Statistics and Application 统计学与应用, 2026, 15(9), 217-226
Published Online September 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.159220

Hans汉斯

加速失效模型下大规模生存数据的分布式模型平均

梅一平，邢玉展，陈晓静，孔泽滢，陈晓林*

曲阜师范大学统计与数据科学学院，山东 济宁

收稿日期：2026年8月20日；录用日期：2026年9月12日；发布日期：2026年9月21日

摘 要

大规模分布式数据的出现，给传统统计分析带来了巨大挑战，也使得分布式统计分析近年来成为统计研究中的热点问题。分布式模型平均方法虽能有效处理大规模数据中的模型不确定性问题，但据作者了解，现有文献仅研究了完全数据的场景，尚未涉及生存数据的分布式模型平均。针对大规模分布式右删失生存数据，利用逆概率删失加权技术，本文提出了加速失效模型下的两种分布式Mallows类型的模型平均方法。两种方法分别采用单轮型估计算法与替代最小二乘的迭代型估计算法进行备选模型的回归系数估计，但在权重选择上均统一采用局部最优权重的简单平均策略。本文通过数值模拟实验验证了所提方法的有效性，并将所提方法用到了实际数据分析。结果表明，所提方法在保持与基于全数据方法相近预测精度的同时，显著提升了计算效率。

关键词

加速失效模型，分布式模型平均，大规模数据，右删失

Distributed Model Averaging for Large-Scale Survival Data under Accelerated Failure Time Models

Yiping Mei, Yuzhan Xing, Xiaojing Chen, Zeying Kong, Xiaolin Chen*

School of Statistics, Qufu Normal University, Jining Shandong

Received: August 20, 2026; accepted: September 12, 2026; published: September 21, 2026

*通讯作者。

文章引用: 梅一平, 邢玉展, 陈晓静, 孔泽滢, 陈晓林. 加速失效模型下大规模生存数据的分布式模型平均[J]. 统计学与应用, 2026, 15(9): 217-226. DOI: 10.12677/sa.2026.159220

Abstract

The emergence of massive distributed data imposes significant challenges to traditional statistical approaches, and also makes distributed statistical learning a hot topic in recent years. Although distributed model averaging methods can effectively handle model uncertainty in these large-scale datas, to the best of our knowledge, existing relevant literature has only investigated the scenarios with complete data and has not yet covered survival data. For the massive distributed survival data with rightly censoring responses, we propose two distributed Mallows-type model averaging methods using the technique of inverse probability of censoring weighting under the accelerated failure time model. These two methods differ in the estimation of the coefficient vector of candidate models—one employs a one-round algorithm while the other uses an iterative algorithm based on surrogate least squares—but both adopt a simple averaging strategy based on locally optimal weights for weight aggregation. The effectiveness of the proposed methods is validated through numerical simulation studies, and the methods are also applied to a real data analysis. The results show that the proposed approaches maintain the prediction accuracy similar to that of full-data methods and yield a marked improvement in computational time.

Keywords

Accelerated Failure Time Model, Distributed Model Averaging, Large-Scale Data, Right Censoring

1. 引言

随着科学技术的发展，大规模数据越来越多地出现在电子商务、社交网络、电子健康系统、公共政策、生物医学等诸多领域。与传统数据相比，大规模数据具有样本量大、数据物理分布于不同节点、数据结构复杂、变量维数高、更新速度快等特点。由于计算和存储的限制，一方面，过大的样本量使得人们无法使用传统的统计分析方法和软件分析大规模数据；另一方面，大规模数据经常物理分散在不同节点、数据中心等。因此，大规模数据为统计学的发展带来了新的挑战和机遇。

大规模数据统计建模和推断研究的一个重要内容是如何用传统的、标准的统计模型和方法来拟合大规模数据。这个问题的困难之处有以下三点：第一，由于样本量过大，大规模数据已不能完全存储于单个计算机内存中；第二，分析大规模数据时，传统统计方法需要的计算时间远远超出了可以接受的范围；第三，大规模数据在物理上经常分布式存储于不同节点、数据中心等，且由于隐私保护数据无法共享。分治策略是处理大规模分布式数据的一种非常自然的框架。近年来，基于分治策略的分布式统计学习方法成为统计学研究的热点，且取得了显著进展。相关的研究进展见综述文章 Gao 等(2022) [1]、Li 等(2024) [2]及其中的参考文献。

在实际数据分析中，数据的真实生成模型一般是未知的，实际工作者因此会面临模型不确定性的问题。作为一种集成学习方法，模型平均能有效处理模型的不确定性问题。模型平均方法主要分为贝叶斯模型平均与频率模型平均两大类，其中频率模型平均依据其对模型参数的假定，可进一步划分为固定参数框架、局部渐近框架两种类型。在线性模型下，Hansen (2007) [3]提出了基于 Mallows 准则的最小二乘

