非加性损失下不可共享数据的分布式卷积秩回归

张文、朱丽萍和杨松山

中国人民大学

arXiv:2602.04594v1 [stat.ME] 4 Feb 2026

摘要

我们研究数据分布在多台机器上且损失为非加性 U统计量时的高维秩回归问题，例如卷积秩回归（CRR）。经典的通信高效替代似然（CSL）方法关键依赖于经验损失的可加性，因此无法用于 CRR，因为其全局损失将各机器上的所有样本对耦合在一起。我们提出一种分布式卷积秩回归（DCRR）框架，构造类似的替代损失，并证明该框架在非加性损失下的有效性。我们证明，该替代损失与全数据 CRR 损失具有相同的总体极小值点，且所得估计量在统计上等价于集中式 CRR。在此基础上，我们开发了一个两阶段稀疏 DCRR 程序——先进行迭代式 ℓ_1 惩罚阶段，再进行折叠凹精化——并建立了非渐近误差界、分布式强预言机性质，以及用于一致模型选择的 DHBIC型准则。一项缩放结果表明，机器数量可按$M = o(N/(s^2 \log\ p))$增长，同时仅需 $O(\log\ N)$ 轮通信即可达到集中式预言机速率。模拟和一个大规模真实数据示例表明，与朴素的分而治之方法相比，该方法优势显著，尤其是在误差具有重尾的情况下。

关键词： 分布式学习、元学习、参数函数、最优收敛速率、U统计量。

*

1. 引言

高维回归广泛应用于基因组学、成像和金融等当代领域，其中样本量 N 和环境维度 p都可能很大，且数据可能存储在多台机器上。此外，噪声分布通常具有重尾特征或受到离群值污染，因此基于最小二乘的方法会变得不可靠。基于秩的回归提供了一种稳健的替代方法，不受单调变换和响应变量中离群值的影响；早期的开创性研究包括Jurevckov´a（1971）和Jaeckel（1972），全面论述请参阅Hettmansperger和McKean（2010）。近期，研究者提出了卷积秩回归（CRR），这是一种平滑的秩型 U统计量损失函数，既保留了传统秩估计量的稳健性，又具有可微性和良好的效率（Zhou 等，2024）。在高维情形下，Zhou 等人（2024）为惩罚化卷积秩回归确立了稀疏性预言性质；此外，针对更一般的 M估计量的相关高维稳健回归方法也已得到研究，例如 Wang等人（2020）和Loh（2017）。

尽管取得了这些进展，现有的 CRR 理论和方法本质上仍是集中式的：它们假设所有数据都存储在单台机器上，并且可以计算和求导 CRR 损失中的二次数量级成对项。当 N 较大，且由于内存、隐私或组织方面的限制，数据分布在多台机器上时，这就会成为问题。针对大规模问题的通信高效分布式估计研究日益增多；例如，参见 Zhang 等人（2013b）；Jordan 等人（2019）；Battey 等人（2018）；Fan 等人（2023）及其中的参考文献。许多此类研究聚焦于最小二乘或广义线性模型（Zhang 等人，2013a；Chen 和 Xie，2014；Rosenblatt 和 Nadler，2016；

Wang 等人，2017；Lee 等人，2017；Huang 和 Huo，2019），其中经验损失对观测值具有可加性。在可加损失结构下，分布式稳健回归和分位数回归也得到了研究，包括分布式分位数回归及其一步修正（Chen 等人，2020；Pan 等人，2022），以及基于去偏 Lasso 的稳健估计（Shi 等人，2020）。

这些方法之所以成功，一个关键原因在于，大多数通信高效替代似然（CSL）及相关框架都严重依赖于经验损失函数的可加性假设。具体而言，对于一大类 M-估计量，全数据损失可表示为局部损失的平均值，

$$\mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_m(\boldsymbol{\beta}), \qquad \nabla\mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{\beta}),$$

其中， $\mathcal{L}_m$ 是利用存储在第 m台机器上的数据计算得到的。这一精确等式使我们能够根据局部梯度重构全数据梯度，并在单台主机器上构造一个模拟全局损失的替代损失，从而实现通信效率极高的程序（Zhang 等人，2013b；Jordan等人，2019；Fan等人，2023）。基于这种可加结构，近期研究进一步为高维模型中的分布式程序确立了预言机性质：Gu 等人（2025年）提出了一种协同得分型检验，该检验在一般高维模型下具有检验统计量的预言机性质；Liu 等人（2025年）则开发了通信高效的分布式稀疏学习算法，这些算法对非凸惩罚 M估计量具有预言机性质和几何收敛性。然而，所有这些预言机结果都是在形式为$\frac{1}{N}\sum_{i=1}^{N}\ell(\boldsymbol{\beta};z_i)$的经验风险下推导得出的，并且依赖于全局损失在各台机器间的可加性。据我们所知，目前尚无针对非可加U统计量秩损失下高维估计量的分布式预言机保证，在这种情形下，全局经验损失无法分解为可

—

的平均值 加损失。填补这一空白是本文的主要目标之一。

相比之下，在卷积秩回归中，经验损失是对所有样本对构成的二阶 U-统计量（Zhou 等，2024）。全局 CRR 损失通过跨机器样本对将不同机器上的观测值联系起来，而每个本地端损失仅涉及单台机器内部的样本对。因此，完整的 CRR 损失并非各本地端损失的简单平均，相应的梯度也不满足

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{M} \sum_{m=1}^{M} \nabla \mathcal{L}_m(\boldsymbol{\beta}).$$

CRR 的 U-统计量结构与 CSL 所依赖的可加性假设之间存在这一根本性的不匹配，这意味着现有的 CSL 理论及其变体——包括 Gu 等人（2025）和 Liu 等人（2025）的分布式预言机结果——无法直接应用于分布式秩回归。具体而言，直接尝试用本地端 CRR 梯度的平均值替代 $\nabla \mathcal{L}_N(\boldsymbol{\beta})$ 会引入额外偏差。而对于高维情形下的稀疏 CRR 估计量，能否仍获得集中式最优收敛率和强预言机性质，则远非显而易见。

本文提出了一种分布式稀疏秩回归框架，克服了这一困难，并聚焦于高维估计、变量选择和模型选择。我们的关键观察是，尽管全局 CRR 损失与局部 CRR 损失在样本层面不再具有加性关系，但由于采用相同的核和独立同分布抽样方案，它们共享相同的总体风险函数。具体而言，全局与局部 CRR 损失的总体极限相同，因此它们具有相同的极小化解 $\boldsymbol{\beta}^*$。我们利用这一事实构造了一种新颖的替代损失，它只需要在单个“主节点”上计算 CRR U-统计量损失，并结合从所有机器汇总而来的梯度校正项。这种

替代损失不再与全数据损失完全相等，但我们证明，它具有相同的总体极小值点，并且在适当条件下，其极小化解在高维情形下与集中式 CRR 估计量在统计上等价。从技术上讲，这需要采用新工具，利用次高斯过程和U-统计量的集中不等式（Bousquet，2003年；Ledoux 和 Talagrand，2013年），控制 U-统计量损失的真实全数据梯度与基于局部 U-统计量的梯度近似之间的差异，并确立替代损失在稀疏方向上的受限曲率性质。

在该代理的基础上，我们提出一种用于高维稀疏估计的两阶段分布式卷积秩回归（DCRR）方法。在第一阶段，我们通过一种仅传递局部 CRR 损失梯度的迭代式分布式算法，计算 ℓ_1惩罚的DCRR估计量。在第二阶段，我们借助局部线性近似方案施加折叠凹惩罚，以改进该初始估计量，其思想源于非凸惩罚（Fan和Li，2001；Zhang，2010）以及 I-LAMM 框架（Fan等，2015）。由此得到一个计算上可扩展的估计量，它结合了秩回归的稳健性、CRR的效率，以及折叠凹正则化在高维情形下的统计优势。我们进一步提出用于调参的分布式高维 BIC（DHBIC）准则，并证明其在分布式环境中能够相合地恢复$\beta^\star$ 的真实支撑集。

与现有的高维 CRR 方法（Zhou 等，2024）相比，我们的框架可扩展至海量分布式数据集，且无需计算完整的二次 CRR 损失。与采用可加损失的分布式回归程序（Zhang等人，2013b；Battey 等人，2018年；Fan 等人，2023；Chen 等人，2020年；Pan 等人，2022年；Gu等人，2025年；Liu等人，2025年）相比，我们的方法专为非可加的

U统计量秩损失而设计；据我们所知，它是首个在非可加损失下为高维秩回归提供强预言机保证的通信高效分布式解决方案。

1.1. 我们的贡献

我们将主要贡献概述如下。

一种适用于非加性损失的通信高效 U统计量秩回归框架。 由于全局 CRR 损失涉及跨机器样本对，不再是局部损失的平均值（Zhou 等，2024），这给我们的分析带来了新的挑战。为解决这一问题，我们利用全局 CRR 损失与局部 CRR 损失具有相同总体风险和极小值点这一事实，构造了替代损失，将单台机器上的 CRR 损失与根据局部 CRR 梯度计算的梯度校正项结合起来。我们证明，该替代损失与全数据 CRR 损失具有相同的总体极小值点，并且在适当的稀疏性和正则条件下，其高维极小化解可达到与集中式 CRR 估计量相同的收敛速率。这将 CSL 理念（Zhang等人，2013b；Jordan等人，2019；Fan等人，2023）推广到了广泛的一类基于 U 统计量的秩回归问题，在这类问题中，样本层面的精确可加性并不成立。

高维秩回归的分布式强预言机性质。基于替代损失，我们提出一种两阶段分布式CRR估计量：首先进行 ℓ_1 惩罚，然后通过局部线性近似方案，利用折叠凹惩罚（Fan 和 Li，2001年；Zhang，2010年）对解进行细化（Fan 等人，2015）。我们建立了非渐近误差界，并证明，只要通信轮数适当，折叠凹 DCRR 估计量便具有分布式强预言机性质：它能够以高概率恢复 $\beta^\star$ 的真实支撑集，并达到

