Statistica Sinica Preprint No: SS-2019-0238	
Title	A projection-based consistent test incorporating dimension-reduction in partial linear models
Manuscript ID	SS-2019-0238
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0238
Complete List of Authors	Zhihua Sun Feifei Chen Hua Liang and David Ruppert
Corresponding Author	Hua Liang
E-mail	hliang@gwu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Statistica Sinica

A PROJECTION-BASED CONSISTENT TEST INCORPORATING DIMENSION-REDUCTION IN PARTIALLY LINEAR MODELS

Zhihua Sun[1], Feifei Chen[2], Hua Liang[3] and David Ruppert[4]

[1]*University of Chinese Academy of Sciences*

[2]*Beijing Normal University at Zhuhai*

[3]*George Washington University and* [4]*Cornell University*

Abstract: We propose a projection-based test to check partially linear models. The proposed test achieves a reduction in dimension and, in the presence of multiple linear regressors, behaves as though only a single covariate is present. The test is shown to be consistent and can detect Pitman local alternative hypothetical models. We further derive the asymptotic distributions of the proposed test under the null hypothesis and the local and global alternatives. Most importantly, the test's numerical performance is consistently and remarkably superior to that of its competitors. Real examples are presented for illustration. Although we assume that the nonparametric component of the model has a univariate covariate, our model can be generalized to partially linear additive models, partially linear single-index models, and other models with linear and nonparametric components.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Key words and phrases: Consistent test, curse of dimensionality, dimensionality reduction, empirical process, integrated conditional moment, projection, uncountable moments restriction.

1. Introduction

Regression models with linear and nonparametric components, or "semiparametric regression" are widely used in practice (Ruppert et al., 2003). In particular the linearity assumption of the parametric component means there is a need for methods to check whether these models provide a satisfactory fit to data. Although a number of lack-of-fit tests have been developed, they do not work well when the dimension of the parametric component is even moderately high. To address this problem, we extend the projection-based lack-of-fit test of Escanciano (2006) for parametric models to include semiparametric models. This is the first theoretical study of Escanciano's test applied to models with nonparametric components. We work only with a particular class of semiparametric models, namely partially linear models (PLMs). Extensions to other semiparametric regression models, such as partially linear additive models and partially linear single-index models, are important and straightforward in practice, although an asymptotic study will require more work. Nonetheless, the theory presented here should provide a good starting point for further research.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Consider the PLM

$$Y = X^\top \beta + g(T) + \varepsilon, \tag{1.1}$$

where $Y \in \mathbb{R}^1$ is the response variable, $X \in \mathbb{R}^p$ is a predictor vector, $\beta \in \mathbb{R}^p$ is an unknown parameter vector, $g(T)$ is an unknown smooth function of a univariate predictor T, and $E(\varepsilon^2|X,T) < \infty$.

The PLM is important in the context of semiparametric regression owing to the interpretability of the linear component and the flexibility of the nonparametric part. Various estimation methods for parametric and nonparametric components have been proposed and well studied in the literature (Ma et al., 2006; Speckman, 1988; Engle et al., 1986; Heckman, 1986; Wahba, 1984); for detailed information on estimators and their properties, see Härdle et al. (2000).

A number of methods have been proposed that check the lack-of-fit of a PLM. Define the residual $\varepsilon(U,\beta,g(T)) = Y - \{X^\top\beta + g(T)\}$, where $U^\top = (X^\top, T)$, and consider

$$\mathcal{H}_0 : \Pr\Big\{\mathrm{E}\big\{\varepsilon(U,\beta,g(T))\big|X,T\big\} = 0\Big\} = 1, \quad \text{for some } \beta \text{ and } g(t), \tag{1.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

against the alternative hypothesis:

$$\mathcal{H}_1 : \Pr\Big\{\mathrm{E}\big\{\varepsilon(U,\beta,g(T))\big|X,T\big\}=0\Big\}<1, \quad \text{for all} \quad \beta\in\mathbb{R}^p \text{ and any function } g(t).$$

Fan and Li (1996) developed a U-statistic-based test that is consistent for general semiparametric models and is applicable for PLM diagnosis. Zhu and Ng (2003) developed an empirical process-based test. Both methods have desirable statistical properties such as consistency, and perform well in terms of empirical size and power when the dimension of X is small. However, the performance of the two methods deteriorates as the dimension of the covariates increases, as noted by Xia (2009). This is further corroborated by the results of our simulation studies in Section 6 and the online Supplementary Material. Here, we find that with five-dimensional covariates in the linear part, the U-statistic-based statistic sometimes degenerates, that is, becomes equal to zero, and the empirical process-based statistic yields low empirical size and power. This is not surprising because, for the statistic proposed by Fan and Li (1996), one needs to estimate $E(\varepsilon|X,T)$ nonparametrically, which suffers from the curse of dimensionality. The statistic proposed by Zhu and Ng (2003) involves the term $I(X\le x,T\le t)$, which is equivalent to $\{\prod_{j=1}^{p} I(X_j\le x_j)\}I(T\le t)$. When p becomes larger, this product can easily degenerate for given sample sizes, causing the empirical

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

process-based statistic to degenerate.

To overcome the problems caused by the curse of dimensionality, proposed solutions include avoiding a high-dimensional nonparametric regression or applying a simple indicator weighting function. Important results from these efforts include applications of the integrated conditional moment (ICM) method proposed by Bierens (1982). The principle of the ICM method is to transform the conditional expectation condition of the null hypothesis, (i.e., $\mathrm{E}\{\varepsilon(U,\beta,g(T))|X,T\} = \mathrm{E}\{Y-\{X^\top\beta+g(T)\}|X,T\} = 0$) into an uncountable number of unconditional moment restrictions, specifically that $\mathrm{E}\{\varepsilon(U,\beta,g(T))w(X,T,\mathbf{x})\} = 0$. The weighting function $w(X,T,\mathbf{x})$ is chosen to guarantee that $\mathrm{E}\{\varepsilon(U,\beta,g(T))|X,T\} = 0$ is equivalent to $\mathrm{E}\{\varepsilon(U,\beta,g(T))w(X,T,\mathbf{x})\} = 0$ for all $\mathbf{x}$. Note that the curse of dimensionality occurs more often in model checking than it dose in estimation, because we need to deal with the regression of $\varepsilon(U,\beta,g(T))$ against $(p+1)$ covariates $(X^\top,T)$, even we just check a multiple linear model.

