Statistica Sinica **34** (2024), 1-21
doi:https://doi.org/10.5705/ss.202022.0028

SCALABLE ESTIMATION FOR HIGH VELOCITY SURVIVAL DATA ABLE TO ACCOMMODATE ADDITION OF COVARIATES

Ying Sheng[1], Yifei Sun[2], Charles E. McCulloch[3] and Chiung-Yu Huang*[3]

[1]*Chinese Academy of Sciences,* [2]*Columbia University and* [3]*University of California at San Francisco*

Abstract: With the rapidly increasing availability of large-scale streaming data, there is growing interest in methods that process data in batches without requiring storage of the full data set. In this paper, we propose a hybrid likelihood approach for scalable estimation of the Cox model using individual-level data in the current data batch and summary statistics calculated from historical data. We show that the proposed scalable estimator is asymptotically as efficient as the maximum likelihood estimator calculated using the full data set with low data storage requirements and low loading and computation time. A difficulty with analyzing batches of survival data that is not accommodated in extant methods is that new covariates may become available midway through data collection. To accommodate addition of covariates, we develop a hybrid empirical likelihood approach that incorporates the historical covariate effects evaluated using a reduced Cox model. The extended scalable estimator is asymptotically more efficient than the maximum likelihood estimator obtained using only the data batches that include the additional covariates. The proposed approaches are evaluated using numerical simulations and illustrated with an analysis of Surveillance, Epidemiology, and End Results breast cancer data.

Key words and phrases: Batch processing, hybrid empirical likelihood, scalable estimation.

1. Introduction

In recent years, unprecedented technological advances in data collection systems such as medical devices, health apps, surveillance systems, and wearable sensors have led to a proliferation of large-scale streaming data. The key characteristics of such data include a massive sample size and high velocity, posing challenges in terms of data storage and statistical analysis. For example, continuous glucose monitors that report blood sugar levels as frequently as once per minute are becoming increasingly common and readily available (Vettoretti et al. (2018)). The huge amount of streaming glucose data can provide valuable insight into how well diabetic patients are managing their disease and, in the

*Corresponding author.