Accepted Manuscript

Estimation of partially linear regression models under the partial consistency property

Xia Cui, Ying Lu, Heng Peng

PII: S0167-9473(17)30099-3
DOI: http://dx.doi.org/10.1016/j.csda.2017.05.004
Reference: COMSTA 6464

To appear in: *Computational Statistics and Data Analysis*

Received date : 11 July 2014
Revised date : 12 May 2017
Accepted date : 15 May 2017

Estimation of Partially Linear Regression Models under the Partial Consistency Property

Xia Cui[a], Ying Lu[b,*], Heng Peng[c]

[a]School of Mathematics and Information Science, Guangzhou University, Guangzhou, China. Email: cuixia@gzhu.edu.cn
[b]Department of Applied Statistics, Social Science and Humanities, Steinhardt School of Culture, Education and Human Development,New York, USA. Email: ying.lu@nyu.edu
New York University, New York, USA
[c]Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong. Email: hpeng@math.hkbu.edu.hk

Abstract

Utilizing recent theoretical results in high dimensional statistical modeling, a flexible yet computationally simple approach is proposed to estimate the partially linear models. Motivated by the partial consistency phenomena, the nonparametric component in the partially linear model is modeled via incidental parameters and estimated by a simple local average over small partitions of the support of the nonparametric variables. The proposed least-squares based method seeks to strike a balance between computation burden and efficiency of the estimators while minimizing model bias. It is shown that given inconsistent estimators of the nonparametric component, square root-n consistent estimators of the parameters of the parametric component can be obtained with little loss in efficiency. Moreover, conditional on the parametric estimates, an optimal estimator of the nonparametric component can be obtained using classic nonparametric methods. The statistical inference problems regarding the parametric parameters and a two-population nonparametric testing problem regarding the nonparametric component are considered. The results show that the behavior of the test statistics is satisfactory. To assess the performance of the new method in comparison with other methods, three simulation studies are conducted and a real data set about risk factors of birth weights is analyzed.

[*]Correspondence to: 246 Greene Street, New York, NY 10003. Tel: 1-212-998-5560.

Keywords: Partially linear model, partial consistency, high correlation, categorical data, asymptotic normality, nonparametric testing *AMS 2001 subject classification*: 62J05, 62G08, 62G20

1. Introduction

In statistics, regression analysis is a family of important techniques that estimate the relationship between a continuous response variable Y and covariates X with dimension p, $E(Y|X) = f(X)$. Parametric regression models specify the regression function in terms of a small number of parameters. For example, in linear regression, a linear response surface $E(Y|X) = X^T\beta$ is assumed and determined by the $p \times 1$ vector of β.

Parametric methods are relatively easy to compute and are widely used in statistical practices because the parameters β can be naturally interpreted as the "effects of X on Y". However, the stringent requirement of linearity can increase the risk of model misspecification, which leads to invalid estimates. In contrast, nonparametric methods assume no predetermined functional form and $f(X)$ is estimated entirely using the information from the data.

Various kernel methods or smoothing techniques have been developed to estimate $f(X)$. In general, these methods use local information about $f(X)$ to blur the influence of noise at each data point. The bandwidth, h, determines the width of the local neighborhood and the kernel function determines the contribution of the data points in the neighborhood. The bandwidth h is essential to the nonparametric estimator $\hat{f}(X)$. Smoother estimates of $f(X)$ are produced as h increases and vice versa. As a special case, the local linear model reduces to linear regression when h spans the entire data set with a flat kernel. The choice of h is data-driven and can be computationally demanding as the dimension of X increases. Moreover, nonparametric estimation suffers the curse of dimensionality which requires the sample size to increase exponentially with the dimension of X. In addition, most kernel functions are designed for continuous variables, and it is not natural to incorporate categorical predictors. Hence a fully nonparametric approach is rarely useful to estimate the regression function with multiple covariates.

The partially linear model, as given in equation (1), is one of the most commonly used semi-parametric regression models.

$$Y_i = X_i^T\boldsymbol{\beta} + g(Z_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

This offers an appealing alternative in that it allows both parametric and nonparametric specifications in the regression function. In this model, the covariates are separated into parametric components $X_i = (X_{i1}, \ldots, X_{ip})^T$ and nonparametric components $Z_i = (Z_{i1}, \ldots, Z_{iq})^T$. The parametric part of the model can be interpreted as a linear model, while the nonparametric part frees the rest of the model from stringent structural assumptions. As a result, the estimates of β are also less affected by model bias.

This model has gained great popularity since it was first introduced by Engle, Granger, Rice, and Weiss (1986) and has been widely applied in economics, social and biological sciences. Engle, Granger, Rice, and Weiss (1986) and many others study the penalized least-squares method for partially linear regression models estimation. Robinson (1988) introduces a profile least squares estimator for $\boldsymbol{\beta}$ based on the Nadaraya-Watson kernel estimate of the unknown function $g(\cdot)$. Heckman (1986), Rice (1986), Chen (1988) and Speckman (1988) study the consistency properties of the estimate of $\boldsymbol{\beta}$ under different assumptions. Schick (1996) and Liang and Härdle (1997) extend the root n consistency and asymptotic results for the case of heteroscedasticity. For models with only specification of the first two moments, Severini and Staniswalis (1994) propose a quasi-likelihood estimation method. Härdle, Mammen, and Müller (1998) investigate nonparametric testing problem of the unknown function $g(\cdot)$. Among others, Härdle, Liang, and Gao (2000) provide a good comprehensive reference of the partially linear model.

Most of the above methods are based on the idea of first taking the conditional expectations given Z_i and then subtracting the conditional expectations in both sides of (1). This way, the function $g(\cdot)$ disappears,

$$Y_i - \mathsf{E}(Y_i|Z_i) = \{X_i - \mathsf{E}(X_i|Z_i)\}^T\boldsymbol{\beta} + \varepsilon_i,\ i = 1, \ldots, n. \tag{2}$$

If the conditional expectations were known, $\boldsymbol{\beta}$ could be readily estimated via regression techniques. In practice, the quantities $\mathsf{E}(Y|Z)$ and $\mathsf{E}(X|Z)$ are estimated via nonparametric methods. The estimation of these conditional expectations is very difficult when the dimension of Z_i is high. Without accurate and stable estimates of those conditional expectations, the estimates of $\boldsymbol{\beta}$ can be greatly affected. In fact, Robinson (1988), Andrews (1994) and Li (1996) obtain the root-n consistency of the estimator of $\boldsymbol{\beta}$ under an important bandwidth condition with respect to the nonparametric part: $\sqrt{n}\Big(h^4 + \frac{1}{nh^q}\Big) \to 0$. Clearly, this condition breaks down when $q > 3$.

To circumvent the curse of dimensionality, $g(Z)$ is often specified in terms of additive structure of one-dimensional nonparametric functions, $\sum_{j=1}^{q} g_j(Z_j)$. This is the so-called generalized additive model. In theory, if the specified additive structure corresponds to the underlying true model, every $g_j(\cdot)$ can be estimated with desired one-dimensional nonparametric precision, and $\boldsymbol{\beta}$ can be estimated efficiently with optimal convergence rate. But in practice, estimating multiple nonparametric functions is related to complicated bandwidth selection procedures, which increases computation complexity and makes the results unstable. Moreover, when variables $\{Z_j\}$ are highly correlated, the stability and accuracy of such additive structure in partially linear regression model is problematic (see Jiang, Fan, and Fan, 2010). Lastly, if the additive structure is misspecified, for example, when there are interactions between the nonparametric predictors Z, the model and the estimation of $\boldsymbol{\beta}$ will be biased.

In this paper, a simple least-squares based method is proposed to estimate the parametric component of model (1) without complicated nonparametric estimation. The basic idea is as follows. Since the value of $g(Z)$ at each point is only related to the local properties of $g(\cdot)$, a stepwise function can be used to approximate the function $g(Z)$. The local average approximation can be represented by a set of incidental parameters that are only related to finite local sample points belonging to the same step interval. When the length of step interval is small enough, the approximation error is small and can be ignored. The variance of these incidental parameter estimates is high due to small sample size and these estimates are not consistent. However, due to the partial consistency property (Neyman and Scott, 1948; Lancaster, 2000; Fan et al., 2005), the increases in variance for estimating these incidental parameters are expected to be integrated and have only limited impact on the estimation of the parametric components β in model (1). The parametric parameters $\boldsymbol{\beta}$ can then be estimated using profile least squares. Following the classic results about the partial consistency property (Fan et al., 2005), under moderate conditions, this estimator of $\boldsymbol{\beta}$ has optimal root-n consistency and is almost efficient. Moreover, given a good estimate of $\boldsymbol{\beta}$, an improved estimate of the nonparametric component $g(Z)$ can be obtained. Compared to the classic nonparametric approach, this method is not only easy to compute, it also readily incorporates covariates Z involving both continuous and categorical variables. The statistical inference problems regarding the parametric and nonparametric components are also investigated under the proposed estimating method. Two test statistics are proposed and their limiting distributions are examined.

The rest of the paper is organized as follows. In section 2, following a brief technical review of the partial consistency property, a new estimation method of the parametric component for the partially linear regression model is proposed. The consistency of the parameter estimates are shown when the nonparametric component consists of a univariate variable, one continuous and one categorical variable, or two highly correlated continuous variables. The inference methods of the partially linear regression model are discussed in Section 3. Numerical studies assessing the performance of the proposed method in comparison with existing methods are presented in Section 4. A real data example is analyzed in Section 5. Section 6 includes an in-depth discussion about the implications of the proposed method and further directions. Technical proofs are relegated to the Appendix.

2. Estimating partially linear regression model under partial consistency property

2.1. Review of partial consistent phenomenon

The partial consistency phenomena occurs when a statistical model contains nuisance parameters whose number grows with sample size; although the nuisance parameters themselves cannot be estimated consistently, the rest of the parameters sometimes can be. Neyman and Scott (1948) first studied this problem. Using their terminology, the nuisance parameters are "incidental" since each of them is only related to finite sample points, and the parameters that can be estimated consistently are "structural" because every sample point contains information about them. The partial consistency phenomena appear in mixed effect models, models for longitudinal data, and panel data in econometrics; see Lancaster (2000) etc. In one JASA discussion paper, Fan, Peng, and Huang (2005) formally study the theoretical properties of parameter estimators under partial consistency and apply the results to to microarray normalization. A very general form of regression model is considered,

$$\mathbf{Y}_n = \mathbf{B}_n\boldsymbol{\alpha}_n + \mathbf{Z}_n\boldsymbol{\beta} + \mathbf{M} + \boldsymbol{\epsilon}_n, \quad n = J \times I, \tag{3}$$

where $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^T$, $\mathbf{B}_n = \mathbf{I}_J \otimes \mathbf{1}_I$ is an $n \times J$ design matrix, I is assumed to be a constant and J grows with sample size n. $\mathbf{Z}_n$ is a $n \times d$ random matrix with d being the dimension of $\boldsymbol{\beta}$, $\mathbf{M} = (m(X_1), \ldots, m(X_n))$ is a nonparametric function, and $\boldsymbol{\epsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of i.i.d. errors.

In the above model, $\boldsymbol{\alpha}_n$ is a vector of incidental parameters as its dimension J increases with sample size, $\boldsymbol{\beta}$ and $\mathbf{M}$ are the structural parameters. Fan *et al.* (2005) show that $\boldsymbol{\beta}$ and M can be estimated consistently and nearly efficiently when the value of I is moderately large in the following result,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}(0, \frac{I}{I-1}\sigma^2\Sigma^{-1}),$$

where the factor $I/(I-1)$ is the cost of estimating the nuisance parameters $\boldsymbol{\alpha}_n$.

