An Efficient Estimation Procedure for the Generalized Additive Models with Unknown Link Function [*]

Huazhen Lin Lixian Pan Shaogao Lv
Center of Statistical Research, School of Statistics
Southwestern University of Finance and Economics, China

Wenyang Zhang
Department of Mathematics
The University of York, UK

November 7, 2017

Abstract

The generalised additive models (GAM) are widely used in data analysis. In the application of the GAM, the link function involved is usually assumed to be a commonly used one without justification. Motivated by a real data example with binary response where the commonly used link function does not

[*] Lin's research was partially supported by National Natural Science Foundation of China (No.11571282 and 11528102) and Fundamental Research Funds for the Central Universities (No. JBK140507 and JBK120509) of China. Lv's research is supported partially by National Natural Science Foundation of China (Grant No.11301421). We sincerely thank Professor Ronghua Ruo for his suggestion and help in computation.

work, we propose a generalised additive models with unknown link function (GAMUL) for various types of data, including binary, continuous and ordinal. The proposed estimators are proved to be consistent and asymptotically normal. Semiparametric efficiency of the estimators is demonstrated in terms of their linear functionals. In addition, an iterative algorithm, where all estimators can be expressed explicitly as a linear function of Y, is proposed to overcome the computational hurdle for the GAM type model. Extensive simulation studies conducted in this paper show the proposed estimation procedure works very well. The proposed GAMUL are finally used to analyze a real data set about loan repayment in China, which leads to some interesting findings.

Key words and phrases: Generalized additive model; Local linear smoothing; Quasi-likelihood; Asymptotical properties; Semiparametric efficiency.

1 Introduction

The additive models and the generalised additive models (GAM) are widely used in data analysis to explore the nonlinear effects of the covariates on the response variable. Whilst getting away with the "curse of dimensionality", they are also reasonably flexible. The relevant literature includes Stone (1985, 1986), Breiman and Friedman (1985), Buja *et al.*(1989), Hastie and Tibshirani (1990), Linton and Nielsen (1995), Linton and Härdle (1996), Opsomer and Ruppert (1997), Fan *et al.*(1998), Mammen *et al.*(1999), Opsomer (2000), Linton (2000), Horowitz and Mammen (2004), Nielsen and Sperlich (2005), Mammen and Park (2006), Yu *et al.*(2008) and the reference therein.

When using GAM to fit a data set, people usually assume the link function is a commonly used one without much justification, e.g., the link function is assumed to be a logit function when the response variable is binary, a logarithmic function when the response variable is a count variable. However, in some real data analysis, the commonly used link function may not be appropriate, and the misspecification of the link function results in biased estimators for the component functions. The analysis of the real data set, which stimulates this paper, is the case. The data set is about the delay in loan repayment, and it comes from a financial company in China. For the sake of confidentiality, we call this financial company FCC in this paper. What FCC is interested in is if and how some factors affect the behavior of a customer's repaying his/her loan for buying a smart phone. The factors concerned are Tongdun score (a credit score), price of the mobile phone, loan amount, and income. The data set consists of the information of 2160 customers about the four factors and whether there is a delay in loan repayment (denoted by Y, $Y = 1$ if there is a delay, 0 otherwise). As there is no evidence showing the impacts of the factors concerned on Y are linear, given the sample size of 2160, which is quite decent, we apply GAM to fit the data. Rather than assuming the link function involved is logit function, as people usually do, we estimate the link function based on the data set using the methodology developed in this paper. The estimated link function is presented in Figure 3(e). Figure 3(e) shows the estimated link function is quite different to the commonly used logit function, which suggests the commonly used logit function is not appropriate for this data set. This example provides a compelling case for developing the generalised additive models with unknown link function, which is the purpose of this paper.

Investigating data driven approaches to specify the link function in the generalised linear type models has appeared in literature. See Aranda-Ordaz (1981), Scallan *et al.*(1984), Weisberg and Welsh (1994), Carroll *et al.*(1997), and the references therein for the generalised linear models with unknown link function. Zhang *et al.*(2015) considered the generalised varying-coefficient models with unknown link function. Let Y be the response variable, $\mathbf{X} = (X_1, \cdots, X_d)'$ the vector of covariates, Horowitz and Mammen (2007) proposed a general class of additive regression models, which can be written as

$$Y = m\left\{g_1(X_1) + \cdots + g_d(X_d)\right\} + \varepsilon, \quad i = 1, \cdots, n, \tag{1.1}$$

where $m(\cdot)$, $g_j(\cdot)$, $j = 1, \cdots, d$, are unknown function and ε is an unobserved random variable satisfying $E(\varepsilon|\mathbf{X}) = 0$. They also developed an estimation procedure for model (1.1) based on spline smoothing, and established optimal convergence rate for their estimators. Model (1.1) is an extension of the additive models, however, it is not very suitable for the cases where the response variable is categorical or discrete variable. The categorical or discrete response should also follow the respective distribution; any other predicted values are not logically possible. For example, a researcher may be interested in predicting one of three possible discrete outcomes. In this case, the dependent variable can only take 3 distinct values, and follows a multinomial distribution. The generalised type models provide a distributional framework for the response according to the type of the response. Therefore, there is difference between model (1.1) and the generalised additive model with unknown link function. Whilst the theoretical properties of the estimators proposed in Horowitz and Mammen (2007) are appealing, the implementation of their estimation procedure can be very difficult due to the high dimensional optimization involved, indeed, we found, in our simulation studies, the computational algorithm does not converge for quite a few cases.

In this paper, based on the framework of the generalised type models, we propose the generalised additive models with unknown link function (GAMUL), which is defined through

$$\begin{aligned} \mu = E(Y|\mathbf{X}) &= m\{g_1(X_1) + \cdots + g_d(X_d)\}, \\ Var(Y|\mathbf{X}) &= V(\mu), \end{aligned} \tag{1.2}$$

where $g_j(\cdot),\ j = 1,\ \cdots,\ d$, are unknown additive component functions, $m(\cdot)$ is the unknown link function, and $V(\cdot)$ is a known variance function and determined by the type of data. For example, for binary response, $V(\mu) = \mu(1-\mu)$.

The proposed GAMUL are more general than the generalised additive models on two aspects: (1) the link function is left unspecified in GAMUL, (2) we do not assume Y, given $\mathbf{X}$, follows the exponential family distribution. Zhang, Li and Xia (2015) also considered the generalised models with unknown link function, but under the varying-coefficient framework. Compared with the model addressed by Zhang, Li and Xia (2015), our model has several challenges. Firstly, the functional coefficients in the model in Zhang, Li and Xia (2015) share the same variable, therefore, all of them can be simultaneously locally linearized by one-dimensional kernel smoothing. While different components in our model have different variables, they cannot be simultaneously locally linearized by one-dimensional kernel smoothing. Hence, the local quasi-likelihood method proposed by Zhang, Li and Xia (2015) would suffer from the curse of dimensionality for our model. The difference between the model in Zhang, Li and Xia (2015) and our model is like the difference between the varying coefficient models and the additive models.

To avoid the curse of dimensionality, we use backfitting iterative method, the estimation of each curve at each point depends on all other functions, the theoretical properties of the proposed method are more involved than the local quasi-likelihood method since the proposed method is defined implicitly as the limit of a complicated iterative algorithm.

The proposed iterative estimation procedure is based on the backfitting quasi-likelihood idea. The advantage of the proposed estimation procedure is each step of the estimation procedure only involves one-dimensional smoothing and has a closed form, this has dramatically reduced the computational burden and difficulty in convergence. On theoretical front, we will show the proposed estimators are uniformly consistent, asymptotically normal, and enjoy the semiparametric efficiency, defined by Bickel *et al.*(1993), under some technical conditions. Particularly, compared with the estimation procedure in Horowitz and Mammen (2007) for model (1.1), the proposed estimation procedure is more easy to implement due to the closed form of our estimator, and also is more efficient due to the using of the quasi-likelihood, which utilizes the structure on the conditional variance.

The paper is organised as follows. We begin in Section 2 with a description of the proposed estimation procedure. The asymptotic properties of the proposed estimators are presented in Section 3. The performance of the proposed estimation procedure is also assessed by simulation studies in Section 4. In Section 5, we apply the proposed GAMUL together with the proposed estimation procedure to analyze the real data set, mentioned before, about loan repayment. The analysis reveals some quite interesting findings. A brief discussion about further research in this direction is made in Section 6. Technical proofs are relegated to the Appendix.

2 Estimation procedure

Apparently, models (1.2) are not identifiable. To make (1.2) identifiable, we assume

$$E\{g_j(X_j)\} = 0,\ \ j = 1,\ \cdots,\ d,\ \ \sum_{j=1}^{d} Var\{g_j(X_j)\} = 1,\ \text{ and } Cov(X_1, g_1(X_1)) > 0. \tag{2.1}$$

Following the proof of Proposition 3.1 in Horowitz and Mammen (2007), we can show models (1.2) are identifiable under Condition (2.1). Denote the support of X_j by A_j, we state this in the following proposition.

Proposition 1 *For continuously differentiable functions $m(\cdot)$, $g_j(\cdot), j = 1, \cdots, d$ with bounded support, we assume that the functions $g_j(\cdot)$ are nonconstant for at least two values of $j (1 \le j \le d)$, $\dot{m}(z) > 0$ for $z \in R$, $m\{g_1(x_1) + \cdots + g_d(x_d)\} = \check{m}\{\check{g}_1(x_1) + \cdots + \check{g}_d(x_d)\}$ for any $x_j \in A_j$, $1 \le j \le d$, and Condition (2.1) holds. Then $m(\cdot) \equiv \check{m}(\cdot)$ and $g_j(\cdot) \equiv \check{g}_j(\cdot)$ for $j = 1, \cdots, d$ on the corresponding supports.*

Let $(Y_i,\ \mathbf{X}_i)$, $i = 1,\ \cdots,\ n$, be an i.i.d. sample from $(Y,\ \mathbf{X})$. $(Y,\ \mathbf{X})$ follows the proposed models (1.2) and satisfies the identification condition (2.1), $\mathbf{g}(\cdot) = (g_1(\cdot),\ \cdots,\ g_d(\cdot))'$. In this section, we are going to build an estimation procedure for the unknown functions $m(\cdot)$ and $\mathbf{g}(\cdot)$ in the GAMUL.

