Statistica Sinica **34** (2024), 1-21
doi:https://doi.org/10.5705/ss.202022.0127

RESPONSE VARIABLE SELECTION IN MULTIVARIATE LINEAR REGRESSION

Kshitij Khare and Zhihua Su*

University of Florida

Abstract: In this article, we discuss response variable selection and the subsequent estimation of the regression coefficients in multivariate linear regression. Because of the asymmetric roles of the predictors and responses in a regression, response variable selection differs markedly from the usual predictor variable selection. When a response is inferred to have a coefficient of zero, it should not simply be removed from subsequent estimation. Instead, we should analyze its relationship with the responses that have nonzero coefficients, which we call dynamic responses. If it is correlated with the dynamic responses, given all other responses, it should be retained to improve the estimation efficiency of the nonzero coefficients, as an ancillary statistic. Otherwise, it can be removed from further inference (leading to significant resource savings in high-dimensional settings), and we call it a static response. Therefore, we can classify responses into three categories: dynamic responses, ancillary responses, and static responses. We derive an algorithm to identify these response variables, and provide an estimator of the regression coefficients based on the selection result. Applications using synthetic and real data illustrate the efficacy of the proposed response variable selection procedure in both low- and high-dimensional settings. Lastly, we establish the consistency of the variable selection procedures and the asymptotic properties of the estimators for both the large-sample setting and the high-dimensional small-sample setting.

Key words and phrases: Group sparsity, high-dimensional data, oracle property, response variable selection.

1. Introduction

Consider the standard multivariate linear regression

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{Y} \in \mathbb{R}^r$ is the multivariate response vector, $\mathbf{X} \in \mathbb{R}^p$ contains the predictors, with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$, and the error vector $\boldsymbol{\varepsilon}$ has mean $\mathbf{0}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. The errors and the predictors are independent of each other. We use n to denote the sample size. Furthermore, we assume that $n > p$, because our primary focus is response variable selection. If $n < p$, we can use any predictor variable selection method

*Corresponding author.