

SCAD-L2正则化下广义估计方程的性质分析

刘玥佳, 赵慧秀*

南京理工大学数学与统计学院, 江苏 南京

收稿日期: 2025年12月30日; 录用日期: 2026年1月21日; 发布日期: 2026年2月2日

摘要

广义估计方程(Generalized estimating equations, GEE)因能够有效处理个体内相关性而在纵向数据分析中得到广泛应用。然而, 当纵向数据的协变量高度相关时, 传统变量选择方法往往面临变量选择不稳定的问题。本文将SCAD-L2正则化项融入GEE框架中, 以实现变量选择与参数估计的双重优化。随后, 本文提出一种适用于多重共线性纵向数据的初始值选择策略, 即使用 L_2 惩罚下的GEE估计值作为计算初始值。最后, 本文研究了SCAD-L2惩罚下GEE估计的大样本渐近性质。模拟实验表明, 该方法在纵向数据的参数估计与变量选择中显著优于现有方法, 为复杂相关结构下的纵向数据建模提供了有效方法。

关键词

广义估计方程, 变量选择, 纵向数据分析, Oracle性质, SCAD-L2

Analysis of the Properties of Generalized Estimating Equations under SCAD-L2 Regularization

Yuejia Liu, Huixiu Zhao*

School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing Jiangsu

Received: December 30, 2025; accepted: January 21, 2026; published: February 2, 2026

Abstract

Generalized estimating equations (GEE) have been widely used in longitudinal data analysis due to their ability to effectively account for within-subject correlation. However, when covariates in longitudinal data are highly correlated, traditional variable selection methods often suffer from instability. In this paper, we incorporate the SCAD-L2 regularization into the GEE framework to simulta-

*通讯作者。

neously optimize variable selection and parameter estimation. We then propose an initial value selection strategy for longitudinal data with multicollinearity, which uses the GEE estimator under an L2 penalty as the starting value for computation. Finally, we investigate the large-sample asymptotic properties of the SCAD-L2 penalized GEE estimator. Simulation studies show that the proposed method substantially outperforms existing approaches in both parameter estimation and variable selection, providing an effective tool for modeling longitudinal data with complex correlation structures.

Keywords

Generalized Estimating Equations, Variable Selection, Longitudinal Data Analysis, Oracle Property, SCAD-L2

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着数据来源的不断丰富, 数据维度呈现爆炸式增长, 尤其是在生物医学研究、基因组学和公共卫生等领域, 研究数据越来越多地表现出高维纵向的特征。为了处理对同一受试者进行重复测量所产生的复杂相关结构, Liang 等[1]提出了广义估计方程(generalized estimating equations, GEE), 该方法已成为纵向数据分析的基础工具, 具有较强灵活性, 被广泛应用于各领域的纵向数据研究中[2]。Wang [3]针对具有大规模协变量的聚类二元数据, 系统建立了 GEE 估计量的存在性、一致性与渐近正态性。Xie 等[4]则进一步将 GEE 的渐近理论推广到每个受试者观测次数趋于无穷的情形, 拓展了其理论的适用范围。

在高维纵向数据中, 有效的特征选择能够降低模型复杂度与计算负担, 有助于更深入地理解数据趋势, 从而提高预测精度, 并揭示时间动态性与个体异质性[5] [6]。常见的变量选择方法包括过滤法、包裹法和嵌入法[5] [7], 其中嵌入法能够将变量选择直接融入模型估计过程, 在高维建模中得到广泛应用。纵向数据的特征选择通常基于 GEE 框架引入正则化项(如 LASSO [8]、SCAD [9]或自适应 LASSO [10]), 以同步实现特征选择与参数估计[11]。在此基础上, Xu 等[12]针对参数维数发散条件下的相关二元数据, 引入收缩惩罚并系统分析了正则化 GEE 估计量的一致性与变量选择性质, 为高维纵向二元数据的特征选择提供了重要的理论依据。

近年来, “大 n , 发散 p ” (large- n , diverging- p) 框架下的高维理论研究受到广泛关注[13], 其主要考察协变量维度随样本量趋于无穷时的变量选择问题。Wang 等[14]针对参数维数随样本量增长的广义线性模型, 研究了 Bridge 惩罚下估计量一致性与变量选择一致性, 为高维正则化方法在发散维度情形下的理论分析提供了重要参考。Wang 等[6]研究了 “大 n , 发散 p ” 框架下的惩罚化 GEE (penalized GEE, PGEE), 证明其具有良好的大样本渐近性质与 Oracle 性质[9] [13], 但在多重共线性情形下难以平衡稀疏性与稳定性。

为了解决多重共线性问题, Zou 等[15]提出的 Elastic net 通过结合 L_1 与 L_2 惩罚促进相关变量的联合选择, 随后 SCAD-L2 [16]与自适应 Elastic net [17]等扩展方法相继被提出, 在保持估计稳定性的同时提升了变量选择一致性, 使其更适用于强相关协变量下的高维稀疏建模。在此基础上, Blommaert 等[18]在正态情形且协变量维数固定的设定下, 通过引入非凸惩罚以应对纵向数据的共线性问题。此外, Lin 等[19]提出的 GPGEE 将组惩罚机制引入 GEE, 可以实现已知分组信息时纵向数据的变量选择, 但该方法无法

主动识别纵向数据中隐含的特征关联, 在实际应用中存在局限。

为此, 本文首先将 SCAD-L2 正则化引入 GEE 框架, SCAD-L2 正则化中 SCAD 惩罚能够在实现变量选择的同时减少对大系数的惩罚偏差, 从而提高变量选择的准确性; L_2 惩罚则增强了模型在多重共线性下的数值稳定性, 有助于获得更为稳健的参数估计。因此, SCAD-L2 惩罚能够在稀疏性与稳定性之间取得更好的平衡, 特别适用于协变量高度相关的高维纵向数据。随后, 本文在“大 n , 发散 p ”框架下, 研究了该方法的参数估计一致性与 Oracle 性质。最后通过模拟研究验证了该方法在估计精度、变量选择准确性及稳定性等方面的综合优势。

2. 模型与方法

2.1. 基于 SCAD-L2 的高维 GEE 方法

2.1.1. 方法框架与理论基础

GEE 是一种用于纵向数据的估计方法, 其核心思想是通过刻画响应变量的边际均值函数, 并结合工作相关矩阵近似描述组内相关结构, 从而获得稳健且有效的参数估计。对于第 i 个个体的第 j 次观测, 记响应变量为 Y_{ij} , 协变量为 X_{ij} , 其中 $i=1, \dots, n$, $j=1, \dots, m_i$ 。这里, m_i 表示个体 i 的重复测量次数, 为简化理论推导, 通常假设各个体的测量次数相同, 即 $m_i = m$; 协变量的维数记为 p ; 因此可定义个体 i 的响应向量为 $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{im})^T$, 协变量矩阵为 $X_i = (X_{i1}, X_{i2}, \dots, X_{im})^T$ 。GEE 具体表示为

$$\begin{aligned} S(\beta) &= \frac{1}{n} \sum_{i=1}^n D_i^T(\beta) V_i^{-1}(\beta) (Y_i - \mu_i(\beta)) \\ &= \frac{1}{n} \sum_{i=1}^n X_i^T A_i^{1/2}(\beta) \hat{R}^{-1}(\alpha) A_i(\beta)^{-1/2} (Y_i - \mu_i(\beta)) \\ &= 0, \end{aligned} \quad (1)$$

其中 $D_i(\beta)$ 表示个体均值函数对参数的导数矩阵, $V_i(\beta)$ 为个体内的工作协方差矩阵。在指数族分布的假设下 $D_i(\beta) = A_i(\beta) X_i$, $V_i(\beta) = A_i^{1/2}(\beta) R(\alpha) A_i^{1/2}(\beta)$, $A_i(\beta) = \text{diag}\{\sigma_{i1}^2, \sigma_{i2}^2, \dots, \sigma_{im}^2\}$ 是一个由边际均值 $\mu_i(\beta)$ 决定的对角方差函数矩阵, $R(\alpha)$ 是由参数 α 控制的工作相关矩阵, 用于近似同一簇内的相关结构。