与预先已知支撑集的理想神谕估计量相同的估计速率，与集中式高维CRR基准相匹配（Zhou 等，2024年）。不同于近期针对可加损失的分布式预言机结果（Gu 等人，2025年；Liu 等人，2025年），我们的预言机性质是在非可加的 U 统计量秩损失下获得的，此时全局经验损失并非局部损失之和。这需要针对分布式CRR情形提出新的偏差控制和曲率论证。此外，我们的框架允许机器数量 M 随总样本量 N 增长，但需满足温和的高维尺度条件，例如 $M = o\{N/(s^2 \log p)\}$；并且只需 $O(\log N)$ 轮通信，即可达到集中式预言机速率和模型选择相合性。这比朴素的分而治之方案宽松得多，因为后者的统计精度从根本上受局部样本量限制。最后，我们的算法无需为带宽参数 h 调参，该参数只需满足 $h = O(1)$，且不需要进行数据依赖选择。

秩回归的分布式模型选择。我们提出一种用于高维秩回归调参的分布式HBIC型准则（DHBIC），并证明其模型选择相合性。在温和条件下，DHBIC 选择器能够以高概率得到恢复真实支撑集的稀疏模型，同时无需进行样本分割或重复重拟合。结合我们的估计结果，该方法为稀疏高维CRR提供了一种完全分布式的流程，兼具统计准确性和计算效率。

本文其余部分安排如下。第 2 节介绍分布式 CRR 框架，定义 DCRR 替代损失，并提出 ℓ_1 惩罚算法和折叠凹惩罚算法。第 3 节建立非渐近误差界，证明分布式强预言机性质，并论证 DHBIC

模型选择准则的一致性。第4节报告模拟研究和一项真实数据应用，以展示所提方法的数值表现。最后，我们在第5节进行简要讨论。

1.2. 符号

我们介绍一些必要的符号。对于 $a, b \in \mathbb{R}$, $a \wedge b = \min\{a, b\}$ 和 $a \vee b = \max\{a, b\}$。对于矩阵 A， $\lambda_{\min}(A)$ 和 $\lambda_{\max}(A)$ 分别表示特征值绝对值中的最小值和最大值。对于向量 a，$\mathrm{supp}(\mathsf{a}) = \{j \mid a_j \neq 0\}$。对于正整数 p, $[p] = \{1, 2, \ldots, p\}$。对于指标集 $\mathcal{A}$、向量 b 和矩阵 C，$\mathsf{b}_{\mathcal{A}} = (b_j, j \in \mathcal{A})$ 和 $\mathsf{C}_{\mathcal{A}\mathcal{A}} = (c_{ij}, i \in \mathcal{A}, j \in \mathcal{A})$。对于向量 $\mathsf{a} = (a_1, \ldots, a_p)$ 和$q \in [1, \infty)$，将 l_q范数定义为 $\|\mathsf{a}\|_q = (\sum_{j=1}^{p} a_j^q)^{\frac{1}{q}}$。令 $\|\mathsf{a}\|_\infty = \max_j |a_j|$为 l_∞范数，且 $\|\mathsf{a}\|_{\min} = \min_j |a_j|$。对于对称正定矩阵 B，将向量范数定义为 $\|\mathsf{a}\|_{\mathbf{B}} = \|\mathsf{B}^{1/2}\mathsf{a}\|_2$。对于序列 a_n 和另一个非负序列$\{b_n\}$，当存在常数 $C > 0$ 使得 $|a_n| \leq C b_n$时，我们记为 $a_n = O(b_n)$；当 $a_n / b_n \to 0$时，我们记为 $a_n = o(b_n)$。当存在常数$C > 0$ 使得 $a_n \leq C b_n$时，我们记为 $a_n \lesssim b_n$；若 $a_n \lesssim b_n$ 且 $b_n \lesssim a_n$，则记为 $a_n \asymp b_n$。对于两个随机序列 Z_n 和 Z_n'，当对于任意 $\epsilon > 0$，存在常数 $M > 0$和 N，使得对所有 $n \geq N$，$\mathrm{P}(|Z_n / Z_n'| > M) < \epsilon$ 时，我们记为 $Z_n = O_p(Z_n')$；对于任意 $\epsilon > 0$，当 $\lim_{n\to\infty} \mathrm{P}(|Z_n / Z_n'| > \epsilon) = 0$时，我们记为 $Z_n = o_p(1)$。

2. 分布式卷积秩回归

本节介绍分布式卷积秩回归（DCRR）框架。我们首先回顾标准卷积秩回归（CRR），然后介绍分布式替代损失以及稀疏高维估计量。

2.1. 模型与卷积秩回归

我们考虑一种数据存储在 M 台机器上的分布式设定。为方便记号，在平衡设计下介绍我们的方法和理论；在这种设计中，每台机器所持有的观测值数量相同，即 n ，因此总样本量满足 $N = nM$；异质本地样本量的讨论请参见备注 2。假设我们观测到独立同分布数据 $\{(y_i^m, \mathbf{x}_i^m) : i = 1, \ldots, n,\ m = 1, \ldots, M\}$，其中 $\{(y_i^m, \mathbf{x}_i^m)\}_{i=1}^n$ 表示存储在第 m 台机器上的子样本。这里，$y_i^m \in \mathbb{R}$ 是响应变量，$\mathbf{x}_i^m \in \mathbb{R}^p$ 是 p 维协变量向量。令 $\mathbf{X} = (\mathbf{X}1, \ldots, \mathbf{X}p) \in \mathbb{R}^{N \times p}$ 为全局设计矩阵，$\mathbf{X}^m = (\mathbf{X}_1^m, \ldots, \mathbf{X}_p^m) \in \mathbb{R}^{n \times p}$ 为第 m 台机器上的设计矩阵，其中 $\mathbf{X}j = (x_{1j}, \ldots, x_{Nj})^{\mathrm{T}}$ 和 $\mathbf{X}_j^m = (x_{1j}^m, \ldots, x_{nj}^m)^{\mathrm{T}}$ 分别包含第 j 个变量 $j = 1, \ldots, p$ 的观测值。响应向量 $\mathbf{y}$ 和 $\mathbf{y}^m$ 也使用相同的记号。

假设数据由以下线性回归模型生成

$$y_i^m = \mathbf{x}_i^m \boldsymbol{\beta}^* + \epsilon_i^m, \quad i = 1, \ldots, n,\ m = 1, \ldots, M, \tag{2.1}$$

其中 $\{\epsilon_i^m : i = 1、\ldots\ldots、n、\ m = 1、\ldots\ldots、M\}$ 为独立同分布的随机误差，$\boldsymbol{\beta}^* \in \mathbb{R}^p$ 为未知参数向量。不失一般性，我们假设 (2.1) 中的误差均值为零，且与 $\mathbf{x}$ 独立。正如备注 2 所说明的，我们的所有结果均可推广到局部样本量 n_m 不相等这一更一般的情形，只需选择任意一台具有足够大 n_m 的机器作为主机器。

令 $(y', \mathbf{x}')$ 为 $(y, \mathbf{x})$ 的一个独立副本。对于秩回归，真实回归系数 $\boldsymbol{\beta}^*$ 可表示为总体秩损失的极小化解

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} E\big\{|y - y' - (\mathbf{x} - \mathbf{x}')^{\mathrm{T}}\boldsymbol{\beta}|\big\}.$$

然而，非光滑绝对损失给计算带来了巨大挑战

和高维分析。为解决这一问题，Zhou 等人（2024）提出了典范卷积秩回归

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j\neq i} L_h\big\{y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\boldsymbol{\beta}\big\}, \tag{2.2}$$

其中 $L_h(u) = \int_{-\infty}^{\infty} |u-v| \frac{1}{h} K\big(\frac{v}{h}\big)\, dv$ 是由 $L(u) = |u|$ 与 $K_h(u) = \frac{1}{h}K(u/h)$ 卷积得到的光滑凸函数。

我们假设核 K 满足以下性质：(i) 对所有 $t \in \mathbb{R}$，$K(-t) = K(t)$；(ii) 存在 $\delta_0 > 0$ 使得 $\kappa_l := \inf_{t\in[-\delta_0,\delta_0]} K(t) > 0$;$\int_{-\infty}^{\infty} K(t)\mathrm{d}t = 1$；(iv) $\kappa_u := \sup_{t\in\mathbb{R}} K(t)$ (iii) $< \infty$；(v) $\kappa_j := \int_{-\infty}^{\infty} |t|^j K(t)\mathrm{d}t < \infty$ 对于 $j = 1,\ 2$；(vi) 存在 $\alpha_0 \in (0,1]$ 与 $L_0 > 0$ 使得 $|K(x) - K(y)| \le L_0|x-y|^{\alpha_0}$对所有 $x, y \in \mathbb{R}$。典型例子包括高斯核 $K(u) = \frac{1}{\sqrt{2\pi}} \exp\{-u^2/2\}$以及 Epanechnikov 核 $K(u) = \frac{3}{4}(1-u^2)\mathrm{I}(-1 \le u \le 1)$。

令 $\mathcal{L}_h(\boldsymbol{\beta}) = E\big\{L_h\big(y - y' - (\mathbf{x}-\mathbf{x}')^{\mathrm{T}}\boldsymbol{\beta}\big)\big\}$，并将其极小化解记为 $\boldsymbol{\beta}_h^*$。Zhou 等人（2024）证明，对于任意 $h > 0$, $\boldsymbol{\beta}_h^* = \boldsymbol{\beta}^*$，因此平滑在总体层面不会引入渐近偏差。

2.2. 分布式*CRR*与替代损失

在分布式系统中，式（2.2）中的完整 CRR损失无法直接获得，因为它涉及各机器上的所有 $N(N-1)$ 对。我们将全局和本地端经验损失定义为

$$\mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j\neq i} L_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\boldsymbol{\beta}\big),$$

$$\mathcal{L}_m(\boldsymbol{\beta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} L_h\big(y_i^m - y_j^m - (\mathbf{x}_i^m - \mathbf{x}_j^m)^{\mathrm{T}}\boldsymbol{\beta}\big), \quad m = 1, \ldots, M.$$

