Statistica Sinica **33** (2023), 1-27
doi:https://doi.org/10.5705/ss.202020.0303

MODEL CHECKING IN LARGE-SCALE DATA SET VIA STRUCTURE-ADAPTIVE-SAMPLING

Yixin Han[1], Ping Ma[2], Haojie Ren[3] and Zhaojun Wang[1]

[1]*Nankai University,* [2]*University of Georgia and* [3]*Shanghai Jiao Tong University*

Abstract: Lack-of-fit testing is often essential in many statistical/machine learning applications. Despite the availability of large-scale data sets, the challenges associated with model checking when some resource budgets are limited are not yet well addressed. In this paper, we propose a design-adaptive testing procedure for checking a general model when only a limited number of data observations are available. We derive an optimal sampling strategy, called *Structure-Adaptive-Sampling*, to select a small subset from a large pool of data. With this subset, the proposed test possesses the asymptotically best power. Numerical results on both synthetic and real-world data confirm the effectiveness of the proposed method.

Keywords and phrases: Dimension reduction, kernel smoothing, large-scale data set, nonparametric lack-of-fit tests, optimal sampling, semiparametric modelling.

1. Introduction

The emergence of big data has provided statisticians with both unprecedented opportunities and challenges. One of the key challenges is that applying statistical methods directly to super-large data using conventional computing approaches is prohibitive, which calls for the development of new tools. Recently, statistical analysis and inference in large-scale data sets have garnered much attention. As a result, computationally scalable methods have been proposed to reduce the computation and storage effort from various aspects of applications. These include the divide-and-conquer procedures (Battey et al. (2018); Jordan, Lee and Yang (2019); Zhao, Zou and Wang (2017, 2019)), subsampling strategies (Kleiner et al. (2014); Wang, Zhu and Ma (2018)), and online learning methods (Balakrishnan and Madigan (2008); Schifano et al. (2016)). Most of the aforementioned works usually assume a parametric model, typically a linear or a logistic regression model. Therefore, it is necessary to check that a given regression model is not misspecified, such that the subsequent planning, analysis,

Corresponding author: Zhaojun Wang, School of Statistics and Data Science, LPMC & KLMDASR, Nankai University, Tianjin 300071, P.R. China. E-mail: zjwang@nankai.edu.cn.