文章编号:2095-6134(2023)01-0012-09

高维生存分析数据在带有测量误差情形下的变量选择方法*

张家睿†[1,2],吴耀华[1]

(1 中国科学技术大学管理学院, 合肥 230026; 2 香港大学浙江科学技术研究院, 杭州 310000)

(2020 年 12 月 30 日收稿; 2021 年 3 月 8 日收修改稿)

Zhang J R, Wu Y H. Variable selection method for high-dimensional survival error-in-variable data[J]. Journal of University of Chinese Academy of Sciences, 2023, 40(1): 12-20. DOI: 10.7523/j.ucas.2021.0016.

摘 要 对带有删失的生存数据的分析是高维稀疏回归分析的一个重要组成部分。然而,过去的大量相关工作都是建立在干净原始数据这一基础之上的,实践中面对的往往都是缺失数据或带有测量误差的数据,因此对此类数据的研究实用性更强。而在已有的高维生存分析数据相关文献中,关于带有测量误差情形下变量选择的研究还略显空白。在此背景下,提出一种基于伪得分函数和最近邻半正定投影的方法,对带有测量误差的高维可加风险模型进行变量选择,并且通过随机模拟和实际数据分析验证了该方法可以取得很好的效果。

关键词 变量选择;高维;可加风险模型;测量误差

中图分类号:O212.1 **文献标志码**:A **DOI**:10.7523/j.ucas.2021.0016

Variable selection method for high-dimensional survival error-in-variable data

ZHANG Jiarui[1,2], WU Yaohua[1]

(1 *School of Management, University of Science and Technology of China, Hefei* 230026, *China*;
2 *Zhejiang Institute of Research and Innovation, University of Hong Kong, Hangzhou* 310000, *China*)

Abstract Analysis with censored survival data plays an important role in high-dimensional sparse modeling. Much theoretical and applied work is based on clean data. However, we often face corrupted data with missing data or error-in-variable data and as a result analysis on error-in-variable data is more useful. While in the known literature, relatively few work has been done on high-dimensional survival data variable selecting with measurement error. In this situation, we propose a new method to select variables in high-dimensional additive hazards model with error-in-variable data, which combines the pseudoscore function and the nearest positive semi-definite projection. Our numerical studies and real data analysis show that the method has good performance and can select the nonzero coefficients successfully.

Keywords variable selection; high-dimensional; additive hazard model; error-in-variable data

* 国家自然科学基金(72071187,11671374,71731010,71921001)资助

† 通信作者,E-mail:zjrt46@mail.ustc.edu.cn

在过去 10 年里分子生物学试验技术的进展给我们带来了丰富的生物医学数据,举例来说,DNA 显微序列可以用来测量一个细胞中成千上万的基因。这种类型的数据中样本维度 p 比样本量 n 要大得多,对于传统的统计推断方法来说是一个巨大的挑战,有很多经典的推断方法在这种情况下变得不适用。这种情形下有效的变量选择方法就变得尤为重要。比较著名的高维数据变量选择方法有 Lasso[1],SCAD[2] 和 MCP[3] 等。

当研究关于患者生存状态的医疗数据时,将高维的生物医疗数据和患者的生存状态数据结合起来分析是一个很有效的方法。因此近些年来也有很多关于高维生存分析模型的变量选择方法,比如 Bradic 等[4] 关于高维 Cox 模型的正则化方法,Gorst-Rasmussen 和 Scheike[5] 关于高维单指数模型的筛选方法,Lin 和 Lyu[6] 关于高维可加模型的正则化方法等等。高维生存分析模型还广泛地应用到信用风险分析,比如 Fan 等[7]。

由于在实际生活中,我们经常会遇到带有测量误差的数据,所以对于带有测量误差数据的分析方法也是一个重要的研究方向,对于高维线性模型有 Loh 和 Wainwright[8] 以及 Datta 和 Zou[9] 的相关工作;对于变系数模型,有刘智凡等[10] 的工作。对于带有测量误差的生存分析数据的变量选择方法,代表文章有 Song 和 Wang[11] 关于工具变量的工作,Chen 和 Yi[12] 关于 Cox 模型左截断右删失数据的工作。高维生存分析模型由于其计算复杂度较高以及理论性质较为复杂,所以对于带有测量误差的高维生存分析数据的工作随着近些年大数据的迅速发展才逐步出现在视野之中。具有代表性的文章有 Chen 和 Yi[13] 关于高维生存分析图模型的工作以及 Chen 等[14] 关于高维 Cox 模型利用纠正似然函数的工作。本文选择同样具有重要应用的可加风险模型作为基础,结合处理高维线性模型的正则化方法对带有测量误差的生存分析数据进行分析。

1　研究背景

本文所采用的模型为高维可加风险模型,结合高维线性模型测量误差处理办法对带有测量误差的生存分析数据进行分析。下面对高维可加风险模型和高维线性模型测量误差处理方法分别进行介绍。

1.1　高维可加风险模型

对于生存分析数据的变量选择技术的发展已经不拘泥于 Cox 模型,可加风险模型便是除 Cox 模型以外的一种重要替代方式。可加风险模型假设失效时间为 T 的风险函数和 p 维的协变量 $\boldsymbol{X}(\cdot)$ 有如下形式的关系

$$\lambda(t \mid \boldsymbol{X}) = \lambda_0(t) + \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}(t), \tag{1}$$

其中:$\lambda_0(\cdot)$ 是一个不确定的基线风险函数,$\boldsymbol{\beta}_0$ 是一个 p 维的回归系数。令 C 为删失时间,则定义删失失效时间为 $\mathrm{CFT} = C \wedge T$,令 $\mathrm{CFT} = t_1, \cdots, t_n$,失效指数定义为 $\delta = I(T \leqslant C)$,其中 $I(\cdot)$ 为指示函数,令 $\boldsymbol{X}(t) = (X_1(t), \cdots, X_p(t))$ 并且假设给定 $\boldsymbol{X}$ 观察到的数据为 $(\mathrm{CFT}, \delta, \boldsymbol{X}(\cdot))$,风险函数由式(1)给出。

