

基于半参数变换模型的右删失生存数据纵向联邦统计学习

张欣然, 李建波*

江苏师范大学数学与统计学院, 江苏 徐州 221116

DOI:10.61369/ASDS.2026070008

摘要 : 处于大数据时代的今天, 分布式存储是医疗领域常见的生存数据存储模式, 这给传统的生存分析模型带来了新的挑战。基于此, 本文基于半参数变换模型研究右删失生存数据的纵向联邦统计学习问题。首先, 基于全部样本构建全局似然函数, 通过引入各机构本地特征线性组合中间变量, 将全局似然函数转化为各机构可联合优化的似然函数; 然后, 提出了一套相应的 ADMM 优化算法, 允许各机构与中心服务器之间进行隐私数据信息传输, 同时保持了集中式极大似然估计的效率; 最后统计大量模拟例子和实例分析说明了所研究方法的有效性和合理性。

关键词 : 半参数变换模型; 似然函数; 纵向联邦学习; 右删失数据

Vertical Federated Statistical Learning for Right-Censored Survival Data Based on Semiparametric Transformation Models

Zhang Xinran, Li Jianbo*

School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, Jiangsu 221116

Abstract : In the current big data era, distributed storage has become a common storage mode for survival data in the medical field, posing new challenges to traditional survival analysis models. Against this background, this paper investigates the vertical federated statistical learning of right-censored survival data based on semiparametric transformation models. Firstly, a global likelihood function is constructed using all samples. By introducing intermediate variables of linear combinations of local features from each institution, the global likelihood function is transformed into a jointly optimizable form for all participating institutions. Secondly, a corresponding ADMM optimization algorithm is proposed, which enables privacy-preserving information transmission between local institutions and the central server while maintaining the efficiency of centralized maximum likelihood estimation. Finally, extensive simulation experiments and real data analysis verify the effectiveness and rationality of the proposed method.

Keywords : semiparametric transformation model; likelihood function; vertical federated learning; right-censored data

引言

在现代临床医学领域, 不同医疗机构出于数据隐私保护、管理策略等原因, 各自独立存储患者数据, 数据存储往往呈现出分布式模式。此类数据一般包括疾病诊断时间、治疗方案、生存状态及随访时间等, 不仅在物理存储空间上分散, 而且由于各机构数据收集标准、患者群体特征等差异, 在数据结构和分布上也存在一定的异质性, 为传统统计推断方法带来了新的挑战: 一方面, 传统统计推断方法往往基于数据集中式假设, 难以直接应用于分布式环境, 强行将数据集可能导致严重的数据隐私泄露风险; 另一方面, 分布式数据的异质性使得传统统计推断方法的假设条件往往不再成立, 从而降低了模型的准确性和可靠性。分布

式存储数据一般包含横向和纵向两种存储模式, 其中横向分布式存储是指多个机构同时存储具有相同特征的不同样本, 而纵向分布式存储是指多个机构同时存储同一研究群体的不同特征的数据。如何融合分布式数据进行联合建模已成为当前生存分析领域的研究重点问题, 相应的融合统计推断方法称为横向和纵向联邦统计学习方法。

右删失生存数据是临床医学的常见数据类型, Cox 比例风险模型^[1]是处理此类数据最为广泛的模型。记 $T \in \mathbf{R}^+$ 为寿命时间变量, 该模型具有如下形式:

基金项目: 江苏省研究生科研训练项目 (KYCX25_3159); 江苏省学位与研究生教育教学改革课题 (JGKT25_C075); 江苏师范大学研究生课程思政教学研究示范中心项目。

作者简介: 张欣然, 江苏师范大学应用统计2023级硕士研究生, 研究方向为生存分析; Email: 2933125272@qq.com。

通讯作者: 李建波, 江苏师范大学数学与统计学院教授, 统计学博士, 研究领域包括生存分析、应用统计、统计学习等; Email: lijianbo@jsnu.edu.cn。

$$\lambda(t|Z) = \lambda_0(t) \exp\{Z^T \beta\}, \quad (1)$$

其中, $\lambda(t|Z)$ 为寿命 T 的条件风险函数, $\lambda_0(t)$ 为未知基准风险函数, $Z \in R^d$ 为协变量向量, $\beta \in R^d$ 为回归系数. 在数据集中式场合下, 关于模型 (1), 众多统计学者提出了一系列有效的估计方法, 比如极大似然估计估计方法^[2]、经验似然估计估计方法^[3]等. 注意到模型 (1) 假设调整的风险比是常数, 这在实际问题中往往很难满足. 基于此, 学者们进一步提出了具有如下形式的半参数变换模型^[4]:

$$\ln H(T) = -Z^T \beta + \varepsilon, \quad (2)$$

其中, $H(\cdot)$ 为一未知的严格单调递增函数, 且满足 $\lim_{t \rightarrow 0^+} H(t) = 0$, ε 为具有已知分布的随机误差项. 变换模型 (2) 具有高度的数据建模灵活性和易解释性, 包含模型 (1) 作为特殊情况, 通过简单推导, 当 ε 服从标准极值分布, 那么模型 (2) 即为模型 (1). 在数据集中情形下, 关于模型 (2) 的参数估计包括估计方程方法^[4]、极大似然估计方法^[5]、分层贝叶斯方法^[6]. 然而, 这些方法在分布式数据存储情形下, 难以直接推广应用. 目前关于生存数据的联邦统计学习方法的研究主要围绕比例风险回归模型 (1) 展开. 比如, Lu 等提出了 WebDISCO 横向联邦统计学习框架^[7], 该框架通过中心服务器聚合关键中间统计量构建全局评分函数和 Hessian 矩阵, 实现全局参数估计; 为了进一步降低通信成本, Duan 等提出了一种基于代理似然的 ODACH 算法^[8], 实现了通信高效的无损分布式参数推断. Dai 等基于 Cox 模型 (1) 提出了纵向分布式生存数据的 ADMM 统计推断方法^[9]; 而关于更一般的模型 (2) 的联邦学习还未见报道. 考虑模型 (2) 数据建模的高度灵活性和纵向生存数据的普遍性, 本文将推广 Dai 等进一步研究基于模型 (2) 的右删失数据纵向联邦统计学习问题^[10].

一、理论方法

记 $C \in R^+$ 为删失时间, $Y = \min\{T, C\}$ 为观测事件时间, $\delta = I(T \leq C)$ 为删失示性函数, 则观测总体为 (Y, δ, Z) . 假设现有 K 个机构共同存储 n 个个体数据, 第 i 个个体观测数据为 $\{y_i, \delta_i, Z_i\}$. 在纵向分布式结构下, 响应变量 (y_i, δ_i) 通常由中心服务器持有; 而全局协变量观测 Z_i 分散存储于 K 个不同的机构中, 即 Z_i 被如下划分分别存在 K 个机构中:

$$Z_i = (z_{i1}^T, z_{i2}^T, \dots, z_{iK}^T)^T,$$

其中 $z_{ik} \in R^{p_k}$ 为第 i 个个体在第 k 个机构存储的部分协变量数据, 从而有 $\beta^T Z_i = \sum_{k=1}^K \beta_k^T z_{ik}$, 相应的对数似然函数为:

$$l(\beta, H) = \sum_{i=1}^n \left\{ \delta_i \ln \lambda \left(\ln H(y_i) + \sum_{k=1}^K \beta_k^T z_{ik} \right) + \delta_i \ln H'(y_i) - \delta_i \ln H(y_i) - \Lambda_\varepsilon \left(\ln H(y_i) + \sum_{k=1}^K \beta_k^T z_{ik} \right) \right\}, \quad (3)$$

其中, $\lambda_\varepsilon(\cdot)$ 为随机误差项 ε 的风险函数; $\Lambda_\varepsilon(\cdot)$ 为随机误差项 ε 的累计风险函数.

从非参数似然估计的角度, 我们使用递增不减的阶梯函数来逼近未知变换函数 $H(\cdot)$. 记 $r = \sum_{i=1}^n \delta_i$, 且相应递增排序可观测失效寿命为 $t_{(1)} < t_{(2)} < \dots < t_{(r)}$, 则变换函数可进行如下逼近:

$$H(t) = \sum_{m=1}^r \alpha_m 1\{t \geq t_{(m)}\},$$

其中 α_m 表示变换函数在第 m 个失效时间点 $t_{(m)}$ 上的跳跃. 于是估计变换函数 $H(\cdot)$ 转化为估计 α_m 的问题.

考虑数据纵向分布式存储环境下, 各机构仅拥有部分协变量特征数据, 且由于数据隐私问题而不能暴露或访问其他机构个体水平数据, 从而无法直接优化目标函数 (3). 为此, 引入中间参数 $\theta_{ik} = \beta_k^T z_{ik}$, 仅允许各机构与中心服务器之间传输与 θ_{ik} 的相关信息, 实现隐私保护的目和全局参数估计. 借鉴增广拉格朗日乘数法, 并充分利用样本信息, 优化问题 (3) 转化为在

$$\beta_k^T z_{ik} - \theta_{ik} = 0, k = 1, \dots, K, i = 1, \dots, n.$$

约束下, 极大化如下目标函数:

$$L(\beta, H, \theta, \eta) = -\sum_{i=1}^n \left\{ \delta_i \ln \lambda_\varepsilon \left(\ln H(y_i) + \sum_{k=1}^K \theta_{ik} \right) + \delta_i \ln H'(y_i) - \delta_i \ln H(y_i) - \Lambda_\varepsilon \left(\ln H(y_i) + \sum_{k=1}^K \beta_k^T z_{ik} \right) \right\} + \sum_{i=1}^n \sum_{k=1}^K \eta_{ik} (\beta_k^T z_{ik} - \theta_{ik}) + \frac{\rho}{2} \sum_{i=1}^n \sum_{k=1}^K \|\beta_k^T z_{ik} - \theta_{ik}\|^2, \quad (4)$$

其中, $\theta = (\theta_{11}, \theta_{12}, \dots, \theta_{nK})^T$, $\eta_{ik} > 0$ 为对偶变量, 控制每个机构的局部变量和全局变量的一致性; $\rho > 0$ 是增广项系数, 约束局部变量和全局变量之间的差异. 下面将分别对目标函数 (4) 中的参数分别进行交替分布式优化. 记 $\beta_i^{(p)}$ 为第 i 个机构在第 p 轮迭代的 β_i 估计.