当 N 较大时，在分布式环境中无法计算全局损失 $\mathcal{L}_N$；而每个 $\mathcal{L}_m$ 都可以在第 m 台本地端机器上计算。

对于具有如下形式的可加经验损失的一般 M 估计量

$$\mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_m(\boldsymbol{\beta}).$$

Jordan 等人（2019）提出了通信高效替代似然（CSL）框架。在最简单的形式下，CSL 构造替代损失

$$\widetilde{\mathcal{L}}(\boldsymbol{\beta}) = \mathcal{L}_1(\boldsymbol{\beta}) - \langle \boldsymbol{\beta}, \nabla\mathcal{L}_1(\boldsymbol{\beta}_0) - \nabla\mathcal{L}_N(\boldsymbol{\beta}_0)\rangle\,, \tag{2.3}$$

其中，$\boldsymbol{\beta}_0$ 是 $\boldsymbol{\beta}^*$ 的初始估计量，且将第一台机器用作中心端。这一构造关键依赖于可加性 $\nabla\mathcal{L}_N(\boldsymbol{\beta}) = M^{-1}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{\beta})$，从而可以根据局部梯度重构全局梯度。

然而，在卷积秩回归中，经验损失是基于所有样本对的二阶 U 统计量。全局 CRR 损失 $\mathcal{L}_N$ 通过跨机器样本对将不同机器上的观测值耦合起来，而每个 $\mathcal{L}_m$ 仅涉及机器内样本对。因此，

$$\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{\beta}) \neq \nabla\mathcal{L}_N(\boldsymbol{\beta}),$$

因此，CSL 所依据的可加性假设不再成立。这使得现有 CSL 理论不适用于分布式 CRR，也引出了一个问题：我们能否仍然构造一种通信高效的替代方法，使其达到与集中式 CRR 估计量相同的统计性能?

受（2.2）的形式和（2.3）中的 CSL 思想启发，我们提出分布式

卷积秩回归（DCRR）估计量

$$\widehat{\boldsymbol{\beta}}_d = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \mathcal{L}_1(\boldsymbol{\beta}) - \Big\langle \boldsymbol{\beta}, \nabla\mathcal{L}_1(\boldsymbol{\beta}_0) - \frac{1}{M}\sum_{m=1}^{M} \nabla\mathcal{L}_m(\boldsymbol{\beta}_0) \Big\rangle \Big\}, \tag{2.4}$$

该方法仅在中心机器上使用 CRR损失，并汇总来自本地机器的梯度校正。定义分布式卷积秩损失

$$\mathcal{L}_d(\boldsymbol{\beta};\boldsymbol{\beta}_0) := \mathcal{L}_1(\boldsymbol{\beta}) - \Big\langle \boldsymbol{\beta}, \nabla\mathcal{L}_1(\boldsymbol{\beta}_0) - \frac{1}{M}\sum_{m=1}^{M} \nabla\mathcal{L}_m(\boldsymbol{\beta}_0) \Big\rangle .$$

根据构造， $\mathcal{L}_d(\cdot;\,\boldsymbol{\beta}_0)$ 不等于不可行的全数据损失 $\mathcal{L}_N$，并且由于 U 统计量结构，$\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{\beta})$ 与 $\nabla\mathcal{L}_N(\boldsymbol{\beta})$ 并不一致。尽管如此，我们将证明它们具有相同的总体极小值点，且由此得到的估计量与集中式 CRR 具有相同的极限分布和预言机速率。

令 $\boldsymbol{\beta}_d^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} E\{\mathcal{L}_d(\boldsymbol{\beta};\boldsymbol{\beta}_0)\}$。以下引理表明，我们的替代损失的总体极小值点与真实参数 $\boldsymbol{\beta}^*$ 一致。

引理 1. 对于任意 $h>0$ 和任意 $M\geq 1$，我们有 $\boldsymbol{\beta}_d^* = \boldsymbol{\beta}^*$。

这一结论直接源于总体 CRR 损失满足的等式 $\boldsymbol{\beta}_h^* = \boldsymbol{\beta}^*$，并为我们的替代损失构造提供了总体层面的理论依据。

接下来，我们考察 DCRR估计量的渐近分布。以下引理表明，在初始估计量满足一个温和条件的情况下， $\boldsymbol{\beta}_d$ 与集中式 CRR 估计量渐近等价。

引理2. 对每个 $h>0$，若 $\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|_2 = o_p(M^{-1/2})$，则

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\, \frac{E\big[E\{L_h'(\epsilon-\epsilon')\mid\epsilon\}^2\big]}{E\{L_h''(\epsilon-\epsilon')\}^2}\,\mathrm{cov}(\mathbf{x})^{-1}\right),$$

其中 (ϵ, ϵ') 是误差对的独立副本，且 $\mathrm{cov}(\mathbf{x}) = \Sigma$ 是 $\mathbf{x}$ 的协方差矩阵。

注释 1. 引理*2*表明，在初始条件 $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 = o_p(M^{-1/2})$，下，*DCRR*估计量 $\boldsymbol{\beta}_d$ 与全局卷积估计量 $\widehat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}} \mathcal{L}_N(\boldsymbol{\beta})$具有相同的渐近方差。令 $\widehat{\boldsymbol{\beta}} = \arg\min^{\boldsymbol{\beta}}{}_N{}^1(_{N-1)} \sum_{i=1}^N \sum_{j\neq i} |y_i - y_j - (\mathbf{x}_i - \mathbf{x}j)^T\boldsymbol{\beta}|$ 为非光滑秩估计量。则渐近相对效率满足

$$\mathrm{ARE}(\widehat{\boldsymbol{\beta}}_d, \widehat{\boldsymbol{\beta}}) = \frac{E\big[E\{L_h'(\epsilon - \epsilon') \mid \epsilon\}^2\big]}{12\left\{\int f^2(x)\,dx\right\}^2 E\{L_h''(\epsilon - \epsilon')\}^2},$$

其中， $f(\cdot)$ 是 ϵ的密度。根据 *Zhou* 等人（*2024*）的定理*3*，我们有 $\lim_{h\to 0^+} \mathrm{ARE}(\boldsymbol{\beta}_d, \boldsymbol{\beta}) = 1$，这意味着，当 h 选得足够小时， $\boldsymbol{\beta}_d$ 的估计效率可以任意接近 $\boldsymbol{\beta}$ 的估计效率。结合第 *3* 节的高维结果，这表明 *DCRR* 仅使用本地端 U-统计量和一轮通信，即可达到集中式 *CRR* 的效率。

2.3. 稀疏 *DCRR* 与迭代算法

在高维情形下，维度 p 可能很大，并且可能随 N增长，而我们通常假设 $\boldsymbol{\beta}^*$ 中只有相对较少的一部分坐标对结果有不可忽略的影响。令 $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ 为 $\boldsymbol{\beta}^*$ 的支撑集，$s = |\mathcal{A}| \geq 1$ 为其大小。我们允许 p 和 s 均随 n趋于无穷，并假设 s的阶小于 n。

为纳入稀疏性，我们考虑惩罚DCRR估计量

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \mathcal{L}_d(\boldsymbol{\beta}; \boldsymbol{\beta}_0) + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\}, \tag{2.5}$$

其中 $p_\lambda(\cdot)$ 是一个调节参数为 λ 的惩罚函数，例如 ℓ_1 惩罚项或

折叠凹惩罚。我们现在介绍一种两阶段迭代算法：第一阶段进行 ℓ_1-惩罚迭代，随后第二阶段基于折叠凹惩罚进行细化。这些估计量的理论性质将在第 3 节中介绍。

在第一阶段，我们首先使用带有 l_1惩罚项的 DCRR 进行初始迭代。我们将 ℓ_1 惩罚的 DCRR 迭代算法设计如下：

$$\widehat{\boldsymbol{\beta}}^k := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \mathcal{L}_1(\boldsymbol{\beta}) - \Big\langle \boldsymbol{\beta}, \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}^{k-1}) - \frac{1}{M}\sum_{m=1}^{M} \nabla\mathcal{L}_m(\widehat{\boldsymbol{\beta}}^{k-1}) \Big\rangle + \lambda \sum_{j=1}^{p} |\beta_j| \Big\}, \quad (2.6)$$

对于 $k = 1, \ldots, k_1$，其中 $k_1 \geq 1$ 为最大迭代次数。在第 k 次迭代中，我们将前一次估计 $\widehat{\boldsymbol{\beta}}^{k-1}$ 作为 DCRR 损失函数中的初始值。中心机器将 $\widehat{\boldsymbol{\beta}}^{k-1}$ 广播给所有本地机器，每台本地机器计算梯度 $\nabla\mathcal{L}_m(\widehat{\boldsymbol{\beta}}^{k-1})$，中心端汇总这些梯度以构造替代损失。每次迭代产生的通信成本为 $O(pM)$ 量级，因此，带 ℓ_1 惩罚项的 DCRR 算法具有通信高效性。

受折叠凹惩罚估计量强预言机性质的启发（Fan和Li，2001；Zhang，2010），我们采用基于折叠凹惩罚项的第二阶段细化，以提高估计精度并实现分布式强预言机保证。

对于折叠凹惩罚 DCRR，我们对 $p_\lambda(\cdot)$ 采用局部线性近似（LLA），并考虑如下形式的惩罚项 $p_\lambda(v) = \lambda^2 q(v/\lambda)$，其中 $v \geq 0$，且 $p: [0,\infty) \to [0,\infty)$ 满足：(i) $p(\cdot)$ 在 $[0,\infty)$ 上单调不减且为凹函数，且 $p(0) = 0$；(ii) $p'(\cdot)$ 几乎处处存在，在 $(0,\infty)$ 上单调不增，$0 \leq p'(v) \leq 1$，且 $\lim_{v\downarrow 0} q'(v) = 1$；(iii) 对某个 $\alpha_1 > 0$，$p'(\alpha_1) = 0$。特殊情形包括 SCAD 惩罚（Fan 和 Li，2001）和 MCP 惩罚（Zhang，2010）。