采用常用的计数手段,定义观察到的失效计数序列为 $N_i(t) = I(t_i \leqslant t, \delta_i = 1)$,风险中指数为 $Y_i(t) = I(t_i \geqslant t)$,计数过程鞅为

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda_0(s) + \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_i(s)\mathrm{d}s, \tag{2}$$

后文也将用 $N(t)$, $Y(t)$ 和 $M(t)$ 来代表这些计数过程的广义形式。

Lin 和 Ying[15] 采用一种有如下形式的伪得分方程来对可加风险模型进行分析:

$$\boldsymbol{U}_0(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{X}_i(t) - \bar{\boldsymbol{X}}(t) \{\mathrm{d}N_i(t) - Y_i(t)\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i(t)\mathrm{d}t\}, \tag{3}$$

其中 $\boldsymbol{\beta} \in \mathbb{R}^p$,并且

$$\bar{\boldsymbol{X}}(t) = \sum_{j=1}^{n} Y_j(t)\boldsymbol{X}_j(t) / \sum_{j=1}^{n} Y_j(t), \tag{4}$$

τ 是最大的跟踪时间(生存时间和删失时间的最大值)。这个估计函数关于回归系数是线性的,令

$$\boldsymbol{b}_0 = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{X}_i(t) - \bar{\boldsymbol{X}}(t)\mathrm{d}N_i(t), \tag{5}$$

和

$$\boldsymbol{V}_0 = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)\{\boldsymbol{X}_i(t) - \bar{\boldsymbol{X}}(t)\}^{\otimes 2}\mathrm{d}t, \tag{6}$$

其中 $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$,通过一些代数变换,可以写出如下等式

$$\boldsymbol{U}_0(\boldsymbol{\beta}) = \boldsymbol{b}_0 - \boldsymbol{V}_0\boldsymbol{\beta}. \tag{7}$$

在没有测量误差的情况下,$\boldsymbol{V}_0$ 是半正定的,式(7)两边关于 $\boldsymbol{\beta}$ 积分就可以得到损失函数

$$L(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{V}_0\boldsymbol{\beta} - \boldsymbol{b}_0{}^{\mathrm{T}}\boldsymbol{\beta}, \tag{8}$$

Leng 和 Ma[16] 以及 Martinussen 和 Scheike[17] 都建议用上述损失函数配合正则化方法对可加风险模型(1)进行变量选择。本文的相关工作也是在此基础上进行。

1.2 高维线性模型测量误差数据的处理方法

为了进一步构建更深层次的讨论,假设观察到的是被污染的协变量矩阵

$$\boldsymbol{Z}(\cdot) = (z_{ij}(\cdot))_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}, \tag{9}$$

而不是真实的协变量矩阵 $\boldsymbol{X}(\cdot)$。有很多种造成测量误差的途径,在加法测量误差设定中, $z_{i,j}(\cdot) = x_{i,j}(\cdot) + a_{i,j}$, 其中 $\boldsymbol{A}(\cdot) = (a_{i,j})$ 是加法测量误差。在乘法测量误差设定中, $z_{i,j}(\cdot) = x_{i,j}(\cdot)m_{i,j}$, 其中 $m_{i,j}$ 就是乘法测量误差。缺失数据可以看作乘法测量误差的一个特殊形式, $m_{i,j} = I(x_{i,j}(\cdot)$ 没缺失)。

不失一般性,用 Lasso 算法来举例说明测量误差的影响,对于线性模型 $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ 来说,Lasso 算法是最小化

$$1/(2n)\ \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{10}$$

这等价于最小化

$$\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma\beta} - \boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1, \tag{11}$$

其中: $\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{\rho} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, λ 为 Lasso 算法的惩罚参数。假设测量误差是加法误差且服从均值是 0、方差是 τ^2 的正态分布, τ 是已知的常数。Loh 和 Wainwright[8] 建议使用无偏估计

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} - \tau^2\boldsymbol{I}, \tilde{\boldsymbol{\rho}} = \frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}, \tag{12}$$

然后解决下面的优化问题来得到 $\boldsymbol{\beta}$ 的估计:

$$\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \tilde{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1, \tag{13}$$

然而 $\hat{\boldsymbol{\Sigma}}$ 在有测量误差的情况下经常会出现负的特征值(高维情形下更加常见),给优化上述问题带来了困难。为解决这个问题, Loh 和 Wainwright[8] 采用如下方法(简记为 NCL):

$$\hat{\boldsymbol{\beta}} \in \underset{\|\boldsymbol{\beta}\|_1\leqslant R}{\operatorname{argmin}} \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \tilde{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1. \tag{14}$$

其中 R 是一个跟稀疏度有关的常数。Datta 和 Zou[9] 提出一种最近邻正定投影矩阵的算法来解决上述问题,对于任意方阵 $\boldsymbol{K}$:

$$(\boldsymbol{K})_+ = \underset{\boldsymbol{K}_1\geqslant 0}{\operatorname{argmin}} \|\boldsymbol{K} - \boldsymbol{K}_1\|_{\max}. \tag{15}$$

其中 $\|\boldsymbol{K}\|_{\max} = \max\limits_{i,j}|K_{i,j}|$, 本文在附录 A 中给出了最近邻正定投影矩阵的详细计算方法。定义 $\tilde{\boldsymbol{\Sigma}} = (\hat{\boldsymbol{\Sigma}})_+$, 则它们的 CoCoLasso 估计量为

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \tilde{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1. \tag{16}$$

由定义可知, $\tilde{\boldsymbol{\Sigma}}$ 一直是半正定矩阵,为解决上述优化问题,可以重写优化函数为

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{Z}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{17}$$

其中: $\tilde{\boldsymbol{Z}}$ 是 $\tilde{\boldsymbol{\Sigma}}$ 的 Cholesky 分解, $\tilde{\boldsymbol{y}}$ 是 $\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{y}} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}$ 的解。式(17)就是 Lasso 算法的优化函数,我们已经有很多现成的优化算法来求解。

