超高维竞争风险模型的特征筛选

李二倩^{1,2}, 梅波^{1,2}, 田茂再^{1,2,3,4,5*}

1. 中国人民大学应用统计科学研究中心, 北京 100872;

2. 中国人民大学统计学院, 北京 100872;

3. 新疆财经大学新疆社会经济统计研究中心, 乌鲁木齐 830012;

4. 新疆财经大学统计与信息学院, 乌鲁木齐 830012;

5. 兰州财经大学统计学院, 兰州 730020

E-mail: li2qian@ruc.edu.cn, meibo119@126.com, mztian@ruc.edu.cn

收稿日期: 2017-04-15; 接受日期: 2018-03-08; 网络出版日期: 2018-04-20; * 通信作者

中国人民大学科学研究基金 (中央高校基本科研业务费专项资金) (批准号: 17XNH090) 资助项目

摘要 在生存分析中, 已有一些文献提出处理普通时间事件数据的 Cox 模型的超高维变量选择方法. 然而, 对于个体处在多个互斥事件的风险下, 即存在竞争风险情形, 并不能直接应用这些方法. 一个分析竞争风险数据的常用模型就是比例子分布风险 (proportional subdistribution hazard, PSH) 模型. 本文基于确定联合筛选 (sure joint screening, SJS) 和惩罚近似对数部分似然, 对于超高维的 PSH 模型提出了两阶段变量选择方法, 并证明了第一步特征筛选方法的确定筛选性质 (sure screening property), 即选出的变量集合以概率 1 渐近地包含实际的显著变量. 本文通过 Monte Carlo 模拟展现了方法的性能和表现, 并与确定独立筛选 (sure independence screening) 方法进行了比较. 最后将方法应用到一个关于膀胱癌的公开数据集的分析中.

关键词 竞争风险 特征筛选 超高维变量 部分似然逼近 生存分析

MSC (2010) 主题分类 00A73, 62N99

1 引言

在生存分析或对于时间事件数据的分析中, 往往事件的发生或失效除了我们感兴趣的原因外, 还可能是因为某个特定原因或者一些其他原因发生, 也即存在竞争风险的情形. 在以往的许多文献中, 这种失效通常被视为删失的情形, 然而, 这种有信息的删失可能存在疑虑, 如当协变量对于竞争风险也有影响时 (参见文献 [1]). 特定原因的边际失效概率, 即累计发病率函数 (cumulative incidence function, CIF) 或者称为子分布 (subdistribution), 在直观上更具有吸引力和易解释性, 从而在近些年变得更加重要和流行. Fine 和 Gray^[2] 对这一子分布提出了一种半参数比例风险模型, 即 PSH 模型, 成为分析具有竞争风险的时间事件数据的一种很受欢迎的模型.

英文引用格式: Li E Q, Mei B, Tian M Z. Feature screening based on ultrahigh dimensional competing risks models (in Chinese). Sci Sin Math, 2018, 48: 1061–1086, doi: 10.1360/N012017-00069

在许多涉及竞争风险的实际应用中, 常常遇到存在高维或超高维变量的情形. 例如, Dyrskjot 等^[3]分析了基因表达信号预测非肌肉侵略性的膀胱癌结果, 然而事件中除了癌症恶化和膀胱癌导致的死亡之外, 还有因为其他原因而死亡的竞争风险; 预测变量中除了包含临床变量, 还包括远远超过样本量的基因变量数. 在这种情形下, 有效识别出对于 CIF 有显著效应的变量成为关键之处, 对于高维或超高维的 PSH 模型进行变量选择变得非常必要.

有关生存数据的变量选择方法的文献十分丰富. 例如, Faraggi 和 Simon^[4]提出了对删失生存数据的 Bayes 变量选择方法. 而一些文献研究了变量选择问题与 Cox 比例风险模型这一生存分析中流行的模型的结合. Fan 等^[5]讨论了 Cox 比例风险模型的几种变量选择方法, 包括惩罚方法和确定独立筛选^[6] (sure independence screening, SIS) 和迭代 SIS (iterative SIS, ISIS) 方法. Zhao 和 Li^[7]提出并在理论上验证了一种在删失数据分析中选择重要变量的原则方法. Song 等^[8]研究了高维生存数据的删失秩独立筛选方法. Yang 等^[9]对于具有超高维协变量的 Cox 模型提出了一种基于潜在有效预测变量的联合似然的特征筛选过程. Liu 等^[10]选择性地介绍了超高维数据的特征筛选方法. Shao 等^[11]提出了对二元分类响应数据的特征筛选方法. Wang 等^[12]提出了对超高维数据的不依赖于模型的条件独立特征筛选方法. Lin 等^[13]介绍了一种对于删失生存数据的稳健的不依赖于模型的特征筛选方法.

一些文献探索了有关竞争风险模型的高维变量选择的方法. 例如, Binder 等^[1]提出了对于带有竞争风险的高维时间事件数据的提升 (boosting) 方法. Kuk 和 Varadhan^[14]对 PSH 模型提出了基于 AIC (Akaike information criteria)、BIC (Bayesian information criteria) 和对竞争风险数据的 BIC 的逐步选择方法. Ha 等^[15]对于聚类竞争风险, 提出了在 PSH 脆弱模型下的惩罚分层似然的变量选择方法, 并利用局部二次逼近 (local quadratic approximation, LQA) 算法对方法进行实现. Tapak 等^[16]对于带竞争风险的高维时间事件数据进行了变量选择方法的比较, 包括 LASSO (least absolute shrinkage and selection operator, 参见文献 [17])、弹性网和基于似然的提升方法, 并将其应用到膀胱癌数据中. Fu 等^[18]结合变量选择的常用方法, 对竞争风险回归变量选择的惩罚方法及它们的群组版本进行了总结, 其中包括 LASSO、Zou^[19]提出的适应性 LASSO (adaptive LASSO)、Fan 和 Li^[20]提出的 SCAD (smoothly clipped absolute deviation) 和 Zhang^[21]提出的极小极大凹惩罚 (minimax concave penalty, MCP). 然而, 尚未有文献对于竞争风险模型的超高维变量选择进行探讨, 这就是本文想要研究的内容.

本文基于 PSH 模型下的对数部分似然 $\ell(\beta)$ 及其前两阶导数, 对部分对数似然进行 Taylor 展开并得到其近似, 并通过硬阈值法得到 L_0 约束下最大化近似对数部分似然问题的解, 从而得到第一阶段联合特征筛选方法粗糙筛选出的 q 个变量; 在第二阶段中, 我们最大化惩罚对数部分似然来做更精细的变量选择. 我们证明了对于竞争风险模型, 这种确定联合筛选方法在一些正则条件下具有确定筛选性质, 并且这种两步变量选择方法得到的估计量具有渐近地相合性和 oracle 性质^[18]. 在数值模拟中, 我们对于基于两种不同的协变量结构和两种真实系数生成的连续数据和存在离散变量的数据分别进行了不同样本量、维度和相关系数的模拟, 并进行了在不同初步粗糙筛选变量个数 q 情形下模拟表现的探索. 我们使用但不限于 LASSO 惩罚和 BIC 准则, 有关不同惩罚函数和其他准则 (如交叉验证 (cross validation, CV) 准则等) 可参见文献 [18]. 本文还对于一个公开可得的膀胱癌数据进行了实证分析.

本文剩余部分结构如下. 第 2 节介绍了处理竞争风险数据的 PSH 模型, 给出了对数部分似然及其前两阶导数. 第 3 节建立了对于超高维 PSH 模型进行变量选择的过程, 包括基于近似对数部分似然的确定联合筛选方法和惩罚近似对数部分似然的变量选择方法. 第 4 节证明了对于 PSH 模型中的对数部分似然, 所提方法的算法具有上升性质, 第一阶段筛选的确定筛选性质和第二阶段估计量的相合性与 oracle 性质. 第 5 节为数值研究, 通过 Monte Carlo 模拟验证了提出方法的性能和表现, 并用实证研究说明了方法的实用性. 第 6 节总结了全文.

2 模型

在竞争风险模型的设定下, 假设存在 K 个失效的原因, 记为 $\iota \in \{1, 2, \dots, K\}$. 令 T 和 C 分别表示生存时间和删失时间, 那么在通常考虑的右删失的情形下, 我们观测到 $X = \min(T, C)$, 删失指示变量为 $\delta = I(T \leq C)$, 其中 $I(\cdot)$ 为示性函数. 记 $\mathbf{Z}$ 为 $p \times 1$ 有界协向量, 并假设 $\{X_i, \delta_i, \delta_i \iota_i, \mathbf{Z}_i\}$ ($i = 1, 2, \dots, n$) 为 n 个独立同分布的观测样本. 高维或超高维度的假设意味着 $p \gg n$.

不失一般性, 记我们感兴趣的原因为原因 1, Fine 和 Gray^[2] 对由原因 1 造成失效的累积发病率函数在协变量的条件下进行建模, $F_1(t; \mathbf{Z}) = P\{(T \leq t, \iota = 1) | \mathbf{Z}\}$, 提出了基于子分布风险形式的比例子分布风险 (PSH) 模型. 其中重新叙述了文献 [22] 中子分布风险的定义如下:

$$\begin{aligned} \lambda_1(t; \mathbf{Z}) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P\{(t \leq T \leq t + \Delta t, \iota = 1) | (T \geq t) \cup ((T \leq t) \cap (\iota \neq 1)), \mathbf{Z}\} \\ &= \frac{\{dF_1(t; \mathbf{Z})/dt\}}{\{1 - F_1(t; \mathbf{Z})\}}. \end{aligned}$$

由比例风险的设定, 有 $\lambda_1(t; \mathbf{Z}) = \lambda_{10}(t) \exp(\mathbf{Z}^T \boldsymbol{\beta}_0)$, 其中 $\lambda_{10}(t)$ 为时间 t 的未指定的非负函数, $\boldsymbol{\beta}_0$ 为对于我们感兴趣的事件 (即 $\iota = 1$ 情形) 的 $p \times 1$ 阶真实的回归系数.

对于一般右删失数据, 应用逆概率删失加权 (inverse probability of censoring weighting, IPCW) 方法 (参见文献 [23]), 文献 [2] 给出了似然方程, 文献 [18] 给出了如下的对数部分似然:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\infty \left[\mathbf{Z}_i^T \boldsymbol{\beta} - \log \left\{ \sum_j w_j(u) Y_j(u) \exp(\mathbf{Z}_j^T \boldsymbol{\beta}) \right\} \right] \times w_i(u) dN_i(u), \quad (2.1)$$

其中 $N_i(t) = I(T_i \leq t, \iota_i = 1)$, $Y_i(t) = 1 - N_i(t-)$, $w_i(t)$ 为时间相依的权重; 对于个体 i 在时间 t , IPCW 权重定义为 $w_i(t) = I(C_i \geq T_i \wedge t) \hat{G}(t) / \hat{G}(X_i \wedge t)$, 其中 $G(t) = P(C \geq t)$ 为删失变量 C 的生存函数, $\hat{G}(t)$ 为 $G(t)$ 的 Kaplan-Meier 估计. 在给定的时间 t , 如果个体是右删失或者因为感兴趣的事件而失效, 则 $w_i(t) Y_i(t) = 0$; 如果因为竞争风险而失效, 则 $w_i(t) Y_i(t)$ 在 0 与 1 之间且随时间递减; 否则, $w_i(t) Y_i(t) = 1$. 另一种对数部分似然的写法为

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n I(\delta_i \iota_i = 1) \left[\mathbf{Z}_i^T \boldsymbol{\beta} - \log \left\{ \sum_{j \in R_i} w_j(X_i) \exp(\mathbf{Z}_j^T \boldsymbol{\beta}) \right\} \right], \quad (2.2)$$

其中风险集为 $R_i = \{j : (X_j \geq X_i) \cup ((X_j \leq X_i) \cap (\delta_j = 1) \cap (\iota_j \neq 1))\}$.

于是得到对数部分似然相对于 $\boldsymbol{\beta}$ 的前两阶导数为

$$\begin{aligned} \ell'(\boldsymbol{\beta}) &= \sum_{i=1}^n I(\delta_i \iota_i = 1) \left[\mathbf{Z}_i^T - \frac{\sum_{j \in R_i} w_j(X_i) \mathbf{Z}_j^T \exp(\mathbf{Z}_j^T \boldsymbol{\beta})}{\sum_{j \in R_i} w_j(X_i) \exp(\mathbf{Z}_j^T \boldsymbol{\beta})} \right], \\ -\ell''(\boldsymbol{\beta}) &= \sum_{i=1}^n I(\delta_i \iota_i = 1) \left[\frac{\sum_{j \in R_i} w_j(X_i) \mathbf{Z}_j \mathbf{Z}_j^T \exp(\mathbf{Z}_j^T \boldsymbol{\beta})}{\sum_{j \in R_i} w_j(X_i) \exp(\mathbf{Z}_j^T \boldsymbol{\beta})} \right. \\ &\quad \left. - \frac{\{\sum_{j \in R_i} w_j(X_i) \mathbf{Z}_j \exp(\mathbf{Z}_j^T \boldsymbol{\beta})\} \{\sum_{j \in R_i} w_j(X_i) \mathbf{Z}_j \exp(\mathbf{Z}_j^T \boldsymbol{\beta})\}^T}{(\sum_{j \in R_i} w_j(X_i) \exp(\mathbf{Z}_j^T \boldsymbol{\beta}))^2} \right]. \end{aligned} \quad (2.3)$$

3 变量选择

假设 $\boldsymbol{\beta}_0$ 为真实的回归系数, 并满足稀疏性, 记 $S = \{j : \beta_{0j} \neq 0\}$ 为潜在模型的指标集, $s = |S|$ 为 $\boldsymbol{\beta}_0$ 非零分量的个数, 那么稀疏性的假设使得 s 有界且很小. 我们的目标是选出所有重要的协变量,

即识别真实系数 β_0 的显著分量. 对于超高维变量选择, 我们分为两步进行: 第一步, 基于近似对数部分似然的确定联合筛选方法选出 $q > s$ 个粗糙的变量; 第二步, 通过惩罚近似对数部分似然方法对变量进行进一步选择.

3.1 特征筛选

对于 PSH 模型, 我们利用文献 [9] 提出的确定联合筛选方法, 求解约束对数部分似然

$$\hat{\beta}_q = \arg \max_{\beta} \ell(\beta), \quad \text{约束为 } \|\beta\|_0 \leq q, \quad (3.1)$$

其中 $\|\beta\|_0$ 表示 β 的 L_0 范数, 定义为 β 的非零元素的数量; q 是预先指定的大于 β_0 非零个数的数值, 即 $q > s$. 由于在高维情形下几乎不可能直接解决约束最大化问题 (3.1), 我们将 $\ell(\gamma)$ 在 γ 邻域中的点 β 处进行 Taylor 展开, 从而得到近似的对数部分似然函数如下:

$$\ell(\gamma) \approx \ell(\beta) + (\gamma - \beta)^T \ell'(\beta) + \frac{1}{2}(\gamma - \beta)^T \ell''(\beta)(\gamma - \beta),$$

其中 $\ell'(\beta) = \partial \ell(\gamma) / \partial \gamma |_{\gamma=\beta}$, $\ell''(\beta) = \partial^2 \ell(\gamma) / \partial \gamma \gamma^T |_{\gamma=\beta}$. 当 $p < n$ 且 $\ell''_p(\beta)$ 可逆时, 计算 $\ell''_p(\beta)$ 的逆的计算复杂度为 $O(p^3)$, 当 p 较大时, 计算复杂度显然是相当高的. 而对于高维或超高维, $p > n$ 时, $\ell''_p(\beta)$ 是不可逆的.