Our estimation is quasi-likelihood based coupled with kernel smoothing. It is easy to see the log quasi-likelihood function of $m(\cdot)$ and $\mathbf{g}(\cdot)$ is

$$Q(\mathbf{g},\ m) = \sum_{i=1}^{n} L(\mu_i,\ Y_i) \tag{2.2}$$

with $L(\mu_i,\ Y_i)$ being defined through

$$\partial L(\mu_i,\ Y_i)/\partial \mu_i = V^{-1}(\mu_i)\Big\{Y_i - m\Big(\sum_{j=1}^{d} g_j(X_{ij})\Big)\Big\}.$$

The proposed estimation procedure is a backfitting procedure. To make the presentation more clear, we first present the estimation procedure for $m(\cdot)$ when $\mathbf{g}(\cdot)$ given, then the estimation procedure for $g_j(\cdot)$ when $m(\cdot)$ and $g_k(\cdot)$, $k \neq j$, are given. Finally, we present the proposed iterative algorithm.

2.1 Estimation of $m(\cdot)$ when $\mathbf{g}(\cdot)$ is given

By the Taylor's expansion, for any given u, we have

$$m\Big(\sum_{j=1}^{d} g_j(X_{ij})\Big) \approx m(u) + \dot{m}(u)\Big\{\sum_{j=1}^{d} g_j(X_{ij}) - u\Big\} \tag{2.3}$$

when $\sum_{j=1}^{d} g_j(X_{ij})$ is in a small neighbourhood of u. (2.3) together with (2.2) leads to that the quasi-likelihood estimator of $\mathbf{m} = (m_1,\ m_2)' \equiv (m(u),\ \dot{m}(u))'$ is the solution of the equation

$$S_m(\mathbf{m};\mathbf{g}) \hat{=} \sum_{i=1}^{n} \Big\{Y_i - W_i(u;\ \mathbf{g})'\mathbf{m}\Big\} \frac{W_i(u;\mathbf{g})}{V(\mu_i)} K_{h_m}\Big(\sum_{j=1}^{d} g_j(X_{ij}) - u\Big) = 0 \tag{2.4}$$

where $W_i(u;\ \mathbf{g}) = (1,\ \sum_{j=1}^{d} g_j(X_{ij}) - u)'$, and h_m is a bandwidth. Hence, the estimator for $(m(u),\ \dot{m}(u))'$ is

$$\begin{pmatrix} \hat{m}(u) \\ \hat{\dot{m}}(u) \end{pmatrix} = \Big[\sum_{i=1}^{n} W_i(u;\mathbf{g}) W_i(u;\mathbf{g})' K_{h_m}\Big\{\sum_{j=1}^{d} g_j(X_{ij}) - u\Big\}/V(\mu_i)\Big]^{-1} \\ \times \sum_{i=1}^{n} W_i(u;\mathbf{g}) K_{h_m}\Big\{\sum_{j=1}^{d} g_j(X_{ij}) - u\Big\} Y_i/V(\mu_i). \tag{2.5}$$

2.2 Estimation of $g_j(\cdot)$ when $m(\cdot)$ and $\{g_k(\cdot), k \neq j\}$ are given

Applying the Taylor's expansion to $g_j(\cdot)$, for any given x, we have

$$g_j(X_{ij}) \approx g_j(x) + \dot{g}_j(x)(X_{ij} - x)$$

when X_{ij} is in a small neighbourhood of x. This together with (2.2) leads to that the quasi-likelihood estimator of $\delta_j = (\zeta_j,\ \gamma_j)' = (g_j(x),\ \dot{g}_j(x))'$ is the solution of the equation

$$U_j(\delta_j; \mathbf{g}, m, \dot{m}) = 0 \tag{2.6}$$

where

$$U_j(\delta_j; \mathbf{g}, m, \dot{m}) \equiv \frac{1}{n}\sum_{i=1}^{n}\Big[Y_i - m\Big\{\sum_{k\neq j} g_k(X_{ik}) + \zeta_j + \gamma_j(X_{ij}-x)\Big\}\Big]\Upsilon_{ij}(x)$$
$$\times \dot{m}\Big\{\sum_{k=1}^{d} g_k(X_{ik})\Big\}K_{h_j}(X_{ij}-x)/V(\mu_i),$$

$\Upsilon_{ij}(x) = (1, X_{ij}-x)'$, and h_j is a bandwidth.

The nonlinearity of $m(\cdot)$ demands an iterative algorithm, such as Newton-Raphson iteration, in the calculation of $\delta_j = (g_j(x), \dot{g}(x))'$. This may lead to considerably intensive computation because the iteration has to be repeated over xs varying in the support of X_{ij} and js varying in $\{1, \cdots, d\}$ for each given $m(\cdot)$ and $\dot{m}(\cdot)$. To avoid such expensive computation, in this paper, we apply the Taylor's expansion to $m(\cdot)$ at $\sum_{k=1}^{d} g_k(X_{ik})$, which results in

$$\begin{aligned} & m\Big\{\sum_{k\neq j} g_k(X_{ik}) + \zeta_j + \gamma_j(X_{ij}-x)\Big\} \\ = \;& m\Big[\sum_{k=1}^{d} g_k(X_{ik}) - \Big\{g_j(X_{ij}) - \zeta_j - \gamma_j(X_{ij}-x)\Big\}\Big] \\ \approx \;& m\{\sum_{k=1}^{d} g_k(X_{ik})\} - \dot{m}\{\sum_{k=1}^{d} g_k(X_{ik})\}\Big\{g_j(X_{ij}) - \zeta_j - \gamma_j(X_{ij}-x)\Big\}. \end{aligned} \tag{2.7}$$

Plugging (2.7) into (2.6) and solving the equation, we have the estimator of $(g_j(x),\ \dot{g}_j(x))'$ as

$$\begin{pmatrix} \hat{g}_j(x) \\ \hat{\dot{g}}_j(x) \end{pmatrix} = \Big(\sum_{i=1}^{n} \rho^2(\mathbf{X}_i)\Upsilon_{ij}(x)\Upsilon_{ij}(x)' K_{h_j}(X_{ij}-x)/V(\mu_i)\Big)^{-1} \times$$
$$\sum_{i=1}^{n}\Big[Y_i - m\{\sum_{k=1}^{d} g_k(X_{ik})\} + \rho(\mathbf{X}_i)g_j(X_{ij})\Big]\Upsilon_{ij}(x)\rho(\mathbf{X}_i)K_{h_j}(X_{ij}-x)/V(\mu_i),$$

where $\rho(\mathbf{X}_i) = \dot{m}\{\sum_{k=1}^{d} g_k(X_{ik})\}$.

2.3 Iterative algorithm for $m(\cdot)$ and $\mathbf{g}(\cdot)$

To start the proposed iterative algorithm, we need an initial value. We have two choices. One is using the method proposed Horowitz and Mammen (2007, HM). Although the HM estimator ignores the heteroscedasticity and may be inefficient, the estimator is consistent. Our simualtion studies also show the HM serves well as an initial value. In addition, we also can start with an initial estimator by combining B-spline approximation for $m(\cdot)$ and the algorithm proposed by Hastie and Tibshirani (1990, p.141) if the HM estimator fails to convergy.

We are now ready to present the proposed iterative algorithm for $m(\cdot)$ and $\mathbf{g}(\cdot)$. Let $g_j^{(r-1)}(\cdot)$, $m^{(r-1)}(\cdot)$ and $\dot{m}^{(r-1)}(\cdot)$ be the estimators of $g_j(\cdot)$, $m(\cdot)$ and $\dot{m}(\cdot)$ obtained after the $(r-1)$th iteration, respectively, and

$$\mu_i^{(r-1)} = m^{(r-1)}\Big\{\sum_{j=1}^d g_j^{(r-1)}(X_{ij})\Big\}, \quad \rho^{(r-1)}(\mathbf{X}_i) = \dot{m}^{(r-1)}\{\sum_{j=1}^d g_j^{(r-1)}(X_{ij})\}$$

In the rth iteration, we update the estimators as follows

- For each given j, $j = 1, \cdots, d$, we apply the estimation procedure in Section 2.2 to estimate $g_j(\cdot)$ and standardise the obtained estimator. Specifically, let $X_{(j1)} < \cdots < X_{(j,d_j)}$ be the distinct points in $\{X_{ij} : i = 1, \ldots, n\}$. For each x, $x \in \{X_{(j1)}, \cdots, X_{(j,d_j)}\}$, we first estimate $g_j(x)$ and $\dot{g}_j(x)$ by

$$\begin{pmatrix} \hat{g}_j(x) \\ \hat{\dot{g}}_j(x) \end{pmatrix} = \Big(\sum_{i=1}^n \big\{\rho^{(r-1)}(\mathbf{X}_i)\big\}^2 \Upsilon_{ij}(x)\Upsilon_{ij}(x)' K_{h_j}(X_{ij}-x)/V(\mu_i^{(r-1)})\Big)^{-1}$$
$$\times \sum_{i=1}^n \Big(\Big[Y_i - \mu_i^{(r-1)} + \rho^{(r-1)}(\mathbf{X}_i) g_j^{(r-1)}(X_{ij})\Big]$$
$$\times \Upsilon_{ij}(x)\rho^{(r-1)}(\mathbf{X}_i) K_{h_j}(X_{ij}-x)/V(\mu_i^{(r-1)})\Big), \qquad (2.8)$$

  then standardise the obtained estimator. We use the standardised estimator $g_j^{(r)}(X_{ij})$ to update the estimator of $g_j(X_{ij})$ obtained in the $(r-1)$th iteration, where

$$g_j^{(r)}(X_{ij}) = \frac{\hat{g}_j(X_{ij}) - Avg(\hat{g}_j)}{\Big\{\sum_{j=1}^d Var(\hat{g}_j)\Big\}^{1/2}}, \quad i = 1, \ldots, n, \ j = 1, \ldots, d,$$

with

$$Avg(\hat{g}_j) = \frac{1}{n}\sum_{i=1}^{n} \hat{g}_j(X_{ij}), \quad Var(\hat{g}_j) = \frac{1}{n-1}\sum_{i=1}^{n} \left\{\hat{g}_j(X_{ij}) - Avg(\hat{g}_j)\right\}^2,$$

and the restriction of $\sum_{i=1}^{n} \{X_{i1} - Avg(X_1)\}\, g_1^{(r)}(X_{i1}) > 0$.