为方便起见，我们将 ℓ_1-惩罚 DCRR 算法的第 k 次迭代记为 $\widehat{\boldsymbol{\beta}}^{(1,k)}$。设第二阶段的初始估计量为 $\widehat{\boldsymbol{\beta}}^{(2,0)} = \widehat{\boldsymbol{\beta}}^{(1,k_1)}$，并设 k_2 为第二阶段的最大迭代次数。利用 LLA 思想，对于 $t \geq 2$ 和 $k = 1$、.....、k_t，我们将第 t 阶段的估计量 $\widehat{\boldsymbol{\beta}}^{(t,k)}$ 定义为

$$\widehat{\boldsymbol{\beta}}^{(t,k)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \mathcal{L}_d(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}^{(t,k-1)}) + \sum_{j=1}^{p} p'_{\lambda_{t,k}}\big(|\widehat{\beta}_j^{(t,0)}|\big)|\beta_j| \Big\}, \tag{2.7}$$

其中 $\widehat{\boldsymbol{\beta}}^{(t,0)} = \widehat{\boldsymbol{\beta}}^{(t-1,k_t)}$。在实践中，我们主要关注 $k_t = 1$ 取$t \geq 2$的情形，这已足以实现第 3 节所述的分布式强预言机性质。相应的简化第二阶段算法总结于算法1中。

算法1 两阶段 DCRR 迭代算法

1：输入：带宽 h、最大迭代次数 k_1、最大阶段数 T、初始值 $\widehat{\boldsymbol{\beta}}^0 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\mathcal{L}_1(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$。 2：对于 $k = 1, \ldots, k_1$，执行 3：中心端：将当前迭代值 $\widehat{\boldsymbol{\beta}}^{(1,k-1)}$ 发送给所有本地机器； 4：本地端：每台机器 m 计算本地梯度 $\nabla\mathcal{L}_m(\widehat{\boldsymbol{\beta}}^{(1,k-1)})$ 并将其发送给中心端； 5：中心端：根据 (2.6) 更新估计量。 6：结束循环 7：对于 $t = 2, \ldots, T$，执行 8：中心端：设置 $\widehat{\boldsymbol{\beta}}^{(t,0)} = \widehat{\boldsymbol{\beta}}^{(t-1)}$； 9：中心端：将当前迭代值 $\widehat{\boldsymbol{\beta}}^{(t,0)}$ 发送给所有本地机器； 10：本地端：每台机器 m 计算本地梯度 $\nabla\mathcal{L}_m(\widehat{\boldsymbol{\beta}}^{(t,0)})$ 并将其发送给中心端； 11：中心端：更新估计量，依据 12：结束循环 13：返回 $\widehat{\boldsymbol{\beta}}^{(T)}$。

$$\widehat{\boldsymbol{\beta}}^{(t)} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\mathcal{L}_1(\boldsymbol{\beta}) - \left\langle \boldsymbol{\beta}, \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}^{(t,0)}) - \frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\widehat{\boldsymbol{\beta}}^{(t,0)})\right\rangle + \sum_{j=1}^{p} p'_\lambda\big(|\beta_j|\big)|\beta_j|\right\}.$$

3. 稀疏 DCRR 的理论性质

本节将建立所提出的稀疏 DCRR 估计量的非渐近误差界和预言机性质。我们首先介绍正则性假设与符号，然后研究两阶段 DCRR 迭代算法的收敛性、分布式预言机性质，以及通过分布式 HBIC 进行调参。

3.1. 假设与符号

我们首先介绍关于误差分布和设计的正则条件。

假设 1 (误差分布**)**。 令 $g(\cdot)$ 为$\zeta_{ij} = \epsilon_i - \epsilon_j$的概率密度函数。我们假设 g 是利普希茨连续的：存在常数 $L_1 > 0$，使得对所有 x、$y \in \mathbb{R}$，$|g(x) - g(y)| \le L_1|x - y|$。这意味着 μ_0：$=$上确界$_{t\in\mathbb{R}}\, g(t) < \infty$。此外，我们假设存在正常数 δ_1、μ_1，使得对所有 $t \in [-\delta_1, \delta_1]$，$g(t) \ge \mu_1$。

假设 *2*（设计）。协变量向量 *x* 的分量有界且均值为零：$\|x\|_\infty \le b_1$ 几乎必然成立，且 $E(x) = 0$。令 *x′* 为 *x* 的一个独立副本，并定义 $z = \Sigma^{-1/2}(x - x')$，其中 $\Sigma = E(xx^T)$。我们假设 *z* 为次高斯，即存在常数 $v_1 \ge 1$，使得对于所有 $z' \in \mathbb{R}^p$和 $t \ge 0$，$P\big(|z^T z'| > v_1\|z'\|_2 t\big) \le 2e^{-t^2/2}$。

假设 3（受限特征值条件）。设 $\Sigma = E(\mathbf{x}\mathbf{x}^T)$。存在常数 $\rho > 0$ 使得 最小值 $_{\mathbf{v}\in\mathbb{R}^p}$ $\|\mathbf{v}\|_\Sigma^2$ $\|\mathbf{v}\|_2^2 \ge \rho$，其中 $\|\mathbf{v}\|_\Sigma^2 = \mathbf{v}^T\Sigma\mathbf{v}$。特别地，$\lambda_{\min}(\Sigma) \ge \rho$ 且 $\lambda_{\max}(\Sigma) < \infty$。

假设 1 是关于误差差值密度的 Lipschitz 条件，略强于 Zhou 等人（2024）中的 H¨older 条件，但能简化我们的论述。假设 2 是随机设计中的标准假设；我们默认使用中心化协变量，因此无需截距。次高斯条件则用于推导高维集中界。假设 3 是一种受限特征值型条件，使我们能够通过$\|\mathbf{v}\|_2 \le \rho^{-1/2}\|\mathbf{v}\|_\Sigma$，利用 Σ-范数控制 ℓ_2 范数。事实上，我们只要求该条件在下文引入的某些锥集上成立。

对于 $d,\ l>0$，我们定义局部椭圆区域和 ℓ_1 锥

$$\Theta(d)=\{\mathbf{v}\in\mathbb{R}^p:\ \|\mathbf{v}\|_\Sigma\le d\},\qquad \Lambda(l)=\{\mathbf{v}\in\mathbb{R}^p:\ \|\mathbf{v}\|_1\le l\|\mathbf{v}\|_\Sigma\}.$$

我们还为折叠凹惩罚 DCRR 估计量引入了收缩因子。

$$\gamma=\widetilde{C}_3\Big(\sqrt{\frac{s+x}{nh}}+\sqrt{\frac{s+x}{Nh}}\Big),\qquad \lambda^*=C_2\sqrt{\frac{\log p}{N}},\tag{3.1}$$

其中 $\widetilde{C}_3=24C_3v_1^2\kappa^{-1}\rho^{-1/2}$ 和 $C_3=258e\kappa_u\mu_0$，且 $x>0$ 是集中界中出现的调节参数；$C_2=4\sqrt{2}(1+c_0^2)^{1/2}b_1$，其中某个 $c_0>0$。我们还假设存在 $\alpha_0>0$，使得

$$\phi=\frac{\big[1+\{p'(\alpha_0)/2\}^2\big]^{1/2}}{\alpha_0\kappa\rho^{1/2}}\in(0,1),$$

并定义

$$l=\Big\{2+\frac{2}{p'(\alpha_0)}\Big\}\Big(\frac{c^2+1}{\rho}\Big)^{1/2}s^{1/2},$$

其中 $c>0$ 满足 $p'(\alpha_0)(c^2+1)^{1/2}/2+1=\alpha_0\kappa\rho^{1/2}c$。这些常数将用于刻画迭代算法的收缩行为。

备注 2（不同的局部样本量）。更一般地，假设第 m 台机器包含 n_m 个观测值，且满足 $N=\sum_{m=1}^M n_m$，并令 $\mathcal{L}_m$ 表示基于 n_m 个样本的相应局部 *CRR* 损失。若选择局部样本量为 $n_\star=n_{m_\star}$ 的任意索引 $m_\star$ 作为主机器，则 *DCRR* 替代损失和算法的定义与之前完全相同，只需将 $\mathcal{L}_1$ 替换为 $\mathcal{L}_{m_\star}$。第 *3* 节中的所有论证均可沿用，只需将定理 *3–1* 的条件中的 n 替换为 $n_\star$。特别是，关键要求是 $n_\star$ 满足与 n 相同的高维尺度，例如 s^2 对数 $p/n_\star=o$（1）以及 $M=o\{N/$（s^2 对数 p）$\}$。由于 最大值$_{1\le m\le M}n_m\ge N/M$，我们总可以选择

局部样本量最大的机器作为主节点，从而确保在全局缩放假设成立时满足这些条件。

3.2. 两阶段 *DCRR* 迭代算法的收敛性

现在我们研究两阶段 DCRR 迭代算法的非渐近误差界。为简单起见，令

$$r_{t,k} = \|\widehat{\boldsymbol{\beta}}^{(t,k)} - \boldsymbol{\beta}^*\|_\Sigma, \qquad r_{t,0} = \|\widehat{\boldsymbol{\beta}}^{(t,0)} - \boldsymbol{\beta}^*\|_\Sigma,$$

并定义指标集 $\mathcal{T}_{t,k} = \mathcal{A} \cup \{j \in [p] : \lambda_j^{(t,k)} < p'(\alpha_0)\lambda_{t,k}\}$，其中 $\lambda^{(t,k)} = (\lambda_1^{(t,k)}, \ldots, \lambda_p^{(t,k)})^\top$ 满足 $\lambda_j^{(t,k)} = p'_{\lambda_{t,k}}(|\widehat{\beta}_j^{(t,0)}|)$。我们假设 $|\mathcal{T}_{t,k}| \leq (c^2+1)s$，这是关于初始估计量的稀疏性条件，且由 ℓ_1-惩罚估计量自然满足。令 $\boldsymbol{\omega}^* = \frac{1}{M}\sum_{m=1}^M \nabla \mathcal{L}_m(\boldsymbol{\beta}^*)$，并将其在 $\mathcal{A}$ 上的限制记为 $\boldsymbol{\omega}^*_{\mathcal{A}}$。