为了节省处理 $p \times p$ 阶 Hesse 矩阵 $\ell''(\beta)$ 奇异性的计算成本, 使用 $\ell(\gamma)$ 的如下近似:

$$g(\gamma | \beta) = \ell(\beta) + (\gamma - \beta)^T \ell'(\beta) - \frac{u}{2}(\gamma - \beta)^T \mathbf{W}(\gamma - \beta),$$

其中 u 为界定的刻度常数, $\mathbf{W}$ 为对角矩阵 $\text{diag}\{-\ell''(\beta)\}$. 换句话说, 我们用 $u \text{diag}\{\ell''_p(\beta)\} = -u\mathbf{W}$ 来逼近 $\ell''_p(\beta)$.

注 3.1 Xu 和 Chen^[24] 对广义线性模型中的特征筛选方法使用了似然函数 $\ell(\gamma)$ 的线性渐近表达式 $\ell(\beta) + (\gamma - \beta)^T \ell'(\beta)$, 并引入了正则项 $-u\|\gamma - \beta\|^2$, 这一情形相当于 $\mathbf{W} = \mathbf{I}_p$ 的特殊情形. 这里采用 $u \text{diag}\{\ell''(\beta)\}$ 来逼近 $\ell''(\beta)$, 一方面, 通过 Taylor 展开式可知, 当 γ 在 β 邻域时, $g(\gamma | \beta)$ 与 $\ell(\gamma)$ 之间只差 $O(\|\gamma - \beta\|^2)$.

可以证明, 在一定条件下, 对于所有 γ , $g(\beta | \beta) = \ell_p(\beta)$ 且 $g(\gamma | \beta) \leq \ell_p(\gamma)$, 这意味着上升性质, 详情见定理 4.1 及其证明, 这样使得最后得到的估计量逼近理论的估计结果. 由于 $\mathbf{W}$ 是对角矩阵, 对于任意给定 β , $g(\gamma | \beta)$ 是 γ_j ($j = 1, 2, \dots, p$) 的可加函数, 这使得我们对于给定的 β 和 q , 如下约束最大化问题可以求出显示解:

$$\max_{\gamma} g(\gamma | \beta) \quad \text{约束为 } \|\gamma\|_0 \leq q.$$

注意到 $g(\gamma | \beta)$ 的极大值点为 $\tilde{\gamma} = \beta + u^{-1}\mathbf{W}^{-1}\ell'(\beta)$. 记 $\tilde{\gamma}_j$ 的次序统计量为 $|\tilde{\gamma}_{(1)}| \geq \dots \geq |\tilde{\gamma}_{(p)}|$. 通过硬阈值法则, 约束最大化问题的解为

$$\hat{\gamma}_j = \tilde{\gamma}_j I\{|\tilde{\gamma}_j| > |\tilde{\gamma}_{(q+1)}|\} =: H(\tilde{\gamma}_j; q).$$

根据上一节定义的 $\ell'(\beta)$ 和 $\ell''(\beta)$, 我们可以得到对于竞争风险模型的特征筛选方法.

第 1 步 设定初始值 β 为 $\beta^{(0)}$.

第 2 步 对于 $l = 0, 1, 2, \dots$, 执行如下迭代第 2a 和 2b 步直到算法收敛:

第 2a 步 计算 $\tilde{\gamma}^{(l)} = (\tilde{\gamma}_1^{(l)}, \dots, \tilde{\gamma}_p^{(l)})^T = \beta^{(l)} + u_l^{-1} \mathbf{W}^{-1}(\beta^{(l)}) \ell'(\beta^{(l)})$, 及

$$\tilde{\beta}^{(l)} = (H(\tilde{\gamma}_1^{(l)}; q), \dots, H(\tilde{\gamma}_p^{(l)}; q))^T =: \mathbf{H}(\tilde{\gamma}^{(l)}; q).$$

设定 $S_l = \{j: \tilde{\beta}_j^{(l)} \neq 0\}$ 为 $\tilde{\beta}^{(l)}$ 的非零指标集.

第 2b 步 更新 β 为 $\beta^{(l+1)} = (\beta_1^{(l+1)}, \dots, \beta_p^{(l+1)})^T$. 如果 $j \notin S_l$, 则设定 $\beta_j^{(l+1)} = 0$; 否则, 设定 $\{\beta_j^{(l+1)}: j \in S_l\}$ 为子模型 S_l 的极大对数部分似然估计.

记上述算法收敛后得到的 q 个变量的集合为 $\hat{\mathcal{I}}$, 系数估计为 $\hat{\beta}_{\hat{\mathcal{I}}}$.

注 3.2 竞争风险模型的 SJS 方法采用在系数的 L_0 限制下的对数部分似然作为目标函数进行最大化, 来研究对竞争风险数据的超高维变量选择问题, 利用 Taylor 展开对对数部分似然进行逼近, 解决了目标函数过于复杂不易进行优化的难点; 使用经典的硬阈值法则求解近似带约束目标函数的最大化问题, 从而将较小的系数收缩为零, 实现变量选择的目标. 这一方法从思想上区别于传统的 SIS 方法, 并非从各个变量的边际效应入手, 而是对于联合似然进行处理, 从而对于变量之间存在较高相关性, 或者响应变量与协变量之间边际无关但联合相关的情形, 有更好的表现.

3.2 惩罚变量选择

在第 1 步选择 q 个粗糙的变量之后, 可以在 $\hat{\mathcal{I}}$ 上采用最大化如下惩罚近似对数部分似然部分的方法来进行进一步更加精细的变量选择:

$$Q(\beta) = \ell(\beta) - \sum_{j=1}^q p_{\lambda_n}(|\beta_j|), \quad (3.2)$$

其中 $p_{\lambda_n}(\cdot)$ 为选取的恰当的惩罚函数, λ_n 为调节参数.

令 $\mathbf{Z}^T = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)$ 表示由 n 个协变量向量组成的设计矩阵, 对于给定的 β , 记 $\eta = \mathbf{Z}\beta$ 为维数为 $n \times 1$ 的向量, 根据 $\ell(\beta)$ 的定义可知, 它也是 η 的函数, 我们不改变对数似然的记号, 仍记为 $\ell(\eta) := \ell(\beta; \mathbf{Z})$. 定义 $\ell(\eta)$ 对 η 的梯度向量和 Hesse 矩阵分别为 $\mathbf{u} = \partial \ell / \partial \eta$ 和 $\mathbf{H} = -\partial^2 \ell / \partial \eta \partial \eta^T$.

根据文献 [25, 第 8 章], 构造伪响应变量 $\mathbf{y} = \eta + \mathbf{H}^{-1}\mathbf{u}$. 对于任意处于 γ 邻域的 β , $\ell(\gamma)$ 也可以表示为 $\ell(\mathbf{Z}\gamma)$, 将 $\ell(\gamma)$ 和 $\ell(\mathbf{y})$ 在 η 附近进行二阶 Taylor 展开,

$$\ell(\gamma) \approx \ell(\eta) + (\mathbf{Z}\gamma - \eta)^T \mathbf{u} - \frac{1}{2}(\mathbf{Z}\gamma - \eta)^T \mathbf{H}(\mathbf{Z}\gamma - \eta), \quad (3.3)$$

$$\ell(\mathbf{y}) \approx \ell(\eta) + (\mathbf{y} - \mathbf{Z}\beta)^T \mathbf{u} - \frac{1}{2}(\mathbf{y} - \eta)^T \mathbf{H}(\mathbf{y} - \eta). \quad (3.4)$$

令 (3.4) - (3.3), 且由 $\mathbf{y}$ 的定义得到 $\eta = \mathbf{y} - \mathbf{H}^{-1}\mathbf{u}$, 以及 $\mathbf{H}$ 的对称性, 可以推出

$$\begin{aligned} -\ell(\gamma) &\approx -\ell(\mathbf{y}) + (\mathbf{y} - \mathbf{Z}\gamma)^T \mathbf{u} + \frac{1}{2}\{(\mathbf{Z}\gamma - \eta)^T \mathbf{H}(\mathbf{Z}\gamma - \eta) - (\mathbf{H}^{-1}\mathbf{u})^T \mathbf{H}(\mathbf{H}^{-1}\mathbf{u})\} \\ &= -\ell(\mathbf{y}) + (\mathbf{y} - \mathbf{Z}\gamma)^T \mathbf{u} + \frac{1}{2}\{(\mathbf{Z}\gamma - \mathbf{y} + \mathbf{H}^{-1}\mathbf{u})^T \mathbf{H}(\mathbf{Z}\gamma - \mathbf{y} + \mathbf{H}^{-1}\mathbf{u}) - \mathbf{u}^T \mathbf{H}^{-1}\mathbf{u}\} \\ &= -\ell(\mathbf{y}) + (\mathbf{y} - \mathbf{Z}\gamma)^T \mathbf{u} + \frac{1}{2}\{(\mathbf{Z}\gamma - \mathbf{y})^T \mathbf{H}(\mathbf{Z}\gamma - \mathbf{y}) \\ &\quad + 2(\mathbf{Z}\gamma - \mathbf{y})^T \mathbf{u} + \mathbf{u}^T \mathbf{H}^{-1}\mathbf{u} - \mathbf{u}^T \mathbf{H}^{-1}\mathbf{u}\} \\ &= -\ell(\mathbf{y}) + \frac{1}{2}(\mathbf{Z}\gamma - \mathbf{y})^T \mathbf{H}(\mathbf{Z}\gamma - \mathbf{y}). \end{aligned} \quad (3.5)$$

因为 $\ell(\mathbf{y})$ 是 β 的函数, 与 γ 无关, 因此, 对于 γ 邻域内的 β , 有

$$\min_{\gamma} -\ell(\gamma) \approx \min_{\gamma} \frac{1}{2} (\mathbf{Z}\gamma - \mathbf{y})^T \mathbf{H} (\mathbf{Z}\gamma - \mathbf{y}). \quad (3.6)$$

用 $\mathbf{H}$ 的对角部分 $\text{diag}\{h(\boldsymbol{\eta})_i, i = 1, \dots, n\}$ 作为 $\mathbf{H}$ 的替代 (参见文献 [26]), 以及 $u(\boldsymbol{\eta})_i$ 和 $h(\boldsymbol{\eta})_i$ 的表达

$$u(\boldsymbol{\eta})_i = I(\delta_i \iota_i = 1) - \sum_{k \in E_i} I(\delta_k \iota_k = 1) \frac{w_i(X_k) e^{\eta_i}}{\sum_{j \in R_k} w_j(X_k) e^{\eta_j}}, \quad (3.7)$$

$$h(\boldsymbol{\eta})_i = \sum_{k \in E_i} I(\delta_k \iota_k = 1) \left\{ \frac{w_i(X_k) e^{\eta_i}}{\sum_{j \in R_k} w_j(X_k) e^{\eta_j}} - \frac{(w_i(X_k) e^{\eta_i})^2}{(\sum_{j \in R_k} w_j(X_k) e^{\eta_j})^2} \right\}, \quad (3.8)$$

其中 E_i 为集合 $\{k : (\delta_i \iota_i \in \{2, \dots, K\}) \cup ((X_k \leq X_i) \cap (\delta_i \iota_i \notin \{2, \dots, K\}))\}$, 那么由 (3.6), 在第 1 步特征筛选得到 $\hat{\mathcal{I}}$ 后, 求解 (3.2) 的最大化可以近似等价求解如下惩罚近似负对数部分似然的最小化问题来实现:

$$\min_{\gamma_{\hat{\mathcal{I}}}} \frac{1}{n} \sum_{i=1}^n h(\boldsymbol{\eta}_{\hat{\mathcal{I}}})_i (y_{\hat{\mathcal{I}},i} - \mathbf{Z}_{\hat{\mathcal{I}},i}^T \gamma_{\hat{\mathcal{I}}})^2 + \sum_{m \in \hat{\mathcal{I}}} p_{\lambda}(\gamma_m), \quad (3.9)$$

其中 $\mathbf{Z}_{\hat{\mathcal{I}},i}$ 表示 $\mathbf{Z}_i$ 中列指标位于 $\hat{\mathcal{I}}$ 中的子向量, $\gamma_{\hat{\mathcal{I}}}$ 表示分量指标在 $\hat{\mathcal{I}}$ 中的系数向量. $p_{\lambda}(\cdot)$ 表示某个惩罚函数, 如 LASSO、适应性 LASSO、SCAD 或 MCP 等. 其他记号如下:

$$\mathbf{Z}_{\hat{\mathcal{I}}}^T = (\mathbf{Z}_{\hat{\mathcal{I}},1}, \dots, \mathbf{Z}_{\hat{\mathcal{I}},n}), \quad (3.10)$$

$$\boldsymbol{\eta}_{\hat{\mathcal{I}}} = \mathbf{Z}_{\hat{\mathcal{I}}}^T \boldsymbol{\beta}_{\hat{\mathcal{I}}}, \quad (3.11)$$

$$u_{\hat{\mathcal{I}},i} = I(\iota_i = 1) - \sum_{k \in E_i} I(\iota_k = 1) \frac{w_i(X_k) e^{\eta_{\hat{\mathcal{I}},i}}}{\sum_{j \in R_k} w_j(X_k) e^{\eta_{\hat{\mathcal{I}},j}}}, \quad (3.12)$$

$$h(\boldsymbol{\eta}_{\hat{\mathcal{I}}})_i = \sum_{k \in E_i} I(\iota_k = 1) \left\{ \frac{w_i(X_k) e^{\eta_{\hat{\mathcal{I}},i}}}{\sum_{j \in R_k} w_j(X_k) e^{\eta_{\hat{\mathcal{I}},j}}} - \frac{(w_i(X_k) e^{\eta_{\hat{\mathcal{I}},i}})^2}{(\sum_{j \in R_k} w_j(X_k) e^{\eta_{\hat{\mathcal{I}},j}})^2} \right\}, \quad (3.13)$$

$$\mathbf{y}_{\hat{\mathcal{I}}} = \boldsymbol{\eta}_{\hat{\mathcal{I}}} + \mathbf{D}_{\hat{\mathcal{I}}}^{-1} \mathbf{u}_{\hat{\mathcal{I}}}, \quad (3.14)$$

其中 $\mathbf{u}_{\hat{\mathcal{I}}} = (u_{\hat{\mathcal{I}},1}, \dots, u_{\hat{\mathcal{I}},n})^T$, $\mathbf{D}_{\hat{\mathcal{I}}} = \text{diag}(h(\boldsymbol{\eta}_{\hat{\mathcal{I}}})_1, \dots, h(\boldsymbol{\eta}_{\hat{\mathcal{I}}})_n)$.

最后得到稀疏回归系数 β_0 的估计 $\hat{\beta}$ 的迭代算法如下:

对于固定的调节参数 λ ,

(i) 设定初始值 $\boldsymbol{\beta}_{\hat{\mathcal{I}}}^{(0)} = \mathbf{0}$;

(ii) 基于 $\boldsymbol{\beta}_{\hat{\mathcal{I}}}^{(l)}$ ($l = 0, 1, \dots$) 计算 $\boldsymbol{\eta}_{\hat{\mathcal{I}}}^{(l)}$ 、 $\mathbf{u}_{\hat{\mathcal{I}}}^{(l)}$ 、 $\mathbf{D}_{\hat{\mathcal{I}}}^{(l)}$ 和 $\mathbf{y}_{\hat{\mathcal{I}}}^{(l)}$;

(iii) 使用坐标下降算法最小化函数 (3.9), 得到了更新的 $\boldsymbol{\beta}_{\hat{\mathcal{I}}}^{(l+1)}$;

(iv) 令 $l = l + 1$, 重复 (ii) 和 (iii) 步直到 $\hat{\beta}_{\hat{\mathcal{I}}}$ 收敛.