Several weight functions have been proposed, including the exponential weighting function (Bierens, 1982), linear indicator weighting function (Stute and Zhu, 2002; Escanciano, 2006), logistic weighting function (Lee et al., 2001), and trigonometric weighting function (Bierens and Ploberger, 1997). Some weighting functions lead to inconsistent model checking meth-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

ods and different weighting functions lead to different power properties. Furthermore, theoretically, there is no best choice among these weighting functions in term of power, because, as shown by Bierens and Ploberger (1997), they all lead to asymptotic admissible tests. Note that the statistics proposed by Fan and Li (1996) and Zhu and Ng (2003) are special cases of the ICM test corresponding to weighting functions $\mathrm{E}\{\varepsilon(U,\beta,g(T))|X,T\}$ and $I(X\leq x,T\leq t)$, respectively; unfortunately, they may suffer from the curse of dimensionality.

A popular choice of weighting function is the linear indicator weighting function $I(U^\top W\leq u)$, where $u\in\mathbb{R}^1$, U is a vector of predictors, and W is a projection direction. This function avoids both high-dimensional problems and having to use a multiple integration to calculate a Crámer–von Mises type test statistic (see Section 2). For example, Stute and Zhu (2005, 2002) and Xia et al. (2004) applied this weighting function to check generalized linear models and single-index models. Ma et al. (2014) applied a similar idea to check partially linear single-index models.

A critical step when using the ICM method with the linear indicator weighting function is the selection of the projection direction, W. This direction should ideally ensure (1) the equivalence of the null hypothesis and the weighted unconditional moment conditions, (2) the consistency of

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

the associated tests, (3) outstanding power performance under the alternatives, and (4) computational expediency. Stute and Zhu (2002) and Xia et al. (2004) chose a vector of regression parameters as the projection direction, and then weakened the testing problem to that of testing the independence of the residuals and a linear combination of regressors (Escanciano, 2006). Ma et al. (2014) chose a fixed projection direction by estimating a single-index model. Because only one fixed direction is considered, the tests proposed in Xia et al. (2004), Stute et al. (2008), and Ma et al. (2014) may be *inconsistent*, except under specific conditions.

Xia (2009) also developed a projection-based testing procedure for parametric and semiparametric models by projecting the fitted residuals onto a direction via a single-index model. The proposed method is applicable for general settings and reduces the dimensionality. However, asymptotic distributions under the null hypothesis are not available, making it difficult to control type-I errors.

To overcome these limitations, we also use projections, but allow the direction to vary such that the null hypothesis is equivalent to an infinite collection of weighted unconditional moment restrictions. Recall that $U = (X^\top, T)^\top$ and $\mathrm{E}\{\varepsilon(U, \beta, g(T))|U\} = 0$ a.e. if and only if $\mathrm{E}\{\varepsilon(U, \beta, g(T))|U^\top W\} = 0$ a.e. for every unit $(p+1)$-vector W (Lavergne

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

and Patilea, 2008; Bierens, 1990; Stinchcombe and White, 1998). Therefore, if $\mathrm{E}\{\varepsilon(U,\beta,g(T))|U\} \neq 0$ a.e., then the set $\{W : \mathrm{E}\{\varepsilon(U,\beta,g(T))|U^\top W\} \neq 0\}$ has a Lebesgue measure larger than zero. Thus, it is critical that the test statistic contains as many projection directions as possible, which ensures that $\mathrm{E}\{\varepsilon(U,\beta,g(T))|U^\top W\} \neq 0$ if $\mathrm{E}\{\varepsilon(U,\beta,g(T))|U\} \neq 0$.

This observation motivates us to assume that (1) W is independent of the response variable, covariates, and model error, and (2) W follows a uniform distribution on the unit ball in $\mathbb{R}^{p+1}$ such that every possible projection is considered. Therefore, the corresponding test can detect a deviation from the null hypothesis in any direction. As a result, the proposed statistic is consistent against all alternatives. Because $U^\top W$ in $I(U^\top W \leq u)$ is scalar, the test avoids the curse of dimensionality. Furthermore, we show that the proposed test is independent of the data-sgeneration process (see the discussion following Theorem 3) and can detect the alternative hypothesis, which approaches to the null hypothesis at the rate n^{-r} with $0 \leq r \leq 1/2$. To avoid complexity in calculating the critical value, we suggest a robust bootstrap method. Extensive numerical experiments, including two real examples, confirm our theoretical findings and demonstrate the superiority of the test.

Note that the proposed procedure can be treated as an extension of

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Escanciano (2006) to include partially linear models. However, such an extension, while important, is by no means straightforward. Escanciano (2006) focused mainly on parametric models and required an asymptotic expansion(A3(b)), for the estimators of the parameters that does not hold for the estimators of the parametric and nonparametric components in the PLM. Furthermore, in a comparison of the projection test with the existing methods, Escanciano (2006) uses simulations to show the power gain of the projection test. In this study, we focus on both the superior power of the proposed procedure and its dimension-reduction characteristics.

The rest of this paper is organized as follows. In Section 2, we develop an empirical-process testing statistic using projection for (1.2). The asymptotic properties of the testing statistic under the null and alternative hypothetical models are shown in Sections 3 and 4, respectively. In Section 5, we develop a wild bootstrap method to calculate the critical value. Simulation studies and real-data analyses are conducted in Section 6. The assumptions and estimations of β and $g(\cdot)$ are given in the Appendix. The proofs of the main results and additional simulation results are presented in the online Supplemental Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

2. The proposed test

Let $\{(Y_i, X_i, T_i), i = 1, \ldots, n\}$ be a sample from (Y, X, T), and let $\hat{\beta}_n$ and $\hat{g}_n(\cdot)$ be the estimators of β and $g(\cdot)$, respectively; see Appendix A.2 for the definitions. Write $\hat{\varepsilon}(U_i, \hat{\beta}_n, \hat{g}_n(T_i)) = Y_i - \{X_i^\top \hat{\beta}_n + \hat{g}_n(T_i)\}$, where $U_i^\top = (X_i^\top, T_i)$. Define

$$M_{n,pro}(u, W) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\varepsilon}(U_i, \hat{\beta}_n, \hat{g}_n(T_i)) I\left(U_i^\top W \leq u\right),$$

for $(u, W) \in \Pi$, where $\Pi = \mathbb{R}^1 \times \mathbb{S}^{p+1}$ and W is uniformly distributed on $\mathbb{S}^{p+1} = \{w \in \mathbb{R}^{p+1} : \| w \| = 1\}$, the unit ball in $\mathbb{R}^{p+1}$.

