Statistica Sinica Preprint No: SS-2020-0303	
Title	Model Checking in Large-Scale Dataset via Structure-Adaptive-Sampling
Manuscript ID	SS-2020-0303
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0303
Complete List of Authors	Yixin Han, Ping Ma, Haojie Ren and Zhaojun Wang
Corresponding Author	Zhaojun Wang
E-mail	zjwang@nankai.edu.cn

Statistica Sinica: Preprint
doi: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]*School of Statistics and Data Science, LPMC & KLMDASR, Nankai University, Tianjin, P.R. China*

[2]*Department of Statistics, University of Georgia, Athens, GA, USA*

[3]*School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, P.R. China*

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

Corresponding author: zjwang@nankai.edu.cn (Zhaojun Wang)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

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 et al., 2019; Zhao et al., 2017, 2019), subsampling strategies (Kleiner et al., 2014; Wang et al., 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, and inference can proceed in a creditable way. This study focuses on lack-of-fit checking for parametric and semiparametric models in a large-scale data set setting.

Suppose Y is the response and $\mathbf{X} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ is the p-dimensional covariate. We consider the general model in Xia (2009)

$$Y = G\left(\mathbf{X}; \boldsymbol{\beta}, \mathbf{g}\right) + \varepsilon, \tag{1.1}$$

where $\mathbf{g} = (g_1, \ldots, g_q)^\top$ are unknown smooth functions of $\mathbf{X}$, $G(\cdot)$ is known up to a parameter vector $\boldsymbol{\beta}$, and ε is a random error with $\mathbb{E}\left(\varepsilon \mid \mathbf{X}\right)$=0. This model includes many parametric and semiparametric models as special cases, such as the generalized additive models (Hastie and Tibshirani, 1986), partially linear models (Speckman, 1988), single-index or multi-index models (Hardle et al., 1993; Xia et al., 2002), and varying coefficient models (Hastie and Tibshirani, 1993). Specifically, the generalized

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

additive models and single-index models admit the forms of $Y = g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) + \varepsilon$ and $Y = g(\mathbf{X}^\top \boldsymbol{\beta}) + \varepsilon$, respectively.

The cared model checking problem can be formulated as the following test:

$$\begin{aligned} \mathbb{H}_0 &: \mathbb{E}\left(Y \mid \mathbf{X}\right) = G\left(\mathbf{X}; \boldsymbol{\beta}_0, \mathbf{g}_0\right), \text{ for some } \boldsymbol{\beta}_0 \in \Theta, \mathbf{g}_0 \in \mathcal{G}, \\ \mathbb{H}_1 &: \mathbb{E}\left(Y \mid \mathbf{X}\right) \neq G\left(\mathbf{X}; \boldsymbol{\beta}, \mathbf{g}\right), \text{ for any } \boldsymbol{\beta} \in \Theta, \mathbf{g} \in \mathcal{G}, \end{aligned} \tag{1.2}$$

where $G(\mathbf{X}; \boldsymbol{\beta}_0, \mathbf{g}_0)$ is a prespecified model with unknown $\boldsymbol{\beta}_0$ and $\mathbf{g}_0$, and Θ and $\mathcal{G}$ are the parameter and function spaces, respectively.

In this paper, we aim to answer the question that "given a limited budget or resources, how can a practitioner optimally use this budget to test (1.2) in a large-scale data set analysis". There are usually two types of limited budgets. On the one hand, computing capacity limits how much data can be processed. Because model checking is very likely one of the most preliminary steps in data analysis, practitioners are typically reluctant to expend much computational effort. Several lack-of-fit tests for small and moderate sample sizes have been proposed. Here, nonparametric smoothing-based tests and their variants, such as those of Hardle and Mammen (1993), Zheng (1996), and Fan and Huang (2005), are very popular, owing to their efficiency and flexibility; see González-Manteiga and Crujeiras (2013) and Guo and Zhu (2017) for comprehensive reviews. However, the computational complexity and memory required by these methods are typically *quadratic* in the sample size, which may greatly hamper their applicability to large-scale data sets applications. On the other hand, despite the availability of large-scale data sets, in many applications, collecting responses or labels for all data points is impossible due to measurement constraints or costs (Wang et al.,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

2017; Ren et al., 2020), especially at the beginning of the data processing. As a result, these constraints often require that we select a small subset from a large pool of given design points $\mathbf{X}$, and use the limited budget to obtain the corresponding responses Y. For example, in the problem of speech recognition, one may easily get large amounts of unlabeled audio data, but the accurate labeling of speech utterances is extremely time-consuming and requires trained linguists. Annotation at the word level can take 10 times longer than the actual audio (Tur et al., 2005).

When proven statistical methods are no longer applicable because of the two types of limited resources, a natural and appealing method of extracting useful information from the data is the subsampling method (Kleiner et al., 2014; Ma and Sun, 2015). Many existing works on subsampling take uniform samples from the full data. However, a nonuniform sampling strategy may achieve better performance. For example, in the estimation problem of linear models, Ma et al. (2015) and Ma and Sun (2015) propose a so-called algorithmic leveraging with a nonuniform sampling probability to draw a more informative subsample data set. Other recent developments include the works of Wang et al. (2019), Yao and Wang (2019), Yu et al. (2020), and Ai et al. (2021). However, the challenges associated with designing an efficient testing procedure for model checking are not yet well addressed.

In this paper, we propose a new design-adaptive testing procedure for problem (1.2) when a computation or measurement budget is imposed. The main idea is to select the most informative sample points from the full data, and then to construct a computationally tractable test statistic based on the observations of those selected

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

points. We derive an optimal sampling strategy, called *Structure-Adaptive-Sampling* (SAS), with which the proposed test possesses the asymptotically best power. An initial step is needed to obtain raw estimations of the quantities involved in the optimal design criterion. The estimated designs with plug-in estimators are shown to perform as well as the theoretical oracle design from asymptotic viewpoints. The SAS procedure addresses a key question in a general semiparametric framework: how to use limited resources to implement efficient lack-of-fit tests. Our simulation results clearly demonstrate the superiority of the proposed procedure over existing methods.

The remainder of our paper is structured as follows. In Section 2, we construct the optimal sampling designs and discuss the asymptotic justifications. Some practical guidelines are given in Section 3. Numerical studies and a real-world example are conducted in Section 4. Section 5 concludes the paper, and theoretical proofs are provided in the Appendix.

2. Methodology

Assume that there are total N available data points or observable subjects $\mathcal{X} = \{\mathbf{X}_j^a\}_{j=1}^N \in \mathbb{R}^p$. Given a measurement constraint, only n samples $\mathcal{S} = \{\mathbf{X}_i, Y_i\}_{i=1}^n$ can be obtained, or, similarly, the computational budget only allows us to deal with one data set of size n, where Y_i is the response and $n \ll N$. For the data set $\mathcal{S}$, we independently sample $\mathbf{X}_i$ from $\mathcal{X}$ with replacement and then observe its corresponding response Y_i. We start with the test construction on $\mathcal{S}$ given the full data.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

2.1 Test construction

Denote the residual as $\epsilon = Y - G(\mathbf{X}; \boldsymbol{\beta}_0, \mathbf{g}_0)$. Then, test problem (1.2) amounts to assessing whether or not $\mathbb{E}(\epsilon \mid \mathbf{X}) = \mathbb{E}\{Y - G(\mathbf{X}; \boldsymbol{\beta}_0, \mathbf{g}_0) \mid \mathbf{X}\}$ is equal to zero. A standard way is to construct a test statistic based on an estimated $\mathbb{E}(\epsilon \mid \mathbf{X})$ with fitted residuals under the null hypothesis. See, for example, Hardle and Mammen (1993), Stute et al. (1998), Dette (1999), and Fan et al. (2001). However, because of the difficulty of nonparametrically estimating the function $\mathbb{E}(\epsilon \mid \mathbf{X})$ or $\mathbb{E}(Y \mid \mathbf{X})$ when $p > 2$, the efficiency of those methods drops rapidly as the dimension p of the covariates increases.

To this end, we consider a structured alternative model as $\mathbb{E}(\epsilon \mid \mathbf{X}) = M\left(\boldsymbol{\theta}^\top \mathbf{X}\right)$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ is one projection direction with $\|\boldsymbol{\theta}\|_2 = 1$, and $M(\cdot)$ is an unknown smooth function. Thus, the alternative $\mathbb{E}(\epsilon \mid \mathbf{X}) \neq 0$ is equivalent to $\mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right) \neq 0$. It is also clear that $\mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right) = 0$ because $\mathbb{E}(\epsilon \mid \mathbf{X}) = 0$ under the null hypothesis. Then, test problem (1.2) can be formulated as follow:

$$\mathbb{H}_0 : \mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right) = 0 \quad \text{versus} \quad \mathbb{H}_1 : \mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right) = M(\boldsymbol{\theta}^\top \mathbf{X}) \neq 0, \tag{2.1}$$

where $M(\boldsymbol{\theta}^\top \mathbf{X}) \neq 0$, for some $\boldsymbol{\theta}^\top \mathbf{X} \in \Omega \subset \mathbb{R}$. Intuitively, $\mathbb{E}\left\{\epsilon \mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right)\right\}$ is a good choice to measure the derivation between $\mathbb{H}_0$ and $\mathbb{H}_1$. Under $\mathbb{H}_0$, $\mathbb{E}\left\{\epsilon \mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right)\right\} = 0$, whereas under $\mathbb{H}_1$, $\mathbb{E}\left\{\epsilon \mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right)\right\} = \mathbb{E}\left\{\mathbb{E}^2\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right)\right\} > 0$. Thus, our test can be built using $\mathbb{E}\left\{\epsilon \mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right)\right\}$, which is a popular quantity in the context of model specification tests (Zheng, 1996; Guerre and Lavergne, 2005).

