对数经验似然值函数为

$$\ell(\beta) = -\log\{L(\beta)\} - n\log(n)$$

我们定义惩罚经验似然估计量 $\hat{\beta}$ 是使

$$\log\{L(\beta)\} - n\sum_{i=1}^{p} p_\tau(|\beta_i|) \qquad (2.1)$$

的最大者. 其中 $p_\tau(t)$ 是调整参数为 τ 的惩罚函数. 关于惩罚函数 $p_\tau(t)$ 的例子参见 Fan 和 Li[7];Zou[14];Fan 和 Lv[16] 及 Lv 和 Fan[22].

因为 SCAD 惩罚具备上述所提到的三条性质，足见它是一个好的惩罚函数，近来，SCAD 惩罚也受到大家的广泛青睐，故本文我们将运用 SCAD 惩罚.

基于正规方程可以构造广义线性模型的经验似然. 设 $(X_1,Y_1),\ldots,(X_n,Y_n)$ 是一个独立同分布样本族，定义辅助变量

$$U_i(\beta) = \frac{G'(X_i^T\beta)}{V(G(X_i^T\beta))} X_i\{Y_i - G(X_i^T\beta)\},$$

定义 β 的经验似然如下：

$$L(\beta) = \sup\{\prod_{i=1}^{n}\omega_i : \omega_i \geq 0, \sum_{i=1}^{n}\omega_i U_i(\beta) = 0\}. \qquad (2.2)$$

根据拉格朗日乘子法，

$$\omega_i = \frac{1}{n}\frac{1}{1+\lambda_\beta^T U_i(\beta)}, \text{ 其中 } \lambda_\beta \text{ 满足 } n^{-1}\sum_{i=1}^{n}\frac{U_i(\beta)}{1+\lambda_\beta^T U_i(\beta)} = 0. \qquad (2.3)$$

惩罚经验似然估计量 $\hat{\beta}$ 可以被定义为式子（2.4）的最小值

$$\ell_p(\beta) = \sum_{i=1}^{n}\log\{1+\lambda_\beta^T U_i(\beta)\} + n\sum_{i=1}^{p} p_\tau(|\beta_i|), \qquad (2.4)$$

其中， $p_\tau(t)$ 是关于优化参数 τ 的惩罚函数，在这篇文字章中，惩罚（SCAD）的一阶导数满足（1.4）式. 对于 $a > 2$（Fan and Li[7]）. 根据 Fan and Li[7] 中的方法，本文设 $a = 3.7$. （1.4）惩罚是区间接近于零且除常数之外的一个样条函数，所以，它可以缩小一个较小估计量的值为零且同时不影响大的估计值.

设 $\Im=\{j:\beta_{0j}\neq 0\}$ 是 β_0 的真值平均向量的非零元素的集合，且它的基数为 $|\Im|=d$，其中，当 $n\to\infty$ 时，可以得到 $d\le p$. 记 $\beta=(\beta_1^T,\beta_2^T)^T$，其中 $\beta_1\in\Re^d$，$\beta_2\in\Re^{p-d}$，从而真值参数 $\beta_0=(\beta_{10}^T,0)^T$. 相应的，$X$ 的协方差矩阵可以记作 $\Sigma=\begin{pmatrix}\Sigma_{11}\Sigma_{12}\\ \Sigma_{21}\Sigma_{22}\end{pmatrix}$.

为得到本文的主要结果还需以下正则条件：

（A.1）Σ 的特征值满足 $C_1\le\gamma_1(\Sigma)\le\gamma_2(\Sigma)\le\cdots\le\gamma_p(\Sigma)\le C_2$，其中，$C_2>C_1>0$；

（A.2）$n\to\infty$，$p\to\infty$，$p^{1-2\delta}/n^{1/2-2\delta}\to 0$，其中，$\delta$ 在结果证明的引理 4.1 中有详细的说明，且 $d\le p$；

（A.3）$\{\varepsilon_i\}_{i=1}^n$ 是独立同分布的，且 $E(\varepsilon_i^{4k})<\infty$，$k\ge 1$；

（A.4）$G(\cdot)$，$V(\cdot)$ 是连续可导的，$G(\cdot)$ 的一阶导数和二阶导数都是有界的，即存在常数 $0<a<b<\infty$，使得 $a\le G'(\cdot),G''(\cdot)\le b$，此外，$a\le V(\cdot),V'(\cdot)\le b$；

（A.5）当 $n\to\infty$ 时，优化参数 τ 满足 $\tau\to 0$ 和 $\tau(n/p)^{1/2-\delta}\to\infty$，其中，$\delta$ 满足条件 A.2 中的规定，非零元满足 $\min_{j\in\Im}|\beta_{0j}|/\tau\to\infty$；

(A.6) $\max_{j\in\Im}p'_\tau(|\beta_{0j}|)=o\{(np)^{-1/2}\}$，$\max_{j\in\Im}p''_\tau(|\beta_{0j}|)=o(p^{-1/2+\delta}n^{-\delta})$.

条件 A.3，A.4 作为常规的条件被用在结果的证明中，当条件 A2 成立时，p,d 都可以是发散的，条件 A.5 对惩罚参数 τ 具有微弱的影响. 这个假设是常规的，它可以确保影响的加强. 条件 A.6 可以控制惩罚中对非零元的影响. 对于 SCAD 惩罚（1.4），由于 $\max_{j\in\Im}p'_\tau(|\beta_j|)=0$，且当 n 足够大时，在 A.5 成立的条件下，A.6 是成立的；参见 Fan and Li [7].

2.2 主要结果

定理 2.1 在正则条件 A.1-A.6 下，当 $n\to\infty$ 时，

1.（稀疏性）依概率 1，有 $\hat\beta_2=0$；

2.（渐近正态性）$\sqrt{n}W_nI_B^{-1/2}(\hat\beta_1-\beta_{10})\xrightarrow{d}N(0,T)$，

其中，$W_n\in\Re^{q\times d}$ 满足 $W_nW_n^T\to T$，为了进行计数过程，定义 $I_p=(H_1^T,H_2^T)$，其中 $H_1\in\Re^{p-d\times p}$. 对于固定的 q，$T\in\Re^{q\times q}$ 且 $I_B=\Gamma_{11}-\Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}$，其中 $\Gamma_{ij}=H_i\Sigma^{-1}H_j^T$.