—

引理 3。 假设假设 *1–3* 和条件 $|\mathcal{T}_{t,\,k}| \leq (c^2+1)s$ 成立。令 $h \gtrsim \{s\log p/n + \sqrt{x/n}\}$ 和 $\lambda_{t,\,k} \asymp \gamma_1 r_{t,\,k-1} + \lambda^*$，那么对于任意 $t \geq 2$ 和 $1 \leq k \leq k_t$，第二阶段估计量 $\widehat{\boldsymbol{\beta}}^{(t,k)}$ 满足：$\widehat{\boldsymbol{\beta}}^{(t,k)} \in \Lambda(l)$，$\mathrm{supp}(\widehat{\boldsymbol{\beta}}^{(t,k)}) \lesssim s$；并且

$$r_{t,k} \leq \gamma r_{t,k-1} + \phi r_{t,0} + \kappa^{-1}\rho^{-1/2}\Big\{\big\|\boldsymbol{\omega}^*_{\mathcal{A}}\big\|_2 + \big\|p'_{\lambda_{t,k}}\big((|\boldsymbol{\beta}^*_{\mathcal{A}}| - \alpha_0\lambda_{t,k})_+\big)\big\|_2\Big\}. \tag{3.2}$$

其成立的概率至少为 $1 - 10e^{-x}$。

注3。 由(3.2)，我们得到

$$r_{t,k} \leq \{\gamma^k + \phi(1-\gamma^k)/(1-\gamma)\}r_{t,0} + \frac{1-\gamma^k}{1-\gamma}\,\kappa^{-1}\rho^{-1/2}\Big\{\big\|\boldsymbol{\omega}^*_{\mathcal{A}}\big\|_2 + \big\|p'_{\lambda_{t,k}}\big((|\boldsymbol{\beta}^*_{\mathcal{A}}| - \alpha_0\lambda_{t,k})_+\big)\big\|_2\Big\}.$$

这表明，实现超越 ℓ_1惩罚估计量的进一步收缩，关键在于存在 $k \geq 1$，使得 $\gamma^k + \phi(1-\gamma^k)/(1-\gamma) < 1$，这等价于

$\gamma+\phi<1$。当 $s\log\, p/n=o(1)$时，取 $x=c_0\log\, p$与 $h=O(1)$会使 $\gamma+\phi<1$成为自然选择。因此，在尺度 $s\log\, p/n=o(1)$下，我们的算法对带宽 h是免调参的。(3.2) 中的三个附加项分别反映了初始误差$r_{t,0}$、预言误差 $\|\boldsymbol{\omega}^*_{\mathcal{A}}\|_2$以及由惩罚项引入的偏差。正如补充材料中引理 *B5* 所证明的，$\|\boldsymbol{\omega}^*_{\mathcal{A}}\|_2=O_p\big(\{(s+x)/N\}^{1/2}\big)$。

为简化第二阶段，我们聚焦于 $k_t=1$ for $t\geq 2$ 的情形。在此情形下，令 $\widehat{\boldsymbol{\beta}}^{(t)}$ for $\widehat{\boldsymbol{\beta}}^{(t,1)}$，并选择 $\lambda=\lambda_{t,\ 1}\asymp(\log\ p/N)^{1/2}$。

定理 1。 假设假设 *1–3*、*beta-min* 条件 $\|\boldsymbol{\beta}^*_{\mathcal{A}}\|_{\min}>(\alpha_0+\alpha_1)\lambda$、$x'\lesssim s\log\, p$，以及 $s^2\log\, p/n=o(1)$ 均成立。令 $h=O(1)$ 且 $\lambda\asymp(\log\, p/N)^{1/2}$。则对于第 $(t+1)$ 个估计量 $\widehat{\boldsymbol{\beta}}^{(t+1)}$,

$$\|\widehat{\boldsymbol{\beta}}^{(t+1)}-\boldsymbol{\beta}^*\|_\Sigma\lesssim\sqrt{\frac{s+x'}{N}},$$

以至少 $1-10tp^{-c_0}-e^{-x'}$ 的概率成立，前提是 $t\gtrsim{}^{\log\{\log p+\log(c_0+1)\}\ \log(1/\bar{\gamma})}$，其中 $\bar{\gamma}=\gamma+\phi$。

注 4. 定理 *1* 为最终的两阶段估计量给出了预言机速率。事实上，我们可以证明，经过足够多次 ℓ_1 迭代后，当 $k_1=O_p\{\log(1/(s^{1/2}\lambda^*))\}$时，第一阶段估计量可得到一个具有近预言机速率 $(s\log\, p/N)^{1/2}$ 的近预言估计量，并满足 $\Big\{\widehat{\boldsymbol{\beta}}^{(1,k_1)}-\boldsymbol{\beta}^*\in\Lambda(l),\ \mathrm{supp}(\widehat{\boldsymbol{\beta}}^{(1,k_1)})\lesssim s\Big\}$。证明的更多细节见补充材料。随后，折叠凹第二阶段将误差改善至预言机速率 $(s/N)^{1/2}$，其中对数因子被吸收到 x' 中。根据引理 *1*，平滑不会带来额外偏差。在 *beta-min* 条件下，项 $\big\|p'_{\lambda_{t,k}}\big(|\boldsymbol{\beta}^*_{\mathcal{A}}|-\alpha_0\lambda_{t,k})_+\big)\big\|_2$ 趋于零。

3.3. 分布式预言机性质

众所周知，折叠凹惩罚估计量在集中式设定中具有强预言机性质。然而，在我们的分布式设定中，全样本预言机估计量不可行，因为无法获得全局 CRR 损失 $\mathcal{L}_N$。因此，难以直接要求估计量相对于全样本折叠凹 CRR 估计量具有强预言机性质。为此，我们引入一个基于替代损失 $\mathcal{L}_d$ 定义的分布式预言机估计量，并研究实际估计量与该分布式预言机估计量之间的距离。

对于 $t \geq 2$，我们将第 t 个分布式预言机估计量定义为

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ora},t} = \arg\min_{\boldsymbol{\beta}_{\mathcal{A}^c}=0} \Big\{ \mathcal{L}_1(\boldsymbol{\beta}) - \Big\langle \boldsymbol{\beta}, \sum_{m=1}^{M} \nabla \mathcal{L}_m(\widehat{\boldsymbol{\beta}}^{(t-1)}) - \nabla \mathcal{L}_1(\widehat{\boldsymbol{\beta}}^{(t-1)}) \Big\rangle \frac{1}{M} \Big\}. \tag{3.3}$$

定理 2. 在定理 *1* 的条件下，第$(t+1)$个分布式预言机估计量满足

$$\|\widehat{\boldsymbol{\beta}}^{ora,t+1} - \boldsymbol{\beta}^*\|_\Sigma \lesssim \sqrt{\frac{s+x'}{N}},$$

以至少 $1 - 10tp^{-c_0} - e^{-x'}$ 的概率，前提是 $t \gtrsim \log\{\log p + \log(c_0+1)\} / \log(1/\gamma)$。

定理 2 表明，分布式预言机估计量达到了与基于全体数据的集中式预言机估计量相同的预言机速率$O_p\{(s/N)^{1/2}\}$。在我们的设定中，这一预言机基准并非平凡结果，因为全局 CRR 损失是非加性的，无法分解为局部损失之和。在下一个定理中，我们将证明，我们实际采用的折叠凹 DCRR 估计量以高概率与分布式预言机估计量一致，从而具有分布式强预言机性质。

我们对协方差结构施加一个额外的不可表示型条件。**假设 4**（不可表示条件）。存在一个常数 $A_0 \geq 0$ 使得 $\max_{j \in \mathcal{A}^c} \|\Sigma_{j\mathcal{A}}(\Sigma_{\mathcal{A}\mathcal{A}})^{-1}\|_1 \leq A_0$，其中 $\Sigma_{j\mathcal{A}} = E(\mathbf{x}_j \mathbf{x}_{\mathcal{A}}^T)$ 和 $\Sigma_{\mathcal{A}\mathcal{A}} = E(\mathbf{x}_{\mathcal{A}} \mathbf{x}_{\mathcal{A}}^T)$。

注 5（关于假设 4）。 *Lasso* 的经典不可表示条件通常要求 $\max_{j\in\mathcal{A}^c}\left\|\Sigma_{j\mathcal{A}}(\Sigma_{\mathcal{A}\mathcal{A}})^{-1}\right\|_1 \le 1-\eta$ 对某个 $\eta>0$ 成立，也就是说，该常数严格小于 1。相比之下，假设 *4* 只要求这一数值被某个有限常数 $A_0 \ge 0$ 所界定，而我们的分析并未采用更强的要求 $A_0<1$。

定理 *3.* 假设假设*1–4*以及 *beta-min* 条件 $\|\beta^*_{\mathcal{A}}\|_{\min} \ge (\alpha_0+\alpha_1)\lambda$ 成立， $n \gtrsim max\{N/(s^3 \log p)\ N/\{s(s+\log\ p)\}\}$ 且 $s^2 \log p/n = o(1)$。令 $h = O(1)$ 且 $\lambda \asymp (\log\ p/N)^{1/2}$。则分布式强预言机性质 $\widehat{\beta}^{t+1} = \widehat{\beta}^{ora,t+1}$以至少 $1-12tp^{-c_0}-p^{-1}$ 的概率成立，条件是 $t \gtrsim log\ \left(|\mathcal{A}^{(1)}|^{1/2}\right)$，其中$\mathcal{A}^{(1)} = supp(\widehat{\beta}^{(2,0)}) \cup \mathcal{A}$。

备注 6. 定理 *3* 证明，经过对数数量的折叠凹精化步骤后，第二阶段*DCRR*估计量以高概率与分布式预言机估计量一致。结合定理 *2*，这为高维*DCRR*带来了分布式强预言机性质。这与文献中现有的分布式预言机结果存在本质差异，后者依赖于可加经验损失，其中全局目标函数是局部目标函数之和。在我们的设定中，全局 *CRR* 损失在不同机器之间并不可加，但所提出的 *DCRR* 程序仍然能够达到与集中式预言机估计量相同的预言机速率和模型选择性能。