模型平均方法，并证明了权重选择的最优性。Hansen (2007) [3]关于频率模型平均方法中最优权重选择的研究是最优模型平均领域的奠基性工作。该工作不仅激发了后续大量相关研究，也系统性地推动了整个模型平均领域的发展。过去的二十年里，最优模型平均方法的适用性已显著拓展，覆盖了广泛的统计学习模型和多种数据类型，包括广义线性模型(Zhang, 2016) [4]、非线性回归模型(Feng 等，2022) [5]、分位数回归模型(Lu 和 Su，2015) [6]、高维数据(Ando 和 Li，2014) [7]、缺失数据(Wei 等，2021) [8]等。生存数据是一类与生存时间密切相关的复杂数据。该类数据的复杂之处在于由于某些原因，例如病人退出试验、试验的终止等，生存时间往往被右删失。生存数据是一类与事件发生时间密切相关的复杂数据类型，其复杂性主要源于观测过程中的右删失现象，例如因受试者退出试验或研究终止等原因，导致确切的生存时间无法被完整观测。近年来，模型平均的研究也被扩展到生存分析领域，例如 He 等(2020) [9]、Liang 等(2022) [10]、He 等(2023) [11]等。

近年来，大规模数据的快速增长推动了对大规模数据模型平均方法的研究。Fang 等(2018) [12]针对线性模型与广义线性模型开展了分布式模型平均的初步研究，但该研究并没有给出严格的理论性质的证明。在线性回归模型下，Zhang 等(2023) [13]引入了两种基于最小二乘 Mallows 准则的模型平均方法，并严格证明了权重选择的渐近最优性等理论性质。Zhang 等(2023) [13]工作的局限性在于需要候选模型为嵌套结构。Xia 等(2025) [14]进一步拓展了 Zhang 等(2023) [13]的工作，允许候选模型为非嵌套结构。Song 等(2026) [15]研究了非参数多元可加模型下的分布式 Mallows 模型平均方法。与现有其他文献不同，Zhang 等(2025) [16]首次在局部渐近框架下对分布式模型平均进行了研究。尽管已有部分文献对分布式模型平均进行了研究，但据作者所知，目前尚未有文献探讨面向大规模生存数据的分布式模型平均问题。

针对大规模分布式右删失生存数据，本文融合逆概率删失加权技术、分布式估计算法与 Mallows 模型平均，提出了两种加速失效模型下的分布式模型平均方法。基于逆概率删失加权的最小二乘损失函数，两种方法首先分别采用单轮型算法与基于替代最小二乘的迭代型算法，对备选模型的回归系数进行估计。在此基础上，进一步在局部节点构造逆概率删失加权的 Mallows 类型权重准则函数，并通过简单平均的方式，对各局部节点得到的最优模型平均权重进行聚合，最终用于估计或预测。数值模拟实验结果表明，本文所提出的分布式模型平均方法在风险水平上与基于全数据的方法近似相等，但计算所需时间显著降低。本文还将新的方法应用到了一个实际数据分析中。

2. 数据与模型

设有 N 个独立个体，第 i 个个体的生存时间和(可数无穷维)预测变量分别为 T_i 和 $\boldsymbol{x}_i=\left(x_{i1},x_{i2},\cdots\right)^{\mathrm{T}}$，其服从加速失效模型

$$\log\left(T_i\right)=\mu_i+\varepsilon_i,\quad i=1,2,\cdots,N, \tag{1}$$

其中 $\mu_i=\mu\left(\boldsymbol{x}_i\right)=\sum_{j=1}^{\infty}\beta_j x_{ij}$，$\beta_j$ 为回归系数，ε_i 为随机误差且满足 $E\left(\varepsilon_i\mid\boldsymbol{x}_i\right)=0$，$D\left(\varepsilon_i\mid\boldsymbol{x}_i\right)=\sigma^2$。考虑模型(1)的 M 个线性近似模型，第 m 个备选近似模型为

$$\log\left(T_i\right)=\sum_{j=1}^{p_m}x_{ij}\beta_{j,m}+e_{i,m}=\boldsymbol{x}_{i,m}^{\mathrm{T}}\beta_m+e_{i,m},$$

其中 $\boldsymbol{x}_{i,m}=\left(x_{i1,m},x_{i2,m},\cdots,x_{ip_m,m}\right)^{\mathrm{T}}$ 为第 m 个候选模型的协变量，p_m 为协变量的维数，$e_{i,m}$ 为备选模型的随机误差。

在生物医学等领域的具体研究中，由于研究时间有限、受访者失访等原因，个体的生存时间往往被右删失了。记第 i 个个体的右删失时间为 V_i，则实际观测时间为生存时间和删失时间中的较小者。为后续叙述的方便，记 $Y_i=\log\left(T_i\right)$，$C_i=\log\left(V_i\right)$，$Z_i=\min\left\{Y_i,C_i\right\}$。令 $\delta_i=I\left(Y_i\le C_i\right)$ 为删失指示变量，其中 $I\left(\cdot\right)$

为示性函数。记全部观测个体的数据集为 $\mathcal{D}=\{(Y_i,\boldsymbol{x}_i,\delta_i),i=1,\cdots,N\}$。假设数据集 $\mathcal{D}$ 中的数据被存储于 K 个局部计算机。不失一般性，假设每个局部计算机存储的数据量相同，记为 n，则 $N=Kn$。记第 k 个局部计算机中的数据集为 $\mathcal{D}_k=\{(Y_{ki},\boldsymbol{x}_{ki},\delta_{ki})\,|\,i=1,\cdots,n\}$。