- We use the estimation in Section 2.1 to estimate $(m(\cdot),\ \dot{m}(\cdot))$. Specifically, let $U_{(1)} < \cdots < U_{(d_m)}$ be the distinct points in

  $$\Big\{ \sum_{j=1}^{d} g_j^{(r)}(X_{ij}) :\ i = 1,\ \cdots,\ n \Big\}.$$

  For each u, $u \in \{U_{(1)},\ \cdots,\ U_{(d_m)}\}$, we estimate $(m(u),\ \dot{m}(u))'$ by

  $$\begin{pmatrix} \hat{m}(u) \\ \hat{\dot{m}}(u) \end{pmatrix} = \Big[\sum_{i=1}^{n} W_i(u;\mathbf{g}^{(r)}) W_i(u;\mathbf{g}^{(r)})' K_{h_m}\Big\{\sum_{j=1}^{d} g_j^{(r)}(X_{ij}) - u\Big\} / V(\mu_i^{(r-1)})\Big]^{-1} \times \sum_{i=1}^{n} W_i(u;\mathbf{g}^{(r)}) K_{h_m}\Big\{\sum_{j=1}^{d} g_j^{(r)}(X_{ij}) - u\Big\} Y_i / V(\mu_i^{(r-1)}). \tag{2.9}$$

  Let $m^{(r)}(U_{(k)}) = \hat{m}(U_{(k)})$ and $\dot{m}^{(r)}(U_{(k)}) = \hat{\dot{m}}(U_{(k)})$, $k = 1,\ \cdots,\ d_m$, we use $(m^{(r)}(U_{(k)}),\ \dot{m}^{(r)}(U_{(k)}))$ to update the estimator of $(m(\cdot),\ \dot{m}(\cdot))$ obtained in $(r-1)$th iteration.

Repeat the iteration until convergence. In practice, the convergence is defined as $\sup_{j,x} |g_j^{(r)}(x) - g_j^{(r-1)}(x)| < a_0$ and $\sup_u |m^{(r)}(u) - m^{(r-1)}(u)| < a_0$, where a_0 is a prespecified small number.

Remark *The proposed iterative estimation procedure is easy to implement as there is a closed form at each step. It also converges very quickly because only one-dimensional smoothing is involved at each step. The local convergence of the algorithm is shown in Appendix B, provided that the link function $m(\cdot)$ is appropriately smooth. The proposed iterative estimation procedure also applies to the models (1.1), proposed in Horowitz and Mammen (2007), and it is more efficient than the estimation proposed there since the proposed method takes the heteroscedasticity into account and uses the backfitting procedure. The efficiency is confirmed by the extensive simulation studies* *in Section 4. Moreover, on theoretical ground, the proposed iterative estimation procedure is shown to be semiparametrically efficient when the conditional distribution is an exponential family distribution, see Theorem 3 in Section 3.*

3 Asymptotic properties

The proposed estimation procedure is defined implicitly as the solution to a complicated iterative algorithm, as Yu, Park and Mammen (2008) rightly pointed out, asymptotic properties for such estimation procedure are difficult to establish. Appealing some advanced techniques, we will show the proposed estimators are uniformly consistent and asymptotically normal. In addition to that, we will also show the proposed estimators enjoy the semiparametric efficiency defined by Bickel *et al.*(1993) when the conditional distribution of the response variable is an exponential family distribution. This underlines the combined advantages of quasi-likelihood idea and iterative backfitting algorithm.

Let $\mathbf{e}$ be a d-dimensional vector with elements 1, and f_j be the density function of X_{ij}, and $f_{\boldsymbol{\zeta}}$ be the density of the random variable $\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X})$ associated with $\boldsymbol{\zeta}$. The following assumptions are required for the asymptotic properties of the proposed estimators.

(A1) The kernel function $K(\cdot)$ is a symmetric density function with compact support $[-1,1]$ and a bounded derivative.

(A2) $\mathbf{X}_i$ is bounded with compact subset of $\mathbb{R}^d$. For notational simplicity, assume that $\mathbf{X}_i \in [-1,1]^d$.

(A3) The second derivatives of functions $g_j(\cdot), j = 1, \cdots, d$ and $m(\cdot)$ on $[-1,1]$ are bounded. Without the loss of generality, $\|g_j\|_\infty \leq 1, j = 1, \cdots, d$ and $\|m\|_\infty \leq 1$. The variance function $V(\cdot)$ is continuously differential and bounded away from zero on $[-1,1]$.

(A4) The conditional distribution of Y_i has subexponential tails. That is, there are constants C, $M > 0$ such that

$$E[|Y_i|^\ell|\mathbf{X}_i] \le C\ell!M^\ell, \qquad \forall\, 2 \le \ell \le \infty.$$

(A5) Denote $\mathbf{m} = (m_1,\ m_2)'$ and

$$\begin{aligned}\mathbf{u}_j(\boldsymbol{\zeta}, \mathbf{m}; x) &= E\left[\{m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)) - m_1(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))\}\tfrac{m_2(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))}{V(m_1\{\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i)\})}|X_{ij} = x\right] f_j(x),\\ \mathbf{s}_m(\boldsymbol{\zeta}, m_1; u) &= E\left[\{m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)) - m_1(u)\}/V(m_1\left(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i)\right))|\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i) = u\right] f_{\boldsymbol{\zeta}}(u).\end{aligned}$$

Define $\mathbf{u}(\boldsymbol{\zeta}, \mathbf{m}; \mathbf{x}, u) = \left(\{\mathbf{u}_j(\boldsymbol{\zeta}, \mathbf{m}; x_j)\}_{j=1}^d, \mathbf{s}_m(\boldsymbol{\zeta}, m_1; u)\right)'$. Then, one shall assume that $\mathbf{u}(\boldsymbol{\zeta}, \mathbf{m}; \mathbf{x}, u) = 0$ has a unique root over $\boldsymbol{\zeta} \in \mathcal{C}_d$, $m_1 \in \mathcal{C}_1$, where $\mathcal{C}_d$ and $\mathcal{C}_1$ are defined in Appendix A.

(A6) $h_j \to 0$ and $nh_j/(\log n) \to \infty, j = 1, \cdots, d, m,$ as $n \to \infty$.

(A7) Ψ^{-1} and $(\mathrm{H}_{\mathbf{g}} - \mathrm{H}_{\mathbf{m}} \circ \mathrm{H}_{mg})^{-1}$ exist and are bounded uniformly, where Ψ is an operator-type matrix, and $\mathrm{H}_{\mathbf{g}}$ and $\mathrm{H}_{\mathbf{m}}$ are vector-valued operators, and all of them are defined in Appendix A.

These conditions are used for deriving the convergence properties, asymptotic representation and efficiency for the proposed estimators. Conditions (A1)-(A4) are regular conditions for the kernel function, covariates, the functions of interest and the distribution. $\mathbf{u}_j(\boldsymbol{\zeta}, \mathbf{m}; x), j = 1, \cdots, d$ and $\mathbf{s}_m(\boldsymbol{\zeta}, m_1; u)$ in essential are the population version of quasi-score function for $g_j(x), j = 1, \cdots, d$ and $m(u)$, respectively. The proposed estimator converges to the root of quasi-score function. Hence, Condition (A5) ensures the proposed estimator converges to a determined value. Condition (A6) is the most often used condition for the bandwidths. Condition (A7) ensures the existence of asymptotic variance of the proposed estimator for $\mathbf{g}(\mathbf{x})$ and $m(u)$.

For the sake of convenience, without any confusion, we denote h_m by h_{d+1}, and for any $\mathbf{x} = (x_1,\ \cdots,\ x_d)' \in [-1,\ 1]^d$, we denote $(\widehat{g}_1(x_1) - g_1(x_1),\ \cdots,\ \widehat{g}_d(x_d) - g_d(x_d))'$ by $\widehat{\mathbf{g}}(\mathbf{x}) - \mathbf{g}(\mathbf{x})$.

Theorem 1 *Under the Conditions (A1)-(A6), when $n \longrightarrow \infty$, we have*

$$\sup_{x\in[-1,1]} |\widehat{g}_j(x) - g_j(x)| \longrightarrow 0, \quad j = 1, \ \cdots, \ d,$$

and

$$\sup_{u\in[-1,1]} |\widehat{m}(u) - m(u)| \to 0,$$

in probability.

Theorem 1 shows the proposed estimator $\widehat{g}_j(\cdot)$, $j = 1, \ \cdots, d$, is uniformly convergent, and so is the proposed estimator of $\widehat{m}(\cdot)$

Theorem 2 *Under the Conditions (A1)-(A7), we have*

$$\Psi \left(\begin{array}{c} \widehat{\mathbf{g}}(\mathbf{x}) - \mathbf{g}(\mathbf{x}) \\ \widehat{m}(u) - m(u) \end{array} \right) = (nH)^{-1/2}\mathbf{M}(\mathbf{x}, u)^{-1/2}\boldsymbol{\varphi} + H^2\mathrm{B}(\mathbf{x}, u) + o_p[\sum_{k=1}^{d+1}\{h_k^2 + (nh_k)^{-1/2}\}],$$

uniformly on $\mathbf{x} \in [-1, \ 1]^d$, $u \in [-1, \ 1]$, where Ψ is an operator-type matrix and defined in Appendix A, $H = diag(h_1, \cdots, h_{d+1})$, $\boldsymbol{\varphi}$ is a random vector following the standard normal distribution, both $\mathrm{B}(\mathbf{x}, u)$ and $\mathbf{M}(\mathbf{x}, u)$ are defined in Appendix A.

