Statistica Sinica **32** (2022), 2095-2117
doi:https://doi.org/10.5705/ss.202020.0427

PRIOR KNOWLEDGE GUIDED ULTRA-HIGH DIMENSIONAL VARIABLE SCREENING WITH APPLICATION TO NEUROIMAGING DATA

Jie He and Jian Kang

University of Michigan

Abstract: Variable screening is a powerful and efficient tool for dimension reduction under ultrahigh-dimensional settings. However, most existing methods overlook useful prior knowledge in specific applications. In this work, from a Bayesian modeling perspective, we develop a unified variable screening procedure for linear regression models. We discuss different constructions of posterior mean screening (PMS) statistics to incorporate different types of prior knowledge according to specific applications. With non-informative prior specifications, PMS is equivalent to the high-dimensional ordinary least-square projection (HOLP). We establish the screening consistency property for PMS with different types of prior knowledge. We show that PMS is robust to prior misspecifications. Furthermore, when the prior knowledge provides correct information on the true parameter settings, PMS can substantially improve the selection accuracy over that of the HOLP and other existing methods. We illustrate our method using extensive simulation studies and an analysis of neuroimaging data.

Key words and phrases: Linear regression, posterior mean screening, prior knowledge, screening consistency.

1. Introduction

Modern technologies have produced a vast amount of high-throughput data, in which the number of variables far outweighs the sample size. This has motivated the development of feature learning and screening methods: a powerful and efficient tool for dimension reduction (Fan and Fan (2008); Fan and Song (2010); Bühlmann and van de Geer (2011); Zhao and Li (2012)) in regression.

The pioneering work on variable screening was that on sure independence screening (SIS) (Fan and Lv (2008)), which has been extended to generalized linear models (Fan and Fan (2008); Fan, Samworth and Wu (2009); Fan and Song (2010)), generalized additive models (Fan, Feng and Song (2011)), quantile regression (He, Wang and Hong (2013); Ma, Li and Tsai (2017)), and the propor-

Corresponding author: Jian Kang, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA. E-mail: jiankang@umich.edu.