根据上述的记号，在 k 个局部计算机上，数据在正确模型和近似模型下的回归关系分别为

$$\log(T_{ki})=\sum_{j=1}^{\infty}\beta_j x_{kij}+\varepsilon_{ki}$$

和

$$\log(T_{ki})=\boldsymbol{x}_{ki,m}^{\mathrm{T}}\beta_m+e_{ki,m}。$$

3. 分布式模型平均

3.1. 基于单轮型估计的分布式模型平均

单轮型分布式估计是处理大规模数据的一种高效策略，各节点仅需与中心机器进行一次通信即可完成参数聚合。本节基于逆概率删失加权最小二乘损失函数，提出基于单轮型(one-shot, OS)估计的分布式模型平均。

对于第 m 个候选模型，在第 k 个局部计算机上，基于逆概率删失加权的最小二乘估计方法，回归系数向量 β_m 估计的损失函数为

$$Q_k(\beta_m)=\sum_{i=1}^{n}\frac{\delta_i}{1-G(Y_{ki})}\left(Y_{ki}-\boldsymbol{x}_{ki,m}^{\mathrm{T}}\beta_m\right)^2,\tag{2}$$

其中 G 是 C_i 的累积分布函数。但是，由于 G 是未知的，上述的损失函数并不能直接使用。一个直观的做法是使用第 k 个局部计算机上的数据估计 G。注意到，使用全局数据来估计 G 只需要一轮向量的传输，通信效率的损失非常少。因此，首先把各局部计算机上的观测时间和删失指示变量传输到中心计算机，其次在中心计算机上构建 G 的全局估计 $\hat{G}_N$，最后把 $\hat{G}_N$ 再传输给各局部计算机。具体地，令 $\hat{G}_N$ 是 G 的 Kaplan-Meier 估计，其中

$$\hat{G}_N(y)=1-\prod_{i=1}^{N}\left(1-\frac{1}{\sum_{j=1}^{N}I\{Y_j\geq Y_i\}}\right)^{(1-\delta_i)I(Y_i\leq y)}。\tag{3}$$

将(3)式代入(2)即得可行的局部损失函数

$$Q_k(\beta_m)=\sum_{i=1}^{n}\frac{\delta_i}{1-\hat{G}_N(Y_{ki})}\left(Y_{ki}-\boldsymbol{x}_{ki,m}^{\mathrm{T}}\beta_m\right)^2。\tag{4}$$

记 n_k 为第 k 个局部计算机上未删失观测值的数量，不失一般性，假设前 n_k 个观测值未删失，则(4)式可改写为

$$Q_k(\beta_m)=\sum_{i=1}^{n_k}\frac{1}{1-\hat{G}_N(Y_{ki})}\left(Y_{ki}-\boldsymbol{x}_{ki,m}^{\mathrm{T}}\beta_m\right)^2。\tag{5}$$

令 $\boldsymbol{Y}_{k,un}=(Y_{k1},Y_{k2},\cdots,Y_{kn_k})^{\mathrm{T}}$，$\boldsymbol{X}_{k,un}=(\boldsymbol{x}_{k1},\boldsymbol{x}_{k2},\cdots,\boldsymbol{x}_{kn_k})^{\mathrm{T}}$，$\hat{\boldsymbol{D}}_{k,un}=\mathrm{diag}(\hat{d}_1,\hat{d}_2,\cdots,\hat{d}_{n_k})$，其中 $\hat{d}_i=1/\left[1-\hat{G}_N(Y_{ki})\right]$，则 β_m 的逆概率删失加权估计为

$$\hat{\beta}_{k,m}=\left(\boldsymbol{X}_{k,un,m}^{\mathrm{T}}\hat{\boldsymbol{D}}_{k,un}\boldsymbol{X}_{k,un,m}\right)^{-1}\boldsymbol{X}_{k,un,m}^{\mathrm{T}}\hat{\boldsymbol{D}}_{k,un}\boldsymbol{Y}_{k,un},$$

其中 $\boldsymbol{X}_{k,un,m}$ 是由 $\boldsymbol{X}_{k,un}$ 的前 m 个列构成的子矩阵。对第 m 个候选模型，每个局部计算机将回归系数估计传输到中心计算机，通过简单聚合得到回归系数的单轮型估计 $\hat{\beta}_{os,m}=K^{-1}\sum_{k=1}^{K}\hat{\beta}_{k,m}$。

给定各候选模型的回归系数估计和新的预测向量 $\boldsymbol{x}=\left(x_1,x_2,\cdots\right)^{\mathrm{T}}$，其回归均值的加权平均预测值为

$$\hat{\mu}_{os}\left(\boldsymbol{w}\right)=\sum_{m=1}^{M}w_m\hat{\mu}_{os,m}=\sum_{m=1}^{M}w_m\boldsymbol{x}_m^{\mathrm{T}}\hat{\beta}_{os,m},$$

其中 $\hat{\mu}_{os,m}=\boldsymbol{x}_m^{\mathrm{T}}\hat{\beta}_{os,m}$，$\boldsymbol{x}_m=\left(x_{1,m},x_{2,m},\cdots,x_{p_m,m}\right)^{\mathrm{T}}$，$\boldsymbol{w}=\left(w_1,w_2,\cdots,w_M\right)^{\mathrm{T}}$ 为权重向量，且 $\boldsymbol{w}\in\mathcal{X}=\left\{\boldsymbol{w}\in[0,1]^M:\sum_{m=1}^{M}w_m=1\right\}$。对于 $\hat{\mu}_{os}\left(\boldsymbol{w}\right)$，一个关键的问题是如何获得数据驱动的权重向量，本文借鉴 Zhang 等(2020) [17]的方法选择权重。