最终使用由 $\hat{S}$ 表示 $\hat{\beta}_{\hat{\mathcal{I}}}$ 的非零分量的指标集, 即真实系数非零指标集 S 的估计, 以及系数估计 $\hat{\beta}$.

4 渐近性质

在描述部分似然和竞争风险模型的正则性条件之前, 首先给出一些记号. 大多数符号改编自文献 [2, 27], 前者对于竞争风险的分分布提出了一种半参数成比例风险模型; 后者针对 Cox 模型引入计

数过程, 并对部分似然估计建立了相合性和渐近正态性. 记 $N_i(t) = I(T_i \leq t, \iota_i = 1)$, $Y_i(t) = 1 - N_i(t-)$, $R_i(t) = r_i(t)Y_i(t)$ (其中 $r_i(t) = I(C_i \geq T_i \wedge t)$) 也可以表示成 $R_i(t) = I\{[(C_i \wedge T_i) \geq t] \cup [(T_j \leq t) \cap (\iota_i \neq 1) \cap (C_i \geq t)]\}$, 通过简单运算可知这两个记号是等价的, 指示个体 i 是否处于风险集中. 结合 IPCW 中赋予的权重, 重新定义 $\tilde{R}_i(t) = w_i(t)Y_i(t)$, 其中如之前定义的个体 i 的相关的时间相依的权重为 $w_i(t) = r_i(t)\hat{G}(t)/\hat{G}(X_i \wedge t)$. $\tilde{R}_i(t)$ 衡量了个体 i 在时间 t 时处在原因 1 的风险下的大小. 假设不会在同一时间出现两个 $N_i(t)$, 也就是在时刻 t , $N_i(t)$ 要么为 0, 要么为 1. 为简单起见, 我们处理有限的时间区间 $[0, \tau]$. 在 PSH 模型中, 随机过程如局部鞅或预测过程的一些性质与样本空间 $(\Omega, \mathcal{F}, \mathcal{P})$ 的子 σ -代数、一个右连续非降族 $(\mathcal{F}(t) : t \in [0, \tau])$ 密切相关, $(\mathcal{F}(t) : t \in [0, \tau])$ 也称为滤波. 在本节中, 定义 $\Lambda_{10}(t) = \int_0^t \lambda_{10}(u)du$.

使用计数过程表示, (3.1) 也可以记作

$$U_2(\beta) = \sum_{i=1}^n \int_0^\infty \left[\mathbf{Z}_i(u) - \frac{\sum_j w_j(u)Y_j(u)\mathbf{Z}_j(u) \exp\{\mathbf{Z}_j^T(u)\beta\}}{\sum_j w_j(u)Y_j(u) \exp\{\mathbf{Z}_j^T(u)\beta\}} \right] w_i(u) dM_i^1(u, \beta), \quad (4.1)$$

其中

$$M_i^1(t, \beta) = N_i(t) - \int_0^t Y_i(u) \lambda_{10}(u) \exp\{\mathbf{Z}_i^T(u)\beta\} du, \quad (4.2)$$

针对完全数据的滤波 $\mathcal{F}^1$ 而言是鞅, 但对于自然的滤波却并不是鞅; 这里 $\mathcal{F}^1$ 是一列右连续的非降子 σ 代数 $\sigma\{N_i(u), Y_i(u)\mathbf{Z}_i(u), u \leq t, i = 1, 2, \dots, n\}$. 为了辅助证明, 我们引入文献 [2] 中介绍的完全删失的情形. 这里的完全删失数据, 指删失仅来源于管理上的不再跟进, 即患者在分析数据之前没有失效. 在这一条件下, 即使对于分析时间之前死亡的个体, 也能观测到潜在的删失时间. 在这种删失机制下的竞争风险集的定义不变, 而对数部分似然记为

$$\ell_{1*}(\beta) = \frac{1}{n} \sum_{i=1}^n \int_0^\infty \left[\mathbf{Z}_i(u)\beta - \frac{\sum_j Y_j^*(u)\mathbf{Z}_j(u) \exp\{\mathbf{Z}_j^T(u)\beta\}}{\sum_j Y_j^*(u) \exp\{\mathbf{Z}_j^T(u)\beta\}} \right] dN_i(u). \quad (4.3)$$

将对数部分似然对 β 求导可以得到具有鞅形式的估计函数

$$U_{1*}(\beta) = \sum_{i=1}^n \int_0^\infty \left[\mathbf{Z}_i(u) - \frac{\sum_j Y_j^*(u)\mathbf{Z}_j(u) \exp\{\mathbf{Z}_j^T(u)\beta\}}{\sum_j Y_j^*(u) \exp\{\mathbf{Z}_j^T(u)\beta\}} \right] dM_i^{1*}(u, \beta), \quad (4.4)$$

其中

$$Y_i^*(t) = I(C_i \geq t)\{1 - N_i(t-)\}, \quad (4.5)$$

且

$$M_i^{1*}(t, \beta) = \int_0^t I(C_i \geq u) dN_i(u) - \int_0^t Y_i^*(u) \lambda_{10}(u) \exp\{\mathbf{Z}_i^T(u)\beta\} du, \quad (4.6)$$

相对于完全删失滤波 $\mathcal{F}^{1*}(t)$ 满足鞅的定义, 这里 $\mathcal{F}^{1*}(t)$ 表示一系列递增的 σ 代数 $\sigma\{I(C_i \geq u), I(C_i \geq u)N_i(u), Y_i^*(u), Y_i^*(u)\mathbf{Z}_i(u), u \leq t, i = 1, 2, \dots, n\}$.

我们参见文献 [2] 中的记号, 对于 $k = 0, 1, 2$, 记

$$\mathbf{s}^{(k)}(\beta, t) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Y_i(t) \mathbf{Z}_i(t)^{\otimes k} \exp\{\mathbf{Z}_i^T(t)\beta\}, \quad \mathbf{a}^{(k)}(\beta, t) = E[\mathbf{S}_1^{(k)}(\beta, t)],$$

$$\begin{aligned} S_1^{(k)}(\beta, t) &= \frac{1}{n} \sum_{i=1}^n Y_i(t) \mathbf{Z}_i(t)^{\otimes k} \exp\{\mathbf{Z}_i^T(t)\beta\}, \\ \hat{S}_2^{(k)}(\beta, t) &= \frac{1}{n} \sum_{i=1}^n w_i(t) Y_i(t) \mathbf{Z}_i(t)^{\otimes k} \exp\{\mathbf{Z}_i^T(t)\beta\}, \end{aligned}$$

且

$$\mathbf{E}(\beta, t) = \frac{S_1^{(1)}(\beta, t)}{S_1^{(0)}(\beta, t)}, \quad \mathbf{V}(\beta, t) = \frac{S_1^{(2)}(\beta, t)}{S_1^{(0)}(\beta, t)} - \mathbf{E}(\beta, t)^{\otimes 2}, \quad (4.7)$$

其中 $\mathbf{Z}_i^{\otimes 0} = 1$, $\mathbf{Z}_i^{\otimes 1} = \mathbf{Z}_i$, $\mathbf{Z}_i^{\otimes 2} = \mathbf{Z}_i \mathbf{Z}_i^T$. 注意到 $S_1^{(0)}(\beta, t)$ 是一个标量, $S_1^{(1)}(\beta, t)$ 和 $\mathbf{E}(\beta, t)$ 为 p 维向量, $S_1^{(2)}(\beta, t)$ 和 $\mathbf{V}(\beta, t)$ 为 $p \times p$ 矩阵.

令 $t_1 < \dots < t_N$ 为排序的观测失效时间. 定义

$$\begin{aligned} Q_j &= \sum_{i=1}^n \int_0^{t_j} \left[\mathbf{Z}_i(u) - \frac{\sum_j Y_j^*(u) \mathbf{Z}_j(u) \exp\{\mathbf{Z}_j^T(u)\beta\}}{\sum_j Y_j^*(u) \exp\{\mathbf{Z}_j^T(u)\beta\}} \right] \times dM_i^{1*}(u, \beta) \\ &= \sum_{i=1}^n \int_0^{t_j} \left[\mathbf{Z}_i - \frac{\sum_{i \in R_j} \mathbf{Z}_i \exp\{\mathbf{Z}_i^T \beta\}}{\sum_{i \in R_j} \exp\{\mathbf{Z}_i^T \beta\}} \right] dM_i^{1*}, \end{aligned}$$

其中 $E[Q_j | \mathcal{F}_{j-1}^{1*}] = Q_{j-1}$, 即 $E[Q_j - Q_{j-1} | \mathcal{F}_{j-1}^{1*}] = 0$, 这里 $\mathcal{F}_{j-1}^{1*} = \mathcal{F}^{1*}(t_{j-1})$. 令 $b_j = Q_j - Q_{j-1}$, 那么 $(b_j)_{j=1,2,\dots}$ 为在 $(\Omega, \mathcal{F}^{1*}, \mathcal{P})$ 上的一列有界鞅的差分, 即 b_j 几乎确定 (a.s.) 有界且对于 $j = 1, 2, \dots$, 有 $E[b_j | \mathcal{F}_{j-1}^{1*}] = 0$. 上述结论可以通过如下局部有界预测过程:

$$\mathbf{H}_i^*(u) = \mathbf{Z}_i(u) - \frac{\sum_j Y_j^*(u) \mathbf{Z}_j(u) \exp\{\mathbf{Z}_j^T(u)\beta_0\}}{\sum_j Y_j^*(u) \exp\{\mathbf{Z}_j^T(u)\beta_0\}} \quad (4.8)$$

和如下引理得到.

引理 4.1 令 N 为计数过程, 满足对于任意 t , $EN(t) < \infty$. 令 $\{\mathcal{F}_t : t \geq 0\}$ 为一个右连续的滤波, 使得

- (1) $M = N - A$ 为一个 $\mathcal{F}_t$ -鞅, 其中 $A = \{A(t) : t \geq 0\}$ 为一个递增 $\mathcal{F}_t$ -预测过程, 满足 $A(0) = 0$;
- (2) H 为有界的 $\mathcal{F}_t$ -预测过程,

那么如下给定的过程 L :

$$L(t) = \int_0^t H(u) dM(u) \quad (4.9)$$

是一个 $\mathcal{F}_t$ -鞅.

下面介绍一些证明定理需要满足的一些正则条件:

(D1) (有界区间) $\Lambda_{10}(\tau) = \int_0^\tau \lambda_{10}(t) dt < \infty$.

(D2) (渐近稳定性) 存在 β_0 的紧邻域 $\mathcal{B}$ 及定义在 $\mathcal{B} \times [0, \tau]$ 上的标量 $\mathbf{a}^{(0)}$ 、向量 $\mathbf{a}^{(1)}$ 和矩阵 $\mathbf{a}^{(2)}$, 使得对 $k = 0, 1, 2$, 有

$$\sup_{t \in [0, \tau], \beta \in \mathcal{B}} \|\mathbf{S}_1^{(k)}(\beta, t) - \mathbf{s}^{(k)}(\beta, t)\| \xrightarrow{P} 0.$$

(D3) (Linderberg 条件) 存在 $\delta > 0$ 使得

$$n^{-1/2} \sup_{i, t} |\mathbf{Z}_i| R_i(t) I\{\mathbf{Z}_i^T \beta_0 > -\delta |\mathbf{Z}_i|\} \xrightarrow{P} 0.$$

(D4) (渐近正则条件) 令 $\mathcal{B}$ 、 $\mathbf{a}^{(0)}$ 、 $\mathbf{a}^{(1)}$ 和 $\mathbf{a}^{(2)}$ 为条件 (D2) 中定义, 并定义 $\mathbf{e} = \mathbf{a}^{(1)}/\mathbf{a}^{(0)}$ 且 $\mathbf{v} = \mathbf{a}^{(2)}/\mathbf{a}^{(0)} - \mathbf{e}^{\otimes 2}$, 对于所有的 $\beta \in \mathcal{B}$, $t \in [0, \tau]$, 有

$$\mathbf{a}^{(1)}(\beta, t) = \frac{\partial \mathbf{a}^{(0)}(\beta, t)}{\partial \beta}, \quad \mathbf{a}^{(2)}(\beta, t) = \frac{\partial^2 \mathbf{a}^{(0)}(\beta, t)}{\partial \beta^2},$$

其中 $\mathbf{a}^{(0)}(\cdot, t)$ 、 $\mathbf{a}^{(1)}(\cdot, t)$ 和 $\mathbf{a}^{(2)}(\cdot, t)$ 为 $\beta \in \mathcal{B}$ 在 $t \in [0, \tau]$ 上一致的连续函数, $\mathbf{a}^{(0)}$ 、 $\mathbf{a}^{(1)}$ 和 $\mathbf{a}^{(2)}$ 在 $\mathcal{B} \times [0, \tau]$ 上有界; $\mathbf{a}^{(0)}$ 在 $\mathcal{B} \times [0, \tau]$ 上有界且不为零, 且矩阵

$$\mathbf{A} = \int_0^\tau \mathbf{v}(\beta_0, t) \mathbf{a}^{(0)}(\beta_0, t) \lambda_{10}(t) dt$$

为正定的.

(D5) 函数 $\mathbf{S}_1^{(0)}(\beta_0, t)$ 和 $\mathbf{a}^{(0)}(\beta_0, t)$ 在 $[0, \tau]$ 上有界且不为 0.

(D6) 存在常数 $C_1, C_2 > 0$, 使得 $\max_{ij} |z_{ij}| < C_1$, $\max_i |\mathbf{Z}_i^T \beta_0| < C_2$.

(D7) $\{b_j\}$ 为一列鞅差, 存在非负常数 c_j 使得对于每个实数 t , 有

$$\mathbb{E}\{\exp(tb_j) \mid \mathcal{F}_{j-1}^{1*}\} \leq \exp\left(\frac{c_j^2 t^2}{2}\right), \quad \text{a.s.} \quad j = 1, 2, \dots, N.$$

对于每个 j , 这些 c_j 的最小值记为 $\eta(b_j)$,

$$|b_j| \leq K_j, \quad \text{a.s.} \quad j = 1, \dots, N,$$

且对于 $b_{j1} < \dots < b_{jk}$ ($k = 1, 2, \dots$), 有 $\mathbb{E}\{b_{j1}, \dots, b_{jk}\} = 0$.

注意有关 $\mathbf{a}^{(0)}$ 、 $\mathbf{a}^{(1)}$ 和 $\mathbf{a}^{(2)}$ 的偏导数条件通过 $\mathbf{S}_1^{(0)}$ 、 $\mathbf{S}_1^{(1)}$ 和 $\mathbf{S}_1^{(2)}$ 满足, 且 $\mathbf{A}$ 自动是半正定的. 此外, 条件中的区间 $[0, \tau]$ 也可以被集合 $\{t : \lambda_{10}(t) > 0\}$ 所代替.

条件 (D1)–(D5) 是比例风险模型性质证明中的一般正则条件 (参见文献 [27]), $\mathbf{S}_1^{(k)}(\beta_0, t)$ 一致收敛到 $\mathbf{a}^{(k)}(\beta_0, t)$. 条件 (D6) 是对一般的经验过程应用集中不等式通常使用的条件. 例如, Huang 等 [28] 在讨论比例风险模型的 LASSO 估计时使用了有界协变量假定. 在稀疏假定下, 真实参数 β_0 的非零个数有界且较小, 这一条件也是合理的. 条件 (D7) 对于部分似然的得分函数 $\ell'(\beta)$ 的渐近性质是必要的, 因为得分函数不能被表示为独立随机向量的和, 但是它可以被表示为鞅的差分序列的和; 事实上, 这是通过 $U_{1*}(\beta)$ 对 $\ell'(\beta)$ 的收敛性得到的一个渐近的结论.