W_n 表示由一个不同维数向量 β_1 到一个固定维数 q 的一个投影. 投影的极限分布可以被描述为一个多维的正态分布.

定理 2.1 阐述了广义线性回归中惩罚经验似然的 oracle 性质. 可以看到，应用经验似

然方法对β进行估计会受到$\beta_2=0$的约束. 在定理 2.1 的证明中同样可以看到（2.2）的最大值$\tilde{\beta}$受限于$\beta_2=0$. 在定理的第二部分得到$\tilde{\beta}$服从渐近正态分布.

关于非零回归系数β_1的假设检验如下：

$$H_0: L_n\beta_{10}=0 \quad \text{vs} \quad H_1: L_n\beta_{10}\neq 0,$$

其中，L_n是一个$q\times d$矩阵，对于固定的有限的q，有$L_nL_n^T=I_q$. 惩罚经验似然比统计量可以构建如下：

$$\tilde{\ell}(L_n)=-2\{\ell_p(\hat{\beta})-\min_{\beta:L_n\beta_1=0}\ell_p(\beta)\}. \tag{2.5}$$

在下面定理中，给出了经验似然比统计量的性质.

定理 2.2 在无效假设和定理 2.1 的条件下，当$n\to\infty$时，$\tilde{\ell}(L_n)\xrightarrow{d}\chi_q^2$. 因此，$L_n\beta$的$(1-\alpha)$-水平置信区间构建为

$$V_\alpha=[\beta:-2\{\ell_p(\hat{\beta})-\min_{\beta:L_n\beta_1=0}\ell_p(\beta)\}\le\chi_{q,1-\alpha}^2], \tag{2.6}$$

其中，$\chi_{q,1-\alpha}^2$是χ_q^2分布的$1-\alpha$分位数. 此外，当$n\to\infty$时，有$P(L_n\beta_{10}\in V_\alpha)\to 1-\alpha$.

对于广义线性模型，定理 2.2 对假设检验和构建数据的不受其它任何限制的置信区间再次给出了方便的方法. 在高维数据分析中，这种方法是可行的.

3 模拟与实例分析

3.1 算法

在数据分析中，优化参数τ选择对变量选择起着重要的作用. 我们用 BIC(Bayesian information criterion)来选择优化参数τ，

$$BIC_\tau = 2\ell_p(\beta_\tau) + C_n \cdot \log(n) \cdot df_\tau, \qquad (3.1)$$

其中，β_τ是关于优化参数τ的β的惩罚经验似然估计；df_τ是β_τ中非零系数的个数；C_n是一个比例因子，当$p \to \infty$时，它依一个较慢的速率（Tang and Leng[23]）趋于无穷. 当p固定时，简记$C_n = 1$，同时，$C_n = \max\{\log\log p, 1\}$成为一个较好的选择. 在本文中，我们计划使用非线性优化算法迭代和 LQA（local quadratic approximation）算法求（2.4）的惩罚经验似然比的最小值. 设$p_\tau(|\beta_j|) = n\sum_{i=1}^{p} p_\tau(|\beta_i|)$，特别地，

$$p_\tau(|\beta_j|) \approx p_\tau(|\beta_j^{(k)}|) + \frac{1}{2}\{p'_\tau(|\beta_j^{(k)}|)/|\beta_j^{(k)}|\}(\beta_j^2 - \beta_j^{(k)2}),$$

其中，$\beta_j^{(k)}$是β_j中第k个估计值，$k = 0,1,\ldots$利用在 Owen[24]中讨论的算法通过非线性优化获得最小值. 重复步骤直至收敛. 由于（2.4）的第k个最小值$\hat{\beta}_j^{(k)}$是非常接近于零的，因此我们将$\hat{\beta}_j^{(k)}$设为零.

3.2 数值模拟

本部分讨论在 Logistic 模型下，应用惩罚经验似然对有限样本的值进行估计。通过对一组真实数据的分析，可以更加详尽的表面 SCAD 惩罚的性质.

例 3.1 （Logistic 回归）在本例中，产生 500 个数据的集合，每一个数据有n个观察值都来自 Logistic 回归模型，

$$P(Y = 1|X) = \frac{\exp(\beta^T X)}{1 + \exp(\beta^T X)},$$

其中，$\beta \in \Re^p$给定，$\beta_0 = (3,1.5,0,0,2,0,\ldots,0)$. 设$\sigma = 1$，考虑两种情况，即$n = 100, p = 10$和$n = 200, p = 20$. 协变量$X$服从$p$维多元正态分布，且第$i$个和第$j$个元的相关系数$\rho^{|i-j|}$设为$\rho = 0.5$. 表 3.1 给出了模拟的结果. 表 1 中的“MRME”列代表平均蒙特卡罗模拟.

根据表 3.1，可以看到 SCAD 惩罚要比经验似然的效果好，并且优于 Lasso 惩罚，SCAD 惩罚的变量选择结果是最好的. 根据表 3.2，可以看到β_j，$j = 1,2,5$的所有估计值都是非常

接近真值的，当参数的数量随着样本量的增加而逐渐增加时，SCAD 惩罚的性质和 Oracle 性质几乎同样好.

表 3.1　Logistic 回归的相关模型误差

Tab.3.1 Relative model errors for Logistic regression

(n, p)	Method	MRME	Correct	Incorrect
(100,10)	EL	0.6830	5.945	0.098
	Lasso	0.5750	6.223	0.003
	SCAD	0.2784	6.912	0
	Oracle	0.2046	7	0
(200,20)	EL	0.5021	14.343	0
	Lasso	0.3132	13.472	0
	SCAD	0.1836	16.952	0.08
	Oracle	0.1235	17	0

表 3.2　模拟研究：比较 EL，Lasso 和 SCAD

Tab.3.2 Simulation study:comparison of EL,Lasso and SCAD

(n, p)	Estimator	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_5$
(100,10)	EL	2.8765	1.4732	1.9543
	Lasso	2.9986	1.4949	1.9723
	SCAD	3.0020	1.4967	1.9969
	Oracle	3.0000	1.4999	1.9999
(200,20)	EL	3.0186	1.4854	1.9692
	Lasso	3.0125	1.4732	2.0021
	SCAD	3.0034	1.4963	1.9983
	Oracle	3.0023	1.4978	1.9995

