Statistica Sinica **32** (2022), 1881-1909
doi:https://doi.org/10.5705/ss.202020.0495

SEMI-STANDARD PARTIAL COVARIANCE VARIABLE SELECTION WHEN IRREPRESENTABLE CONDITIONS FAIL

Fei Xue and Annie Qu

Purdue Univeristy and University of California Irvine

Abstract: Traditional variable selection methods could fail to be sign consistent when irrepresentable conditions are violated. This is especially critical in high-dimensional settings when the number of predictors exceeds the sample size. In this paper, we propose a new semi-standard partial covariance (SPAC) approach that is capable of reducing the correlation effects from other covariates, while fully capturing the magnitude of the coefficients. The proposed SPAC is effective in choosing covariates that have direct effects on the response variable, while eliminating predictors that are not directly associated with the response, but are highly correlated with the relevant predictors. We show that the proposed SPAC method with the Lasso penalty or the smoothly clipped absolute deviation (SCAD) penalty possesses strong sign consistency in high-dimensional settings. Numerical studies and a post-traumatic stress disorder data application confirm that the proposed method outperforms the existing Lasso, adaptive Lasso, SCAD, Peter–Clark-simple algorithm, and factor-adjusted regularized model selection methods when the irrepresentable conditions fail.

Key words and phrases: Irrepresentable condition, Lasso, model selection consistency, partial correlation, smoothly clipped absolute deviation.

1. Introduction

Variable selection is an important model-building tool for selecting covariates relevant to the response variable, which is fundamental for the construction of a sparse model when the number of relevant covariates is much smaller than the total number of observed covariates. This is especially crucial under high dimensionality, where the number of covariates far exceeds the number of observations. For high-dimensional data, traditional regularization variable selection methods (Tibshirani (1996); Fan and Li (2001); Zou and Hastie (2005); Yuan and Lin (2006); Zou (2006); Candes and Tao (2007); Zhang (2010)) are effective in achieving model selection and parameter estimation simultaneously

Corresponding author: Annie Qu, Department of Statistics, University of California Irvine, Irvine, CA 92697, USA. E-mail: aqu2@uci.edu.