首先, 对于特征筛选算法, 类似文献 [9], 定理 4.1 证明了算法的收敛性.

定理 4.1 假设条件 (D1)–(D4) 成立. 记

$$\rho^{(l)} = \sup_{\tilde{\beta}} [\lambda_{\max}\{\mathbf{W}^{-1/2}(\beta^{(l)})\{-\ell''(\tilde{\beta})\}\mathbf{W}^{-1/2}(\beta^{(l)})\}],$$

其中 $\lambda_{\max}(\mathbf{A})$ 表示矩阵 $\mathbf{A}$ 的最大的特征值. 如果 $u_l \geq \rho^{(l)}$, 那么,

$$\ell(\beta^{(l+1)}) \geq \ell(\beta^{(l)}),$$

其中 $\beta^{(l)}$ 如特征筛选算法中第 2b 步中定义.

证明 将 $\ell(\gamma)$ 在 $\gamma = \beta$ 处 Taylor 展开, 那么有

$$\ell(\gamma) = \ell(\beta) + \ell'(\beta)(\gamma - \beta) + \frac{1}{2}(\gamma - \beta)^T \ell''(\tilde{\beta})(\gamma - \beta),$$

其中 $\tilde{\beta}$ 落在 γ 与 β 之间. 由于

$$\begin{aligned} (\gamma - \beta)^T \{-\ell''(\tilde{\beta})\}(\gamma - \beta) &= \text{tr}[(\gamma - \beta)^T \{-\ell''(\tilde{\beta})\}(\gamma - \beta)] \\ &= (\gamma - \beta)^T \mathbf{W}^{1/2} \mathbf{W}^{-1/2} \{-\ell''(\tilde{\beta})\} \mathbf{W}^{-1/2} \mathbf{W}^{1/2} (\gamma - \beta) \\ &\leq \text{tr}[(\gamma - \beta)^T \mathbf{W}^{1/2} \lambda_{\max}[\mathbf{W}^{-1/2}(\beta) \{-\ell''(\tilde{\beta})\} \mathbf{W}^{-1/2}(\beta)] \mathbf{I}_p \mathbf{W}^{1/2}(\beta) (\gamma - \beta)] \\ &= (\gamma - \beta)^T \mathbf{W}(\beta) (\gamma - \beta) \lambda_{\max}[\mathbf{W}^{-1/2}(\beta) \{-\ell''(\tilde{\beta})\} \mathbf{W}^{-1/2}(\beta)], \end{aligned}$$

其中 $\mathbf{I}_p$ 为 $p \times p$ 单位矩阵, 因此, 如果 $u > \lambda_{\max}[\mathbf{W}^{-1/2}(\beta) \{-\ell''(\tilde{\beta})\} \mathbf{W}^{-1/2}(\beta)] \geq 0$, 因为 $-\ell''(\beta)$ 为非负定的, 那么,

$$\ell(\gamma) \geq \ell(\beta) + \ell'(\beta)(\gamma - \beta) - \frac{u}{2}(\gamma - \beta)^T \mathbf{W}(\beta) (\gamma - \beta).$$

因此有 $\ell(\gamma) \geq g(\gamma | \beta)$, 并且由 $g(\gamma, \beta)$ 的定义可得 $\ell(\beta) = g(\beta | \beta)$. 若记 $\beta_*^{(l+1)} = \arg \max_{\gamma} g(\gamma | \beta^{(l)})$, 对于 $\|\gamma\|_0 \leq q$, 在定理 4.1 的条件下, 有

$$\ell(\beta_*^{(l+1)}) \geq g(\beta_*^{(l+1)} | \beta^{(l)}) \geq g(\beta^{(l)} | \beta^{(l)}) = \ell(\beta^{(l)}).$$

第二个不等式可从 $\beta_*^{(l+1)}$ 的定义以及 $\|\beta_*^{(l+1)}\|_0 = \|\beta^{(l)}\|_0 = q$ 得到. 又由 $\beta^{(l+1)}$ 的定义, 可得 $\|\beta^{(l+1)}\|_0 = q$ 且 $\ell(\beta^{(l+1)}) \geq \ell(\beta_*^{(l+1)}) \geq \ell(\beta^{(l)})$. 定理 4.1 得证. $\square$

定理 4.1 说明选择合适的 u_l 可以保证算法的上升性质, 并提供了在实践中关于如何选择 u_l 的一些参照.

另一个需要注意的性质是文献 [6] 中的确定筛选性质. 在详细介绍之前, 我们重新做一些记号和说明. 用 $\mathcal{I}$ 表示 $\{1, \dots, p\}$ 的任意子集, 对应了协变量为 $\mathbf{Z}_{\mathcal{I}} = \{Z_j, j \in \mathcal{I}\}$ 的子模型和相应的系数 $\beta_{\mathcal{I}} = \{\beta_j, j \in \mathcal{I}\}$. 令 $|\mathcal{I}|$ 表示模型 $\mathcal{I}$ 的大小. 重述真正的模型 $S = \{j : \beta_{0j} \neq 0, 1 \leq j \leq p_n\}$, 其中 $|S| = \|\beta_0\|_0 = s$, $p_n \gg n$ 为协变量的维度. 那么特征筛选的目的就是得到一个子集 $\hat{\mathcal{I}}$ 使得

$$P(S \subset \hat{\mathcal{I}}) \rightarrow 1, \quad \text{当 } n \rightarrow \infty.$$

假设一个选择方法选择了 p 个特征中的 q 个, 使得 $|S| = s < q$, 那么定义

$$\mathcal{I}_+^q = \{\mathcal{I} : S \subset \mathcal{I}; |\mathcal{I}| \leq q\} \quad \text{和} \quad \mathcal{I}_-^q = \{\mathcal{I} : S \not\subset \mathcal{I}; |\mathcal{I}| \leq q\}$$

为过拟合模型和欠拟合模型的集合, 其中 $|\mathcal{I}|$ 表示索引集 $\mathcal{I}$ 的大小. 我们讨论在如条件 (C1)–(C3) 等一些条件下 $\hat{\beta}_q$ 的渐近属性, 其中 p, q 和 β_0 可以依赖于样本大小 n .

(C1) 存在 $w_1, w_2 > 0$ 及某些非负常数 τ_1 和 τ_2 使得 $\tau_1 + \tau_2 < 1/2$ 且

$$\min_{j \in S} |\beta_{0j}| \geq w_1 n^{-\tau_1} \quad \text{及} \quad s < q \leq w_2 n^{\tau_2}.$$

(C2) 对于某个 $0 \leq \kappa < 1 - 2(\tau_1 + \tau_2)$, 有 $\log p = O(n^\kappa)$.

(C3) 存在常数 $c_1 > 0$ 和 $\delta_1 > 0$, 使得对于足够大的 n , 有

$$\lambda_{\min}[-n^{-1} \ell''(\beta_{\mathcal{I}})] \geq c_1,$$

其中 $\beta_{\mathcal{I}} \in \{\beta : \|\beta_{\mathcal{I}} - \beta_{0\mathcal{I}}\|_2 \leq \delta_1\}$, $\mathcal{I} \in \mathcal{I}_+^{2q}$, $\lambda_{\min}[\mathbf{A}]$ 表示矩阵 $\mathbf{A}$ 的最小特征值.

条件 (C1) 说明了满足所提方法的确定筛选性质的一些要求. 第一个是 β_0 的稀疏性, 这使得我们可以用 $|\hat{\mathcal{I}}| = q > s$ 来进行确定筛选. 此外, 它要求 β_0 中的最小分量不会下降太快, 因此可以从渐近序列中将显著的变量检测出来. 同时, (C3) 的约束确保了对于 $|\mathcal{I}| \leq q$, $S \subset \mathcal{I}$ 的可识别性. 条件 (C2) 假定 p 最多以一个指数的速率随 n 发散; 同时, 它表明协变量的数量可以充分大于样本大小. 我们通过以下定理得到近似对数部分似然估计的确定筛选属性.

定理 4.2 假设条件 (C1)–(C3) 和 (D1)–(D7) 成立. 令 $\hat{\mathcal{I}}$ 为通过 (3.1) 得到的大小为 q 的模型, 则有

$$P(S \subset \hat{\mathcal{I}}) \rightarrow 1, \quad \text{当 } n \rightarrow \infty.$$

确定筛选性质是筛选方法的一个吸引人的属性, 因为它确保通过筛选过程选择的模型中保留了真正有效的预测变量.

证明 令 $\hat{\beta}_{\mathcal{I}}$ 为部分似然 $\ell(\beta)$ 基于子模型 $\mathcal{I}$ 的 $\beta_{\mathcal{I}}$ 的极大部分似然估计. 要证定理 4.2, 只要证明 $P\{\hat{\mathcal{I}} \in \mathcal{I}_+^q\} \rightarrow 1$, 而这只需要证明随着 $n \rightarrow \infty$, 有

$$P\left\{\max_{\mathcal{I} \in \mathcal{I}_+^q} \ell(\hat{\beta}_{\mathcal{I}}) \geq \min_{\mathcal{I} \in \mathcal{I}_+^q} \ell(\hat{\beta}_{\mathcal{I}})\right\} \rightarrow 0.$$

对于任意 $\mathcal{I} \in \mathcal{I}_+^q$, 定义 $\mathcal{I}' = \mathcal{I} \cup S \in \mathcal{I}_+^{2q}$. 在条件 (C1) 下, 将 β_0 增广到 $\beta_{0,\mathcal{I}'}$, 用 0 增补 $\mathcal{I}'/S$ 的分量, 考虑 $\beta_{0,\mathcal{I}'}$ 附近的 $\beta_{\mathcal{I}'}$, 满足 $\|\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}\| = w_1 n^{-\tau_1}$ 对于某个 $w_1, \tau_1 > 0$. 那么, 当 n 足够大时, $\beta_{\mathcal{I}'}$ 就落在了 $\beta_{0,\mathcal{I}'}$ 的一个小邻域中, 于是就可以应用条件 (C3). 从条件 (C3) 可得到

$$\begin{aligned} \ell(\beta_{\mathcal{I}'}) - \ell(\beta_{0,\mathcal{I}'}) &= [\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}]^T \ell'(\beta_{0,\mathcal{I}'}) + \frac{1}{2} [\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}]^T \ell''(\tilde{\beta}_{\mathcal{I}'})(\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}) \\ &\leq [\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}]^T \ell'(\beta_{0,\mathcal{I}'}) - \frac{n}{2} [\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}]^T \lambda_{\min}[-n^{-1} \ell''(\tilde{\beta}_{\mathcal{I}'})][\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}] \\ &\leq [\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}]^T \ell'(\beta_{0,\mathcal{I}'}) - \frac{c_1}{2} n \|\beta_{\mathcal{I}'} - \beta_{0,\mathcal{I}'}\|_2^2 \\ &\leq w_1 n^{-\tau_1} \|\ell(\beta_{0,\mathcal{I}'})\|_2 - \frac{c_1}{2} w_2^2 n^{1-2\tau_1}, \end{aligned} \quad (4.10)$$

其中最后一步不等式可通过 Cauchy-Schwarz 不等式得到, $\tilde{\beta}_{\mathcal{I}'}$ 为 $\beta_{\mathcal{I}'}$ 与 $\beta_{0,\mathcal{I}'}$ 之间的某个值. 因此,

$$\begin{aligned} P\{\ell(\beta_{\mathcal{I}'}) - \ell(\beta_{0,\mathcal{I}'}) \geq 0\} &\leq P\left\{\|\ell(\beta_{0,\mathcal{I}'})\|_2 - \left(\frac{c_1 w_1}{2}\right) n^{1-\tau_1} \geq 0\right\} \\ &= P\left\{\|\ell(\beta_{0,\mathcal{I}'})\|_2 \geq \left(\frac{c_1 w_1}{2}\right) n^{1-\tau_1}\right\} \\ &= P\left\{\sum_{j \in \mathcal{I}'} [\ell'_j(\beta_{0,\mathcal{I}'})]^2 \geq \left(\frac{c_1 w_1}{2}\right)^2 n^{2-2\tau_1}\right\}. \end{aligned}$$

又因为

$$\left\{\sum_{j \in \mathcal{I}'} [\ell'_j(\beta_{0,\mathcal{I}'})]^2 \geq \left(\frac{c_1 w_1}{2}\right)^2 n^{2-2\tau_1}\right\} \subset \bigcup_{j \in \mathcal{I}'} \left\{[\ell'_j(\beta_{0,\mathcal{I}'})]^2 \geq \left(\frac{c_1 w_1}{2}\right)^2 n^{2-2\tau_1}\right\},$$

所以有

$$P\{\ell(\beta_{\mathcal{I}'}) - \ell(\beta_{0,\mathcal{I}'}) \geq 0\} \leq P\left\{\sum_{j \in \mathcal{I}'} [\ell'_j(\beta_{0,\mathcal{I}'})]^2 \geq \left(\frac{c_1 w_1}{2}\right)^2 n^{2-2\tau_1}\right\}$$

$$\begin{aligned} &\leq \mathbb{P}\left(\bigcup_{j \in \mathcal{I}'} \left\{ [\ell'_j(\beta_{0, \mathcal{I}'})]^2 \geq \left(\frac{c_1 w_1}{2}\right)^2 n^{2-2\tau_1} \right\}\right) \\ &\leq \sum_{j \in \mathcal{I}'} \mathbb{P}\left\{ [\ell'_j(\beta_{0, \mathcal{I}'})]^2 \geq (2q)^{-1} \left(\frac{c_1 w_1}{2}\right)^2 n^{2-2\tau_1} \right\}. \end{aligned}$$

通过 (C1), 有 $q \leq w_2 n^{\tau_2}$, 那么类似上面的推导有

$$\begin{aligned} \mathbb{P}\left\{ [\ell'_j(\beta_{0, \mathcal{I}'})] \geq (2q)^{-1/2} \left(\frac{c_1 w_1}{2}\right) n^{1-\tau_1} \right\} &\leq \mathbb{P}\left\{ \ell'_j(\beta_{0, \mathcal{I}'}) \geq (2w_2 n^{\tau_2})^{-1/2} \left(\frac{c_1 w_1}{2}\right) n^{1-\tau_1} \right\} \\ &= \mathbb{P}\{\ell'_j(\beta_{0, \mathcal{I}'}) \geq cn^{1-\tau_1-0.5\tau_2}\} \\ &= \mathbb{P}\{\ell'_j(\beta_{0, \mathcal{I}'}) \geq ncn^{-\tau_1-0.5\tau_2}\}, \end{aligned} \quad (4.11)$$

其中 $c = c_1 w_1 / (2\sqrt{2w_2}) > 0$ 为常数. 回顾 (4.1), 我们已经知道 $\ell_p(\beta)$ 的梯度为 $\ell'_p(\beta) := U_2(\beta)$, 且

$$\begin{aligned} U_2(\beta) &= \sum_{i=1}^n \int_0^\infty \left\{ \mathbf{Z}_i - \frac{\sum_j w_j(s) Y_j(s) \mathbf{Z}_j \exp\{\mathbf{Z}_j^\top \beta\}}{\sum_j w_j(s) Y_j(s) \exp\{\mathbf{Z}_j^\top \beta\}} \right\} w_i(s) dN_i(s) \\ &= \sum_{i=1}^n \int_0^\infty \left[\mathbf{Z}_i(u) - \frac{\sum_j w_j(u) Y_j(u) \mathbf{Z}_j(u) \exp\{\mathbf{Z}_j^\top(u) \beta\}}{\sum_j w_j(u) Y_j(u) \exp\{\mathbf{Z}_j^\top(u) \beta\}} \right] w_i(u) dM_i^1(u, \beta) \\ &= \sum_{i=1}^n \int_0^\infty \left[\mathbf{Z}_i(u) - \frac{\hat{\mathbf{S}}_2^{(1)}(\beta, u)}{\hat{\mathbf{S}}_2^{(0)}(\beta, u)} \right] w_i(u) dM_i^1(u, \beta). \end{aligned} \quad (4.12)$$