(一) 本地更新 β_i

为了实现分布式计算的解耦问题, 首先引入机构 i 对样本 i 的局部线性预测中间变量:

$$\sigma_{il} = \beta_l^T z_{il}.$$

固定其他参数为上一轮迭代取值, 通过极小化下式更新参数

β_i 的估计:

$$L_i(\beta_i) = \sum_{i=1}^n \left\{ \frac{\rho}{2} \left(\beta_i^T z_{il} - \theta_{il}^{(p-1)} + \frac{\eta_{il}^{(p-1)}}{\rho} \right)^2 + \Lambda_\varepsilon \left(\ln H^{(p-1)}(y_i) + \sigma_{il}^{(p-1)} + \beta_i^T z_{il} \right) \right\},$$

其中, $\sigma_{i,l-1}^{(p-1)} = \sum_{k \neq l}^K \sigma_{ik}^{(p-1)} = \sum_{k=1}^K \sigma_{ik}^{(p-1)} - \beta_l^{(p-1)T} z_i$. 上式本质上是岭回归惩罚问题, 有效增强了局部 Hessian 矩阵的正定性及算法的数值稳定性. 从而可使用如下牛顿迭代法更新 β_i :

$$\beta_i^{(p,q)} = \beta_i^{(p,q-1)} - \left[\frac{\partial^2 L_i}{\partial \beta_i \partial \beta_i} \Big|_{\beta_i^{(p,q-1)}} \right]^{-1} \cdot \left(\frac{\partial L_i}{\partial \beta_i} \Big|_{\beta_i^{(p,q-1)}} \right),$$

其中 q 表示内循环迭代的轮次,

$$\frac{\partial L_i}{\partial \beta_i} = \sum_{i=1}^n \left[\rho \left(\beta_l^T z_{il} - \theta_{il}^{(p-1)} + \frac{\eta_{il}^{(p-1)}}{\rho} \right) z_{il} + \lambda_\varepsilon \left(\ln H^{(p-1)}(y_i) + \sigma_{i,l-1}^{(p-1)} + \beta_l^T z_{il} \right) z_{il} \right],$$

$$\frac{\partial^2 L_i}{\partial \beta_i \partial \beta_i^T} = \sum_{i=1}^n \left[\rho z_{il} z_{il}^T + \lambda'_e \left(\ln H^{(p-1)}(y_i) + \sigma_{ik}^{(p-1)} + \beta_i^T z_{il} \right) z_{il} z_{il}^T \right].$$

(二) 全局更新 θ

在中心服务器，给定其他参数最近更新值 $\eta_{ik}^{(p-1)}, \theta_{ik}^{(p-1)}, H^{(p-1)}$ ，以及接受本轮由各个本地机构汇总的中间变量 $\sigma_{ik}^{(p)}$ ，通过极小化下式更新参数 θ ：

$$L(\theta) = -\sum_{i=1}^n \delta_i \ln \lambda_e \left(\ln H^{(p-1)}(y_i) + \sum_{k=1}^K \theta_{ik}^{(p)} \right) + \sum_{i=1}^n \sum_{k=1}^K \frac{\rho}{2} \left(\theta_{ik}^{(p)} - \sigma_{ik}^{(p)} - \frac{\eta_{ik}^{(p-1)}}{\rho} \right)^2. \quad (5)$$

为了提高优化速度，引入 $\bar{\theta}_i = \frac{1}{K} \sum_{k=1}^K \theta_{ik}$ ，运用 ADMM 算法，

以上优化问题转变为以下带有 $\bar{\theta}_i = \frac{1}{K} \sum_{k=1}^K \theta_{ik}$ 约束的优化问题：

$$\min_{\theta_{ik}, \bar{\theta}_i} L(\bar{\theta}_i, \theta_{ik}) = -\sum_{i=1}^n \delta_i \ln \lambda_e \left(\ln H^{(p-1)}(y_i) + K \bar{\theta}_i^{(p)} \right) + \sum_{i=1}^n \sum_{k=1}^K \frac{\rho}{2} \left(\theta_{ik}^{(p)} - \sigma_{ik}^{(p)} - \frac{\eta_{ik}^{(p-1)}}{\rho} \right)^2.$$

经过简单计算，该优化等价于以下带有约束的优化问题：

$$\begin{aligned} \min_{\theta_{ik}, \bar{\theta}_i} L(\bar{\theta}_i, \theta_{ik}) &= \sum_{i=1}^n \sum_{k=1}^K \left\{ \frac{\rho}{2} \left[\theta_{ik}^{(p)} - \left(\sigma_{ik}^{(p)} + \frac{1}{\rho} \eta_{ik}^{(p-1)} \right) \right]^2 - \frac{\rho}{2} \left(\sigma_{ik}^{(p)} + \frac{1}{\rho} \eta_{ik}^{(p-1)} \right)^2 \right\} \\ &\quad - \sum_{i=1}^n \delta_i \ln \lambda_e \left(\ln H(y_i)^{(p-1)} + K \bar{\theta}_i^{(p)} \right), \\ s.t. \bar{\theta}_i &= \frac{1}{K} \sum_{k=1}^K \theta_{ik}, i=1, \dots, n. \end{aligned}$$