Given the projection direction $\boldsymbol{\theta}$, a nonparametric estimation of $\mathbb{E}\left\{\epsilon \mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X}\right)\right\}$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

based on $\mathcal{S} = \{\mathbf{X}_i, Y_i\}_{i=1}^n$ is

$$V_f = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\widehat{\epsilon}_i \widehat{\epsilon}_j K_h\left(\omega_i - \omega_j\right)}{\sqrt{f\left(\omega_i\right)} \sqrt{f\left(\omega_j\right)}},$$

where $\omega_i = \boldsymbol{\theta}^\top \mathbf{X}_i$ is followed by a density $f(\cdot)$, $\widehat{\epsilon}_i = Y_i - G(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}})$ are the fitted residuals, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{g}}$ are estimates of $\boldsymbol{\beta}$ and $\mathbf{g}$, respectively, and $K_h(\cdot) = K\left(\cdot/h\right)/h$ denotes a one-dimensional kernel function with a bandwidth h. For notational simplicity, we only emphasize the dependence of the test statistic V_f on the density $f(\cdot)$. Under some mild conditions, it can be shown that V_f is asymptotically normal under $\mathbb{H}_0$ (Zheng, 1996; Guerre and Lavergne, 2005); that is,

$$T_f := nh^{1/2} V_f / \sigma_V \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1), \tag{2.2}$$

where $\sigma_V^2 = 2\sigma^4 |\Omega| \int K^2(u) \mathrm{d}u$ is the asymptotic variance under the null. In the latter, $|\Omega|$ is the cardinality of Ω and $\sigma^2 = \mathbb{E}(\epsilon^2 \mid \boldsymbol{\theta}^\top \mathbf{X})$. The arrow $\xrightarrow{\mathcal{L}}$ should be understood as convergence in distribution. A large value of T_f would lead to a rejection of the null.

Note that $\boldsymbol{\theta}$ plays an important role in dimension reduction in the test statistic V_f. In practice, $\boldsymbol{\theta}$ can be specified by the user or estimated using some dimension reduction techniques. See Section 3 for a detailed discussion on how to determine $\boldsymbol{\theta}$. Similar projection-based tests have been proposed in the literature. For example, Fan and Huang (2001) reduced the dimension based on $\mathbf{X}$ alone, and Xia (2009) proposed projecting the fitted residuals along a direction that adapts to the systematic departure of the residuals from the hypothetical pattern with cross-validation in the single-index model. A more closely related work is Guo et al. (2016), who checked the single-index models based on a joint estimation of the dimension reduction matrix. Other works include Zhu et al. (2017) and Tan et al. (2018).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

2.2 Optimal sampling strategy

To implement the proposed method, it is important to specify the sampling density $f(\omega)$. The conventional choice is to let $f(\omega)$ be the uniform distribution (Stute and Zhu, 2002; Guo et al., 2016), which corresponds to simple uniform sampling. However, uniform sampling may not be *optimal* in that informative subsamples are not be selected. Under certain local alternatives, the asymptotic distribution of T_f is also normal, but with a positive mean, which depends on $f(\omega)$. This implies that choosing an appropriate sampling density $f(\omega)$ will maximize the power of this test.

Next, we provide a result that sheds lights on how to determine the optimal sampling density. First, we need the following to facilitate the derivation. Let $(\boldsymbol{\beta}^*, \mathbf{g}^*) = \arg\min_{(\boldsymbol{\beta},\mathbf{g})\in\Theta\otimes\mathcal{G}} \mathbb{E}\left\{Y - G(\mathbf{X};\boldsymbol{\beta},\mathbf{g})\right\}^2$. Under $\mathbb{H}_0$, it is clear that $(\boldsymbol{\beta}^*, \mathbf{g}^*) = (\boldsymbol{\beta}_0, \mathbf{g}_0)$.

Assumption 1. *(Moments condition) For $\kappa = 4 + \gamma$ with small enough $\gamma > 0$, $\mathbb{E}(\varepsilon^\kappa \mid \mathbf{X}) \leq C_1 < \infty$, where $C_1 > 0$ is a fixed constant.*

Assumption 2. *(Density function) The density function of ω, $f(\omega)$, is continuous on the compact support $\omega \in \Omega$, satisfying $0 < \inf_{\omega\in\Omega} f(\omega) \leq \sup_{\omega\in\Omega} f(\omega) < \infty$.*

Assumption 3. *(Kernel function) $K(\cdot)$ is a continuous, nonnegative, bounded, and symmetric kernel function with a bounded first-order derivative.*

Assumption 4. *(Model) The semiparametric model can be approximated using a first-order Taylor expansion,*

$$G(\mathbf{X};\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) = G(\mathbf{X};\boldsymbol{\beta},\mathbf{g}) + \nabla G_{\boldsymbol{\beta}}^\top(\mathbf{X};\widetilde{\boldsymbol{\beta}},\widetilde{\mathbf{g}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \nabla G_{\mathbf{g}}^\top(\mathbf{X};\widetilde{\boldsymbol{\beta}},\widetilde{\mathbf{g}})(\widehat{\mathbf{g}} - \mathbf{g}),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

where $\nabla G_{\mathbf{g}}(\mathbf{X};\boldsymbol{\beta},\mathbf{g}) = \left\{\nabla G_{g_1}(\mathbf{X};\boldsymbol{\beta},\boldsymbol{z}),\ldots,\nabla G_{g_q}(\mathbf{X};\boldsymbol{\beta},\boldsymbol{z})\right\}^\top$, *with* $\nabla G_{g_k}(\mathbf{X};\boldsymbol{\beta},\boldsymbol{z}) = \partial G(\mathbf{X};\boldsymbol{\beta},\boldsymbol{z})/\partial z_k$. *Here,* $\widetilde{g}_k(\nu)$ *lies between* $g_k(\nu)$ *and* $\widehat{g}_k(\nu)$, *for* $k=1,\ldots,q$, *and* $\widetilde{\boldsymbol{\beta}}$ *lies between* $\boldsymbol{\beta}$ *and* $\widehat{\boldsymbol{\beta}}$. *And,* $\nabla G_{\boldsymbol{\beta}}(\mathbf{X};\boldsymbol{\beta},\mathbf{g}) = \partial G(\mathbf{X};\boldsymbol{\beta},\mathbf{g})/\partial\boldsymbol{\beta}$. *Further,* $\nabla G_{\boldsymbol{\beta}}$ *and* $\nabla G_{\mathbf{g}}$ *are Lipschitz continuous, and* $0 < \max_{\boldsymbol{\beta}\in\Theta,\mathbf{g}\in\mathcal{G}}\left(\mathbb{E}\left\{\nabla G_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta},g_k)\right\}^2, \mathbb{E}\left\{\nabla G_{\mathbf{g}}(\mathbf{X}_i;\boldsymbol{\beta},g_k)\right\}^2\right) \le C_2 < \infty$, *for* $i=1,\ldots,n$, *with some positive constant* C_2.

Assumption 5. *(Asymptotic representation) Suppose* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p(n^{-1/2})$. *Assume that all link functions,* $g_1,\ldots,g_q$, *own a common compact support* Γ, *and their estimators admit the following asymptotic expansions:*

$$\sup_{\nu\in\Gamma}\left|\widehat{g}_k(\nu) - g_k^*(\nu) - R_k(\nu)b^2 - n^{-1}H_k(\nu)\sum_{i=1}^{n}\phi_k(\mathbf{X}_i)Q_b\left(u_{ki}-\nu\right)\varepsilon_i\right| = O_p(n^{-1/2}),$$

for $k=1,\ldots,q$, *where* u_{ki} *is a measurable function of* $\mathbf{X}_i$, $R(\cdot)$, $H(\cdot)$, *and* $\phi(\cdot)$ *are bounded continuous functions. For some positive integer* $r \ge 2$, *the* r*th derivative of* $g_k(\cdot)$ *is bounded.* $Q_b(\cdot)$ *is a bounded, symmetric, and* r*th order continuously differentiable kernel function with smoothing parameter* b, *satisfying* $\int Q_b(u)\mathrm{d}u = 1$, $\int u^i Q_b(u)\mathrm{d}u = 0$, *and* $\int u^r Q_b(u)\mathrm{d}u \neq 0$, *for* $0 \le i < r$, $b\to 0$, *and* $nb\to\infty$ *as* $n\to\infty$.

Assumption 6. *(Bandwidth) The testing bandwidth* h *satisfies* $h\to 0$, $nh\to\infty$, *and* $nh^{1/2}b^{2r}\to 0$ *as* $n\to\infty$, *with positive integer* $r\ge 2$.

Remark 1. Assumptions 1−3 are standard in kernel-based methods, though some of them may not be the weakest possible. For instance, we only require that $f(\cdot)$ be Lipschitz continuous and bounded away from zero if some other conditions are imposed. Assumption 4 is a regularity condition on the semiparametric model. This is reasonable because we cannot expect our procedure to work well if $G(\cdot)$ is not in a

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

regular form. The formulation is quite mild and can be satisfied by all commonly used semiparametric models. Assumption 5 sets theoretical requirements for the estimates of the model, and is fulfilled by most semiparametric estimation methods and models, including the partially linear model (Speckman, 1988), additive model (Horowitz and Mammen, 2004), varying coefficient model (Fan and Zhang, 1999), and single-index model (Ichimura, 1993), under the condition that the nonparametric part $\mathbf{g}$ is a twice continuously differentiable function on Ω; see also Xia (2009). Assumption 6 gives the bandwidth requirements for implementing the test V_f with an asymptotic normal calibration. The optimal rate of the nonparametric estimation $n^{-1/5}$ appears to be not allowed if we set $b = h$ for simplicity and consider the case $r = 2$. This is very common in nonparametric specification tests, and certain under-smoothing is necessary (see Fan et al., 2001). □

Consider the local alternative model as $Y = G(\mathbf{X}; \boldsymbol{\beta}, \mathbf{g}) + \delta_n l(\omega) + \varepsilon$. Thus, the corresponding local alternatives become

$$\mathbb{H}_{1n} : M(\omega) = \delta_n l(\omega) \quad \text{for} \quad \omega \in \Omega, \tag{2.3}$$

where $M(\omega) = \mathbb{E}\left(\epsilon \mid \boldsymbol{\theta}^\top \mathbf{X} = \omega\right)$, $\delta_n \to 0$ as $n \to \infty$, and $l(\omega)$ is continuously differentiable on Ω satisfying $\mathbb{E}\left\{l^2(\omega)\right\} < \infty$ and bounded away from zero almost everywhere. Under (2.3), $\delta_n l(\omega)$ characterizes the model difference from the null to the alternative, which goes to zero as $n \to \infty$. Then, we have the following result.