在第 k 个局部计算机上，定义如下的 Mallows 类型的权重选择函数

$$L_k\left(\boldsymbol{w}\right)=\left\|\boldsymbol{Y}_{k,un}-\hat{\mu}_k\left(\boldsymbol{w}\right)\right\|^2+\phi_N\hat{\sigma}_k^2\sum_{m=1}^{M}w_m p_m,$$

其中 $\hat{\mu}_k\left(\boldsymbol{w}\right)=\sum_{m=1}^{M}w_m\hat{\mu}_{k,m}$，$\hat{\mu}_{k,m}=\boldsymbol{X}_{k,un,m}^{\mathrm{T}}\hat{\beta}_{os,m}$，$\hat{\sigma}_k^2=\left\|\boldsymbol{Y}_k-\hat{\mu}_{k,m^*}\right\|^2\Big/\left(n-p_{m^*}\right)$，$m^*=\arg\max_{1\le m\le M}p_m$。基于损失函数 $L_k\left(\boldsymbol{w}\right)$，计算第 k 个局部计算机上的最优权重 $\hat{\boldsymbol{w}}_k=\arg\min_{\boldsymbol{w}\in\mathcal{X}}L_k\left(\boldsymbol{w}\right)$，并将 K 个最优权重传输给中心计算机，进而通过简单加权平均的方式得到权重的全局估计量 $\hat{\boldsymbol{w}}=K^{-1}\sum_{k=1}^{K}\hat{\boldsymbol{w}}_{\boldsymbol{k}}$。新观测的预测向量 $\boldsymbol{x}$ 的最终预测值为

$$\hat{\mu}_{os}\left(\hat{\boldsymbol{w}}\right)=\sum_{m=1}^{M}\hat{w}_m\hat{\mu}_{os,m}=\sum_{m=1}^{M}\hat{w}_m\boldsymbol{x}_m^{T}\hat{\beta}_{\mathrm{os},m}。$$

3.2. 基于梯度增强迭代估计的分布式模型平均

单轮型分布式估计方法虽然具有较高的通讯效率，但是当局部计算机中数据的样本量较小时，局部估计偏差可能影响最终的估计。梯度增强的分布式迭代估计算法是单轮型分布式估计的重要改进，它在局部计算和全局聚合之间迭代。特别地，在该类算法的每一轮迭代中，每个节点计算机都会基于自身样本，以及通过通信获取的来自其它所有局部计算机的梯度信息，最小化一个修正后的损失函数，从而得到该轮迭代的局部估计(Jordan 等(2019) [18]，Fan 等(2023) [19])。受 Jordan 等(2019) [18]和 Shamir 等(2014) [20]的启发，本小节考虑基于梯度增强迭代(gradient-enhanced iterative, GEI)估计的分布式模型平均。

对于第 m 个候选模型，在第 k 个局部计算机上，从(5)式可知局部损失函数 $Q_k\left(\beta_m\right)$ 的梯度函数为

$$\nabla Q_k\left(\beta_m\right)=-\boldsymbol{X}_{k,un,m}^{\mathrm{T}}\hat{\boldsymbol{D}}_{k,un}\left(\boldsymbol{Y}_{k,un}-\boldsymbol{X}_{k,un,m}\beta_m\right).$$

第 m 个候选模型回归系数估计的迭代过程如下：给定第 l 次迭代的全局估计值 $\hat{\beta}_m^{(l)}$，中心计算机首先将其传输给各局部计算机，并局部计算梯度函数 $\nabla Q_k\left(\hat{\beta}_m^{(l)}\right)$；其次，中心计算机收集各局部计算机的梯度函数值，计算 $\nabla Q\left(\beta_m\right)=K^{-1}\sum_{k=1}^{K}\nabla Q_k\left(\beta_m\right)$，再将 $\nabla Q\left(\beta_m\right)$ 分别传播至各局部计算机；再次，在第 k 个局部计算机上定义梯度增强的替代损失函数

$$\tilde{Q}_k\left(\beta_m\right)=Q_k\left(\beta_m\right)-\beta_m^{\mathrm{T}}\left[\nabla Q_k\left(\hat{\beta}_m^{(l)}\right)-\nabla Q\left(\hat{\beta}_m^{(l)}\right)\right],$$

接下来求解第 k 个局部计算机的下一次迭代估计 $\hat{\beta}_{k,m}^{(l+1)}=\arg\min_{\beta_m}\tilde{Q}_k\left(\beta_m\right)$；最后，各局部计算机将局部估计传输给中心计算机，中心计算机计算全局回归系数估计的第 $l+1$ 次迭代值，即

$$\hat{\beta}_m^{(l+1)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_{k,m}^{(l+1)} = \hat{\beta}_m^{(l)} - \frac{1}{K}\sum_{k=1}^{K}\left[\left(\boldsymbol{X}_{k,un,m}^{\mathrm{T}}\hat{\boldsymbol{D}}_{k,un}\boldsymbol{X}_{k,un,m}\right)^{-1}\left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{X}_{k,un,m}^{\mathrm{T}}\hat{\boldsymbol{D}}_{k,un}\hat{\varepsilon}_{km}^{(l)}\right)\right],$$

其中 $\hat{\varepsilon}_{km}^{(l)} = \boldsymbol{Y}_{k,un} - \boldsymbol{X}_{k,un,m}\hat{\beta}_m^{(l)}$。假设经过 L 次迭代后，迭代过程收敛，则第 m 个候选模型回归系数的估计为 $\hat{\beta}_{iter,m} = \hat{\beta}_m^{(L)}$。