在高维情形下, 协变量个数可能随着样本量的增加而同步增长, 因此对估计方程的求解与数值实现提出更高要求。为此, Wang 等[8]提出了 PGEE, 通过在估计方程中引入稀疏惩罚同时实现参数估计与变量选择。其统一形式为

$$U(\beta) = S(\beta) - q_\lambda(\beta) = 0,$$

其中 $q_\lambda(\cdot)$ 为惩罚函数的一阶导数。

基于上述建模框架, 本文进一步将 SCAD-L2 组合惩罚引入 GEE 框架, 其估计方程形式为

$$U(\beta) = S(\beta) - q_{\lambda_1, \lambda_2}(\beta) = 0, \quad (2)$$

其中 $S(\beta)$ 来自公式(1), 惩罚函数导数部分为 SCAD-L2 惩罚函数的导数, 形如

$$q_{\lambda_1, \lambda_2}(\beta) = f'_{\lambda_1, \text{SCAD}}(\beta) + f'_{\lambda_2, L_2}(\beta),$$

其中,

$$\begin{aligned} f'_{\lambda_1, \text{SCAD}}(\theta) &= \lambda_1 \left\{ I(\theta \leq \lambda_1) + \frac{(a\lambda_1 - \theta)_+}{(a-1)\lambda_1} I(\theta > \lambda_1) \right\}, \\ f'_{\lambda_2, L_2}(\theta) &= \lambda_2 \theta, \end{aligned}$$

此处 α , λ_1 , λ_2 为调节参数, 本文遵循 Fan [13] 的推荐取 $\alpha = 3.7$, λ_1 , λ_2 的选择通过网格搜索与交叉验证来确定。

2.1.2. 参数估计与迭代算法

为实现 SCAD-L2 惩罚下 GEE 的求解, 本文遵循 PGEE 的算法框架, 将牛顿-拉夫逊方法(Newton-Raphson method)与极小化-极大化(Majorization-Minimization)算法相结合[20], 形成一种稳定高效的混合求解策略。具体的参数更新方程为

$$\hat{\beta}^k = \hat{\beta}^{k-1} + \left[H(\hat{\beta}^{k-1}) + E(\hat{\beta}^{k-1}) \right]^{-1} \left[S(\hat{\beta}^{k-1}) - E(\hat{\beta}^{k-1}) \right],$$

其中 $H(\hat{\beta}^{k-1}) = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{A}_i^{1/2}(\hat{\beta}^{k-1}) \hat{\mathbf{R}}^{-1} \mathbf{A}_i^{1/2}(\hat{\beta}^{k-1}) \mathbf{X}_i$ 为未加惩罚的 GEE 在当前参数估计值处的局部二阶近似,

而对角矩阵 $E(\hat{\beta}^{k-1}) = \text{diag} \left\{ \frac{q_{\lambda_1, \lambda_2}(|\hat{\beta}_1| + \epsilon)}{\epsilon + |\hat{\beta}_1|}, \dots, \frac{q_{\lambda_1, \lambda_2}(|\hat{\beta}_p| + \epsilon)}{\epsilon + |\hat{\beta}_p|} \right\}$ 用于控制上界, 此处 $|\hat{\beta}_j| + \epsilon$ 表示略大于 $|\hat{\beta}_j|$ 的值,

$\epsilon > 0$ 为任意给定的小正数常数, 通常取 10^{-6} , 该设计用于规避除以零的情形以保障数值稳定性。

2.1.3. 稳健性初始值选择

在迭代估计过程中, 常用的策略是使用独立工作相关结构假设下的 GEE 估计值作为迭代初始值, 但在多重共线性情形下信息矩阵接近奇异, 导致该初始值方差膨胀甚至无法求解。为增强初始估计的数值稳定性并促进算法收敛, 本文采用 L_2 正则化方法构建参数初始估计, 该方法能够有效缓解共线性导致的不适定性, 并为后续的非凸惩罚估计提供稳健的初始点。初始值估计方程形如

$$\tilde{U}_n(\beta_n) = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T (\mathbf{Y}_i - \mu(\beta_n)) - \lambda_{n2} \beta_n = 0 \quad (3)$$

2.1.4. 计算开销与双参数调参复杂度

与仅含 SCAD 罚项的 PGEE 相比, SCAD-L2 惩罚下 GEE 中叠加了 L_2 正则化对应的 Ridge 项, 该项等价于在更新方程中加入对角型的稳定化项或对参数向量施加线性收缩, 其边际计算成本远小于构造与求解涉及 p 维参数的线性系统及工作相关结构更新等核心步骤, 因此, 在固定调参下模型拟合中, 额外计算负担一般较为有限。

SCAD-L2 惩罚下 GEE 的额外计算开销主要源于调参空间由一维扩展至二维所导致的模型重复拟合次数增加。对 PGEE 而言, 惩罚选择通常集中在 λ_1 网格上, 采用 B 折交叉验证的总体模型拟合次数约为 $B \cdot L_{\lambda_1}$ 。但在同时包含稀疏化 SCAD 与稳定化 L_2 两类惩罚强度时, 需要在二维网格 (λ_1, λ_2) 上进行搜索, 使得总体拟合次数增加为 $B \cdot L_{\lambda_1} \cdot L_{\lambda_2}$, 该“乘法型”增长使得在 B 较大或网格较密时, 总耗时会较 PGEE 明显上升。

综合来看, 尽管 SCAD-L2 惩罚下 GEE 因采用基于 L_2 惩罚 GEE 的初始值而增加了一次前置拟合开销, 但在强多重共线性情形下, 该初始值可显著提升数值稳定性, 减少非凸惩罚迭代中的失败重启、震荡与额外迭代, 从而以小幅前置代价换取更稳健的收敛。

2.2. 大样本渐近性质

本节在“大 n , 发散 p ”框架下研究 SCAD-L2 惩罚下 GEE 方法的变量选择与参数估计的渐近性质。记真实的系数向量为 $\beta_{n0} = (\beta_{n10}^T, \beta_{n20}^T)^T$, 其中 $\beta_{n20} = 0$, β_{n10} 对应维度为 s_n 的非零系数。以下给出建立渐近性质所需的正则性条件:

(A1) 存在有限常数 $C_X > 0$, 使得 $\sup_{i,j} \|\mathbf{X}_{ij}\| \leq C_X$ 。

(A2) 真实参数 $\boldsymbol{\beta}_{n0}$ 严格位于紧集 $\mathcal{B} \subset \mathbb{R}^{p_n}$ 内; 估计量 $\boldsymbol{\beta}_n$ 也被限制在 $\mathcal{B}$ 中。

(A3) 存在固定常数 $0 < \kappa_1 \leq \kappa_2 < \infty$, 使得

$$\kappa_1 \leq \lambda_{\min} \left(n^{-1} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{R}_0^{-1} \mathbf{X}_i \right) + \lambda_{n2} \leq \lambda_{\max} \left(n^{-1} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{R}_0^{-1} \mathbf{X}_i \right) + \lambda_{n2} \leq \kappa_2。$$

(A4) 真实的簇内相关矩阵 $\mathbf{R}_0$ 与固定的参考矩阵 $\bar{\mathbf{R}}$ 均为正定矩阵, 且其特征值均被正数和有限常数所界定, 即远离 0 和 $+\infty$ 。此外, $\|\hat{\mathbf{R}}^{-1} - \bar{\mathbf{R}}^{-1}\| = O_p(\sqrt{s_n/n})$, 该式表明估计矩阵 $\hat{\mathbf{R}}^{-1}$ 与参考矩阵 $\bar{\mathbf{R}}^{-1}$ 的差距在概率意义下以 $\sqrt{s_n/n}$ 的速率收敛。

(A5) 存在常数 $c_1 > 0$ 和 $\delta > 0$, 对所有 i , 有 $E(\|\epsilon_i(\boldsymbol{\beta}_{n0})\|^2) \leq c_1$, 其中 $\epsilon_i(\boldsymbol{\beta}) = \mathbf{A}_i(\boldsymbol{\beta})^{-1/2}(\mathbf{Y}_i - \mu_i(\boldsymbol{\beta}))$ 。

(A6) $s_n(\log n)^4 = o(n\lambda_{n1}^2)$, $\log(p_n) = o(n\lambda_{n1}^2/(\log n)^2)$ 且 $p_n s_n^4 (\log n)^6 = o(n^2 \lambda_{n1}^2)$ 。

正则性条件(A1)~(A6)在很大程度上继承了 PGEE 的设定并做出了必要修正, 这些条件共同保证了高维纵向数据下估计量选择一致性、估计一致性与渐近正态性。特别地, (A3)是为应对多重共线性引入的关键修正: 与 PGEE 要求信息矩阵最小特征值有正下界不同, 当协变量高度相关时, 信息矩阵可能接近奇异, 故而通过引入 L2 惩罚项 λ_{n2} 作为扰动来保证理论分析中的稳定下界。此外, (A5)对标准化残差施加了矩条件, 用于在高维设定下保证中心极限定理的适用性并控制随机误差项的尾部行为。