Theorem 1. *Suppose Assumptions 1–6 hold. Under the local alternatives* (2.3) *with* $\delta_n = \left(nh^{1/2}\right)^{-1/2}$*, the test based on* T_f *reaches its asymptotic best power if the sampling density* $f(\omega) \propto M^2(\omega)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Under the local alternative (2.3), $\mathbb{E}(T_f) = \mathbb{E}_f\left\{l^2(\omega)\right\}/\sigma_V$ and the asymptotic variance σ_V^2 does not depend on $f(\omega)$ and $l(\omega)$. As a result, we can obtain an explicit power expression that depends on $\mathbb{E}_f\left\{l^2(\omega)\right\}$. Theorem 1 enlightens us to construct a locally most powerful test by choosing the sampling distribution $f(\omega)$ as a linear function of $M^2(\omega)$. A similar optimization technique, Cauchy's inequality, is adopted in Wang et al. (2018) to derive the nonuniform sampling strategy under some optimality criterion. However, it is an estimation problem, and its focus is to estimate the sampling probabilities when full data are available, which is quite different to our testing procedure. In our work, we only need to sample the data point $\mathbf{X}$ with sampling density $f(\cdot)$, and then observe its corresponding response Y. Because the sampling density is related to the underlying model structure, we call this the *Structure-Adaptive-Sampling* (SAS) procedure.

2.3 Estimation of optimal density $f(\cdot)$

The optimal sampling density depending on $M(\omega)$ contains the unknown function $l(\omega)$ in Theorem 1. As a result, we cannot apply an exact $f(\omega)$ directly in a practical implementation. Hence, it is necessary to obtain *raw but informative* estimates using a pilot study. Then, an approximately optimal sampling plan can be achieved. Assume we have the data set $\mathcal{S}_0 = \{\mathbf{X}_{0i}, Y_{0i}\}_{i=1}^{n_0}$ for the pilot study, where $n_0 \leq n$ and $\{\mathbf{X}_{0i}\}_{i=1}^{n_0}$ are uniformly sampled from $\mathcal{X}$. Given $\boldsymbol{\theta}$, let $\{\widehat{\epsilon}_{0i}\}_{i=1}^{n_0}$ be the fitted residuals based on

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

$\mathcal{S}_0$. Then, a consistent estimator of $M(\omega)$ is

$$\widehat{M}(\omega) = \frac{\sum_{i=1}^{n_0} K_{h_f}(\omega_{0i} - \omega)\,\widehat{\epsilon}_{0i}}{\sum_{i=1}^{n_0} K_{h_f}(\omega_{0i} - \omega)}, \tag{2.4}$$

where $\omega_{0i} = \boldsymbol{\theta}^\top \mathbf{X}_{0i}$, and h_f is a prespecified bandwidth that satisfies $h_f \to 0$ and $n_0 h_f / \log n_0 \to \infty$ as $n_0 \to \infty$. By Theorem 1, the optimal distribution $f(\omega)$ can be estimated by

$$\widehat{f}(\omega) = \frac{\max\{\xi_{n_0}, \widehat{M}^2(\omega)\}}{\int \max\{\xi_{n_0}, \widehat{M}^2(\omega)\} d\omega}, \tag{2.5}$$

where ξ_{n_0} is fixed as $O\{(\log n_0)^c/(n_0 h_f)\}$, for $c > 1$. The use of ξ_{n_0} in (2.5) ensures that the estimated density is bounded away from zero, especially under $\mathbb{H}_0$, where $M(\omega) = 0$ for any $\omega \in \Omega$. By our theoretical results given in the Appendix, $\sup_\omega |M(\omega)| = o_p(\xi_{n_0}^{1/2})$ under $\mathbb{H}_0$, and $\widehat{f}(\omega)$ degenerates to a uniform distribution with probability tending to one, that is, $\widehat{f}(\omega) = 1/|\Omega|$.

The next result shows that the effect of replacing $M(\omega)$ by appropriate estimators can be asymptotically negligible, and the efficiency of the locally most powerful test can still be achieved under some mild conditions.

Theorem 2. *Assume Assumptions 1–6 and* $(n_0/n)(h_f^2/h)^{1/2}/(\log n_0)^c \to \infty$ *all hold. Under the local alternatives* $\mathbb{H}_{1n}$ (2.3)*, the power function of our SAS test with* $\widehat{f}(\omega)$ *can be approximated by*

$$\Pi_f = \Phi\left(\left\{\int l^4(\omega)\mathrm{d}\omega / \int l^2(\omega)\mathrm{d}\omega\right\}/\sigma_V - z_\alpha\right),$$

where $\Phi(\cdot)$ *denotes the cumulative distribution function of the standard normal, and* z_α *is the corresponding upper-*α *quantile.*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Note that $\mathbb{E}_f\{l^2(\omega)\}$ is the alternative mean of the test statistic V_f and satisfies $\mathbb{E}_f\{l^2(\omega)\} \leq \int l^4(\omega)\mathrm{d}\omega / \int l^2(\omega)\mathrm{d}\omega$. Thus, using the estimated density $\widehat{f}(\omega)$ in (2.5) yields a more powerful test.

Furthermore, a more compelling result is that the optimal sampling strategy could be more prominent when the signal function $M(\omega)$ exhibits a sparse pattern. Consider the following sequence of "singular" local alternatives:

$$\mathbb{H}'_{1n} : M(\omega) = \delta'_n l(\omega) \quad \text{for} \quad \omega \in \Omega_n, \tag{2.6}$$

where $\Omega_n \subset \Omega$ satisfying $|\Omega_n| \approx a_n$, with $a_n \to 0$ a deterministic sequence, and $l(\omega)$ bounded away from zero on Ω_n almost everywhere. The main feature of these "singular" local alternatives is that they have narrow spikes and change rapidly as the sample size n increases. In other words, the alternatives in (2.6) can be regarded as sparse/high-frequency alternatives, whereas those in (2.3) can be viewed as dense/low-frequency alternatives.

Corollary 1. *Consider the "singular" local alternative* (2.6) *with* $\delta'_n = (nh^{1/2}a_n^{1/2})^{-1/2}$, $a_n/h \to \infty$, *and* $\inf_{\omega\in\Omega_n} l(\omega) \geq l_{\min} > 0$. *Suppose the conditions in Theorem 2 hold. If* $(n_0/n)(h_f^2/h)^{1/2} / \left\{a_n^{1/2}(\log n_0)^c\right\} \to \infty$, *the asymptotic power of our sampling test with* $\widehat{f}(\omega)$ *is not smaller than* $\Phi\left(\mu(f,\Omega_n)/\sigma_V(f,\Omega_n) - z_\alpha\right)$, *where* $\mu(f,\Omega_n) = |\Omega_n|^{-1/2} l_{\min}^2$ *and* $\sigma_V^2(f,\Omega_n) = 2\sigma^4 |\Omega_n| \int K^2(u)du$.

The condition $a_n/h \to \infty$ in Corollary 1 ensures that our test works well if the sparse signal size of Ω_n goes to zero more slowly than h does as n increases. The advantage of the proposed test under (2.6) is that the conditions imposed on the estimating bandwidth h_f are more relaxed in Corollary 1. That is, if $a_n = h^{1/2}$, the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

optimal rate of the nonparametric bandwidth $h_f = O(n_0^{-1/5})$ can be allowed, as long as the testing bandwidth h satisfies $nh^{3/4}/n_0^{4/5} \to 0$.

2.4 The SAS-based testing procedure

Our procedure for model checking is summarized as follows.

Algorithm 1 Model checking via structure-adaptive-sampling

Step 1 (Initialization) Specify $K(\cdot)$, n_0, n, h_f, h, and $\boldsymbol{\theta}$;

Step 2 (Pilot study) Estimate model (1.1) based on $\mathcal{S}_0 = \{\mathbf{X}_{0i}, Y_{0i}\}_{i=1}^{n_0}$, and compute the fitted residuals $\{\widehat{\epsilon}_{0i}\}_{i=1}^{n_0}$, where $\mathbf{X}_{0i}$ are uniformly sampled from $\mathcal{X}$;

Step 3 (Sampling) Obtain $\widehat{f}(\omega)$ using (2.5) with $\widehat{M}(\omega)$ in (2.4) based on $\mathcal{S}_0$ and $\{\widehat{\epsilon}_{0i}\}_{i=1}^{n_0}$; sample n data points $\mathbf{X}_i$ with $\widehat{f}(\omega)$ from $\mathcal{X}$ and obtain the corresponding Y_i as $\mathcal{S} = \{\mathbf{X}_i, Y_i\}_{i=1}^{n}$;

Step 4 (Test) Compute $\{\widehat{\epsilon}_i\}_{i=1}^{n}$ based on $\mathcal{S}$, and build $T_{\widehat{f}}$ using (2.2); further, reject $\mathbb{H}_0$ if $T_{\widehat{f}} > z_\alpha$.

To generate n samples at Step 3, we could estimate the sampling probabilities using (2.5) for N possible data points in $\mathcal{X}$, and then draw n sampling points $\{\mathbf{X}_i\}_{i=1}^{n}$ by a multinomial distribution. This sampling procedure is fast to implement with the computational complexity $O(Nn_0h_f)$. Note that the computational complexity for the estimation of $G(\cdot)$ is generally linear in the sample size; thus, the computing time of the pilot study is $O(n_0)$. Therefore, the computation of our SAS testing procedure is $O(n_0 + Nn_0h_f + n^2h)$, where $O(n^2h)$ is the complexity of computing $T_{\widehat{f}}$ in Step 4.