2 带有测量误差的高维可加风险模型的变量选择方法

2.1 简化伪得分方程

在第 1 节中已经介绍了 Lin 和 Ying[15] 的伪得分方程的具体形式,下面将在协变量 $\boldsymbol{X}$ 期望值为 0 的前提下简化该伪得分方程,提出一种全新的更加容易计算且符合实际情况的损失函数。首先定义

$$\boldsymbol{b} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{X}_i(t)\,\mathrm{d}N_i(t), \tag{18}$$

以及

$$\boldsymbol{V} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}Y_i(t)(\boldsymbol{X}_i(t))^{\otimes 2}\mathrm{d}t, \tag{19}$$

则有

$$\begin{aligned}\boldsymbol{b} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{X}_i(t)\,\mathrm{d}N_i(t)\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}Y_i(t)(\boldsymbol{X}_i(t))^{\otimes 2}\mathrm{d}t + \\ &\quad \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{X}_i(t)\,\mathrm{d}M_i(t).\end{aligned} \tag{20}$$

接着定义

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{X}_i(t)\,\mathrm{d}M_i(t), \tag{21}$$

由于 $\boldsymbol{X}$ 的期望为 0,所以容易得到 $E(\boldsymbol{U}(\boldsymbol{\beta})) = 0$, 在如上定义的基础上,类似于式(7),有

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{b} - \boldsymbol{V\beta}, \tag{22}$$

式(22)对 $\boldsymbol{\beta}$ 积分即可得到期望为 0 时的损失函数

$$L(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{V\beta} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\beta}, \tag{23}$$

综上所述即为简化版本的损失函数,我们将基于这个损失函数进行变量选择。

2.2　两种测量误差数据的变量选择方法

2.2.1　加法测量误差

假设观测到的设计矩阵 $\boldsymbol{Z}(\cdot)$ 被加法测量误差污染,即 $z_{i,j}(\cdot)=x_{i,j}(\cdot)+a_{i,j}$, 其中 $\boldsymbol{A}(\cdot)=(a_{i,j})$。同时假设 $\boldsymbol{A}$ 的行是独立同分布的,均值是 0,协方差矩阵是 $\boldsymbol{\Sigma}_A$, 次高斯参数是 τ^2。假设 $\boldsymbol{\Sigma}_A$ 是已知的,则 $\boldsymbol{V}$ 和 $\boldsymbol{b}$ 的无偏估计分别为

$$\hat{\boldsymbol{V}}_{\text{add}}=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}Y_i(t)\{(\boldsymbol{Z}_i(t))^{\otimes 2}-\boldsymbol{\Sigma}_A\}\,\mathrm{d}t, \tag{24}$$

和

$$\hat{\boldsymbol{b}}_{\text{add}}=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{Z}_i(t)\,\mathrm{d}N_i(t). \tag{25}$$

可以看出 $\hat{\boldsymbol{V}}_{\text{add}}$ 有负的特征值会导致最小化损失函数成为非凸优化,基于 Datta 和 Zou[9] 的方法,可以得到相应的凸损失函数

$$\tilde{f}_{\text{add}}(\boldsymbol{\beta})=(1/2)\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{V}}_{\text{add}}\boldsymbol{\beta}-\hat{\boldsymbol{b}}_{\text{add}}^{\mathrm{T}}\boldsymbol{\beta}+\lambda\|\boldsymbol{\beta}\|_1, \tag{26}$$

其中 $\tilde{\boldsymbol{V}}_{\text{add}}=(\hat{\boldsymbol{V}}_{\text{add}})_+$, 在此基础上,运用第 1 节提到的方法应用 Lasso 求解最小化上述损失函数即可。即 $\hat{\boldsymbol{\beta}}_{\text{add}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\tilde{f}_{\text{add}}(\boldsymbol{\beta})$。

2.2.2　乘法测量误差

假设测量误差是乘法测量误差,观测到的数据为 $z_{i,j}(\cdot)=x_{i,j}(\cdot)m_{i,j}$, 其中 $m_{i,j}$ 就是乘法测量误差。在矩阵写法中,有 $\boldsymbol{Z}(\cdot)=\boldsymbol{X}(\cdot)\odot\boldsymbol{M}$, 其中 $\odot$ 代表矩阵或向量的元素相乘,假设 $\boldsymbol{M}$ 的行是独立同分布的,均值是 $\boldsymbol{\mu}_M$, 协方差矩阵为 $\boldsymbol{\Sigma}_M$, 次高斯参数为 τ^2。假设 $\boldsymbol{\Sigma}_M+\boldsymbol{\mu}_M\boldsymbol{\mu}_M^{\mathrm{T}}$ 是严格正定的,并且采用无偏估计

$$\hat{\boldsymbol{V}}_{\text{mult}}=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}Y_i(t)\{(\boldsymbol{Z}_i(t))^{\otimes 2}\,/\!/\,(\boldsymbol{\Sigma}_M+\boldsymbol{\mu}_M\boldsymbol{\mu}_M^{\mathrm{T}})\}\,\mathrm{d}t. \tag{27}$$

以及

$$\hat{\boldsymbol{b}}_{\text{mult}}=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{Z}_i(t)\,/\!/\,\boldsymbol{\mu}_M\mathrm{d}N_i(t). \tag{28}$$

其中//代表向量或者矩阵对应元素相除。和加法测量误差模型类似,乘法测量误差下无偏估计矩阵也有可能不是正定的,所以基于 Datta 和 Zou[9] 的方法,可以得到相应的凸损失函数:

$$\tilde{f}_{\text{mult}}(\boldsymbol{\beta})=(1/2)\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{V}}_{\text{mult}}\boldsymbol{\beta}-\hat{\boldsymbol{b}}_{\text{mult}}^{\mathrm{T}}\boldsymbol{\beta}+\lambda\|\boldsymbol{\beta}\|_1. \tag{29}$$