Our projection-based test statistic is

$$\mathcal{T}_{n,pro} = \int_{-\infty}^{\infty} \int_{\mathbb{S}^{p+1}} \{M_{n,pro}(u, w)\}^2 F_{nw}(du) dw,$$

where $F_{nw}(u) = 1/n \sum_{i=1}^{n} I(U_i^\top w \leq u)$ and W has been integrated out. When the test statistic is sufficiently large, we reject the null hypothesis. The estimated empirical process $M_{n,pro}(u, w)$ is actually the cumulative sum of the estimated model error, and $\mathcal{T}_{n,pro}$ is a Crámer-von Mises type test statistic.

Note that the test statistic $\mathcal{T}_{n,pro}$ is equal to the summation

$$\mathcal{T}_{n,pro} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \hat{\varepsilon}(U_i, \hat{\beta}_n, \hat{g}_n(T_i)) \hat{\varepsilon}(U_j, \hat{\beta}_n, \hat{g}_n(T_j)) A_{ijl},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

where $A_{ijl} = \int I(U_i^\top w \le U_l^\top w) I(U_j^\top w \le U_l^\top w) dw$. By an argument of Escanciano (2006), we have

$$A_{ijl} = C_q \left| \pi - \arccos \left\{ \frac{(U_i - U_l)^\top (U_j - U_l)}{|U_i - U_l| \, |U_j - U_l|} \right\} \right|,$$

with $C_q = \pi^{(q/2)-1}/\Gamma(q/2+1)$, where $\Gamma(\cdot)$ is the gamma function and $q = p + 1$. Thus, the calculation of the statistic can be transformed to a calculation of a summation. This avoids the multiple integration in Härdle and Mammen (1993) and makes the implementation easier.

3. Asymptotic properties under the null hypothesis

We now study the asymptotic properties of the projection-based test statistic under the null hypothesis. We show that, for fixed w, the estimated empirical process $M_{n,pro}(u, w)$, $-\infty < u < \infty$, converges to a centered Gaussian process, and $\mathcal{T}_{n,pro}$ converges to an integrated squared Gaussian process. Let $g_1(t) = E(X|T = t)$, $g_2(t) = \mathrm{E}(Y|T = t)$, $\widetilde{X} = X - g_1(T)$, $\Gamma(u, w) = \mathrm{E}\{\widetilde{X}^\top I(U^\top W \le u)|W = w\}$, $\Sigma = \mathrm{E}(\widetilde{X}\widetilde{X}^\top)$, and $\Psi_u(U, Y, \varepsilon, W) = \varepsilon[I(U^\top W \le u) - \mathrm{E}\{I(U^\top W \le u|T, W)\}] - \Gamma(u, W)\Sigma^{-1}\varepsilon\widetilde{X}$. We have the following result.

Theorem 1. *Suppose that Conditions (C1)–(C5) in the Appendix hold. Under the null hypothesis (1.1), for any given nuisance parameter* $W =$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

$w \in \mathbb{S}^{p+1}$, the estimated empirical process $M_{n,pro}(u,w)$, $\infty < u < \infty$, converges in distribution to $M_{pro}(u,w)$, $\infty < u < \infty$, in the Skorohod space $S[-\infty,\infty]$, where $M_{pro}(u,w)$ is a centered Gaussian process with covariance function

$$\mathrm{cov}\{M_{pro}(u_1,w), M_{pro}(u_2,w)\} = \mathrm{E}\{\Psi_{u_1}(U,Y,\varepsilon,W)\Psi_{u_2}(U,Y,\varepsilon,W)|W=w\}.$$

For the test statistic $\mathcal{T}_{n,pro}$, we have

$$\mathcal{T}_{n,pro} \xrightarrow{L} \int \{M_{pro}(u,w)\}^2 F_w(du)dw,$$

where F_w is the conditional distribution of $U^\top W$, given W.

In $\mathcal{T}_{n,pro}$, if the weighting function is taken to be one, then $\mathcal{T}_{n,pro}$ reduces to a score-type statistic. However, this score-type test cannot detect an alternative that satisfies $\mathrm{E}\left[Y - \left\{X^\top\beta + g(T)\right\}\right] = 0$ a.e., but $\mathrm{E}[Y - \{X^\top\beta + g(T)\}|X,T] \neq 0$ a.e.

4. Analysis of the asymptotic power

In the following, we investigate the power behavior of the statistic under local and global alternatives. We consider the local alternative with a deviation of a nonlinear measurable function of (X,T) from the null hypothesis; that is,

$$\mathcal{H}_{1n} : \Pr\left\{Y = X^\top\beta + g(T) + n^{-1/2}D(X,T) + \varepsilon\right\} = 1, \qquad (4.3)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

where $\mathrm{E}(\varepsilon|X,T)=0$, and $D(X,T)$ cannot take the form of $X^\top\beta+g(T)$ for any β and $g(T)$ and is a measurable function of (X,T) satisfying with $0<\mathrm{E}\{D^2(X,T)\}<\infty$. Let $\Omega(u,w)=\mathrm{E}\{\widetilde{D}(X,T)I(U^\top W\le u)|W=w\}-\Gamma(u,w)\Sigma^{-1}\,\mathrm{E}\{\widetilde{X}\widetilde{D}(X,T)\}$, with $\widetilde{D}(X,T)=D(X,T)-\mathrm{E}\{D(X,T)|T\}$. Then, we have the following result.

Theorem 2. *Under Conditions (C1)–(C5) in the Appendix and the alternatives in (4.3), we have*

$$\mathcal{T}_{n,pro}\xrightarrow{L}\int\{M_{pro}(u,w)+\Omega(u,w)\}^2F_w(du)dw,$$

where $M_{pro}(u,w)$ is defined in Theorem 1.

Compared with the results of Theorem 1, Theorem 2 indicates that there is an additional component $\Omega(u,w)$ in the asymptotic distribution of the statistic $\mathcal{T}_{n,pro}$ under the local alternatives in (4.3). The quantity $\Omega(u,w)$ reflects the distance between the null and the alternative hypotheses. Therefore, the proposed statistic can detect a local alternative that approaches the null hypothetical model at the parametric rate. Such a detection cannot be achieved if one uses the local test methods (Härdle and Mammen, 1993; Li and Wang, 1998).

We further consider the following global alternative hypotheses:

$$\mathcal{H}_{1n}:\Pr\left\{Y=X^\top\beta+g(T)+D(X,T)+\varepsilon\right\}=1. \tag{4.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

We have the following results.