We use a numerical example to demonstrate the performance of our SAS test. We

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

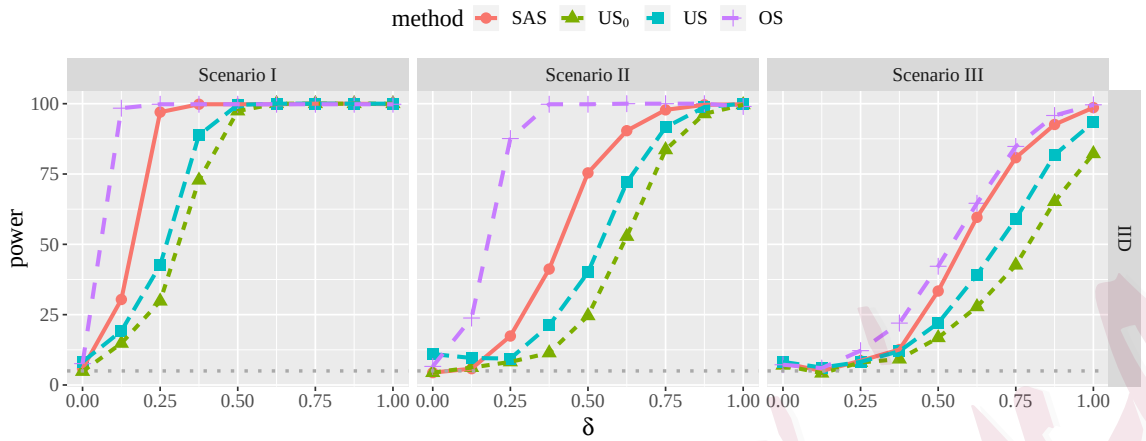

Figure 1: Empirical sizes and powers (%) for the tests with different sampling methods under Scenarios I−III when errors follow $\mathcal{N}(0,1)$ independently and the design points $\mathbf{X}$ are from $\mathcal{N}(0,\mathbf{I}_p)$. The gray dotted line is the significance level $\alpha = 0.05$.

sample $n = 1000$ data points and $n_0 = 300$ pilot samples from $\mathcal{X}$ with size $N = 10^5$, and consider three alternative models. See Scenarios I−III in Section 4 for details. Figure 1 compares the power curves of our proposed procedure, SAS, with those of three other sampling-based tests. These tests differ only in their sampling strategies. Specifically, OS uses the so-called "oracle" density $M^2(\omega)/\int M^2(\omega)d\omega$ as if $M(\omega)$ is known, and US_0 and US use a uniform distribution to sample n and $n + n_0$ points, respectively, for building T_f. The pilot study (if needed) for all methods is based on n_0 observations. The improvement of our adaptive-sampling-based test over US_0 and US is clear. The OS approach has superior performance as expected, but the difference between SAS and OS becomes smaller than US_0 and US as the signal δ increases.

In the foregoing discussion, we assume the alternative model is $\mathbb{E}(\epsilon \mid \mathbf{X}) = M(\boldsymbol{\theta}^\top \mathbf{X}) \neq 0$ for any vector $\boldsymbol{\theta}$. In fact, we cannot expect this model to hold exactly. A more realistic

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

assumption is that $\mathbb{E}(\epsilon \mid \mathbf{X}) = M_{\mathbf{B}}(\mathbf{B}^\top \mathbf{X})$, where $\mathbf{B}$ is a $p \times d$ matrix with unknown $d \geq 1$. In such a situation, we may work with a misspecified model. However, as long as the alternative satisfies $\mathbb{E}\left\{M_{\mathbf{B}}\left(\mathbf{B}^\top \mathbf{X}\right) \mid \boldsymbol{\theta}^\top \mathbf{X}\right\} \neq 0$, the proposed test still works well, and its power function can be verified using $\Pi_{M_{\mathbf{B}}} = \Phi\left(\mu(f, \mathbf{B}, \boldsymbol{\theta})/\sigma_V(\mathbf{B}, \boldsymbol{\theta}) - z_\alpha\right)$, where $\sigma_V^2(\mathbf{B}, \boldsymbol{\theta}) = 2\sigma^4 |\Psi| \int K_d^2(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}$ with $\mathbf{B}^\top \mathbf{X} \in \Psi$ and a d-dimensional kernel function $K_d(\cdot)$, and $\mu(f, \mathbf{B}, \boldsymbol{\theta}) = \mathbb{E}_f\left[\mathbb{E}\left\{M_{\mathbf{B}}(\mathbf{B}^\top \mathbf{X})\right\} \mid \boldsymbol{\theta}^\top \mathbf{X}\right]^2$. Note that if $\mathbb{E}\left\{\mathbb{E}(\epsilon \mid \mathbf{X}) \mid \boldsymbol{\theta}^\top \mathbf{X}\right\}$ is a function of $\boldsymbol{\theta}^\top \mathbf{X}$, say $d = 1$, then $\Pi_{M_{\mathbf{B}}}$ reduces to the one given in Theorem 2.

3. Practical guidelines

In this section, several practical issues on implementing the SAS procedure are discussed, including the choices of projection direction and bandwidths as well as the determination of n_0.

3.1 Determination of $\boldsymbol{\theta}$

A key aspect of the implementation of the SAS procedure is the selection of the projection direction $\boldsymbol{\theta}$. Actually, one can assign a fixed $\boldsymbol{\theta}$ based on practical requirements or estimate it using some dimension reduction techniques. For example, we want to check whether there might be a nonlinear relationship between the response Y and the covariate x_1 when the null hypothesis is a linear function. In this case, we can directly set $\boldsymbol{\theta} = (1, 0, \ldots, 0)^\top$.

In general, we can obtain a reliable estimation of $\boldsymbol{\theta}$ in the pilot study with some techniques on sufficient dimension reduction (SDR). To identify the dimension reduction subspace, the literature contains many proposals, such as the classical sliced inverse

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

regression (SIR, Li, 1991), sliced average variance estimation (SAVE, Cook and Weisberg, 1991), directional regression (DR, Li and Wang, 2007), and likelihood acquired directions (Cook and Forzani, 2009). Specifically, Xia et al. (2002) proposed the *minimum average variance estimate* (MAVE) to estimate the reduction space with fewer regularity conditions on the covariates $\mathbf{X}$.

Next, we demonstrate the asymptotic properties of the test statistic $T_{\widehat{f}}$ when $\boldsymbol{\theta}$ is estimated using a pilot study (Step 2 in Algorithm 1).

Assumption 7. *(Projection direction) The estimated projection direction with a sample of size n_0 satisfies $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n_0^{-1/2})$ as $n_0 \to \infty$, where*

$$\boldsymbol{\theta} = \underset{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_2=1}{\arg\min}\, \mathbb{E}\left\{\epsilon - \mathbb{E}(\epsilon \mid \boldsymbol{\alpha}^\top \mathbf{X} = \omega)\right\}^2.$$

Assumption 7 is very mild and typically holds for most SDR methods.

Theorem 3. *Suppose Assumptions 1–7 and $n_0 h^{3/2} \to \infty$ hold. The variance σ_V^2 can be consistently estimated by $\widehat{\sigma}_V^2 = \{n(n-1)\}^{-1} 2h \sum_{i=1}^n \sum_{j\neq i}^n \widehat{\epsilon}_i^2 \widehat{\epsilon}_j^2 K_h^2(\widehat{\omega}_i - \widehat{\omega}_j) \{\widehat{f}(\widehat{\omega}_i)\widehat{f}(\widehat{\omega}_j)\}^{-1}$. We have the following results:*

(i) Under the null hypothesis $\mathbb{H}_0$ (2.1), $T_{\widehat{f}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$;

(ii) Under the local alternative $\mathbb{H}_{1n}$ (2.3) with $(n_0/n)(h_f^2/h)^{1/2}/(\log n_0)^c \to \infty$ and $\delta_n = \left(nh^{1/2}\right)^{-1/2}$, $T_{\widehat{f}} - \{\int l^4(\omega)\mathrm{d}\omega / \int l^2(\omega)\mathrm{d}\omega\}/\widehat{\sigma}_V \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

Theorem 3 can be viewed as a counterpart of the asymptotic normality result given in Zheng (1996) or Guerre and Lavergne (2005) as shown in (2.2), by generalizing a parametric null specification model to a much more generic semiparametric one (1.1).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Owing to the involvement of the estimations of both $\boldsymbol{\theta}$ and $G(\cdot)$, the technical details of our theory are not straightforward and cannot be obtained from those existing works.

Theorem 3 implies that our SAS procedure achieves the best power with $\widehat{\boldsymbol{\theta}}$ from an asymptotic viewpoint, and the results in Theorem 2 can still be achieved. This is further verified by numerical comparisons over a wide range of values of $\boldsymbol{\theta}$ in the Supplementary Material, Table S1. In this paper, we use the MAVE method (Xia et al., 2002) to estimate $\boldsymbol{\theta}$ in the pilot study, owing to its easy implementation with the R package `MAVE`.

3.2 Bandwidth selection

Like many other smoothing-based tests, the performance of the SAS test possibly depends on the bandwidth h in the test statistic $T_{\widehat{f}}$ and on the h_f in the density estimator (2.4). By Corollary 1, the optimal h_f for estimator $\widehat{M}(\omega)$ can achieve the order $O(n_0^{-1/5})$. Thus, we take the empirical bandwidth formula $h_f = 0.5\mathrm{sd}(\omega_0)n_0^{-1/5}$ as a rule of thumb, where $\mathrm{sd}(\omega_0)$ denotes the sample standard deviation of $\omega_0 = \{\boldsymbol{\theta}^\top \mathbf{X}_{0i}\}_{i=1}^{n_0}$.

In contrast to the estimating bandwidth h_f, it is widely known that selecting a bandwidth h for optimal testing power is an open problem (Hart, 1997; Stute et al., 2005). Asymptotically, a range of bandwidths that satisfy Assumption 6 will retain the consistency of the test, whereas a larger bandwidth generally results in better power from Theorem 2. However, in practice, the condition $h \to 0$ restricts h to not being too large, and the condition $nh^{3/4}/n_0^{4/5} \to 0$ ensures that h cannot be too small. Based on our numerical results, the observed significance changes only mildly over a wide range

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

of values of h, and we recommend $h = \left\{h_f(n_0/n)^{1/5}\right\}^{2+\eta}$ with some $\eta > 0$, so that the condition $(n_0/n)(h_f^2/h)^{1/2}/\{a_n^{1/2}(\log n_0)^c\} \to \infty$ in Corollary 1 is roughly valid. This choice works well for a wide range of models and pilot sample sizes, as shown in Section 4.

