Statistica Sinica **36** (2026), 1-21
doi:https://doi.org/10.5705/ss.202023.0070

FREQUENT-VOTING INDEPENDENCE SCREENING FOR DATA OF DIFFERENT TYPES OR DIFFERENT DIMENSIONS

Haeun Moon* and Kehui Chen

Carnegie Mellon University and University of Pittsburgh

Abstract: Modern datasets often include different types of variables with complex features, making variable selection particularly challenging. For example, a measure of dependence with the response variable may not be directly comparable among predictor variables of different types and different dimensions. To address this challenge, this work proposes a frequent-voting based independent screening method for variable selection, which avoids a direct comparison of the dependence measure among different variables. Asymptotic analyses show that the proposed method selects all of the active variables with probability converging to one. We also demonstrate its great finite sample performance through numerical experiments and the application to an ADHD study.

Key words and phrases: Model-free, sure screening, test of independence, variable selection.

1. Introduction

This work is motivated by collaborative projects with psychiatrists where the goal of the study is to find risk factors that are predictive of mental disorders such as Attention Deficit Hyperactivity Disorder (ADHD), major depression, and suicidal behaviors. This kind of study usually has a large pool of candidate risk factors, consisting of genetic, brain imaging, clinical and demographic variables. Therefore, variable selection plays an important role in these studies. The imaging data such as fMRI data and EEG data are time course data observed at many brain regions. Our data starts with aggregated time course data in each brain region (an average over all voxels in the region). The time course data are then transformed to a multivariate vector using frequency analysis and basis expansion. At this point, we select the multivariate vector as a whole. For the gene data analysis, researchers may be interested in selecting relevant gene pathways consisting a group of genes, where the variables (gene pathways) under consideration are multivariate variables. Meanwhile, many clinical data and demographic data are also available in the form of a continuous variable or a categorical variable. In addition, the response variable could be multivariate as well. For example, we may follow subjects for a few time points and produce

*Corresponding author. E-mail: haeunmoon@snu.ac.kr