性质 1:

假设满足正则化条件(A1)~(A3), 当 $n \rightarrow \infty$ 时, 满足 $p_n^2/n \rightarrow 0$ 且 $\lambda_{n2} = O(n^{-1/2})$, 则初始值估计方程 $\tilde{U}_n(\boldsymbol{\beta}_n) = 0$ (3)存在解 $\tilde{\boldsymbol{\beta}}_n$, 且该解满足:

$$\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_p(\sqrt{p_n/n} + \lambda_{n2}\boldsymbol{\beta}_{n0}) \quad (4)$$

定理 1:

假设满足正则化条件(A1)~(A5), 当 $n \rightarrow \infty$ 时, 满足 $p_n^2/n \rightarrow 0$, $\lambda_{n1} \rightarrow 0$ 且 $\lambda_{n2} = O(n^{-1})$, 则方程 $U_n(\boldsymbol{\beta}) = 0$ (2)存在解 $\hat{\boldsymbol{\beta}}_n$, 且该解满足:

$$n^{-1}p_n^3 = o(1) \quad (5)$$

定理 2:

假设满足正则化条件(A1)~(A6), 若 $n^{-1}p_n^3 = o(1)$, $\lambda_{n1} \rightarrow 0$ 且 $\lambda_{n2} = O(n^{-1})$, 记 $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{n1}^T, \hat{\boldsymbol{\beta}}_{n2}^T)^T$ 为方程 $U_n(\boldsymbol{\beta}_n) = 0$ 的解, 以下结论成立:

(1) $P\left(\left|U_{nj}(\hat{\boldsymbol{\beta}}_n)\right| = 0, j=1, \dots, s_n\right) \rightarrow 1$, $P\left(\left|U_{nj}(\hat{\boldsymbol{\beta}}_n)\right| \leq \frac{\lambda_{n1}}{\log n}, j=s_n+1, \dots, p_n\right) \rightarrow 1$;

(2) $P(\hat{\boldsymbol{\beta}}_{n2} = 0) \rightarrow 1$;

(3) 对于任意满足 $\|\alpha_n\| = 1$ 的 $\alpha_n \in \mathbb{R}^{s_n}$, 有 $\alpha_n^T \bar{\mathbf{M}}_{n1}^{-1/2}(\boldsymbol{\beta}_{n0})(\bar{\mathbf{H}}_{n1}(\boldsymbol{\beta}_{n0}) + \lambda_{n2}\mathbf{I}_{n1})(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10}) \sim N(0, 1)$, 其中,

$$\bar{\mathbf{M}}_{n1}(\boldsymbol{\beta}_{n0}) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{X}_{i1}^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0}) \bar{\mathbf{R}}^{-1} \mathbf{R}_0 \bar{\mathbf{R}}^{-1} \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0}) \mathbf{X}_{i1} + \lambda_{n2} \mathbf{I}_{n1},$$

$$\bar{\mathbf{H}}_{n1}(\boldsymbol{\beta}_{n0}) = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{i1}^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0}) \mathbf{R}_0^{-1} \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0}) \mathbf{X}_{i1}.$$

注: 定理 2 确立了估计量的 Oracle 性质, 即能够一致地识别零系数、精确估计非零系数, 并实现非零系数的渐近正态性。 $\bar{\mathbf{M}}_{n1}$ 表示估计方程的渐近协方差矩阵, 用于刻画得分函数的变异性。在理论推导过程中施加了若干条件, 包括对惩罚参数 λ_{n2} 较为严格的速率条件, 这一设定旨在有效控制 L2 惩罚引入

的偏差, 并为渐近正态性的成立提供支持。在实际应用中, 限制条件可以适当放宽。

性质 1 的证明:

证明式(4)等价于证明: 对于任何 $\epsilon > 0$, 存在一个 n , 使得对于所有的 $\Delta > 0$ 均有

$$Pr \left\{ \sup_{\|\beta_n - \beta_{n0}\| = \Delta \left(\sqrt{\frac{p_n}{n}} + \|\lambda_{n2} \beta_{n0}\| \right)} (\hat{\beta}_n - \beta_{n0})^T \tilde{U}_n(\beta_n) < 0 \right\} > 1 - \epsilon,$$

通过应用泰勒展开, 对于介于 β_{n0} 和 β_n 之间的 β_n^* , 得到:

$$\begin{aligned} (\tilde{\beta}_n - \beta_{n0})^T \tilde{U}_n(\beta_n) &= (\beta_n - \beta_{n0})^T \tilde{U}_n(\beta_{n0}) + (\beta_n - \beta_{n0})^T \frac{\partial}{\partial \beta_n} U_n(\beta_n^*)(\beta_n - \beta_{n0}) \\ &= I_{n1} + I_{n2}. \end{aligned} \quad (6)$$

对于 I_{n1} 项, 有:

$$\begin{aligned} I_{n1} &= (\beta_n - \beta_{n0})^T \tilde{U}_n(\beta_{n0}) \\ &\leq \|\beta_n - \beta_{n0}\| \cdot \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T (\mathbf{Y}_i - \mu_i(\beta_{n0})) - \lambda_{n2} \beta_{n0} \right\| \\ &\leq \Delta \left(\sqrt{\frac{p_n}{n}} + \|\lambda_{n2} \beta_{n0}\| \right) \cdot \left(O_p \left(\sqrt{\frac{p_n}{n}} \right) + \|\lambda_{n2} \beta_{n0}\| \right) \\ &= \Delta O_p(\gamma_n^2), \end{aligned}$$

其中 $\gamma_n \triangleq \sqrt{\frac{p_n}{n}} + \|\lambda_{n2} \beta_{n0}\|$ 。第二个不等式成立源于(A1)下能保证 $E \left(\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T (\mathbf{Y}_i - \mu_i(\beta_{n0})) \right\|^2 \right) \leq O_p \left(\frac{p_n}{n} \right)$ 。

对于 I_{n2} 项, 进一步拆分为 I_{n21} 和 I_{n22} :

$$\begin{aligned} I_{n2} &= (\beta_n - \beta_{n0})^T \frac{\partial}{\partial \beta_n} \tilde{U}_n(\beta_n^*)(\beta_n - \beta_{n0}) \\ &= (\beta_n - \beta_{n0})^T \left(-\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T A_i(\beta_n^*) \mathbf{X}_i - \lambda_{n2} I_n \right) (\beta_n - \beta_{n0}) \\ &= (\beta_n - \beta_{n0})^T \left(-\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T A_i(\beta_n) \mathbf{X}_i - \lambda_{n2} I_n \right) (\beta_n - \beta_{n0}) \\ &\quad + (\beta_n - \beta_{n0})^T \left[-\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T (A_i^*(\beta_n) - A_i(\beta_n)) \mathbf{X}_i \right] (\beta_n - \beta_{n0}) \\ &= I_{n21} + I_{n22}. \end{aligned}$$

对于 I_{n22} 项, 由 $p^2 n^{-1} = o(1)$, $\lambda_{n2}^2 p_n = o(1)$ 以及(A3)有:

$$\begin{aligned} I_{n22} &\leq \|\beta_n - \beta_{n0}\|^2 \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m (A_i(\beta_n) - A_i^*(\beta_n)) \mathbf{X}_{ij} \mathbf{X}_{ij}^T \\ &\leq \|\beta_n - \beta_{n0}\|^2 \frac{1}{n} \sup_{i,j} \|\mathbf{X}_{ij}\| \cdot \|\beta_n - \beta_n^*\| \lambda_{\max} \left(\sum \mathbf{X}_i^T \mathbf{X}_i \right) \\ &\leq \Delta^2 \left(\sqrt{p_n/n} + \|\lambda_{n2} \beta_{n0}\| \right)^2 \frac{1}{n} O_p(p_n) O_p \left(\sqrt{p_n/n} + \|\lambda_{n2} \beta_{n0}\| \right) O_p(n) \\ &= \Delta^2 O_p(p_n^2/n^{3/2}) \\ &= \Delta^2 O_p(p_n/n). \end{aligned}$$