3.3 Choice of sample size

In practice, we could uniformly sample n_0 data points as $\mathcal{S}_0$ for the pilot study. Note that there is a trade-off in the selection of n_0 between estimation efficiency and computational complexity. Intuitively, a larger n_0 would attain more accurate estimations of the density function $\widehat{f}(\cdot)$ and projection direction $\boldsymbol{\theta}$ (if needed), but involves a greater computational burden. To satisfy our theoretical requirements, we could roughly consider $n_0 = \lfloor n^{3/5}(\log n)^{c_0} \rfloor$, with some $c_0 > 0$. Our simulations show that reliable estimations could be obtained, and the performance of the SAS method is not affected too much as long as $n_0 \geq 200$ in the pilot study, which seems to be acceptable for a large-scale data set.

Note too that better performance would be expected when n is larger. How large the sample size n is allowed to be depends on practitioners' resource constraints, such as computational power, measurement costs, and processing time.

4. Numerical studies

In this section, we examine the performance of our proposed SAS procedure using Monte Carlo simulation studies and a real data example.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

4.1 Simulation studies

The data are generated by the model $Y = G(\mathbf{X}; \boldsymbol{\beta}, \mathbf{g}) + \varepsilon$, where ε is distributed from $\mathcal{N}(0, 1)$, and the covariates $\mathbf{X}$ are drawn independently from the whole space $\mathcal{X}$. Consider that full data points $\mathcal{X}$ with $N = |\mathcal{X}| = 10^5$ are independent and identically distributed (i.i.d) from $\mathcal{N}(0, \boldsymbol{\Sigma})$. Two classes of $\boldsymbol{\Sigma}$ are explored: one is the identity matrix $\mathbf{I}_p$, and the other has the components $(\boldsymbol{\Sigma})_{ij} = 0.5^{|i-j|}$, for $i, j = 1, \ldots, p$. We denote these two as the "IID" and "COR" cases, respectively. The sample size n is fixed as 1000. Three scenarios for $G(\mathbf{X}; \boldsymbol{\beta}, \mathbf{g})$ are considered:

- Scenario I (Linear Model): $G(\mathbf{X}; \boldsymbol{\beta}, \mathbf{g}) = \boldsymbol{\beta}^\top \mathbf{X} + \delta \cdot 0.4|\boldsymbol{\theta}^\top \mathbf{X}|^3$ under $p = 4$, $\boldsymbol{\beta} = (1, 1, -1, -1)^\top/2$, and $\boldsymbol{\theta} = (1, -1, 0, 0)^\top/\sqrt{2}$;

- Scenario II (Single-index Model): $G(\mathbf{X}; \boldsymbol{\beta}, \mathbf{g}) = 2\exp(-\boldsymbol{\beta}^\top \mathbf{X}) + \delta \cdot 0.4(\boldsymbol{\theta}^\top \mathbf{X})^2$ under $p = 6$, $\boldsymbol{\beta} = (1, -1, 0, 0, 0, 0)^\top/\sqrt{2}$, and $\boldsymbol{\theta} = (0, 0, 1, -1, 0, 0)^\top/\sqrt{2}$;

- Scenario III (Multi-index Model): $G(\mathbf{X}; \boldsymbol{\beta}, \mathbf{g}) = (\boldsymbol{\beta}_1^\top \mathbf{X})^2 + (\boldsymbol{\beta}_2^\top \mathbf{X})^2 + \delta \cdot 0.8 \exp\left(-0.4\boldsymbol{\theta}^\top \mathbf{X}\right)$ under $p = 4$, $\boldsymbol{\beta}_1 = (1, 0, 0, 0)^\top$, $\boldsymbol{\beta}_2 = (0, 1, 0, 0)^\top$, and $\boldsymbol{\theta} = (0, 0, 1, 0)^\top$.

where $\delta \geq 0$, and $\delta = 0$ corresponds to the null hypothesis in (1.2). All simulation results are based on 500 replications.

We fix the target significance level α as 0.05 and evaluate the performance of the SAS procedure by comparing the empirical sizes under the null and the powers under the alternatives. We discuss the choices of kernel functions for the nonparametric test statistics in the Supplementary Material Table S2, and find that the effect of the kernel functions can be ignored. The Epanechnikov kernel function is applied

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

here. As discussed in Section 3, we consider the empirical bandwidth formula $h_f = 0.5\text{sd}(\omega_0)n_0^{-1/5}$ in the density estimator (2.5) and $h = \left\{h_f(n_0/n)^{1/5}\right\}^{2+\eta}$ in the test statistics $T_{\widehat{f}}$ for some $\eta > 0$, where $\text{sd}(\omega_0)$ is the sample standard deviation of $\omega_0 = \{\boldsymbol{\theta}^\top \mathbf{X}_{0i}\}_{i=1}^{n_0}$. Table 1 reports the empirical sizes and powers of the SAS procedure with different bandwidths when $\boldsymbol{\Sigma}$ is from the "IID" case. We observe that three different values of $\eta \in (0, 0.2]$ present similar results: the empirical sizes and powers are not significantly different across all the settings. In addition, our SAS procedure is not affected too much under the null and presents reliable power under the alternatives when the sample size n_0 in the pilot study is larger than 200. Hence, $\eta = 0.1$ and $n_0 = 300$ are used in the rest of the simulations.

We next compare our SAS procedure with several benchmarks. In addition to the aforementioned methods with a uniform sampling strategy, namely US_0 and US, we also consider two other existing methods. The first is from Guo et al. (2016), in which they proposed a dimension reduction model-adaptive local smoothing test for parametric single-index models. We refer to this method as Guo for simplicity. The second is a global testing approach from Stute and Zhu (2002) (SZ), who developed a dimension reduction test and approximated the distribution of the test statistics based on a certain empirical process. To make a fair comparison, we apply Guo and SZ on a subset with $n + n_0$ samples. The covariates $\mathbf{X}$ are uniformly sampled from $\mathcal{X}$, and we use a wild bootstrap of 500 times to mimic their critical values.

Figure 2 compares the empirical sizes and powers of SAS, US_0, and US. Our SAS outperforms US_0 and US uniformly across all settings. This is not surprising, because

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Table 1: Empirical sizes and powers (%) of the SAS procedure with different bandwidths under Scenarios I−III when $\mathbf{\Sigma}$ is from the "IID" case.

		Scenario I			**Scenario II**			**Scenario III**		
n_0	η \ δ	0.00	0.25	0.50	0.00	0.25	0.50	0.00	0.25	0.50
200	0.05	5.8	91.0	100.0	4.2	9.2	58.2	5.0	5.4	26.8
	0.1	5.6	90.8	100.0	4.4	9.2	57.8	4.8	5.4	25.4
	0.2	6.6	90.0	99.6	4.6	9.0	57.2	4.8	5.4	23.4
300	0.05	6.6	97.0	100.0	4.2	16.6	75.2	4.4	9.0	32.2
	0.1	5.8	97.0	99.8	4.4	17.4	75.4	5.2	9.0	31.8
	0.2	5.8	95.6	99.8	4.2	16.8	74.2	4.6	9.0	30.2
400	0.05	7.0	98.0	100.0	4.6	19.8	85.8	4.6	8.8	38.6
	0.1	7.4	97.6	100.0	4.2	19.6	85.6	4.8	9.2	37.4
	0.2	7.2	97.4	100.0	4.0	18.8	84.6	4.8	9.2	35.2

the "optimal" sampling in SAS considers the data structure, as in our theoretical analysis in Section 2.3. US usually exhibits greater power than that of US_0, because the test statistic of the former is based on a larger set with sample size $n+n_0$. However, US performs a little poorly in terms of the empirical size than US_0, since the samples for the tests are dependent on the estimated $\boldsymbol{\theta}$ in the pilot study.

In Figure 3, we further compare the proposed method SAS with Guo from Guo et al. (2016) and SZ from Stute and Zhu (2002) under Scenario I. In most cases, SAS performs effectively and leads to a higher power than that of the competitors.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

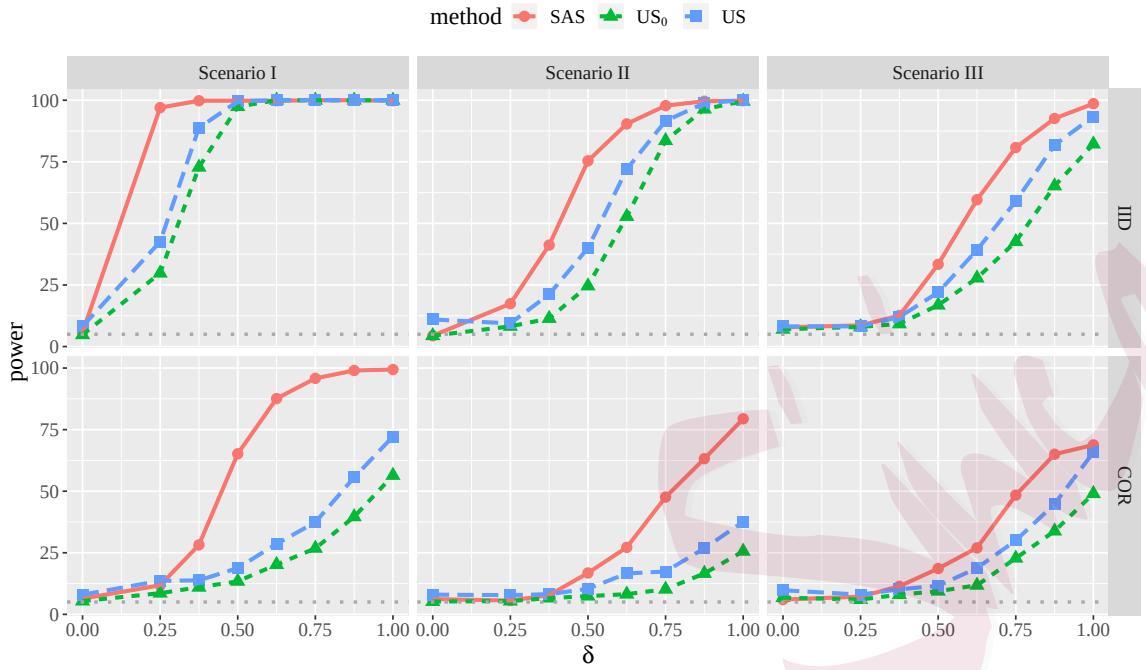

Figure 2: Empirical sizes and powers (%) for SAS, US_0, and US under Scenarios I–III.