由 Chebyshev 不等式, 可得

$$\begin{aligned} \mathbb{P}\{\ell'_j(\beta_{0, \mathcal{I}'}) \geq ncn^{-\tau_1-0.5\tau_2}\} &= \mathbb{P}\{\exp(\ell'_{1*j}(\beta_{0, \mathcal{I}'})) \geq \exp(cn^{1-\tau_1-0.5\tau_2})\} \\ &\leq \frac{\mathbb{E} \exp(\ell'_j(\beta_{0, \mathcal{I}'}))}{\exp(cn^{1-\tau_1-0.5\tau_2})}. \end{aligned}$$

为了证明 $\mathbb{E} \exp \ell'_j(\beta_{0, \mathcal{I}'})$ 的有界性, 我们首先引入文献 [29, 定理 2.3.6], 表述为如下引理:

引理 4.2 $Y_n \xrightarrow{L} Y$ 的充分必要条件为, 对于所有有界连续函数 f , 有

$$\mathbb{E} f(Y_n) \rightarrow \mathbb{E} f(Y). \quad (4.13)$$

由文献 [2] 中的证明过程可知,

$$\frac{1}{n} [U_2(\beta) - U_{1*}(\beta)] = \frac{1}{n} \sum_{i=1}^n \int_0^\infty \left\{ \mathbf{Z}_i(u) - \frac{\mathbf{s}^{(1)}(\beta, u)}{\mathbf{s}^{(0)}(\beta, u)} \right\} \{\tilde{w}_i(u) - I(C_i \geq u)\} dM_i^1(u, \beta) + o_p(1)$$

连续且由一个有界函数所决定, 从而对于落在 β_0 紧邻域中的 β , 依概率收敛到 0. 由文献 [29, 定理 2.3.5] 可知, $\frac{1}{n} [U_2(\beta) - U_{1*}(\beta)] \xrightarrow{L} 0$, 那么连续函数 $f(x) = \exp(2x)$ 在紧致支撑集上有界, 因此可以应用引理 4.2, 得到

$$\mathbb{E} \exp \left(\frac{1}{n} [U_2(\beta) - U_{1*}(\beta)] \right) \rightarrow \exp(0) = 1,$$

可以写作 $\mathbb{E} \exp(\frac{2}{n} [U_2(\beta) - U_{1*}(\beta)]) = 1 + o(1)$. 而由 Cauchy-Schwarz 不等式, 有

$$\mathbb{E} \exp \left(\frac{1}{n} U_2(\beta) \right) = \mathbb{E} \exp \left(\frac{1}{n} (U_2(\beta) - U_{1*}(\beta)) + \frac{1}{n} U_{1*}(\beta) \right)$$

$$\begin{aligned}
&= \mathbb{E} \exp \left(\frac{1}{n} (U_2(\beta) - U_{1*}(\beta)) \right) \exp \left(\frac{1}{n} U_{1*}(\beta) \right) \\
&\leq \frac{1}{2} \left(\mathbb{E} \left(\frac{2}{n} (U_2(\beta) - U_{1*}(\beta)) \right) + \mathbb{E} \exp \left(\frac{2}{n} U_{1*}(\beta) \right) \right) \\
&= \frac{1}{2} \left(1 + o(1) + \mathbb{E} \exp \left(\frac{2}{n} U_{1*}(\beta) \right) \right).
\end{aligned}$$

那么, 接下来考虑 $\mathbb{E} \exp(\frac{2}{n} U_{1*}(\beta))$ 的收敛性, 这可以利用 $\frac{1}{n} U_{1*}(\beta)$ 鞅的性质.

注意到, $\mathbb{E}(Q_j | \mathcal{F}_{j-1}^{1*}) = Q_{j-1}$ 对于 Q_j 的每个分量都成立, 且 $Q_N = U_{1*}(\beta) = n\ell_p(\beta) = \sum_{j=1}^N (Q_j - Q_{j-1}) = \sum_{i=1}^N b_i$, 鞅差满足 $|Q_j - Q_{j-1}| \leq \max_{u,i} |b_i(u)| \leq 1$. 接下来用到文献 [30, 引理 1]:

引理 4.3 如果 $\{x_n\}$ 为一列随机变量, 且满足条件 [M], 且 $K_n = 1$, 对于所有的 n , 那么对于任意实数 t , 有

$$\mathbb{E} \left\{ \exp \left(t \sum_{k=1}^n b_{nk} x_k \right) \right\} \leq \exp \left(\frac{t^2}{2} \sum_{k=1}^n b_{nk}^2 \right),$$

其中 $\{b_{nk}, k = 1, 2, \dots, n; n = 1, 2, \dots\}$ 为一列实数, 条件 [M] 如下:

$$|x_n(\omega)| \leq K_n, \quad \text{a.s.} \quad n = 1, 2, \dots$$

且对于 $i_1 < i_2 < \dots < i_k$ ($k = 1, 2, \dots$), 有 $\mathbb{E}\{x_{i_1} x_{i_2} \dots x_{i_k}\} = 0$.

那么, 由条件 (D7), 可以得到

$$\begin{aligned}
\mathbb{E} \exp \left[\frac{2}{n} U_{1*}(\beta) \right] &= \mathbb{E} \exp \left[\frac{2}{n} Q_N \right] = \mathbb{E} \exp \left[\frac{2}{n} \sum_{i=1}^N b_j \right] \\
&\leq \exp \left(\frac{1}{2} \left(\frac{2}{n} \right)^2 \sum_{k=1}^N 1^2 \right) = \exp \left(\frac{2N}{n^2} \right),
\end{aligned}$$

而 $2N/n^2 \rightarrow 0$, 从而, $\exp(2N/n^2) = O(1)$, 那么有

$$\mathbb{E} \exp \left(\frac{1}{n} U_2(\beta) \right) = \frac{1}{2} \left(1 + o(1) + \mathbb{E} \exp \left(\frac{2}{n} U_{1*}(\beta) \right) \right) = O(1),$$

此外, 由文献 [2] 中完全删失情形下的渐近性知, $\frac{1}{n} U_{1*}(\beta)$ 连续, 且在 β_0 的紧致邻域中依概率收敛到某个 β 的有界确定函数, 从而, $\mathbb{E} \exp(\frac{1}{n} U_{1*}(\beta)) = O(1)$, 那么存在常数 $c' > 0$, 使得对于足够大的 n , 有 $\mathbb{E} \exp(\frac{1}{n} U_2(\beta)) \leq \exp(c')$.

故而,

$$\begin{aligned}
\mathbb{P}\{\ell'_j(\beta_{0,I'}) \geq ncn^{-\tau_1-0.5\tau_2}\} &\leq \frac{\mathbb{E} \exp(\ell'_j(\beta_{0,I'}))}{\exp(cn^{1-\tau_1-0.5\tau_2})} \\
&= O(1) \exp(-cn^{1-\tau_1-0.5\tau_2}) \\
&\leq c' \exp(-n^{1-\tau_1-0.5\tau_2}),
\end{aligned}$$

其中 $c' > 0$ 为常数.

同时,

$$\mathbb{P}\{\ell'_j(\beta_{0,I'}) \leq -ncn^{-\tau_1-0.5\tau_2}\} = \mathbb{P}\{-\ell'_j(\beta_{0,I'}) \geq ncn^{-\tau_1-0.5\tau_2}\}$$

$$\leq \frac{\mathbb{E} \exp(-\ell'_j(\beta_{0,I'}))}{\exp(cn^{1-\tau_1-0.5\tau_2})},$$

其中

$$\begin{aligned} \mathbb{E} \exp(-\ell'_j(\beta_{0,I'})) &= \mathbb{E} \exp\left(-\frac{1}{n}U_2(\beta)\right) \\ &= \mathbb{E} \exp\left(\frac{1}{n}(U_{1*}(\beta) - U_2(\beta)) + \left(-\frac{1}{n}U_{1*}(\beta)\right)\right) \\ &\leq \mathbb{E} \exp\left(\frac{2}{n}(U_{1*}(\beta) - U_2(\beta))\right) + \mathbb{E} \exp\left(-\frac{2}{n}U_{1*}(\beta)\right). \end{aligned}$$

容易看到, $\mathbb{E} \exp(\frac{2}{n}(U_{1*}(\beta) - U_2(\beta))) \rightarrow 1$, 令引理 4.3 中的 $t = -\frac{1}{n}$, 可推出

$$\begin{aligned} \mathbb{E} \exp\left[-\frac{2}{n}U_{1*}(\beta)\right] &= \mathbb{E} \exp\left[-\frac{2}{n}Q_N\right] = \mathbb{E} \exp\left[-\frac{2}{n}\sum_{i=1}^N b_i\right] \\ &\leq \exp\left(\frac{1}{2}\left(-\frac{2}{n}\right)^2 \sum_{k=1}^N 1\right) = \exp\left(\frac{2N}{n^2}\right) \rightarrow 1, \end{aligned}$$

从而也有

$$\mathbb{P}\{\ell'_j(\beta_{0,I'}) \leq -ncn^{-\tau_1-0.5\tau_2}\} \leq c' \exp(-n^{1-\tau_1-0.5\tau_2}). \quad (4.14)$$

由于 $q \leq w_2 n^{\tau_2}$, 从而,

$$-(2q)^{-1/2} \left(\frac{c_1 w_1}{2}\right) n^{1-\tau_1} \leq -(2w_2)^{-1/2} n^{-\tau_2} \left(\frac{c_1 w_1}{2}\right) n^{1-\tau_1} = -cn^{-\tau_1-0.5\tau_2}, \quad (4.15)$$

$$\mathbb{P}\left\{\ell'_j(\beta_{0,I'}) \leq -q^{-1/2} \left(\frac{c_1 w_1}{2}\right) n^{1-\tau_1}\right\} \leq \mathbb{P}\{\ell'_j(\beta_{0,I'}) \leq -cn^{-\tau_1-0.5\tau_2}\} \leq c' \exp(-n^{1-\tau_1-0.5\tau_2}). \quad (4.16)$$

结合不等式 (4.14) 和 (4.16) 推出

$$\mathbb{P}\{\ell_p(\beta_{I'}) \geq \ell_p(\beta_{0,I'})\} \leq 4qc' \exp(-n^{1-\tau_1-0.5\tau_2}).$$

因此, 通过 Bonferroni 不等式、条件 (C1) 和 (C2), 有

$$\begin{aligned} \mathbb{P}\left\{\max_{I' \in \mathcal{I}_-^q} \ell_p(\beta_{I'}) \geq \ell_p(\beta_{0,I'})\right\} &\leq \mathbb{P}\left\{\bigcup_{I' \in \mathcal{I}_-^q} \ell_p(\beta_{I'}) \geq \ell_p(\beta_{0,I'})\right\} \\ &\leq \sum_{I' \in \mathcal{I}_-^q} \mathbb{P}\{\ell_p(\beta_{I'}) \geq \ell_p(\beta_{0,I'})\} \\ &\leq 4c' qp^q \exp\{-n^{1-\tau_1-0.5\tau_2}\} \\ &= 4c' \exp\{\log q + q \log p - n^{1-\tau_1-0.5\tau_2}\} \\ &\leq 4c' \exp\{\log w_2 + \tau_2 \log n + w_2 n^{\tau_2} \tilde{c} n^\kappa - n^{1-\tau_1-0.5\tau_2}\} \\ &= 4c' w_2 \exp\{\tau_2 \log n + w_2 \tilde{c} n^{\tau_2+\kappa} - n^{1-\tau_1-0.5\tau_2}\} \\ &= c' a_1 \exp\{\tau_2 \log n + a_2 n^{\tau_2+\kappa} - n^{1-\tau_1-0.5\tau_2}\} \\ &= o(1), \quad \text{当 } n \rightarrow \infty \end{aligned} \quad (4.17)$$

对于某些正常数 $a_1 = 4w_2$ 和 $a_2 = w_2\tilde{c}$ 成立. 最后一个等式可由 $\kappa < 1 - 2(\tau_1 + \tau_2) < 1 - \tau_1 - 1.5\tau_2$ 推导得到. 具体而言, 由 L'Hopital 法则, 有

$$\begin{aligned} & \lim_{n \rightarrow \infty} \exp\{\tau_2 \log n + a_2 n^{\tau_2 + \kappa} - n^{1 - \tau_1 - 0.5\tau_2}\} \\ &= \lim_{n \rightarrow \infty} \frac{n^{\tau_2}}{\exp(n^{1 - \tau_1 - 0.5\tau_2} - a_2 n^{\tau_2 + \kappa})} \\ &= \lim_{n \rightarrow \infty} \frac{\tau_2 n^{\tau_2 - 1}}{\exp(n^{1 - \tau_1 - 0.5\tau_2} - a_2 n^{\tau_2 + \kappa})[(1 - \tau_1 - 0.5\tau_2)n^{-\tau_1 - 0.5\tau_2} - a_2(\tau_2 + \kappa)n^{\tau_2 + \kappa - 1}]} \\ &= \lim_{n \rightarrow \infty} \frac{\tau_2}{\exp(n^{1 - \tau_1 - 0.5\tau_2} - a_2 n^{\tau_2 + \kappa})[(1 - \tau_1 - 0.5\tau_2)n^{1 - \tau_1 - 1.5\tau_2} - a_2(\tau_2 + \kappa)n^\kappa]} \\ &= \lim_{n \rightarrow \infty} \frac{\tau_2}{\exp(n^{1 - \tau_1 - 0.5\tau_2} - a_2 n^{\tau_2 + \kappa})n^\kappa[(1 - \tau_1 - 0.5\tau_2)n^{1 - \tau_1 - 1.5\tau_2 - \kappa} - a_2(\tau_2 + \kappa)]} \\ &= 0. \end{aligned}$$

由条件 (C3) 可知, $\ell_p(\beta_{I'})$ 关于 $\beta_{I'}$ 是凹函数, (4.17) 对于任意 $\beta_{I'}$ 使得 $\|\beta_{I'} - \beta'_{0,I}\| \geq w_1 n^{-\tau_1}$ 也成立. 原因是对于 $\|\beta_{I'} - \beta_{0,I'}\| \geq w_1 n^{-\tau_1}$, 存在 $0 \leq \alpha \leq 1$, 使得 $\|\beta_{I'} - \beta_{0,I'}\| = \frac{1}{1-\alpha} w_1 n^{-\tau_1}$, 即

$$\|(1-\alpha)\beta_{I'} + \alpha\beta_{0,I'} - \beta_{0,I'}\| = w_1 n^{-\tau_1}, \quad (4.18)$$

那么, 由 $\ell_p(\cdot)$ 的凹性, 有

$$\begin{aligned} & \mathbb{P}\left\{\max_{I \in \mathcal{I}_-^q} \ell_p((1-\alpha)\beta_{I'} + \alpha\beta_{0,I'}) \geq \ell_p(\beta_{I'})\right\} \\ & \leq \mathbb{P}\left\{\max_{I \in \mathcal{I}_-^q} (1-\alpha)\ell_p(\beta_{I'}) + \alpha\ell_p(\beta_{0,I'}) \geq \ell_p(\beta'_{0,I})\right\} \\ & = \mathbb{P}\left\{\max_{I \in \mathcal{I}_-^q} \ell_p(\beta_{I'}) \geq \ell_p(\beta'_{0,I})\right\} = o(1). \end{aligned} \quad (4.19)$$

对于任意 $I \in \mathcal{I}_-^q$, $\hat{\beta}_I$ 为极大部分似然估计, 令 $\check{\beta}_{I'}$ 为 $\hat{\beta}_I$ 的增广, 使得其中指标落在 I'/I 中的元素为零 (即 $I' = \{I \cup (S/I)\} \cup (I'/S)$), 则 $\ell_p(\check{\beta}_{I'}) = \ell_p(\hat{\beta}_I)$. 因为对于任意 $I \in \mathcal{I}_+^q$, 有

$$\ell_p(\hat{\beta}_I) \geq \ell_p(\beta_{0,I'}) = \ell_p(\beta_0),$$

也即 $\min_{I \in \mathcal{I}_+^q} \ell_p(\hat{\beta}_I) \geq \ell_p(\beta_{0,I'})$, 那么, 要证

$$\mathbb{P}\left\{\max_{I \in \mathcal{I}_-^q} \ell_p(\hat{\beta}_I) \geq \min_{I \in \mathcal{I}_+^q} \ell_p(\hat{\beta}_I)\right\} \rightarrow 0, \quad (4.20)$$

由于

$$\begin{aligned} \mathbb{P}\left\{\max_{I \in \mathcal{I}_-^q} \ell_p(\hat{\beta}_I) \geq \min_{I \in \mathcal{I}_+^q} \ell_p(\hat{\beta}_I)\right\} & \leq \mathbb{P}\left\{\max_{I \in \mathcal{I}_-^q} \ell_p(\hat{\beta}_I) \geq \ell_p(\beta_{0,I'})\right\} \\ & = \mathbb{P}\left\{\max_{I \in \mathcal{I}_-^q} \ell_p(\check{\beta}_{I'}) \geq \ell_p(\beta_{0,I'})\right\}, \end{aligned} \quad (4.21)$$

那么, 由 (4.19), 只要证明 $\|\check{\beta}_{I'} - \beta_{0,I'}\| \geq w_1 n^{-\tau_1}$ 即可.