接下来分析 I_{n21} 项:

$$\begin{aligned}
 I_{n21} &= -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T A_i(\boldsymbol{\beta}_n) \mathbf{X}_i + \lambda_{n2} I_n \right) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\
 &\leq -\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 \left(\lambda_{\min}(A_i(\boldsymbol{\beta}_n)) \lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{X}_i \right) + \lambda_{n2} \right) \\
 &\leq -\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 \left(\lambda_{\min}(A_i(\boldsymbol{\beta}_n)) \lambda_{\min}(R_0) \lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T R_0^{-1} \mathbf{X}_i \right) + \lambda_{n2} \right) \\
 &\leq -\Delta^2 \gamma_n^2 \cdot \left(\left(a_A a_R \lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T R_0^{-1} \mathbf{X}_i \right) + \lambda_{n2} \right) + \lambda_{n2} (1 - a_A a_R) \right) \\
 &\leq -C \cdot \Delta^2 \gamma_n^2,
 \end{aligned}$$

其中 a_A 与 a_R 分别表示 $\lambda_{\min}(A_i(\boldsymbol{\beta}_n))$ 与 $\lambda_{\min}(R_0)$, 且 $0 < a_A a_R < 1$ 由标准化数据的协方差矩阵保证。

最后, 泰勒展开式(6)由 I_{n21} 项支配, 即对于所有充分大的 Δ , I_{n21} 项具有较大的负值, 由此性质 1 得证。

定理 1 的证明:

证明式(5)等价于证明: 对于任何 $\epsilon > 0$, 存在一个 n , 使得对于所有的 $\Delta > 0$, 均有

$$\Pr \left\{ \sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| = \Delta \left(\sqrt{\frac{p_n}{n}} + |\lambda_{n2} \boldsymbol{\beta}_{n0}| \right)} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})^T \hat{U}_n(\boldsymbol{\beta}_n) < 0 \right\} > 1 - \epsilon,$$

通过应用泰勒展开, 对于介于 $\boldsymbol{\beta}_{n0}$ 和 $\boldsymbol{\beta}_n$ 之间的 $\boldsymbol{\beta}_n^*$, 得到:

$$\begin{aligned}
 (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T U_n(\boldsymbol{\beta}_n) &= (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T U_n(\boldsymbol{\beta}_{n0}) + (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \frac{\partial}{\partial \boldsymbol{\beta}_n} U_n(\boldsymbol{\beta}_n^*) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\
 &= I_{n1} + I_{n2},
 \end{aligned} \tag{7}$$

其中, $I_{n1} = (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \bar{U}_n(\boldsymbol{\beta}_n) + (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T (U_n(\boldsymbol{\beta}_{n0}) - \bar{U}_n(\boldsymbol{\beta}_n)) = I_{n11} + I_{n12}$ 。

对于 I_{n11} 项, 有:

$$\begin{aligned}
 I_{n11} &= (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T (\bar{S}_n(\boldsymbol{\beta}_n) - q_{scad, \lambda_{n1}}(\boldsymbol{\beta}_{n0}) - \lambda_{n2} \boldsymbol{\beta}_{n0}) \\
 &\leq \|\boldsymbol{\beta}_{n0} - \boldsymbol{\beta}_n\| \cdot (\|\bar{S}_n(\boldsymbol{\beta}_{n0})\| + \|\lambda_{n2} \boldsymbol{\beta}_{n0}\|) \\
 &\leq \Delta \left(\sqrt{\frac{p_n}{n}} + \|\lambda_{n2} \boldsymbol{\beta}_{n0}\| \right) \cdot O_p \left(\sqrt{\frac{p_n}{n}} + |\lambda_{n2} \boldsymbol{\beta}_{n0}| \right) \\
 &= \Delta \gamma_n O_p(\gamma_n),
 \end{aligned}$$

其中 $\gamma_n = \sqrt{p_n/n} + \|\lambda_{n2} \boldsymbol{\beta}_{n0}\|$ 且 $E(\|\bar{S}(\boldsymbol{\beta}_{n0})\|^2) \leq O_p \left(\frac{p_n}{n} \right)$ 。由 SCAD 惩罚的性质可知, 由于在 I_{n12} 时有 $\lambda_{n1} \rightarrow 0$, 对于足够大的 I_{n12} , 当 $\theta > a\lambda_{n1}$ 时, I_{n12} 与 I_{n12} 成立, 因此惩罚项对真实非零参数的影响可以忽略。

对于 I_{n12} 项, 由 Wang 的引理 3.1 [3] 可知 $\|U_n(\boldsymbol{\beta}_{n0}) - \bar{U}_n(\boldsymbol{\beta}_n)\| = O \left(\frac{p_n}{n} \right)$, 因此 $I_{n12} \leq \Delta \gamma_n O \left(\frac{p_n}{n} \right)$ 。

由于 $-\frac{\partial}{\partial \boldsymbol{\beta}_n} U_n(\boldsymbol{\beta}_n^*) \triangleq D_n(\boldsymbol{\beta}_n^*) + q'_{scad, \lambda_{n1}}(\boldsymbol{\beta}_n^*) + \lambda_{n2} I$, 其中 $D_n(\boldsymbol{\beta}_n) = -\frac{\partial}{\partial \boldsymbol{\beta}_n} S_n(\boldsymbol{\beta}_n)$, 因此, 对于 I_{n2} 项有:

$$\begin{aligned}
I_{n2} &= -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left(\bar{D}_n(\boldsymbol{\beta}_n^*) + q'_{scad, \lambda_{n1}}(\boldsymbol{\beta}_n^*) + \lambda_{n2} I_n \right) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\
&\quad - (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left(D_n(\boldsymbol{\beta}_n^* \boldsymbol{\beta}_{n0}) - \bar{D}_n(\boldsymbol{\beta}_n^*) \right) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\
&\triangleq I_{n21} + I_{n22},
\end{aligned}$$

由前述可知 $q'_{scad, \lambda_{n1}}(\boldsymbol{\beta}_n^*) = 0$, 因此,

$$\begin{aligned}
I_{n21} &= -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left(\bar{H}_n(\boldsymbol{\beta}_{n0}) + \lambda_{n2} I_n \right) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) - (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left(\bar{H}_n(\boldsymbol{\beta}_n^*) - \bar{H}_n(\boldsymbol{\beta}_{n0}) \right) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\
&\quad - (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left(\bar{D}_n(\boldsymbol{\beta}_n^*) - \bar{H}_n(\boldsymbol{\beta}_n^*) \right) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\
&\triangleq I_{n21}^a + I_{n21}^b + I_{n21}^c,
\end{aligned}$$

易证 $I_{n21}^b = \Delta^2 o_p(p_n/n)$ 且 $I_{n21}^c = \Delta^2 o_p(p_n/n)$ 。结合 $\bar{H}_n(\boldsymbol{\beta}_{n0}) = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T A_i^{1/2}(\boldsymbol{\beta}_{n0}) \bar{R}^{-1} A_i^{1/2}(\boldsymbol{\beta}_{n0}) \mathbf{X}_i$ 与正则化条件

(A1)~(A4), 对于 I_{n21}^a , 可以得到

$$\begin{aligned}
I_{n21}^a &= -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T A_i^{1/2}(\boldsymbol{\beta}_{n0}) R_0^{-1} A_i^{1/2} \mathbf{X}_i + \lambda_{n2} I_n \right) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\
&\quad - (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T A_i^{1/2}(\boldsymbol{\beta}_{n0}) [\bar{R}^{-1} - R_0^{-1}] A_i^{1/2} \mathbf{X}_i \right) (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\
&\leq -\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 \left[\lambda_{\min}(A_i(\boldsymbol{\beta}_{n0})) \cdot \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T R_0^{-1} \mathbf{X}_i + \lambda_{n2} \right] \\
&\quad - \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 \left[\lambda_{\min}([\bar{R}^{-1} - R_0^{-1}]) \cdot \lambda_{\min}(A_i(\boldsymbol{\beta}_{n0})) \cdot \lambda_{\min}\left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{X}_i\right) \right] \\
&\leq -\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 \left[a_A \lambda_{\min}\left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T R_0^{-1} \mathbf{X}_i\right) + \lambda_{n2} \right] - \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 C_2 \\
&\leq -\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 (C_1 + C_2) \\
&\leq -C \Delta^2 \gamma_n^2,
\end{aligned}$$

其中 $a_A = \lambda_{\min}(A_i(\boldsymbol{\beta}_n))$, $C_1 > 0$ 且 $C_2 \geq 0$, $C_1 = \min(a_A, \lambda_{n2})$ 。