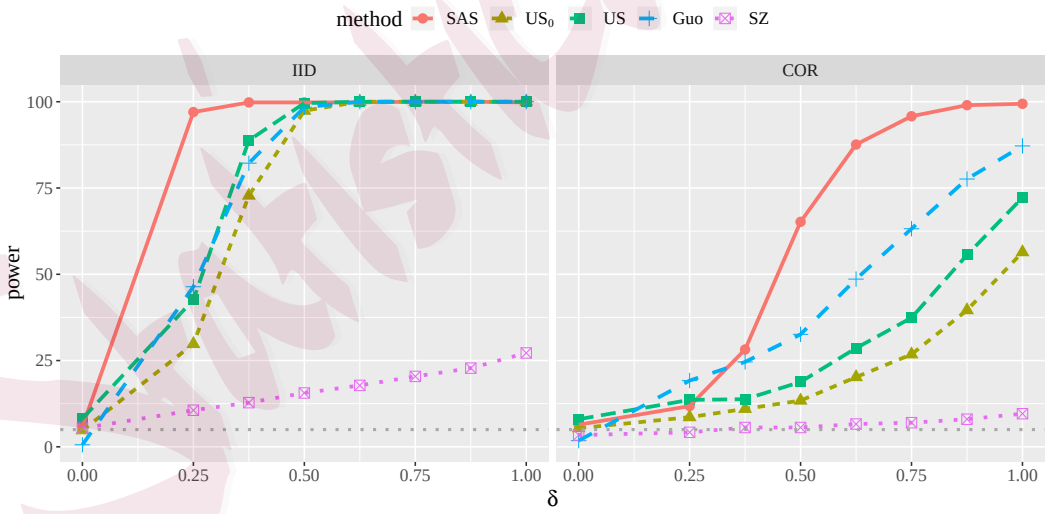

Figure 3: Empirical sizes and powers (%) for SAS, US_0, US, Guo in Guo et al. (2016), and SZ in Stute and Zhu (2002) under Scenario I.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

SZ does not perform well. Because the local smoothing based methods work better than the global testing procedure SZ for local alternatives, as suggested by existing numerical studies in the literature. We also observe that the SAS test tends to be more conservative than the Guo method slightly when the signal δ is very small, especially under the "COR" case. This is because Guo uses $n+n_0$ samples for testing. However, the power of SAS increases quickly as δ increases, and its adaptive-sampling advantage becomes remarkable.

In what follows, we consider one general varying coefficient model.

- Scenario IV (Varying coefficient model): $G(\mathbf{X};\boldsymbol{\beta},\mathbf{g}) = \boldsymbol{\beta}_1(X_1)X_2+\boldsymbol{\beta}_2(X_1)X_3+\delta\cdot 0.3\left(\boldsymbol{\theta}^\top\mathbf{X}\right)^3$ under $p=4$, $\boldsymbol{\beta}_1(X_1)=\sin(X_1)+\cos(X_1)$, $\boldsymbol{\beta}_2(X_1)=2X_1(1-X_1)$, and $\boldsymbol{\theta}=(0,1,1,1)^\top/\sqrt{3}$. $X_1\sim\mathcal{U}(0,1)$, and $X_2,X_3,X_4\sim\mathcal{N}(0,\mathbf{I}_p)$, where $\mathcal{U}(0,1)$ is the standard uniform distribution.

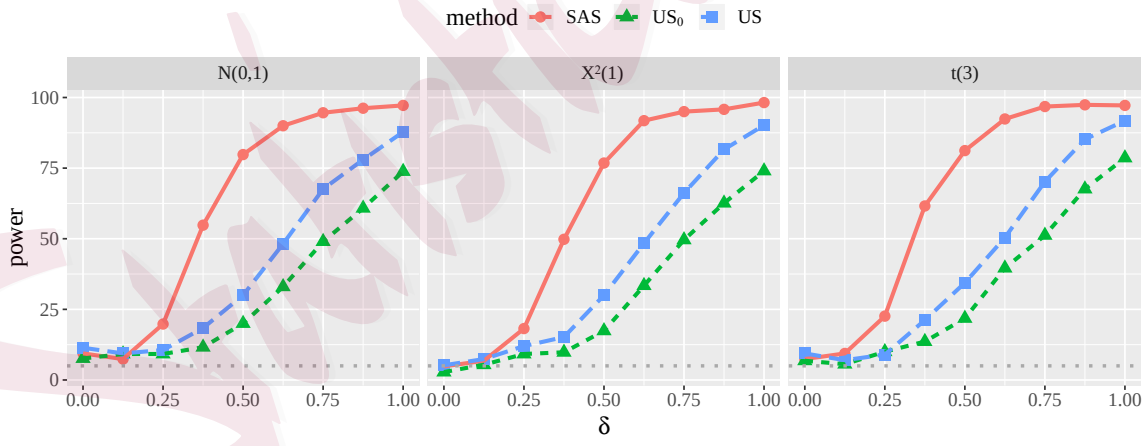

Figure 4: Empirical sizes and powers (%) for SAS, US_0, and US under Scenario IV.

In Scenario IV, we examine the robustness of the proposed test when the errors ε_i are from different distributions. Figure 4 shows the empirical sizes and powers for

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

SAS, US_0, and US under three standardized error distributions, $\mathcal{N}(0,1)$, $\chi^2(1)$, and $t(3)$. Again, it can be seen that SAS outperforms US_0 and US uniformly, and there are no significant differences among the power curves under different error distributions.

Finally, to further investigate the computational benefit of our SAS procedure in large-scale data sets, the computing time is reported in Table 2. As the full sample

Table 2: Computing times (seconds) for SAS, US_0, and US when ε_i are i.i.d from $\mathcal{N}(0,1)$ and the full sample size N is from 10^3 to 10^6 under Scenario IV.

δ	Method \ N	10^3	10^4	10^5	10^6
0.00	SAS	0.237	0.787	3.913	29.907
	US_0	0.215	0.251	0.239	0.259
	US	0.414	0.471	0.310	0.350
	FULL	0.426	7.794	750.735	77033.543
0.25	SAS	0.233	0.487	3.405	30.336
	US_0	0.196	0.222	0.230	0.252
	US	0.268	0.316	0.307	0.330
	FULL	0.432	7.845	790.955	78946.865
0.5	SAS	0.233	0.484	3.407	30.164
	US_0	0.196	0.222	0.234	0.254
	US	0.267	0.316	0.305	0.324
	FULL	0.433	7.832	795.734	81033.674

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

size N increases, the subsampling methods take significantly less computing time compared to the full data approach. We also observe that the computing time of the SAS procedure increases roughly linearly as the sample size N increases, but the time based on the full sample increases quadratically, which is consistent with our theoretical computing cost analysis. It is not surprising that US_0 and US run fast, since no sampling distribution needs to be estimated.

4.2 Real data analysis

Wave energy converters (WECs) are of interest to governments and industry as a way of complementing other renewable energy sources (Neshat et al., 2018) because they have advantages in terms of high availability of resources. However, although huge amounts of information for WECs can be recorded and monitored using buoys, analyzing all the collected data incurs a heavy computing burden. To overcome this challenge, we apply the proposed SAS to the "Wave Energy Converters Data Set." This data set includes all 216000 measuring sample points under three real wave scenarios from the southern coast of Australia, and is available from UCI Machine Learning Repository `http://archive.ics.uci.edu/ml/datasets/Wave+Energy+Converters`.

The interest in this problem is to study the relationship between the total absorbed wave power output (Y) and the WECs positions ($\mathbf{X}$). Here, we focus on the first three WEC positions, that is, $p = 6$. Since the power output is often positive but large, we take the logarithm transformation to the response Y. We randomly select 70% of the full data as the training set ($N = 151200$) and use the rest as the validation set. For the pilot study, $n_0 = 300$ samples are uniformly sampled from the training set, and are

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

used to estimate the projection direction $\boldsymbol{\theta}$ and the optimal sampling distribution $f(\cdot)$.

In this application, we would like to check whether the linear model (LM) and multi-index model (MIM) are sufficient to describe the relationship between the response and the covariates. Specifically, we denote the MIM with two indices as MIM2. MIM3 and MIM4 are defined similarly. In Table 3, we compare the estimated p-values of SAS, US_0, and US under different model assumptions.

Table 3: The estimated p-value and MSPE of WEC data set.

	p-value			**MSPE**
Model	**SAS**	**US_0**	**US**	
LM	0.000	0.000	0.000	164.686
MIM2	0.005	0.097	0.055	0.201
MIM3	0.970	0.444	0.848	0.195
MIM4	0.928	0.607	0.525	0.200

It is clear that all three methods suggest a nonlinear relationship between wave power and the WECs positions, as shown by the p-values being equal to zero under the linear model assumption. Consider the significance level α of 0.05. We find that US_0 and US perform similarly, and both tend to choose MIM2 for model construction. However, SAS suggests MIM3 is more reliable because it has a much smaller p-value 0.005 than those of the other two methods under the model assumption MIM2. This is probably because the full data consists of three real wave scenarios. The result can be verified further by comparing the mean squared prediction errors (MSPE) on the test data. In the last column of Table 3, the MSPEs under MIM3 are much smaller

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

than those of the other models.

5. Conclusion

Model checking in large-scale data sets is an important preliminary step for statistical analysis and machine learning. In this paper, we present the *structure-adaptive-sampling* (SAS) test in a general semiparametric framework to overcome the large-scale data set computational bottleneck for a limited budget or resources. The SAS procedure selects the most informative samples with an optimal sampling procedure. It is shown that our proposed method can asymptotically achieve the locally best power. The asymptotic and numerical results demonstrate the advantages of the proposed procedure in terms of testing and computation by comparing it with the uniform sampling strategy and other existing model checking approaches.

In general, the SAS procedure can be readily extended to many other testing problems with sampling techniques in modern large-scale data sets analysis, such as clustering several regression curves or data sets, as long as the design point sampling is allowed in the process. Our analysis shows that the covariate correlation plays an important role in the performance of SAS. Though our results reveal that the superiority of the proposed method is valid under certain correlations, it would be of interest to incorporate the correlation information into the testing procedure in an efficient way. In addition, we only use a small pilot data set to estimate the projection direction $\boldsymbol{\theta}$ when no further information is given. Additional research is required to obtain a more accurate estimation of the projection direction for test power improvement.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Acknowledgement

The authors thank the editor, associate editor, and anonymous referees for their valuable comments that improved the paper significantly. The authors also thank Professor Changliang Zou for his constructive suggestions and unselfish assistance. Han and Wang's research were supported by the National Natural Science Foundation of China Grants No.11931001, 11690015, 11925106, 11971247, the National Science Foundation of Tianjin 18JCJQJC46000, and the Fundamental Research Funds for the Central Universities of Nankai University 63201162. Ma's research was partially supported by the U.S. National Science Foundation under grants DMS-1903226, DMS-1925066, and the U.S. National Institute of Health under grant R01GM122080. Ren's work was partially sponsored by the Shanghai Sailing Program.