固定其他参数，仅保留关于 θ_{ik} 的部分，那么问题转为以下关于 θ_{ik} 的二次规划问题：

$$\min_{\theta_{ik}} \sum_{i=1}^n \sum_{k=1}^K \frac{\rho}{2} \left[\theta_{ik}^{(p)} - \left(\sigma_{ik}^{(p)} + \frac{\eta_{ik}^{(p-1)}}{\rho} \right) \right]^2, s.t. \bar{\theta}_i = \frac{1}{K} \sum_{k=1}^K \theta_{ik}.$$

运用拉格朗日乘数法，有：

$$F = \sum_{i=1}^n \sum_{k=1}^K \frac{\rho}{2} \left[\theta_{ik}^{(p)} - \left(\sigma_{ik}^{(p)} + \frac{\eta_{ik}^{(p-1)}}{\rho} \right) \right]^2 + \sum_{i=1}^n \gamma_i \left(\bar{\theta}_i^{(p)} - \frac{1}{K} \sum_{k=1}^K \theta_{ik}^{(p)} \right),$$

其中， $\sigma_{ik} = \beta_k^T z_{ik}$ 为汇总的中间结果， $\gamma_i > 0$ 为该等式约束的拉格朗日乘子。

对 F 关于 $\theta_{ik}^{(p)}$ 求偏导并令其为 0，可得：

$$\frac{\partial F}{\partial \theta_{ik}^{(p)}} = \rho \left[\theta_{ik}^{(p)} - \left(\sigma_{ik}^{(p)} + \frac{\eta_{ik}^{(p-1)}}{\rho} \right) \right] - \frac{\gamma_i}{K} = 0.$$

于是 $\theta_{ik}^{(p)}$ 可更新为：

$$\theta_{ik}^{(p)} = \sigma_{ik}^{(p)} + \frac{\eta_{ik}^{(p-1)}}{\rho} + \frac{\gamma_i}{\rho K}.$$

将上式代入原约束条件 $\bar{\theta}_i^{(p)} = \frac{1}{K} \sum_{k=1}^K \theta_{ik}^{(p)}$ ，有：

$$\bar{\theta}_i^{(p)} = \frac{1}{K} \sum_{k=1}^K \left(\sigma_{ik}^{(p)} + \frac{\eta_{ik}^{(p-1)}}{\rho} \right) + \frac{\gamma_i}{\rho K}.$$

于是，有

$$\theta_{ik}^{(p)} = \bar{\theta}_i^{(p)} + \sigma_{ik}^{(p)} + \frac{\eta_{ik}^{(p-1)}}{\rho} - \frac{1}{K} \sum_{k=1}^K \left(\sigma_{ik}^{(p)} + \frac{\eta_{ik}^{(p-1)}}{\rho} \right). \quad (6)$$

从而，关于 θ_{ik} 的估计问题可以转变为关于 $\bar{\theta}_i$ 的更新问题，即将 (6) 式代入到 (5) 式可得：

$$\min_{\bar{\theta}_i} L(\bar{\theta}_i) = -\sum_{i=1}^n \delta_i \ln \lambda_e \left(\ln H(y_i) + K \bar{\theta}_i \right) + K \rho \sum_{i=1}^n \left[\frac{\bar{\theta}_i^2}{2} - \left(\bar{\sigma}_i + \frac{\bar{\eta}_i}{\rho} \right) \bar{\theta}_i \right].$$

于是，可通过如下牛顿迭代法更新 $\bar{\theta}_i$ 估计：

$$\bar{\theta}_i^{(p,q)} = \bar{\theta}_i^{(p,q-1)} - \left[\frac{\partial^2 L}{\partial \bar{\theta}_i^2} \Big|_{\bar{\theta}_i^{(p,q-1)}} \right]^{-1} \cdot \left(\frac{\partial L}{\partial \bar{\theta}_i} \Big|_{\bar{\theta}_i^{(p,q-1)}} \right),$$

其中 q 表示牛顿迭代的轮次，

$$\frac{\partial L}{\partial \bar{\theta}_i} = -K \delta_i \frac{\lambda'_e \left(\ln H(y_i) + K \bar{\theta}_i \right)}{\lambda_e \left(\ln H(y_i) + K \bar{\theta}_i \right)} + K \rho \left(\bar{\theta}_i - \bar{\sigma}_i - \frac{\bar{\eta}_i}{\rho} \right),$$

$$\begin{aligned} \frac{\partial^2 L}{\partial \bar{\theta}_i^2} &= \frac{K^2 \delta_i}{\left(\lambda_e \left(\ln H(y_i) + K \bar{\theta}_i \right) \right)^2} \left[\left(\lambda'_e \left(\ln H(y_i) + K \bar{\theta}_i \right) \right)^2 \right. \\ &\quad \left. - \lambda'_e \left(\ln H(y_i) + K \bar{\theta}_i \right) \lambda_e \left(\ln H(y_i) + K \bar{\theta}_i \right) \right] + K \rho. \end{aligned}$$

