Annals of the Institute of Statistical Mathematics (2025) 77:155–190
https://doi.org/10.1007/s10463-024-00912-x

Model free feature screening for large scale and ultrahigh dimensional survival data

Yingli Pan[1,2] · **Haoyu Wang**[3] · **Zhan Liu**[1]

Received: 19 August 2023 / Revised: 20 July 2024 / Accepted: 5 August 2024 /
Published online: 19 October 2024

Abstract

This paper provides a novel perspective on feature screening in the analysis of high-dimensional right-censored large-p-large-N survival data. The research introduces a distributed feature screening method known as Aggregated Distance Correlation Screening (ADCS). The proposed screening framework involves expressing the distance correlation measure as a function of multiple component parameters, each of which can be estimated in a distributed manner using a natural U-statistic from data segments. By aggregating the component estimates, a final correlation estimate is obtained, facilitating feature screening. Importantly, this approach does not necessitate any specific model specification for responses or predictors and is effective with heavy-tailed data. The study establishes the consistency of the proposed aggregated correlation estimator $\widetilde{\omega}_j$ under mild conditions and demonstrates the sure screening property of the ADCS. Empirical results from both simulated and real datasets confirm the efficacy and practicality of the ADCS approach proposed in this paper.

Keywords Distributed feature screening · Large-p-large-N survival data · Aggregated distance correlation · Sure screening property

✉ Zhan Liu
eleen_20040109@163.com

Yingli Pan
849123147@qq.com

Haoyu Wang
371693579@qq.com

[1] Hubei Key Laboratory of Applied Mathematics, Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China

[2] Guangxi Key Laboratory of Machine Vision and Intelligent, Wuzhou 543002, China

[3] School of Economics and Management, Tongji University, Shanghai 200092, China