其中 $\tilde{\boldsymbol{V}}_{\text{mult}}=(\hat{\boldsymbol{V}}_{\text{mult}})_+$。然后运用第 1 节提到的方法应用 Lasso 求解最小化上述损失函数即可。即 $\hat{\boldsymbol{\beta}}_{\text{mult}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\tilde{f}_{\text{mult}}(\boldsymbol{\beta})$。对于缺失数据,可以将其视为乘法测量误差,具体办法在第 1 节已经详细介绍。

3　理论性质

在这一节中给出并推导估计量的 l_1 和 l_2 误差界。记我们的估计量为 CoCo 估计量。首先定义近邻条件:

条件 3.1　近邻条件:假设 $\hat{\boldsymbol{V}}$ 和 $\hat{\boldsymbol{b}}$ 由一系列参数 θ 确定,则存在依赖于 $\boldsymbol{\beta}_S$,θ 和 σ^2 的全局常数 C 和 c,正函数 ς 和 ε_0 使得对每个 $\varepsilon\leqslant\varepsilon_0$,$\hat{\boldsymbol{V}}$ 和 $\hat{\boldsymbol{b}}$ 满足如下概率条件:

$$P(|\hat{V}_{ij}-V_{ij}|\geqslant\varepsilon)\leqslant C\exp(-cn\varepsilon^2\varsigma^{-1}), \tag{30}$$

$$P(|\hat{b}_j-b_j|\geqslant\varepsilon)\leqslant C\exp(-cn^3\varepsilon^2\varsigma^{-1}). \tag{31}$$

对所有 $1\leqslant i,j\leqslant p$ 成立。其中集合 $S=\{1,2,\cdots,s\}$ 是回归系数 $\boldsymbol{\beta}$ 的支撑集。

同样也需要和线性模型下一样的特征值限制条件:

条件 3.2　协方差阵特征值限制条件

$$0<\Omega=\min_{\boldsymbol{x}\neq 0,\,\|\boldsymbol{x}_{S^C}\|_1\leqslant 3\|\boldsymbol{x}_S\|_1}\frac{\boldsymbol{x}'\boldsymbol{V}\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2}. \tag{32}$$

条件 3.2 是一个在高维线性模型变量选择中比较常见的假设。下面给出 CoCo 估计量的统计误差界:

定理 3.1　在式(30)、式(31)和式(32)成立的前提下,对于 $\lambda\leqslant\min(\varepsilon_0,12\varepsilon_0\|\boldsymbol{\beta}_S\|_\infty)$ 和 $\varepsilon\leqslant\min(\varepsilon_0,\Omega/64s)$, 下式至少以概率 $1-p^2C\exp(-cn^3\lambda^2\varsigma^{-1})-p^2C\exp(-cn\varepsilon^2\varsigma^{-1})$ 成立:

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\leqslant C\lambda\sqrt{s}/\Omega,\ \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1\leqslant C\lambda s/\Omega. \tag{33}$$

其中

$$\hat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{V}}\boldsymbol{\beta}-\hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{\beta}+\lambda\|\boldsymbol{\beta}\|_1. \tag{34}$$

对于两种测量误差 $\tilde{\boldsymbol{V}}$ 和 $\hat{\boldsymbol{b}}$ 的计算方式在 2.2 节中已详细介绍。定理 3.1 为 Datta 和 Zou[9] 文章中定理 1 的平行定理,同样也给出了估计量 l_1

和 l_2 的误差界。本文在固定设计的前提下进行理论部分的展开,即 $\frac{1}{n}\| x_j \|_2^2 = 1$ 对 $1 \leqslant j \leqslant p$ 成立。且假设观察时间的最大值 $t_n < \infty$,加法测量误差 $| a_{ij} | \leqslant L$。下面将对加法测量误差和乘法测量误差对条件 3.1 的满足给出理论上的保证。

引理 3.1 $\hat{\boldsymbol{V}}_{\text{add}}$ 和 $\hat{\boldsymbol{b}}_{\text{add}}$ 满足条件 3.1 中的近邻条件式(30)和式(31),$\varsigma = \max(\tau^2 t_n^2, \tau^4 t_n^2, L^4)$ 且 $\varepsilon_0 = c\tau^2 t_n$。其中 τ 为次高斯参数。

引理 3.1 说明加法测量误差的计算方法满足近邻条件。下面将对乘法测量误差进行说明。为了保证乘法测量误差的计算方法也满足近邻条件,需要添加额外的正则化条件如下:

$$\max_{i,j} | X_{ij} | = X_{\max} < \infty,\ \min \mu_M = \mu_{\min} > 0,$$
$$\min_{i,j} E(m_1 m_1') = M_{\min} > 0,\ \max \mu_M = \mu_{\max} < \infty. \tag{35}$$

则接下来有

引理 3.2 存在依赖于 $\boldsymbol{\beta}_S, \theta, \sigma^2$ 和正则化条件式(35)的 ς 和 ε_0 使得 $\hat{\boldsymbol{V}}_{\text{mult}}$ 和 $\hat{\boldsymbol{b}}_{\text{mult}}$ 满足条件 3.1 中的近邻条件式(30)和式(31)。

引理 3.2 说明了乘法测量误差的计算方法满足近邻条件。将引理 3.1,引理 3.2 和定理 3.1 结合有

推论 3.1 在条件 3.2 成立的前提下,定理 3.1 的结果对加法测量误差估计量 $\hat{\boldsymbol{\beta}}_{\text{add}}$ 和乘法测量误差估计量 $\hat{\boldsymbol{\beta}}_{\text{mult}}$ 成立。

推论 3.1 给出了加法测量误差估计方法和乘法测量误差估计方法的理论保证,确定了估计量 l_1 和 l_2 的误差界,下面将通过随机模拟实验和实际数据分析来验证我们的理论结果。