A. Appendix: Proofs

In this Appendix, we prove the main theoretical results in our paper. Before we present the proofs of the main results, we first state several essential lemmas whose proofs can be founded in the Supplementary Material. Denote $\Upsilon_j^* = G(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}}) - G(\mathbf{X}; \boldsymbol{\beta}^*, \mathbf{g}^*)$ and $\boldsymbol{\Upsilon}^* = (\Upsilon_1^*, \ldots, \Upsilon_n^*)^\top$. Under $\mathbb{H}_0$, $\Upsilon_j^* = G(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}}) - G(\mathbf{X}; \boldsymbol{\beta}_0, \mathbf{g}_0)$ due to $(\boldsymbol{\beta}^*, \mathbf{g}^*) = (\boldsymbol{\beta}_0, \mathbf{g}_0)$. For notational simplicity, we denote the following form as

$$W_n(\boldsymbol{s}, \boldsymbol{t}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{s_i t_j}{\sqrt{f(\omega_i) f(\omega_j)}} K_h(\omega_i - \omega_j),$$

where $\boldsymbol{s} = (s_1, \ldots, s_n)^\top$ and $\boldsymbol{t} = (t_1, \ldots, t_n)^\top$ are two sequences. For example, our test statistic can be written as $V_f(\boldsymbol{\theta}) = W_n(\widehat{\boldsymbol{\varepsilon}}, \widehat{\boldsymbol{\varepsilon}})$, where $\widehat{\boldsymbol{\varepsilon}} = (\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n)^\top$, $\widehat{\varepsilon}_i = Y_i - G(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Lemma A.1. *Suppose the Assumptions 1–3 and 6 hold. Denote* $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. *With a given* $\boldsymbol{\theta}$, *then under* $\mathbb{H}_0$, *we have* $nh^{1/2}W_n(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon})/\sigma_V \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

Lemma A.2. *Suppose the Assumptions 4 and 5 hold, then*

$$\sup_{\nu\in\Gamma}\left|G\left(\mathbf{X};\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}(\nu)\right) - G\left(\mathbf{X};\boldsymbol{\beta}^*,\mathbf{g}^*(\nu)\right)\right| = O_p\left\{b^2 + n^{-1/2} + (nb/\log n)^{-1/2}\right\}.$$

Lemma A.3. *Suppose the Assumptions 1–6 hold. With a given* $\boldsymbol{\theta}$, *then under* $\mathbb{H}_0$, *we have: (i)* $W_n(\boldsymbol{\varepsilon}, \boldsymbol{\Upsilon}^*) = o_p(n^{-1}h^{-1/2})$; *(ii)* $W_n(\boldsymbol{\Upsilon}^*, \boldsymbol{\Upsilon}^*) = o_p(n^{-1}h^{-1/2})$.

Lemma A.4. *Suppose the Assumptions 1–6 hold. Given* $L(\cdot)$ *is a continuously differentiable function, which satisfies* $|L(\mathbf{X})| \leq \varphi(\mathbf{X})$ *for all* $\mathbf{X} \in \mathbb{R}^p$ *and* $\mathbb{E}\{\varphi^2(\mathbf{X})\} < \infty$. *Denote* $\mathbf{L} = \{L(\mathbf{X}_i)\}_{i=1}^n$. *With a given* $\boldsymbol{\theta}$, *then under* $\mathbb{H}_0$, *the following result holds*

$$W_n(\boldsymbol{\varepsilon}, \mathbf{L}) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{\varepsilon_i L(\mathbf{X}_j)}{\sqrt{f(\omega_i)f(\omega_j)}}K_h(\omega_i - \omega_j) = O_p(n^{-1/2}).$$

Lemma A.5. *Suppose Assumptions 1–3 hold. With a given* $\boldsymbol{\theta}$, *then for the density estimator* $\widehat{f}(\omega)$ *in* (2.5) *from the pilot study, we have*

$$\sup_{\omega\in\Omega}|\widehat{f}(\omega) - f(\omega)| = O_p\left(h_f^2 + \sqrt{\frac{\log n_0}{n_0 h_f}}\right).$$

We give a sketch of our proofs. We first derive the asymptotic distribution of our SAS test statistic with a given dimension reduction direction $\boldsymbol{\theta}$. Next, the method of deriving the optimal sampling distribution $f(\cdot)$ is conducted in Theorem 1. Then, we give the power function for our SAS test in Theorem 2. At last, we extend our sampling strategy to the "singular" signal case in Corollary 1. The relevant proof of estimated dimension reduction direction is delineated in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Proposition A.1. *Suppose Assumptions 1–6 hold. Under the local alternatives* $\mathbb{H}_{1n}$ (2.3) *with* $\delta_n = \left(nh^{1/2}\right)^{-1/2}$ *and a given* $\boldsymbol{\theta}$*, we have* $T_f - \mathbb{E}_f\left\{l^2(\omega)\right\}/\sigma_V \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

Proof. Our SAS test statistic is $T_f = nh^{1/2}V_f(\boldsymbol{\theta})$. Note that $V_f(\boldsymbol{\theta})$ has the following decomposition

$$\begin{aligned} V_f(\boldsymbol{\theta}) =& \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\epsilon_i}{\sqrt{f(\omega_i)}}K_h(\omega_i-\omega_j)\frac{\epsilon_j}{\sqrt{f(\omega_j)}} \\ &- \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\epsilon_i}{\sqrt{f(\omega_i)}}K_h(\omega_i-\omega_j)\frac{\Upsilon_j^*}{\sqrt{f(\omega_j)}} \\ &+ \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\Upsilon_i^*}{\sqrt{f(\omega_i)}}K_h(\omega_i-\omega_j)\frac{\Upsilon_j^*}{\sqrt{f(\omega_j)}} \\ =:& W_n(\boldsymbol{\epsilon},\boldsymbol{\epsilon}) - 2W_n(\boldsymbol{\epsilon},\boldsymbol{\Upsilon}^*) + W_n(\boldsymbol{\Upsilon}^*,\boldsymbol{\Upsilon}^*), \end{aligned}$$

where $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^\top$ and $\epsilon_i = Y_i - G(\mathbf{X}_i;\boldsymbol{\beta}^*,\mathbf{g}^*) = \delta_n l(\omega_i) + \varepsilon_i$ under $\mathbb{H}_{1n}$.

Similarly, we can write the first term as $W_n(\boldsymbol{\epsilon},\boldsymbol{\epsilon}) = W_n(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}) + 2\delta_n W_n(\boldsymbol{\varepsilon},\boldsymbol{l}) + \delta_n^2 W_n(\boldsymbol{l},\boldsymbol{l})$ where $\boldsymbol{l} = (l(\omega_1),\ldots,l(\omega_n))^\top$. Note that $W_n(\boldsymbol{l},\boldsymbol{l})$ is a U-statistic of order two with kernel $H_n(\omega_i,\omega_j)$, say $H_n(\omega_i,\omega_j) = l(\omega_i)l(\omega_j)K_h(\omega_i-\omega_j)\left\{f(\omega_i)f(\omega_j)\right\}^{-1/2}$, and

$$\begin{aligned} \mathbb{E}\left\{H_n(\omega_1,\omega_2)\right\} &= \mathbb{E}\left[\mathbb{E}\left\{H_n(\omega_1,\omega_2)|\omega_1\right\}\right] \\ &= \frac{1}{h}\int K(u)l(\omega)l(\omega-hu)\sqrt{f(\omega)f(\omega-hu)}h\mathrm{d}u\mathrm{d}\omega \\ &= \int K(u)l^2(\omega)f(\omega)\mathrm{d}u\mathrm{d}\omega + o(1) \\ &= \mathbb{E}_f\left\{l^2(\omega)\right\} + o(1). \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

$$\begin{aligned}\mathbb{E}\left\{H_n^2(\omega_1,\omega_2)\right\} &= \frac{1}{h^2}\int \frac{l^2(\omega_i)l^2(\omega_j)}{f(\omega_i)f(\omega_j)}K^2\left(\frac{\omega_i-\omega_j}{h}\right)f(\omega_i)f(\omega_j)\mathrm{d}\omega_i\mathrm{d}\omega_j \\ &= \frac{1}{h}\int K^2(u)\mathrm{d}u\cdot\int l^2(\omega_j+hu)l^2(\omega_j)\mathrm{d}\omega_j \\ &= \frac{1}{h}\int K^2(u)\mathrm{d}u\cdot\int l^4(\omega_j)\mathrm{d}\omega_j + o(1/h) \\ &= O(1/h) = o(n).\end{aligned}$$

By Lemma S.2, we can get $W_n(\boldsymbol{l},\boldsymbol{l}) = \mathbb{E}_f\left\{l^2(\omega)\right\} + o_p(1)$. Combining this conclusion with the results in Lemma A.1 and Lemma A.3, we have $nh^{1/2}W_n(\boldsymbol{\epsilon},\boldsymbol{\epsilon})/\sigma_V - \mathbb{E}_f\left\{l^2(\omega)\right\}/\sigma_V \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

Next, we consider the second term $W_n(\boldsymbol{\epsilon},\boldsymbol{\Upsilon}^*) = W_n(\boldsymbol{\varepsilon},\boldsymbol{\Upsilon}^*) + \delta_n W_n(\boldsymbol{l},\boldsymbol{\Upsilon}^*)$. By Lemma A.2 and

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{l(\omega_i)}{\sqrt{f(\omega_i)}}\frac{1}{\sqrt{f(\omega_j)}}K_h(\omega_i-\omega_j) = \mathbb{E}_f\left\{l(\omega)\right\} + o_p(1),$$

we claim that $W_n(\boldsymbol{l},\boldsymbol{\Upsilon}^*) = O_p(n^{-1/2})O_p\left\{b^2 + n^{-1/2} + (nb/\log n)^{-1/2}\right\}$. As a consequence, we have $W_n(\boldsymbol{\epsilon},\boldsymbol{\Upsilon}^*) = o_p(n^{-1}h^{-1/2})$ by Lemma A.3.