类似于第 2.1 节，给定各候选模型的回归系数估计和新的预测向量 $\boldsymbol{x} = (x_1, x_2, \cdots)^{\mathrm{T}}$，其回归均值的加权平均预测值为

$$\hat{\mu}_{iter}(\boldsymbol{w}) = \sum_{m=1}^{M} w_m \hat{\mu}_{iter,m} = \sum_{m=1}^{M} w_m \boldsymbol{x}_m^{\mathrm{T}}\hat{\beta}_{iter,m},$$

其中 $\hat{\mu}_{iter,m} = \boldsymbol{x}_m^{\mathrm{T}}\hat{\beta}_{iter,m}$，$\boldsymbol{x}_m = (x_{1,m}, x_{2,m}, \cdots, x_{p_m,m})^{\mathrm{T}}$，$\boldsymbol{w} = (w_1, w_2, \cdots, w_M)^{\mathrm{T}}$ 为权重向量。在第 k 个局部计算机上，定义权重选择函数

$$L_k^*(\boldsymbol{w}) = \left\|\boldsymbol{Y}_{k,un} - \hat{\mu}_k^*(\boldsymbol{w})\right\|^2 + \phi_N\hat{\sigma}_k^2\sum_{m=1}^{M} w_m p_m,$$

其中 $\hat{\mu}_k^*(\boldsymbol{w}) = \sum_{m=1}^{M} w_m\hat{\mu}_{k,m}^*$，$\hat{\mu}_{k,m}^* = \boldsymbol{X}_{k,un,m}^{\mathrm{T}}\hat{\beta}_{iter,m}$，$\hat{\sigma}_k^2 = \left\|\boldsymbol{Y}_k - \hat{\mu}_{k,m^*}^*\right\|^2 \Big/ \left(n - p_{m^*}\right)$，$m^* = \arg\max_{1\le m\le M} p_m$。基于损失函数 $L_k^*(\boldsymbol{w})$，计算第 k 个局部计算机上的最优权重 $\hat{\boldsymbol{w}}_k^* = \arg\min_{\boldsymbol{w}\in\mathcal{W}} L_k^*(\boldsymbol{w})$，并将 K 个最优权重传输给中心计算机，进而通过简单加权平均的方式得到权重的全局估计量 $\hat{\boldsymbol{w}}^* = K^{-1}\sum_{k=1}^{K}\hat{\boldsymbol{w}}_k^*$。新观测的预测向量 $\boldsymbol{x}$ 的最终预测值为

$$\hat{\mu}_{iter}(\hat{\boldsymbol{w}}^*) = \sum_{m=1}^{M}\hat{w}_m^*\hat{\mu}_{iter,m} = \sum_{m=1}^{M}\hat{w}_m^*\boldsymbol{x}_m^{\mathrm{T}}\hat{\beta}_{iter,m}。$$

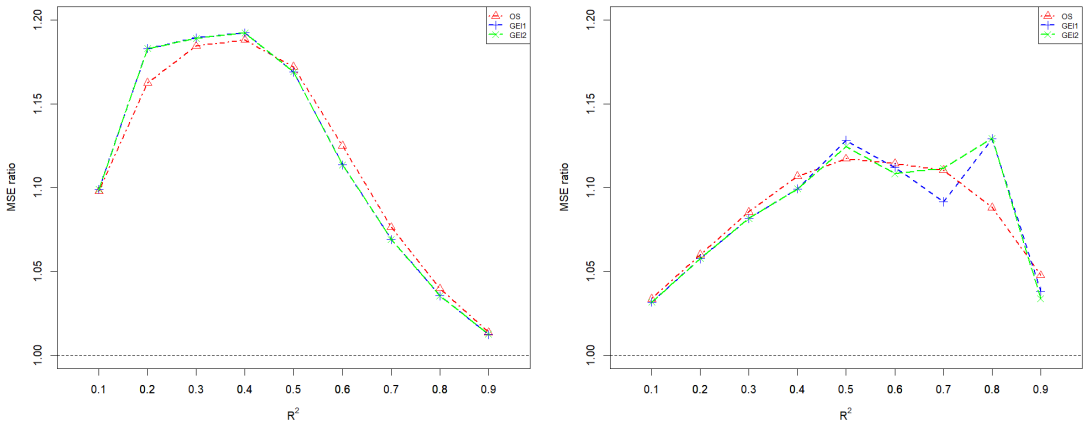
4. 数值模拟

本节通过数值模拟评估所提出的分布式模型平均方法在有限样本下的表现，并与基于全部数据的模型平均方法进行比较。具体地，考虑如下的四种模型平均方法：1) FULL，基于全部数据的模型平均方法；2) OS，单轮型(OS)分布式模型平均；3) GEI1，基于局部加权最小二乘估计初始化的梯度增强迭代(GEI)分布式模型平均；4) GEI2，基于 OS 全局估计初始化的梯度增强迭代(GEI)分布式模型平均。

