CONSISTENT SCREENING PROCEDURES
IN HIGH-DIMENSIONAL BINARY CLASSIFICATION

Hangjin Jiang\textsuperscript{1,2}, Xingqiu Zhao\textsuperscript{3}, Ronald C.W. Ma\textsuperscript{2}, Xiaodan Fan\textsuperscript{2}

\textsuperscript{1}Zhejiang University, \textsuperscript{2}The Chinese University of Hong Kong and \textsuperscript{3}The Hong Kong Polytechnic University

Abstract: We consider variable screening in high-dimensional binary classification. First, we propose nonparametric test statistics for the problem of the two-sample distribution comparison. These test statistics combine the merits of the chi-squared and Kolmogorov–Smirnov statistics, and provide new insights into the equality test of the unspecified distributions underlying the two independent samples. Based on our new statistics, we propose a marginal screening procedure and a pairwise joint screening procedure for detecting important variables in high-dimensional binary classification. Both screening procedures have the consistent screening property, which is stronger than the sure screening property of most existing methods. The marginal screening procedure is much more powerful than other methods over a broad range of cases, and the pairwise joint screening procedure provides a way of detecting variables with a joint effect, but no marginal effect. Extensive simulations and a real-data application show the effectiveness and advantages of the proposed methods.

Key words and phrases: Binary classification, consistency, non-parametric test, Two-sample distribution comparison, variable screening.

1. Introduction

Variable screening aims to screen important variables out of thousands of candidates, and is a fundamental statistical problem in many applied areas. For example, in case-control disease studies, researchers want to find important disease factors out of numerous environmental, clinical, epigenetic, or gene expression variables. For continuous responses, many variable screening methods have been proposed; see, for example, Fan and Li (2001), Fan and Lv (2008), Fan, Feng and Song (2012), Hall and Miller (2012), Huang and Zhu (2016), Li, Zhong and Zhu (2012b), Li et al. (2012a), and the references therein. Fewer methods have been proposed for the binary response case, and include the marginal t-test screening (Fan and Fan (2008)), maximum marginal likelihood screening

Corresponding author: Xiaodan Fan, Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. E-mail: xfan@cuhk.edu.hk