最后, 泰勒展开式(7)由 I_{n21}^a 支配, 对于所有充分大的 Δ , I_{n21}^a 是大且负的, 因此定理 1 得证。

定理 2 的证明:

对于定理 2 中的(1)与(2), Wang 等[8]已在 SCAD 惩罚下的 GEE 框架中给出了完整的理论证明, 其所依赖的正则条件与本文研究内容无冲突, 为避免赘述, 此处不再重复证明。下文将重点证明性质(3)。

首先证明 $\alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\boldsymbol{\beta}_{n10}) \bar{U}_{n1}(\boldsymbol{\beta}_{n10}) \sim N(0, 1)$, 现有:

$$\begin{aligned}
&\alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\boldsymbol{\beta}_{n10}) \bar{U}_{n1}(\boldsymbol{\beta}_{n10}) \\
&= \alpha_{n1}^T \bar{M}_{n1}^{-1/2} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T A_i^{1/2}(\boldsymbol{\beta}_{n10}) \bar{R}^{-1} A_i^{-1/2}(\boldsymbol{\beta}_{n10}) (\mathbf{Y}_i - \mu_i(\boldsymbol{\beta}_{n10})) + \lambda_{n2} \boldsymbol{\beta}_{n10} \right\} \\
&= \sum_{i=1}^n Z_{ni},
\end{aligned}$$

定义 $Z_{ni} = \frac{1}{n} \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\boldsymbol{\beta}_{n10}) \mathbf{X}_i^T A_i^{1/2}(\boldsymbol{\beta}_{n10}) \bar{R}^{-1} \varepsilon_i(\boldsymbol{\beta}_{n10}) + \lambda_{n2} \boldsymbol{\beta}_{n10} \triangleq I_{n1} + I_{n2}$, 再根据柯西 - 施瓦茨不等式, 有:

$$Z_{ni}^2 = \|I_{n1} + I_{n2}\|^2 \leq \|I_{n1}\|^2 + \|I_{n2}\|^2 + 2|I_{n1} \cdot I_{n2}| \leq 2(\|I_{n1}\|^2 + \|I_{n2}\|^2),$$

由于 $\|I_{n2}\|^2 = \lambda_{n2}^2 \|\beta_{n10}\|^2 = O(1/n^2)$, 且

$$\begin{aligned}\|I_{n1}\|^2 &\leq \frac{1}{n^2} \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) \mathbf{X}_i^T A_i^{1/2}(\beta_{n10}) \bar{R}^{-1} \|\varepsilon_i(\beta_{n10})\|^2 \bar{R}^{-1} A_i^{1/2}(\beta_{n10}) \mathbf{X}_i \bar{M}_{n1}^{-1/2}(\beta_{n10}) \alpha_{n1} \\ &\leq \|\varepsilon_i(\beta_{n10})\|^2 \lambda_{\max}(A_i(\beta_{n10})) \lambda_{\max}(\bar{R}^{-1}) \cdot \frac{1}{n^2} \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) \mathbf{X}_i^T \bar{R}^{-1} \mathbf{X}_i \bar{M}_{n1}^{-1/2}(\beta_{n10}) \alpha_{n1} \\ &= C \cdot \psi_{ni} \cdot \|\varepsilon_i(\beta_{n10})\|^2.\end{aligned}$$

定义其中 $\psi_{ni} \triangleq \frac{1}{n^2} \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) \mathbf{X}_i^T \bar{R}^{-1} \mathbf{X}_i \bar{M}_{n1}^{-1/2}(\beta_{n10}) \alpha_{n1}$ 且满足

$$\psi_{ni} \leq \frac{1}{n^2} \lambda_{\max}(\mathbf{X}_i^T \bar{R}^{-1} \mathbf{X}_i) \cdot \lambda_{\max}(\bar{M}_{n1}^{-1}(\beta_{n10})) = \frac{1}{n^2} \frac{\lambda_{\max}(\mathbf{X}_i^T \bar{R}^{-1} \mathbf{X}_i)}{\lambda_{\min}(\bar{M}_{n1}(\beta_{n10}))}.$$

进一步, 矩阵 $\bar{M}_{n1}(\beta_{n10})$ 的最小特征值可以估计为

$$\begin{aligned}b_{n1}^T \bar{M}_{n1}(\beta_{n10}) b_{n1} &= b_{n1}^T \left(\frac{1}{n^2} \sum_{i=1}^n \mathbf{X}_i^T A_i^{-1/2}(\beta_{n10}) \bar{R}^{-1} R_0 \bar{R}^{-1} A_i^{1/2}(\beta_{n10}) \mathbf{X}_i + \lambda_{n2} I_{n1} \right) b_{n1} \\ &\geq b_{n1}^T \left(\min_i \lambda_{\min}(A_i(\beta_{n10})) \frac{1}{n^2} \sum_{i=1}^n \mathbf{X}_i^T R_0^{-1} \mathbf{X}_i + \lambda_{n2} I_{n1} \right) b_{n1} \\ &\geq \frac{c}{n} \lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T R_0^{-1} \mathbf{X}_i \right) + \lambda_{n2}\end{aligned}$$

由正则化假设(A4), 可得 $\|\bar{R}^{-1} R_0 \bar{R}^{-1} - R_0^{-1}\| = o(1)$. 因此,

$$\psi_{ni} = \frac{1}{n^2} \frac{\lambda_{\max}(\mathbf{X}_i^T \bar{R}^{-1} \mathbf{X}_i)}{\frac{c}{n} \lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T R_0^{-1} \mathbf{X}_i \right) + \lambda_{n2}} = O(p_n/n),$$

最终可以得到

$$\begin{aligned}Z_{ni}^2 &\leq 2(\|I_{n1}\|^2 + \|I_{n2}\|^2) \\ &= 2\left(C \cdot \psi_{ni} \cdot \|\varepsilon_i(\beta_{n10})\|^2 + o(1)\right) \\ &= C \cdot \psi_{ni} \cdot \|\varepsilon_i(\beta_{n10})\|^2.\end{aligned}$$

由(A5)中 $E(\|\varepsilon_{ni}\|^2) = c_1$, 根据 Lindeberg 条件, 有

$$\sum_{i=1}^n E[Z_{ni}^2 I(|Z_{ni}| > \varepsilon)] \leq \sum_{i=1}^n C \psi_{ni} E\left[\|\varepsilon_i(\beta_{n0})\|^2 I\left(\|\varepsilon_i(\beta_{n0})\|^2 > \frac{\varepsilon^2}{C \psi_{ni}}\right)\right].$$

因此 $\alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) \bar{U}_{n1}(\beta_{n10}) \sim N(0,1)$ 得证. 最后, 重新整理原式有

$$\begin{aligned}&\alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) \bar{U}_{n1}(\beta_{n10}) \\ &= \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) U_{n1}(\beta_{n10}) + \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) [\bar{U}_{n1}(\beta_{n10}) - U_{n1}(\beta_{n10})] \\ &= \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) [D_{n1}(\beta_{n1}^*) + f_{\lambda'_{n1}}(\beta_{n10}) + \lambda_{n2} I_{n1}] (\hat{\beta}_{n1} - \beta_{n10}) + \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) [\bar{U}_{n1}(\beta_{n10}) - U_{n1}(\beta_{n10})] \\ &= \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) [\bar{H}_{n1}(\beta_{n10}) + f_{\lambda'_{n1}}(\beta_{n10}) + \lambda_{n2} I_{n1}] (\hat{\beta}_{n1} - \beta_{n10}) \\ &\quad + \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) [D_{n1}(\beta_{n1}^*) - \bar{H}_{n1}(\beta_{n10})] (\hat{\beta}_{n1} - \beta_{n10}) + \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) [\bar{U}_{n1}(\beta_{n10}) - U_{n1}(\beta_{n10})].\end{aligned}$$

文献[3]指出,

$$\sup_{\|\beta_n - \beta_{n10}\| \leq \Delta \gamma_n} \left| \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) (D_{n1}(\beta_{n1}^*) - \bar{H}_{n1}(\beta_{n10})) (\hat{\beta}_n - \beta_{n10}) \right| = o_p(1),$$

$$\sup_{\|\beta_n - \beta_{n10}\| \leq \Delta \gamma_n} \left| \alpha_{n1}^T \bar{M}_{n1}^{-1/2}(\beta_{n10}) (\bar{U}_{n1}(\beta_{n10}) - U_{n1}(\beta_{n10})) \right| = o_p(1),$$