2.2. Estimating partially linear model under partial consistency

First the proposed strategy is applied to estimate a partially linear regression model with a uni-dimensional nonparametric component,

$$Y_i = X_i^T\boldsymbol{\beta} + g(Z_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{4}$$

where $g(\cdot)$ is an unknown function, $Z_i \in R^1$ is a continuous random variable, and other assumptions for the model are similar as those imposed on the model (3). Without loss of generality and for convenience of theoretical analysis, Z_i are assumed to be i.i.d random variables and follow $[0, 1]$ uniform distribution, and are sorted as $0 \le Z_1 \le Z_2 \ldots \le Z_n \le 1$ based on their realized values. Note that this condition is quite mild: If Z_i doesn't follow a $[0, 1]$ uniform distribution, one can consider a monotonic transformation $Z_i^* = F(Z_i), i = 1, 2, \ldots, n$ where $F(\cdot)$ is the distribution function of Z_i and Z_i^* follows a uniform distribution. In this case, one can simply investigate the proposed method based on Z_i^*.

The support of Z_i is then partitioned into $J = n/I$ sub-intervals such that the jth interval covers I different random variables with closely realized values from $z_{(j-1)I+1}$ to z_{jI}. If the density of Z_i is smooth enough, these sub-intervals should be narrow, the values of $g(\cdot)$ over the same sub-interval should be close, and $g(Z_{(j-1)I+1}) \approx g(Z_{(j-1)I+2}) \cdots \approx g(Z_{jI}) \approx \alpha_j$ where $\alpha_j = \frac{1}{I}\sum_{i=1}^{I} g(Z_{(j-1)I+i})$. In this case the nonparametric part of model (4) can be reformulated in terms of partially consistent observations and rewritten in the form of the model (3)

$$\mathbf{Y}_n = \mathbf{B}_n\boldsymbol{\alpha}_n + \mathbf{X}_n^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n^*, \quad n = J \times I, \tag{5}$$

with $\varepsilon_{(j-1)I+i}^* = \varepsilon_{(j-1)I+i} + [g(Z_{(j-1)I+i}) - \frac{1}{I}\sum_{i=1}^{I} g(Z_{(j-1)I+i})]$. It is easy to see that the second term in $\varepsilon_{(j-1)I+i}^*$ is the approximation error. Normally when

I is a small constant, it is of order $O(1/J)$ or $O(1/n)$, and much smaller than ε. Hence the approximation error can be ignored and it is expected that, similar to (3), $\boldsymbol{\beta}$ in the model (4) or (5) can be estimated almost efficiently even when $g(\cdot)$ in (4) is not estimated consistently.

Model (5) can be easily estimated by profile least squares,

$$\sum_{j=1}^{J}\sum_{i=1}^{I}(Y_{(j-1)I+i} - X_{(j-1)I+i}^{T}\boldsymbol{\beta} - \alpha_j)^2. \tag{6}$$

The estimates of $\boldsymbol{\beta}$ and α_j can be expressed as follows,

$$\begin{cases} \hat{\boldsymbol{\beta}} = \Big\{\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}^{T}\Big\}^{-1} \\ \qquad \times\Big\{\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}\{Y_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}Y_{(j-1)I+i}\}\Big\}, \\ \hat{\alpha}_j = \frac{1}{I}\sum_{i=1}^{I}\Big\{Y_{(j-1)I+i} - X_{(j-1)I+i}^{T}\hat{\boldsymbol{\beta}}\Big\}. \end{cases} \tag{7}$$

Under the model (4) or (5), the following theorem for the above profile least squares estimator of $\boldsymbol{\beta}$ is derived.

Theorem 1. *Under regularity conditions (a)—(d) in the Appendix, for the profile least squares estimator of $\boldsymbol{\beta}$ defined in (7),*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1}\sigma^2\Sigma^{-1}), \tag{8}$$

where $\Sigma = \mathsf{E}\Big[\{X - \mathsf{E}(X|Z)\}\{X - \mathsf{E}(X|Z)\}^T\Big]$.

Similar to the treatment of the least square estimator for linear regression models, and noting that the degrees of freedom of (5) is approximately $(I-1)/I\cdot n$, one can readily estimate the variance of $\hat{\boldsymbol{\beta}}$ using the sandwich formula as shown in (9).

$$\begin{aligned} \mathsf{Var}(\hat{\boldsymbol{\beta}}) \quad = \quad & \hat{\sigma}^2\Big\{\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\} \\ & \times\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}^{T}\Big\}^{-1}, \end{aligned} \tag{9}$$

where

$$\hat{\sigma}^2 = \frac{I}{I-1} \cdot \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{I} (Y_{(j-1)I+i} - X_{(j-1)I+i}\hat{\boldsymbol{\beta}} - \hat{\alpha}_j)^2.$$

Furthermore, $\hat{\beta}$ can be plugged back into equation (2) to obtain an updated nonparametric estimate of $g(Z)$ based on

$$Y_i^* = Y_i - X_i^T \hat{\beta},$$

using standard nonparametric techniques. Since $\hat{\beta}$ is a root-n consistent estimator of β, it is expected that the updated nonparametric estimator $\hat{g}(Z)$ will converge to $g(Z)$ at the optimal nonparametric convergence rate.

2.3. Extension to multivariate nonparametric $g(Z)$

Case I: This simple method of approximating one-dimensional function $g(Z_i)$ can be readily extended to the multivariate case when Z consists of one continuous variable and several categorical variables. Note that without loss of generality, multiple categorical variables can be expressed as one K-level categorical variable. Hence, a partially linear model

$$Y_i = X_i^T \boldsymbol{\beta} + g(Z_i^d, Z_i^c) + \varepsilon_i,\ i = 1, \ldots, n, \tag{10}$$

where $Z_i^c \in R^1$ and Z_i^d is a K-level categorical variable.

To approximate $g(Z^d, Z^c)$, the data is first split into K subsets given the categorical values of Z_i^d, then the kth $(0 \leq k \leq K)$ subset of the data will be further partitioned into sub-intervals of I data points with adjacent values of Z^c. Based on the partition, model (10) can still be written in the form of (5). The profile least squares as shown above can be used to estimate $\boldsymbol{\beta}$ and lead to the following corollary.

Corollary 1. *Under the model (10) and regularity conditions (a)—(e), for the profile least squares estimator of $\boldsymbol{\beta}$ defined in (7),*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1}\sigma^2 \Sigma^{-1}), \tag{11}$$

where $\Sigma = \mathsf{E}\left[\{X - \mathsf{E}(X|Z)\}\{X - \mathsf{E}(X|Z)\}^T\right]$.

Case II: The simple approximation can also be easily applied to continuous bivariate variable $Z = (Z_1, Z_2) \in R^2$. The partition is now over the bivariate support of Z. In the extreme case when the two components of $Z = (Z_1, Z_2)$ are independent from each other, the approximation error based on the partition is of order $o(1/\sqrt{n})$, the same as the model error. Hence in theory the root-n consistency of $\boldsymbol{\beta}$ can be established.

Below a corollary is outlined where the two components of Z are highly correlated so one only need to partition the support of Z according to one component. First it is assumed that

$$\Delta_{si} \equiv Z_{1i} - Z_{2i} \to 0, \quad i = 1, \cdots, n, \tag{12}$$

a similar condition as in Jiang, Fan, and Fan (2010). Under assumption (12) with $\Delta_{si} = o(1)$, it is sufficient to partition the observations into subintervals of I data points according to the order of $Z_{1i}, i = 1, \ldots, n$. If $g(\cdot)$ satisfies some regular smoothness conditions, given subinterval j, $g(\mathbf{Z}_{(j-1)I+i})$ can be approximated by local average, denoted by α_j. Again the model can be represented in the form of (5), which leads to Corollary 2.

Corollary 2. *Under the model (10) where Z_{1i} and Z_{2i} are highly correlated continuous variables and satisfy the condition (12), and the regularity conditions (a)—(d), for the profile least squares estimator of $\boldsymbol{\beta}$ defined in (7),*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1}\sigma^2\Sigma^{-1}), \tag{13}$$

where $\Sigma = \mathsf{E}\left[\{X - \mathsf{E}(X|Z)\}\{X - \mathsf{E}(X|Z)\}^T\right]$.

The results of the theorem and corollaries are similar to the results of Fan *et al.* (2005) except replacing the unconditional asymptotic covariance matrix of the estimate by the conditional covariance matrix. The proofs of the theorem and corollaries are deferred to the Appendix.

2.4. Choice of I

As shown in the theoretical results, the proposed estimate $\hat{\boldsymbol{\beta}}$ is root-n consistent with asymptotic variance $\frac{I}{I-1}\sigma^2\Sigma^{-1}$. The loss in efficiency is determined by a factor of $I/(I-1)$ and it can be controlled by specifying a desired I value. But in principle, I should not be too large in order to control the approximation error. In general the I value is suggested to be no more than $O(\log_2 n)$. For example, for a sample size of 400, I should be no more than 10 ($log_2(400) = 8.6439$). As a matter of fact, as the simulation examples will show (see next section), $I = 4$ or 5 is good enough for a wide range of sample sizes and various nonlinear forms.

For small to moderate sample sizes, to strike a balance between reducing model bias (prefer smaller I) and minimizing the impact of many incidental parameters (prefer larger I), one can consider model averaging. First, $\hat{\boldsymbol{\beta}}_I$ is estimated under a series of I ranging from 2 to $\log_2 n$. Based on the theoretical analysis, every estimate $\hat{\boldsymbol{\beta}}_I$ can be rewritten as $\hat{\boldsymbol{\beta}}_I = \boldsymbol{\beta} + \epsilon_I$ where ϵ_I is a noise vector with mean zero and covariance matrix $\frac{I}{I-1}\Sigma^{-1}$. An updated $\hat{\beta}_{\text{average}}$ can be calculated as follows,

$$\hat{\boldsymbol{\beta}}_{\text{average}} = \left(\sum_{i=1}^{K} \frac{I_i - 1}{I_i} \hat{\boldsymbol{\beta}}_{I_i} \right) \Bigg/ \left(\sum_{i=1}^{K} \frac{I_i - 1}{I_i} \right),$$

and it is expected to be more stable than a single $\hat{\beta}_I$. This proposed method does not require additional tuning parameter selection. As shown in the theoretical results, the impact of different values of I is gauged explicitly by the inflation factor $I/(I-1)$. This is distinctly different from classic nonparametric curve estimation methods. For nonparametric methods, the bandwidth h is a very important tuning parameter which needs to be determined by computationally intensive cross-validation methods. Hence, computationally the proposed method has a clear advantage.

Remark 1: $g(Z)$ is proposed to be approximated by simply averaging observations within the local neighborhood. This method to some extent resembles kernel methods with small bandwidth, or the I-nearest neighbor averages in nonparametric estimation. However, the proposed method does not require a kernel density function nor complicated bandwidth selection, so it can be viewed as a nonparametric method that is completely model free. Theorem 1 and the two corollaries demonstrate that the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ based on partially consistent estimation of $g(Z)$ is almost as efficient as the estimator of $\boldsymbol{\beta}$ based on classic method for partially linear model, while the latter requires a consistent estimates of $g(Z)$. Theorem 1 shows that the parametric estimates based simply on a naive approximation of $g(Z)$ can still obtain optimal root-n consistency. In the extreme case, the consistent estimate of $\boldsymbol{\beta}$ can be obtained when the number of observations per subinterval, I, is as small as 2. This proposed method can also be modified in various ways such as by considering overlapping partition and local average to improve the efficiency of the proposed estimator.