通过条件 (C1), 可以看到

$$\|\check{\beta}_{I'} - \beta_{0,I'}\|_2 = \|\check{\beta}_{S \cup (I/S) \cup (I'/I)} - \beta_{0, S \cup (I'/S)}\|_2 \geq \|\check{\beta}_{(S/I) \cup (I'/I)} - \beta_{0,S}\|_2$$

$$\geq \|\beta_{0,S/\mathcal{I}} - \beta_{0,S}\|_2 \geq \min_{j \in \mathcal{I}'/\mathcal{I}} |\beta_j| \geq w_1 n^{-\tau_1}.$$

综上有

$$P\left\{\max_{\mathcal{I} \in \mathcal{I}_+^q} \ell_p(\hat{\beta}_{\mathcal{I}}) \geq \min_{\mathcal{I} \in \mathcal{I}_+^q} \ell_p(\hat{\beta}_{\mathcal{I}})\right\} \leq P\left\{\max_{\mathcal{I} \in \mathcal{I}_+^q} \ell_p(\check{\beta}_{\mathcal{I}'}) \geq \ell_p(\beta_{0,\mathcal{I}'})\right\} = o(1).$$

定理 4.2 得证. $\square$

在我们证明最后的估计量的 oracle 性质之前, 我们先做一些记号. 令

$$\begin{aligned} \beta_{0,q} &= (\beta_{01}, \dots, \beta_{0q})^T = (\beta_{01}, \dots, \beta_{0s}, \dots, \beta_{q0})^T = (\beta_{10}^T, \beta_{20}^T)^T, \\ \beta_{10} &= (\beta_{01}, \dots, \beta_{0s})^T, \\ \beta_{20} &= (\beta_{0(s+1)}, \dots, \beta_{0q}). \end{aligned}$$

不失一般性, 假设 $\beta_{20} = \mathbf{0}$. 回顾 s 为 β_{10} 的分量个数, 即真实非零分量的个数. 记

$$\begin{aligned} a_n &= \max\{p'_{\lambda_n}(|\beta_{0j}|) : \beta_{0j} \neq 0\}, \\ b_n &= \max\{p''_{\lambda_n}(|\beta_{0j}|) : \beta_{0j} \neq 0\}. \end{aligned} \quad (4.22)$$

我们将证明存在一个惩罚部分似然估计量以速率 $O_p(n^{-1/2} + a_n)$ 收敛, 并且满足 oracle 性质.

定理 4.3 假设条件 (C1)–(C3) 和 (D1)–(D7) 成立. 如果 $b_n \rightarrow 0$, 那么存在 (3.2) 中 $Q(\beta)$ 的局部极大值 $\hat{\beta}$ 使得 $\|\hat{\beta} - \beta_{0,q}\| = O_p(n^{-1/2} + a_n)$, 其中 a_n 如 (4.22) 给出.

证明 设 $\hat{\beta}_{\hat{\mathcal{I}}}$ 为在 $\hat{\mathcal{I}}$ 上最大化 (3.2) 得到 $\beta_{\hat{\mathcal{I}}}$ 的估计量, 其中 $\hat{\mathcal{I}}$ 为第一步特征筛选得到的基数为 q 的 S 的估计量. 我们要证明 $\hat{\beta}_{\hat{\mathcal{I}}}$ 的相合性, 即证明 $\hat{\beta}_{\hat{\mathcal{I}}} - \beta_{0,q} = O_p(n^{-1/2})$, 其中 $\beta_{0,q} = (\beta_{01}, \dots, \beta_{0s}, \dots, \beta_{0q})$ 表示将真实系数向量扩充到 q 个分量的系数向量. $\hat{\beta}_{\hat{\mathcal{I}}}$ 是我们最终得到的估计量, 将其简记为 $\hat{\beta}$ 并不会引起记号上的模糊. 根据文献 [31] 中 $O_p(\cdot)$ 的定义, 对于任意 $\epsilon > 0$, 存在足够大的 C_ϵ 和 n_ϵ , 使得对于所有 $n > n_\epsilon$, 有

$$P\{\|\sqrt{n}(\hat{\beta} - \beta_{0,q})\| \leq C_\epsilon\} \geq 1 - \epsilon. \quad (4.23)$$

由定理 4.2 知, $P(S \subset \hat{\mathcal{I}}) = 1 - o(1)$; 由 $o(1)$ 定义知, 对于任意 $\epsilon/2 > 0$, 存在足够大的 $n_{1\epsilon}$, 使得对于所有的 $n > n_{1\epsilon}$, 有

$$P(S \subset \hat{\mathcal{I}}) \geq 1 - \frac{\epsilon}{2}. \quad (4.24)$$

另一方面, 假设对于任意指标集合 $\mathcal{I} \subset \mathcal{I}_+^q$, 即 $|\mathcal{I}| = q$, $S \subset \mathcal{I}$, 定义 $\hat{\beta}_{\mathcal{I}}$ 为在指标集 $\mathcal{I}$ 上最大化 (3.2) 得到的估计量, 记为 $\hat{\beta}^*$. 而 $S \subset \hat{\mathcal{I}}$ 等价于 $\hat{\mathcal{I}} \subset \mathcal{I}_+^q$. 由 $\mathcal{I} \subset \mathcal{I}_+^q$ 的任意性, 有 $\{\|\sqrt{n}(\hat{\beta}^* - \beta_{0,q})\| \leq C_\epsilon\} \subset \{\|\sqrt{n}(\hat{\beta} - \beta_{0,q})\| \leq C_\epsilon \mid S \subset \hat{\mathcal{I}}\}$, 即

$$P\{\|\sqrt{n}(\hat{\beta}^* - \beta_{0,q})\| \leq C_\epsilon\} \leq P\{\|\sqrt{n}(\hat{\beta} - \beta_{0,q})\| \leq C_\epsilon \mid S \subset \hat{\mathcal{I}}\}. \quad (4.25)$$

假设我们证明了 $\hat{\beta}^*$ 的相合性, 表示为 $\hat{\beta}^* - \beta_{0,q} = O_p(n^{-1/2})$, 则由定义知, 对于任意 $\epsilon/2 > 0$, 存在足够大的 C_ϵ 和 $n_{2,\epsilon}$, 使得对于所有 $n > n_{2,\epsilon}$, 有

$$P\{\|\sqrt{n}(\hat{\beta}^* - \beta_{0,q})\| \leq C_\epsilon\} \geq 1 - \frac{\epsilon}{2}, \quad (4.26)$$

从而,

$$\begin{aligned}
 P\{\|\sqrt{n}(\hat{\beta} - \beta_{0,q})\| \leq C_\epsilon\} &= P\{\|\sqrt{n}(\hat{\beta} - \beta_{0,q})\| \leq C_\epsilon \mid S \subset \hat{\mathcal{I}}\}P\{S \subset \hat{\mathcal{I}}\} \\
 &\quad + P\{\|\sqrt{n}(\hat{\beta} - \beta_{0,q})\| \leq C_\epsilon \mid S \not\subset \hat{\mathcal{I}}\}P\{S \not\subset \hat{\mathcal{I}}\} \\
 &\geq P\{\|\sqrt{n}(\hat{\beta}^* - \beta_{0,q})\| \leq C_\epsilon\}P\{S \subset \hat{\mathcal{I}}\} \\
 &\geq \left(1 - \frac{\epsilon}{2}\right)\left(1 - \frac{\epsilon}{2}\right) \geq 1 - \epsilon,
 \end{aligned} \tag{4.27}$$

对于所有的 $n \geq n_\epsilon = \max\{n_{1,\epsilon}, n_{2,\epsilon}\}$ 都成立. 也就是说, 只要我们证明了 $\hat{\beta}^*$ 的相合性对于任意 $\mathcal{I} \subset \mathcal{I}_+^q$ 均成立, 那么 $\hat{\beta}$ 也具有相应的相合性. 在接下来的证明中, 简单地记 $\beta_0 := \beta_{0,q}$, 因为 $\beta_{0j} = 0$, $j = q+1, \dots, p$, 这样的记号并不影响证明和结论.

考虑球 $B(C) = \{\beta_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$, 边界为 $\partial B(C)$, 其中 $C > 0$ 为常数, $\alpha_n = n^{-1/2} + a_n$, 那么由条件 (L3), 有 $\alpha_n = O_p(n^{-1/2})$. 要证 $\hat{\beta}^* - \beta_0 = O_p(n^{-1/2})$, 其中 $\hat{\beta}^*$ 为通过在 $\mathcal{I} \subset \mathcal{I}_+^q$ 上最小化 (3.2) 得到的估计量, 那么只要证明对于任意给定 $\epsilon > 0$, 存在一个足够大的常数 C 使得

$$P\left\{\sup_{\partial B(C)} Q(\beta_0 + \alpha_n \mathbf{u}) < Q(\beta_0)\right\} \geq 1 - \epsilon. \tag{4.28}$$

这说明以概率 $1 - \epsilon$ 在 $B(C)$ 中存在 $Q(\beta)$ 的一个局部最大值点, 即存在一个局部极大值点使得

$$\|\hat{\beta}^* - \beta_0\| = O_p(\alpha_n) = O_p(n^{-1/2}).$$

记 $D_n(\mathbf{u}) = Q(\beta_0 + \alpha_n \mathbf{u}) - Q(\beta_0)$, 则有

$$\begin{aligned}
 D_n(\mathbf{u}) &\leq \ell(\beta_0 + \alpha_n \mathbf{u}) - \ell(\beta_0) - \sum_{j=1}^s \{p_{\lambda_n}(|\beta_{0j} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{0j}|)\} \\
 &=: I_1 - I_2,
 \end{aligned} \tag{4.29}$$

那么由 Taylor 展开, 有

$$\begin{aligned}
 I_1 &= \ell(\beta_0 + \alpha_n \mathbf{u}) - \ell(\beta_0) \\
 &= \ell'(\beta_0)^T \alpha_n \mathbf{u} - \frac{1}{2} \alpha_n^2 \mathbf{u}^T \{\Omega(\beta_0) + o_p(1)\} \mathbf{u} \\
 &= -\frac{1}{2} \alpha_n^2 \mathbf{u}^T \{\Omega(\beta_0) + o_p(1)\} \mathbf{u} + O_p(n^{-1/2} \alpha_n \|\mathbf{u}\|),
 \end{aligned} \tag{4.30}$$

其中

$$\Omega(\beta_0) = \lim_{n \rightarrow \infty} n^{-1} \frac{d(-U_2(\beta))}{d\beta} \bigg|_{\beta_0} = \lim_{n \rightarrow \infty} n^{-1} \frac{d(-U_{1*}(\beta))}{d\beta} \bigg|_{\beta_0},$$

其定义也可在文献 [2] 中找到. 最后一个等式可以由 $\ell'(\beta_0) = \frac{1}{n} U_2(\beta_0)$ 和 $n^{-1/2} U_2(\beta_0)$ 的渐近性质得到. 注意到 $\Omega(\beta_0)$ 是正定的, 因此, (4.30) 中第一部分的阶数为 $C^2 \alpha_n^2$, 而因为 $n^{-1/2} \alpha_n = O_p(\alpha_n^2)$, 从而第二部分阶数为 $C^2 \alpha_n^2$. 因此, 当选择足够大的 C 时, I_1 由 (4.30) 中的第一部分一致决定.

另一方面, 通过 Taylor 展开和 Cauchy-Schwarz 不等式, 有

$$I_2 = \sum_{j=1}^s p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)$$

$$\begin{aligned}
&= \sum_{j=1}^s p'_{\lambda_n}(|\beta_{0j}|) \text{sgn}(\beta_{0j}) \alpha_n u_j + p''_{\lambda_n}(|\beta_{j0}|) \alpha_n^2 u_j^2 (1 + o_p(1)) \\
&\leq \sqrt{s} a_n \alpha_n \|\mathbf{u}\| + b_n \alpha_n^2 \|\mathbf{u}\|^2 \\
&= O_p(C \alpha_n^2) + o_p(C^2 \alpha_n^2).
\end{aligned}$$

最后等式的第二项是因为 $b_n \rightarrow 0$, 从而当 C 足够大时, $I_1 - I_2$ 由 $-\frac{1}{2} \alpha_n^2 \mathbf{u}^T \{\boldsymbol{\Omega}(\boldsymbol{\beta}_0)\} \mathbf{u}$ 决定, 其符号是负的, 即 $D_n(\mathbf{u}) < 0$. 定理 4.3 得证. $\square$

从定理 4.3 看出, 选择合适的 λ_n , 存在一个 $\sqrt{n}$ - 相合的惩罚部分似然估计量, 只要

$$a_n = O(n^{-1/2}).$$

记

$$\boldsymbol{\Lambda} = \text{diag}\{p''_{\lambda_n}(|\beta_{01}|), \dots, p''_{\lambda_n}(|\beta_{0s}|)\}, \quad (4.31)$$

$$\mathbf{b} = (p'_{\lambda_n}(|\beta_{01}|) \text{sgn}(\beta_{01}), \dots, p'_{\lambda_n}(|\beta_{0s}|) \text{sgn}(\beta_{0s}))^T, \quad (4.32)$$

其中 s 为 β_{10} 的分量个数.