Finally, it can be checked that $W_n(\boldsymbol{\Upsilon}^*,\boldsymbol{\Upsilon}^*) = o_p(n^{-1}h^{-1/2})$ by Lemma A.3, from which the assertion of the proposition holds. □

Proof of Theorem 1.

By Proposition A.1, it implies that the asymptotic power function of proposed test based on T_f only depends on $\mathbb{E}_f\left\{l^2(\omega)\right\}/\sigma_V$. Note that $\sigma_V^2 = 2\sigma^4|\Omega|\int K^2(u)\mathrm{d}u$ is a constant not depending on the density $f(\omega)$ under a prespecified kernel function $K(\cdot)$. According to Cauchy-Schwarz inequality, we have

$$\mathbb{E}_f\left\{l^2(\omega)\right\} = \int l^2(\omega)f(\omega)\mathrm{d}\omega \leq \left\{\int l^4(\omega)\mathrm{d}\omega\right\}^{1/2}\left\{\int f^2(\omega)\mathrm{d}\omega\right\}^{1/2},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

and the equality holds if and only if $f(\omega) = \rho l^2(\omega)$, where ρ is some constant. Thus, if the sampling distribution $f(\omega)$ is proportional to $l^2(\omega)$, i.e. $f(\omega) = l^2(\omega)/\int l^2(\omega)\mathrm{d}\omega$, we can maximize the asymptotic power function. That is the locally best power will be reached by choosing $f(\omega) = l^2(\omega)/\int l^2(\omega)\mathrm{d}\omega$. □

Proof of Theorem 2.

The test statistic is $T_{\widehat{f}}$ with estimated $\widehat{f}(\omega)$ from the pilot study. Under the local alternatie $\mathbb{H}_{1n}$, the power is $\Pi_f = \Pr(T_{\widehat{f}} > z_\alpha)$. By Proposition A.1, we have

$$\Pr(T_{\widehat{f}} > z_\alpha) - \Phi\left(-z_\alpha + \mathbb{E}_{\widehat{f}}\{l^2(\omega)\}/\sigma_V\right) \xrightarrow{\mathrm{P}} 0.$$

Using the uniform convergence rate given in Lemma A.5, we have

$$\mathbb{E}_{\widehat{f}}\left\{l^2(\omega)\right\} - \int l^4(\omega)\mathrm{d}\omega / \int l^2(\omega)\mathrm{d}\omega \xrightarrow{\mathrm{P}} 0,$$

provided that the order of signal of strength is larger than that the maximum noise level say $(nh^{1/2})^{-1/2}/\sqrt{\log n_0/n_0 h_f} \to \infty$. The condition $(n_0/n)(h_f^2/h)^{1/2}/(\log n_0)^c \to \infty$ implies that the result holds. □

Proof of Corollary 1.

For simplicity, we assume there is only one signal region Ω_n, and the proof for the case with more than one region is similar. By condition $(n_0/n)(h_f^2/h)^{1/2}/\left\{a_n^{1/2}(\log n_0)^c\right\} \to \infty$, it suffices to show that the Nadaraya-Watson estimator of $M(\omega)$ in (2.4) is still a uniformly consistent one except for the boundary under "singular" local alternatives

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

(2.6) (Ren et al., 2020), say

$$\sup_{\omega\in\Omega_n\setminus\mathbb{B}_h}\left|\widehat{M}(\omega)-M(\omega)\right|=O_p\left(h_f^2\delta_n'+\sqrt{\frac{a_n\log n_0}{n_0h_f}}\right),$$
$$\sup_{\omega\in(\Omega\setminus\Omega_n)\setminus\mathbb{B}_h}\left|\widehat{M}(\omega)\right|=O_p\left(\sqrt{\frac{a_n\log n_0}{n_0h_f}}\right),$$

where $\mathbb{B}_h$ denotes a one-dimensional interval with radius h that is around the boundary of Ω_n. The rate of bias term is obvious as $\epsilon_i=\delta_n l(\omega_i)+\varepsilon_i$ under (2.6), and the rate of variance term is reasonable because it uniformly takes maximum over $\omega\in\Omega_n$. Then, we have

$$\int_{\Omega_n\cup\mathbb{B}_h}f(\omega)\mathrm{d}\omega=1+O_p\left(\sqrt{\frac{a_n\log n_0}{n_0h_f}}\right),\quad \Pr\left\{f(\omega_i)<\xi_{n_0}^{1/2},\forall\omega_i\notin\Omega_n\cup\mathbb{B}_h\right\}\to 1,$$

where ξ_{n_0} is the threshold defined in (2.5). Hence, we have

$$\Pr\left(T_{\widehat{f}}>z_\alpha|\mathcal{X}\right)-\Phi\left(\mu(f,\Omega_n)/\sigma_V(f,\Omega_n)-z_\alpha\right)\xrightarrow{\mathrm{P}}0,$$

where $\sigma_V^2(f,\Omega_n)\approx 2\sigma^4|\Omega_n|\int K^2(u)\mathrm{d}u$, and

$$\begin{aligned}\mu(f,\Omega_n)&=nh^{1/2}{\delta_n'}^2\int_{\Omega_n\cup\mathbb{B}_h}l^2(\omega)f(\omega)\mathrm{d}\omega\\&\approx nh^{1/2}{\delta_n'}^2\int_{\Omega_n\setminus\mathbb{B}_h}l^4(\omega)\mathrm{d}\omega/\int_{\Omega_n\setminus\mathbb{B}_h}l^2(\omega)\mathrm{d}\omega\\&\geq nh^{1/2}{\delta_n'}^2\int_{\Omega_n}l^2(\omega)l_{\min}^2\mathrm{d}\omega/\int_{\Omega_n}l^2(\omega)\mathrm{d}\omega\\&=a_n^{-1/2}l_{\min}^2.\end{aligned}$$

Thus, the power of our SAS is not small than $\Phi\left(\mu(f,\Omega_n)/\sigma_V(f,\Omega_n)-z_\alpha\right)$. Finally, by the assumption that $a_n\to 0$, the asymptotic power converges to 1. □

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

B. Supplementary Material

The online Supplementary Material contains the proofs of several technical lemmas, the proof of the estimated dimension reduction direction, and additional simulation results.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

References

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, **6**:363–392.

Balakrishnan, S. and Madigan, D. (2008). Algorithms for sparse linear classifiers in the massive data setting. *Journal of Machine Learning Research*, **9**:313–337.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, **46**(3):1352–1382.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**(485):197–208.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, **86**(414):328–332.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *The Annals of Statistics*, **27**(3):1012–1040.

Fan, J. and Huang, L.-S. (2001). Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association*, **96**(454):640–652.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**(6):1031–1057.

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics*, **29**(1):153–193.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**(5):1491–1518.

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, **22**(3):361–411.

Guerre, E. and Lavergne, P. (2005). Data-driven rate-optimal specification testing in regression models. *The Annals of Statistics*, **33**(2):840–870.

Guo, X., Wang, T., and Zhu, L. (2016). Model checking for parametric single-index models: a dimension reduction model-adaptive approach. *Journal of the Royal Statistical Society, Series B*, **78**(5):1013–1035.

Guo, X. and Zhu, L. (2017). A review on dimension-reduction based tests for regressions. In *From Statistics to Mathematical Finance*, pages 105–125. Springer.

Hardle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, **21**(1):157–178.

Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, **21**(4):1926–1947.

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests*. Springer Series in Statistics. Springer New York, New York, NY.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, **1**:297–318.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, **55**(4):757–779.

Horowitz, J. L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics*, **32**(6):2412–2443.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, **58**(1-2):71–120.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, **114**(526):668–681.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society, Series B*, **76**(4):795–816.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**(479):997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**(414):316–327.

Ma, P., Mahoney, M. W., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, **16**(1):861–911.

Ma, P. and Sun, X. (2015). Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**(1):70–76.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Neshat, M., Alexander, B., Wagner, M., and Xia, Y. (2018). A detailed comparison of meta-heuristic methods for optimising wave energy converter placements. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1318–1325.

Ren, H., Zou, C., Chen, N., and Li, R. (2020). Large-scale datastreams surveillance via pattern-oriented-sampling. *Journal of the American Statistical Association*, pages 1–15.

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics*, **58**(3):393–403.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**(3):413–436.

Stute, W., Manteiga, W. G., and Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, **93**(441):141–149.

Stute, W. and Zhu, L.-X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics*, **29**(3):535–545.

Stute, W., Zhu, L.-X., et al. (2005). Nonparametric checks for single-index models. *The Annals of Statistics*, **33**(3):1048–1083.

Tan, F., Zhu, X., and Zhu, L. (2018). A projection-based adaptive-to-model test for regressions. *Statistica Sinica*, **28**(1):157–188.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Tur, G., Hakkani-Tür, D., and Schapire, R. E. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, **45**(2):171–186.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, **114**(525):393–405.

Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, **113**(522):829–844.

Wang, Y., Yu, A. W., and Singh, A. (2017). On computationally tractable selection of experiments in measurement-constrained regression models. *Journal of Machine Learning Research*, **18**(1):5238–5278.

Xia, Y. (2009). Model checking in regression via dimension reduction. *Biometrika*, **96**(1):133–148.

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B*, **64**(3):363–410.

Yao, Y. and Wang, H. (2019). Optimal subsampling for softmax regression. *Statistical Papers*, **60**(4):235–249.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2020). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, pages 1–12.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0303

Zhao, Y., Zou, C., and Wang, Z. (2017). An adaptive lack of fit test for big data. *Statistical Theory and Related Fields*, 1(1):59–68.

Zhao, Y., Zou, C., and Wang, Z. (2019). A scalable nonparametric specification testing for massive data. *Journal of Statistical Planning and Inference*, **200**:161–175.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, **75**(2):263–289.

Zhu, X., Guo, X., and Zhu, L. (2017). An adaptive-to-model test for partially parametric single-index models. *Statistics and Computing*, **27**(5):1193–1204.