Remark 2: As the dimension of the continuous components of Z increases, Z can be ordered according to the first principle component of Z or other covariates (Fan and Huang, 2001). In practice, as shown by Cheng and Wu (2013), the high dimensional continuous random vector Z can often be represented by a low dimensional manifold. Hence it is expected that in situations when Z can be expressed in a low dimensional manifold, the partition of Z can be done within the manifold effectively and the results of this paper should still apply. Nevertheless, further investigations are needed to ascertain the necessary conditions for the generalization of the proposed method.

3. Statistical inference for partially linear regression model

3.1. Statistical inference for parametric component

In this section, the statistical inference problem with respect to the proposed estimator of $\boldsymbol{\beta}$ is investigated. The following testing problem for $\boldsymbol{\beta}$ is considered,

$$H_0^1 : A\boldsymbol{\beta} = 0, \quad \text{vs} \quad H_1^1 : A\boldsymbol{\beta} \neq 0, \tag{14}$$

where A is a $k \times p$ matrix. A profile likelihood ratio or profile least square ratio test statistic (Fan and Huang, 2005) will be defined and we will investigate whether this test statistic is almost efficient and has an easy-to-work limiting distribution.

Let $\hat{\boldsymbol{\beta}}_0$ be the estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\alpha}}_{n0}$ be the estimator of $\boldsymbol{\alpha}_n$ in (5) under the null hypothesis H_0^1. The residual sum of squares (RSS) under the null hypothesis is $\mathrm{RSS}_0 = n^{-1} \sum_{j=1}^{J} \sum_{i=1}^{I} \hat{\varepsilon}^2_{(j-1)I+i,0}$, where $\hat{\varepsilon}_{(j-1)I+i,0} = Y_{(j-1)I+i} - \hat{\alpha}_{j0} - X^T_{(j-1)I+i}\hat{\boldsymbol{\beta}}_0$. Similarly, let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\alpha}}_{n1}$ be the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_n$ in (5) under the alternative hypothesis. The RSS under H_1^1 is $\mathrm{RSS}_1 = n^{-1} \sum_{j=1}^{J} \sum_{i=1}^{I} \hat{\varepsilon}^2_{(j-1)I+i,1}$, where $\hat{\varepsilon}_{(j-1)I+i,1} = Y_{(j-1)I+i} - \hat{\alpha}_{j1} - X^T_{(j-1)I+i}\hat{\boldsymbol{\beta}}_1$. Following Fan and Huang (2005), a profile least-squares based test statistic is defined,

$$T_n^1 = (RSS_0 - RSS_1)/RSS_1 \tag{15}$$

For linear regression with normal errors, the same test statistic has been shown to have a Chi-square distribution (Fan and Huang, 2005). Under the regularity conditions and the null hypothesis, the theorem below prescribes the asymptotic distribution of T_n^1.

Theorem 2. *Under regularity conditions (a)—(e) in the Appendix, and given the profile least-squares estimator $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ defined above,*

$$\frac{I-1}{I} \cdot nT_n^1 \xrightarrow{\mathcal{L}} \chi_k^2 \quad as \quad n \to \infty. \tag{16}$$

The results in Theorem 2 demonstrate that classic hypothesis testing results can still be applied to the parametric component in (4) with partially consistent nonparametric component estimators. The constant $I/(I-1)$ is the price to be paid for introducing high dimensional nuisance parameters in the model.

In practice, for finite sample size, the following bootstrap procedure can be used to calculate the p-value of the proposed testing statistic T_n^1 under the null hypothesis.

Bootstrap algorithm for T_n^1

1. Generate the residuals $\{\hat{\varepsilon}^*_{(j-1)I+i}, j = 1, \cdots, J; i = 1, \cdots, I\}$ by uniformly resampling from $\{\hat{\varepsilon}_{(j-1)I+i,0}\}$, then centralize $\{\hat{\varepsilon}^*_{(j-1)I+i}\}$ to be mean zero.
2. Define the bootstrap sample under the null hypothesis:
$$Y^*_{(j-1)I+i} = X^T_{(j-1)I+i}\hat{\boldsymbol{\beta}}_0 + \hat{\alpha}_{j,0} + \hat{\varepsilon}^*_{(j-1)I+i}.$$
3. Calculate the bootstrap test statistic T_n^{1*} based on the bootstrap sampling sample
$\Big\{(Y^*_{(j-1)I+i}, X_{(j-1)I+i}, Z_{(j-1)I+i}), j = 1, \cdots, J;\ i = 1, \cdots, I\Big\}$.
4. Repeat steps 1-3 to obtain N replicates of bootstrap samples and compute $T_n^{1*,b}$ for each sample $b = 1, \ldots, N$. The p-value of the test can be calculated based on the relative frequency of the events $\{T_n^{1*,b} \geq T_n^1\}$.

3.2. Statistical inference for nonparametric component when categorical data are involved

Corollary 1 establishes the root-n consistency property of the parametric component $\boldsymbol{\beta}$ when the nonparametric component is of the form $Z_i =$

(Z_i^d, Z_i^c) where Z_i^d is a K-level categorical variable and Z_i^c is a continuous variable in R^1.

Given an almost efficient estimate $\hat{\boldsymbol{\beta}}$, $Y_i^* = Y_i - X_i^T \hat{\boldsymbol{\beta}} = g(Z_i^d, Z_i^c) + \varepsilon_i^*$. The nonparametric function $g(Z_i) = g(Z_i^d, Z_i^c)$ can be expressed in terms of a series of univariate functions conditioning on the values of Z_i^d, $g(Z_i^c | Z_i^d = k), k = 1, \ldots, K$. Each of these univariate functions can be estimated using kernel method based on the split data associated with different Z_i^c values. Those estimates can be defined as $\hat{g}(Z_i^c | Z_i^d = k)$. Naturally one likes to test the equivalence of these univariate functions.

Motivated by the real example in Section 5, the following testing problem when $K = 2$ is considered:

$$\begin{aligned} &H_0^2 : g(Z_i^c | Z_i^d = 0) = g(Z_i^c | Z_i^d = 1) \ \text{ almost everywhere,} \\ &H_1^2 : g(Z_i^c | Z_i^d = 0) \neq g(Z_i^c | Z_i^d = 1) \ \text{ on a set with positive measure.} \end{aligned} \tag{17}$$

The above testing problem resembles a two-population nonparametric testing problem. For such a testing problem, Racine, Hart, and Li (2006) suggest a quadratic distance test statistic. However, the quadratic distance statistics are not sensitive to the local changes. Based on L_∞ norm and the idea from Fan and Zhang (2000), the following statistic in the context of the partially linear model is proposed,

$$T_n^2 = (-2\log h)^{1/2} \left[\sup_{Z^c} \frac{|\hat{g}(Z^c | Z^d = 1) - \hat{g}(Z^c | Z^d = 0)|}{\sqrt{\widehat{\mathsf{Var}}\{\hat{g}(Z^c | Z^d = 1) - \hat{g}(Z^c | Z^d = 0)\}}} - d_n \right], \tag{18}$$

where h is the chosen bandwidth parameter when estimating $g(Z^c | Z^d)$ and

$$d_n = (-2\log h)^{1/2} + \frac{1}{(-2\log h)^{1/2}} \log\left\{ \frac{\int K'^2(t)\,dt}{4\pi \int K^2(t)\,dt} \right\},$$

with $K(\cdot)$ is a kernel function satisfying $\int K(t)\,dt = 1$ and $\int t^2 K(t)\,dt > 0$.

Notice that $\hat{g}(Z^c | Z^d = 1)$ and $\hat{g}(Z^c | Z^d = 0)$ are estimated by different samples, hence $\hat{g}(Z^c | Z^d = 1)$ and $\hat{g}(Z^c | Z^d = 0)$ can be assumed to be independent. The variance can be calculated as follows,

$$\mathsf{Var}\{\hat{g}(Z^c | Z^d = 1) - \hat{g}(Z^c | Z^d = 0)\} = \mathsf{Var}\{\hat{g}(Z^c | Z^d = 0)\} + \mathsf{Var}\{\hat{g}(Z^c | Z^d = 1)\},$$

where $\mathsf{Var}\{\hat{g}(Z^c | Z^d = 0)\}$ and $\mathsf{Var}\{\hat{g}(Z^c | Z^d = 1)\}$, which can be estimated using standard nonparametric procedures.

Given the level of the test, H_0^2 will be rejected when T_n^2 is greater than the critical value. In general, critical values can be determined by the asymptotical distribution of test statistic under the null hypothesis. However, for this kind of nonparametric testing problem the test statistic tends to converge to its asymptotic distribution very slowly (Racine et al., 2006). The best way to approximate the null hypothesis distribution for the above testing statistic is by bootstrapping. Following the idea of Racine, Hart, and Li (2006), a simple bootstrap procedure is used to approximate the null hypothesis distribution of T_n^2.

Bootstrap algorithm for T_n^2:

1. Randomly select Z_i^{d*} from $\{Z_i^d, i = 1, \cdots, n\}$ with replacement, and call $\{Y_i, X_i, Z_i^{d*}, Z_{i2}\}$ the bootstrap sample.

2. Use the bootstrap sample to compute the bootstrap statistic T_n^{2*}, which is the same as T_n^2 except that Z_i^c is replaced by Z_i^{c*} values.

3. Repeat steps 1 and 2 to obtain N replicates of bootstrap samples and $T_n^{2*,b}, b = 1, \ldots, N$. The p-values is based on the relative frequency of the event $\{T_n^{2*,b} \geq T_n^2\}$ in the replications of the bootstrap sampling.

The distribution of T_n^2 under H_0^2 is asymptotically approximated by the bootstrap distribution of T_n^{2*}. Now let $Q_{1-\alpha}(T_n^{2*})$ be the $(1-\alpha)$th quantile of the bootstrapped test statistic distribution, the empirical $(1-\alpha)$ confidence band for $\{\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\}$ can be constructed as follows,

$$\Big[\{\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\} - \Delta_\alpha(Z^c),\\ \hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\} + \Delta_\alpha(Z^c)\Big], \tag{19}$$

where

$$\Delta_\alpha(Z_2)\\ =\{d_n + Q_{1-\alpha}(T_n^{2*})(-2\log h)^{-1/2}\}\sqrt{\widehat{\mathsf{Var}}\{\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\}}.$$

4. Numerical studies

Three simulation examples are conducted to examine the effectiveness of the proposed estimation methods and testing procedures for the partially linear regression model. The first example is a simple partial linear regression model with a uni-dimensional nonparametric component. The second example involves highly correlated bivariate nonparametric components, while the third example involves nonparametric components that are mixed with one categorical and one continuous variable.

The average absolute estimation error, $\mathrm{ASE}(\hat{\beta}) = \sum_{l=1}^{p} |\hat{\beta}_l - \beta_l|$ is computed to assess estimation accuracy of the parametric components, The robust standard deviation estimate (RSD) of $\hat{\boldsymbol{\beta}}$ is calculated using $(Q_3 - Q_1)/1.349$ where Q_1 and Q_3 are the 25% and 75% percentiles, respectively. The limiting distributions of the test statistics T_n^1 and T_n^2 under the null hypothesis and the power curve of each test are simulated in order to understand the behavior of the proposed test statistics. The performance of the proposed estimation and inference methods at varying sample sizes and varying sizes of the subintervals, I, is examined and compared with alternative methods.

For comparison purposes, the simulation examples are also estimated using alternative methods with available R packages. Packages "gam" and "mgcv" are used to fit generalized additive model. In "gam", local linear fitting is used with a default bandwith value 0.5 (about 50 data points). In "mgcv" the optimal bandwith is selected via General Cross Validation method. Package "NP" is used to fit nonparametric regression and packge "locfit" is used for nonparametric curve fitting. The generalized cross validation method is used to select the optimal bandwidth whenever it is applicable.