4 实验及结果分析

本文的方法简记为 CoCo,Loh 和 Wainwright[8] 的方法记为 NCL,在随机模拟实验和实际数据分析中将对两种方法进行比较。

4.1 随机模拟

4.1.1 加法测量误差模型

从可加风险模型中产生数据,设定 $\lambda_0 = 5$,回归系数为

$$\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, \cdots, 0). \tag{36}$$

样本量 $n = 100$,样本维度 $p = 200$,$\boldsymbol{X}$ 的行独立同分布,均值为 0,协方差矩阵为 $\boldsymbol{\Sigma}_X$,考虑两种情形下的 $\boldsymbol{\Sigma}_X$:自回归($\boldsymbol{\Sigma}_{X,ij} = 0.5^{|i-j|}$)和复合对称($\boldsymbol{\Sigma}_{X,ij} = 0.5 + I(i = j) * 0.5$),删失时间服从 U(0,2)的均匀分布使得删失率维持在 20% 左右。首先生成 $3n \times p$ 的 $\boldsymbol{X}$,然后从中选出 n 个满足 $\lambda_0 + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X} > 0$ 的样本作为实验数据。加法测量误差为矩阵 $\boldsymbol{A}$,观测数据由 $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{A}$ 生成,$\boldsymbol{A}$ 的行是服从 $N(0, \tau^2 I)$ 的独立同分布变量,其中 $\tau = 0.25$、0.5 和 0.75。

实验中,参照 Datta 和 Zou[9] 的方法,运用 5 折的交叉验证方法得到惩罚参数 λ 及 CoCo 估计量的值。对于 NCL 的算法,参考 Duchi 等[18] 的工作,其中给出了详细的计算过程。首先根据式(14)得到一个 NCL 算法的初始估计量,即为基于 $\tilde{\boldsymbol{y}}$ 和 $\tilde{\boldsymbol{Z}}$ 的 Lasso 估计量。其次,依旧需要 $\| \boldsymbol{\beta}_S \|_1$ 的信息来调整参数 R。由于真实系数的稀疏度信息无法提前得知,所以用一个简单的 5 折交叉验证从 $[R_{\max}/500, 2R_{\max}]$ 的 100 均分点中选取合适的 R,其中 $R_{\max}$ 代表初始估计量的 l_1 范数。记录 C 和 IC 分别代表选对的系数数量和错误的数量,还记录均方误差(MSE)以及其标准差(se)。总共进行了 100 次实验取平均数作为最后的结果。在表 1 中进行展示。

表 1 展示了 CoCo 和 NCL 两种方法分别在自回归和复合对称条件下的 100 次重复实验的结果,可以看出在两种情形下本文方法的选对数量和估计的均方误差方面都比 NCL 方法要好。

4.1.2 乘法测量误差模型

与加法测量误差模拟类似,依旧从可加风险模型中产生数据,$\lambda_0 = 5$,回归系数,样本量和样本维度都保持不变,$\boldsymbol{X}$ 的行独立同分布,均值为 0,协方差矩阵为 $\boldsymbol{\Sigma}_X$,依旧考虑 $\boldsymbol{\Sigma}_X$ 在自回归和复合对称两种条件下的情形,并且与加法测量误差

表 1 加法测量误差两种方法的结果

Table 1 The results of two methods under additive error-in-variable data

τ		0.25		0.5		0.75	
		CoCo	NCL	CoCo	NCL	CoCo	NCL
AR	C	3	3	2.96	2.88	2.83	2.47
	IC	13.98	8.59	12.71	4.22	10.79	3.96
	MSE	1.61	1.70	3.06	4.49	5.43	6.08
	se(MSE)	0.95	0.95	1.84	1.61	3.56	3.16
CS	C	2.87	2.70	2.57	2.21	2.00	1.53
	IC	12.6	16.65	10.98	7.67	9.89	5.81
	MSE	3.82	4.23	6.55	7.43	10.12	10.92
	se(MSE)	2.30	2.54	3.24	3.60	4.48	4.07

中的设定保持一致。删失时间服从 U(0,2)的均匀分布使得删失率维持在 20% 左右,首先生成 $3n\times p$ 的 $\boldsymbol{X}$,然后从中选出 n 个满足 $\lambda_0+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}$ 的作为实验数据。乘法测量误差矩阵为 $\boldsymbol{M}=((m_{i,j}))$,观测数据由 $\boldsymbol{Z}(\cdot)=\boldsymbol{X}(\cdot)\odot\boldsymbol{M}$ 生成,$\log(m_{i,j})$ 是服从 $N(0,\tau^2\boldsymbol{I})$ 的独立同分布变量,其中 $\tau=0.25$、0.5 和 0.75。与上一个随机模拟实验一样,依旧采用 5 折的交叉验证方法来估计 CoCo 估计量和 NCL 的参数 R。同样记录 C 和 IC 分别代表选对的系数数量和错误的数量,还记录均方误差(MSE)以及其标准差(se)。总共进行 100 次实验取平均数作为最后的结果,在表 2 中展示。

表 2 展示了乘法测量误差中,CoCo 和 NCL 两种方法分别在自回归和复合对称条件下的 100 次重复实验结果,可以看出在两种情形下本文方法的选对数量和估计的均方误差都比 NCL 方法要好。但是随着测量误差变大,CoCo 和 NCL 方法的估计精确度都会有明显下降。

4.2　实际数据分析

在这一节中,会分析 van Houwelingen Hans C. 等[19]乳腺癌和基因的关系数据。这个数据包含了从荷兰乳腺研究中心的 295 个女性乳腺癌患者样本,诊断时间为 1984—1995 年。295 个患者的随访时长中位数为 6.7 a。所有的肿瘤都由一个包含 24 885 个基因的 cDNA 序列来描述。van't Veer L J 等[20]在这个数据只有 78 个肿瘤的初始工作中,运用 Rosetta 误差模型根据 p 值大小选取了 4 919 个基因。本文的数据沿用这一方法,并在此基础上根据 Gorst-Rasmussen 和 Scheike 的 ISIS 方法[5]选择了 500 个和生存时间相关程度最高的基因,因此样本维度是 $p=500$。类似于前面的随机模拟实验,人工地加入加法测量误差来产生被污染的协变量,控制加法测量误差的参数 $\tau=0.75/\sqrt{n}$,这是因为变量经过了标准化处理使得每一列的 L_2 范数是 1。

