JMLR: Workshop and Conference Proceedings 44 (2015) 90-102 NIPS 2015

The 1st International Workshop "Feature Extraction: Modern Questions and Challenges"

A Computationally Efficient Method for Estimating Semi-parametric Regression Functions

Xia Cui CUIXIA@GZHU.EDU.CN
School of Mathematics and Information Science, Guangzhou University, Guangzhou, China

Ying Lu YING.LU@NYU.EDU
Center for the Promotion of Research Involving Innovative Statistical Methodology,
Steinhardt School of Education, Culture and Human Development, New York University

Heng Peng HPENG@MATH.HKBU.EDU.HK
Department of Mathematics, Hong Kong Baptist University, Hong Kong

Editor: Dmitry Storcheus

Abstract

Bias reduction is an important condition for effective feature extraction. Utilizing recent theoretical results in high dimensional statistical modeling, we propose a model-free yet computationally simple approach to estimate the partially linear model $Y = X\beta + g(Z) + \varepsilon$. Based on partitioning the support of Z, a simple local average is used to approximate the response surface $g(Z)$. The model can be estimated via least squares and no tuning parameter is needed. The proposed method seeks to strike a balance between computation burden and efficiency of the estimators while minimizing model bias. The desired theoretical properties of the proposed estimators are established. Moreover, since the proposed method bypasses data-driven bandwith selection of traditional nonparametric methods, it avoids the further efficiency loss due to computation burden.

Keywords: Nonparametric estimation, computational efficiency, least squares estimation, low model dependency

AMS 2001 subject classification: 62J05, 62G08, 62G20

1. Introduction

Regression analysis is a family of important techniques that estimate the relationship between a continuous response variable Y and covariates X with dimension p, $Y = f(X) + \epsilon$. Parametric regression models such as linear regression are easy to estimate and interpret but the requirement of a strict functional form can increase the risk of model misspecification. In contrast, nonparametric methods such as kernel methods or smoothing techniques have been developed to estimate flexible form of $f(X)$. The estimation of such models requires a data driven bandwith parameter h that can be computationally demanding as the dimension of X increases. Moreover, A fully nonparametric approach is rarely useful as it suffers the curse of dimensionality which requires the sample size to increase exponentially with the dimension of X. Semi-parametric regression models such as the partially linear model

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

offer an appealing alternative. 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 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 (for examples, Robinson (1988) and Severini and Staniswalis (1994)). For more details, Härdle, Liang, and Gao (2000) provide a good comprehensive reference of the partially linear model.

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 β can be estimated efficiently with optimal convergent 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 β will be biased.

In this paper, we propose a simple least squares based method 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 simple stepwise function can be used to approximate the function $g(Z)$. Such local average approximation can be represented by a set of incidental parameters that are only related to finite local sample points falling within the same step interval. When the length of step interval is small enough, the approximation error can be ignored. The increasing of variance due to estimating those incidental parameters is expected to be integrated and its effect on the parametric vector estimate β in model (1) can be almost ignored.

2. The method

The key motivation behind this method is the partial consistency propertyLancaster (2000) and Fan, Peng, and Huang (2005)) First consider a partially linear regression model with one-dimensional nonparametric component,

$$Y_i = X_i^T \beta + g(Z_i) + \varepsilon_i, \ i = 1, \ldots, n, \tag{2}$$

where $g(\cdot)$ is an unknown function, $Z_i \in R^1$ is a continuous random variable. Without loss of generality and for convenience of theoretical analysis, we assume that Z_i are i.i.d random variables and follow $[0,1]$ uniform distribution, and is sorted as $0 \leq Z_1 \leq Z_2 \ldots \leq Z_n \leq 1$ based on their realized values.[1]

Next we can partition the support of Z_i 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 and 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 $\alpha_j = \frac{1}{I}\sum_{i=1}^{I} g(Z_{(j-1)I+i})$. Then the nonparametric part of model (2) can be reformulated in terms of partially consistent observations and rewritten in the the following form,

$$\mathbf{Y}_n = \mathbf{B}_n \alpha_n + \mathbf{X}_n^T \beta + \varepsilon_n^*, \quad n = J \times I \tag{3}$$

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 β in the model (2) or (3) can be estimated almost efficiently even when $g(\cdot)$ in (2) is not estimated consistently. Model (3) can be easily estimated by profile least squares,

$$\sum_{j=1}^{J}\sum_{i=1}^{I}(Y_{(j-1)I+i} - X^T_{(j-1)I+i}\beta - \alpha_j)^2. \tag{4}$$

We have the following theorem for the above profile least squares estimate of β under the model (2) or (3).