数值模拟数据的生成过程如下：首先，独立生成 N 个 p 维预测向量 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}}$，其中 $p = 50$，$x_{i1} = 1$，x_{ij}，$j = 2, \cdots, p$，是独立从标准正态分布生成的随机数；其次，基于下述模型生成生存时间，$\log(T_i) = \boldsymbol{x}_i^{\mathrm{T}}\beta + \varepsilon_i$，$i = 1, 2, \cdots, N$。其中 $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^{\mathrm{T}}$，$\beta_j = 1.5j^{-1}$，$j = 1, \cdots, p$，误差 ε_i 独立同分布于正态分布 $N(0, \sigma^2)$；最后，令 C_0 是某个正常数，从 $(0, C_0)$ 上的均匀分布生成 N 个对数删失时间 C_i，进而得到 $Z_i = \min\{\log(T_i), C_i\}$ 和删失指示变量 $\delta_i = I(Y_i \le C_i)$。在数据生成过程中，选择不同的 σ^2，使得决定系数 $R^2 = \mathrm{var}(\boldsymbol{x}_i^{\mathrm{T}}\beta) \Big/ \left[\mathrm{var}(\boldsymbol{x}_i^{\mathrm{T}}\beta) + \sigma^2\right]$ 分别取到 0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9；选择不同的常数 C_0，使得在不同的情形下删失率达到 40%和 60%。

在预测向量 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}}$ 中，我们选取其前 p_1 个分量进行建模。具体地，构造 $M = p_1$ 个嵌套的子模型，其中第 m 个候选模型的协变量由 $\boldsymbol{x}_i$ 的前 m 个分量构成。注意到，第 1 个候选模型仅包含截距项；第 M 个候选模型包含全部选取的 p_1 个分量，也就是候选模型中最大的模型。基于 $H = 100$ 次的数据

重复，我们汇报每个方法平均计算时间(time，单位：秒)和平均均方预测误差(mean squared prediction error, MSPE) $\mathrm{MSPE}=H^{-1}\sum_{h=1}^{H}N^{-1}\sum_{i=1}^{N}\left(\mu_i-\hat{\mu}_i\right)^2$，其中 μ_i 和 $\hat{\mu}_i$ 分别是第 i 个个体真实的和估计的条件均值。

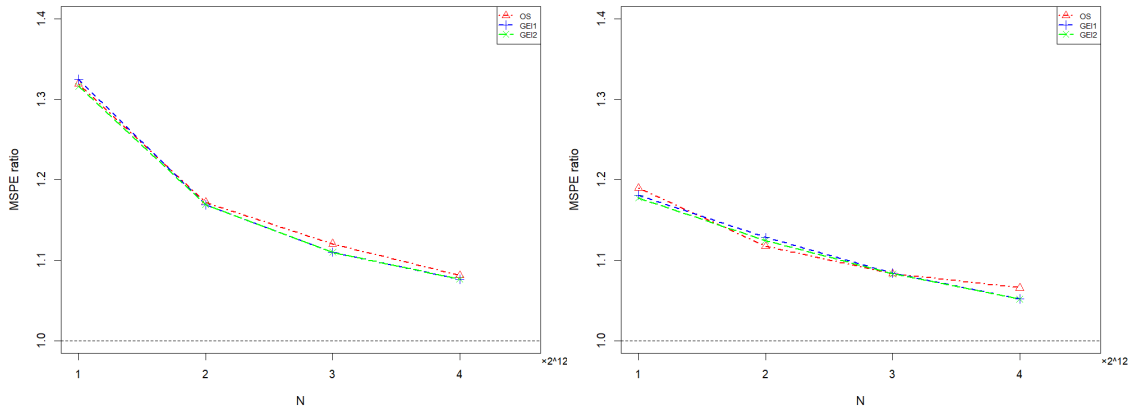


(a) 删失率 40%　　(b) 删失率 60%

Figure 1. Comparison of MSPE ratios for various methods versus the FULL method under different values of R^2
图 1. 不同 R^2 下各方法与 FULL 方法的 MSPE 比值的比较

首先，基于不同决定系数 R^2 的设定，比较本文所提方法与 FULL 方法的有限样本差异，见图 1。设定 $N=2^{13}$，$K=4$，$p_1=20$，$T=6$，图 1 展示了不同 R^2 下本文所提方法的 MSPE 和 FULL 方法的 MSPE 的比值。比值曲线越接近于 1 表明所提方法的有限样本效果越好。从图 1 的结果可以看出，在各种 R^2 的取值下，本文提出方法都展现了良好的有限样本性质，三种方法的 MSPE 都非常接近 FULL 方法的 MSPE。与此同时，梯度增强迭代型方法在多数情况下略优于单轮型方法，但并未呈现出稳定且一致的优势。

其次，进一步比较本文所提方法与 FULL 方法在不同样本量下的有限样本差异，见图 2。设定 $K=4$，$p_1=20$，$T=6$，$R^2=0.5$，图 2 展示了不同样本量 N 下 OS、GEI1 和 GEI2 的 MSPE 和 FULL 方法的 MSPE 的比值。从图 2 的结果可以看出，在各种 N 的取值下，本文提出方法都展现了良好的有限样本性质，三种方法的 MSPE 都非常接近 FULL 方法的 MSPE。



(a) 删失率 40%　　(b) 删失率 60%

Figure 2. Comparison of MSPE ratios of various methods relative to the FULL method under different sample sizes
图 2. 不同样本量下各方法与 FULL 方法的 MSPE 比值的比较

Table 1. Comparison of average computation time for different methods (Unit: seconds)
表 1. 各方法平均计算耗时的比较(单位：秒)