Example 1. *Consider the following partially linear regression model*

$$Y_i = X_i^T \beta + g(Z_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

where $\beta = (1, 3, 0, 0, 0, 0)$ *and* $g(Z_i) = 3\sin(2Z_i) + 10\delta I(0 < Z_i < 0.1) + \delta I(Z_i \geq 0.1)$. (X_i, Z_i)*s are i.i.d. draws from a multivariate normal distribution with mean zero and the covariance matrix*

$$\begin{pmatrix} 1.0 & \rho & \cdots & \rho \\ \rho & 1.0 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1.0 \end{pmatrix}.$$

with $\rho = 0.5$. ε_i*s are i.i.d. and follow the standard normal distribution.* δ *is set to vary among* $0, 3$ *and* 6.

Figure 1 illustrates the nonparametric component of this model. It can be seen that the value of δ determines the jump in the nonparametric function. Classical nonparametric method does not estimate function with jump accurately, hence the estimate of $\boldsymbol{\beta}$ will be affected too.

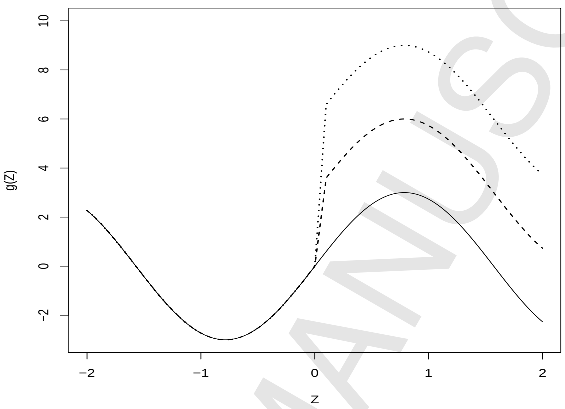

Figure 1: Simulation example 1. Plot of the function $g(Z_i)$: $\delta = 0$, Solid line; $\delta = 3$, dashed line; and $\delta = 6$, dotted line.

In this simulation example, 400 simulated samples are produced to evaluate the performance of the proposed estimators for the parametric components. The results are compared with those produced by function `gam` in R package "gam" that fits Generalized Additive Models. As suggested by Table 1, when I is set to be moderately large at 4 or 5, the ASEs of the proposed estimators are generally comparable with those are produced by the `gam` function in R. As sample size increases and when the nonparametric function is not smooth ($\delta \neq 0$), the ASEs of the proposed estimators are often smaller than the other methods. In these cases, the estimated standard errors of the ASE are also much smaller, suggesting that the proposed method produces more stable estimators than `gam`. This set of simulations also show that the improved estimators based on averaging results from different choices of I can be a good alternative, especially when sample size is relatively small.

In the extreme case when $I = 2$, the ASE decreases with sample size and it is only about 1.3 times that of function `gam,`suggesting the empirical model variance of the proposed method is about 1.7 times as large. Similar results in the other two numerical studies together suggest that although in theory the

Table 1: Average estimation errors for simulation example 1 (estimated standard errors in parentheses)

Method	Proposed Method						GAM
$\delta = 0$	I=2	I=4	I=5	I=10	I=20	Average	
n=100	0.977 (0.357)	0.781 (0.290)	0.800 (0.291)	0.846 (0.302)	1.075 (0.353)	0.784 (0.284)	0.723 (0.248)
n=200	0.650 (0.207)	0.538 (0.160)	0.528 (0.169)	0.516 (0.182)	0.583 (0.175)	0.496 (0.167)	0.507 (0.173)
n=400	0.470 (0.158)	0.374 (0.121)	0.361 (0.120)	0.349 (0.104)	0.346 (0.107)	0.345 (0.107)	0.355 (0.128)
$\delta = 3$							
n=100	0.957 (0.357)	0.834 (0.305)	0.815 (0.285)	0.904 (0.296)	1.225 (0.400)	0.821) (0.275)	0.789 (0.269)
n=200	0.676 (0.241)	0.539 (0.199)	0.509 (0.184)	0.518 (0.191)	0.601 (0.208)	0.495 (0.182)	0.543 (0.174)
n=400	0.460 (0.149)	0.380 (0.112)	0.365 (0.119)	0.349 (0.114)	0.350 (0.117)	0.344 (0.108)	0.401 (0.133)
$\delta = 6$							
n=100	1.006 (0.312)	0.852 (0.313)	0.847 (0.310)	1.017 (0.351)	1.420 (0.486)	0.855 (0.303)	0.934 (0.294)
n=200	0.624 (0.220)	0.530 (0.182)	0.527 (0.174)	0.544 (0.169)	0.687 (0.222)	0.517 (0.153)	0.625 (0.197)
n=400	0.458 (0.151)	0.377 (0.117)	0.363 (0.117)	0.359 (0.111)	0.379 (0.124)	0.348 (0.117)	0.465 (0.163)

proposed method has an efficiency loss by a factor of $I/(I-1)$, in practice the kernel based methods also suffer efficiency loss due to computational complexity that is not captured in theoretical results, therefore the empirical difference between the two methods is not as big as the theoretical results suggest.

Setting $\delta = 0$, the behavior of the proposed estimators in terms of efficiency and power is examined. In the left panel of Figure 2, the theoretical results given in Theorem 1 is nicely illustrated by a linear relationship between $\log(MSE) - \log I/(I-1)$ and the logarithm of the sample size. In the right panel of Figure 2, the empirical null distributions of the proposed test statistic T_n^1 testing

$$H_0^1 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0,$$

are compared with χ_4^2 (solid line). It can be seen that as sample size increases, the estimated density of T_n^1 converges to its asymptote given in Theorem 2.

It is worth noting that the estimation was carried out by setting the size of the subintervals I to be as small as 2.

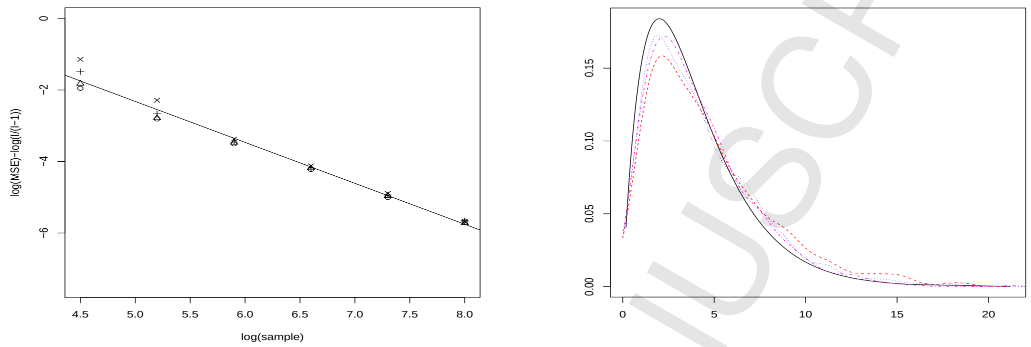

Figure 2: Simulation example 1. Left: Plots of MSEs of β: I=2 ($\circ$), I=5($\triangle$), I=10(+), I=20($\times$). The slope of the regression line between $\log(\text{MSE})-\log(I/(I-1))$ and log(Sample Size) is -1.12615. Right: Estimated density of the scaled test statistic $I/(I-1)nT_n^1$ for $n=100$ (long-dash), 200 (dot) and 400 (dot-dash) with the χ_4^2 distribution (solid) when $I=2$. The number of simulated samples is 1000.

The power of the proposed test statistic T_n^1 is examined by considering the alternative:

$$H_1^1 : \beta_3 = \Delta, \quad \beta_l = 0 \ \text{ for } \ l \geq 4.$$

Δ takes values from the set $(0,1)$. When $\Delta = 0$, the alternative hypothesis becomes the null hypothesis. To assess the bootstrap procedures proposed in section 3.1, 1000 bootstrap samples are generated and the p-value of the test for each simulated sample is calculated. Figure 3 illustrates the behavior of the power functions with respect to different Δ values and I values. Two sample sizes are considered, the left panel $n = 100$, and the right panel $n = 200$. Though the small value of I increases the variance of the estimator, the power of the test T_n^1 is not compromised. As shown in Figure 3, the power curves are similar for different values of I. The simulation results further confirm that the profile least squares test statistic T_n^1 is a useful tool for the linear testing problem in the partially linear regression model under partial consistency.

Example 2. *Consider the following generalized additive model,*

$$Y_i = X_i^\top \beta + g_1(Z_{1i}) + g_2(Z_{2i}) + g_3(Z_{3i}) + \varepsilon_i, \ i = 1, \ldots, n,$$

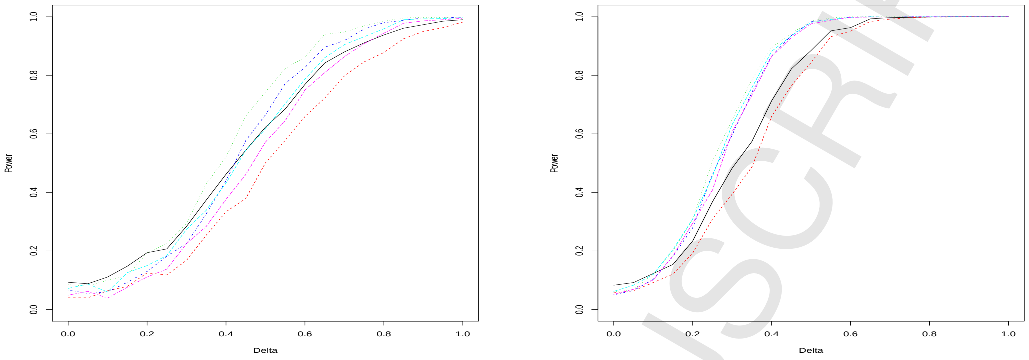

Figure 3: Simulation example 1. Left: Power of T_n^1 for $n = 100$, Right Power of T_n^1 for $n = 200$. Solid line (I=2), Dot line (I=5) and Long dash line (I=10) are power curves based on scaled $\chi^2(4)$ distribution. Short dash line (I=2), dot-Short dash line (I=5), dot-long dash line (I=10) are power curves based on the bootstrap algorithm for T_n^1.

where $\beta = (1.5, 0.3, 0, 0, 0, 0)^\top$. *The functions* g_1, g_2, g_3 *are:*

$$g_1(Z_{1i}) = -5\sin(2Z_{1i}), \quad g_2(Z_{2i}) = (Z_{2i})^2 - 2/3, \quad g_3(Z_{3i}) = Z_{3i}.$$

X_i*s follow a multivariate normal distribution with mean vector zero and the same covariance matrix as in Example 1.* Z_i*s are constructed to be highly correlated,*

$$\begin{aligned} Z_{1i} &= X_{1i} + u_{1i}, \\ Z_{2i} &= Z_{1i} + n^{-1/2}u_{2i}, \\ Z_{3i} &= Z_{1i} + n^{-1/2}u_{3i}, \end{aligned}$$

where n *is the sample size and* u_{is} *(*$s = 1, 2, 3$*) are* $N(0, 1)$ *disturbance terms that are drawn independently from covariates* X*. The correlation among* Z*s goes up as sample size increases. Lastly, the error term of the model* $\varepsilon_i \sim N(0, 1)$.

As in Example 1, 400 simulation examples are generated to evaluate the performance and running time of the proposed estimating method in comparison with the R function `gam`. As indicated by Table 2, as sample size increases, the proposed method outperforms the `gam` package even when $I = 2$. In general, one can see that the proposed method is not sensitive to the choice of I as long as it is not chosen to be too large a value relative to the sample size. Given a fixed sample size, larger I will yield smaller number of subintervals and lead to coarser approximation of the nonparametric function, but with shorter running time. Compared to `gam`,the computational efficiency of the proposed method is quite evident.