Theorem 3. *Under Conditions (C1)–(C5) in the Appendix and the alternatives in (4.4), we have* $\mathcal{T}_{n,pro} \longrightarrow \infty$ *as* $n \to \infty$.

Theorem 3 shows that the statistic $\mathcal{T}_{n,pro}$ diverges to infinity under the global alternative hypothesis in (4.4). Therefore, it has asymptotic power one and is consistent. Note that the results of Theorems 1–3 do not depend on distributional assumptions on the model error, but do allow for error heteroscedasticity.

We also consider the following local alternative hypothetical models:

$$\mathcal{H}_{1n} : \Pr\left\{Y = X^\top \beta + g(T) + n^\alpha D(X,T) + \varepsilon\right\} = 1. \tag{4.5}$$

Theorem 4. *Under Conditions (C1)–(C5) in Appendix and the alternatives in (4.5), with* $-1/2 < \alpha < 0$, *we have* $\mathcal{T}_{n,pro} \longrightarrow \infty$ *as* $n \to \infty$.

When $-1/2 < \alpha < 0$, the convergence rate of model (4.5) to the null hypothetical model is between those of models (4.3) and (4.4). Theorems 1, 3, and 4 show that the test can detect alternative models converging to the null model with rates n^α, for $-1/2 \leq \alpha \leq 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

5. A bootstrap option for critical value calculation

Theorem 1 gives the asymptotic distribution of the statistic $\mathcal{T}_{n,pro}$ under the null hypothesis. An immediate concern is that this distribution may be case-dependent, which complicates the calculation of the critical value. To overcome this potential difficulty, we suggest using the bootstrap method to determine the critical value.

To begin bootstrapping, generate an independent and identically distributed (i.i.d.) random variable sequence $\{V_i, i = 1, \ldots, n\}$ with mean zero and variance one, that also satisfies the condition that $|V_i| \leq c$ for some finite constant c. Let $Y_i^* = X_i^\top \hat{\beta}_n + \hat{g}_n(T_i) + [Y_i - \{X_i^\top \hat{\beta}_n + \hat{g}_n(T_i)\}]V_i$. Then, calculate the statistic $\mathcal{T}_{n,pro}$, denoted by $\mathcal{T}^*_{n,pro}$, based on the bootstrap sample $\{(Y_i^*, X_i, T_i), i = 1, \ldots, n\}$. Repeat the above process B times and obtain $\mathcal{T}^*_{n1,pro}, \ldots, \mathcal{T}^*_{nB,pro}$. Then, calculate the $1-\alpha$ empirical quantile of the bootstrap statistic based on $\{\mathcal{T}^*_{n1,pro}, \ldots, \mathcal{T}^*_{nB,pro}\}$, which is taken as the α-level critical value.

Note that for the bootstrap procedure, it is not necessary to estimate any new quantities, such as the influential function. In addition, the testing procedure is data-driven. Given only the sample $\{(Y_1, X_1, T_1), \ldots, (Y_n, X_n, T_n)\}$, the proposed testing procedure using the bootstrap-generated critical value can determine whether the partially linear model fits the data adequately,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

without any other information on the data-generation process.

For the bootstrap testing statistic $\mathcal{T}^*_{n,pro}$, we have the following result.

Theorem 5. *Under the null hypothesis (1.1) or alternative hypothesis (4.4), if Conditions (C1)–(C5) in the Appendix are satisfied, the conditional distribution of $\mathcal{T}^*_{n,pro}$ converges in distribution to the limiting null distribution of $\mathcal{T}_{n,pro}$, given $\{(Y_1, X_1, T_1), \ldots, (Y_n, X_n, T_n), \ldots\}$.*

Theorem 5 shows that the bootstrap test statistic has the same asymptotic distribution as that of the proposed test. By repeatedly generating series of i.i.d. random variables $\{V_i, i = 1, \ldots, n\}$, we can obtain a series of bootstrap test statistics that can be viewed as a sample coming from the population $\mathcal{T}_{n,pro}$. Then, we can calculate the empirical quantile of the distribution of $\mathcal{T}_{n,pro}$. The critical value determined using this method approximates the theoretical value, regardless of whether the data are from the null hypothetical model (1.1) or the alternative hypothetical model (4.4).

6. Simulations and real data analyses

6.1 Simulation studies

In this section, we report simulation results to evaluate the finite sample performance of the proposed method. For the comparisons, four tests (i.e.,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Fan and Li's test, T_n^u; Zhu and Ng's test, T_n^s, Xia's test, T_n^{Xia}; and the proposed test, $\mathcal{T}_{n,Pro}$) were evaluated. Two settings with were considered, namely, with two-dimensional, and 20-dimensional covariates in the linear part were considered. Additional simulation results for the settings with five- and 10-dimensional linear covariates are presented in the online Supplemental Material. In the estimation procedure (see Appendix A.2), we used a Gaussian kernel and bandwidth $h_n = 1.06 \min(\text{std}(T), 3\hat{Q}r/4)\, n^{-1/3}$, where std$(T)$ and $\hat{Q}r$ are the sample standard deviation and interquantile of $\{T_1, \ldots, T_n\}$, respectively. This choice of bandwidth is a combination of a rule of thumb and an undersmoothing method. We considered three different sample sizes: $n = 60, 100$, and 200. All simulation results are based on 1000 replications. For each replication, the bootstrap process was repeated 300 times. The nominal level was set to 0.05 and 0.1.

Example 1. We consider candidate models with two-dimensional linear covariates and possible interaction between the linear covariates:

$$Y = \beta_1 X_1 + \beta_2 X_2 + g(T) + C X_1 X_2/2 + \varepsilon \tag{6.6}$$

with $X_1, X_2 \sim \mathcal{U}(0, \pi)$, $g(T) = \exp(T^2 - 2T)$, $T \sim \mathcal{U}(0, 1)$, $\varepsilon \sim \mathcal{N}(0, 1)$, and $\beta_1 = 2, \beta_2 = 3$. To examine the empirical size and power of each test, we took $C = 0, 0.2, 0.4, 0.6, 0.8, 1.0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Example 2. We consider candidate models with 20-dimensional linear covariates:

$$Y = X^\top \beta + g(T) + C \sum_r \log(X_r^2 + T^2) + \varepsilon, \tag{6.7}$$

where $X = (X_1, \ldots, X_{20})^\top$, $g(T) = T^2$, $T \sim \mathcal{N}(0, 1)$, $\varepsilon \sim \mathcal{N}(0, 0.5)$, and $\beta = 1_{20}$ (a 20-dimensional vector of ones). Let X follow a multivariate normal distribution $\mathcal{N}_{20}(0, \Sigma)$, with $\Sigma = (\sigma_{jj'})$ and $\sigma_{jj'} = 0.1, j, j' = 1, \ldots, 20$. We used $C = 0, 0.1, 0.2, 0.3, 0.4, 0.5$.