3.3 实例

例 3.2 (数据分析) 本例考虑 Cameron and Trivedi[25] 中的数据‘doctor visits’. 这个数据的最初来源是澳大利亚卫生调查 1977-78. 应用泊松回归方法对这个数据进行变量选择. 数据包括 5190 个成年人的看医生的信息. 我们研究在过去两个星期内，十二个协变量对去看医生的数量的影响：X_1：Sex（1= 女性，0= 男性）．X_2：Age（年龄数除以 10）．X_3：Agesq（年龄的乘方）．X_4：Income（收入，澳大利亚的年度收入除以 1000 美元）．X_5：Levyplus，（1= 被私人健康保险覆盖）．X_6：Freepoor，（1= 由政府资助的低收入人）．X_7：Freerepa，（1= 由于年老，伤残抚恤金，财产状况而由政府免费资助的人）．X_8：Iliness，（过去两周内疾病的数目）．X_9：Actdays，（在过去两周内活动性减少的天数）．X_{10}：Hscore，（使用戈德保的方法计算一般的调查问卷的得分）．X_{11}：Chcond1，（ 1= 慢性疾病但不限制其活动，0= 其他）．X_{12}：Chcond2，（1= 慢性疾病且限制其活动，0= 其他）.

根据表 3.3，很容易看出 SCAD 惩罚可以获得最简单的模型. 它选择模型的效果几乎和 EBIC（Variyath et al.[26]）相同. 而且，SCAD 惩罚估计值的标准误差在所有其他的方法中是最小的. 所有的结果表明 SCAD 惩罚是一个可行的方法.

表 3.3　不同模型下回归系数的估计和标准误差

Tab.3.3 Estimates of regression coefficients and standard errors for different models identified

Variable	El	Lasso	SCAD	EBIC
Intercept	−2.2243(0.1897)	−2.1320(0.1812)	−2.1006(0.1799)	−2.1048(0.0618)
X_1	0.1573(0.0578)	0.2378(0.0543)	0.2512(0.0542)	0.2625(0.0526)
X_2	1.0557(1.0006)	0.9168(0.8932)	0	0
X_3	−0.8476(1.0632)	−0.6740(1.0452)	0	0
X_4	−0.2048(0.0778)	−0.1711(0.0819)	0	0
X_5	0.1230(0.0705)	0.0950(0.0867)	0	0
X_6	−0.4312(0.1778)	0	0	0
X_7	0.0798(0.0921)	0	0	0
X_8	0.1993(0.0162)	0.1864(0.0172)	0.2187(0.0196)	0.2321(0.0198)
X_9	0.1268(0.0048)	0.1334(0.0042)	0.1278(0.0049)	0.1358(0.0049)
X_{10}	0.0284(0.0125)	0.0274(0.0117)	0	0
X_{11}	0.1141(0.0666)	0	0	0
X_{12}	0.1416(0.0831)	0.1686(0.1040)	0	0

4 主要结果证明

在下面的引理中，λ_β 如上文的定义，其中 β 是 β_0 的一个足够大的领域. 在证明定理 1 之前提出以下引理.

引理 4.1 设 $a_n=(p/n)^{1/2-\delta}$，$D_n=\{\beta:\|\beta-\beta_0\|\le ca_n\}$ 其中 $\delta,c>0$ 是常数且 δ 满足 $p^{1-\delta}/n^{1/2-\delta}\to 0$，因此对于 $\beta\in D_n$，有 $\|\lambda_\beta\|=O_p(a_n)$.

证明：为简单起见，在以下定理中假设 $E(X_i)=0$. 通过对随机变量的中心化，它可以扩展到其他情况 $E(X_i)=\mu$. 设 $\lambda_\beta=\rho\theta$，其中 $\rho=\|\lambda_\beta\|$，θ 是单位向量，设 $\varepsilon_i=Y_i-G(X_i^T\beta_0)$ 是随机误差，其均值为 0，方差为 σ^2.

$$\rho\{\theta^T T_n(\beta)\theta-\max_{1\le i\le n}\|U_i(\beta)\|n^{-1}\left|\sum_{i=1}^n\theta^T U_i(\beta)\right|\}\le n^{-1}\left|\sum_{i=1}^n\theta^T U_i(\beta)\right|,$$

其中

$$T_n(\beta)=n^{-1}\sum_{i=1}^n U_i(\beta)U_i^T(\beta).$$

因为

$$U_i(\beta_0)=\frac{G'(X_i^T\beta_0)}{V(X_i^T\beta_0)}X_i\varepsilon_i,$$

ε_i 不依赖 X_i 的支配，且满足条件 A.3，由条件 A.4，得到

$$\left\|n^{-1}\sum_{i=1}^n U_i(\beta_0)\right\|=O_p\{(p/n)^{1/2}\}.$$

因此

$$n^{-1}\left|\sum_{i=1}^n\theta^T U_i(\beta_0)\right|=O_p\{(p/n)^{1/2}\}.$$

由

$$T_n(\beta_0)=n^{-1}\sum_{i=1}^n\frac{G'^2(X_i^T\beta_0)}{V^2(G(X_i^T\beta_0))}X_iX_i^T\varepsilon_i^2,$$

可以得到

$$\theta^T T_n(\beta_0)\theta = O_p(1).$$

从而，由条件 A.3 有

$$m^{4k} = E(\varepsilon_i^{4k}) < \infty,$$

由条件 A.4，可以得到

$$\begin{aligned}
&\max_i \|U_i(\beta_0)\| \\
&= \max\{\frac{G'(X_i^T\beta_0)}{V(G(X_i^T\beta_0))}\}\{\max_i(X_iX_i^T\varepsilon_i^2)^k - m^{2k}E(X_iX_i^T)^k + m^{2k}E(X_iX_i^T)^k\}^{1/2k} \\
&= \frac{b}{a}[\mathrm{var}\{(X_iX_i^T\varepsilon_i^2)^k\}\max_i\{\frac{(X_iX_i^T\varepsilon_i^2)^2 - m^{2k}E(X_iX_i^T)^k}{\mathrm{var}\{(X_iX_i^T\varepsilon_i^2)^k\}} + m^{2k}E(X_iX_i^T)^k]^{1/2k} \\
&= O(p^{1/2}) + o_p\{p^{-(2k-1)/(4k)}n^{1/4k}\}.
\end{aligned}$$