Theorem 2 shows the asymptotic bias of the proposed estimator $(\widehat{\mathbf{g}}(\mathbf{x})', \ \widehat{m}(u))'$ is of order $h^2 = \max_{1\leq j\leq d+1} h_j^2$, and the asymptotic variance is of order $(nh)^{-1}$. As a consequence, the theoretical optimal bandwidth for the estimator is of order $n^{-1/5}$, and the convergence rate of the estimator is of order $n^{-2/5}$.

Theorem 2 implies the following Corollary 1.

Corollary 1 *Under the Conditions (A1)-(A7). For any given $\mathbf{x} \in [-1, \ 1]^d$ and $u \in [-1, 1]$, if $nh_k^5 = O(1)$ for any $k = 1, \ \cdots, \ d+1$, we have*

$$(nH)^{1/2} \left\{ \left(\begin{array}{c} \widehat{\mathbf{g}} - \mathbf{g} \\ \widehat{m} - m \end{array} \right) (\mathbf{x}, u) - H^2\Psi^{-1}(\mathrm{B})(\mathbf{x}, u) \right\} \to N(0, \mathbf{V}(\mathbf{x}, u)),$$

where

$$\mathbf{V}(\mathbf{x}, u) = [\Psi^{-1}(\mathbf{M}^{-1/2})(\mathbf{x}, u)][\Psi^{-1}(\mathbf{M}^{-1/2})(\mathbf{x}, u)]'.$$

Linear functionals are important because any smooth function f can be approximated by an expansion of the orthonormal basis functions $\psi_0, \psi_1, \cdots$ (e.g. Fourier basis), in which the coefficients are of form such linear functionals. Estimators for the function $f(\cdot)$ are obtained by substituting estimators for the coefficients in a truncated or a tapered expansion of $f(\cdot)$ in the orthonormal basis $\psi_0, \psi_1, \cdots$. Hence, we determine the efficiency of the functional estimator via the coefficients, that has the form of linear functionals. In the appendix, if the conditional distribution of Y_i given X_i is an exponential family distribution, we prove our estimator $\hat{\tau} = \sum_{j=1}^{d} \int_{-1}^{1} \widehat{g}_j(x)\psi_j(x)dx + \int_{-1}^{1} \widehat{m}(u)\psi_m(u)du$ for the linear functionals $\tau = \sum_{j=1}^{d} \int_{-1}^{1} g_j(x)\psi_j(x)dx + \int_{-1}^{1} m(u)\psi_m(u)du$ has the same asymptotic variance with that of the maximum likelihood estimator for τ under some parametric submodel. This is actually semiparametrically efficient in the sense of Bickel et al. (1993). Theorem 3 presents the semiparametric efficiency, which is a justification for the optimality of the proposed approach. Let

$$\mathcal{D} = \{\psi(z) \text{ has a continous derivative on } [-1,1] \text{ and } \int_{-1}^{1} \psi(z)dz = 0\}.$$

Theorem 3 *Under the conditions (A1)-(A7). When* $nh_k^4 \to 0$, $h_k^2 h_j^{-1} \log(n) \to 0$ *and* $nh_k h_j/(\log(n))^2 \to \infty$ *for any* $k, j = 1, \cdots, d+1$, *for any functions* $\psi_j(x)$, $j = 1, \cdots, d$, *and* $\psi_m(u)$, *if* $\psi_j \in \mathcal{D}$, $j = 1, \cdots, d$, *and* $\psi_m(u)$ *has a continuous derivative, we have*

$$\sum_{j=1}^{d} \int_{-1}^{1} (\widehat{g}_j - g_j)(x)\psi_j(x)dx + \int_{-1}^{1} (\widehat{m} - m)(u)\psi_m(u)du \to N(0, \sigma_v^2).$$

In particular, $\sum_{j=1}^{d} \int_{-1}^{1} \widehat{g}_j(x)\psi_j(x)dx + \int_{-1}^{1} \widehat{m}(u)\psi_m(u)du$ *is an efficient estimator of* $\sum_{j=1}^{d} \int_{-1}^{1} g_j(x)\psi_j(x)dx + \int_{-1}^{1} m(u)\psi_m(u)du$ *if the conditional distribution of* Y_i *given* X_i *is an exponential family distribution, where* σ_v^2 *is defined in Appendix A.*

Theorem 3 imply that the estimator of the parameter $\sum_{j=1}^{d} \int g_j(x)\psi_j(x)dx + \int m(u)\psi_m(u)du$ is $\sqrt{n}-$consistent when $h_k = o(n^{-1/4})$. The requirement of undersmoothing to gain $\sqrt{n}$-consistent estimators is common in semi-parametric regression (Carroll, et al., 1997; Hastie and Tibshirani, 1990).

Theorem 3 shows the proposed estimators enjoy the semiparametric efficiency. Essentially, the using of likelihood function is the key for the semiparametric efficiency. To illuatrate it, we take the estimation of $\mathbf{m} = (m(u), \dot{m}(u))'$ as an example. Substituting (2.3) into the quasi-likelihood function, which indeed is the full likelihood function when the conditional distribution is an exponential family distribution, has the form of

$$\begin{aligned} Q(\mathbf{g},\ m) &= \sum_{i=1}^{n} L(\mu_i,\ Y_i) K_{h_m}(R_i - u) + \sum_{i=1}^{n} L(\mu_i,\ Y_i)\left\{1 - K_{h_m}(R_i - u)\right\} \\ &\approx \sum_{i=1}^{n} L(\bar{\mu}_i,\ Y_i) K_{h_m}(R_i - u) + \sum_{i=1}^{n} L(\mu_i,\ Y_i)\left\{1 - K_{h_m}(R_i - u)\right\}, \end{aligned} \tag{3.1}$$

where $R_i = \sum_{j=1}^{d} g_j(X_{ij})$ and $\bar{\mu}_i = m(u) + \dot{m}(u)(R_i - u)$. The μ_i in the second term of (3.1) can not be approximated by the linear function $\bar{\mu}_i = m(u) + \dot{m}(u)(R_i - u)$ because R_i is out of the neighborhood of u, which indicated by the weight $1 - K_{h_m}(R_i - u)$. Differentiating the likelihood function $Q(\mathbf{g},\ m)$ with respect $\mathbf{m} = (m(u), \dot{m}(u))'$, and setting the derivatives to zero leads to the following score equations:

$$\sum_{i=1}^{n} \Big(Y_i - \bar{\mu}_i\Big) \frac{W_i(u; \mathbf{g})}{V(\bar{\mu}_i)} K_{h_m}(R_i - u) = 0. \tag{3.2}$$

Noting that $V(\bar{\mu}_i) \approx V(\mu_i)$ when R_i is in the neighborhood of u, indicated by the weight $K_{h_m}(R_i - u)$, the proposed estimating equation (2.4), $S_m(\mathbf{m}; \mathbf{g}) = 0$, is exactly the same with the socre equation (3.2) for estimating $\mathbf{m}$, implying the efficiency of the estimator for $\mathbf{m}$.

4 Numerical studies

In this section, we are going to use four simulated examples to demonstrate how well the proposed estimation procedure works. We will also compare the proposed estimation procedure with the method in Horowitz and Mamman (2007), denoted by HM, which is designed for model (1.1).

We define the bias, standard deviation (SD) and root mean integrated squared error (RMISE) of an estimator $\hat{f}(\cdot)$ of $f(\cdot)$ as

$$\text{bias} = \left(\int \left[E\left\{ \hat{f}(v) \right\} - f(v) \right]^2 dv \right)^{1/2}, \quad \text{SD} = \left(\int var\left\{ \hat{f}(v) \right\} dv \right)^{1/2}$$

and

$$\text{RMISE} = \left(\text{bias}^2 + \text{SD}^2 \right)^{1/2},$$

respectively, and use them to assess the accuracy of the estimator $\hat{f}(\cdot)$.

The kernel function used in the proposed estimation procedure is the Epanechnikov kernel for all simulated examples in this section and the real data analysis in Section 5. For each simulated example, we assess the accuracy of the proposed estimation procedure for sample size n = 200, 400, 800, or 1600 and for each case, we compute the bias, sd and RMISE of an obtained estimator based on 200 simulations. We consider the following four settings.

Example 1. (*Binary Case*). In models (1.2), we set $d = 2$, and

$$g_1(x) = \sin(\pi x), \quad g_2(x) = \frac{1}{2}(x+1)^2 - \frac{2}{3}.$$

We generate X_{i1} and X_{i2}, $i = 1, \cdots, n$, from the uniform distribution $U[-1, 1]$, and $Y_i = I\{g_1(X_{i1}) + g_2(X_{i2}) > U_i\}$, where U_i is generated by a mixed normal $0.5N(-2/3 + 0.05, 0.5^2) + 0.5N(-2/3 - 0.05, 0.5^2)$ and I is the indicator function. Hence given X_{i1} and X_{i2}, Y_i has the Bernoulli distribution $B(1, p_i)$ with

$$p_i = E(Y_i|X_{i1}, X_{i2}) = m\left\{ g_1(X_{i1}) + g_2(X_{i2}) \right\},$$

$$m(x) = 0.5\Phi\left\{ \frac{x + 2/3 - 0.05}{0.5} \right\} + 0.5\Phi\left\{ \frac{x + 2/3 + 0.05}{0.5} \right\}.$$

It is clear $m(\cdot)$ is not the commonly used logit function for binary response.

