Statistica Sinica **33** (2023), 1-23
doi:https://doi.org/10.5705/ss.202020.0421

FEATURE SCREENING VIA DISTANCE CORRELATION FOR ULTRAHIGH DIMENSIONAL DATA WITH RESPONSES MISSING AT RANDOM

Linli Xia^{1,2} and Niansheng Tang^{1}

^{1}*Yunnan University and* ^{2}*Tongren University*

Abstract: This study examines the feature screening problem for ultrahigh-dimensional data with responses missing at random. A two-step procedure is proposed to screen important features. The first step screens the significant covariates associated with the missing indicators via the fused mean-variance filter. The second step screens the important predictors associated with the response by fusing the distance correlation and a nonparametric imputation technique. The proposed feature screening procedure has the following merits: (i) it is model free, because it does not depend on a special model structure or distribution assumption; (ii) it avoids resampling on the conditional function of the missing value because a kernel smoothing technique is adopted to implement the nonparametric conditional mean imputation; (iii) it is not sensitive to a misspecification of the propensity score function because it does not impose a special model on the respondent probability. Under some regularity conditions, the sure screening property is shown. A modified maximum ratio criterion is proposed to select the tuning parameter. Simulation studies are conducted to investigate the finite-sample performance of the proposed feature screening procedure. Finally, an example is used to illustrate the proposed methodologies.

Key words and phrases: Distance correlation, missing at random, nonparametric imputation, sure screening property, ultrahigh dimensional data.

1. Introduction

Ultrahigh-dimensional data are often encountered in fields of modern scientific research such as signal processing, biomedical imaging and functional magnetic resonance imaging, and finance. Here, the number of candidate predictors p may increase at an exponential rate of the sample size n, while only a small number of predictors contribute to the response when there is sparsity among the candidate predictors. Under the “larger p smaller n” data framework, various penalized variable selection procedures have been developed to reduce the dimensionality to a number below the sample size by effectively distinguishing

Corresponding author: Niansheng Tang, Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kun-ming 650500, China. E-mail: nstang@ynu.edu.cn.