因此

$$\max_{1\le i\le n}\|U_i(\beta_0)\| n^{-1}\left|\sum_{i=1}^n \theta^T U_i(\beta_0)\right| = o_p(1).$$

并且可以推断出 $\|\lambda_{\beta_0}\| = O_p\{(p/n)^{1/2}\}$，对于其他的 $\beta \in D_n$，可以得到

$$n^{-1}\left|\sum_{i=1}^n \theta^T U_i(\beta)\right| = n^{-1}\left|\sum_{i=1}^n \theta^T\{U_i(\beta_0) + K_1X_iX_i^T(\beta-\beta_0) + o_p(1)\}\right|.$$

其中，

$$K_1 = \frac{G''(X_i^T\beta)\varepsilon_i - G'(X_i^T\beta)V'(G(X_i^T\beta))\varepsilon_i - G'^2(X_i^T\beta)}{V^2(G(X_i^T\beta))},$$

因为

$$\begin{aligned}
&T_n(\beta) - T_n(\beta_0) \\
&= -\frac{2}{n}\sum_{I=1}^n X_iX_i^TX_i^T[Y_i - G(X_i^T\beta_0)]\{\frac{G'^2(X_i^T\beta_0)V(G(X_i^T\beta_0)) + V'(G(X_i^T\beta_0))G'^2(X_i^T\beta_0)[Y_i - G(X_i^T\beta_0)]}{V^3(G(X_i^T\beta_0))} \\
&- \frac{G''(X_i^T\beta_0)G'(X_i^T\beta_0)[Y_i - G(X_i^T\beta_0)]}{V^2(G(X_i^T\beta_0))}\}(\beta-\beta_0) + o_p(1) \\
&= -\frac{2}{n}\sum_{i=1}^n K_2X_iX_i^T\{X_i^T(\beta-\beta_0)\}\varepsilon_i + o_p(1),
\end{aligned}$$

其中

$$K_2=\frac{G'^2(X_i^T\beta_0)V(G(X_i^T\beta_0))+V'(G(X_i^T\beta_0))G'^2(X_i^T\beta_0)\varepsilon_i}{V^3(G(X_i^T\beta_0))}-\frac{G''(X_i^T\beta_0)G'(X_i^T\beta_0)\varepsilon_i}{V^2(G(X_i^T\beta_0))},$$

由条件 A.4，得到

$$\theta^T T_n(\beta)\theta=\theta^T T_n(\beta_0)\theta+o_p(1)=O_p(1).$$

通过 Chen et al.[27] 的引理 2，因为 $E(X_i)=0$，得到

$$\max_i\|X_i\|=O_p\{p^{1/2}\}+o_p\{p^{-(2k-1)/(4k)}n^{1/(4k)}\},\beta\in D_n.$$

$$\max_i\left\|\frac{G'(X_i^T\beta)}{V(G(X_i^T\beta))}X_iX_i^T(\beta-\beta_0)\right\|\le\frac{b}{a}\cdot ca_n\max_i\|X_i\|^2=o_p(1).$$

由条件 A.2,

$$\max_i\|U_i(\beta)\|n^{-1}\left|\sum_{i=1}^n\theta^TU_i(\beta)\right|=o_p(1).$$

引理 4.1 意味着

$$\max_i\left|\lambda_\beta^TU_i(\beta)\right|=o_p(1).$$

引理 4.2 当 $n\to\infty$ 时，依概率 1，（2.4）中的 $\ell_p(\beta)$ 在 D_n 中有最小值.

证明： 由引理 4.1 和经验似然的定义，对于 $\beta\in D_n$，

$$0=Q_{1n}(\beta,\lambda_\beta)=n^{-1}\sum_{i=1}^n\frac{U_i(\beta)}{1+\lambda_\beta^TU_i(\beta)}$$

可以扩展为

$$0=n^{-1}\sum_{i=1}^nU_i(\beta)-\lambda_\beta^TT_n(\beta)+r_n,$$

其中的余项

$$r_n=n^{-1}\sum_{i=1}^nU_i(\beta)\{\lambda_\beta^TU_i(\beta)\}^2(1+\xi_i)^{-3},|\xi_i|\le\left|\lambda_\beta^TU_i(\beta)\right|.$$

这意味着

$$\lambda_\beta=T_n^{-1}(\beta)\overline{U}(\beta)+T_n^{-1}(\beta)r_n,$$

其中

$$\overline{U}(\beta)=n^{-1}\sum_{i=1}^{n}U_i(\beta).$$

将 λ_β 替代成 $\ell(\beta)$，构建方程

$$2\ell(\beta)=n\overline{U}(\beta)T_n^{-1}(\beta)\overline{U}(\beta)-nr_n^TT_n^{-1}(\beta)r_n+2/3\sum_{i=1}^{n}\{\lambda_\beta^TU_i(\beta)\}^3(1+\xi_i)^{-4}. \qquad (A.1)$$

对于 $\beta\in\partial D_n$，其中 ∂D_n 表示 D_n 的边界，记

$$\beta=\beta_0+ca_n\theta_\beta,$$

其中 θ_β 是单位向量，得到分解式

$$2\ell(\beta)=T_0+T_1+T_2,$$

其中

$$T_0=n\overline{U}^T(\beta)T_n^{-1}(\beta_0)\overline{U}(\beta_0),\ T_1=n(\beta-\beta_0)^TT_n^{-1}(\beta),$$

$$T_2=n\{\overline{U}^T(\beta_0)(T_n^{-1}(\beta)-T_n^{-1}(\beta_0))\overline{U}(\beta_0)-2\overline{U}^T(\beta_0)T_n^{-1}(\beta_0)\overline{U}(\beta_0)\}$$

$$-nr_n^TT_n^{-1}(\beta)r_n+2/3\sum_{i=1}^{n}\{\lambda_\beta^TU_i(\beta)\}^3(1+\xi_i)^{-4}.$$