最后将牛顿迭代估计 $\bar{\theta}_i^{(p)}$ 代入式 (6)，即可得到各机构的局部辅助变量 $\theta_{ik}^{(p)}$ 的估计。同时，为了便于本地机构在下一步骤中更新对偶变量，中心服务器还需要聚合计算全局预测量中间结果的均值 $\bar{\sigma}_i^{(p)}$ 与对偶变量的均值 $\bar{\eta}_i^{(p-1)}$ ，即

$$\bar{\sigma}_i^{(p)} = \frac{1}{K} \sum_{k=1}^K \sigma_{ik}^{(p)},$$

$$\bar{\eta}_i^{(p-1)} = \frac{1}{K} \sum_{k=1}^K \eta_{ik}^{(p-1)}.$$

接着，中心服务器将更新后的 $\theta_{ik}^{(p)}$ 、 $\bar{\theta}_i^{(p)}$ 以及 $\bar{\eta}_i^{(p-1)}$ 发送至各个本地机构，即完成本轮的全局辅助变量更新。

(三) 本地更新 η

各本地机构持有中心服务器下发的 $\bar{\sigma}_i^{(p)}$ 、 $\bar{\eta}_i^{(p-1)}$ 、 $\bar{\theta}_i^{(p)}$ 后，即可在各机构本地服务器更新对偶变量 η 。根据 ADMM 算法的对偶上升步骤，第 k 个机构的对偶变量更新公式为：

$$\eta_{ik}^{(p)} = \eta_{ik}^{(p-1)} + \rho \left(\sigma_{ik}^{(p)} - \theta_{ik}^{(p)} \right).$$

将 $\theta_{ik}^{(p)}$ 的估计 (6) 代入上式，并结合平均值定义

$$\bar{\sigma}_i^{(p)} = \frac{1}{K} \sum_{k=1}^K \sigma_{ik}^{(p)} \text{ 与 } \bar{\eta}_i^{(p-1)} = \frac{1}{K} \sum_{k=1}^K \eta_{ik}^{(p-1)}, \text{ 可得:}$$

$$\begin{aligned} \eta_{ik}^{(p)} &= \eta_{ik}^{(p-1)} + \rho \left[\sigma_{ik}^{(p)} - \left(\bar{\theta}_i^{(p)} + \sigma_{ik}^{(p)} + \frac{\eta_{ik}^{(p-1)}}{\rho} - \bar{\sigma}_i^{(p)} - \frac{\bar{\eta}_i^{(p-1)}}{\rho} \right) \right] \\ &= \eta_{ik}^{(p-1)} + \rho \sigma_{ik}^{(p)} - \rho \bar{\theta}_i^{(p)} - \rho \sigma_{ik}^{(p)} - \eta_{ik}^{(p-1)} + \rho \bar{\sigma}_i^{(p)} + \bar{\eta}_i^{(p-1)} \\ &= \bar{\eta}_i^{(p-1)} + \rho \left(\bar{\sigma}_i^{(p)} - \bar{\theta}_i^{(p)} \right). \end{aligned}$$

因此，对偶变量的更新公式为：

$$\eta_{ik}^{(p)} = \bar{\eta}_i^{(p-1)} + \rho \left(\bar{\sigma}_i^{(p)} - \bar{\theta}_i^{(p)} \right).$$

(四) 全局更新 α

在中心服务器更新阶梯函数的跳跃参数。将(3)式中的变换函数替换为阶梯函数,并注意到 $\sum_{k=1}^K \sigma_{ik}^{(p)} = \sum_{k=1}^K \beta_k^T z_{ik}$, 那么关于 α_m , 令其为零, 得

$$\begin{aligned} \frac{\partial l}{\partial \alpha_m} &= -\frac{\sum_{i=1}^n \delta_i 1\{y_i = t_{(m)}\}}{\alpha_m} \\ &+ \sum_{i=1}^n \frac{1\{y_i \geq t_{(m)}\}}{H^{(p-1)}(y_i)} \left\{ \lambda_\varepsilon \left(\ln H^{(p-1)}(y_i) + \sum_{k=1}^K \sigma_{ik}^{(p)} \right) \right. \\ &\quad \left. - \delta_i \frac{\lambda'_\varepsilon \left(\ln H^{(p-1)}(y_i) + \sum_{k=1}^K \sigma_{ik}^{(p)} \right)}{\lambda_\varepsilon \left(\ln H^{(p-1)}(y_i) + \sum_{k=1}^K \sigma_{ik}^{(p)} \right)} + \delta_i \right\} = 0. \end{aligned}$$

从而有

$$\alpha_m^{(p)} = \frac{\sum_{i=1}^n \delta_i 1\{y_i = t_{(m)}\}}{\psi(t_{(m)}; \sigma^{(p)}, H^{(p-1)})},$$

其中,

$$\begin{aligned} \psi(t, \sigma^{(p)}, H^{(p-1)}) &= \sum_{i=1}^n \frac{1\{y_i \geq t\}}{H^{(p-1)}(y_i)} \left\{ \lambda_\varepsilon \left(\ln H^{(p-1)}(y_i) + \sum_{k=1}^K \sigma_{ik}^{(p)} \right) \right. \\ &\quad \left. - \delta_i \frac{\lambda'_\varepsilon \left(\ln H^{(p-1)}(y_i) + \sum_{k=1}^K \sigma_{ik}^{(p)} \right)}{\lambda_\varepsilon \left(\ln H^{(p-1)}(y_i) + \sum_{k=1}^K \sigma_{ik}^{(p)} \right)} + \delta_i \right\}. \end{aligned}$$