(N, K, R^2, p_1)	FULL	OS	GEI1	GEI2
$(2^{12}, 2, 0.5, 20)$	2.366	1.208	1.182	1.855
$(2^{13}, 2, 0.5, 20)$	5.542	2.804	2.773	2.757
$(2^{13}, 4, 0.5, 20)$	5.725	1.418	1.424	1.381
$(2^{13}, 4, 0.9, 20)$	5.714	1.415	1.419	1.379
$(2^{13}, 4, 0.9, 30)$	12.003	3.062	2.996	3.012

最后，本文结合多个不同的参数设置，综合评估各方法的计算耗时，见表 1。表 1 给出了各方法在一些情形下平均计算所需的时间。由表 1 结果可见，本文所提方法所需的平均计算时间显著低于基于完整数据的模型平均方法。并且，这种计算上的优势会随着样本量和备选模型规模的增大进一步凸显。此外，梯度增强迭代型方法和单轮型方法的计算效率相近，无明显差异。

5. 实例分析

美国国家健康与营养调查(National Health and Nutrition Examination Survey, NHANES)是由美国疾病控制与预防中心下属的国家健康统计中心(National Center for Health Statistics, NCHS)主导的一项旨在评估美国成人和儿童健康与营养状况的权威调查项目。该调查始于 20 世纪 60 年代初期，自 1999 年以来，NCHS 持续开展 NHANES，并搭建了相应的 NHANES 数据库。NHANES 收集的数据包括参与者的人口学特征、体检指标、重要的生化指标以及死亡随访信息等。本文聚焦于 NHANES 的 2009 年至 2018 年的五个连续调查周期数据，评估所提方法在实际数据分析中的预测能力和计算效率，并与基于完整数据的方法进行比较。

通过对数据进行预处理，我们得到一个包含 25313 个观测值的数据集，其删失率为 92%。除生存时间外，每个观测值包含 20 个预测变量，按与生存时间的皮尔逊相关系数的绝对值从大到小排序如下：红细胞分布宽度、平均红细胞血红蛋白浓度、平均红细胞血红蛋白含量、年龄、单核细胞计数、收缩压第三次读数、平均收缩压、家庭受访者年龄、收缩压第二次读数、最大充气水平、单核细胞百分比、嗜碱性粒细胞绝对值、平均血小板体积、嗜碱性粒细胞百分比、骨密度状态、种族、总胆固醇、胆固醇、糖化血红蛋白、红细胞计数。基于这些预测变量，类似于数值模拟研究，构建 20 个嵌套的备选模型用于模型平均。

Table 2. Comparison of four methods under NHANES data
表 2. NHANES 数据下四种方法的比较

		FULL	OS	GEI1	GEI2
$K=2$	MSPE	0.143	0.148	0.181	0.169
	Times (s)	0.719	0.358	0.348	0.350
$K=3$	MSPE	0.143	0.152	0.267	0.197
	Times (s)	0.719	0.242	0.237	0.238

在方法评估阶段，首先将数据集随机划分为训练集和测试集，其中训练集和测试集的样本量分别为 24,313 和 1000；其次，在训练集上进行模型参数和权重的估计，然后在测试集上进行预测和评估；最后，

把这个过程重复 100 次，计算平均的平方预测误差和计算耗时。表 2 给出了实际数据分析得到的平均平方预测误差和计算时间。从表 2 结果可见，本文所提方法的计算效率明显高于基于完整数据的方法；OS 方法的 MSPE 与 FULL 方法基本一致，GEI 方法的 MSPE 稍高于 FULL 方法。在梯度增强迭代分布式模型平均方法中，当分组数 $K=3$ 时的 MSPE 要高于 $K=2$ 时。究其原因，主要在于数据集的删失率较高，随着分组数的增加，各局部子集的样本量急剧缩减，致使逆概率删失加权估计的不稳定性显著加剧。

6. 结论与展望

针对大规模分布式生存数据，在加速失效模型下，本文提出了两种 Mallows 类型的分布式模型平均方法：基于单轮型估计的分布式模型平均和基于梯度增强迭代估计的分布式模型平均。本文不仅通过数值模拟验证了所提方法在有限样本下的表现，还进一步利用实际数据检验了其实用性。数值模拟与实际数据验证结果表明，本文提出的两种分布式模型平均方法均具有优良的预测性能。

在估计删失时间的分布函数时，本文采用了全局 Kaplan-Meier 估计的策略，旨在兼顾估计效率与计算可行性。具体而言，该策略能够基于全体样本提高估计的精度，且其传输向量的计算负担较轻，其所带来的时间成本是可以接受的。然而，在数据分布式存储且无法共享的场景下，全局估计策略将不再适用，需要转而寻求能充分利用分布式特性的估计方法，例如利用局部数据直接估计删失时间的分布，或采用分布式迭代算法等。此外，本文使用了最基本的 Kaplan-Meier 估计。事实上，通过引入更精细的统计手段，如核光滑技术，可以构造出具有更优良性质的删失时间的分布式估计。

据作者了解，现有针对生存数据的分布式模型平均的研究仍较为匮乏。本文仅考虑了生存时间右删失的情形，然而在实际医学、公共卫生等应用中，生存时间往往伴随更复杂的删失机制，例如信息删失、区间删失等。此外，Cox 比例风险率模型是生存分析研究中应用最广泛的模型，但其框架下的分布式模型平均尚未见相关研究。诸如此类的科学问题还很多，因此，生存数据的分布式模型平均还存在诸多亟待解决的科学问题，具备广阔的研究前景。