当 $n\to\infty$ 时，可以得到

$$T_1=c^2na_n^2\theta_\beta T_n^{-1}(\beta)\theta_\beta\geq c^2na_n^2\gamma_p^{-1}(T_n(\beta))=O_p(na_n^2),na_n^2\to\infty,T_2/T_1\xrightarrow{p}0$$

且

$$2\ell(\beta_0)-T_0=o_p(1).$$

这意味着对于任意给定的 C，当 $n\to\infty$ 时，

$$P[\{2\ell(\beta)-2\ell(\beta_0)\}>C]\to1.$$

从而，当 n 足够大时，

$$\ell_p(\beta)-\ell_p(\beta_0)=\ell(\beta)-\ell(\beta_0)+n\sum_j\{p_\tau(|\beta_j|)-p_\tau(|\beta_{0j}|)\}\geq\ell(\beta)-\ell(\beta_0),$$

根据条件 A.5 和 SCAD 惩罚的性质，上式中的最后一个不等式成立. 惩罚项中 $j\in\Im$，当 n 足够大时，有 $p_\tau(|\beta_{0j}|)=p_\tau(|\beta_j|)$. 因此，依概率收敛于 1，$\ell_p(\beta)>\ell_p(\beta_0)$ 对于 $\beta\in\partial D_n$，引理 4.2 成立.

定理 2.1 的证明：

根据引理 4.2，$\ell_p(\beta)$ 中的 $\hat{\beta}$ 满足 $\left\|\hat{\beta}-\beta_0\right\| \le ca_n$. 再一次表明对于任何 $j \notin \Im$，依概率收敛于 1，

$$\frac{1}{n}\frac{\partial \ell_p(\beta)}{\partial \beta_j} < 0 \quad \text{对于} \ \beta_j \in (0, \varepsilon_n)$$

和

$$\frac{1}{n}\frac{\partial \ell_p(\beta)}{\partial (\beta_j)} > 0 \ \text{对于} \ \beta_j \in (-\varepsilon_n, 0) \ \text{和} \ \varepsilon_n = ca_n.$$

首先，得到

$$\frac{1}{n}\frac{\partial \ell_p(\beta)}{\partial \beta_j} = n^{-1}\sum_{i=1}^{n}\frac{G''(X_i^T\beta)^2}{V(G(X_i^T\beta))}\cdot\frac{-\lambda_\beta^T X_i X_{ij}}{1+\lambda_\beta^T U_i(\beta)} + p_\tau'(\left|\beta_j\right|)sign(\beta_j) =: I_j + II_j.$$

由引理 4.1 中 $\max_i\left|\lambda_\beta^T X_i\right| = o_p(1)$ 和条件 A.4，当 n 足够大时，得到

$$\begin{aligned}
&\max_{j\notin\Im}\left|I_j\right| \le \max_i\left|\lambda_\beta^T X_i\right| \max_{j\notin\Im}\left|n^{-1}\sum_{i=1}^{n}\frac{G'(X_i^T\beta)^2}{V(G(X_i^T\beta))}X_{ij}\right| \\
&\le \max_i\left|\lambda_\beta^T X_i\right|\left\|n^{-1}\sum_{i=1}^{n}X_i\right\|\left|\frac{G'(X_i^T\beta)^2}{V(G(X_i^T\beta))}\right| \\
&= o_p\{(p/n)^{1/2}\}.
\end{aligned}$$

这意味着

$$P(\max_{j\notin\Im}\left|I_j\right| > \tau/2) \to 0.$$

此外，注意到

$$\left|\beta_j\right| \le ca_n, \ j \notin \Im \quad \text{和} \quad \tau/a_n \to \infty.$$

因此当 n 足够大且 $j \notin \Im$ 时，有

$$p'(\left|\beta_j\right|) = \tau,$$

因此

$$II_j = p'(\left|\beta_j\right|)sign(\beta_j) = \tau sign(\beta_j).$$

所以，β_j 支配 $\partial\ell_p/\partial\beta_j$ 的渐近性，对于 $j \in \Im$. 即，当 $n \to \infty$ 时，对于任何 $j \notin \Im$，依概率

收敛于 1. 因此 $\hat{\beta}_2 = 0$ 依概率 1，且得到了定理 2.1 的第一部分.

接下来证明定理 2.1 的第二部分. 由定理 2.1 的第一部分和惩罚经验似然的定义，为简单起见，设 $\sigma^2 = 1$，因为 $\hat{\beta}_2 = 0$ 依概率收敛于 1，考虑 $\ell(\beta)$ 的最小化受限于 $H_2\beta = 0$. 通过拉格朗日乘子法，这相当于一个新的目标函数的最小化过程.

$$\tilde{\ell}(\beta,\lambda,\nu) = n^{-1}\sum_{i=1}^{n}\log\{1+\lambda^T U_i(\beta)\} + \sum_{j=1}^{d} p_\tau(|\beta_j|) + \nu^T H_2\beta. \tag{A.2}$$

由引理 4.1，$\beta \in D_n$，$\|\lambda\|$ 和 $\|\nu\|$ 都是 $o_p(1)$ 且

$$0 = \tilde{Q}_{1n}(\beta,\lambda,\nu) = n^{-1}\sum_{i=1}^{n} U_i(\beta)/\{1+\lambda^T U_i(\beta)\}.$$

定义

$$\tilde{Q}_{2n}(\beta,\lambda,\nu) = -n^{-1}\sum_{i=1}^{n}\frac{G'(X_i^T\beta)^2}{V(G(X_i^T\beta))}\cdot\frac{-\lambda_\beta^T X_i X_{ij}}{1+\lambda_\beta^T U_i(\beta)} + b(\beta) + H_2^T\nu$$

和

$$\tilde{Q}_{3n}(\beta,\lambda,\nu) = H_2\beta,$$

其中

$$b(\beta) = \{p'_\tau(|\beta_1|)sign(\beta_1), p'_\tau(|\beta_2|)sign(\beta_2),\ldots, p'_\tau(|\beta_d|)sign(\beta_d),0,\ldots,0\}^T.$$