表 2　乘法测量误差两种方法的结果

Table 2　The results of two methods under multiplicative error-in-variable data

τ		0.25 CoCo	0.25 NCL	0.5 CoCo	0.5 NCL	0.75 CoCo	0.75 NCL
AR	C	2.99	2.99	2.95	2.33	2.82	1.96
	IC	13.25	6.32	13.88	7.77	15.46	7.81
	MSE	1.63	1.86	3.41	7.59	5.95	13.30
	se(MSE)	1.36	1.03	2.90	4.63	2.81	13.22
CS	C	2.87	2.84	2.65	2.31	2.17	1.64
	IC	12.08	4.86	11.75	6.11	13.69	7.47
	MSE	3.35	3.54	5.60	7.33	9.95	11.69
	se(MSE)	2.02	2.61	2.89	3.86	5.11	4.98

为了检验我们方法的有效性,将 295 个样本随机分成包含 235 个样本的训练集和 60 个样本的验证集并重复 100 次,在每一次实验中,都采用随机模拟实验中的两种方法,即 CoCo 和 NCL,用训练集训练模型参数并用验证集来筛选表现最好的估计量。计算

$$\mathrm{SEE}=\left|\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\beta}-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\beta}\right| \tag{37}$$

作为检验两种方法效果的指标。具体的结果展示在表 3 中。从表 3 中可以看出我们的方法依旧有比较高的预测精确度,这也和随机模拟实验的结果相符。我们方法的指标相比 NCL 方法要好一些,并且变量选择的数量上也比较相近。

表 3　加法测量误差情形下两种方法应用在乳腺癌数据中的结果

Table 3　The results of two methods in breast cancer data under additive measurement error

Measure	CoCo	NCL
SEE	10.32	18.32
se(SEE)	3.75	5.31
size	8.69	5.97

5　结论

本文提出一种针对高维可加风险模型中带有测量误差情况下的变量选择方法。在已知的生存分析数据相关文献中,尚未有针对测量误差数据的变量选择方法。本文基于高维线性模型测量误差数据的估计方法,重构了高维可加风险模型,并给出了加法和乘法两种测量误差模型的变量选择算法。简化伪得分方程的形式更加简洁且实用性强。随机模拟实验和实际数据分析的相关结果证实了本文方法的有效性和精确性。

在未来的工作中,我们将致力于将简化伪得分方程应用于高维可加风险模型的变量选择中。同时也会对 Cox 模型,加速失效模型等其他生存分析模型中的测量误差数据利用最近邻半正定投影的方法进行变量选择方面的探索。

参考文献

[1]　Tibshirani R. Regression shrinkage and selection via the lasso

[J]. Journal of the Royal Statistical Society: Series B (Methodolo gieal), 1996, 58(1): 267-288. DOI:10.1111/j.2517-6161.1996.tb02080.x.

[2] Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties [J]. Journal of the American Statistical Association, 2001, 96(456): 1348-1360. DOI:10.1198/016214501753382273.

[3] Zhang C H. Nearly unbiased variable selection under minimax concave penalty [J]. Annals of Statistics, 2010, 38(2): 894-942.

[4] Bradic J, Fan J Q, Jiang J C. Regularization for cox's proportional hazards model with np-dimensionality [J]. Annals of Statistics, 2011, 39(6): 3092-3120. DOI: 10.1214/11-AOS911.

[5] Gorst-Rasmussen A, Scheike T. Independent screening for single-index hazard rate models with ultrahigh dimensional features[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2013, 75(2): 217-245. DOI:10.1111/j.1467-9868.2012.01039.x.

[6] Lin W, Lyu J C. High-dimensional sparse additive hazards regression [J]. Journal of the American Statistical Association, 2013, 108(501): 247-264. DOI: 10.1080/01621459.2012.746068.

[7] Fan J Q, Lyu J C, Qi L. Sparse high-dimensional models in economics[J]. Annual Review of Economics, 2011, 3: 291-317. DOI:10.1146/annurer-economics-061109-080451.

[8] Loh P L, Wainwright M J. High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity [J]. The Annals of Statistics, 2012, 40(3): 1637-1664.

[9] Datta A, Zou H. CoCoLasso for high-dimensional error-in-variables regression[J]. The Annals of Statistics, 2017, 45(6): 2400-2426.

[10] 刘智凡, 王妙妙, 谢田法,等. 工具变量辅助的变系数测量误差模型的估计[J]. 中国科学院大学学报, 2018, 35(1):1-9. DOI:10.7523/j.issn.2095-6134.2018.01.001.

[11] Song X, Wang C Y. Proportional hazards model with covariate measurement error and instrumental variables[J]. Journal of the American Statistical Association, 2014, 109(508):1636-1646. DOI:10.1080/01621459.2014.896805.

[12] Chen L P, Yi G Y. Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error[J]. Annals of the Institute of Statistical Mathematics, 2021, 73(3): 481-517. DOI: 10.1007/s10463-020-00755-2.

[13] Chen L P, Yi G Y. Analysis of noisy survival data with graphical proportional hazards measurement error models[J]. Biometrics, 2021, 77(3): 956-969. DOI: 10.1111/biom.13331.

[14] Chen B J, Yuan A, Yi G Y. Variable selection for proportional hazards models with high-dimensional covariates subject to measurement error[J]. The Canadian Journal of Statistics, 2020,49(2):397-420DOI:10.1002/cjs.11568.

[15] Lin D Y, Ying Z L. Semiparametric analysis of the additive risk model [J]. Biometrika, 1994, 81(1): 61-71. DOI:10.1093/biomet/81.1.61.

[16] Leng C L, Ma S G. Path consistent model selection in additive risk model via Lasso[J]. Statistics in Medicine, 2007, 26(20): 3753-3770. DOI:10.1002/sim.2834.

[17] Martinussen T, Scheike T H. Covariate selection for the semiparametric additive risk model[J]. Scandinavian Journal of Statistics, 2009, 36(4): 602-619.