完整算法总结如表1。

表1: VERTILEM 算法实现

步骤	操作过程
1	初始化: 给定初始参数 $p=1$, 局部参数 $\beta_k^{(0)} = 0$, 辅助变量 $\bar{\theta}_i^{(0)} = 0$, 对偶变量 $\eta_k^{(0)} = 0$, 以及基准变换函数初始化 $H^{(0)}(y_i) = 1$. 各机构本地更新参数 $\beta_k^{(p)}$: 对于第 k 个机构, $k=1, \dots, K$ 的第 p 轮次更新 2.1 本地储存数据: 上一轮旧值 $\beta_k^{(p-1)}, \eta_k^{(p-1)}$, 本地特征数据 z_{ik}
2	2.2 接收来自中心服务器的信息: $\theta_k^{(p-1)}, H^{(p-1)}, \bar{\sigma}_i^{(p-1)}$ 2.3 本地机构利用已知信息通过牛顿迭代极小化目标函数得到 $\beta_k^{(p)}$ 2.4 第 k 个机构发送更新后的中间结果 $\sigma_{ik}^{(p)} = \beta_k^{(p)T} z_{ik}$ 至中心服务器 全局更新辅助变量 $\theta_k^{(p)}$: 3.1 中心服务器储存数据: 上一轮旧值 $\theta_k^{(p-1)}$ 3.2 接收来自各本地机构发送的中间结果: $\sigma_{ik}^{(p)}$
3	3.3 中心服务器利用已知信息更新标量均值 $\bar{\theta}_i^{(p)}$ 3.4 中心服务器更新分配给各机构的局部辅助变量 $\theta_k^{(p)}$ 3.5 中心服务器聚合计算均值 $\bar{\sigma}_i^{(p)}$ 与 $\bar{\eta}_i^{(p-1)}$ 3.6 中心服务器下发 $\theta_k^{(p)}, \bar{\sigma}_i^{(p)}, \bar{\theta}_i^{(p)}, \bar{\eta}_i^{(p-1)}$ 至对应的本地机构服务器
4	机构本地更新 $\eta_k^{(p)}$: 对于第 k 个机构, $k=1, \dots, K$ 的第 p 轮次更新 4.1 接收来自中心服务器的信息: $\bar{\eta}_i^{(p-1)}, \bar{\sigma}_i^{(p)}, \bar{\theta}_i^{(p)}$ 4.2 本地机构利用已知信息进行对偶上升更新得到 $\eta_k^{(p)}$
5	全局更新 $H^{(p)}$: 5.1 中心服务器利用聚合结果 $\sigma^{(p)}$ 更新阶梯函数参数 $\alpha^{(p)}$, 进而得到 $H^{(p)}$ 5.2 中心服务器发送 $H^{(p)}$ 至各个本地服务器中
6	收敛检查: 当 $\ \bar{\theta}^{(p)} - \bar{\theta}^{(p-1)}\ _k > \varepsilon$ 且 $\ \bar{\sigma}^{(p)} - \bar{\sigma}^{(p-1)}\ _k > \varepsilon$ 时, 设置 $p=p+1$ 并重复 Step 2 - Step 5; 否则算法收敛并终止.

二、数值研究

(一) 数值模拟

假设协变量维度为 $d=6$, 有 $K=3$ 个参与方, 每方拥有 $d_k=2$ 个特征数据, 从以下模型中模拟样本量为 $n=200,500,1000$ 的随机样本:

于是得到

$$H^{(p)}(y_i) = \sum_{m=1}^r \alpha_m^{(p)} 1\{y_i \geq t_{(m)}\}..$$

上述更新过程在中心服务器基于各机构上传的中间结果 $\sigma_{ik}^{(p)}$ 进行, 全程无需调用各机构的原始协变量 z_{ik} 与局部参数 β_k , 从而在机制上严格保证了联邦生存分析架构的隐私安全性。

(五) ADMM 算法

总结以上过程, 网络相应信息传输示意图如图1所示。

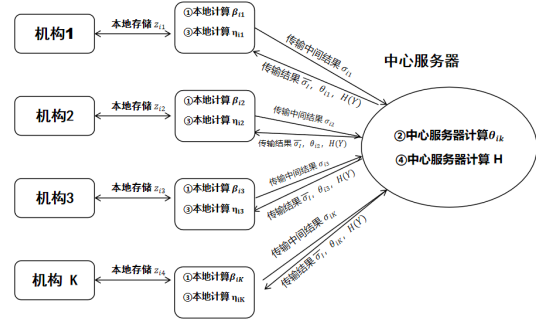


图1: 信息传输示意图

$$\ln H(T) = -\beta^T Z_i + \varepsilon_i,$$

其中 $z_{ij} \sim N(0,1)$, $\varepsilon_i \sim \text{Logistic}(0,1)$, $H(y)=y, \beta=(0.6,0.8,1.0,1.2,1.4,1.6)$. 删失时间 C_i 服从指数分布, 考虑三个删失率 (C_r): 10%、20%、30%。重复模拟 $S=200$ 次, 使用以下指标对参数估计进行评估。

$$\text{RMSE}(\hat{\beta}_j) = \sqrt{\frac{1}{S} \sum_{s=1}^S (\beta_j^{(s)} - \beta_j)^2}.$$
$$\text{MAE}(\hat{\beta}_j) = \frac{1}{S} \sum_{s=1}^S |\beta_j^{(s)} - \beta_j|.$$
$$\text{SD}(\hat{\beta}_j) = \sqrt{\frac{1}{S-1} \sum_{s=1}^S (\beta_j^{(s)} - \bar{\beta}_j)^2},$$