在 $(\beta_0,0,0)^T$ 处展开 $\tilde{Q}_{jn}(\beta,\lambda,\nu)$ 如下：

$$\begin{pmatrix} -\tilde{Q}_{1n}(\beta_0,0,0) \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -\tilde{\Sigma} & -\tilde{\Sigma} & 0 \\ -\tilde{\Sigma} & 0 & H_2^T \\ 0 & H_2 & 0 \end{pmatrix}\begin{pmatrix} \hat{\lambda}-0 \\ \hat{\beta}-\beta_0 \\ \hat{\nu}-0 \end{pmatrix} + R_n, \tag{A.3}$$

其中

$$\tilde{\Sigma} = \frac{G'(X_i^T\beta)^2}{V(G(X_i^T\beta))}\Sigma,$$

$$R_n = \sum_{k=1}^{5} R_n^{(k)}, R_n^{(1)} = (R_{1n}^{T(1)}, R_{2n}^{T(1)},0)^T, R_{1n}^{(1)} \in \Re^p,$$

且

$$R_{2n}^{(1)} \in \Re^p.$$

且 $R_{jn}^{(1)}, j=1,2,$ 的第 k 个元被定义为

$$R_{jn,k}^{(1)}=\frac{1}{2}(\hat{\eta}-\eta_0)^T\frac{\partial^2 Q_{jn,k}(\eta^*)}{\partial\eta\partial\eta^T}(\hat{\eta}-\eta_0),$$

$$\eta=(\beta,\lambda)^T,\quad \eta^*=(\beta^*,\lambda^*)$$

满足

$$\left\|\beta^*-\beta_0\right\|\le\left\|\hat{\beta}-\beta_0\right\|\text{和}\left\|\lambda^*\right\|\le\left\|\hat{\lambda}\right\|,$$

$$R_n^{(2)}=\{0,b^T(\beta_0),0\}^T,$$

$$R_n^{(3)}=\{0,\{b'(\beta_0)(\hat{\beta}-\beta_0)\}^T,0\}^T,$$

$$R_n^{(4)}=\{\{(T_n(\beta_0)-\tilde{\Sigma})\hat{\lambda}\}^T,0,0\}^T$$

和

$$R_n^{(5)}=\{\{(T_n(\beta_0)-\tilde{\Sigma})\hat{\lambda}\}^T,\{(T_n(\beta_0)-\tilde{\Sigma})(\hat{\beta}-\beta_0)\}^T,0\}^T.$$

根据 Chen et al.[27] 中的定理 3，在条件 A.2 下有

$$\left\|R_n^{(1)}\right\|=o_p\{(1/n)^{-1/2}\},$$

和在条件 A.6 中 SCAD 惩罚函数的定义下，可以看到

$$R_n^{(k)}=o_p\{(1/n)^{-1/2}\},k=1,\ldots,4.$$

因为

$$\left\|T_n(\beta_0)-\tilde{\Sigma}\right\|=O_p\{(p/n)^{1/2}\},\left\|\hat{\beta}-\beta_0\right\|=O_p(a_n)$$

和

$$\left\|\hat{\lambda}_\beta\right\|=O_p(a_n),\left\|R_n^{(5)}\right\|=o_p\{(1/n)^{-1/2}\}.$$

定义

$$K_{11}=-\tilde{\Sigma},K_{12}=(-\tilde{\Sigma},0),K_{21}=K_{12}^T,K_{22}=\begin{pmatrix}0 & H_2^T\\ H_2 & 0\end{pmatrix}$$

和

$$K=\begin{pmatrix}K_{11} & K_{12}\\ K_{21} & K_{22}\end{pmatrix},$$

设$\upsilon=(\beta^T,\nu^T)^T$, 由(A.3)的逆，有

$$\begin{pmatrix}\hat{\lambda}-0\\ \hat{\upsilon}-\upsilon_0\end{pmatrix}=K^{-1}\{\begin{pmatrix}-\tilde{Q}_{1n}(\beta_0,0,0)\\ 0\\ 0\end{pmatrix}+R_n\},\tag{A.4}$$

其中

$$R_n=\sum_{k=1}^{5}R_n^{(k)}$$

和

$$\|R_n\|\le\sum_{i=1}^{5}\|R_n^{(k)}\|=o_p\{(1/n)^{-1/2}\}.$$

应用分块矩阵的逆，得到

$$K^{-1}=\begin{pmatrix}K_{11}^{-1}&0\\0&0\end{pmatrix}+\begin{pmatrix}-K_{11}^{-1}K_{12}\\I\end{pmatrix}A^{-1}\begin{pmatrix}-K_{21}K_{11}^{-1}I\end{pmatrix},$$

其中

$$A=K_{22}-K_{21}K_{11}^{-1}K_{12}=\begin{pmatrix}\tilde{\Sigma}&H_2^T\\H_2&0\end{pmatrix}.$$

且

$$\hat{\upsilon}-\upsilon_0=A^{-1}K_{21}K_{11}^{-1}\tilde{Q}_{1n}(\beta_0,0,0)+o_p\{(1/n)^{-1/2}\},$$

其中

$$A^{-1}=\begin{pmatrix}\tilde{\Sigma}^{-1}-\tilde{\Sigma}^{-1}H_2^T(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}H_2\tilde{\Sigma}^{-1}&\tilde{\Sigma}^{-1}H_2^T(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}\\(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}H_2\tilde{\Sigma}^{-1}&-(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}\end{pmatrix}.$$

这意味着

$$\hat{\beta}-\beta_0=\{\tilde{\Sigma}^{-1}-\tilde{\Sigma}^{-1}H_2^T(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}H_2\tilde{\Sigma}^{-1}\}(\overline{U}(\beta)+R_{1n}),\tag{A.5}$$

其中R_{1n}是向量$K^{-1}R_n$中相应的元，且

$$\|R_{1n}\|=o_p\{(1/n)^{-1/2}\}.$$

β_1中的非零元可以被扩展为

$$\hat{\beta}-\beta_{10}=\{H_1\tilde{\Sigma}^{-1}-H_1\tilde{\Sigma}^{-1}H_2^T(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}H_2\tilde{\Sigma}^{-1}\}(\overline{U}(\beta)+R_{1n}). \quad (A.6)$$