We calculated the proportions of times the null hypothesis was rejected among the 1000 replicates. This yields the empirical size under the null hypothesis (i.e., $C = 0$) and the empirical power under the alternative hypothesis (i.e., $C \neq 0$). We report the rejection proportions of the tests in Figures 1 and 2, where $\mathcal{T}_{n,Pro}$, T_n^s, T_n^u, and T_n^{Xia} denote the proposed test (solid line with filled diamond), Zhu and Ng's test (dotted line with filled circle), Fan and Li's test (dashed line with filled square), and Xia's test (dot-dash line with filled triangle), respectively. The thin horizontal line indicates the nominal level of 0.05 or 0.1.

In Example 1, the empirical sizes of T_n^s and $\mathcal{T}_{n,Pro}$ are close to the nominal levels, while the empirical sizes of T_n^u and T_n^{Xia} are lower than the nominal levels. With regard to the power curves, $\mathcal{T}_{n,Pro}$ clearly performs best, followed by T_n^{Xia}, T_n^s, and T_n^u in a consistent order for all configura-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

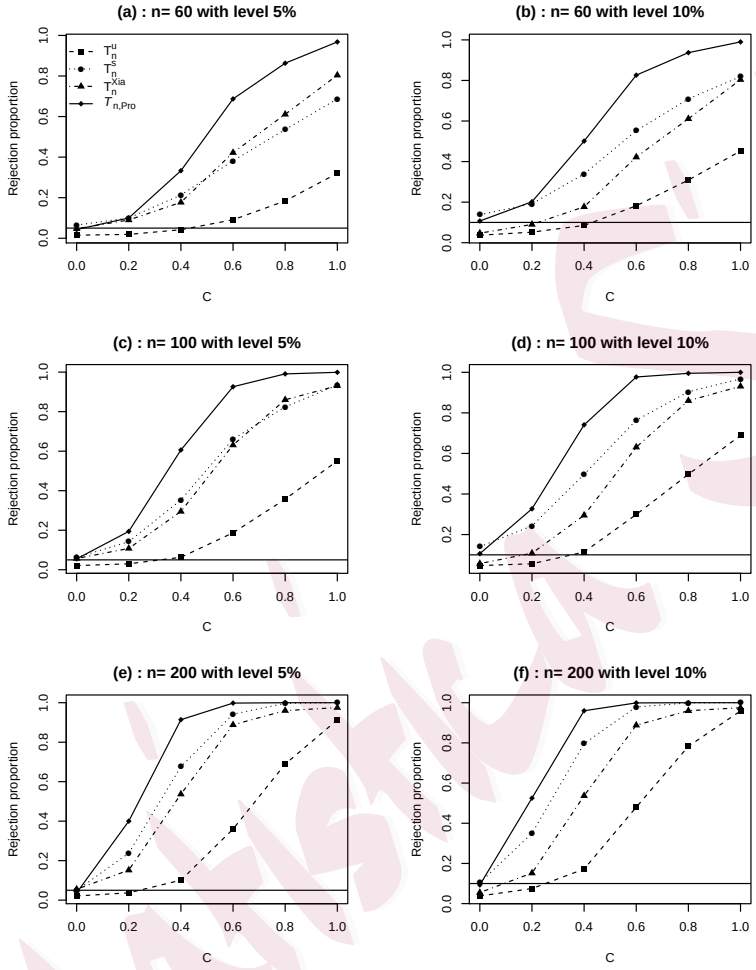

Figure 1: Simulation results for model (6.6) in Example 1. Rejection proportions of four methods against C with different sample sizes and test levels 0.05, 0.1.

tions.

In Example 2, the performance of $\mathcal{T}_{n,Pro}$ is still very promising, while the other tests almost crash. Specifically, T_n^u is always equal to zero, which

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

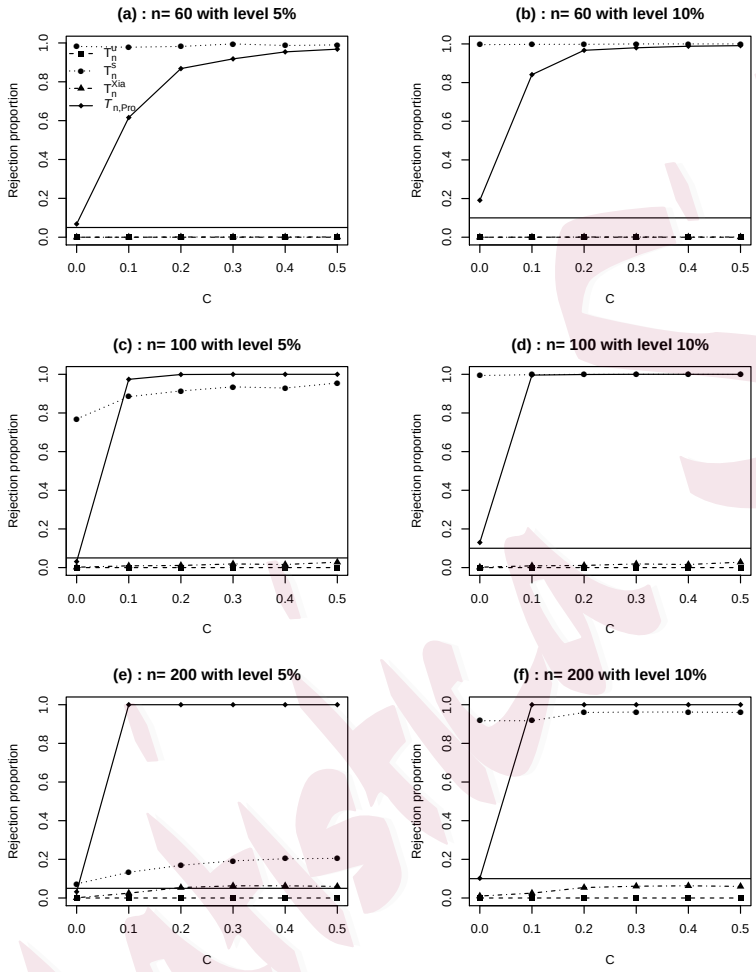

Figure 2: Simulation results for model (6.7) in Example 2. The legend is the same as that in Figure 1.

