Statistica Sinica **35** (2025), 1-20
doi:https://doi.org/10.5705/ss.202022.0375

SPARSE AND DEBIASED ADAPTIVE HUBER REGRESSION IN DISTRIBUTED DATA: AGGREGATED AND COMMUNICATION-EFFICIENT APPROACHES

Wei Ma[1], Junzhuo Gao[1], Lei Wang*[1] and Heng Lian[2]

[1]*Nankai University and* [2]*City University of Hong Kong*

Abstract: Distributed estimation and statistical inference for linear models have drawn much attention recently, but few studies focus on robust learning in the presence of heavy-tailed/asymmetric errors and high-dimensional covariates. Based on adaptive Huber regression to achieve the bias-robustness tradeoff, two classes of sparse and debiased lasso estimators are proposed using aggregated and communication-efficient approaches. To be specific, an aggregated ℓ_1-penalized and a multi-round ℓ_1-penalized communication-efficient adaptive Huber estimators are respectively proposed in the first stage to handle the distributed data with high-dimensional covariates and heavy-tailed/asymmetric errors. To correct the biases caused by the lasso penalty, a unified debiasing framework based on the decorrelated score equations is considered in the second stage. In the third stage, hard-thresholding is used to produce the sparse and debiased lasso estimators. The convergence rates and asymptotic properties of the proposed two estimators are established. The finite-sample performance is studied through simulations and a real data application to Communities and Crime Data Set is also presented to illustrate the validity and feasibility of the proposed estimators.

Key words and phrases: Asymptotic normality, convergence rates, debiased lasso, decorrelated score, multi-round, thresholding.

1. Introduction

With the advancement of science and technology, massive data with large sample size and high-dimensional covariates are stored independently in many different sites, and referred to as *distributed data*. Due to the limitation of storage, computing capability and personal privacy in practice, traditional methods by processing all data simultaneously in one central site are not practical for distributed data. To overcome this problem, distributed estimation and statistical inference have drawn much attention in modern statistical learning recently. The aggregated/divide-and-conquer (Chen and Xie, 2014; Battey et al., 2018; Volgushev, Chao and Cheng, 2019) and communication-efficient surrogate likelihood (CSL) (Wang et al., 2017; Jordan, Lee and Yang, 2019) are the two well-known methods for dealing with distributed data. The aggregated method

*Corresponding author. E-mail: lwangstat@nankai.edu.cn