设

$$I_B=H_1\tilde{\Sigma}^{-1}H_1^T-H_1\tilde{\Sigma}^{-1}H_2^T(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}H_2\tilde{\Sigma}^{-1}H_1^T, Y_{ni}=\frac{1}{\sqrt{n}}Z_{ni},$$

其中

$$Z_{ni}=W_nI_B^{-1/2}\{H_1\tilde{\Sigma}^{-1}-H_1\tilde{\Sigma}^{-1}H_2^T(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}H_2\tilde{\Sigma}^{-1}\}(U_i(\beta)+R_{1n}).$$

可以直接证明出

$$P(\|Y_{ni}\|>\varepsilon)\le n^{-1}\varepsilon^{-2}E\|Z_{ni}\|^2=O(1/n), E(\|Y_{ni}\|^4)=n^{-2}E(Z_{ni}^TZ_{ni})^2=O(p^2/n^2).$$

因此，

$$\sum_{i=1}^{n}E\|Y_{ni}\|^2I(\|Y_{ni}\|>\varepsilon)\le n\sqrt{E\|Y_{n1}\|^4}\sqrt{P(\|Y_{n1}\|>\varepsilon)}\to 0.$$

当$W_nW_n^T\to T$时，由 Lindeberg-Feller's 中心极限定理（Van Der Vaart[28]），有

$$\sqrt{n}W_nI_B^{-1/2}\{H_1\tilde{\Sigma}^{-1}-H_1\tilde{\Sigma}^{-1}H_2^T(H_2\tilde{\Sigma}^{-1}H_2^T)^{-1}H_2\tilde{\Sigma}^{-1}\}\overline{U}(\beta)\xrightarrow{d}N(0,T).$$

最后，由

$$\left\|\sqrt{n}W_nI_B^{-1/2}H_1R_{1n}\right\|^2\le n\gamma_d(W_nW_n^T)\gamma_1^{-1}(I_B)\gamma_1^{-2}(k)\|R_n\|^2=o_p(1).$$

这样就完成了定理 2.1 的证明.

定理 2.2 的证明：

首先，我们给出$\ell(\hat{\beta})$的渐近展式，其中$\hat{\beta}$是（2.4）的最小值，即用经验似然比估计惩罚经验似然的估计值. 设$Z_i=\hat{\lambda}^TU_i(\beta)$.

由引理 4.1 有

$$\max_i\left|\hat{\lambda}^TU_i(\beta)\right|=o_p(1),$$

根据泰勒展式和定理 2.1 中的证明结果，得到

$$\ell(\hat{\beta})=\sum_{i=1}^{n}Z_i-\sum_{i=1}^{n}Z_i^2/2+\sum_{i=1}^{n}Z_i^3/\{3(1+\xi_i)^4\}+o_p(1) \quad (A.7)$$

其中$|\xi_i|<\left|\hat{\lambda}^TU_i(\beta)\right|$，且由惩罚函数得到$o_p(1)$. 根据引理 4.2，得到$\beta\in D_n$的展式为

$$\lambda_\beta = T_n^{-1}(\beta)\overline{U}(\beta) + T_n^{-1}(\beta)r_n,$$

其中

$$r_n = n^{-1}\sum_{i=1}^{n}[U_i(\beta)\{\lambda_\beta^T U_i(\beta)\}^2(1+\xi_i)^{-3}]$$

和

$$|\xi_i| \le |\lambda_\beta^T U_i(\beta)|,$$

用 Z_i 代替（A.5）中 $\hat{\lambda}$ 和 $\hat{\beta}$ 的展式，得到

$$2\ell(\hat{\beta}) = n\overline{U}^T(\beta)H_2^T(H_2\tilde{\Sigma}H_2^T)^{-1}H_2\overline{U}(\beta) + o_p(1) \qquad (A.8)$$

其次，将 $\ell_p(\hat{\mu})$ 在无效假设 $H_0 : L_n\beta_1 = 0$ 下展开，因为 $L_nL_n^T = I_q$, 存在 $\tilde{H}_2$ 使得

$$\tilde{H}_2\beta = 0 \quad 和 \quad \tilde{H}_2\tilde{H}_2^T = I_{p-d+q}.$$

定义

$$\overline{U}(\beta) = n^{-1}\sum_{i=1}^{n}\frac{G'(X_i^T\beta)}{V(G(X_i^T\beta))}X_i\varepsilon_i,$$

接下来重复定理 2.1 的证明，在无效假设条件下构建 β 的估计量可以获得(A.9)的最小值

$$\tilde{\ell}_p(\beta,\lambda,\nu) = \sum_{i=1}^{n}\log\{1+\lambda^T U_i(\beta)\} + n\sum_{i=1}^{n}p_\tau(|\beta_i|) + \nu^T\tilde{H}_2\beta. \qquad (A.9)$$

表明（A.9）的最小值为 $(\breve{\beta},\breve{\lambda},\breve{\nu})$, 根据定理 2.1 的第一部分，依概率 1 有 $\breve{\beta} = 0.$

所以，依概率 1 有

$$n\{\sum_{i=1}^{p}p_\tau(|\hat{\beta}_j|) - \sum_{i=1}^{p}p_\tau(|\breve{\beta}_j|)\} = 0,2$$

由 H_2 在(A.3)中替换为 $\tilde{H}_2$ 和（A.7）的展式，构建

$$2\ell(\hat{\beta})_{L_n\beta_1=0} = 2\ell(\breve{\beta}) = n\overline{U}^T(\beta)\tilde{H}_2^T(\tilde{H}_2\tilde{\Sigma}\tilde{H}_2^T)^{-1}\tilde{H}_2\overline{U}(\beta) + o_p(1). \qquad (A.10)$$

结合（A.8）和（A.10），得到

$$\tilde{\ell}(L_n) = n\overline{U}^T(\beta)\tilde{\Sigma}^{-1/2}\{P_1 - P_2\}\tilde{\Sigma}^{-1/2}\overline{U}(\beta) + o_p(1)$$

其中

$$P_1 = \tilde{\Sigma}^{1/2}\tilde{H}_2^T(\tilde{H}_2\tilde{\Sigma}\tilde{H}_2^T)^{-1}\tilde{H}_2\tilde{\Sigma}^{1/2} \quad 和 \quad P_2 = \tilde{\Sigma}^{1/2}H_2^T(H_2\tilde{\Sigma}H_2^T)^{-1}H_2\tilde{\Sigma}^{1/2}.$$