causes the empirical size and power to be zero. Furthermore, T_n^s and its bootstrap version may degenerate to zero, which results in large empirical sizes. Though T_n^{Xia} is free from any degeneration, its power curve indicates that it does not perform well when the dimension of X is moderate or large.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Table 1: Failure times among the 1000 replicates for the four tests in model (6.7) with different sample sizes and different C-values.

	n=60		n=100		n=200			
C	F^u	F^s	F^u	F^s	F^u	F^s	F^{Xia}	F^{Pro}
0.0	1000	995	1000	998	1000	946	0	0
0.5	1000	996	1000	987	1000	953	0	0
1.0	1000	999	1000	985	1000	949	0	0
2.0	1000	997	1000	989	998	946	0	0
3.0	1000	994	1000	983	1000	950	0	0
4.0	1000	999	1000	980	1000	945	0	0

F^u, F^s, F^{Xia}, and F^{Pro}: corresponding to the tests by Fan and Li (1996), Zhu and Ng (2003), Xia (2009), and the proposed test, respectively.

We report the failure times of the tests T_n^u, T_n^s, T_n^{Xia}, and $\mathcal{T}_{n,Pro}$ under Example 2 in Table 1. There were no failures for the four tests in Example 1. In Example 2, T_n^u and T_n^s almost always degenerated in all configurations. This may explain why the power curves of these two tests in Figure 2 are so flat.

Overall, the proposed test performs best, with satisfactory empirical size and power. Most importantly, the proposed test is free from the curse of dimensionality. This feature becomes more significant with higher-dimensional covariates.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

6.2 Real-data analyses

Additive models are perhaps the most realistic, parsimonious option when the relationship between the dependent variable and the covariates may not be linear. On the other hand, if some nonparametric components can be simplified to linear components, the estimation can be more efficient and easy to interpret. In this case, partially linear models are preferable to additive models. In the real-data analysis, our preliminary exploration indicates that the relationship between the dependent variable and the covariates is not linear. However, whether a partially linear model can parsimoniously reflect this relationship is unclear. We therefore apply the proposed method.

In this section, we apply the proposed test $\mathcal{T}_{n,Pro}$ and the three tests T_n^u, T_n^s, and T_n^{Xia} used in the simulation studies to analyze two real data sets. We test whether the partially linear model in (1.1) can adequately fit the data sets. The choices of the kernel function and bandwidth are the same as those in the simulation studies, in principle.

Example 3. (Analysis of hitters' salary data) In this example, we apply the four tests to analyze hitters' salary data, which were analyzed previously by Xia et al. (2002). After removing 59 missing values from the original data set of 322 observations, we were left with 263 observations. The annual salary in 1987 served as the response variable Y. We treated home runs during

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

their entire career up to 1986 (`CHmRun`) as the nonlinear component, and the following 15 covariates as linear components: times at bat in 1986 (`AtBat`), hits in 1986 (`Hits`), home runs in 1986 (`HmRun`), runs in 1986 (`Runs`), runs batted in 1986 (`RBI`), walks in 1986 (`Walks`), years in major leagues (`Years`), times at bat during their entire career up to 1986 (`CAtBat`), hits during their entire career up to 1986 (`CHits`), runs during their entire career up to 1986 (`CRuns`), runs batted in during their entire career up to 1986 (`CRBI`), walks during their entire career up to 1986 (`CWalks`), put-outs (`PutOuts`), assistances (`Assists`), and errors (`Errors`). For numerical convenience, all predictors were standardized to have mean zero and variance one.

Having conducted 5000 bootstrap replications, we obtained the p-values based on T_n^u, T_n^s, and $\mathcal{T}_{n,Pro}$ to be 0.3756, 0.2538, and 0.0170, respectively. We also had $SCV_n = 0.9820 < TSS_n = 0.9962$ for the test T_n^{Xia}. Here, SCV_n and TSS_n are the single-indexing cross-validation values and the average residual sum of squares, respectively, that is, $\sum_{i=1}^n (\widehat{\varepsilon}_i - \sum_{j=1}^n \widehat{\varepsilon}_j/n)^2/n$. See Xia (2009) for the calculation of SCV_n and TSS_n. Therefore, the proposed method and Xia's method both suggest that we should reject the null hypothesis, while the tests of Fan and Li (1996) and Zhu and Ng (2003) suggest not rejecting the null hypothesis of the partially linear model. Based on our simulation results, we prefer to reject the null hypothesis.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

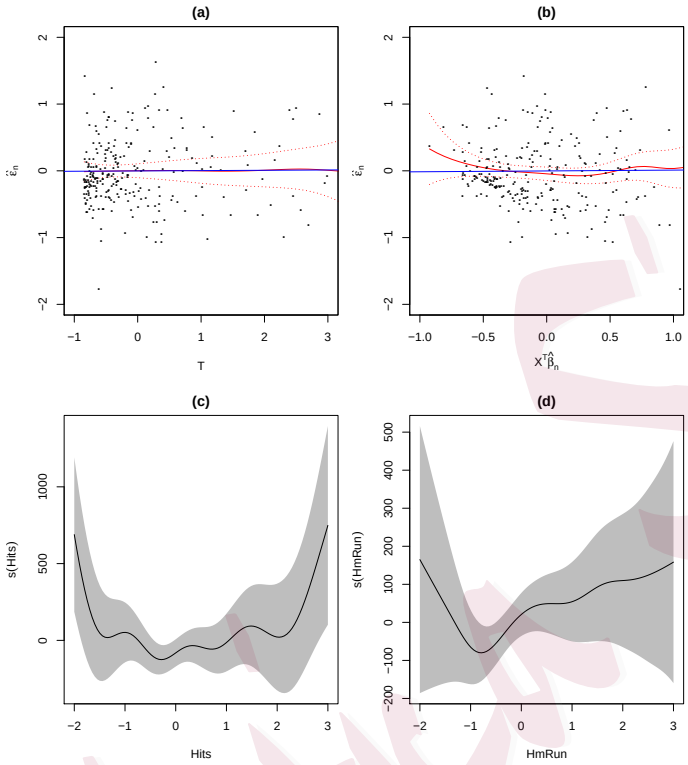

Figure 3: Results for the hitters salary data. The estimated residuals $\hat{\varepsilon}_n$ versus `CHmRun` (a) and $X^\top \hat{\beta}_n$ (b) along the nonparametric estimated curves with 95% confidence bands. The estimated curves of the salary against hits in 1986 (c) and home runs in 1986 (d) via the gam fitting.