其中, $\bar{\beta}_j = \frac{1}{S} \sum_{s=1}^S \beta_j^{(s)}$ 。详细模拟结果见表2-3。

表2: 不同样本量和删失率下的参数估计

样本量 / 删失率	β_1	β_2	β_3	β_4	β_5	β_6
真实值	0.6000	0.8000	1.0000	1.2000	1.4000	1.6000
$n=1000$, Cr=0.1	0.6044	0.8026	0.9981	1.2058	1.3893	1.5950
$n=1000$, Cr=0.2	0.5966	0.8106	1.0014	1.2073	1.3975	1.6012
$n=1000$, Cr=0.3	0.5930	0.8025	1.0027	1.1976	1.4031	1.6047
$n=500$, Cr=0.1	0.5970	0.8029	1.0064	1.2060	1.4044	1.5988
$n=500$, Cr=0.2	0.6063	0.8037	1.0060	1.2008	1.4035	1.5998
$n=500$, Cr=0.3	0.6059	0.8039	1.0091	1.2053	1.3996	1.6041
$n=200$, Cr=0.1	0.5967	0.8161	0.9995	1.2143	1.4352	1.5980
$n=200$, Cr=0.2	0.6062	0.8167	1.0056	1.2114	1.4339	1.6121
$n=200$, Cr=0.3	0.6099	0.8102	1.0163	1.1992	1.4269	1.6103

表3: 不同样本量和删失率下的估计性能指标

n	Cr	RMSE	MAE	SD
200	0.1	0.1460	0.1147	0.1455
	0.2	0.1425	0.1121	0.1418
	0.3	0.1485	0.1174	0.1482
500	0.1	0.0901	0.0718	0.0902
	0.2	0.0891	0.0707	0.0892
	0.3	0.0928	0.0731	0.0928
1000	0.1	0.0615	0.0486	0.0614
	0.2	0.0631	0.0507	0.0630
	0.3	0.0644	0.0515	0.0644

由表 2-3 可看出, 随着样本量的增大, 删失比例的降低, 估计精度越来越好。特别地, 当样本量达到 1000 时, 各个回归系数估计的偏, RME、MAE、SD 均非常小, 在误差容许范围内。

为了说明我们所提出的纵向联邦学习方法 (分布式 ADMM) 的有效性与合理性, 本文还汇报数据集集中情形下 (即 $K=1$) 的极大似然估计结果, 并与我们所提出方法结果进行比较。此时, 仅考虑样本量 $n=1000$ 的结果。结果汇总如表 4 所示。

表4: 分布式 ADMM 与集中式估计在不同删失率下的性能比较 ($n=1000$)

删失率	方法	估计精度		运行时间
		RMSE	MAE	时间 (s)
10%	分布式 ADMM ($K=3$)	0.0615	0.0486	244.1
	集中式估计 ($K=1$)	0.0518	0.0425	1.93
20%	分布式 ADMM ($K=3$)	0.0631	0.0507	243.9
	集中式估计 ($K=1$)	0.0553	0.0462	1.44
30%	分布式 ADMM ($K=3$)	0.0644	0.0515	229.4
	集中式估计 ($K=1$)	0.0605	0.0504	1.39

由表 4 可看出, 分布式 ADMM 算法与传统的集中式估计方法在不同删失率下均表现相当, 具体来看, 在 10% 删失率下, 分布式的 RMSE 为 0.0615, 仅比集中式的高出不到 0.01; 当删失率增大到 30% 时, 两者的精度差距也没有扩大。这说明本文所提出的方法在实现隐私保护的情况下, 结果与集中式算法的估计效率相当, 但是付出的时间成本较高, 这与中心服务器与本地机构信息通讯造成的正常后果。