Table 2: Average estimation errors (estimated standard errors in parentheses) and running time (second) for simulation example 2

Method	Proposed Method						GAM
	I=2	I=4	I=5	I=10	I=20	Average	
ASE, n=100	1.372 (0.523)	1.343 (0.482)	1.358 (0.532)	1.606 (0.765)	2.364 (0.851)	1.352 (0.520)	1.123 (0.374)
Running Time	5.51	2.95	2.89	2.09	1.14	14.58	17.68
ASE, n=200	0.814 (0.332)	0.738 (0.279)	0.731 (0.277)	0.899 (0.408)	1.103 (0.403)	0.758 (0.282)	0.829 (0.285)
Running Time	10.27	5.73	4.87	2.77	2.02	25.66	26.36
ASE, n=400	0.511 (0.173)	0.459 (0.155)	0.465 (0.158)	0.504 (0.171)	0.620 (0.222)	0.449 (0.160)	0.563 (0.192)
Running Time	20.94	11.78	8.70	4.95	3.14	49.51	39.58

Moreover, one thousand simulation examples and the same number of bootstrap samples are used to study the properties of T_n^1 for the same testing problem as in Example 1. The left panel of Figure 4 presents the empirical null distribution of $(I-1)/InT_n^1$. Similar to in Simulation Example 1, the empirical null distribution is a reasonable approximation of the asymptotical null distribution χ_4^2. This is true for various values of I.

Compared with the results in Example 1, additional nonparametric components increase the estimation variability for the proposed method as well as for the method of GAM. ASE and standard errors are larger in Example 2. As shown in the right panel of Figure 4, the power of T_n^1 for the same testing problem also reduces. However, the proposed method is more robust to the high correlation situation as it is able to produce more efficient results than `gam` when sample size increases.

Example 3. *Consider the following model,*

$$Y_i = X_i^T \beta + g(Z_i^d, Z_2^c) + \varepsilon_i, \; i = 1, \ldots, n,$$

where

$$g(Z_i^d, Z_i^c) = (Z_i^c)^2 + 2Z_i^c + 2\delta Z_i^d e^{-16 Z_i^{c2}}.$$

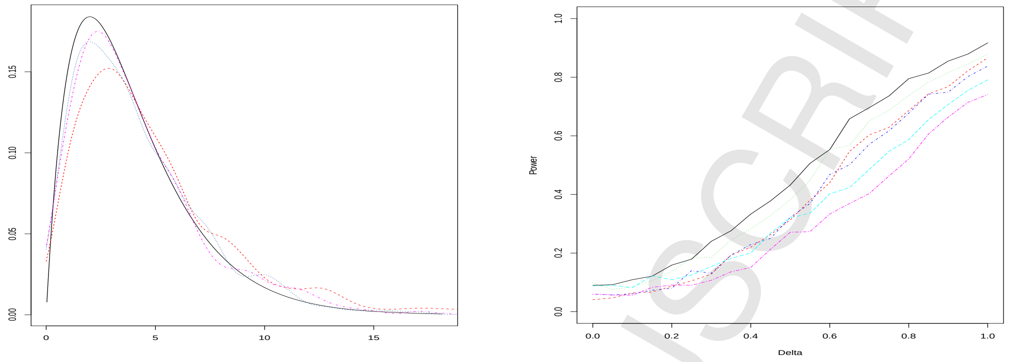

Figure 4: Simulation example 2. Left: Estimated density of the test statistic $I/(I-1)nT_n^1$ for $n = 100$ (long-dash) , $n = 200$ (dot) and $n = 400$ (dot-dash) with the χ_4^2 distribution (solid) when $I = 2$. Right: Power of T_n^1. Solid line (I=2), dot line (I=5) and long dash line (I=10) are power curves of T_n^1 based on χ_4^2 distribution. Short dash line (I=2), dot-short dash line (I=5), dot-long dash line (I=10) are power curves based on the bootstrap algorithm for T_n^1.

and $\beta = (3.5, 1.3, 0, \cdots, 0)^\top$. *The covariates* $X_i, i = 1, \ldots, n$ *are independently generated from Bernoulli distribution with equal probability being 0 or 1. The categorical variable* Z_i^d *is a Bernoulli variable independent of* X_i *with* $P(Z_i^d = 1) = 0.7$. *Variable* Z_i^c *is continuous and sampled from a uniform distribution on* $[-1, 1]$ *and is independent of* X_i *and* Z_i^d. *The error term* $\varepsilon_i \sim N(0, 0.2^2)$.

For the purpose of comparison, R package `np` is used to estimate the bivariate function $g(Z_i^d, Z_i^c)$ nonparametrically. In addition, package `mgcv` is used to estimate a “pseudo” model (MGCV1) with an additive nonparametric structure specified as below,

$$g(Z_i^d, Z_i^c) = \delta Z_i^d + g(Z_i^c) + \varepsilon_i, \ i = 1, \ldots, n,$$

and the semi-correct model (MGCV2)

$$g(Z_i^d, Z_i^c) = \delta Z_i^d + g_1(Z_i^c) + g_2(Z_i^c) Z_i^d + \varepsilon_i, \ i = 1, \ldots, n.$$

The true nonparametric components with $\delta = 0.25$ are plotted in the left panel of Figure 5.

Table 3 summarizes the simulation results comparing the performance of the proposed method under various choices of I, nonparametric method

("NP"), generalized linear models (two specifications–"MGCV1" and "MGCV2") in terms of running time, and estimation error (ASE) of the parametric parameter $\boldsymbol{\beta}$. To examine the impact of smoothness of the curves on estimation, three δ values, 0.25, 0.5 and 1.0 are examined.

First the nonparametric approach ("NP") attempts to optimize the estimation of both β and the bivariate function $g(Z_1, Z_2)$, therefore in general it is expected to have little model bias and therefore small estimation error. However, since it involves complicated tuning parameter selection, one may expect greater computational burden and longer running time. The two GAM models suffer different degrees of model misspecification. "MGCV2" attempts to fit a more flexible and also more complex form of nonparametric component than "MGCV1", hence it is expected to perform better than "MGCV1" in terms of estimation error but also with longer running time. In comparison, the proposed method can be viewed as a poor man's nonparametric method that aims to avoid model misspecifcation and computational complexity, at the cost of predetermined increase in estimation variance.

As shown in the Table 3, the proposed method demonstrates superb computational efficiency. At any I value, the proposed method is much faster than the `np` function. The running time of the proposed method only increases linearly with sample size and it is not sensitive to the lack of smoothness of the nonparametric part. In contrast, one can see that the computational burden for the `np` function increases geometrically and it also increases as the nonparametric part is less smooth (larger δ). On the other hand, the ASE of the proposed weighted average estimate is comparable with `np` especially when δ becomes large, while the proposed method based on a single I in general produces slightly bigger ASE than the `np` method but in a factor less than $I/(I-1)$. Compared to MGCV1, the proposed method is more robust to the misspefication of the models and has similar running time. Since MGCV2 mimics a unrealistic situation when the complex structure of the true nonparametric component is known, it produces smaller ASE than the proposed method but at the cost of longer running time.

The results pertaining to the comparison between NP and the two GAM models (MGCV1 and MGCV2) are surprising and interesting. Although both GAM models are considered to be misspeficied, they tend to outperform `np` results in terms of estimation error in β. Compared to either NP or the proposed approach, the GAM approach offers a more parsimonious representation of the nonparametric component that helps to ameliorate the cost of model misspecification and improve the performance of the estima-

Table 3: Average estimation errors (estimated standard errors in parentheses) and running time (second) for example 3

Method	Proposed Method				NP	MGCV1	MGCV2
$\delta = 0.25$	I=2	I=5	I=10	Average			
ASE, n=100	0.312 (0.099)	0.287 (0.091)	0.372 (0.115)	0.264 (0.085)	0.251 (0.100)	0.216 (0.070)	0.203 (0.067)
Running Time	4.68	2.21	1.30	10.67	245.12	3.19	12.75
ASE, n=200	0.203 (0.060)	0.165 (0.058)	0.183 (0.060)	0.163 (0.051)	0.157 (0.053)	0.151 (0.049)	0.143 (0.042)
Running Time	9.61	4.65	2.47	21.43	861.95	3.27	14.38
ASE, n=400	0.139 (0.044)	0.108 (0.032)	0.105 (0.032)	0.104 (0.032)	0.102 (0.030)	0.104 (0.031)	0.093 (0.030)
Running Time	19.62	7.96	3.88	40.92	3324.06	7.20	20.14
$\delta = 0.50$							
ASE, n=100	0.306 (0.096)	0.297 (0.106)	0.372 (0.119)	0.265 (0.096)	0.270 (0.100)	0.255 (0.073)	0.207 (0.064)
Running Time	4.87	1.97	1.29	10.74	256.36	3.09	12.68
ASE, n=200	0.193 (0.063)	0.165 (0.050)	0.181 (0.058)	0.157 (0.049)	0.157 (0.052)	0.171 (0.055)	0.135 (0.050)
Running Time	9.31	4.52	2.56	21.45	896.29	3.53	14.23
ASE, n=400	0.135 (0.041)	0.105 (0.031)	0.108 (0.029)	0.103 (0.029)	0.101 (0.031)	0.119 (0.039)	0.095 (0.030)
Running Time	19.40	8.29	3.92	41.34	3482.14	7.52	20.43
$\delta = 1.00$							
ASE, n=100	0.314 (0.112)	0.318 (0.104)	0.436 (0.139)	0.293 (0.095)	0.317 (0.136)	0.371 (0.132)	0.210 (0.071)
Running Time	5.06	2.33	1.28	11.14	279.61	2.82	12.68
ASE, n=200	0.204 (0.065)	0.173 (0.054)	0.204 (0.064)	0.164 (0.044)	0.167 (0.057)	0.247 (0.089)	0.137 (0.043)
Running Time	9.13	4.22	2.31	20.86	1007.14	3.21	14.52
ASE, n=400	0.139 (0.043)	0.107 (0.032)	0.109 (0.032)	0.106 (0.031)	0.107 (0.032)	0.173 (0.051)	0.097 (0.030)
Running Time	20.13	7.81	4.20	41.80	4050.64	6.71	19.56

tion of the parametric components. As shown in Figure 5 (left panel), the degree of misspecification is determined by δ. When δ is small, MGCV1 is reasonably close to the true nonparametric function. Only when the model is severely misspecified, does the NP method have an advantage over MGCV1.

On the other hand, MGCV2 approach uses the structure of the real model, so it can be regarded as an oracle situation which should be optimal with the smallest ASE. When the sample becomes large, the performance of the proposed method is close to the results of MGCV2, which is consistent to the theoretical results in Section 2. However, the running time for MGCV2 is considerably longer than MGCV1 and the proposed method with single I value.

Next the equivalence of the two nonparametric components associated with $Z^d = 0, 1$ is tested,

$$\begin{aligned} H_0^2 &: g(Z^c|Z^d = 0) = g(Z^c|Z^d = 1), \\ H_1^2 &: g(Z^c|Z^d = 0) \neq g(Z^c|Z^d = 1). \end{aligned}$$

In this simulation example, $g(Z_i^c|Z_i^d = 0) = (Z_i)^{c2} + 2Z_i^c$ and $g(Z_i^c|Z_i^d = 1) = (Z_i)^{c2} + 2Z_i^c + 2\delta \exp(-16(Z_i^{c2}))$. To explore the relationship between effect size and power of the proposed test statistic T_n^2, the value of δ is set to vary from 0 to 0.25.