定理 4.4 (Oracle 性质) 假设惩罚函数 $p_{\lambda_n}(\cdot)$ 满足如下条件:

- (L1) $\lambda_n \rightarrow 0, \sqrt{n} \lambda_n \rightarrow \infty$;
- (L2) $\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0+} p'_{\lambda_n}(\theta) / \lambda_n > 0$;
- (L3) $a_n = O(n^{-1/2})$,

那么在定理 4.3 的条件下, 依概率趋近于 1, 定理 4.3 中的 $\sqrt{n}$ - 相合的极大值点 $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ 满足:

- (i) (稀疏性) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;
- (ii) (渐近正态性)

$$\sqrt{n}(\boldsymbol{\Omega}_1(\boldsymbol{\beta}_{10}) + \boldsymbol{\Lambda})\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\boldsymbol{\Omega}_1(\boldsymbol{\beta}_{10}) + \boldsymbol{\Lambda})^{-1} \mathbf{b}\} \rightarrow N\{\mathbf{0}, \boldsymbol{\Sigma}_1\},$$

其中 $\boldsymbol{\Omega}_1(\boldsymbol{\beta}_{10})$ 为 $\boldsymbol{\Omega}(\boldsymbol{\beta}_0)$ 的前 $s \times s$ 阶子矩阵, $\boldsymbol{\Sigma}$ 为文献 [2] 中关于得分函数 $n^{-1/2} U_2(\boldsymbol{\beta}_0)$ 渐近收敛到 Gauss 过程的协方差矩阵, $\boldsymbol{\Sigma}_1$ 为其前 $s \times s$ 阶子矩阵.

证明 对于 (i) 部分稀疏性的证明, 我们将其表述成如下引理:

引理 4.4 假设条件 (L1)–(L3) 和定理 4.3 的条件成立, 那么以趋近于 1 的概率, 对于任意给定的满足 $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_p(n^{-1/2})$ 的 $\boldsymbol{\beta}_1$, 以及任意常数 C , 有

$$Q\{(\boldsymbol{\beta}_1^T, \mathbf{0})^T\} = \max_{\|\boldsymbol{\beta}_2\| \leq C n^{-1/2}} Q\{(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T\},$$

其中 $\boldsymbol{\beta}_{10} = (\beta_{01}, \dots, \beta_{0s})^T$ 为 $\boldsymbol{\beta}_0$ 的前 s 个分量.

只要证明随着 $n \rightarrow \infty$, 以趋近于 1 的概率, 对于任意满足 $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$ 和 $\|\boldsymbol{\beta}_2\| \leq C n^{-1/2}$ 的 $\boldsymbol{\beta}_1$ 和 $\boldsymbol{\beta}_2$, $\partial Q(\boldsymbol{\beta}) / \partial \beta_j$ 与 β_j 符号相反, 对于 $\beta_j \in (-C n^{-1/2}, C n^{-1/2})$, $j = s+1, \dots, q$. 由 (4.30), 令 $f(\boldsymbol{\beta}) = -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T (-\ell''(\boldsymbol{\beta}_0))(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, 从而有

$$\begin{aligned}
n\ell(\boldsymbol{\beta}) &= n\ell(\boldsymbol{\beta}_0) + nf(\boldsymbol{\beta}) + n\ell'(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o_p(n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2) \\
&= n\ell(\boldsymbol{\beta}_0) + nf(\boldsymbol{\beta}) + O_p(\sqrt{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|),
\end{aligned}$$

且有 $f(\beta_0) = 0$, $f'(\beta_0) = 0$, 且 $f(\beta) = -\frac{1}{2}(\beta - \beta_0)^T(\Omega(\beta_0) + o_p(1))(\beta - \beta_0)$. 那么由 Taylor 展开, 有

$$f'(\beta) = f'(\beta_0) + f''(\beta_0)(\beta - \beta_0) + O_p(\|\beta - \beta_0\|^2),$$

从而,

$$\begin{aligned} \frac{\partial nQ(\beta)}{\partial \beta_j} &= \frac{\partial n\ell(\beta)}{\partial \beta_j} - np'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \\ &= n\ell'_j(\beta_0) + n(f''_j(\beta_0))^T(\beta - \beta_0) + O_p(n\|\beta - \beta_0\|^2) - np'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \\ &= O_p(n^{1/2}) + O_p(n\|\beta - \beta_0\|) + O_p(n\|\beta - \beta_0\|^2) - np'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \\ &= O_p(n^{1/2})(1 - n^{1/2}p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)), \end{aligned}$$

其中 $f''_j(\beta_0)$ 为 $f''(\beta_0)$ 的第 j 行. 由条件 (L1) 和 (L2) 有 $\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0+} n^{1/2}p'_{\lambda_n}(\theta) \rightarrow \infty$, 因此, 最后一个等式由 $-n^{1/2}p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)$ 决定, 从而, $\frac{\partial nQ(\beta)}{\partial \beta_j}$ 与 β_j 符号相反得证. 从而稀疏性得证.

对于 (ii) 部分渐近正态性, 我们首先用定理 4.3 的证明可知, 存在一个 $\hat{\beta}_1$ 为 $Q\{(\beta_1^T, \mathbf{0})^T\}$ 的 $\sqrt{n}$ -相合的局部极大值点, 满足如下的似然方程:

$$\left. \frac{\partial Q(\beta)}{\partial \beta_j} \right|_{\beta=(\hat{\beta}_1^T, \mathbf{0})^T} = 0, \quad j = 1, \dots, s. \quad (4.33)$$

从而由 Taylor 展开, 对于 $j = 1, \dots, s$, 成立

$$\begin{aligned} \left. \frac{\partial \ell(\beta)}{\partial \beta_j} \right|_{\beta=(\hat{\beta}_1^T, \mathbf{0})^T} - p'_{\lambda_n}(|\hat{\beta}_j|) &= \ell'_j(\beta_0) - \{\Omega_j(\beta_0) + o_p(1)\}^T(\hat{\beta} - \beta_0) - (p'_{\lambda_n}(|\beta_{0j}|)\text{sgn}(\beta_{0j}) \\ &\quad + \{p''_{\lambda_n}(|\beta_{0j}|) + o_p(1)\}(\hat{\beta}_j - \beta_{0j})), \end{aligned}$$

其中 $\Omega_j(\beta_0)$ 表示 $\Omega(\beta_0)$ 的第 j 行. 由文献 [2] 中关于 $n^{-1/2}U_2(\beta_0)$ 的渐近性可知, 若记 $U_{21}(\beta)$ 为得分函数 $U_2(\beta)$ 的前 s 个分量, 则可知随着 $n \rightarrow \infty$, 有 $n^{-1/2}U_{21}(\beta_0) \rightarrow N(\mathbf{0}, \Sigma_1)$. 那么由 Slutsky 定理, 可得

$$\sqrt{n}(\Omega_1(\beta_{10}) + \Lambda)\{\hat{\beta}_1 - \beta_{10} + (\Omega_1(\beta_{10}) + \Lambda)^{-1}\mathbf{b}\} \rightarrow N\{\mathbf{0}, \Sigma_1\}.$$

定理 4.4 得证. □

5 数值研究

5.1 Monte Carlo 模拟

本节通过 Monte Carlo 方法进行数值模拟来研究超高维度竞争风险模型的变量选择方法的性能, 包括 CR-SIS (sure independent screening for competing risks) 和 CR-SJS (sure joint screening for competing risks) 两种方法的比较. 为了公平比较, 令 CR-SIS 的模型大小与 CR-SJS 的模型大小相同. 在模拟中, 预测变量 $\mathbf{Z}$ 从 p 维正态分布中产生, 均值为 0, 协方差矩阵为 $\Sigma = (\sigma_{ij})$. 考虑两种常用的协方差结构:

(S1) Σ 为对称矩阵, 即对于 $i \neq j$, 有 $\sigma_{ij} = \rho$; 对于 $i = j$, 有 $\sigma_{ij} = 1$. ρ 分别取 0.25、0.50 和 0.75.

(S2) Σ 具有自回归结构, 即 $\sigma_{ij} = \rho^{|i-j|}$. 同样考虑 ρ 分别取 0.25、0.50 和 0.75.

真实系数 β 的产生机制有如下两种:

(b1) $\beta_1 = \beta_2 = \beta_3 = 5$, $\beta_4 = -15\rho$, 其他的 β_j 等于 0.

(b2) $\beta_j = (-1)^U(a + |V_j|)$, 对于 $j = 1, 2, 3, 4$, 其中 $a = 4 \log n / \sqrt{n}$, $U \sim \text{Bernoulli}(0.4)$, $V_j \sim N(0, 1)$.

在 (S1) 和 (b1) 的设定下, 当 $\rho \neq 0$ 时, Z_4 是与生存时间联合相依但边际独立的. 这一设定对于边际 SIS 是一种挑战.

在模拟中, 考虑样本量 $(n, p) = (100, 500)$ 和 $(n, p) = (200, 800)$ 两种情形, 并对每种情形进行 100 次 Monte Carlo 模拟. 利用如下两种标准进行比较:

P_s : 在 100 次重复中对于给定样本量 m 某个单独的有效预测变量 Z_s 被选出来的比例.

P_a : 在 100 次重复中对于给定样本量 m 所有有效预测变量 Z_s 被选出来的比例.

除了上述 4 种情形之外, 我们还对于存在离散协变量的情形进行了模拟, 其中 $Z_1 \sim \text{Pois}(1)$, $Z_5 \sim \text{Bernoulli}(0.7)$, $(Z_2, Z_3, Z_4, Z_6, \dots, Z_p) \sim N(0, \Sigma)$, $\Sigma_{ij} = \rho^2$, $\beta_0 = (5, 5, -5, 5, 0, \dots, 0)^T$. 并将模拟 500 次的结果放在了表 1 中.

我们根据文献 [2] 中的方法来产生竞争风险模型的数据. 令 p 个协变量 $\mathbf{z} = (z_1, \dots, z_p)$, 对应的系数为: 对于原因 1, $\beta_1 = (\beta_{11}, \dots, \beta_{1p})^T$; 对于原因 2, $\beta_2 = -\beta_1$. 原因 1 的 CIF 为

$$F_1(t; \mathbf{z}) = P(T \leq t, \iota = 1 | \mathbf{z}) = 1 - [1 - p\{1 - \exp(-t)\}]^{\exp(\beta_1^T \mathbf{z})}, \quad (5.1)$$

p 的值设置为 0.3. 由 $P(\iota = 2 | \mathbf{z}) = 1 - P(\iota = 1 | \mathbf{z})$, 以及将条件 CIF, $P(T \leq t | \iota = 2, \mathbf{z})$ 设为速率为 $\exp(\beta_2^T \mathbf{z})$ 的指数分布, 可以得到原因 2 的 CIF. 删失时间是从指数分布 $\exp(10)$ 独立生成的.

文献 [9] 指出, SJS 方法的表现依赖于如下的因素: 协方差矩阵的结构、 β 的取值、特征或变量的维数和样本量 n , 且在生存数据分析中, 统计方法的表现依赖于删失率. 于是, 我们将模拟中不同情形对应的删失率呈现在表 2 中, 包括不同的 ρ 、不同的协方差矩阵结构, 以及不同的 β . 删失率的范围从 20% 到 36%.

表 1 存在离散协变量时 P_s 和 P_a 的比例

n	p	ρ	CR-SJS					CR-SIS				
			P_s				P_a	P_s				P_a
			Z1	Z2	Z3	Z4	ALL	Z1	Z2	Z3	Z4	ALL
100	500	0.25	0.942	0.980	0.970	0.994	0.942	0.570	0.880	0.238	0.870	0.220
100	500	0.50	0.948	0.990	0.980	0.982	0.948	0.586	0.800	0.004	0.824	0.004
100	500	0.75	0.930	0.946	0.928	0.952	0.906	0.666	0.690	0.000	0.682	0.000
200	800	0.25	0.992	0.990	0.990	1.000	0.990	0.874	0.990	0.390	1.000	0.390
200	800	0.50	0.980	0.997	1.000	0.994	0.980	0.880	0.974	0.000	1.000	0.000
200	800	0.75	0.930	0.952	0.934	0.950	0.912	0.900	0.972	0.000	0.982	0.000

表 2 删失率

Σ	$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
	(b1)	(b2)	(b1)	(b2)	(b1)	(b2)
(S1)	0.23	0.35	0.24	0.29	0.29	0.26
(S2)	0.23	0.36	0.20	0.31	0.22	0.27

表 3 报告了当 $\mathbf{Z}$ 的协方差矩阵为 (S1) 结构时, 重要预测变量的 P_s 和 P_a 的模拟表现. 注意到, 对于 (S1) 和 (b1) 的设计, Z_4 是联合相依而边际独立于生存时间的, 对于所有的 $\rho \neq 0$ 成立. 这一设定对于筛选方法是一个很大的挑战, 特别是边际独立筛选的方法. 从表 3 中也可以看出, 在 $n = 100$, $p = 500$ 和 $n = 200$, $p = 800$ 的情形下, CR-SIS 都很难准确识别出 Z_4 , 从而具有极低的 P_a , 而 CR-SJS 则表现更好一些, P_a 在大部分情形下都很接近于 1.

对于 (b2) 的 β 的设定, 并没有类似 (b1) 中的边际独立而联合相依于响应变量的预测变量. 表 3 清晰地显示出 CR-SIS 和 CR-SJS 是如何被样本量、预测变量维数和 ρ 的值所影响的. 总体而言, 对于 (S1) 的情形, 无论是 (b1) 还是 (b2) 的 β , CR-SJS 的表现确实比 CR-SIS 要好, 哪怕是对于较小的样本量. CR-SJS 的表现随着样本量的增加而更好. 这与理论分析中 CR-SJS 具有确定筛选性质相一致. 表 3 中同样显示, 尽管 CR-SIS 的表现也随着样本量的增加得到了很大的改善, 然而 CR-SJS 的表现相对稳定许多.

表 4 展现了对于协方差结构为 (S2) 的模拟结果, 在这种结构下, $Z_1, \dots, Z_4$ 中并没有与响应变量边际独立而联合相依的预测变量, 理论上 CR-SIS 的表现应当都很不错. 然而, 从表 4 中看到, 在 (b2) 的条件下, CR-SIS 的确表现不错, 然而在 (b1) 的情形时, CR-SIS 的表现仍然欠佳, 尽管随着样本量的增加, CR-SIS 的表现有了很大改善. 与之对比的是, 无论是 (b1) 还是 (b2), CR-SJS 均有较好表现. 比较两种协方差结构和系数结构, 可以总结出, CR-SJS 筛选效果大部分情形都优于 CR-SIS, 且表现更加稳定, 因此在实际应用中, 显然 CR-SJS 应当是更优的选择.

表 1 展示了对于我们设定的存在离散变量情形的模拟结果, 删失率约 26%. 可以看到, 当存在离散变量时, 对于其中部分协变量存在相关性时, CR-SJS 具有较好的筛选性质, 而 CR-SIS 的效果并不好. 当 n 增加时, CR-SJS 的表现均得到了一些改善, 而 CR-SIS 的改善则不大.