最终可以得到:

$$n = 200, \rho = 0.5$$

至此定理 2 证毕。

3. 模拟实验

本节通过广泛的模拟研究评估并比较广义估计方程(GEE)、惩罚广义估计方程(PGEE)及 SCAD-L2 惩罚 GEE (简称 SLGEE)的性能表现, 惩罚系数 λ_{n1} 与 λ_{n2} 的选取通过网格搜索结合交叉验证实现。在模型拟合中, 为考察工作相关设定对估计与变量选择的影响, 对 GEE、PGEE 和 SLGEE 分别采用三种工作相关结构进行计算: 独立结构(indep)、可交换结构(exch)以及一阶自回归结构(AR-1), 每次模拟可得到“方法 $\times$ 工作相关结构”的多组结果, 用于衡量不同工作相关矩阵假设下模型估计与变量选择的稳定性。由于 MM 算法固有的数值扰动, 所估系数不会精确为零, 因此将绝对值小于 0.001 的系数判定为非活跃变量, 并将其从最终模型中剔除。

实验设计涵盖两类建模情形: 其一为高斯型响应, 采用恒等连接函数; 其二为二分类响应, 采用 Logistic 连接函数。对模拟实验结果采用多种指标评估模型性能: 均方误差(MSE), 计算公式为 $MSE = \|\hat{\beta}_n - \beta_{n0}\|^2 / p$; 过拟合数, 即真实零系数被错误选中的数量; 欠拟合数, 即真实非零系数被错误排除的数量; 识别准确率, 即正确识别系数的比例; 选择稳定性, 通过 Jaccard 指数量化, 即重复模拟中非零系数集交集与并集的比值。

3.1. 实验 1

相关正态响应数据的生成模型为

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \varepsilon_{ij}, i = 1, \dots, n; j = 1, \dots, m,$$

令 $m = 4$, 则 $\mathbf{X}_{ij} = (x_{ij,1}, \dots, x_{ij,200})^T$ 为包含 200 个协变量的向量, 真实参数设为

$$\boldsymbol{\beta}_{n0} = \left(\underbrace{3, \dots, 3}_{15}, \underbrace{1.5, \dots, 1.5}_{15}, \underbrace{2, \dots, 2}_{15}, \underbrace{0, \dots, 0}_{155} \right)^T,$$

即前 45 个系数非零, 其余 155 个为零。随机误差 ε_i 来自多元正态分布, 均值为 0, 协方差矩阵为 $EX(\rho)$, 即对角元为 1、非对角元均为 ρ 的对称矩阵。实验设置 $\rho = 0.5$ 和 $\rho = 0.8$ 两种情形, 以反映不同程度的组内相关性。对任意 i , 协变量的生成方式为

$$\begin{aligned} \mathbf{X}_{jk} &= Z_{1,j} + \varepsilon_{jk}^x, Z_{1,j} \sim N(0, EX(\rho)), j = 1, \dots, 4; k = 1, \dots, 15, \\ \mathbf{X}_{jk} &= Z_{2,j} + \varepsilon_{jk}^x, Z_{2,j} \sim N(0, EX(\rho)), j = 1, \dots, 4; k = 16, \dots, 30, \\ \mathbf{X}_{jk} &= Z_{3,j} + \varepsilon_{jk}^x, Z_{3,j} \sim N(0, EX(\rho)), j = 1, \dots, 4; k = 31, \dots, 45, \\ \mathbf{X}_{jk} &\sim N(0, EX(\rho)), j = 1, \dots, 4; k = 46, \dots, 200, \end{aligned}$$

其中 $\varepsilon_{jk}^x \sim_{i.i.d.} N(0, 0.001), k = 1, \dots, 100$ 。

为了更直观地看出不同参数组合对模型的影响, 图 1 展示了实验 1 中通过二维网格搜索与五折交叉验证选取惩罚参数 λ_1, λ_2 的热力图。对每一个网格点均计算平均交叉验证误差, 并选择使其最小的参数组

合作为最优惩罚参数。结果显示, 交叉验证误差关于 λ_1 呈现明显的“先下降后上升”的趋势: 当 λ_1 过小时惩罚不足易导致过拟合, CV 误差较大; 当 λ_1 过大时, 真实信号被过度收缩从而产生欠拟合, CV 误差亦上升。对 λ_2 而言, 小幅 L2 惩罚有助于提高估计稳定性并改善预测误差, 但当 λ_2 增大至 0.1 或 1 时会引发过度收缩, 导致 CV 误差显著增大。最终在所设网格内选得 $(\lambda_1^*, \lambda_2^*) = (0.3, 0.01)$, 最小 CV 误差为 966.29。

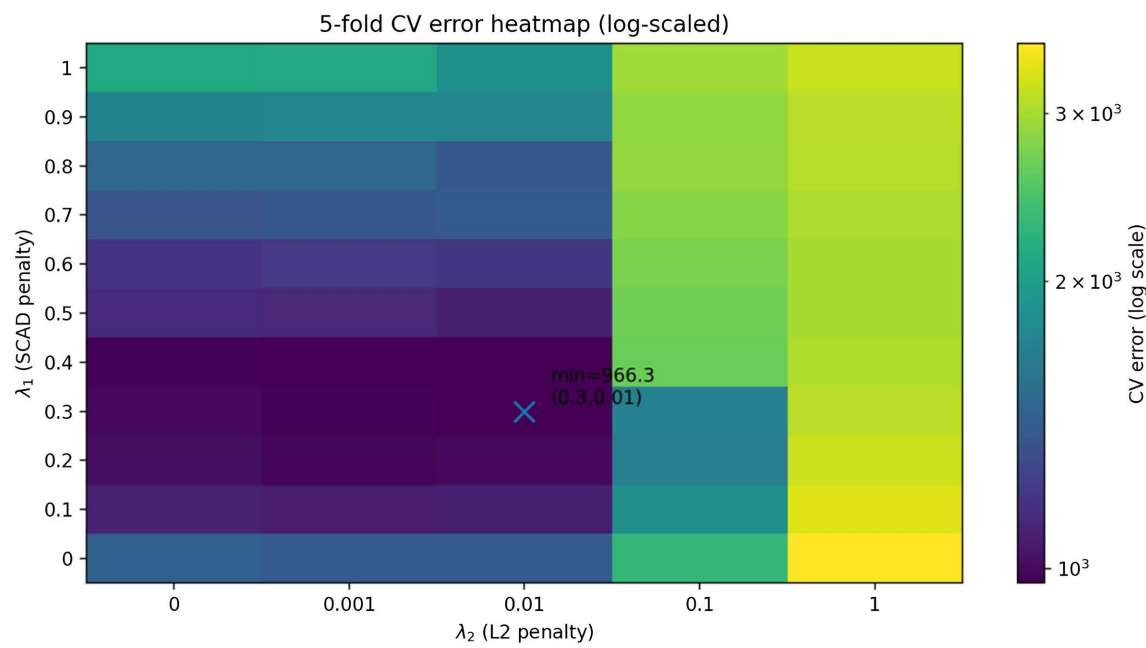


Figure 1. Heat map of five-fold cross-validation error for the SCAD-L2 penalized GEE over a two-dimensional parameter grid
图 1. SCAD-L2 化惩罚 GEE 在二维参数网格下的五折交叉验证误差热力图

表 1 和表 2 汇报了在相关正态连续型响应数据中, GEE、PGEE 与 SLGEE 三种方法在样本量 $n = 200$ 与 $n = 600$ 下 50 次重复实验的模拟结果。结果显示, 相较于不具备变量选择能力的传统 GEE 以及仅依赖单一惩罚机制的 PGEE 方法, SLGEE 在所有实验设定下均表现出更优的统计性能。具体而言, SLGEE 的均方误差在各方法中始终最低, 略小于 0.1, 且在不同工作相关结构下结果保持一致, 表明其估计过程具有较好的稳健性。在变量选择方面, SLGEE 能够在有效控制过拟合的同时显著降低真实信号变量的遗漏率, 使变量识别准确率稳定提升至 97%~99%, Jaccard 相似系数保持在 0.94 以上, 显示出其高维纵向数据情形下兼顾选择准确性与稳定性的优势。随着样本量的增加, PGEE 和 GEE 的均方误差均呈下降趋势, 而 SLGEE 即使在样本量较小的情况下也保持较低且稳定的均方误差, 表现出更优的估计性能。

Table 1. Simulation results for GEE, PGEE, and SLGEE in Simulation Study 1 with $n = 200$
表 1. 模拟实验 1 中 $n = 200$ 时 GEE、PGEE 和 SLGEE 的模拟结果