To calculate T_n^2, $\hat{\boldsymbol{\beta}}$ is first estimated using equation (5), then it is removed from the model such that

$$Y_i^* = g(Z_i^c, Z_i^d) + \varepsilon_i^*, \quad i = 1, \ldots, n,$$

where $Y_i^* = Y_i - X_i\hat{\boldsymbol{\beta}}$ and $\varepsilon_i^* = \varepsilon_i + X_i\boldsymbol{\beta} - X_i\hat{\boldsymbol{\beta}}$. Next we use R package `locfit` to select a bandwidth h and actually use $0.8h$ to get slightly undersmoothed estimates of $\hat{g}(Z_i^c|Z_i^d = 0)$ and $\hat{g}(Z_i^c|Z_i^d = 1)$ and their variance estimates. The test statistic T_n^2 is calculated by plugging these estimators into formula (18). The p-values associated with T_n^2 are calculated using the bootstrap procedure suggested in Section 3. One thousand bootstrap samples are used to approximate the null distribution of T_n^2. This procedure is repeated 400 times to calculate the power of the test statistic under the alternative models defined by various δ values from 0 to 0.25.

The empirical distribution and bootstrapped distribution of T_n^2 under the null hypothesis when $\delta = 0$ and under the alternative hypothesis with different values of δ at 0.083, 0.167 and 0.25 are shown in the middle graph of Figure 5. It can be seen that the bootstrapped distributions under different alternative models provide fairly good approximations to the real null hypothesis distribution of the proposed test statistics. It suggests that the asymptotical null distribution of the proposed test statistics for the two-population nonparametric testing problem is a model free test statistic. In the right panel of Figure 5, the power curves of T_n^2 under various δ values and different sample sizes are shown. The estimates of $\boldsymbol{\beta}$ have little impact on the power curves and such impact is only through sample size. As a two-population nonparametric test, it is not too surprising to see that the power of this test is relatively low for small sample size. But as the sample size doubles, the power function picks up quickly even for small effect size when $\delta = 0.1$.

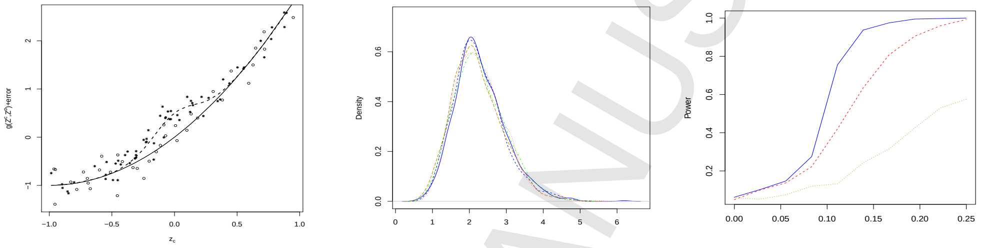

Figure 5: Simulation for example 3. Left: Scatterplot of $g(Z^c, Z^d)$+error vs Z_2 overlaid by solid-blue line: $g(Z^c, Z^d = 0)$ and dash-red line: $g(Z^c, Z_d = 1)$. Middle: Estimated density of the empirical and bootstrap null distribution of nT_n^2 for $n = 200$ and $I = 5$: solid blue line ($\delta = 0$) is the empirical null distribution. Dash-blue line ($\delta = 0$), dot-red line ($\delta = 0.083$), dot-dash green line ($\delta = 0.167$) and long-dash dark golden red line ($\delta = 0.25$) are bootstrapped estimation of null distribution. Right: the power function evaluated at $I = 5$ and different δ values with different sample sizes $n = 100$ (dot line), $n = 200$ (dash line) and $n = 400$ (solid line).

5. Real data application: correlates of birth weight

Low birth weight is an important biological indicator since it is associated with both birth defects and infant mortality. A woman's physical condition and behavior during pregnancy can greatly affect the birth weight of the newborn. In this section, the proposed methods are applied to a classic example studying the determinants of birth weights (Hosemer and Lemeshow, 2000). This dataset is part of a larger study conducted at Bay State Medical Center in Springfield, Massachusetts. The dataset contains variables (see below) that are believed to be associated with low birth weight in the obstetrical literature. The goal of the analysis is to determine whether these variables are risk factors in the clinical population being served by Bay State Medical Center.

- MOTH_AGE: Mother's age (years)
- MOTH_WT: Mother's weight (pounds)
- Black: Mother's race being black ('White' is the reference group)
- Other: Mother's race being other than black or white
- SMOKE: Mother's smoking status (1=Yes, 0=No)
- PRETERM: Any history of premature labor (1=Yes, 0=No)
- HYPER: History of hypertension (1=Yes, 0=No)
- URIN_IRR: History of urinary irritation (1=Yes, 0=No)
- PHYS_VIS: Number of physician visits
- BIRTH_WT Birth weight of newborn (grams)

First the data set is analyzed using a linear regression model to estimate the relationship between various factors and birth weight. Shown in Table 4 (OLS-1 model), mother's race (Black vs White, Other vs White), history of pregnancy hypertension and history of urinary irritation have significantly negative impact on birth weights of newborns, while mother's weight is positively related to birth weight. Perhaps surprisingly, mother's age is not a significant predictor of baby's birth weight(p-value=0.30). To check the linearity assumption, standardized residuals are plotted against each of the two continuous predictors, mother's age and weight. Left panel of Figure 6 shows that linearity is an adequate assumption for mother's weight and this relationship is not different between smokers and nonsmokers. But the residual diagnostics (graph not shown) indicate that the relationship between mother's age and birth weight is not linear and the relationship could potentially vary by mother's smoking status.

Then the analysis is expanded to 1) a linear regression with interaction term between age and smoking (the OLS-2 model), and 2) a generalized additive model (GAM) that specifies a nonparametric term with respect to mother's age. Under the OLS-2 model, the baseline age effect is insignificant.

Although the interaction term improves the model fit slightly, it is deemed insignificant (p-value=0.12). Under the GAM model, the nonparametric term of age is also tested insignificant (p-value=0.56). The conclusions about the effects of other variables on birth weights are similar compared to the OLS-1 model.

Table 4: Estimated effects of correlates of birth weight and their standard errors

	OLS-1	OLS-2	GAM	PL
Intercept	3026.9(308.2)	2741.9(357.0)	3044.2(309.0)	2482.0(388.2)
MOTH_WT	4.6(1.7)	4.5(1.7)	4.5(1.7)	5.6(2.0)
Black	-482.2(146.8)	-431.5(149.7)	-480.1(147.4)	-295.2(175.2)
Other	-327.5(112.6)	-302.2(113.3)	-320.1(112.9)	-203.6(132.8)
PRETERM	-179.5(133.8)	-169.8(133.4)	-166.4(134.2)	-220.0(153.5)
HYPER	-584.4(197.6)	-588.4(196.8)	-582.2(198.1)	-651.7(232.2)
URIN_IRR	-492.3(134.6)	-526.1(135.8)	-508.2(134.9)	-510.2(153.6)
PHYS_VIS	-7.0(45.4)	-0.7(45.4)	-12.2(45.5)	-14.7(52.8)
MOTH_AGE	-10.4(9.9)	1.8(12.6)	—(—)	—(—)
SMOKE	-312.5(104.5)	402.1(468.5)	-321.3(104.7)	—(—)
MOTH_AGE × SMOKE	—(—)	-30.6(19.6)	—(—)	—(—)
R^2	0.251	0.261	0.255	0.391

To model the nonlinear relationship between age and birth weight as well as its interaction with mother's smoking status, a partially linear model with a bivariate nonparametric components is specified as,

$$\begin{aligned}\text{BirthWT} &= \beta_0 + \beta_1\text{MOTH\_WT} + \beta_2\text{Black} + \beta_3\text{Other}\\ &\quad +\beta_4\text{PRETERM} + \beta_5\text{HYPER} + \beta_6\text{URIN\_IRR}\\ &\quad +\beta_7\text{PHYS\_VIS} + g(\text{MOTH\_AGE}, \text{SMOKE}) + \varepsilon.\end{aligned}$$

This model is estimated using the method proposed in Section 2.3. Since mother's age is recorded by a series of discrete values from 14 to 36 years, the support of $g(\text{MOTH\_AGE}, \text{SMOKE})$ is partitioned annually and according to mother's smoking status, the nonparametric response curve is then estimated for each group at every distinct age using available sample points (instead of using fixed cell size). The parameter estimates of the parametric components with standard errors are given in the last column of Table 4 (PL).

Given the parametric components, $g(\text{MOTH\_AGE}, \text{SMOKE})$ is estimated using the local polynomial regression methods via `locfit`.The fitted curves (after removing the parametric components) are shown in the right panel of Figure 6. This figure reveals that the response curves between age and birth weight are quite different for smoking and nonsmoking mothers. It can be seen that among non-smoking mothers, age is not particularly associated with birth weight. However, for smoking mothers, the birth weight decreases quite dramatically as mother's age increases. The gap is as wide as over 400 grams of birth weight between nonsmoking and smoking mothers who are 30 years and older. Similar as in the simulation studies, in the local polynomial regression (`locfit)`,a quadratic term is used and the optimal bandwidth is chosen via generalized cross validation.

The following one-sided nonparametric test is conducted to compare the two response curves between smokers and nonsmokers,

$$\begin{aligned} &H_0^2 : g(\text{MOTH\_AGE}, \text{Smoke}) = g(\text{MOTH\_AGE}, \text{Nonsmoke}), \\ &\quad \text{almost everywhere}, \\ &H_1^2 : g(\text{MOTH\_AGE}, \text{Smoke}) < g(\text{MOTH\_AGE}, \text{Nonsmoke}), \\ &\quad \text{on a set with positive measure}. \end{aligned}$$

Based on expression (16), the test statistic T_n^2 for the above test is 3.36, and the bootstrap p-value is 0.029, suggesting that the response curve of age among smokers is lower than that of non-smokers. This result and the right panel of Figure 6 together suggest that the PL model provides a better specification for the relationship between mother's age, smoking status and birth weight.

The estimates of the parameter components of the PL model also exhibit some interesting changes compared with other models. It can been that the racial gap in birth weight narrows. Controlling for other factors, on average babies born to Black mothers are 295 grams lighter than those born to White mothers. This difference is much smaller compared to the previous models. In addition, the effect of "Black" now is only marginally significant (p-value=0.1) and the effect of "Other" becomes insignificant (p-value=0.147). The effect sizes and significance values of other covariates remain about the same.