We show scatter plots of the estimated residuals $\hat{\varepsilon}_n$ versus `CHmRun` and $X^\top \hat{\beta}_n$ in Figure 3 (a) and (b), and a nonparametric regression of salary against `Hits` and `HmRun` via the gam fitting in Figure 3 (c) and (d). Both curves show significant nonlinear patterns. This evidence further supports the finding that a partially linear model is not adequate to fit this data set,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

supporting the conclusion of the proposed method and of Xia's method.

Example 4. (Analysis of body fat data) We studied a body fat data set, available at http://lib.stat.cmu.edu/datasets/bodyfat, with 249 observations after removing three outliers from the original data set. The logarithm of the percentage of body fat serves as the response variable Y. There are 11 predictors in the linear part: age (`Age`), weight (`Weight`), height (`Height`), chest circumference (`CChest`), abdomen circumference (`CAbdomen`), hip circumference (`CHip`), thigh circumference (`CThigh`), ankle circumference (`CAnkle`), bicep (extended) circumference (`CBiceps`), forearm circumference (`CForearm`), and wrist circumference (`CWrist`).

Following the editor's suggestion, we apply our method to a partially linear model with a two-dimensional nonparametric component; that is, we consider

$$Y = X^\top \beta + g(\texttt{CKnee}, \texttt{CNeck}) + \varepsilon.$$

The procedure and theory are still valid for this situation, with mild additional assumptions, although we focus only on univariate T. For numerical convenience, all predictors were standardized with mean zero and variance one.

Based on 5000 bootstrap replications, we obtained the p-values based on T_n^s and $\mathcal{T}_{n,Pro}$ to be 0.1022 and 0.0058, respectively, while the test based

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

on the U-statistic T_n^u degenerated. For the test T_n^{Xia}, we obtained $SCV_n = 1.2615 > TSS_n = 0.9960$ and, therefore, the null hypothetical partially linear model should not be rejected. Thus, the tests T_n^s and T_n^{Xia} suggest not rejecting the null hypothesis, and the test T_n^u is not applicable. However, the proposed test suggests that we reject the null hypothesis, which means that the hypothesized partially linear model does not adequately fit this body fat data.

To investigate whether the above results seem sensible, we plot the estimated surface of the nonparametric function $g(T)$ with $T = (T_1, T_2)^\top = (\texttt{CKnee}, \texttt{CNeck})^\top$ in Figure 4 (a), indicating that it is difficult to find a suitable form to model the function $g(T)$. We also provide scatter plots of the estimated residuals $\hat{\varepsilon}_n$ versus `CKnee`, `CNeck`, and $X^\top\hat{\beta}_n$ in Figure 4 (b), (c), and (d), respectively. Figures 4 (b) and (c) indicate that the nonparametric model for knee circumference and neck circumference fits reasonably well. However, Figure 4 (d) shows a nonlinear trend between the residuals and $X^\top\hat{\beta}_n$, which casts suspicion on the model adequacy. To explore this further, we depict the estimated effects for ankle circumference (`CAnkle`) in Figure 4 (e) and bicep (extended) circumference (`CBiceps`) in Figure 4 (f). The evidence of nonlinear patterns indicates that a partially linear model is not adequate to fit this data set. As shown, the proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

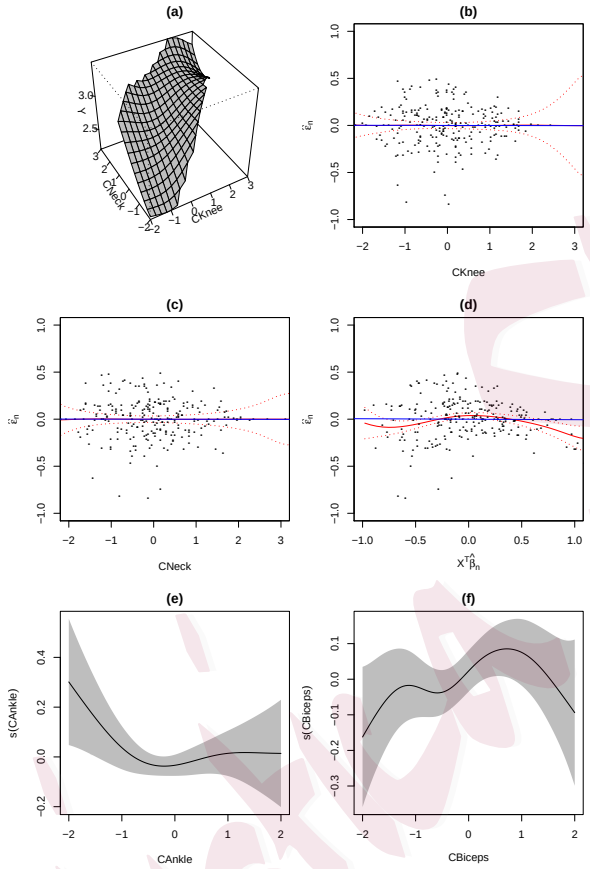

Figure 4: Results for the body fat data. The estimated surface of the nonparametric function $g(\texttt{CKnee}, \texttt{CNeck})$ (a). The estimated residuals $\hat{\varepsilon}_n$ versus `CKnee` (b), `CNeck` (c), and $X^\top \hat{\beta}_n$ (d) along the nonparametric estimated curves with 95% confidence bands. The estimated curves of log(fat) against ankle circumference (e) and bicep (extended) circumference (f) via the gam fitting.

test is more powerful than the tests T_n^s and T_n^{Xia}.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

7. Conclusion

We have proposed a projection-based method for checking the adequacy of a PLM. The method is consistent and reduces dimensionality, which may be of interest in dealing with high-dimensional observations. In summary, the proposed procedure is computationally expedient, theoretically reliable, intuitively appealing, and practically useful. We have shown both theoretically and numerically that the proposed procedure has advantages over the existing methods. However, note that we do not claim that the proposed method will always be best. Different circumstances may favor other methods, based on the assertion of Bierens and Ploberger (1997) that the aforementioned four weighting functions and the simple indicator function lead to asymptotic admissible tests. However, our overall numerical comparison suggests that the proposed procedure is very promising.