Example 2. (*Poisson Case*). We still set $d = 2$, in models (1.2), but

$$g_1(x) = sin(\pi x), \quad g_2(x) = \Phi(3x) - 0.5, \quad m(x) = (3x + 8.5)^2.$$

X_{i1} and X_{i2}, $i = 1, \cdots, n$, are still generated in the same way as that in Example 1. However, Y_i, given X_{i1} and X_{i2}, is independently generated from a Poisson distribution with mean

$$m\left\{g_1(X_{i1}) + g_2(X_{i2})\right\}.$$

Example 3. (*Normal Distribution Case*). This example is to compare the proposed estimation procedure, when applied to model (1.1), with the one in Horowitz and Mamman (2007) which is developed only for model (1.1). To make the comparison more convincing, we generate data from the similar setup as that used in Horowitz and Mamman (2007) for simulation. Specifically, d is still set to be 2, and

$$g_1(x) = \sin(-4\pi x), \quad g_2(x) = \Phi(3x) - 0.5, \quad m(x) = -(3x + 3.5)^2.$$

X_{i1} and X_{i2}, $i = 1, \cdots, n$, are generated in the same way as that in Example 1. Y_i, given X_{i1} and X_{i2}, is generated through

$$Y_i = m\left\{g_1(X_{i1}) + g_2(X_{i2})\right\} + U_i$$

where U_i is generated from the normal distribution $N(0,\ 0.5^2)$, and independent of X_{i1} and X_{i2}.

Example 4. (*Normal Distribution Case with Four Components*). This example is similar to the example 3 with normal distribution and $m(x) = -(3x + 17/3)^2$ except for $d = 4$ components,

$$g_1(x) = \sin(\pi x), \quad g_2(x) = \Phi(3x) - 0.5, \quad g_3(x) = x^2 - 1/3, \quad g_4(x) = \cos(\pi x).$$

X_{i1}, X_{i2}, X_{i3} and X_{i4}, $i = 1, \cdots, n$, are generated in the same way as that in Example 3. Y_i, given X_{i1}, X_{i2}, X_{i3} and X_{i4}, is generated through

$$Y_i = m\left\{g_1(X_{i1}) + g_2(X_{i2}) + g_3(X_{i3}) + g_4(X_{i4})\right\} + U_i$$

where U_i is generated from the normal distribution $N(0,\ 0.5^2)$, and independent of X_{i1}, X_{i2}, X_{i4} and X_{i4}.

We apply either the proposed estimation procedure or HM to the simulated data. The biases, SDs, and RMISEs of the estimators of the functions $g_j(\cdot), j = 1, 2, 3, 4$ and $m(\cdot)$, obtained by either approach with its optimal smoothing parameter that minimizing RMISE over several pre-specified value of smoothing parameter, are presented in Tables 1 to 4. The summaries are based on convergent cases in the 200 replicates. For Example 1, the bandwidths $h_1 = 0.3, h_2 = 0.3, h_m = 0.9$ are used for the proposed method, 4 interior knots and smoothing parameter 10^{-5} are used for the HM method. For Example 2, the bandwidths $h_1 = 0.2, h_2 = 0.2, h_m = 0.6$ are used for the proposed method, 2 interior knots and smoothing parameter 5×10^{-5} are used for the HM method. For Example 3, the bandwidths $h_1 = 0.05, h_2 = 0.1, h_m = 0.35$ are used for the proposed method, 17 knots and smoothing parameter 5×10^{-7} are used for the HM method. For Example 4, the bandwidths $h_1 = 0.15, h_2 = 0.2, h_2 = 0.2, h_3 = 0.3, h_4 = 0.15, h_m = 0.3$ are used for the proposed method, 3 knots and smoothing parameter 10^{-5} are used for the HM method.

The NOC in Tables 1 to 4 is the total number of the simulations where convergence has attained in the 200 simulations. Tables 1 to 4 show the proposed estimation procedure is always convergent except when sample size $n = 200$, even for that case, there are only 7 and 11 simulations for 2 and 4 components respectively where convergence is not attained. On the other hand, there are quite a few cases where HM does not converge, especially when sample size is 200 and the response is Poisson variable, there are more than half cases where HM fails to converge. Tables 1 to 4 also show the estimators obtained by the proposed estimation procedure have much smaller RMISE than that obtained by HM. So, we can conclude the proposed estimation procedure works much better than HM.

To have a more visible idea about how well the proposed estimation procedure works, and how better it compared with HM, we choose Example 1 and sample size $n = 400$ as an example. For the proposed estimation procedure, we plot out in Figure 1 the average of the estimates of each unknown function across the 200 simulations, and superimpose its 95% pointwise confidence interval on it. We do exactly the same for HM, and also present the results in Figure 1. Figure 1 shows the average of

Table 1: The simulation results for Example 1 with binary response

		N=400			N=800			N=1600		
		$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$	$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$	$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$
Prop.	RMSE	0.122	0.122	0.024	0.095	0.095	0.022	0.083	0.079	0.021
	Bias	0.070	0.068	0.018	0.070	0.065	0.018	0.069	0.064	0.019
	SD	0.099	0.102	0.016	0.064	0.069	0.012	0.045	0.047	0.008
	NOC		200			200			200	
HM	RMSE	0.192	0.208	0.048	0.164	0.164	0.041	0.132	0.139	0.030
	Bias	0.133	0.137	0.018	0.119	0.111	0.023	0.115	0.096	0.021
	SD	0.139	0.156	0.045	0.112	0.121	0.034	0.065	0.102	0.022
	NOC		186			198			198	

Table 2: The Simulation Results for Example 2 with Poisson response

		n=200			n=400			n=800		
		$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$	$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$	$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$
Prop.	RMISE	0.110	0.041	1.038	0.039	0.020	0.212	0.029	0.015	0.172
	Bias	0.020	0.005	0.265	0.016	0.007	0.165	0.016	0.007	0.152
	SD	0.108	0.040	1.004	0.031	0.018	0.133	0.025	0.013	0.079
	NOC		193			200			200	
HM	RMISE	0.064	0.037	0.905	0.048	0.024	0.764	0.033	0.019	0.636
	Bias	0.011	0.007	0.277	0.013	0.008	0.323	0.014	0.009	0.363
	SD	0.063	0.036	0.862	0.046	0.023	0.693	0.030	0.016	0.522
	NOC		96			181			198	

the estimates, obtained by the proposed estimation procedure, is very close to the true function, and more than that obtained by HM close to the true function. This suggests the proposed estimators have smaller bias than the estimators obtained by HM. Furthermore, it is evident, from the 95% pointwise confidence interval in Figure

Table 3: The Simulation Results for Example 3 with normal response

		n=200			n=400			n=800		
		$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$	$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$	$\hat{g}_1$	$\hat{g}_2$	$\hat{m}$
Prop.	RMISE	0.052	0.047	1.796	0.040	0.031	1.318	0.027	0.023	1.012
	Bias	0.013	0.008	0.785	0.015	0.005	0.724	0.014	0.006	0.707
	SD	0.050	0.047	1.615	0.037	0.031	1.101	0.024	0.022	0.725
	NOC		200			200			200	
HM	RMISE	0.071	0.067	3.812	0.051	0.044	2.629	0.035	0.031	1.816
	Bias	0.016	0.013	1.460	0.010	0.008	0.725	0.007	0.003	0.251
	SD	0.070	0.066	3.521	0.050	0.044	2.527	0.034	0.031	1.799
	NOC		190			196			200	

1, that the proposed estimators have smaller standard deviation than the estimators obtained by HM.

5 Application to a microfinance data

With the advancement of mobile phone technology, the functionality of a smartphone has gone far beyond its traditional role of communication. It acts as a portable computer in many cases, and plays an important role in people's daily life. A modern and fashionable smart phone typically costs around 5000 RMB in China, which is not affordable for some Chinese with lower income. On the other hand, fashion pursuit and keeping up with the Joneses are a habit of some Chinese, which makes some people appeal to personal loan for a fancy smartphone. Whilst the financial service providers, the loan companies, make big profit from such kind of loans, they are also inflicted by the loss from loan default or delay of repayment from time to time. Therefore, credit check becomes very important for financial service providers.

In this paper, we define the credit score of a person as the probability of this

Table 4: The Simulation Results for Example 4 with four components

		n=200				
		$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_3$	$\hat{g}_4$	$\hat{m}$
Prop.	RMISE	0.036	0.020	0.016	0.032	0.290
	Bias	0.007	0.003	0.005	0.007	0.169
	SD	0.035	0.020	0.015	0.032	0.236
	NOC			189		
HM	RMISE	0.039	0.023	0.018	0.041	0.568
	Bias	0.005	0.002	0.003	0.018	0.070
	SD	0.039	0.023	0.017	0.037	0.563
	NOC			193		
		n=400				
		$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_3$	$\hat{g}_4$	$\hat{m}$
Prop.	RMISE	0.027	0.013	0.012	0.025	0.226
	Bias	0.009	0.003	0.005	0.006	0.157
	SD	0.025	0.013	0.011	0.024	0.162
	NOC			200		
HM	RMISE	0.029	0.015	0.012	0.031	0.344
	Bias	0.007	0.003	0.003	0.016	0.022
	SD	0.028	0.014	0.012	0.026	0.343
	NOC			197		
		n=800				
		$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_3$	$\hat{g}_4$	$\hat{m}$
Prop.	RMISE	0.018	0.010	0.008	0.018	0.197
	Bias	0.009	0.004	0.005	0.007	0.165
	SD	0.016	0.009	0.007	0.016	0.107
	NOC			200		
HM	RMISE	0.020	0.011	0.009	0.025	0.245
	Bias	0.009	0.004	0.004	0.016	0.075
	SD	0.018	0.010	0.008	0.018	0.233
	NOC			200		