基金项目

教育部人文社会科学研究规划基金项目(21YJA910002)，国家级大学生创新训练计划项目(202510446018)，国家社会科学基金一般项目(25BTJ038)。

参考文献

[1] Gao, Y., Liu, W., Wang, H., Wang, X., Yan, Y. and Zhang, R. (2022) A Review of Distributed Statistical Inference. *Statistical Theory and Related Fields*, **6**, 89-99. https://doi.org/10.1080/24754269.2021.1974158

[2] Li, X., Gao, Y., Chang, H., Huang, D., Ma, Y., Pan, R., *et al*. (2024) A Selective Review on Statistical Methods for Massive Data Computation: Distributed Computing, Subsampling, and Minibatch Techniques. *Statistical Theory and Related Fields*, **8**, 163-185. https://doi.org/10.1080/24754269.2024.2343151

[3] Hansen, B.E. (2007) Least Squares Model Averaging. *Econometrica*, **75**, 1175-1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x

[4] Zhang, X., Yu, D., Zou, G. and Liang, H. (2016) Optimal Model Averaging Estimation for Generalized Linear Models and Generalized Linear Mixed-Effects Models. *Journal of the American Statistical Association*, **111**, 1775-1790. https://doi.org/10.1080/01621459.2015.1115762

[5] Feng, Y., Liu, Q., Yao, Q. and Zhao, G. (2022) Model Averaging for Nonlinear Regression Models. *Journal of Business & Economic Statistics*, **40**, 785-798. https://doi.org/10.1080/07350015.2020.1870477

[6] Lu, X. and Su, L. (2015) Jackknife Model Averaging for Quantile Regressions. *Journal of Econometrics*, **188**, 40-58. https://doi.org/10.1016/j.jeconom.2014.11.005

[7] Ando, T. and Li, K. (2014) A Model-Averaging Approach for High-Dimensional Regression. *Journal of the American Statistical Association*, **109**, 254-265. https://doi.org/10.1080/01621459.2013.838168

[8] Wei, Y., Wang, Q. and Liu, W. (2021) Model Averaging for Linear Models with Responses Missing at Random. *Annals of the Institute of Statistical Mathematics*, **73**, 535-553. https://doi.org/10.1007/s10463-020-00759-y

[9] He, B., Liu, Y., Wu, Y., Yin, G. and Zhao, X. (2020) Functional Martingale Residual Process for High-Dimensional Cox Regression with Model Averaging. *Journal of Machine Learning Research*, **21**, 1-37.

[10] Liang, Z., Chen, X. and Zhou, Y. (2022) Mallows Model Averaging Estimation for Linear Regression Model with Right Censored Data. *Acta Mathematicae Applicatae Sinica, English Series*, **38**, 5-23. https://doi.org/10.1007/s10255-022-1054-z

[11] He, B., Ma, S., Zhang, X. and Zhu, L. (2023) Rank-Based Greedy Model Averaging for High-Dimensional Survival Data. *Journal of the American Statistical Association*, **118**, 2658-2670. https://doi.org/10.1080/01621459.2022.2070070

[12] Fang, F., Yin, X. and Zhang, Q. (2018) Divide and Conquer Algorithms for Model Averaging with Massive Data. *Journal of Systems Science and Mathematical Sciences*, **38**, 764-776. (In Chinese)

[13] Zhang, H., Liu, Z. and Zou, G. (2023) Least Squares Model Averaging for Distributed Data. *Journal of Machine Learning Research*, **24**, 1-59.

[14] Xia, X., He, S. and Pang, N. (2025) Communication-Efficient Model Averaging Prediction for Massive Data with Asymptotic Optimality. *Statistical Papers*, **66**, 1-45. https://doi.org/10.1007/s00362-025-01664-3

[15] Song, M., Qu, T., Zhao, Z. and Zou, G. (2026) Optimal Distributed Model Averaging for Multivariate Additive Model. *Journal of Systems Science and Complexity*, **39**, 309-333. https://doi.org/10.1007/s11424-025-5054-y

[16] Zhang, Y., Chen, X. and Xing, Y. (2025) Distributed Focused Information Criterion and Focused Frequentist Model Averaging for Massive Data. Statistica Sinica. https://www3.stat.sinica.edu.tw/ss_newpaper/SS-2025-0060_na.pdf

[17] Zhang, X., Zou, G., Liang, H. and Carroll, R.J. (2020) Parsimonious Model Averaging with a Diverging Number of Parameters. *Journal of the American Statistical Association*, **115**, 972-984. https://doi.org/10.1080/01621459.2019.1604363

[18] Jordan, M.I., Lee, J.D. and Yang, Y. (2019) Communication-Efficient Distributed Statistical Inference. *Journal of the American Statistical Association*, **114**, 668-681. https://doi.org/10.1080/01621459.2018.1429274

[19] Fan, J., Guo, Y. and Wang, K. (2023) Communication-Efficient Accurate Statistical Estimation. *Journal of the American Statistical Association*, **118**, 1000-1010. https://doi.org/10.1080/01621459.2021.1969238

[20] Shamir, O., Srebro, N. and Zhang, T. (2014) Communication-Efficient Distributed Optimization Using an Approximate Newton-Type Method. *Proceedings of the* 31*st International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, **32**, 1000-1008.