(二) 实例分析

在本部分, 将我们所研究的方法和集中式极大似然估计应用于经典的 UARU 药物滥用数据, 主要用来评价不同的治疗方案对于药物滥用复发时间的影响, 该数据集被广泛用于生存分析研究。剔除缺失值后共有 574 个有效记录。包含生存时间和删失示性函数及 8 个关键协变量, 经标准化和独热编码后扩展为 10 个特征: TIME (生存时间) 表示患者从入组到首次再次滥用药物的间隔天数 (响应变量), STATUS (删失指示): 表示是否确切观测到患者再次滥用药物事件 (1 表示发生复吸, 0 表示数据删失), AGE (年龄) 表示患者入组治疗时的生理年龄, BECK (贝克抑郁评分) 表示患者入组时测定的抑郁程度量表评分, NDT (既往治疗次数): 表示患者既往接受药物滥用治疗的累计次数, SITE (治疗地点) 表示患者接受治疗的具体临床中心机构, TREAT (治疗时长) 表示患者接受的是短期还是长期治疗方案, RACE (种族) 表示患者的自我认同种族分类, HERCOC (混合吸毒史) 表示患者是否有海洛因与可卡因混合使用的历史, IVHX (静脉注射史) 表示患者过往静脉注射吸毒的历史状态 (如从不、曾经或近期)。将 10 个特征属性按照业务逻辑纵向切分到 $K=3$ 个参与机构, 机构 1 持有人口学特征 (AGE, RACE) 和部分药物使用史数据; 机构 2 持有临床评估特征 (BECK、NDT) 和静脉注射史; 机构 3 持有治疗干预相关信息 (如 TREAT, SITE)。为了评估我们研究的方法的合理性, 我们还将基于全部数据集集中式训练得到的估计量作为基准真值, 记为 $\hat{\beta}_{con}$; 将通过 VERTILEM 算法在分布式环境下得到的估计量记为 $\hat{\beta}_{dist}$ 。为了度量二者数值上的差别, 使用均方误差、绝对误差、最大绝对误差、最大相对误差率来评价参数估计效率。

表 5 展示了分布式算法与集中式基准方法在 UARU 数据集上的评估指标的对比结果。实验结果表明, 分布式算法与集中式算法有着类似的估计精度。具体而言, 两组估计量之间的均方误差 (MSE) 仅为 1.86×10^{-11} , 最大绝对误差被严格控制在 1.02×10^{-5} 的水平。

表5：UARU 数据集下分布式估计与集中式基准的误差度量	
评价指标 (Metric)	数值 (Value)
均方误差 (MSE)	1.8570×10^{-11}
绝对误差和 (SAD)	3.7750×10^{-5}
最大绝对误差 (MAD)	1.0169×10^{-5}
最大相对误差率 (MDR)	1.2019×10^{-5}

表6列出了部分核心风险因子的回归系数对比。从表6中可以看出，无论是对于连续变量还是分类变量，分布式估计值 $\hat{\beta}_{dist}$ 与集中式基准估计值 $\hat{\beta}_{cent}$ 在精度上高度一致。

表6：核心风险因子的回归系数对比			
变量 (Variable)	分布式估计 ($\hat{\beta}_{dist}$)	集中式基准 ($\hat{\beta}_{cent}$)	差异
AGE	1.74273	1.74273	-2.54×10^{-61}
BECK	-0.59771	-0.59771	2.81×10^{-8}
NDT	-9.11761	-9.11762	2.43×10^{-6}
IVHX (Level 1)	0.61040	0.61041	-4.06×10^{-6}
IVHX (Level 2)	0.52303	0.52303	-1.09×10^{-6}

两种方法得到的估计结果均表明，既往治疗次数（NDT）和年龄（AGE）的回归系数绝对值较大，这与临床戒毒干预的实际经验相符：患者过往的干预历史及其生理特征，通常是预测其远

期复吸风险的关键因子。对于静脉注射史（IVHX）等包含层级递进
的分类变量，同样说明了不同用药历史之间的微小风险差异。

三、结论

在生存数据纵向分布式存储情形下，我们研究了变换模型的联邦统计学习问题。首先，基于极大似然估计构建全局似然函数；然后通过引入机构节点特征线性组合中间变量，将全局似然函数的优化问题转化为中心服务器协调下的联合优化，既实现了隐私信息的保护，也保证了与数据集中式的极大似然估计类似的估计效果；最后通过模拟研究与案例分析说明了我们所提出的纵向联邦学习方法具有较好的合理性和科学性。

在实际问题中，往往还存在着左删失数据、区间删失数据等数据类型，同时
在各个机构样本量都较大时，中心服务器难以处理大规模的中间变量的信息传
输，关于这些问题的联邦统计学习问题将是我们未来进一步研究的方向。

参考文献

[1]Cox DR. Regression models and life-tables[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1972, 34(2): 187–202.

[2]Cox DR. Partial likelihood[J]. Biometrika, 1975, 62(2): 269–276.

[3]Ren JJ, Zhou M. Full likelihood inferences in the Cox model: an empirical likelihood approach[J]. Annals of the Institute of Statistical Mathematics, 2011, 63(5): 1005–1018.

[4]Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data[J]. Biometrika, 1995, 82(4): 835–845.

[5]Zeng D, Lin YD. Maximum likelihood estimation in semiparametric regression models with censored data[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2007, 69(4): 507–564.

[6]Achcar JA, Barili E, Martinez EZ. Survival analysis of critically ill patients with cancer: use of semiparametric (transformation models) under a hierarchical Bayesian approach[J]. Brazilian Journal of Biometrics, 2025, 43(3): e–43722.

[7]Lu CL, Wang S, Ji Z, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing[J]. Journal of the American Medical Informatics Association, 2015, 22(6): 1212–1219.

[8]Duan R, Boland MR, Liu Z, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm[J]. Journal of the American Medical Informatics Association, 2020, 27(3): 430–437.

[9]Dai W, Jiang X, Bonomi L, et al. VERTICOX: Vertically distributed Cox proportional hazards model using the alternating direction method of multipliers[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(2): 996–1010.