Statistica Sinica **36** (2026), 1-20
doi:https://doi.org/10.5705/ss.202022.0391

PENALIZED REGRESSION WITH MULTIPLE LOSS FUNCTIONS AND VARIABLE SELECTION BY VOTING

Guorong Dai*[1], Ursula U. Müller[2] and Raymond J. Carroll[2]

[1]*Fudan University and* [2]*Texas A&M University*

Abstract: We consider a sparse linear model with a fixed design matrix in a high dimensional scenario. We introduce a new variable selection procedure called "voting", which combines the results from multiple regression models with different penalized loss functions to select the relevant predictors. A predictor is included in the final model if it receives enough votes, i.e. is selected by most of the individual models. By employing multiple different loss functions our method takes various properties of the error distribution into account. This is in contrast to the standard penalized regression approach, which typically relies on just one criterion. When that single criterion is not met the standard approach is likely to fail, whereas our method is still able to identify the underlying sparse model. Working with the voting procedure reduces the number of predictors that are incorrectly selected, which simplifies the structure and improves the interpretability of the fitted model. We prove model selection consistency and illustrate the advantages of our method numerically using simulated and real data sets.

Key words and phrases: High dimensional data, linear model, model selection consistency, sparse estimators.

1. Introduction

In the past few decades variable selection has attracted much attention in the statistical community because of its usefulness in analyzing modern data sets in which the number of features can be very high, often greatly exceeding the number of observations. Variable selection is used to determine the covariates that have an effect on an outcome and that should be included in the model as relevant predictors. This can substantially improve the simplicity and interpretability of a fitted statistical model. We refer interested readers to Desboulets (2018) for a detailed overview of this topic in a variety of regression settings. In this article we consider the variable selection problem in a high dimensional linear model with a deterministic design matrix, where the number of predictors can exceed the sample size. We further assume that the model is sparse, i.e. only a fraction of the predictors significantly affects the response. Our goal is to identify this fraction of important predictors and to exclude those with no influence.

*Corresponding author. E-mail: guorongdai@fudan.edu.cn