Statistica Sinica **35** (2025), 1-25
doi:https://doi.org/10.5705/ss.202022.0321

OPTIMAL MODEL AVERAGING FOR SINGLE-INDEX MODELS WITH DIVERGENT DIMENSIONS

Jiahui Zou, Wendun Wang, Xinyu Zhang* and Guohua Zou

Capital University of Economics and Business,
Erasmus University Rotterdam and Tinbergen Institute,
Chinese Academy of Sciences and Capital Normal University

Abstract: This paper offers a new approach to address the model uncertainty in (potentially) divergent-dimensional single-index models (SIMs). We propose a model-averaging estimator based on cross-validation, which allows the dimension of covariates and the number of candidate models to increase with the sample size. We show that when all candidate models are misspecified, our model-averaging estimator is asymptotically optimal with its squared loss asymptotically identical to that of the infeasible best possible averaging estimator. In a different situation where correct models are available in the model set, the proposed method assigns all weights to the correct models asymptotically. We also propose averaging regularized estimators and prescreening methods to deal with high-dimensional covariates. We illustrate the method via simulations and two empirical applications.

Key words and phrases: Asymptotic optimality, cross-validation, model averaging, single-index model, model screening.

1. Introduction

A linear regression model is a common tool to analyze the relationship between a response variable of interest y and a vector of covariates $\mathbf{x}$ in diversified fields. However, in many applications, such a relationship is nonlinear (Naik and Tsai, 2001; Liang, Wang and Carroll, 2007). A natural extension to relax linearity is to consider a single-index model (SIM) that enables y to depend on $\mathbf{x}$ via an unknown and possibly nonlinear link function g, i.e., $y = g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + \epsilon$, where $\boldsymbol{\beta}$ is a vector of unknown parameters, and ϵ is the disturbance term. With an unknown link function, this model is more flexible than linear regression models, while maintaining relative ease of interpretation (Horowitz, 1998). It also avoids the curse of dimensionality in many nonparametric models. Various approaches have been proposed to estimate the SIM, e.g., average derivative estimation (Powell, Stock and Stoker, 1989), nonlinear least squares (Ichimura, 1993), and profile least squares (Liang et al., 2010). All of these methods require correct specification of the covariates. However, this knowledge is often unavailable in practice, especially when there are many covariates, so researchers are exposed to a potentially large

*Corresponding author. E-mail: E-mail: xinyu@amss.ac.cn