根据定理*3*和定理*1*的条件，机器数量 M 可以以合理的速度发散： $M = o\left(\frac{N}{s^2 \log p}\right)$ 且 $M \lesssim \max\{s^3 \log p, s(s+\log p)\}$。选择 k_1 和 T，使定理*3*和定理*1*中的收缩条件得到满足，并令 $R = k_1 + T - 1$ 表示梯度通信总轮数。具体而言，当机器数量 M发散时，*DCRR*仍能达到集中式高维*CRR*预言机速率并实现精确支持集恢复，同时主节点与

本地机器之间只需进行*log log log* 轮梯度通信。换言之，只要每台机器拥有足够的观测值，所提出的分布式估计量在估计误差和支撑集恢复方面的表现，就如同我们能够集中访问全部 N 个样本一样，同时只需进行对数级数量的通信轮数。

3.4. 调参：分布式 *HBIC*

在实践中，需要采用数据驱动方法来选择调优参数。Zhou 等人（2024）提出的 HBIC 并不直接适用于分布式系统，因为它需要访问全数据 CRR 损失。因此，我们针对 DCRR 框架提出了分布式高维 BIC（DHBIC）准则。

对于 $t \geq 2$，定义

$$\mathrm{DHBIC}(\lambda) = \log\left\{\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_m(\widehat{\boldsymbol{\beta}}_\lambda^{(t)})\right\} + |\operatorname{supp}(\widehat{\boldsymbol{\beta}}_\lambda^{(t)})|\frac{C_N \log p}{n},$$

其中 $\widehat{\boldsymbol{\beta}}_\lambda^{(t)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\mathcal{L}_d(\boldsymbol{\beta};\, \widehat{\boldsymbol{\beta}}^{(t-1)}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)\right\}$。我们通过 $\widehat{\lambda} = \arg\min_{\lambda\in\mathcal{M}^{(t)}} \mathrm{DHBIC}(\lambda)$ 选择调节参数，其中 $\mathcal{M}^{(t)} = \{\lambda > 0 : \ |\operatorname{supp}(\widehat{\boldsymbol{\beta}}_\lambda^{(t)})| \leq K_N\}$，且 $K_N > s$ 可以发散。

定理 4。 在定理*3*的条件下，假设 $E|\epsilon - \epsilon'| < \infty$，$K_N = o\{(N/\text{对数 } p) \wedge C_N\}$，$C_N s$ 对数 $p/N = o(1)$，并且

$$\Big(\frac{C_N s^{1/2}\log p}{N}\Big)^{1/2} \vee \Big\{\frac{C_N s^{1/2} K_N \log p}{N}\Big\} = o\big(\|\boldsymbol{\beta}^*_{\mathcal{A}}\|_{\min}\big).$$

那么，只要 $t \gtrsim$对数 $\big(|\mathcal{A}^{(1)}|^{1/2}\big)$，*DHBIC* 选择器以至少 $1 - 12tp^{-c_0} - C_N N^{-1/2} - 4p^{-1}$ 的概率满足 $\operatorname{supp}(\widehat{\boldsymbol{\beta}}_{\widehat{\lambda}}^{(t+1)}) = \mathcal{A}$。

注 7. 定理 *4* 保证，*DHBIC* 选择的调节参数能够得到一个折叠凹 *DCRR* 估计量，并一致地恢复真实支撑集 $\mathcal{A}$，从而

保持定理 *3* 的分布式强预言机性质。与集中式设置下的 *HBIC* 类似，*DHBIC* 无需样本拆分或重复重拟合，因此计算复杂度较低。实践中，可以选择 $C_N \asymp \log(\log N)$。

4. 数值研究

本节评估所提出的分布式卷积秩回归（DCRR）方法的有限样本性能。模拟旨在展示：(i) 与集中式 CRR 相比，DCRR 的估计精度和变量选择性能；(ii) 折叠凹精化带来的改进及其与分布式强预言机性质的联系；(iii) 基于秩的方法在重尾误差下的稳健性；以及 (iv) 简单分而治之策略的局限性。随后，我们将这些方法应用于一个真实数据示例。

4.1. 模拟设计

我们考虑线性模型 $y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon$，其中 $\boldsymbol{\beta}^* = (\sqrt{3},\ \sqrt{3},\ \sqrt{3}, \bar{0}, \bar{\ldots},\ 0)^{\mathrm{T}} \in \mathbb{R}^p$，因此真实支撑集大小为 $s = 3$。协变量向量 $\mathbf{x}$ 服从均值为零的多元正态分布 $\mathcal{N}(0, \Sigma)$，其相关结构为自回归相关结构$\Sigma_{ij} = 0.5^{|i-j|}$, $1 \le i,\ j \le p$，记为 AR(0.5)。我们将环境维数设为$p = 1000$。总样本量为 $N = nM$，每台机器的局部样本量为 $n = 100$，并考虑 $M = 5$ 台和 $M = 15$ 台机器，分别对应 $N = 500$ 和 $N = 1500$。

为考察对重尾误差的稳健性，我们从尾部厚度逐渐增加的三种分布中生成误差项 ϵ：(1) 标准正态分布：$\epsilon \sim \mathcal{N}(0, 1)$；(2)$t$自由度为 4、缩放至单位方差的分布：$\epsilon \sim \sqrt{2}\, t(4)$；(3)

标准柯西分布： $\epsilon \sim$ 柯西分布(0,1)。因此，设计逐步从性质良好的高斯噪声过渡到中度重尾和极重尾情形。

在所有模拟中，我们在 CRR 损失中使用 Epanechnikov 核 $K(u) = \frac{3}{4}(1-u^2)$I($-1 \leq u \leq 1$)，并将带宽固定为 $h = 1$，这符合第 3 节中的理论条件。对于每种 (ϵ, M) 组合，我们生成 100 次独立重复试验。

我们比较以下估计量：

类别	估计量描述
集中式 CRR	CRR-LASSO：使用全样本进行 ℓ_1-惩罚化 CRR； CRR-SCAD：使用全样本的 SCAD 惩罚 CRR。
分布式DCRR（所提方法）	DCRR-LASSO：采用 ℓ_1 惩罚的第一阶段估计量； DCRR-SCAD（$T = 2$）：采用 $T = 2$ 次改进的折叠凹DCRR； DCRR-SCAD（$T = 6$）：采用 $T = 6$ 次改进的折叠凹DCRR。
分而治之 CRR（基线方法）	DC-CRR-LASSO：单轮分而治之 CRR-LASSO； DC-CRR-SCAD：采用 SCAD 惩罚项的类似估计量。
Oracle 方法	CRR-ORA：集中式 Oracle CRR 估计量； DCRR-ORA ($T = 2$) / ($T = 6$)：支撑集受限于 $\mathcal{A}$ 的 Oracle DCRR。

为评估估计和变量选择性能，我们计算： ℓ_1 估计-
$\widehat{\ }$ $\widehat{\ }$误差： $E\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1$； ℓ_2 估计误差：
$E\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$；FP：假阳性（所选噪声变量）的平均数量；FN：假阴性（遗漏的真实信号）的平均数量。所有性能指标均基于 100 次重复进行平均，并报告均值，括号内为蒙特卡洛标准误（标准误）。对于基于 CRR 的方法，调优参数通过 HBIC（Zhou 等，2024）（集中式 CRR）或第 3 节提出的 DHBIC（DCRR）进行选择。对于 DC-CRR 方法，在每台本地机器上独立应用 HBIC。

4.2. 模拟结果

表1–3 分别总结了 $\epsilon \sim \mathcal{N}(0,\ 1)$、$\epsilon \sim \sqrt{2}\,t(4)$ 和 $\epsilon \sim$ 柯西分布$(0,\ 1)$ 情形下的结果。在每种情形中，就各项指标而言表现最佳的方法以粗体突出显示。

表1：在 AR(0.5)设计和 $\epsilon \sim \mathcal{N}(0,\ 1)$ 下，对 $M = 5$和$M = 15$ 进行模拟结果比较。我们比较了集中式 CRR、分布式 CRR（DCRR）和分而治之 CRR（DC-CRR），每种方法均采用 ℓ_1 或 SCAD 惩罚项，并纳入 CRR 和 DCRR 的预言机估计量。报告的是 100 次重复的均值（标准误）。

M	方法	ℓ_1	ℓ_2	FP	FN
$M = 5$	CRR-LASSO	0.34(0.01)	0.22(0.01)	0.01(0.01)	0.00(0.00)
	CRR-SCAD	0.14(0.01)	0.09(0.01)	0.00(0.00)	0.00(0.00)
	DCRR-LASSO	0.32(0.01)	0.21(0.01)	0.03(0.02)	0.00(0.00)
	DCRR-SCAD（T=2）	0.15(0.01)	0.10(0.01)	0.00(0.00)	0.00(0.00)
	DCRR-ORA（T=2）	0.15(0.01)	0.10(0.01)	0.00(0.00)	0.00(0.00)
	CRR-ORA	0.13(0.01)	0.08(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-SCAD（T=6）	0.13(0.01)	0.09(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-ORA（T=6）	0.13(0.01)	0.09(0.00)	0.00(0.00)	0.00(0.00)
	DC-CRR-LASSO	0.81(0.01)	0.50(0.01)	0.78(0.08)	0.00(0.00)
	DC-CRR-SCAD	0.33(0.02)	0.21(0.01)	0.82(0.08)	0.00(0.00)
$M = 15$	CRR-LASSO	0.20(0.01)	0.13(0.00)	0.03(0.02)	0.00(0.00)
	CRR-SCAD	0.09(0.00)	0.06(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-LASSO	0.17(0.00)	0.11(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-SCAD (T=2)	0.09(0.00)	0.06(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-ORA（T=2）	0.09(0.00)	0.06(0.00)	0.00(0.00)	0.00(0.00)
	CRR-ORA	0.08(0.00)	0.05(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-SCAD（T=6）	0.08(0.00)	0.05(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-ORA（T=6）	0.08(0.00)	0.05(0.00)	0.00(0.00)	0.00(0.00)
	DC-CRR-LASSO	0.79(0.01)	0.48(0.00)	2.38(0.14)	0.00(0.00)
	DC-CRR-SCAD	0.30(0.01)	0.19(0.01)	2.47(0.17)	0.00(0.00)

