Statistica Sinica **36** (2026), 1-22
doi:https://doi.org/10.5705/ss.202023.0309

NEW FEATURE SCREENING METHODS FOR MASSIVE INTERVAL-CENSORED FAILURE TIME DATA

Huiqiong Li[1], Zhimiao Cao[1], Jianguo Sun*[2] and Niansheng Tang[1]

[1] *Yunnan University and* [2] *University of Missouri-Columbia*

Abstract: Screening important features has become one of the important tasks in statistical analysis and correspondingly, various screening procedures have been proposed for various types of studies or data including both complete and incomplete data. However, these methods would be computationally costly or even infeasible when one faces massive health databases with both high dimensionality and huge sample size, which have become increasingly popular for comparative effectiveness and safety studies of medical products. In this paper, we consider such a type of incomplete data, interval-censored failure time data, that have not be discussed before and propose two procedures with the use of distance correlation and orthogonal sampling as well as the the jackknife debiased average technique. The proposed approaches can be easily implemented and their sure screening and rank consistency properties are established. Simulation studies demonstrate that the proposed methods work well for practical situations and they are applied to the SEER breast cancer data.

Key words and phrases: Distance correlation, jackknife debiased average, orthogonal subsampling, rank consistency, sure screening.

1. Introduction

This paper considers the feature screening for massive interval-censored failure time data. By being massive, we assume that the data have both large numbers of features (p) and huge sample sizes (N), while by interval censoring, we mean that the failure time of interest is only observed to belong to an interval instead of being observed exactly (Sun, 2006). It is easy to see that such data naturally occur in many studies such as epidemiological or medical follow-up studies, in particular clinical trials. Two specific examples of them are given by the medicare data in Wang et al. (2021b) and the LEGEND-HTN data in Yang et al. (2024). For the problem, two methods will be developed.

Screening important features has become one of the important tasks in statistical analysis and correspondingly, various model-based or model-free screening procedures have been proposed for various types of studies or data including both complete and incomplete data. Among them, one important contribution was given by Fan and Lv (2008), who proposed a sure independence screening (SIS)

*Corresponding author. E-mail: sunj@missouri.edu