由于 $P_1 - P_2$ 是一个秩为 q 的幂等阵， $P_1 - P_2$ 可以记作 $\Xi_n^T\Xi_n$ ，其中 Ξ_n 是 $q \times p$ 矩阵. 使得 $\Xi_n^T\Xi_n = I_q$ （Fan and Li[7]）. 直接得到

$$\sqrt{n}\Xi_n\tilde{\Sigma}^{-1/2}\overline{U}(\beta) \to N(0, I_q).$$

那么

$$n\overline{U}^T(\beta)\tilde{\Sigma}^{-1/2}(P_1 - P_2)\tilde{\Sigma}^{-1/2}\overline{U}(\beta) \xrightarrow{d} \chi_q^2.$$

这样就完成了定理 2.2 的证明.

结　论

本文主要的结果是在广义线性模型下研究 SCAD 惩罚经验似然方法，在适当的条件下，针对带有发散参数的参数估计和变量选择问题提出了惩罚经验似然方法. 通过选取适当的惩罚函数，可以依概率 1 选择非零系数和非零系数的估计值是渐近正态的，也就是 Oracle 性质.

当前，广义线性模型变量选择的方法研究仍旧不完善，仍然有许多问题有待继续深入研究. 理论分析和数值结果说明在广义线性模型下，惩罚经验似然是一个很有效的方法. 这种方法也可以推广到其他的半参数模型，如：单指标模型，变系数模型等. 在以后的研究学习中可以进行这方面的讨论.

参 考 文 献

[1] Nelder J A, Wedderburn R W M. Generalized linear models[J]. Journal of the Royal Statistical Society, 1972, 135:370-384.

[2] Owen A. Empirical likelihood ratio confidence intervals for a single function[J]. Biometrika. 1988, 75:237-249.

[3] Owen A B. Empirical likelihood ratio confidence regions. Ann. Statist, 1990, 18(1): 90-120.

[4] Wedderburn R W M. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. 1974, Biometrika 61, 439-447.

[5] McCullagh P, Nelder J A. Generalized Linear Models. Chapman and Hall, London. 1990.

[6] Zhao L C, Krishnaiah P R and Bai z D, On Rates of Convergence of Efficient Detection Criteria in Signal Processing with White Noise IEEE Transactions on Infornation Theory, Vol 35, NO. 2, 1989, 380-388.

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

[8] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B, 1996, 58:267-288.

[9] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools (with discussion), Technometrics. 1993, 35:109-148.

[10] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Annals of Stistics. 2004, 32: 407-499.

[11] Fan, J., Peng, H., Nonconcave penalized likelihood with a diverging number of parameters. The Annals of Statistics. 2004, 32:928-961.

[12] Kim Y, Choi H, Oh H, Smoothly clipped absolute deviation on high dimensions. Journal of the American Statistical Association. 2008a, 103: 1656-1673.

[13] Knight K, Fu W J, 2000. Asymptotics for lasso-type estimators. The Annals of Statistics 28, 1356-1378.

[14] Zou H, 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418-1429.

[15] Hastie T, Tibshirani R and Friedman J. (2009). The Elements of Statistical Learning Data Mining, Inference, and Prediction. 2nd Edition. Springer-New York.

[16] Fan, J. and Lv, J. Sure independence screening for ultrahigh dimensional feature space (with discussion). Journal of the Royal Statistical Society. 2008, Series B 70: 849-911.

[17] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, 1996, Series B 58:267-288.

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

[19] Zhang H H and Lu W. Adaptive lasso for Cox's proportional hazard model. 2007, Biometrika 94 691-703.

[20] Candes, E. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). The Annals of Statistics . 2007, 35:2313-2351.

[21] Hjort N L., McKeague I, and Van Keilegom I. Extending the scope of empirical likelihood. The Annals of Statistics. 2009, 37:1079-1111.
[22] Lv J and Fan Y. A unified approach to model selection and sparse recovery using regularized least squares. The Annals of Statistics.2009, 37: 3498-3528.
[23] Tang C Y, Leng C L. Penalized high-dimensional empirical likelihood[J]. Biometrika. 2010, 97:905-920.
[24] Owen A. Empirical Likelihood[M]. Chapman and Hall-CRC. United states of America. 2001.
[25] Cameron A C, Trivedi P K. Regression Analysis of Count Data[M], New York:Cambridge University Press. 1998.
[26] Variyath A M, Chen J and Abraham B. Empirical Likelihood Based Variable Selection[J]. Journal of Statistical Planning and Inference. 2010, 140: 971-981.
[27] Chen S X, Peng L and Qin Y-L. Effects of Data Dimension on Empirical Likelihood[J]. Biometrika. 2009, 96: 711-722.
[28] Van Der Vaart A W. Asymptotic Statistics[M], New York: Cambridge University Press. 1998.

攻读硕士学位期间发表学术论文情况

[1] 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，2012，35（1）：6-9.

[2] 刘琦.广义线性模型的惩罚高维经验似然估计[J].辽宁师范大学学报（增刊），2012，1：33-37.

致　　谢

光阴似箭，三年的研究生学习生活已经接近尾声，在这里我渡过了人生中最美好的时光．首先，我要衷心的感谢我的导师侯文副教授，侯老师渊博的学识、敏锐的思维、严谨的学风、诲人不倦的高尚师德，对我影响深远并给了我诸多的促进和鼓励．在生活上，也给予了我很大的关心和帮助，在此，谨向导师表示崇高的敬意和衷心的感谢！谢谢您对我的谆谆教导和不倦关怀！

辽宁师范大学为我提供了良好的学习环境，对此表示深深的感谢！三年的学习生活一晃而过，但生命不止，奋斗不息．这是个终点，也是一个起点，我要把握好未来，在新的工作、生活中继续前进．

最后，我要深深感谢我的父母，感谢他们的养育之恩，以及在生活上无微不至的关心爱护，同时也谢谢他们对我的理解和支持．他们的祝福和鼓励一直伴随着我，增加了我专心致志投入学习和研究中的决心和信心．