DCRR 与集中式 CRR 的比较。在高斯误差下（表1），对于 $M = 5$和 $M = 15$, DCRR-LASSO 估计量在 ℓ_1 和 ℓ_2 误差方面均与 CRR-LASSO 密切吻合，这证实了替代损失在分布式环境中保留了集中式 CRR 的统计效率。在实践中，我们将 DCRR-LASSO 的最大迭代次数设为 $k_1 = 8$。采用 $T = 2$ 和 $T = 6$ 的折叠凹 DCRR-SCAD 估计量进一步降低了估计误差，并且在估计误差和变量选择方面基本达到集中式 CRR-SCAD 的水平（假阳性和假阴性基本为零）。这与定理1中确立的预言机速率一致。

表 2：AR(0.5)设计和 $\epsilon \sim \sqrt{2}\,t(4)$下，$M = 5$ 和$M = 15$的模拟结果。100 次重复的均值（标准误）。

M	方法	ℓ_1	ℓ_2	FP	FN
$M = 5$	CRR-LASSO	0.56(0.02)	0.37(0.01)	0.06(0.02)	0.00(0.00)
	CRR-SCAD	0.25(0.01)	0.16(0.01)	0.00(0.00)	0.00(0.00)
	DCRR-LASSO	0.55(0.01)	0.35(0.01)	0.02(0.01)	0.00(0.00)
	DCRR-SCAD（T=2）	0.26(0.01)	0.17(0.01)	0.00(0.00)	0.00(0.00)
	DCRR-ORA（T=2）	0.25(0.01)	0.17(0.01)	0.00(0.00)	0.00(0.00)
	CRR-ORA	0.23(0.01)	0.15(0.01)	0.00(0.00)	0.00(0.00)
	DCRR-SCAD（T=6）	0.23(0.01)	0.15(0.01)	0.00(0.00)	0.00(0.00)
	DCRR-ORA（T=6）	0.23(0.01)	0.15(0.01)	0.00(0.00)	0.00(0.00)
	DC-CRR-LASSO	1.40(0.02)	0.85(0.01)	0.98(0.10)	0.00(0.00)
	DC-CRR-SCAD	1.00(0.04)	0.64(0.03)	1.51(0.11)	0.00(0.00)
$M = 15$	CRR-LASSO	0.35(0.01)	0.23(0.00)	0.12(0.03)	0.00(0.00)
	CRR-SCAD	0.15(0.01)	0.10(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-LASSO	0.31(0.01)	0.20(0.00)	0.01(0.01)	0.00(0.00)
	DCRR-SCAD（T=2）	0.15(0.01)	0.10(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-ORA（T=2）	0.15(0.01)	0.10(0.00)	0.00(0.00)	0.00(0.00)
	CRR-ORA	0.11(0.01)	0.08(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-SCAD（T=6）	0.12(0.01)	0.08(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-ORA（T=6）	0.12(0.01)	0.08(0.00)	0.00(0.00)	0.00(0.00)
	DC-CRR-LASSO	1.43(0.01)	0.86(0.01)	2.72(0.14)	0.00(0.00)
	DC-CRR-SCAD	0.94(0.02)	0.60(0.01)	3.38(0.18)	0.00(0.00)

表3：在 AR(0.5)设计和 $\epsilon \sim$ 柯西分布(0，1)下的模拟结果，针对 $M = 5$和 $M = 15$。100 次重复试验的均值（标准误）。

M	方法	ℓ_1	ℓ_2	FP	FN
$M = 5$	CRR-LASSO	0.86(0.09)	0.54(0.05)	0.01(0.01)	0.10(0.05)
	CRR-SCAD	0.47(0.10)	0.29(0.06)	0.00(0.00)	0.10(0.05)
	DCRR-LASSO	0.79(0.08)	0.50(0.05)	0.00(0.00)	0.09(0.05)
	DCRR-SCAD（T=2）	0.50(0.09)	0.32(0.05)	0.00(0.00)	0.09(0.05)
	DCRR-ORA（T=2）	0.31(0.02)	0.21(0.01)	0.00(0.00)	0.00(0.00)
	CRR-ORA	0.24(0.01)	0.16(0.01)	0.00(0.00)	0.00(0.00)
	DCRR-SCAD (T=6)	0.44(0.09)	0.28(0.05)	0.00(0.00)	0.09(0.05)
	DCRR-ORA（T=6）	0.26(0.01)	0.17(0.01)	0.00(0.00)	0.00(0.00)
	DC-CRR-LASSO	3.73(0.07)	2.16(0.04)	0.08(0.03)	0.07(0.03)
	DC-CRR-SCAD	2.79(0.08)	1.71(0.04)	0.11(0.03)	0.02(0.01)
$M = 15$	CRR-LASSO	0.48(0.05)	0.31(0.03)	0.01(0.01)	0.02(0.02)
	CRR-SCAD	0.23(0.05)	0.15(0.03)	0.00(0.00)	0.02(0.02)
	DCRR-LASSO	0.40(0.04)	0.26(0.02)	0.00(0.00)	0.01(0.01)
	DCRR-SCAD（T=2）	0.22(0.04)	0.14(0.02)	0.00(0.00)	0.01(0.01)
	DCRR-ORA（T=2）	0.18(0.01)	0.11(0.01)	0.00(0.00)	0.00(0.00)
	CRR-ORA	0.14(0.01)	0.09(0.00)	0.00(0.00)	0.00(0.00)
	DCRR-SCAD（T=6）	0.19(0.04)	0.12(0.03)	0.00(0.00)	0.01(0.01)
	DCRR-ORA（T=6）	0.15(0.01)	0.10(0.00)	0.00(0.00)	0.00(0.00)
	DC-CRR-LASSO	3.80(0.04)	2.20(0.02)	0.26(0.04)	0.00(0.00)
	DC-CRR-SCAD	2.89(0.05)	1.75(0.03)	0.36(0.06)	0.00(0.00)

折叠凹精化的效果与预言机基准。 在表 1–3 中，预言机 CRR 估计量（CRR-ORA）为可达到的误差提供了下界。预言机 DCRR 估计量 DCRR-ORA($T = 2$) 和 DCRR-ORA($T = 6$) 在数值上与 CRR-ORA 无法区分，表明分布式替代损失不会降低预言机性能。对于非预言机估计量，一次折叠凹精化（$T = 2$）就已使 DCRR-SCAD 非常接近 CRR-SCAD，而进一步精化（$T = 6$）则带来了更多但幅度有限的改进。这与理论图景一致：第一阶段达到近预言机速率，第二阶段则进一步向预言机速率提升。

朴素的分而治之 CRR。 在所有设置下，DC-CRR-LASSO 和 DC-CRR-SCAD 估计量的表现都明显逊于集中式 CRR 和 DCRR，尤其是在 M 较大且误差更重尾时。它们的 ℓ_2 误差比 CRR-SCAD 和 DCRR-SCAD 高出数倍，而且即使假阴性为零，仍会出现不可忽略的假阳性。这证实了对局部 CRR 估计量进行朴素平均无法有效校正局部估计偏差，因此其精度受局部样本量 n限制；这与所提出的 DCRR 程序形成鲜明对比，后者在相同的分布式架构下能够达到全样本速率。

总结。 总体而言，模拟结果表明：(i) DCRR-LASSO 和 DCRR-SCAD 在通信高效的分布式方式下运行，同时实现了与其集中式 CRR 对应方法相当的估计和变量选择性能；(ii) 折叠凹精化使第一阶段估计量向预言机基准改进，这与定理3中的分布式强预言机性质一致；(iii) 朴素的分而治之 CRR 可能明显次优；(iv) DCRR 程序对重尾误差具有稳健性。

4.3. 真实数据应用

二手车价格建模提供了一个典型示例，其中会出现重尾响应分布和海量、天然分布式的数据。车辆价格可能相差数个数量级，少量高端或豪华车辆会产生极端离群值；而交易记录通常由不同的经销商、在线平台和保险公司分别收集和存储。这使得二手车定价成为检验稳健、通信高效的分布式回归方法的理想测试平台。

我们在一个来自 Kaggle 的美国二手车真实数据集上展示所提出的方法（`https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset`）。经过数据清理后，我们保留了 $N_{full} = 332,382$ 条运动型多用途汽车（SUV）记录。响应变量为车辆售价，单位为 1,000 美元。我们考虑 27 个基础预测变量，包括马力、行驶里程、车型年份、发动机排量和制造商等连续变量与分类变量。关于这些变量的详细说明，请参见 Pan 等人（2022）的图1。

为构建设计矩阵，我们为所有分类预测变量创建虚拟变量，并纳入所选预测变量之间的二阶交互项。移除高度共线的列以确保数值稳定性后，我们得到 $p = 604$ 个预测变量（不包括截距）。车辆售价呈右偏分布并表现出重尾特征，因此对于最小二乘法而言，这是一个具有挑战性的问题场景，也是检验基于秩的方法的理想测试平台。

为使集中式 CRR 方法在计算上可行，并模拟从大型总体中重复抽样的情形，我们按如下步骤进行。在每次重复中，从全体数据中随机无放回抽取一个大小为 $N = 2000$ 的子样本。前 $N_{\text{train}} = 1000$ 个观测值作为训练集，其余 $N_{\text{test}} = 1000$ 个作为测试集。对于分布式方法，将 N_{train} 个观测值均匀分配到 $M \in \{25, 20, 10, 8, 5\}$ 台机器上。在每次重复中，我们拟合：

1. 分布式方法：DCRR-LASSO、DCRR-SCAD($T = 2$)、DCRR-SCAD(T =6)、DC-CRR-LASSO 和 DC-CRR-SCAD；

2. 集中式方法：CRR-LASSO、CRR-SCAD。

对于每个拟合模型，我们计算测试集上的 ℓ_1 和 ℓ_2 预测误差，并记录模型大小（MS），其定义为非零系数的数量。所有结果均取自100次重复试验的平均值。调节参数的选择方式与模拟研究相同。表4报告了结果。