方法	均方误差		过拟合数		欠拟合数		识别准确率		Jaccard
	均值	方差	均值	方差	均值	方差	均值	方差	均值
$n = 200, \rho = 0.5$									
GEE.indep	0.303	0.073	152.88	1.35	0	0.00	23.56%	0.67%	0.9791
GEE.exch	0.22	0.049	151.98	1.67	0	0.00	24.01%	0.84%	0.9704

续表

GEE.AR-1	0.249	0.053	151.8	1.94	0.02	0.14	24.09%	0.96%	0.9684
PGEE.indep	0.388	0.077	16.2	5.31	10.36	1.89	86.72%	2.98%	0.4075
PGEE.exch	0.344	0.064	7.72	3.92	10.22	1.73	91.03%	2.21%	0.4996
PGEE.AR-1	0.368	0.059	9.2	4.62	10.58	1.54	90.11%	2.60%	0.4729
SLGEE.indep	0.098	0.021	17.66	5.87	2.94	1.10	89.70%	2.96%	0.5672
SLGEE.exch	0.092	0.022	8.72	4.61	2.34	1.19	94.47%	2.22%	0.7008
SLGEE.AR-1	0.095	0.022	10.56	5.29	2.38	1.26	93.53%	2.63%	0.6656
$n = 200, \rho = 0.8$									
GEE.indep	0.221	0.052	153.16	1.11	0.02	0.14	23.41%	0.56%	0.9816
GEE.exch	0.097	0.023	152.1	1.71	0	0.00	23.95%	0.85%	0.9717
GEE.AR-1	0.117	0.029	152	1.68	0	0.00	24.00%	0.84%	0.9705
PGEE.indep	0.375	0.057	24.24	5.37	10.18	1.30	82.79%	2.81%	0.3693
PGEE.exch	0.244	0.044	7.38	4.30	8.58	1.30	92.02%	2.29%	0.5454
PGEE.AR-1	0.267	0.045	9.9	5.44	8.9	1.34	90.60%	2.87%	0.5022
SLGEE.indep	0.1	0.021	26.1	6.34	3.34	0.72	85.28%	3.26%	0.5035
SLGEE.exch	0.086	0.021	7.52	4.78	1.28	1.05	95.60%	2.42%	0.7373
SLGEE.AR-1	0.087	0.020	10.6	5.65	1.68	1.13	93.86%	2.90%	0.6718

Table 2. Simulation results for GEE, PGEE, and SLGEE in Simulation Study 1 with $n = 600$ **表 2.** 模拟实验 1 中 $n = 600$ 时中 GEE、PGEE 和 SLGEE 的模拟结果

Methods	MSE		Overfitting		Underfitting		ER		Jaccard
	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean
$n = 600, \rho = 0.5$									
GEE.indep	0.093	0.021	150.84	2.23	0	0.00	24.58%	1.12%	0.9596
GEE.exch	0.063	0.015	148.58	2.42	0.02	0.14	25.70%	1.20%	0.9383
GEE.AR-1	0.071	0.018	149.96	2.83	0	0.00	25.02%	1.41%	0.9512
PGEE.indep	0.156	0.042	3.4	1.23	5.38	1.63	95.61%	1.09%	0.7043
PGEE.exch	0.139	0.038	0.48	0.58	5.36	1.59	97.08%	0.87%	0.7935
PGEE.AR-1	0.144	0.040	0.78	0.71	5.4	1.62	96.91%	0.82%	0.7818
SLGEE.indep	0.103	0.019	1.82	1.38	3.12	1.22	97.53%	0.97%	0.8924
SLGEE.exch	0.098	0.020	0.2	0.45	2.34	1.44	98.73%	0.71%	0.9476
SLGEE.AR-1	0.099	0.018	0.34	0.59	2.66	1.27	98.50%	0.65%	0.9463
$n = 600, \rho = 0.8$									
GEE.indep	0.088	0.019	151.98	1.76	0	0.00	24.01%	0.88%	0.9704
GEE.exch	0.027	0.005	149.16	2.46	0	0.00	25.42%	1.23%	0.9439
GEE.AR-1	0.034	0.006	150.04	2.40	0	0.00	24.98%	1.20%	0.9519
PGEE.indep	0.164	0.042	9.16	2.22	5.76	1.42	92.54%	1.26%	0.5806
PGEE.exch	0.114	0.034	0.7	0.86	5.24	1.44	97.03%	0.81%	0.7953

续表

PGEE.AR-1	0.12	0.033	1.1	1.02	5.3	1.40	96.80%	0.81%	0.7796
SLGEE.indep	0.097	0.024	8	2.77	2.76	1.19	94.62%	1.52%	0.7145
SLGEE.exch	0.079	0.021	0.12	0.33	1.42	1.28	99.23%	0.67%	0.9542
SLGEE.AR-1	0.08	0.022	0.32	0.68	1.16	1.20	99.26%	0.64%	0.9515

3.2. 实验 2

相关二元响应数据的边际均值 π_{ij} 满足

$$\Pr(Y_{ij} = 1 | X_{ij}) = \frac{\exp X_{ij}^T \beta}{1 + \exp X_{ij}^T \beta}, i = 1, \dots, n; j = 1, \dots, m,$$

令 $n = 200, m = 10$, $X_{ij} = (x_{ij,1}, \dots, x_{ij,100})^T$ 为包含 100 个协变量的向量, 真实参数设为

$$\beta = \left(\underbrace{0.7, \dots, 0.7}_{10}, \underbrace{-0.7, \dots, -0.7}_{10}, \underbrace{-0.4, \dots, -0.4}_{10}, \underbrace{0, \dots, 0}_{70} \right)^T,$$

协变量的生成方式为

$$\begin{aligned} X_{jk} &= Z_1 + \varepsilon_{jk}^x, \quad Z_{1,j} \sim \text{Uniform}(0, 0.5), j = 1, \dots, 10; k = 1, \dots, 10, \\ X_{jk} &= Z_{2,j} + \varepsilon_{jk}^x, \quad Z_{2,j} \sim \text{Uniform}(0, 0.5), j = 1, \dots, 10; k = 11, \dots, 20, \\ X_{jk} &= Z_{3,j} + \varepsilon_{jk}^x, \quad Z_{3,j} \sim \text{Uniform}(0, 0.5), j = 1, \dots, 10; k = 21, \dots, 30, \\ X_{jk} &\sim \text{Uniform}(0, 0.5), \quad j = 1, \dots, 10; k = 31, \dots, 100, \end{aligned}$$

其中 $\varepsilon_{jk}^x \sim_{i.i.d.} N(0, 0.001)$ 。

表 3 与表 4 给出了相关二元响应数据情形下的模拟结果, 该情形相较于连续型响应具有更强的非线性特征和信息稀疏性。数值结果显示, SLGEE 在该建模场景下仍能够保持良好的估计与选择性能, 其回归系数均方误差在不同样本量设定下均处于较低水平, 仅有 0.03 左右, 表明该方法对二元响应模型具有较好的适应性。在变量选择方面, 随着样本量的增加, SLGEE 能够保持对过拟合与欠拟合的有效控制, 当样本量达到 600 时, 其过拟合变量数接近于 0, 欠拟合变量数稳定在 1~2 个, 使变量识别准确率提升至 95%~98%, Jaccard 相似系数超过 0.85, 显示出较高的选择准确性与稳定性。并且在不同工作相关结构及相关性情形下, SLGEE 的估计精度和变量选择表现变化较小, 体现了其在二元响应纵向数据分析中的稳健性。

Table 3. Simulation results for GEE, PGEE, and SLGEE in Simulation Study 2 with $n = 200$

表 3. 模拟实验 2 中 $n = 200$ 时 GEE、PGEE 和 SLGEE 的模拟结果

Methods	MSE		Overfitting		Underfitting		ER		Jaccard
	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean
$n = 200, \rho = 0.5$									
GEE.indep	0.290	0.019	69.88	0.14	0.13	0.14	30.00%	0.00%	0.9950
GEE.exch	0.281	0.019	69.88	0.14	0.00	0.00	30.13%	0.00%	0.9975
GEE.AR-1	0.287	0.019	69.88	0.14	0.00	0.00	30.13%	0.00%	0.9975
PGEE.indep	0.279	0.019	35.75	0.77	7.63	0.67	56.63%	1.97%	0.4269
PGEE.exch	0.269	0.019	34.25	0.72	7.50	0.60	58.25%	2.11%	0.4154

