Statistica Sinica **36** (2026), 1-20
doi:https://doi.org/10.5705/ss.202023.0274

NEARLY OPTIMAL TWO-STEP POISSON SAMPLING AND EMPIRICAL LIKELIHOOD WEIGHTING ESTIMATION FOR M-ESTIMATION WITH BIG DATA

Yan Fan, Yang Liu, Yukun Liu* and Jing Qin

Shanghai University of International Business and Economics, Soochow University, East China Normal University and National Institutes of Health

Abstract: Subsampling techniques can effectively reduce the computational costs of processing big data. Practical subsampling plans typically involve initial uniform sampling and refined sampling. Subsample-based big data inferences are generally built on the inverse probability weighting (IPW), which may be unstable and cannot incorporate auxiliary information. In this paper, we consider a two-step Poisson sampling, which combines an initial uniform sampling with a second Poisson sampling. Under this sampling plan, we propose an empirical likelihood weighting (ELW) estimation approach to an M-estimation parameter, and then construct a nearly optimal two-step Poisson sampling plan based on the ELW method to improve estimation efficiency of IPW-based optimal subsamplings. Further, we derive methods for determining the smallest sample sizes with which the proposed sampling-and-estimation method produces estimators of guaranteed precision. Our ELW method overcomes the instability of IPW by circumventing the use of inverse probabilities, and utilizes auxiliary information including the size and certain sample moments of big data. We show that the proposed ELW method produces more efficient estimators than IPW, leading to more efficient optimal sampling plans and more economical sample sizes for a prespecified estimation precision. These advantages are confirmed through real data based simulations.

Key words and phrases: Big data, empirical likelihood, two-step Poisson sampling.

1. Introduction

One of the most significant features of big data is its incredibly large volume, which poses serious challenges to its timely processing. Data analytics need to be performed efficiently so that the results are made available to users in a cost-effective and timely manner. A popular and efficient strategy for solving this problem is to draw small-scale subsamples from the big data (original sample) and make statistical inferences based on the subsamples (Drineas, Mahoney and Muthukrishnan, 2006; Drineas et al., 2011). Compared with the original big data, the subsamples are usually much smaller, and so subsample-based inferences

*Corresponding author. E-mail: ykliu@sfs.ecnu.edu.cn