[18] Duchi J, Shalev-Shwartz S, Singer Y, et al. Efficient projections onto the l_1-ball for learning in high dimensions [C]//Proceedings of the 25th International Conference on Machine Learning-ICML, 08. July 5-9, 2008, Helsinki, Finland. New York: ACM Press, 2008: 272-279. DOI:10.1145/1390156.1390191.

[19] van Houwelingen H C, Bruinsma T, Hart A A M, et al. Cross-validated cox regression on microarray gene expression data[J]. Statistics in Medicine, 2006, 25(18): 3201-3216. DOI:10.1002/sim.2353.

[20] van't Veer L J, Dai H Y, van de Vijver M J, et al. Gene expression profiling predicts clinical outcome of breast cancer [J]. Nature, 2002, 415(6871): 530-536. DOI:10.1038/415530a.

附录 A 最近邻半正定投影矩阵的计算方法

采用 ADMM(alternating direction method of multipliers)算法来解决如下优化问题

$$\hat{\boldsymbol{A}} = \underset{\boldsymbol{A}\geqslant\varepsilon\boldsymbol{I}}{\operatorname{argmin}} \| \boldsymbol{A} - \hat{\boldsymbol{\Sigma}} \|_{\max}. \tag{A.1}$$

引进一个额外的变量 B 将上述优化问题重新写作

$$(\hat{\boldsymbol{A}},\hat{\boldsymbol{B}}) = \underset{\boldsymbol{A}\geqslant\varepsilon\boldsymbol{I},\boldsymbol{B}=\boldsymbol{A}-\hat{\boldsymbol{\Sigma}}}{\operatorname{argmin}} \| \boldsymbol{B} \|_{\max}. \tag{A.2}$$

上述优化问题可以采用增广拉格朗日方程求解

$$f(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\Lambda}) = 1/2 \,|\, \boldsymbol{B} \,|_{\max} - \langle \boldsymbol{\Lambda}, \boldsymbol{A} - \boldsymbol{B} - \hat{\boldsymbol{\Sigma}} \rangle + \frac{1}{2\mu} \| \boldsymbol{A} - \boldsymbol{B} - \hat{\boldsymbol{\Sigma}} \|_F^2. \tag{A.3}$$

其中:μ 是惩罚参数,$\boldsymbol{\Lambda}$ 是拉格朗日矩阵,$\langle\cdot\rangle$代表内积,$\|\cdot\|_F$ 代表 Frobenius 范数。用下面的算法来求解该优化问题。

算法 1:最近邻半正定投影矩阵的 ADMM 算法

1. 输入 μ 和初始值 $\boldsymbol{B}_0$ 以及 $\boldsymbol{\Lambda}_0$。
2. 在第 i 步更新:

2.1 (步骤 A) $\boldsymbol{A}_{i+1} = (\boldsymbol{B}_i + \hat{\boldsymbol{\Sigma}} + \mu\boldsymbol{\Lambda}_i)_\varepsilon$,

2.2 (步骤 B) $\boldsymbol{B}_{i+1} = \mathrm{mat}_l(\mathrm{vec}_l(\boldsymbol{A}_{i+1} - \hat{\boldsymbol{\Sigma}} - \mu\boldsymbol{\Lambda}_i)) - l_1(\mathrm{vec}_l(\boldsymbol{A}_{i+1} - \hat{\boldsymbol{\Sigma}} - \mu\boldsymbol{\Lambda}_i, \mu)))$,

2.3 (步骤 Λ) $\boldsymbol{\Lambda}_{i+1} = \boldsymbol{\Lambda}_i - \dfrac{\boldsymbol{A}_{i+1} - \boldsymbol{B}_{i+1} - \hat{\boldsymbol{\Sigma}}}{\mu}$,

3. 重复第 2 步直到收敛。

其中 $\mathrm{vec}_l(\boldsymbol{M})$ 代表将对称阵 $\boldsymbol{M}$ 的下半部分向量化。$\mathrm{mat}_l(\boldsymbol{M})$ 代表 $\mathrm{vec}_l(\boldsymbol{M})$ 的逆过程。如果 $\boldsymbol{M} = \sum_j \lambda_j p_j p_j'$ 代表了矩阵的谱分解,则 $M_\varepsilon = \sum_j \max(\lambda_j, \varepsilon) p_j p_j'$。

附录 B 相关定理的证明

本文在固定设计的前提下来进行理论部分的证明,即 $\frac{1}{n}\|x_j\|_2^2 = 1$ 对 $1 \leqslant j \leqslant p$ 成立。且假设观察时间的最大值 $t_n < \infty$,加法测量误差 $|a_{ij}| \leqslant L$。

定理 3.1 的证明　定理 3.1 为 Datta 和 Zou[9] 中定理 1 的平行定理,证明过程与 Datta 和 Zou[9] 中定理 1 的证明过程类似,故此处省略。我们将详细证明两种测量误差的估计量满足引理 3.1 和引理 3.2。

引理 3.1 的证明　令 $\boldsymbol{\Sigma}_A = ((\sigma_{a,ij}))$,则

$$\hat{V}_{\mathrm{add},jk} - V_{jk} = \sum_{i=1}^n \frac{t_i}{n} a_{ij} x_{ij} + \sum_{i=1}^n \frac{t_i}{n} a_{ik} x_{ik} + \sum_{i=1}^n \frac{t_i}{n}(a_{ik}a_{ij} - \sigma_{a,jk}), \tag{B.1}$$

由于 $\frac{1}{n}\|\sum_{i=1}^n t_i x_{ij}\|_2^2 \leqslant t_n$,则由引理 B.2 可知 $|\sum_{i=1}^n \frac{t_i}{n} a_{ij} x_{ij}|$,$|\sum_{i=1}^n \frac{t_i}{n} a_{ik} x_{ik}|$ 至多以概率 $C\exp(-cn\varepsilon^2/t_n^2\tau^2)$ 大于 $\varepsilon/3$。同理,由引理 B.1,$\sum_{i=1}^n \frac{t_i}{n}(a_{ik}a_{ij} - \sigma_{a,jk})$ 也可以被证明是一个小量,综上所述,$\hat{\boldsymbol{V}}_{\mathrm{add}}$ 满足近邻条件式(30),且

$$\varsigma = \max(t_n^2\tau^4, t_n^2\tau^2), \varepsilon_0 = c\tau^2 t_n. \tag{B.2}$$