6. Discussion

In this paper, based on the concept of partial consistency, a simple estimation method for the partially linear regression model is proposed. The nonparametric component of the model is transformed into a set of artificially created nuisance or incidental parameters. Though these nuisance parameters cannot be estimated consistently, the parametric components of the partially linear model can be estimated consistently and almost efficiently

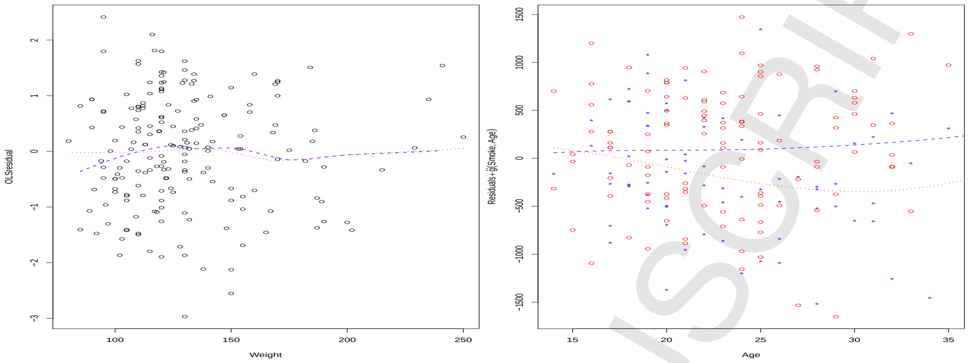

Figure 6: Correlates of birth weight. The left graph plots the residuals (OLS-2) vs mother's weight. The dotted red and dashed blue lines are the lowess fits for smoking mothers and non-smoking mothers, respectively. The right graph plots estimated regression function $g(\text{Age}, \text{Smoke})$ removing the effects of other covariates under the partial consistency PL model. The dotted red and dashed blue lines are the (lcofit) nonparametric estimates of response curves for smoking and non-smoking mothers, respectively.

under this configuration. As long as the sample size is reasonably large, the number of the nuisance parameters used is not too important. The estimation results have been shown to be fairly stable under various "coarseness" of the approximation. The statistical inference with respect to the parametric components via profile likelihood ratio test is also efficient. Generally speaking, the proposed simple estimation method for the partially linear regression model has two advantages that are worth noting. First, it greatly simplifies the computation burden in model estimation with little loss of efficiency. Second, it can be used to reduce the model bias by considering interaction between categorial predictor and continuous predictor, or between two continuous predictors in the nonparametric component of the model.

Though the partially linear regression model is a simple semiparameric model, the theoretical and simulation results offer interesting insights about the "bias-efficiency" tradeoff in semiparametric model estimations: when estimating the nonparametric components, pursing further bias reduction can increase the variance of nonparametric estimation, but it has little effect on the estimation of the parametric components of the model, and the efficiency loss in the parametric part is small. Given a much eased computational burden, such loss in efficiency in the parametric part can be negligible. This study raised an interesting problem in semiparametric estimation: how to balance between the computation burden and the efficiency of the estimators while minimizing model bias. These results can be generalized to estimate more broadly defined semiparametric models utilizing the partial consistency properties to fully exploit the information in the data.

Reference

Andrews, D. W. K., 1994. Asymptotic for semiparametric econometric models via stochastic equicontinuity. Econometrica 62, 43–72.

Chen, H., 1988. Convergence rates for parametric components. Ann. Statist. 16, 135–146.

Cheng, M.-Y., Wu, H.-T., 2013. Local linear regression on manifolds and its geometric interpretation. J. Am. Statist. Ass. 108 (504), 1421–1434.

Engle, R. F., Granger, C. W. J., Rice, J., Weiss, A., 1986. Semiparametric estimates for the relation between weather and electricity sales. J. Am. Statist. Ass. 81, 310–320.

Fan, J., Huang, L. S., 2001. Goodness-of-fit tests for parametric regression models. J. Am. Statist. Ass. 96, 640–652.

Fan, J., Peng, H., Huang, T., 2005. Semilinear high-dimensional model for normalization of microarray data: A theoretical analysis and partial consistency. J. Am. Statist. Ass. 100, 781–798.

Fan, J., Zhang, W., 2000. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. Scandinavian Journal of Statistics 27, 715–731.

Fan, J. Q., Huang, T., 2005. Profile likelihood inferences on semiparametric varying coefficient partially linear models. Bernoulli 11, 1031–1057.

Härdle, W., Liang, H., Gao, J., 2000. Partially Linear Models. Springer Verlag.

Härdle, W., Mammen, E., Müller, M., 1998. Testing parametric versus semiparametric modelling in generalized linear models. J. Am. Statist. Ass. 93, 1461–1474.

Heckman, N. E., 1986. Spline smoothing in a partly linear model. J. R. Statist. Soc. B. 48, 244–248.

Hosemer, D., Lemeshow, S., 2000. Applied Logistic Regression. New York: John Wiley & Sons Inc.

Hsing, T., Carroll, R. J., 1992. An asymptotic theory of sliced inverse regression. Ann. Statist. 20, 1040–1061.

Jiang, J., Fan, Y., Fan, J., 2010. Estimation in additive models with highly or non-highly correlated covariates. Ann. Statist. 38, 1403–1432.

Lancaster, T., 2000. The incidental parameter problem since 1948. Journal of Econometrics 95, 391–413.

Li, Q., 1996. On the root-n-consistent semiparametric estimation of partially linear models. Econ. Lett. 51, 277–285.

Liang, H., Härdle, W., 1997. Asymptotic properties of parametric estimation in partially linear heteroscedastic models. Sonder-forschungsbereich 373 Technical report no 33, Humboldt-Universität zu Berlin.

Neyman, J., Scott, E. L., 1948. Consistent estimates based on partially consistent observations. Econometrica 16 (1), 1–32.

Racine, J. S., Hart, J. D., Li, Q., 2006. Testing the significance of categorical predictor variables in nonparametric regression models. Economet Rev 25 (4), 523–544.

Rice, J., 1986. Convergence rates for partially linear spline models. Stat. Probabil. Lett. 4, 203–208.

Robinson, P. M., 1988. Root-n consistent semiparametric regression. Econometrica 56, 931–954.

Schick, A., 1996. Root-n consistent estimation in partly linear regression models. Stat. Probabil. Lett. 28, 353–358.

Severini, T. A., Staniswalis, J. G., 1994. Quasilikelihood estimation in semiparametric models. J. Am. Statist. Ass. 89, 501–511.

Speckman, P., 1988. Kernel smoothing in partial linear models. J. R. Statist. Soc. B. 50, 413–436.

Zhu, L. X., Ng, K. W., 1995. Asymptotics of sliced inverse regression. Stat Sinica, 727–736.

Appendix: assumptions and proofs

The following conditions to prove the theoretical results are assumed:

(a). $E|\varepsilon|^4 < \infty$ and $E\|X\|^4 < \infty$.

(b). The support of the continuous component of Z is bounded.

(c). The functions $g(z^d, z^c)$, $\mathsf{E}(X|Z^d = z^d, Z^c = z^c)$, the density function of Z, and their corresponding second derivatives with respect to z^c are all bounded.

(d). Σ is nonsingular.

(e). In presence of discrete covariate in Z, assume that for any category, the number of samples lies in this category is large enough and of order n.

For simplicity of presentation, we only discuss the case of $Z = Z^c$ and prove Theorem 1. When Z is of 2-dimension, we mainly consider that one component of Z is discrete or both components in Z are highly correlated. For the former case, according to condition (e) it can be concluded that each category has a sample size of order n. So categories do not affect the following proof which leads to the results of Corollary 1 . For the latter case, assumption (12) implies that the following proof can be easily generalized to obtain Corollary 2. The proofs for both Corollary 1 and Corollary 2 are therefore omitted here.

Proof of Theorem 1. First, based on standard operations in least squares estimation, we can obtain the decomposition $\sqrt{n}(\hat{\beta}-\beta) = R_1 + R_2$, where

$$\begin{aligned} R_1 = \Big\{ & \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I} X_{(j-1)I+i}\} \\ & \times \{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I} X_{(j-1)I+i}\}^{T}\Big\}^{-1} \\ & \times \Big\{\frac{1}{\sqrt{n}}\sum_{j=1}^{J}\sum_{j=1}^{J}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I} X_{(j-1)I+i}\} \\ & \times \{g(Z_{(j-1)I+i}) - \frac{1}{I}\sum_{i=1}^{I} g(Z_{(j-1)I+i})\}\Big\} \\ \equiv {} & R_1^N / R_1^D, \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} R_2 = \Big\{ & \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I} X_{(j-1)I+i}\} \\ & \times \{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I} X_{(j-1)I+i}\}^{T}\Big\}^{-1} \\ & \times \Big\{\frac{1}{\sqrt{n}}\sum_{j=1}^{J}\sum_{j=1}^{J}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I} X_{(j-1)I+i}\} \\ & \times \{\varepsilon_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I} \varepsilon_{(j-1)I+i}\}\Big\} \\ \equiv {} & R_2^N / R_2^D. \end{aligned} \tag{A.2}$$

Hereby we will show that the term R_1 converges to zero in probability as $n \to \infty$ and the asymptotic distribution of R_2 is multivariate normal with zero mean vector and covariance matrix given in (13).

According to the form of R_1, we need to first analyze the numerator R_1^N and the denominator R_1^D respectively. Let $\mathcal{F}_n = \sigma\{Z_1, Z_2, \cdots, Z_n\}$ and observe that conditionally on $\mathcal{F}_n$, $X_{(j-1)I+i}$ are independent of each other. The following is a sketch.

We first analyze R_1^N. Denote $\mathsf{E}(X|Z=z)$ by $m(z)$ and $X - m(Z)$ by e, then

$$\begin{aligned}
R_1^N =& \frac{1}{\sqrt{n}} \sum_{j=1}^{J} \sum_{i=1}^{I} \{m(Z_{(j-1)I+i}) - \frac{1}{I}\sum_{i=1}^{I} m(Z_{(j-1)I+i})\} \\
& \qquad \times \{g(Z_{(j-1)I+i}) - \frac{1}{I}\sum_{i=1}^{I} g(Z_{(j-1)I+i})\} \\
& + \frac{1}{\sqrt{n}} \sum_{j=1}^{J} \sum_{i=1}^{I} \{e_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I} e_{(j-1)I+i}\} \qquad\qquad (A.3) \\
& \qquad \times \{g(Z_{(j-1)I+i}) - \frac{1}{I}\sum_{i=1}^{I} g(Z_{(j-1)I+i})\} \\
& = R_1^{N(1)} + R_1^{N(2)}.
\end{aligned}$$

Notice that $R_1^{N(1)}$ can be expressed using the following summations,

$$\begin{aligned}
R_1^{N(1)} = \frac{1}{\sqrt{n}I^2} \sum_{j=1}^{J} \sum_{i=1}^{I} \sum_{l=1}^{I} \sum_{k=1}^{I} & \{m(Z_{(j-1)I+i}) - m(Z_{(j-1)I+l})\} \\
& \times \{g(Z_{(j-1)I+i}) - g(Z_{(j-1)I+k})\}.
\end{aligned}$$

Parallel to the proof of Hsing and Carroll (1992) and Zhu and Ng (1995), we can show that

$$\begin{aligned}
R_1^{N(1)} \le& \frac{1}{\sqrt{n}I^2} \sqrt{\sum_{j=1}^{J} \sum_{i=1}^{I} \sum_{l=1}^{I} \sum_{k=1}^{I} \|m(Z_{(j-1)I+i}) - m(Z_{(j-1)I+l})\|^2} \\
& \times \sqrt{\sum_{j=1}^{J} \sum_{i=1}^{I} \sum_{l=1}^{I} \sum_{k=1}^{I} |g(Z_{(j-1)I+i}) - g(Z_{(j-1)I+k})|^2} \\
=& O_P(n^{-1/2} I^{-2} n^{\delta}) = o_P(1).
\end{aligned}$$

Here δ is a arbitrarily small positive constant. Let Ω_j denote the sample set lying in the jth partition with $1 \le j \le J$. The last equality obtained from the fact that, under condition (c), $m(\cdot)$ and $g(\cdot)$ have a total variation of order δ,

$$\lim_{n\to\infty}\frac{1}{n^{\delta}}\sup_{\{\Omega_j,1\le j\le J\}}\sum_{i=1}^{I-1}\|m(Z_{(j-1)I+i})-m(Z_{(j-1)I+(i+1)})\|=0,$$

and

$$\lim_{n\to\infty}\frac{1}{n^{\delta}}\sup_{\{\Omega_j,1\le j\le J\}}\sum_{i=1}^{I-1}|g(Z_{(j-1)I+i})-g(Z_{(j-1)I+(i+1)})|=0.$$