续表

PGEE.AR-1	0.276	0.019	35.38	0.75	7.63	0.67	57.00%	2.05%	0.4235
SLGEE.indep	0.026	0.002	6.75	0.82	2.13	0.50	91.13%	0.84%	0.6364
SLGEE.exch	0.026	0.002	4.00	0.86	1.63	0.56	94.38%	1.25%	0.7270
SLGEE.AR-1	0.026	0.002	6.25	0.76	2.00	0.52	91.75%	0.76%	0.6543
$n = 200, \rho = 0.8$									
GEE.indep	0.444	0.037	69.75	0.19	0.00	0.00	30.25%	0.19%	0.9950
GEE.exch	0.418	0.034	69.75	0.28	0.00	0.00	30.25%	0.28%	0.9950
GEE.AR-1	0.443	0.036	70.00	0.00	0.00	0.00	30.00%	0.00%	1.0000
PGEE.indep	0.424	0.031	39.88	1.88	8.00	0.68	52.13%	1.89%	0.4469
PGEE.exch	0.399	0.028	39.00	1.97	7.88	0.71	53.13%	1.95%	0.4381
PGEE.AR-1	0.419	0.029	39.75	2.12	8.25	0.68	52.00%	1.89%	0.4423
SLGEE.indep	0.029	0.002	9.50	0.60	2.13	0.26	88.38%	0.71%	0.5809
SLGEE.exch	0.034	0.001	2.00	0.43	2.25	0.28	95.75%	0.67%	0.7832
SLGEE.AR-1	0.029	0.002	7.13	0.84	2.00	0.30	90.88%	1.08%	0.6263

Table 4. Simulation results for GEE, PGEE, and SLGEE in Simulation Study 2 with $n = 600$ **表 4.** 模拟实验 2 中 $n = 600$ 时 GEE、PGEE 和 SLGEE 的模拟结果

Methods	MSE		Overfitting		Underfitting		ER		Jaccard
	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean
$n = 600, \rho = 0.5$									
GEE.indep	0.081	0.003	69.75	0.19	0.00	0.00	30.25%	0.19%	0.9950
GEE.exch	0.078	0.004	69.88	0.14	0.00	0.00	30.13%	0.14%	0.9975
GEE.AR-1	0.080	0.003	69.63	0.21	0.00	0.00	30.38%	0.21%	0.9925
PGEE.indep	0.071	0.003	13.88	1.49	7.50	0.64	78.63%	1.58%	0.3880
PGEE.exch	0.071	0.004	12.25	1.73	8.13	0.50	79.63%	1.76%	0.3691
PGEE.AR-1	0.071	0.003	13.63	1.57	7.63	0.64	78.75%	1.68%	0.3826
SLGEE.indep	0.022	0.001	0.38	0.21	1.50	0.21	98.13%	0.26%	0.8873
SLGEE.exch	0.025	0.001	0.13	0.14	1.63	0.30	98.25%	0.28%	0.8980
SLGEE.AR-1	0.022	0.001	0.25	0.18	1.75	0.28	98.00%	0.37%	0.8872
$n = 600, \rho = 0.8$									
GEE.indep	0.128	0.006	69.88	0.14	0.25	0.18	29.88%	0.26%	0.9925
GEE.exch	0.114	0.006	70.00	0.00	0.00	0.00	30.00%	0.00%	1.0000
GEE.AR-1	0.124	0.006	69.75	0.19	0.13	0.14	30.13%	0.14%	0.9925
PGEE.indep	0.120	0.007	23.38	1.46	8.75	0.90	67.88%	1.29%	0.3496
PGEE.exch	0.107	0.006	20.50	1.19	8.88	0.78	70.63%	1.42%	0.3379
PGEE.AR-1	0.117	0.007	22.75	1.40	8.88	0.89	68.38%	1.50%	0.3445
SLGEE.indep	0.023	0.001	2.13	0.85	1.63	0.65	96.25%	1.18%	0.8007
SLGEE.exch	0.032	0.001	0.00	0.00	1.88	0.50	98.13%	0.50%	0.8952
SLGEE.AR-1	0.025	0.001	1.25	0.46	1.75	0.47	97.00%	0.74%	0.8313

4. 结论

本文提出了一种将 SCAD-L2 惩罚融入广义估计方程框架的高维变量选择方法, 以在纵向数据的协变量存在强相关性时同时实现参数估计的稀疏性与稳定性。理论分析表明, 在“大 n , 发散 p ”的高维设定下, 该方法的参数估计值满足估计一致性与 Oracle 性质, 从而确保其在高维情境下的理论可靠性。模拟实验进一步从多角度验证了该方法在参数估计精度与变量选择稳定性方面的显著优势, 并显示其在不同相关结构和响应类型下均具有稳健的统计性能, 体现了其在复杂纵向数据分析中的有效性与实用价值。

参考文献

- [1] Liang, K. and Zeger, S.L. (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22. <https://doi.org/10.1093/biomet/73.1.13>
- [2] Zorn, C.J.W. (2001) Generalized Estimating Equation Models for Correlated Data: A Review with Applications. *American Journal of Political Science*, **45**, 470-490. <https://doi.org/10.2307/2669353>
- [3] Wang, L. (2011) GEE Analysis of Clustered Binary Data with Diverging Number of Covariates. *The Annals of Statistics*, **39**, 289-417. <https://doi.org/10.1214/10-aos846>
- [4] Xie, M. and Yang, Y. (2003) Asymptotics for Generalized Estimating Equations with Large Cluster Sizes. *The Annals of Statistics*, **31**, 310-347. <https://doi.org/10.1214/aos/1046294467>
- [5] Guyon, I. and Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, 1157-1182.
- [6] Wang, L., Zhou, J. and Qu, A. (2011) Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. *Biometrics*, **68**, 353-360. <https://doi.org/10.1111/j.1541-0420.2011.01678.x>
- [7] Desboulets, L.D.D. (2018) A Review on Variable Selection in Regression Analysis. *Econometrics*, **6**, Article 45. <https://doi.org/10.3390/econometrics6040045>
- [8] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**, 267-288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- [9] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. <https://doi.org/10.1198/016214501753382273>
- [10] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. <https://doi.org/10.1198/016214506000000735>
- [11] Fu, W.J. (2003) Penalized Estimating Equations. *Biometrics*, **59**, 126-132. <https://doi.org/10.1111/1541-0420.00015>
- [12] Xu, P.R., Fu, W.J. and Zhu, L.X. (2013) Shrinkage Estimation Analysis of Correlated Binary Data with a Diverging Number of Parameters. *Science China Mathematics*, **56**, 359-377. <https://doi.org/10.1007/s11425-012-4564-y>
- [13] Fan, J. and Peng, H. (2004) Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *The Annals of Statistics*, **32**, 928-961. <https://doi.org/10.1214/009053604000000256>
- [14] Wang, M., Song, L. and Wang, X. (2010) Bridge Estimation for Generalized Linear Models with a Diverging Number of Parameters. *Statistics & Probability Letters*, **80**, 1584-1596. <https://doi.org/10.1016/j.spl.2010.06.012>
- [15] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **67**, 301-320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
- [16] Zeng, L. and Xie, J. (2014) Group Variable Selection via SCAD-L2. *Statistics*, **48**, 49-66. <https://doi.org/10.1080/02331888.2012.719513>
- [17] Zou, H. and Zhang, H.H. (2009) On the Adaptive Elastic-Net with a Diverging Number of Parameters. *The Annals of Statistics*, **37**, 1733-1751. <https://doi.org/10.1214/08-aos625>
- [18] Blommaert, A., Hens, N. and Beutels, P. (2014) Data Mining for Longitudinal Data under Multicollinearity and Time Dependence Using Penalized Generalized Estimating Equations. *Computational Statistics & Data Analysis*, **71**, 667-680. <https://doi.org/10.1016/j.csda.2013.02.023>
- [19] Lin, Y., Zhou, J., Kumar, S., Xie, W., G. Jensen, S.K., Haque, R., et al. (2020) Group Penalized Generalized Estimating Equation for Correlated Event-Related Potentials and Biomarker Selection. *BMC Medical Research Methodology*, **20**, Article No. 221. <https://doi.org/10.1186/s12874-020-01103-x>
- [20] Hunter, D.R. and Li, R. (2005) Variable Selection Using MM Algorithms. *The Annals of Statistics*, **33**, 1617-1642. <https://doi.org/10.1214/009053605000000200>