The proposed projection-based methodology is not limited to the PLM, but, in fact, is applicable to more general semiparametric models. While a theoretical investigation in this direction would be challenging,we believe the success of our projection-based method on the PLM makes further research warranted. Future studies should also examine the cases when (i) the number of covariates increases with the sample size, and (ii) the response variable is not continuous.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Supplementary Material

The online Supplementary Material provides proofs for Theorems 1 – 5 and additional simulation studies.

Acknowledgments

The authors would like to thank the two reviewers and the editor, Professor Müller, for their constructive comments and helpful suggestions. Sun's research was supported by the National Natural Science Foundation of China (Grant Nos. 11571340). Liang's research was partially supported by NSF grant DMS-1620898.

Appendix

In what follows, we denote $\widetilde{X} = X - g_1(T)$ and $\widetilde{X}_i = X_i - g_1(T_i)$.

A.1 Assumptions

We begin this section by giving the conditions needed in the proofs of the theorems.

(C1) The functions $g(t)$, $g_1(t) = \mathrm{E}(X|T = t)$ and $g_2(t) = \mathrm{E}(Y|T = t)$ are second-order continuously differentiable and satisfy Lipschitz condition of order 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

(C2) The matrix $\Sigma = \mathrm{E}(\widetilde{X}\widetilde{X}^{\top})$ is positively definite and $\sup_{x,t}\mathrm{E}(Y^2|X = x, T = t) < \infty$.

(C3) (i)The density of T, $f_t(t)$, exists and satisfies

$$0 < \inf_{t\in\mathbb{R}^1} f_t(t) \leq \sup_{t\in\mathbb{R}^1} f_t(t) < \infty;$$

(ii) $f_t(t)$ is second-order continuously differentiable.

(C4) The kernel function $K(\cdot)$ is a bounded kernel function of order 2 with bounded support.

(C5) The bandwidths satisfy $h_n \to 0$, $nh_n \to \infty$ and $nh_n^4 \to 0$ as $n \to \infty$.

Remark 1. Conditions (C1)and (C2) are necessary for the asymptotic normality of the model estimating procedure. Condition (C3) aims at avoiding tedious proofs of the theorems. Conditions (C4)–(C5) are generally needed to obtain the convergence rates of the nonparametric estimates.

A.2 Estimation of β and $g(t)$

Let $S_j(t,h) = 1/n\sum_{i=1}^n (T_i - t)^j K_h(t - T_i)$, $\quad j = 0, 1, 2$, with $K(\cdot)$ being a kernel function, h_n a bandwidth sequence and $K_h(t) = 1/h_n K_h(t/h_n)$. We first estimate the function $g_1(t)$ and $g_2(t)$ by the local linear method:

$$\hat{g}_{1n}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\{S_2(t,h) - S_1(t,h)(T_i - t)\}K_h(t - T_i)X_i}{S_0(t,h)S_2(t,h) - S_1^2(t,h)},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

$$\hat{g}_{2n}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\{S_2(t,h) - S_1(t,h)(T_i - t)\}K_h(t - T_i)Y_i}{S_0(t,h)S_2(t,h) - S_1^2(t,h)}.$$

Then we can estimate β and $g(t)$ as follows:

$$\hat{\beta}_n = \left[\sum_{i=1}^{n}\{X_i - \hat{g}_{1n}(T_i)\}\{X_i - \hat{g}_{1n}(T_i)\}^\top\right]^{-1} \sum_{i=1}^{n}\{X_i - \hat{g}_{1n}(T_i)\}\{Y_i - \hat{g}_{2n}(T_i)\}$$

and

$$\hat{g}_n(t) = \hat{g}_{2n}(t) - \hat{g}_{1n}(t)^\top \hat{\beta}_n.$$

Therefore we can estimate the model error ε for the ith subject by $\hat{\varepsilon}(U_i, \hat{\beta}_n, \hat{g}_n(T_i))$ $= Y_i - \{X^\top \hat{\beta}_n + \hat{g}_n(T_i)\}$.

References

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics 20*, 105–134.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica 58*, 1443–1458.

Bierens, H. J. and W. Ploberger (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica 65*, 1129–1151.

Engle, R. F., C. W. J. Granger, J. Rice, and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association 81*(394), 310–320.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory 22*, 1030–1051.

Fan, Y. and Q. Li (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica 64*, 865–890.

Härdle, W., H. Liang, and J. Gao (2000). *Partially Linear Models*. Heidelberg: Physica-Verlag.

Härdle, W. and E. Mammen (1993). Testing parametric versus nonparametric regression. *Annals of Statistics 21*, 1926–1947.

Heckman, N. E. (1986). Spline smoothing in partly linear models. *Journal of the Royal Statistical Society, Series B 48*, 244–248.

Lavergne, P. and V. Patilea (2008). Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics 143*, 103–122.

Lee, T.-H., H. White, and C. W. J. Granger (2001). Testing for neglected nonlinearity in time series models: a comparison of neural network methods and alternative tests. *Journal of Econometrics 56*, 208–229.

Li, Q. and S. Wang (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics 87*, 145–165.

Ma, S., J. Zhang, Z. Sun, and H. Liang (2014). Integrated conditional moment test for partially linear single index models incorporating dimension-reduction. *Electronic Journal of Statistics 8*, 523–542.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Ma, Y., J.-M. Chiou, and N. Wang (2006). Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika 93*(1), 75–84.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.

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

Stinchcombe, M. B. and H. White (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory 14*(03), 295–325.

Stute, W., W. L. Xu, and L. X. Zhu (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika 95*, 451–467.

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

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

Wahba, G. (1984). Cross validated spline methods for the estimation of multivariate functions from data on functionals. *in statistics: an Appraisal, Proc. 50th Anniversary Conf (eds H.A.David and H.T.David, Ames:Iowa State University Press*, 205–235.

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

Xia, Y., W. Li, H. Tong, and D. Zhang (2004). A goodness-of-fit test for single-index models.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0238

Statistica Sinica 14(1), 1–28.

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

Zhu, L. X. and K. W. Ng (2003). Checking the adequacy of a partial linear model. *Statistica Sinica 13*, 763–781.

School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China

E-mail: (sunzh@ucas.ac.cn)

Center for Statistics and Data Science, Beijing Normal University at Zhuhai, Zhuhai 519087, China

E-mail: (chenfeifei12@mails.ucas.ac.cn)

Department of Statistics, George Washington University, Washington, D.C. 20052, USA

E-mail: (hliang@gwu.edu)

Department of Statistical Science, Cornell University, Ithaca, New York 14853, USA

E-mail: (dr24@cornell.edu)