Theorem 1 *Under regularity conditions (a)—(d) in the Appendix, for the profile least squares estimator of β,*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1}\sigma^2\Sigma^{-1}), \tag{5}$$

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

Similar to the treatment of least square estimator for linear regression models, and noting that the degrees of freedom of (3) is approximately $(I-1)/I \cdot n$, we can estimate the variance of $\hat{\beta}$ using sandwich formula. Furthermore, we can plug $\hat{\beta}$ back into equation (3) and obtain an updated nonparametric estimate of $g(Z)$ based on $Y_i^* = Y_i - X_i\hat{\beta}$ using standard nonparametric techniques. Since $\hat{\beta}$ is a root n consistent estimator of β, we expect the updated nonparametric estimator $\hat{g}(Z)$ will converge to $g(Z)$ at the optimal nonparametric convergence rate.

2.1 Extension to multivariate nonparametric $g(Z)$

Case I: The 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

[1] Note that this condition is indeed quite mild. If Z_i doesn't follow a $[0,1]$ uniform distribution, we 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, we can just investigate the proposed method based on Z_i*.

categorical variables. Note that without loss of generality, we can express multiple categorical variables as one K-level categorical variable. Hence, a partially linear model

$$Y_i = X_i\beta + g(Z_i^d, Z_i^c) + \varepsilon_i,\ i = 1, \dots, n, \tag{6}$$

where $Z_i = (Z_i^d, Z_i^c)$ where $Z_i^c \in R^1$ as specified in (2), Z_i^d is a N-level categorical variable.

To approximate $g(Z^d, Z^c)$ we first split the data into N subsets given the categorical values of Z_i^d, then the kth $(0 \le k \le N)$ 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 (6) can still be written in the form of (3). The profile least squares as shown above can be used to estimate β and we have the following corollary.

Corollary 1 *Under the model (6) and regularity conditions (a)—(e), for the profile least squares estimator of β,*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1}\sigma^2\Sigma^{-1}), \tag{7}$$

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 will need to be done 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 β can be established. Below we outline a corollary that based on the case when the two components of Z are highly correlated so we only need to partition the support of Z according to one component. First we assume

$$\Delta_{si} \equiv Z_{1i} - Z_{2i} \to 0, \quad i = 1, \cdots, n, \tag{8}$$

a similar condition as in Jiang, Fan, and Fan (2010)

Under the assumption (8) with $\Delta_{si} = o(1)$, it is sufficient to partition the observations into subintervals of I data points according to the order of $Z_{1i}, i = 1, \dots, n$. If $g(\cdot)$ satisfies some regular smoothness conditions, given subinterval j, $g(\mathbf{Z}_{(j-1)I+i})$ is approximately equal for $i = 1, \cdots, I$, denoted by α_j. Again the model can be represented in the form of (3) and we have another corollary,

Corollary 2 *Under the model (6) where Z_{1i} and Z_{2i} are highly correlated and satisfy the condition (8), and the regularity conditions (a)—(d), for the profile least squares estimator of β,*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1}\sigma^2\Sigma^{-1}), \tag{9}$$

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

The proofs of the theorem and corollaries are deferred to the Appendix. As the dimension of the continuous components of Z increases, similar as the discussion of Fan and Huang (2001) about ordering multivariate vector, Z can be ordered according to the first principle component of Z or certain covariate. In practice, as shown by Cheng and Wu

(2013) and Mohri et al. (2015), the high dimensional continuous random vector Z can often be represented by a low dimensional manifold. Hence we can expect that for many cases, once Z is expressed in a low dimensional manifold without losing much information, the partition of Z can be done within the manifold effectively and our results should still apply.