对于 $\hat{b}_{\mathrm{add},j} - b_j = \sum_{i=1}^n \frac{\delta_i t_t}{n} a_{ij}$,其中 a_{ij} 为相互独立且均值为 0 的随机变量序列,则由 Hoeffding 不等式有 $|\hat{b}_{\mathrm{add},j} - b_j|$ 至多以概率 $\exp(-c\varepsilon^2 n^3/L^4)$ 大于 ε。综上所述,$\hat{\boldsymbol{b}}_{\mathrm{add}}$ 满足近邻条件式(31),且

$$\varsigma = \max(t_n^2\tau^4, {t_n}^2\tau^2, L^4), \varepsilon_0 = c\tau^2 t_n.$$

综上,引理 3.1 的结论已论证完毕。

引理 3.2 的证明　令 $\boldsymbol{\Sigma}_M = ((\sigma_{m,ij}))$,则有

$$\begin{aligned}\hat{V}_{\mathrm{mult},jk} - V_{jk} &= \frac{1}{n}\sum_{i=1}^n \frac{t_i x_{ij} x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}(m_{ij}m_{ik} - \mu_j\mu_k - \sigma_{m,jk}) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{t_i x_{ij} x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}((m_{ij} - \mu_j)(m_{ik} - \mu_k) - \sigma_{m,jk}) + \\ &\quad \frac{1}{n}\sum_{i=1}^n \frac{t_i x_{ij} x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}(\mu_j(m_{ik} - \mu_k) + \mu_k(m_{ij} - \mu_j)).\end{aligned} \tag{B.3}$$

利用式(9)的正则化条件,有

$$|\hat{V}_{\text{mult},jk} - V_{jk}| \leqslant \left| \frac{1}{nM_{\min}} \sum_{i=1}^{n} t_i x_{ij} x_{ik} ((m_{ij} - \mu_j)(m_{ik} - \mu_k) - \sigma_{m,jk}) \right| +$$

$$\frac{\mu_{\max}}{nM_{\min}} \sum_{i=1}^{n} t_i x_{ij} x_{ik} (m_{ik} - \mu_{ik}) + \frac{\mu_{\max}}{nM_{\min}} \sum_{i=1}^{n} t_i x_{ij} x_{ik} (m_{ik} - \mu_{ik}). \quad (B.4)$$

将上式右边 3 项分别记为 T_1, T_2 和 T_3。对于 T_1，首先有 $\| v \|_\infty \leqslant X_{\max}^2$，其中 $v_i = x_{ij} x_{ik}$，则和引理 3.1 证明过程类似，继续由引理 B.1 可以得到下面结论，对于 $\varsigma = \max(t_n{}^2 \tau^2 X_{\max}^2 \frac{\mu_{\max}^2}{M_{\min}^2}, t_n{}^2 \tau^4 \frac{X_{\max}^4}{M_{\min}^2})$ 和 $\varepsilon_0 = c\tau^2 t_n X^{2_{\min}}_{\max}$，有

$$P(T_1 \geqslant \varepsilon) \leqslant C\exp(-cn\varepsilon^2 \varsigma^{-1}). \quad (B.5)$$

对于 T_2 和 T_3 仿照引理 3.1 的证明过程即可得到类似结论。由于 $\hat{\boldsymbol{b}}_{\text{mult}}$ 相比于 $\hat{\boldsymbol{b}}$ 只是对应元素除去了 μ_M，并且在条件(9)中 μ_M 的各个元素上下界已经做了限制，所以对于 $|\hat{\boldsymbol{b}}_{\text{mult}} - \hat{\boldsymbol{b}}|$ 的界只需简单应用引理 B.2 即可直接得到结论。综上所述，引理 3.2 证明完毕。

引理 B.1 令 $\boldsymbol{z}_i = (x_i, y_i)'$ 代表独立同分布的，均值为 0，协方差矩阵为 $\boldsymbol{\Sigma} = ((\sigma_{ij}))$，次高斯参数为 τ^2 的向量。则存在绝对常数 C 和 c 使得，对于任意的 $\varepsilon \leqslant c\tau^2 \| a \|_\infty$，有

$$P(\frac{1}{n} | \sum_{i=1}^{n} a_i (x_i y_i - \sigma_{12}) | \geqslant \varepsilon) \leqslant C\exp(\frac{-nc\varepsilon^2}{\tau^4 \| a \|_\infty^2}). \quad (B.6)$$

定义 B.1（次高斯随机变量） Z 是一个随机变量，若存在有限的 $\kappa > 0$ 使得 $\kappa = \sup_{p \geqslant 1} p^{-1/2} (E|Z|^p)^{1/p}$，则 Z 被称为次高斯随机变量，κ 被称为 Z 的次高斯范数，记为 $\| Z \|_\varphi$。

引理 B.2 一个次高斯随机变量 Z 满足如下的尾概率界

$$P(|Z| > t) \leqslant 2\exp(-t^2/2\tau^2), \forall t > 0. \quad (B.7)$$

其中满足上式的最小的 τ 被称为次高斯参数，如果 $\boldsymbol{\omega} = (\omega_1, \cdots \omega_n)'$ 并且所有的 ω_i 是独立的 0 均值的随机次高斯变量，则有一个重要的性质

$$\| \boldsymbol{v}' \boldsymbol{\omega} \|_\varphi^2 \leqslant K \| \boldsymbol{v} \|_2^2 \max_i (\| \omega_i \|_\varphi^2). \quad (B.8)$$

引理 B.1 和引理 B.2 均引自 Datta 和 Zou[9] 工作中的证明部分，此处省略相关证明。