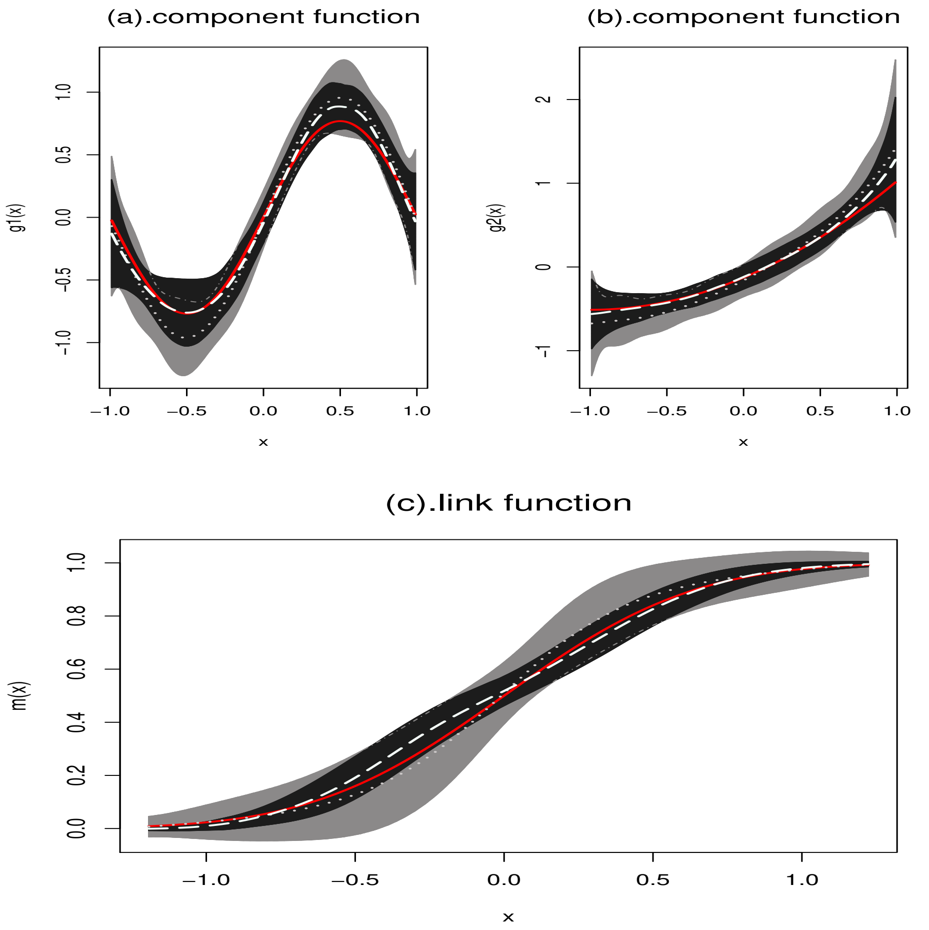

Figure 1: The averaged estimates of component curves (top) and link function (bottom) (dashed—proposed estimator; dotted—HM estimator; black shadow—95% confidence limit of proposed estimator; gray shadow—95% confidence limit of the HM estimator; solid-red—true functions) for Example 1.

person repaying his/her loan in time. We are going to build a credit rating model based on the proposed GAMUL model, thereby, we can estimate the credit score of a loan applicant and decide whether to lend the loan to this person.

The dataset for us to study is a microfinance dataset, collected from a financial service provider, which records the repayment statuses of 2160 loans for buying smartphone and the personal information of the 2160 borrowers. Due to confidentiality, we cannot disclose the identity of this financial service provider. We use Y_i to denote the loan repayment status of the ith borrower, $Y_i = 0$ if the loan is fully repaid in time, $Y_i = 1$ otherwise. The variables of interest are the Tongdun score (denoted by X_{i1} for the ith borrower), the price of the smartphone to buy (in 1000RMB, denoted by X_{i2}), loan amount (in 1000RMB, denoted by X_{i3}), and the logarithm of personal income per month in 1000RMB (denoted by X_{i4}).

To have a basic idea about what the data is like, we plot out the histograms of the four variables of interest in Figure 2.

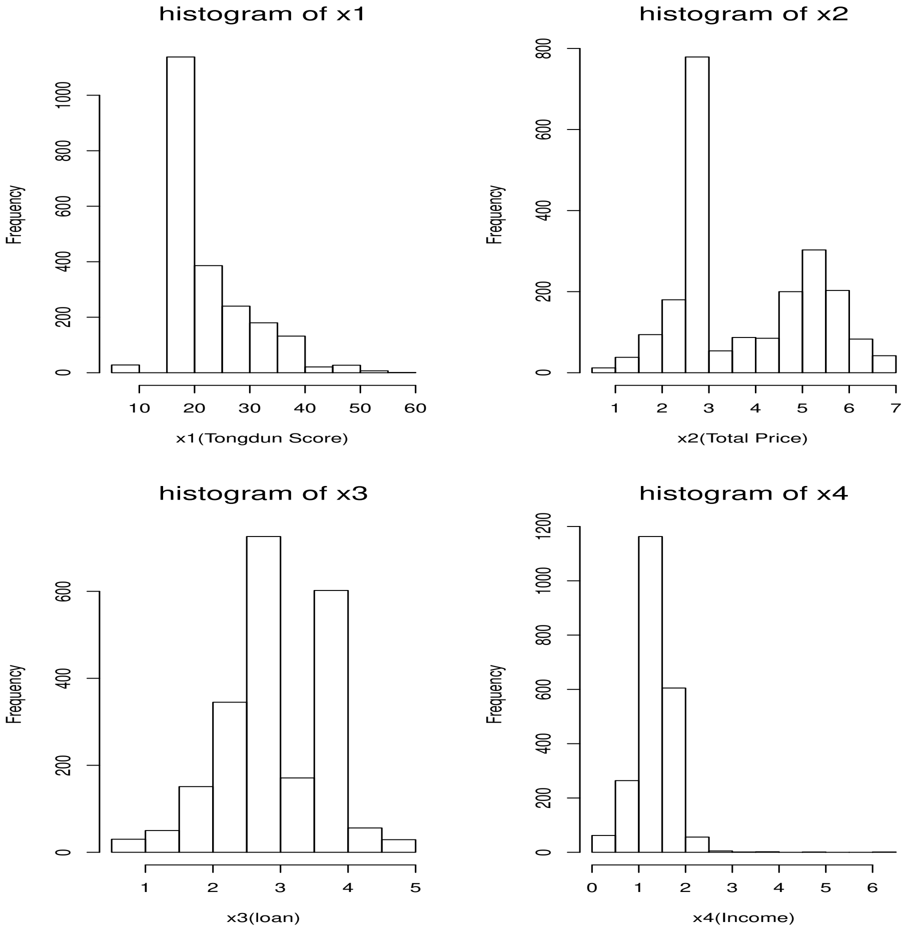

Figure 2: Histogram of four covariates

As the response variable Y_i is binary, a typical approach to analyse such kind of dataset would be the logistic regression with logit link function, which could be too restrictive for our dataset and has the danger of misspecification. The proposed GAMUL model is much more flexible than the commonly used logistic regression. Given the sample size of the dataset is in the order of thousands, we are going to use the proposed GAMUL model to fit the dataset, which is detailed as follows: we assume

$$\begin{cases} p_i &= m\{g_1(X_{i1}) + g_2(X_{i2}) + g_3(X_{i3}) + g_4(X_{i4})\}, \\ \sigma_i^2 &= p_i(1 - p_i) \end{cases}$$

where

$$p_i = E[Y_i|X_i], \quad \sigma_i^2 = Var[Y_i|X_i], \quad X_i = (X_{i1},\ X_{i2},\ X_{i3},\ X_{i4})',$$

$m(\cdot)$ and $g_j(\cdot)$ are unknown functions to be estimated, and

$$E\{g_j(X_{ij})\} = 0, \quad j = 1,\ 2,\ 3,\ 4. \quad \sum_{j=1}^{d} Var\{g_j(X_{ij})\} = 1,$$

with the restriction of $Cov\,\{X_{i1}, g_1(X_{i1})\} > 0$.

Due to non-uniformly distributed covariates, see Figure 2, we use the adaptive bandwidth (Brockmann et al., 1993) in our estimation. Specifically, for each covariate, we select an adaptive bandwidth such that the resulting neighbourhood covers a given portion, denoted by q, of the observations. We apply the K-fold cross-validation (Cai et al., 2000; Fan et al., 2006) to choose q. The number K is usually chosen to be $K = 5$ or $K = 10$. K is set to be 5 for our dataset. Denote the full dataset by $\mathbf{B}$, and denote cross-validation training and test sets by $\mathbf{B} - \mathbf{B}_k$ and $\mathbf{B}_k$, respectively, for $k = 1, \cdots, K$. For each bandwidth q and k, we find the estimator $g^{(-k)}(x)$ of $g(x)$ using the training set $\mathbf{B} - \mathbf{B}_k$, and form the cross-validation criterion based on the predict error $PE(h)$ for the test sets. We then find the optimal bandwidth q that minimizes the criterion $PE(q)$. Specifically, we minimize

$$PE(q) = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in B_k} \frac{1}{m_i} \sum_{j=1}^{m_i} \left[Y_{ij} - \hat{m}^{(-k)} \left\{ \sum_{j=1}^{d} \hat{g}_j^{(-k)}(X_{ij}) \right\} \right]^2,$$

where n_k is the number of the observations in set B_k, and $q = (q_1, \cdots, q_d, q_m)$ for $g_1(\cdot), \cdots, g_d(\cdot)$ and $m(\cdot)$, respectively. The obtained q's are $q_1 = 0.5$ for $g_1(\cdot)$, $q_2 = 0.4$ for $g_2(\cdot)$, $q_3 = 0.4$ for $g_3(\cdot)$, $q_4 = 0.4$ for $g_4(\cdot)$, and $q_m = 0.6$ for the link function $m(\cdot)$.