2.2 Choice of I

As shown in the theoretical results, our proposed estimate $\hat{\beta}$ is $\sqrt{n}$-consistent with the 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 desired I value. But in principle, I shouldn't be too large since we also need to control the approximation error. In general we suggest to choose an I value that is no more than $O(\log_2 n)$. For example, for sample size 400, I should be no more than 10 ($log_2(400) = 8.6439$). As a matter of fact, as our simulation examples show (see next section), $I = 4$ or 5 is good enough for a wide range of sample sizes and various nonlinear forms. In addition, 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), we can consider model averaging. This proposed method doesn't 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 $I/(I-1)$. In practice the optimal value of I need to be determined by cross-validation which is computationally intensive. Hence, computationally the proposed method has a clear advantage.

3. Numerical studies

We conduct three simulation examples to examine the effectiveness of the proposed estimation method. To assess estimation accuracy of the parametric components, we compute the average estimation errors, $\mathrm{ASE}(\hat{\beta}) = \sum_{l=1}^{p} |\hat{\beta}_l - \beta_l|$. For comparison purposes, all the simulations examples are also calculated using available R packages. Packages `gam` and `mgcv` are used to fit generalized additive model. package `NP` is used to fit nonparametric regression and package `locfit` is used for nonparametric curve fitting. Generalized cross validation method is used to select the optimal bandwidth whenever it is applicable.

Example 1 *Consider the following simple 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* $g(Z_i) = 3\sin(2Z_i) + 10\delta I(0 < Z_i < 0.1) + \delta I(Z_i \geq 0.1)$. $(X_i, Z_i), i = 1, \ldots, n$ *are i.i.d. draws from a multivariate normal distribution with mean zero and a covariance matrix where the variance of all the terms are 1 and the pair-wise correlation is* $\rho = 0.5$. $\varepsilon_i, i = 1, \ldots, n$ *are i.i.d. and follow the standard normal distribution.*

We let $\delta = 0, 3$ and 6 in this simulation. The size of δ determines the jump in the nonparametric function. Classical nonparametric method does not estimate functions with jump accurately hence the estimate of β will be affected too. 400 simulated samples are produced to evaluate the performance of the proposed estimators for the parametric components. The results will be 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 ASE of the proposed estimators is generally comparable with that is produced by the `gam` function in R. As sample size increases and when the nonparametric function is not smooth ($\delta \neq 0$), the ASE of the proposed estimators is often smaller. In which cases, the estimated standard errors of the ASE are also much smaller suggesting that our method produces more stable estimators than `gam`.We have also learned from this set of simulations 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`.This implies the empirical model variance of our method is about 1.7 times that of `gam` results. It suggests that in practice the optimal efficient kernel based methods also suffer efficiency loss due to computational complexity that is not captured in theoretical results.

Table 1: Average Estimation Errors for Simulation Example 1 (estimated standard errors in parentheses)

Method	Our 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)
$\delta = 3$							
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)
$\delta = 6$							
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)

Example 2 *Consider the following generalized additive model,*

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

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

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

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

$$Z_{1i} = X_{1i} + u_{1i}; \qquad Z_{2i} = Z_{1i} + n^{-1/2} u_{2i}; \qquad Z_{3i} = Z_{1i} + n^{-1/2} u_{3i} \tag{10}$$

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 estimation method in comparison with the R function `gam`. As indicated by Table 2, as sample size increases, our proposed method outperforms the `gam` package even when $I = 2$. In general, we can see that the proposed method is not

Table 2: The Average Estimation Errors and Running Time (second) for Simulation 2

Method	Our 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 (0.0374)
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

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 our method is quite evident.