表4：二手车数据集的真实数据结果。对于每个机器数量 M，我们报告测试集上的平均 ℓ_1 和 ℓ_2 预测误差以及模型大小（MS）。结果为 100 个随机子样本的平均值；括号内为标准误。

M	方法	ℓ_1	ℓ_2	MS
$M = 25$	DCRR-LASSO	4.64(0.04)	7.26(0.13)	11.13(0.36)
	DCRR-SCAD(T=2)	4.28(0.03)	6.61(0.13)	4.92(0.22)
	DCRR-SCAD(T=6)	**4.18(0.03)**	**6.52(0.13)**	**3.30(0.10)**
	DC-CRR-LASSO	7.76(0.03)	10.97(0.11)	50.89(0.96)
	DC-CRR-SCAD	7.75(0.03)	10.97(0.11)	51.20(0.91)
$M = 20$	DCRR-LASSO	4.38(0.04)	6.96(0.14)	8.73(0.27)
	DCRR-SCAD(T=2)	4.19(0.03)	6.52(0.13)	4.59(0.16)
	DCRR-SCAD(T=6)	**4.10(0.03)**	**6.44(0.13)**	**3.76(0.13)**
	DC-CRR-LASSO	7.73(0.03)	10.94(0.11)	40.49(0.56)
	DC-CRR-SCAD	7.71(0.03)	10.92(0.12)	41.77(0.59)
$M = 10$	DCRR-LASSO	4.29(0.02)	6.87(0.13)	7.43(0.18)
	DCRR-SCAD(T=2)	4.06(0.02)	6.44(0.12)	4.71(0.21)
	DCRR-SCAD(T=6)	**4.03(0.02)**	**6.39(0.12)**	**4.20(0.13)**
	DC-CRR-LASSO	7.47(0.03)	10.60(0.12)	5.39(0.20)
	DC-CRR-SCAD	5.81(0.04)	8.61(0.13)	16.20(0.36)
$M = 8$	DCRR-LASSO	4.23(0.02)	6.80(0.13)	7.87(0.16)
	DCRR-SCAD(T=2)	4.01(0.02)	6.37(0.13)	4.82(0.14)
	DCRR-SCAD(T=6)	**3.96(0.02)**	**6.30(0.13)**	**4.82(0.12)**
	DC-CRR-LASSO	6.84(0.05)	9.81(0.13)	5.95(0.20)
	DC-CRR-SCAD	4.64(0.05)	7.26(0.14)	14.58(0.37)
$M = 5$	DCRR-LASSO	4.19(0.02)	6.76(0.12)	8.56(0.24)
	DCRR-SCAD(T=2)	3.96(0.02)	6.33(0.13)	5.46(0.16)
	DCRR-SCAD(T=6)	**3.96(0.02)**	**6.31(0.13)**	**4.79(0.11)**
	DC-CRR-LASSO	5.15(0.04)	7.87(0.14)	9.24(0.12)
	DC-CRR-SCAD	4.16(0.02)	6.59(0.13)	6.82(0.21)
全局	CRR-LASSO	4.18(0.02)	6.75(0.13)	9.54(0.19)
	CRR-SCAD	3.95(0.02)	6.31(0.13)	6.28(0.13)
	零模型	7.78(0.03)	11.00(0.11)	–

真实数据结果突显了几点。首先，在所有 M 取值下，DCRR-SCAD($T = 2$) 和 DCRR-SCAD($T = 6$) 在预测误差方面始终优于 DCRR-LASSO 和分而治之基线方法，且模型规模通常更小或相当。随着 M 从 25 降至 5（即每个本地端的样本量增加），DCRR-SCAD 的预测误差稳步降低，并逐渐接近集中式 CRR-SCAD 的误差，体现了 DCRR 在实践中的可扩展性和统计效率。

其次，对于所有 M，朴素 DC-CRR 估计量的预测误差都明显大于 DCRR，且模型规模要么非常大，要么极不稳定。这再次证实，在这种情况下，简单平均局部 CRR 估计量并不可行。

总体而言，真实数据示例印证了我们的理论和模拟结果：对于分布式环境中的高维秩回归，所提出的 DCRR 框架能够达到接近集中式方法的性能，显著优于朴素的分而治之方法，并在稳健性、预测准确性和模型稀疏性之间实现了颇具实践价值的折中。

5. 讨论

我们提出了一种通信高效的分布式卷积秩回归框架，用于非可加 U统计量损失下的高维稀疏估计。通过构造将单机 CRR 损失与聚合梯度校正相结合的替代损失，我们证明所得的DCRR估计量在分布式环境中运行时，能够达到与集中式 CRR相同的收敛率。在该替代损失的基础上，我们开发了迭代两阶段 DCRR 算法，建立了分布式强预言机性质，并引入了DHBIC 准则以实现一致的模型选择。数值研究同时使用模拟和

真实数据表明，DCRR能够匹配或紧密逼近集中式 CRR 的性能，显著优于朴素的分而治之基线方法，并在重尾误差下保持稳健性。

若干方向值得进一步研究。首先，尽管我们的理论分析与模拟侧重于估计、变量选择和模型选择，但我们的框架自然能够容纳基于部分惩罚DCRR估计量的低维推断。发展一套系统的分布式推断理论，包括具有有限样本校准的置信区间和假设检验，是未来工作中一个有趣的方向。其次，将现有方法推广到更一般的成对损失函数以及其他基于 U统计量的模型（其中经验损失本质上是非可加的）将具有重要价值。最后，探索DCRR的隐私保护或通信受限变体，并研究其在统计效率与资源使用之间的权衡，将进一步提升分布式秩回归在大规模应用中的实际适用性。

参考文献

Battey, H., Fan, J., Liu, H., Lu, J.和Zhu, Z.（2018年）。“具有统计保证的分布式估计与推断。”《统计年鉴》，*46*(3)，1350–1380。

Bousquet，O.（2003年）。“使用熵方法研究次可加函数的集中不等式。”载于《随机不等式及其应用》，第213–247页。施普林格。

Chen, X., Liu, W., Mao, X., 和 Yang, Z.（2020年）。“非加性损失下不可共享数据的分布式高维分位数损失回归。”《机器学习研究杂志》，*21*(182)，1–43。

Chen, X. 和 Xie, M.g.（2014）。“分析超大规模数据的分而治之方法。”《中华统计学志》，第 1655–1684 页。

Fan, J., Guo, Y. 和 Wang, K.（2023）。《通信高效的精确统计估计》。《美国统计学会会刊》，118(542)，1000–1010。

Fan, J. 和 Li, R.（2001年）。《通过非凹惩罚似然进行变量选择及其预言机性质》。《美国统计学会会刊》，96(456)，1348–1360。

Fan, J., Liu, H., Sun, Q. 和 Zhang, T.（2015）。《面向稀疏学习的 I-LAMM：同时控制算法复杂度和统计误差》。《统计年鉴》，46 2，814–841。

Gu, Y., Yang, H., Yang, S. 和 Zou, H.（2025年）。《利用非共享数据对稀疏高维模型进行协同推断》。

Hettmansperger, T.P. 和 McKean, J.W.（2010年）。《稳健非参数统计方法》。美国佛罗里达州博卡拉顿：CRC出版社。

Huang, C. 和 Huo, X.（2019）。“分布式一步估计量。”《数学规划》，174，41–76。

Jaeckel, L.A.（1972）。“通过最小化残差的离散程度估计回归系数。”《数学统计年鉴》，第1449–1458页。

Jordan, M.I.、Lee, J.D. 和 Yang, Y.（2019）。“通信高效分布式统计推断。”《美国统计学会会刊》。

Jurevckov´a, J.（1971年）。“回归系数的非参数估计。”《数学统计年鉴》，第1328–1338页。

Ledoux, M. 和 Talagrand, M.（2013年）。巴拿赫空间中的概率：等周不等式与随机过程。Springer Science & Business Media。

Lee, J.D., Liu, Q., Sun, Y. 和 Taylor, J.E.（2017）。“通信高效稀疏回归：一种单轮方法。”《机器学习研究杂志》，**18(5)**，1–30。

Liu, W., Mao, X. 和 Tu, J.（2025年）。“具有神谕性质和几何收敛性的通信高效分布式稀疏学习。”《美国统计学会会刊》。doi:10.1080/01621459.2025.2479237。即将发表。

Loh, P.L.（2017）。“高维稳健M估计量的统计一致性与渐近正态性。”《统计年刊》，**45(2)**，866–896。doi: 10.1214/16-AOS1461。

Pan, R., Ren, T., Guo, B., Li, F., Li, G. 和 Wang, H.（2022年）。“关于通过先导抽样和单步更新进行分布式分位数回归的说明。”《商业与经济统计杂志》，**40(4)**，1691–1700。

Rosenblatt, J.D. 和 Nadler, B.（2016）。《分布式统计学习中平均的最优性》。《信息与推断：*IMA*杂志》，**5(4)**，379–404。

Shi, C.、Qu, X. 和 Ren, Z.（2020年）。《使用去偏Lasso的通信高效且稳健的分布式统计推断》。《美国统计学会会刊》，**115(529)**，30–45。

Wang, J., Kolar, M., Srebro, N. 和 Zhang, T.（2017）。《具有稀疏性的高效分布式学习》。载于《国际机器学习会议》，第3636–3645页。PMLR。

Wang, L., Peng, B., Bradic, J., Li, R. 和 Wu, Y.（2020）。《一种免调参的稳健高效高维回归方法》。《美国统计学会会刊》，**115(532)**，1700–1714。

Zhang, C.H.（2010年）。“极小极大凹惩罚下近乎无偏的变量选择。”《统计年鉴》，**38(2)**，894–942。

Zhang, Y.、Duchi, J. 和 Wainwright, M.（2013a）。“分治核岭回归。”收录于《学习理论会议》，第 592–617 页。PMLR。

Zhang, Y.、Duchi, J.C. 和 Wainwright, M.J.（2013b）。“统计优化的通信高效算法。”机器学习研究杂志，**14(1)**，3321–3363。

Zhou, L.、Wang, B. 和 Zou, H.（2024年）。“高维稀疏卷积秩回归。”《美国统计学会会刊》，**119(546)**，1500–1512。