The estimated component functions and link function, by the proposed estimation procedure, are presented in Figures 3(a)-3(e) in solid line along with their 95% confidence intervals (dark shadow). To make a comparison of proposed method with the HM method, we also apply the HM method, with optimal smoothing parameters, to estimate the component functions and link function for our dataset, and superimpose the obtained estimates on their counterparts obtained by the proposed method in Figures 3(a)-3(e) in dashed line along with their 95% confidence intervals (gray shadow).

Figures 3(a)-3(e) show that although the trend of each estimated function obtained by the proposed method is similar to that obtained by the HM method, the HM method produces much wider 95% confidence interval (gray shadow). As a result, the HM method has failed to detect the significant effects of the four variables concerned on the probability of a borrower repaying his/her loan for buying a smartphone, whereas the proposed method has successfully identified that Tongdun score and loan amount significantly affect the probability of a borrower repaying his/her loan.

Importantly, Figure 3(e) shows the estimated link function is quite different to the commonly used logit function. In fact, the logit function (the dotted line in Figure 3(e)) is even not in the 95% confidence interval of the link function. This suggests the commonly used logit function is not appropriate for this data set.

It is easy to see, from Figures 3(a) and 3(c), that neither the 95% confidence interval of $g_1(X_{i1})$ nor the 95% confidence interval of $g_3(X_{i3})$ covers the zero function, this suggests both Tongdun score and loan amount have significant impact on the probability of repaying the loan. Furthermore, Figure 3(a) shows the higher the Tongdun score, the bigger the risk of default or repayment delay, and this risk is linearly increasing when the Tongdun score is less than 45, then tends to steady. This is consistent with empirical observation. Figure 3(c) is quite interesting, it shows that when the loan amount is less than 3900 RMB, the larger the loan the bigger the risk of default or repayment delay, however, when the loan amount is greater than 3900 RMB, the larger the loan the smaller the risk of default or repayment delay. This is because the penalty incurred, as a result of the delay of repayment, does not have big difference when the loan is smaller than a threshold, therefore, the larger the loan, the more likely people delay their repayment. However, when the loan is larger than that threshold, the penalty would have a big jump, which deters people delaying their repayment. That is why the larger the loan the smaller the risk of default or repayment delay when the loan amount is greater than 3900 RMB.

Finally, to examine the prediction accuracy, we randomly divided the data into two subsets: the training set and validation set. We use the model concerned to fit the training set. For each subject in the validation set, we predicted the subjects loan repayment status by the fitted model obtained from the training set. We investigated the performance of the model by examining the squared difference of observed loan repayment status and the prediction of loan repayment status in each of the training set. We take the training sets to be 70%, 80% and 90%, respectively, of full dataset, using the same bandwidth and smoothing parameters mentioned above. The prediction error (PE) for the proposed method and HM method are presented in Table 5, which suggest that regardless whatever the percentage of the training sets is, the proposed method always outperforms the HM method.

Table 5: Prediction error for the Proposed method and HM method

Training Set Rate	Prop.	HM
0.7	0.1523	0.1535
0.8	0.1513	0.1527
0.9	0.1532	0.1542

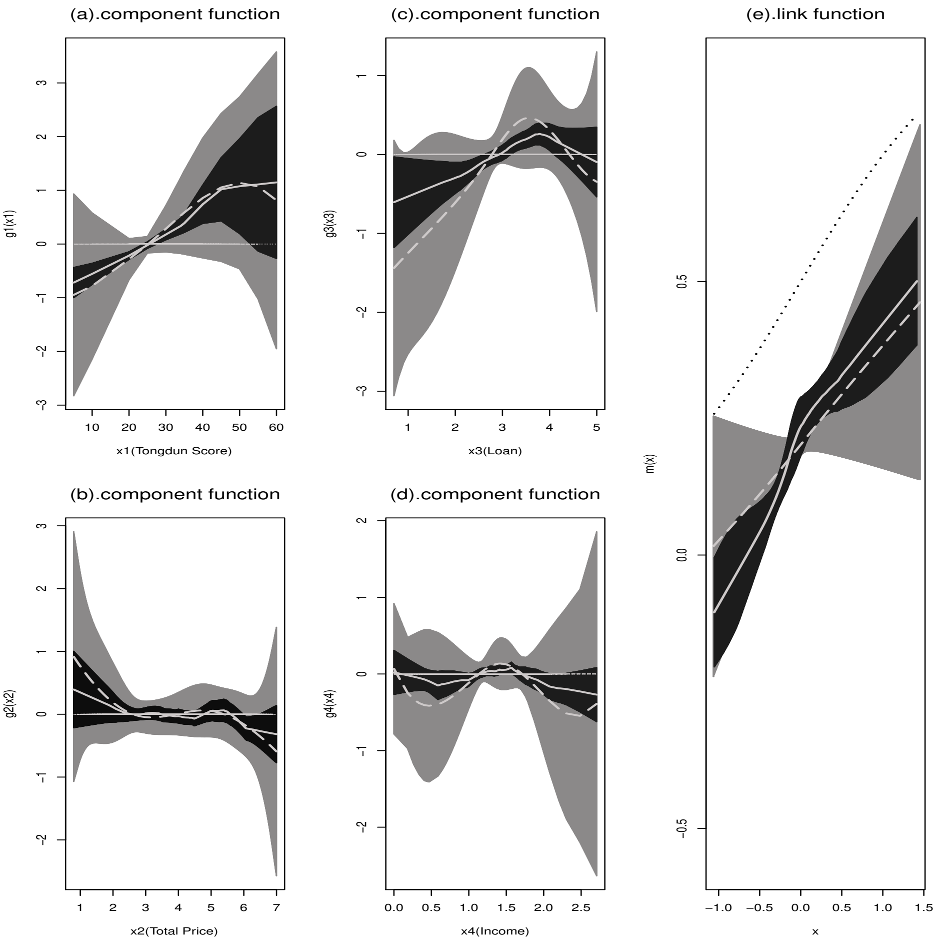

Figure 3: Estimated component and link functions for the mobile phone microfinance data and their 95% confident interval by proposed method (solid line and black area, respectively) with $q_1 = 0.5$, $q_2 = 0.4$, $q_3 = 0.4$, $q_4 = 0.4$, and $q_m = 0.6$ and by the HM estimator and their 95% confident interval with 2 interior knots and smoothing parameter 10^{-3} (dashed line and gray area, respectively). The dotted line in (e) is the logit function.

6 Discussion

In the paper, we propose a generalized additive model for normal and non-normal response. Different from the existing methods (Horowitz, 2001; Horowitz and Mammen, 2007), our method can handle with heteroscedastic variance data, hence is more flexible and efficient. To estimate the component and link functions, we propose quasi-likelihood backfitting method, which involves just one-dimensional kernel hence avoid the problem of the curse of the dimensionality. Moreover, the proposed estimator has a closed form, this has dramatically reduced the computational burden. Finally, the proposed method is shown to be uniformly consistent, asymptotically normal and semiparametrically efficient in terms of Bickel et al. (1993) if the conditional distribution belongs to an exponential family. The simulation study and real data analysis show that our estimator is more efficient and robust than the existing method.

It is straightforward to extend our method to the generalized additive models (1.2) with unknown variance function $V(\cdot)$, but larger amount of information is required because the estimation of variance function involves the second order moment.

In practice, the number of covariates may be large, we need to estimate the component functions and select the significant covariates simultaneously, which is commonly conducted by adding a penalty term to a objective function. However, since the proposed method estimates the component functions point-by-point, it may not be suitable to combine the proposed method with a penalty method to simultaneously estimate and select the covariates. Some spline regression method may be a better choice.

References

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, **68**, 357–363.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models.* Springer-Verlag, New York.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580–619.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.*, **17**, 453–555.

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series models. *J. Amer. Statist. Assoc.* **95**, 941–956.

Carroll, Fan, Gijbels and Wand (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.*, **92**, 477–489.

Chen, K., Guo, S., Sun, L. and Wang, J. (2010). Global partial likelihood for nonparametric proportional hazards models. *J. Amer. Statist. Assoc.*, **105**, 750–760.

Chen, K., Lin, H. Z. and Zhou, Y. (2012). Efficient estimation for the Cox model with varying coefficients. *Biometrika*, **2**, 379-392.

Chiou, J. M. and Müller, H. G. (1998). Quasi-likelihood regression with unknown link and variance functions. *J. Amer. Statist. Assoc.*, **93**, 1376–1387.

Chiou, J. M. and Müller, H. G. (2004). Quasi-likelihood regression with multiple indices and smooth link and variance functions. *Scand. J. Statist.*, **31**, 367–386.

Chen, K., Lin, H. Z. and Zhou, Y. (2012). Efficient estimation for the Cox model with varying coefficients. *Biometrika*, **2**, 379-392.

Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.*, **26**, 943–971.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *Ann. Statist.*, **21**, 157–178.

Härdle, W., and Stoker, T. M. (1989), Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, **84**, 986–995.

Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. Chapman and Hall, London.

Horowitz, J. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, **69**, 499–513.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of a single-index model with discrete covariates. *J. Amer. Statist. Assoc.*, **91**, 1632–1640.

Horowitz, J. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist.*, **32**, 2412–2443.

Horowitz, J., Klemel, J. and Mammen, E. (2006). Optimal estimation in additive regression models. Bernoulli 12 271?98.

Horowitz, J. and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.*, **35**, 2589–2619.

Kauermann, G. and Opsomer, J. D. (2003). Local likelihood estimation in generalized additive models. *Scand. J. Statist.*, **30**, 317–337.

Li, K. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.*, **17**, 1009–1052.

Li, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86**, 316–327.

Linton, O. (1997). Miscellanea efficient estimation of additive nonparametric regression models. *Biometrika*, **84**, 469–473.

Linton, O. (2000). Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory*, **16**, 502–523.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93–100.

Linton, O. and Härdle, W. (1996). Estimation of additive regression models with known links. *Biometrika*, **83**, 529–540.

Mammen, E., Linton, O. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, **27**, 1443–1490.