Example 3 *The model is*

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

$$where \quad g(Z_i^d, Z_i^c) = (Z_i^c)^2 + 2Z_i^c + 0.25 Z_i^d e^{-16 Z_i^{c2}}.$$

and the true parameter β is a 6×1 vector and equals to $(3.5, 1.3, 0, \cdots, 0)^\top$. $X_i, i = 1, \ldots, n$ 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 $P(Z_i^d = 1) = 0.7$. The variable Z_i^c is continuous and sampled from a uniform distribution on $[-1, 1]$ and independent of X_i and Z_i^d. The error term $\varepsilon \sim N(0, 0.2^2)$.

For comparison purpose, we use R package `np` to estimate the bivariate function $g(Z_i^d, Z_i^c)$ nonparametrically. In addition, we also use package `gam` to estimate a "pseudo" model with an additive nonparametric structure specified as below,

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

We can see that the "pseudo" model misspecifies the nonparametric components. It will be interesting to compare the performance of the proposed method, generalized additive model and nonparametric method in terms of estimation of the parametric parameter β.

Again, we produced 400 samples for numerical comparison. Table 3 presents the ASE and estimates of β under three different methods. The `np` method tries to estimate β and the bivariate function $g(Z_1, Z_2)$ simultaneously which involves iterative algorithm and complicated tuning parameter selections. Hence we expect the numerical performance will be compromised to some extent. As the other two simulation studies suggested, our method in general produces slightly bigger ASE than the `np` method but in a factor less than $I/(I-1)$. On the other hand our method produces more precise estimates of β than the nonparametric approach. It is interesting that the GAM approach outperforms the nonparametric approach even under the wrong model specification.

4. Discussion

In this paper, based on the concept of partial consistency, we proposed a simple estimation method to partially linear regression model that has two worth noting advantages. First, it

Table 3: Fitting Results of ASE and Estimation of β for Example 3 based on the proposed method, NP and GAM (estimated standard errors in parentheses)

Method		Our Method				NP	GAM
n		I=2	I=5	I=10	I=20		
100	ASE	0.302 (0.104)	0.298(0.091)	0.367(0.118)	0.505(0.165)	0.254 (0.091)	0.217 (0.078)
	β_1	3.504(0.067)	3.502(0.063)	3.506(0.076)	3.498 (0.112)	3.464 (0.058)	3.499(0.047)
	β_2	1.305 (0.067)	1.294(0.065)	1.302(0.086)	1.310(0.100)	1.291(0.055)	1.302 (0.047)
200	ASE	0.197(0.064)	0.163(0.052)	0.187(0.059)	0.242(0.072)	0.153 (0.052)	0.149(0.048)
	β_1	3.502(0.045)	3.497 (0.033)	3.498(0.042)	3.504 (0.055)	3.486 (0.032)	3.499(0.030)
	β_2	1.300(0.041)	1.299(0.035)	1.303(0.039)	1.297 (0.054)	1.293(0.031)	1.299(0.032)

greatly simplifies the computation burden in model estimation with little loss of efficiency. Second, it can be used to reduce the model bias by allowing more generalized form of nonparametric components in the model, while bias reduction is an important condition for effective feature extraction.

Our results have offered us more 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 the efficient loss in the parametric part is small due to partial consistency property. Moreover, nonparametric estimation involving bandwidth selection can lead to some efficiency loss due to computation cost that is not outlined in the theoretical results. Our 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.

References

M.-Y Cheng and H.-T. Wu. Local linear regression on manifolds and its geometric interpretation. *Journal of the American Statistical Association*, 108(504):1421–1434, 2013.

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

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

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

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

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

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

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

Mehryar Mohri, Afshin Rostamizadeh, and Dmitry Storcheus. Foundations of coupled nonlinear dimensionality reduction. *arXiv preprint arXiv:1509.08880v2*, 2015.

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

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

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

Appendix A.

We need the following conditions to prove our theoretical results:

(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 (8) 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

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

and

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

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 (9).

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

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

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

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

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,

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

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.

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

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 $\{\varepsilon_{(j-1)I+i} - \frac{1}{J}\sum_{j=1}^{J}\varepsilon_{(j-1)I+i}\}$ 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 (9). 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. □