Statistica Sinica **35** (2025), 1-23
doi:https://doi.org/10.5705/ss.202022.00277

OPTIMAL SUBSAMPLING FOR MULTINOMIAL LOGISTIC MODELS WITH BIG DATA

Zhiqiang Ye[1], Jun Yu[2] and Mingyao Ai*[1]

[1]*Peking University and* [2]*Beijing Institute of Technology*

Abstract: To model categorical responses, multinomial logistic regressions with different links and parameter restrictions have widely been adopted based on the relationships among different categories. In this paper, a unified Poisson subsampling method is proposed to approximate efficiently the maximum likelihood estimator for regression parameters when big data are encountered. The asymptotic normality of the estimator generated from the Poisson subsample is established. Based on the derived asymptotic variance, optimal subsampling probabilities are given according to the A-optimality criterion. To mitigate the burden on the calculation of optimal subsampling probabilities, a random projection based procedure is applied. For practical implementation, some robustness issues including model misspecification and full data with possible outliers are further discussed with theoretical backups. The advantages of the proposed methods are illustrated through numerical studies on both simulated and real datasets.

Key words and phrases: Categorical data, Johnson–Lindenstrauss transform, Poisson subsamplin, randomized hadamard transform.

1. Introduction

Extremely large datasets are ubiquitous due to the rapid development of science and technologies. Volume is one of the key concepts associated with big data. Specifically, the quantity of generated and stored data in a big data era is usually larger than terabytes and petabytes. Therefore, it is a common challenge on extracting useful information from massive datasets with limited computing resources.

Many statistical methods, which focus on drawing an inference based on a big data set with a fixed computational budget, have been developed up to now. Subsampling is one of the most popular techniques of achieving a good balance between computational complexity and statistical efficiency. Extensive researches show the great success of subsampling in dealing with massive data in various fields. For example, uniform subsampling was used in Drineas et al. (2011) to approximate ordinary least square estimators in linear regressions. To further improve statistical accuracy, some non-uniform subsampling strategies such as leverage score subsampling (Ma, W.Mahoney and Yu, 2015; Ma et al., 2020),

*Corresponding author. E-mail: myai@math.pku.edu.cn