除了比较 CR-SIS 和 CR-SJS 之外, 我们还进行了对于 q 的选择的探索. 我们以在变量选择中常用的 $q = \lfloor n/(\log n) \rfloor$ 为基准, 分别选择 $\tau = 0.2, 0.4, \dots, 1.0$, 并令 $q = q_\tau = q\tau$, 分别做出了 CR-SJS 和 CR-SIS 对于 P_a 、 P_1 、 P_2 、 P_3 和 P_4 在不同 q 的选择下的变化规律, 放在图 1(a) 中, 以及在不同的协方

表 3 P_s 和 P_a 的比例, 其中 $\Sigma = (1 - \rho)I + \rho\mathbf{1}\mathbf{1}^T$

n	p	ρ	β	CR-SJS					CR-SIS				
				P_s					P_a				
				Z1	Z2	Z3	Z4	ALL	Z1	Z2	Z3	Z4	ALL
100	500	0.25	(b1)	0.99	1.00	1.00	1.00	0.99	0.96	0.98	0.94	0.00	0.00
100	500	0.25	(b2)	0.93	0.92	0.95	0.93	0.81	0.68	0.75	0.72	0.60	0.18
100	500	0.50	(b1)	0.97	1.00	1.00	0.99	0.97	0.53	0.58	0.57	0.00	0.00
100	500	0.50	(b2)	0.92	0.85	0.91	0.90	0.76	0.55	0.56	0.53	0.57	0.06
100	500	0.75	(b1)	0.88	0.88	0.88	0.99	0.72	0.12	0.14	0.25	0.03	0.00
100	500	0.75	(b2)	0.69	0.81	0.69	0.76	0.43	0.44	0.48	0.39	0.48	0.04
200	800	0.25	(b1)	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.00	0.00
200	800	0.25	(b2)	0.99	1.00	0.99	0.99	0.98	0.86	0.94	0.9	0.89	0.63
200	800	0.50	(b1)	1.00	1.00	1.00	1.00	1.00	0.92	0.85	0.88	0.00	0.00
200	800	0.50	(b2)	0.98	0.98	0.99	0.99	0.95	0.80	0.86	0.79	0.82	0.38
200	800	0.75	(b1)	1.00	0.99	1.00	1.00	0.99	0.35	0.34	0.31	0.01	0.01
200	800	0.75	(b2)	0.93	0.88	0.88	0.91	0.71	0.78	0.69	0.77	0.77	0.26

表 4 P_s 和 P_a 的比例, 其中 $\Sigma = (\rho^{|i-j|})$

n	p	ρ	β	CR-SJS					CR-SIS				
				P_s				P_a	P_s				P_a
				Z1	Z2	Z3	Z4	ALL	Z1	Z2	Z3	Z4	ALL
100	500	0.25	(b1)	1.00	0.99	1.00	0.98	0.98	0.99	1.00	0.97	0.18	0.17
100	500	0.25	(b2)	0.88	1.00	0.95	0.94	0.78	0.88	0.99	0.99	0.92	0.79
100	500	0.50	(b1)	0.99	0.99	0.96	0.97	0.96	0.97	0.94	0.38	0.33	0.25
100	500	0.50	(b2)	0.94	0.99	0.96	0.95	0.86	1.00	1.00	1.00	0.99	0.99
100	500	0.75	(b1)	1.00	0.94	0.69	0.98	0.67	0.73	0.57	0.05	0.31	0.03
100	500	0.75	(b2)	0.96	0.96	0.98	0.94	0.84	1.00	0.99	1.00	1.00	0.99
200	800	0.25	(b1)	1.00	1.00	0.96	1.00	0.96	0.96	0.92	0.74	0.42	0.36
200	800	0.25	(b2)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
200	800	0.50	(b1)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.79	0.72	0.72
200	800	0.50	(b2)	1.00	1.00	0.99	0.98	0.97	1.00	1.00	0.99	1.00	0.99
200	800	0.75	(b1)	1.00	1.00	0.87	1.00	0.87	0.89	0.89	0.45	0.62	0.45
200	800	0.75	(b2)	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00

差和系数设定下, P_a 和 P_i ($i = 1, 2, 3, 4$) 对于不同 q 的变化趋势, 在图 1(b) 中表示出. 从图 1-4 中也可以看出大部分情形中, CR-SJS 的表现都比 CR-SIS 要好, 具体表现为模拟中选出重要变量的比例更高. 在 $q > s$ 的基础上, 随着第一阶段的 q 增加, 能够选出重要变量的比例先增加, 然后逐渐缓慢降低, 这可能是由于当 q 较小时, 有可能不能保证 $q > s$; 而当 q 较大时, 一方面收敛算法有可能与初始值的选择有关; 另一方面, 算法中子模型的最大似然估计有可能收敛到局部最大值点, 且在模拟中发现, 程序运行花费的时间逐渐增加, 显然这是由于需要更长时间达到算法收敛.

在实际应用中, 由于真实的显著变量我们无法预知, 为了尽量减小误差, 我们可以选择一系列 q_i , 分别做变量选择, 并且可以通过 BIC 等准则选出最优的 q . 其二是在算法中, 多取几个初始值. 其三, 我们可以借鉴 Fan 等^[32] 分割样本的想法, 将样本随机分成两个子样本, 在取较大的 q 之后, 将第一阶段选出的两个变量集合取交集作为最终的第一阶段变量集合, 再通过惩罚方法来选出最终的变量.

图 1 (a) 在 (S1) 和 (b1) 及 $\rho = 0.25$ 的设定下, CR-SJS 和 CR-SIS 的 P_a 和 P_i ($i = 1, 2, 3, 4$) 相对于不同的比例 τ 的变化; (b) 在不同的协方差和系数及 $\rho = 0.50$ 的设定下, CR-SJS 和 CR-SIS 的 P_a 相对于不同的比例 τ 的变化

图 2 (a) 在不同的协方差和系数及 $\rho = 0.75$ 的设定下, CR-SJS 和 CR-SIS 的 P_a 相对于不同的比例 τ 的变化; (b) 在 (S1) 和 (b1) 的设定下, 对于不同的 ρ , CR-SJS 和 CR-SIS 的 P_a 相对于不同的比例 τ 的变化

图 3 在 (S1) 和 (b1) 的设定下, 对于不同的 ρ , CR-SJS 和 CR-SIS 的 (a) P_1 、(b) P_2 相对于不同的比例 τ 的变化

图 4 在 (S1) 和 (b1) 的设定下, 对于不同的 ρ , CR-SJS 和 CR-SIS 的 (a) P_3 、(b) P_4 相对于不同的比例 τ 的变化

5.2 实证分析

我们使用一个公开的膀胱癌数据来对所提出的方法进行实证分析. 这个数据集来自于文献 [3], 使用来自于之前或目前没有侵入肌肉的 pT_a 和 pT_1 的肿瘤患者样本, 来验证预测恶化的信号. Binder 等^[1] 提出了高维竞争风险模型的提升 (boosting) 方法, 并利用这一数据集对恶化的预测模型进行拟合和评估. 这个数据集包括 404 个样本, 每个样本对应公开可得的预处理过的 1,381 个个人平台微数组特征 (custom platform microarray features), GEO (Gene Expression Omnibus) 系列编号为 No. GSE5479. 此外, Dyrskjot 等^[3] 还提供了样本个体的其他信息, 包括文献 [1] 使用的潜在的重要临床变量: 年龄 (从 27 到 95)、性别、肿瘤阶段 (pT_a vs. pT_1)、WHO (World Health Organization) 评分 (PUNLMP/low vs. high) 和接受的治疗. 除去一些缺失的信息, 我们将分析的数据集中在 $n = 338$ 个观测的子集中. 考虑将临床变量强制视作显著的变量和加入选择变量两种情形, 来分析对于 1,381 个基因组变量进行选择的结果. 我们感兴趣的响应变量是到恶化的时间. 对于 53 个患者, 可以观测到我们感兴趣的事件, 即恶化或者死于膀胱癌. 删失的患者有 215 个. 竞争风险为死于其他或未知原因, 这样的样本有 70 个.

将 5 个临床因素作为变量和 1,381 个基因变量共同选择, 我们使用对于竞争风险模型的确联合筛选和确定独立筛选两种方法进行第一阶段选择 $q = [n/\log(n)] = 58$ 个变量, 第二阶段变量选择均使用惩罚近似部分似然的方法, 惩罚函数使用 LASSO 惩罚, 并使用 BIC 准则选择调节参数. 最终, CR-SJS 和 CR-SIS 方法在第一阶段选出 20 个相同的变量和 38 个相异的变量. 在第二个阶段, 使用 LASSO 惩罚, CR-SJS + LASSO 和 CR-SIS + LASSO 两种方法选出的变量分别为 7 个 (SEQ523、SEQ361、SEQ1037、SEQ1082、SEQ1157、SEQ203 和 SEQ643) 和 5 个 (SEQ370、SEQ766、SEQ1038、SEQ509 和 SEQ1327). 两种方法选出的变量均远远小于文献 [3] 中提出的 88 个基因变量, 略小于文献 [1] 中选出的 8 个基因变量.

表 5 给出了当第一阶段选出的 $q = [n/\log(n)]$ 时, CR-SJS 和 CR-SIS 选出的相同变量和相异变量. 可以看到, 这两种方法选出的第一阶段的变量集合有一定比例的重合, 对于相异变量集合, CR-SIS 可能选择对于响应变量具有较强边际联系的变量, 而 CR-SJS 能更加准确识别出与响应变量联合相依的变量. 然而, 第一阶段选出的变量集合仍然可能含有大量与响应变量无关的变量, 而第二阶段的惩罚方法对变量集合进行了进一步的收缩.

表 5 CR-SIS 和 CR-SJS 选出的 q 个变量

相同变量			SIS				SJS		
SEQ994	SEQ34	SEQ370	SEQ247	SEQ709	SEQ407	SEQ1295	SEQ1171	SEQ1388.1	SEQ1216
Grade	SEQ1037	SEQ790	SEQ334	SEQ152	SEQ248	SEQ537	SEQ1384.1	SEQ262	SEQ1253
SEQ589	SEQ1157	SEQ1172	SEQ1142	SEQ1354	SEQ382	SEQ539	SEQ844	SEQ405	SEQ811
SEQ1262	SEQ1226	SEQ766	SEQ469	SEQ591	SEQ509	SEQ813	SEQ1390	SEQ1389	SEQ108
SEQ279	SEQ295	SEQ369	SEQ767	SEQ785	SEQ818	SEQ1218	SEQ938	SEQ954	SEQ505
SEQ1042	SEQ1227	SEQ246	SEQ782	SEQ828	SEQ807	SEQ842	SEQ414	SEQ802	SEQ1202
SEQ1082	SEQ361	SEQ972	SEQ713	SEQ1327	SEQ1174	SEQ39	SEQ956	SEQ389	SEQ203
SEQ377	SEQ1364	SEQ440	SEQ973	SEQ1339	SEQ1302	SEQ11	SEQ847	SEQ408	SEQ169
SEQ704	SEQ887	SEQ820	SEQ765	SEQ977		SEQ1249	SEQ824	SEQ287	
SEQ523	SEQ643	SEQ1038	SEQ711	SEQ1353		SEQ929	SEQ142	SEQ843	

6 结论

本文旨在处理生存分析中常见的存在竞争风险的超高维变量选择问题. 对于竞争风险的 PSH 模型, 提出了基于确定联合筛选和惩罚近似部分似然的两阶段变量选择方法, 并证明了所提出方法的确筛选性质. 通过模拟分析了方法的表现, 并以一个膀胱癌的高维数据为例展现了方法的实用性, 具有一定的理论意义和实践价值.

致谢 感谢主编、编委及各位审稿人给予的指导与建议.

参考文献

- 1 Binder H, Allignol A, Schumacher M, et al. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, 2009, 25: 890–896
- 2 Fine J P, Gray R J. A proportional hazards model for the subdistribution of a competing risk. *J Amer Statist Assoc*, 1999, 94: 496–509
- 3 Dyrskjot L, Zieger K, Real F X, et al. Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: A multicenter validation study. *Clin Cancer Res*, 2007, 13: 3545–3551
- 4 Faraggi D, Simon R. Bayesian variable selection method for censored survival data. *Biometrics*, 1998, 54: 1475–1485
- 5 Fan J, Feng Y, Wu Y. High-dimensional variable selection for Cox's proportional hazards model. *Inst Math Stat Collect*, 2010, 6: 70–86
- 6 Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B Stat Methodol*, 2008, 70: 849–911
- 7 Zhao S D, Li Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J Multivariate Anal*, 2012, 105: 397–411
- 8 Song R, Lu W, Ma S, et al. Censored rank independence screening for high-dimensional survival data. *Biometrika*, 2014, 101: 799–814
- 9 Yang G, Yu Y, Li R, et al. Feature screening in ultrahigh dimensional Cox's model. *Stat Sinica*, 2016, 26: 881–901
- 10 Liu J Y, Zhong W, Li R Z. A selective overview of feature screening for ultrahigh-dimensional data. *Sci China Math*, 2015, 58: 2033–2054
- 11 Shao L, Yu Y, Zhou Y. Sure feature screening for high-dimensional dichotomous classification. *Sci China Math*, 2016, 59: 2527–2542
- 12 Wang L H, Liu J Y, Li Y, et al. Model-free conditional independence feature screening for ultrahigh dimensional data. *Sci China Math*, 2017, 60: 551–568
- 13 Lin Y, Liu X, Hao M. Model-free feature screening for high-dimensional survival data. *Sci China Math*, 2018, 61, doi: 10.1007/s11425-016-9116-2
- 14 Kuk D, Varadhan R. Model selection in competing risks regression. *Stat Med*, 2013, 32: 3077–3088
- 15 Ha I D, Lee M, Oh S, et al. Variable selection in subdistribution hazard frailty models with competing risks data. *Stat Med*, 2014, 33: 4590–4604
- 16 Tapak L, Saidijam M, Sadeghifar M, et al. Competing risks data analysis with high-dimensional covariates: An application in bladder cancer. *Genomics Proteomics Bioinform*, 2015, 13: 169–176
- 17 Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Stat Methodol*, 1996, 58: 267–288
- 18 Fu Z, Parikh C R, Zhou B. Penalized variable selection in competing risks regression. *Lifetime Data Anal*, 2017, 23: 353–376
- 19 Zou H. The adaptive Lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429
- 20 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 21 Zhang C H. Nearly unbiased variable selection under minimax concave penalty. *Ann Statist*, 2010, 38: 894–942
- 22 Gray R J. A class of K -sample tests for comparing the cumulative incidence of a competing risk. *Ann Statist*, 1988, 16: 1141–1154
- 23 Robins J M, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: *AIDS Epidemiology Methodological Issues*. Boston: Birkhäuser, 1992, 297–331
- 24 Xu C, Chen J. The sparse MLE for ultra-high-dimensional feature screening. *J Amer Statist Assoc*, 2014, 109: 1257–1269
- 25 Hastie T J, Tibshirani R J. *Generalized Additive Models*. Boca Raton: CRC Press, 1990

- 26 Tibshirani R. The Lasso method for variable selection in the Cox model. *Stat Med*, 1997, 16: 385–395
- 27 Andersen P K, Gill R D. Cox's regression model for counting processes: A large sample study. *Ann Statist*, 1982, 10: 1100–1120
- 28 Huang J, Sun T, Ying Z, et al. Oracle inequalities for the Lasso in the Cox model. *Ann Statist*, 2013, 41: 1142–1165
- 29 Lehmann E L. *Elements of Large-Sample Theory*. New York: Springer-Verlag, 1999
- 30 Azuma K. Weighted sums of certain dependent random variables. *Tohoku Math J (2)*, 1967, 19: 357–367
- 31 Boos D D, Stefanski L A. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013
- 32 Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: Beyond the linear model. *J Mach Learn Res*, 2009, 10: 2013–2038

Feature screening based on ultrahigh dimensional competing risks models

Erqian Li, Bo Mei & Maozai Tian

Abstract In survival analysis, variable selection methods have been proposed in the literature for ultrahigh dimensional Cox model which deals with standard time-to-event data. However, in the cases where subjects may suffer from multiple mutually exclusive events, or competing risks, these methods cannot be applied directly. A popular model for data with competing risks turns out to be proportional subdistribution hazard (PSH) model. In this paper, we develop a two-stage variable selection strategy based on sure joint screening (SJS) and penalized approximated log partial likelihood for ultrahigh PSH model, and establish the sure screening property for the proposed method; in other words, with probability tending to 1, the set of selected variable includes all actual significant covariates. The performance of our proposed method is demonstrated by Monte Carlo simulation and compared with sure independence screening for competing data. We also conduct an analysis of a public dataset about bladder cancer to illustrate our method.

Keywords competing risks, feature screening, ultrahigh dimension, partial likelihood approximation, survival analysis

MSC(2010) 00A73, 62N99

doi: 10.1360/N012017-00069