Next we consider $R_1^{N(2)}$. Let $\bar{e}_{(n)}$ and $\bar{e}_1$ be the largest and smallest of the corresponding e_i's, respectively. It is clear that

$$\begin{aligned}R_1^{N(2)}\le&\frac{\bar{e}_{(n)}-\bar{e}_1}{\sqrt{n}I}\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{l=1}^{I}|g(Z_{(j-1)I+i})-g(Z_{(j-1)I+l})|\\=&2\frac{\bar{e}_{(n)}-\bar{e}_1}{\sqrt{n}I}\sum_{j=1}^{J}\sum_{1\le i<l\le I}|g(Z_{(j-1)I+i})-g(Z_{(j-1)I+l})|.\end{aligned}$$

The above argument leads to that

$$\begin{aligned}R_1^{N(2)}\le&2\frac{\bar{e}_{(n)}-\bar{e}_1}{\sqrt{n}I}\sum_{i=1}^{I}\sum_{l=1}^{I}\sum_{j=1}^{n-1}|g(Z_{(j+1)})-g(Z_{(j)})|\\\le&2I\frac{\bar{e}_{(n)}-\bar{e}_1}{\sqrt{n}}\sum_{j=1}^{n-1}|g(Z_{(j+1)})-g(Z_{(j)})|.\end{aligned}$$

Applying Lemma A.1 of Hsing and Carroll (1992), we obtain

$$n^{-1/4}|\bar{e}_{(n)}-\bar{e}_1|\xrightarrow{P}0.$$

Note the fact that total variation of $g(\cdot)$ is of order n^{δ}, we have $R_1^{N(2)}=o_P(1)$. Combining the results about $R_1^{N(1)}$ and $R_1^{N(2)}$, the proof for R_1^N is completed.

Next consider R_1^D and R_2^D. Since $R_1^D = R_2^D$, we only need to show the case of R_1^D. The expectation of R_1^D is calculated as follows.

$$\begin{aligned}\mathsf{E}(R_1^D) =& \mathsf{E}\Big(XX^T - \frac{1}{nI}\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{l=1}^{I} E\{X_{(j-1)I+i}X_{(j-1)I+l}^T\}\Big)\\ =& \mathsf{E}\Big(XX^T - \frac{1}{nI}\sum_{j=1}^{J}\sum_{i=1}^{I} E\{X_{(j-1)I+i}X_{(j-1)I+i}^T\}\\ &- \frac{1}{nI}\sum_{j=1}^{J}\sum_{i\neq l} E\{X_{(j-1)I+i}X_{(j-1)I+l}^T\}\Big)\\ =&(1-\frac{1}{I})\,\mathsf{E}(XX^T) - \frac{1}{nI}\sum_{j=1}^{J}\sum_{i\neq l} E\Big[E\{X_{(j-1)I+i}X_{(j-1)I+l}^T|\mathcal{F}_n\}\Big].\end{aligned}$$

Under the assumption that conditionally on $\mathcal{F}_n$, $X_{(j-1)I+i}$ are independent of each other, we can obtain that

$$E\{X_{(j-1)I+i}X_{(j-1)I+l}|\mathcal{F}_n\} = m(Z_{(j-1)I+i})m(Z_{(j-1)I+l}).$$

This, together with the above analysis, gives

$$\begin{aligned}\mathsf{E}(R_1^D) =&(1-\frac{1}{I})\,\mathsf{E}(XX^T) - \frac{I-1}{nI}\sum_{j=1}^{J}\sum_{i=l}^{I} E\Big[m(Z_{(j-1)I+i})m(Z_{(j-1)I+i})\Big]\\ &- \frac{1}{nI}\sum_{j=1}^{J}\sum_{i\neq l} E\Big[m(Z_{(j-1)I+i})\{m(Z_{(j-1)I+l}) - m(Z_{(j-1)I+i})\}\Big]\\ =&(1-\frac{1}{I})\,\mathsf{E}(XX^T) - \frac{I-1}{nI}\sum_{j=1}^{J}\sum_{i=l}^{I} E\Big[m(Z_{(j-1)I+i})m(Z_{(j-1)I+i})\Big] + o(1)\\ =&(1-\frac{1}{I})E\Big[\{X-\mathsf{E}(X|Z)\}\{X-\mathsf{E}(X|Z)\}^T\Big] + o(1).\end{aligned}$$

The term of order $o(1)$ is obtained following a similar argument of Theorem 2.3 of Hsing and Carroll (1992). This completes the proof for R_1.

We now deal with the term R_2. Observe that given $\{(X_i, Z_i), i = 1, \cdots, n\}$, each term of $\{\varepsilon_{(j-1)I+i} - \frac{1}{J}\sum\limits_{j=1}^{J}\varepsilon_{(j-1)I+i}\}$ has mean zero and is independent of each other. Thus R_2 is asymptotically normal with mean zero. We will show that the limiting variance of R_2 is equal to the covariance matrix given in (13). That is,

$$\begin{aligned}\mathrm{Var}(R_2|\{X_i,Z_i\}) =&(R_2^D)^{-1}\mathrm{Var}(R_2^N|\{X_i,Z_i\})(R_2^D)^{-1}\\ =&\{\mathsf{E}(R_2^D)\}^{-1}\,\mathsf{E}\{\mathrm{Var}(R_2^N|\{X_i,Z_i\})\}\{\mathsf{E}(R_2^D)\}^{-1}+o_P(1),\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Var}(R_2^N|\{X_i,Z_i\})\\ =&\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i}-\frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}\{X_{(j-1)I+i}-\frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}^T\\ &\times\mathsf{E}\left[\{\varepsilon_{(j-1)I+i}-\frac{1}{I}\sum_{i=1}^{I}\varepsilon_{(j-1)I+i}\}^2\Big|\{X_i,Z_i\}\right]\\ =&\frac{\sigma^2}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i}-\frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}\{X_{(j-1)I+i}-\frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}^T\\ \xrightarrow{P}&\sigma^2(1-\frac{1}{I})\,\mathsf{E}\left[\{X-\mathsf{E}(X|Z)\}\{X-\mathsf{E}(X|Z)\}^T\right].\end{aligned}$$

Combining the last two equations, we complete the proof of Theorem 1. □

Proof of Theorem 2. First we show that $\mathrm{RSS}_1=\sigma^2\{1+o_P(1)\}$. By (7), we have

$$\begin{aligned}\mathrm{RSS}_1=&\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{Y_{(j-1)I+i}-\hat\alpha_{j1}-X_{(j-1)I+i}^T\hat{\boldsymbol\beta}_1\}^2\\ =&\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\Big[\{X_{(j-1)I+i}^T-\frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}^T\}(\boldsymbol\beta-\hat{\boldsymbol\beta}_1)\\ &\qquad+\{g(Z_{(j-1)I+i})-\frac{1}{I}\sum_{i=1}^{I}g(Z_{(j-1)I+i})\}\\ &\qquad+\{\varepsilon_{(j-1)I+i}-\frac{1}{I}\sum_{i=1}^{I}\varepsilon_{(j-1)I+i}\}\Big]^2\\ =&I_1+I_2+I_3+I_4+I_5+I_6,\end{aligned}$$

where

$$
\begin{aligned}
I_1 &= \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{\varepsilon_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}\varepsilon_{(j-1)I+i}\}^2,\\
I_2 &= \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}[\{X_{(j-1)I+i}^T - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}^T\}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)]^2,\\
I_3 &= \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{g(Z_{(j-1)I+i}) - \frac{1}{I}\sum_{i=1}^{I}g(Z_{(j-1)I+i})\}^2,\\
I_4 &= \frac{2}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{\varepsilon_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}\varepsilon_{(j-1)I+i}\}\{g(Z_{(j-1)I+i}) - \frac{1}{I}\sum_{i=1}^{I}g(Z_{(j-1)I+i})\},\\
I_5 &= \frac{2}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i}^T - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}^T\}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)\\
&\qquad\qquad \times \{g(Z_{(j-1)I+i}) - \frac{1}{I}\sum_{i=1}^{I}g(Z_{(j-1)I+i})\},
\end{aligned}
$$

and

$$
\begin{aligned}
I_6 = \frac{2}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}&\{X_{(j-1)I+i}^T - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}^T\}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)\\
&\times \{\varepsilon_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}\varepsilon_{(j-1)I+i}\}.
\end{aligned}
$$

Using the same arguments when analyzing R_1 and R_2, it can be shown that

$$
\begin{aligned}
&I_1 = \frac{I-1}{I}\sigma^2\{1 + o_P(1)\}, \qquad I_2 = O_P(n^{-1}), \qquad I_3 = o_P(n^{-1/2}),\\
&I_4 = o_P(n^{-1/4}), \qquad I_5 = o_P(n^{-3/4}), \quad \text{and} \quad I_6 = o_P(n^{-1/2}).
\end{aligned}
$$

Similarly, RSS_0 can be decomposed as

$$\begin{aligned}\mathrm{RSS}_0 &= \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{Y_{(j-1)I+i} - \hat{\alpha}_{j0} - X_{(j-1)I+i}^T\hat{\boldsymbol{\beta}}_0\}^2 \\ &= \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\Big[\{Y_{(j-1)I+i} - \hat{\alpha}_{j1} - X_{(j-1)I+i}^T\hat{\boldsymbol{\beta}}_1\} \\ &\qquad + \{X_{(j-1)I+i}^T - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}^T\}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)\Big]^2 \\ &= \mathrm{RSS}_1 + J_1 + J_2,\end{aligned}$$

with

$$J_1 = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\Big[\{X_{(j-1)I+i}^T - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}^T\}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)\Big]^2,$$

and

$$\begin{aligned}J_2 = \frac{2}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}&\{Y_{(j-1)I+i} - \hat{\alpha}_{j1} - X_{(j-1)I+i}^T\hat{\boldsymbol{\beta}}_1\} \\ &\times \{X_{(j-1)I+i}^T - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}^T\}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0).\end{aligned}$$

From the proof of Theorem 1, it holds that

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}\{X_{(j-1)I+i} - \frac{1}{I}\sum_{i=1}^{I}X_{(j-1)I+i}\}^T \\ &\xrightarrow{P} \frac{I-1}{I}\Sigma.\end{aligned}$$

Furthermore, the estimators for $\boldsymbol{\beta}$ under the null and alternative hypotheses then have the following relation

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_1 - \Sigma^{-1}A^T\{A\Sigma^{-1}A^T\}^{-1}A\hat{\boldsymbol{\beta}}_1 + o_P(\hat{\boldsymbol{\beta}}_1).$$

J_1 can then be written as

$$J_1 = \frac{I-1}{I}\hat{\boldsymbol{\beta}}_1^T A^T\{A\Sigma^{-1}A^T\}^{-1}A\hat{\boldsymbol{\beta}}_1 + o_P(\hat{\boldsymbol{\beta}}_1).$$

This, together with the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$ in Theorem 1 implies that under the null hypothesis $A\hat{\boldsymbol{\beta}}_1 \xrightarrow{\mathcal{L}} N(0, \sigma^2 \frac{I-1}{I} A\Sigma^{-1}A^T)$,

$$nJ_1 \xrightarrow{\mathcal{L}} \sigma^2\chi_k^2.$$

By some calculation, it can be shown that $J_2 = 0$. Thus,

$$n(\mathrm{RSS}_0 - \mathrm{RSS}_1) \xrightarrow{\mathcal{L}} \sigma^2\chi_k^2.$$

Then by Slutsky theorem,

$$nT_n^1 = n\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1} \xrightarrow{\mathcal{L}} \frac{I}{I-1}\chi_k^2. \qquad \square$$