Statistica Sinica **33** (2023), 1-28
doi:https://doi.org/10.5705/ss.202020.0449

WEAK SIGNAL IDENTIFICATION AND INFERENCE IN PENALIZED LIKELIHOOD MODELS FOR CATEGORICAL RESPONSES

Yuexia Zhang, Peibei Shi, Zhongyi Zhu, Linbo Wang and Annie Qu

The University of Texas at San Antonio, Meta, Fudan University, University of Toronto and University of California, Irvine

Abstract: Penalized likelihood models are widely used to simultaneously select variables and estimate model parameters. However, the existence of weak signals can lead to inaccurate variable selection, biased parameter estimation, and invalid inference. Thus, identifying weak signals accurately and making valid inferences are crucial in penalized likelihood models. We develop a unified approach to identify weak signals and make inferences in penalized likelihood models, including the special case when the responses are categorical. To identify weak signals, we use the estimated selection probability of each covariate as a measure of the signal strength and formulate a signal identification criterion. To construct confidence intervals, we propose a two-step inference procedure. Extensive simulation studies show that the proposed procedure outperforms several existing methods. We illustrate the proposed method by applying it to the Practice Fusion diabetes data set.

Key words and phrases: Adaptive Lasso, de-biased method, model selection, post-selection inference.

1. Introduction

In the big data era, massive data are collected with large-dimensional covariates. However, only some of the covariates might be important. To select the important variables and estimate their effects on the response variable, various penalized likelihood models have been proposed, such as the penalized least squares regression model (Tibshirani (1996); Zou and Hastie (2005); Tibshirani et al. (2005); Yuan and Lin (2006); Zou (2006); Zhang (2010)), penalized logistic regression model (Park and Hastie (2008); Zhu and Hastie (2004); Wu et al. (2009)), and penalized Poisson regression model (Lambert and Eilers (2005); Jia, Xie and Xu (2019)).

To achieve model selection consistency or the variable screening property for a high-dimensional problem, a common condition is the "beta-min" condition,

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