Mammen, E. and Park, B. U. (2006). A simple smooth backfitting method for additive models. *Ann. Statist.*, **34**, 2252–2271.

Mccullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2nd ed. Chapman and Hall, London.

Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *J. Multiv. Anal.*, **73**, 166–179.

Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.*, **25**, 186–211.

Pinelis, I. F. and Sakhanenko, A. I. (1985). Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.* **30** (1985), 143–148.

Pollard, D. (1984). Convergence of Stochastic Process. New York: Springer Verlag.

Scallan A J, Gilchrist R and Green M. (1984). Fitting parametric link functions in generalized linear models. *Comput. statist. Data. Anal.*, **84**, 17–22.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, **13**, 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.*, **14**, 590–606.

Tjostheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: Projections. *J. Amer. Statist. Assoc.*, **89**, 1398–1409.

Wang, Y. G. and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika*, **90**, 29–41.

Weisberg, S. and Welsh, A. H. (1994). Adapting for the missing link. *Ann. Statist.*, **22**, 1674–1700.

Yu, K., Park, B. U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.*, **36**, 228–260.

Zhang, W., Li, D. and Xia, Y. (2015). Estimation in generalised varying-coefficient models with unspecified link functions. *J. Econom.*, **187**, 238-255.

7 Appendix A: Notations

Let $\boldsymbol{\zeta}(\mathbf{x}) = (\zeta_1(x_1), \zeta_2(x_2), ..., \zeta_d(x_d))'$, $\gamma(\mathbf{x}) = (\gamma_1(x_1), \gamma_2(x_2), ..., \gamma_d(x_d))'$, $\boldsymbol{\beta}(\mathbf{x}) = (\boldsymbol{\zeta}(\mathbf{x}), \gamma(\mathbf{x}))$, and $\mathbf{m}(u) = (m_1(u), m_2(u))'$. Let $\mathbf{e}$ be a d-dimensional vector with elements 1, $\mathbf{e}_j$ be a d-dimensional vector with the j-th element 1 and the rest $d-1$ elements 0, and $\mathbf{e}_{-j}$ be the vector with the j-th element 0 and the others 1. Let f_j be the density function of X_{ij}, and $f_{\boldsymbol{\zeta}}$ be the density of the random variable $\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X})$ associated with $\boldsymbol{\zeta}$. Define

$$\begin{aligned}\mathcal{C}_d = \{\boldsymbol{\zeta}(\cdot) :\, & \boldsymbol{\zeta}(\cdot) \,\text{is continuous on}\; [-1,1]^d, \text{and satisfying} \\ & \int \zeta_j(x) f_j(x) dx = 1, j = 1, \cdots, d, \int (\zeta_1(x) - 1)^2 f_1(x) dx = 1\}.\end{aligned}$$

and $\mathcal{C}_1 = \{\zeta(\cdot) :\, \zeta(\cdot) \,\text{is continuous on}\; [-1,1]\}$.

The following operators are used in Theorem 2, which are defined by

$$\begin{aligned}&\mathrm{H}_{\mathbf{m}j}(q)(x) = E\left[q\left[\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right] \times \dot{m}\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right)/V\big(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))\big)|X_{ij} = x\right] f_j(x),\ \forall\, q \in \mathcal{C}_1, \\ &\mathrm{H}_{\mathbf{g}j}(\mathbf{q})(x) = E\left[\left[\dot{m}\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right)\right]^2 \mathbf{e}_j'\mathbf{q}(\mathbf{X}_i)/V\big(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))\big)\Big|\, X_{ij} = x\right] f_j(x),\ \forall\, \mathbf{q} \in \mathcal{C}_d, \\ &\mathrm{H}_{mg}(\mathbf{q})(u) = E[\{m(\mathbf{e}'(\mathbf{g}(\mathbf{X}_i) + \mathbf{q}(\mathbf{X}_i))) - m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))\}/V\big(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))\big)\big|\, \mathbf{e}'\mathbf{g}(\mathbf{X}_i) = u] f_{\mathbf{g}}(u).\end{aligned}$$

Based on this, we define vector-valued operators as follows
$\mathrm{H}_{\mathbf{g}}(\mathbf{q})(\mathbf{x}) = \big\{\mathrm{H}_{\mathbf{g}1}(\mathbf{q})(x_1), \cdots, \mathrm{H}_{\mathbf{g}d}(\mathbf{q})(x_d)\big\}'$, $\mathrm{H}_{\mathbf{m}}(q)(\mathbf{x}) = \big\{\mathrm{H}_{\mathbf{m}1}(q)(x_1), \cdots, \mathrm{H}_{\mathbf{m}d}(q)(x_d)\big\}'$.
Define an operator-type matrix $\Psi = \begin{pmatrix} \mathrm{H}_{\mathbf{g}} & \mathrm{H}_{\mathbf{m}} \\ \mathrm{H}_{mg} & I \end{pmatrix}$, where I is the identity operator.

Denote $\mu_2 = \int_0^1 x^2 K(x) dx$, $\mathrm{B}(\mathbf{x}, u) := (B_1(x_1), B_2(x_2), ..., B_d(x_d), B_{d+1}(u))'$,

$$B_j(x_j) = \mu_2 \ddot{g}_j(x_j) f_j(x_j),\ j = 1, 2, ..., d, \quad B_{d+1}(u) = \mu_2 \ddot{m}(u) f_{\mathbf{g}}(u).$$

$\mathbf{M}(\mathbf{x}, u) = (M_{k,j}(\mathbf{x}, u))$ is defined to be a semidefinite matric with the following elements

$$\begin{aligned}&M_{k,j}(\mathbf{x}, u) = E\left[\dot{m}^2\left(\mathbf{e}'\mathbf{g}(X_i)\right)|X_{i,k} = x_k, X_{i,j} = x_j\right] \frac{f_k(x_k)}{\mathrm{H}_{\mathbf{g}j}(\mathbf{e}_j)(x_k)} \frac{f_j(x_j)}{\mathrm{H}_{\mathbf{g}j}(\mathbf{e}_j)(x_j)},\ k, j = 1, ..., d; \\ &M_{k,d+1}(\mathbf{x}, u) = \dot{m}(u) f_{\mathbf{g}}(u)/V(m(u)) \frac{f_k(x_k)}{\widetilde{g}_k(x_k)},\ k = 1, ..., d; \\ &M_{d+1,d+1}(\mathbf{x}, u) = f_{\mathbf{g}}(u)/V(m(u)).\end{aligned}$$

The following notations are used in Theorem 3. Let $\sigma_v^2 = E[\Phi_i \Lambda_i \Phi_i']$, $\Phi_i = (\mathbf{e}'\widetilde{\boldsymbol{\phi}}(X_i), \phi_m(\mathbf{e}'\mathbf{g}(X_i)))'$ with $\widetilde{\boldsymbol{\phi}}(X_i) = \big(\phi_1(X_{i1})/\mathrm{H}_{\mathbf{g}j}(\mathbf{e}_j)(X_{i1}), ..., \phi_d(X_{id})/\mathrm{H}_{\mathbf{g}j}(\mathbf{e}_j)(X_{id})\big)$, and Λ_i is a 2×2 symmetric matrix, which can be expressed as

$$\Lambda_i = \begin{pmatrix} A_{i,11} & A_{i,12} \\ A_{i,12} & A_{i,22} \end{pmatrix},$$

where??

$$A_{i,11} = [Y_i - m(\mathbf{e}'\mathbf{g}(X_i))]^2 \dot{m}^2(\mathbf{e}'\mathbf{g}(X_i)), \quad A_{i,12} = [Y_i - m(\mathbf{e}'\mathbf{g}(X_i))]^2 \dot{m}(\mathbf{e}'\mathbf{g}(X_i)),$$
$$A_{i,22} = [Y_i - m(\mathbf{e}'\mathbf{g}(X_i))]^2 .??$$

Appendix B: Proofs of Theorems

We firstly present two Lemmas, which are needed to prove Theorems.

Lemma 1 *Suppose conditions (A1)-(A3) hold and $g(x, y, z)$ is any bounded and continuous function. Then*

$$\sup_{x \in [-1,1]} |c_n(x) - Ec_n(x)| = O_p((\log n)^{1/2}(nh)^{-1/2}).$$

where $c_n(x) = \frac{1}{n}\sum_{i=1}^n g(X_i, (X_i - x)/h, x)K_h(X_i - x)$.

This Lemma is similar to Lemma 4 of Chen, al et.(2010) and follows from Theorem 37 and Example 38 in Chapter 2 of Pollard (1984).

Note that Theorem 37 of Pollard (1984) requires the bound of the random function, and it is no longer valid to the unbounded random variable, such as the case where Y_i is unbounded. To solve this problem, we need to introduce another concentration inequality for unbounded random variables in Hilbert spaces (Pinelis, 1985).

Lemma 2 *Let ξ_i $(i = 1, ..., n)$ be independent random variables with values in a Hilbert space such that $\mathbb{E}\xi_i = 0$. If for some constants $M, V > 0$, the bound $\mathbb{E}\|\xi_i\|^\ell \leq \frac{1}{2}\ell! M^{\ell-2}V$ holds for every $2 \leq \ell < \infty$, then there holds*

$$Prob\left\{\|\sum_{i=1}^{n}\xi_i\| \geq \varepsilon\right\} \leq 2\exp\left\{-\frac{\varepsilon^2}{2(\varepsilon M + Vm)}\right\} \quad \forall \varepsilon > 0.$$

Proof of Theorem 1.

For any vector functions $\boldsymbol{\beta}(\cdot)$ and $\mathbf{m}(\cdot)$, set

$$\begin{aligned}
U_j(\boldsymbol{\beta}, \mathbf{m}; x) &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - m_1\left\{\mathbf{e}'_{-j}\zeta(\mathbf{X}_i) + \zeta_j(x) + \gamma_j(x)(X_{ij} - x)\right\}\right]\Upsilon_{ij}(x) \\
&\qquad \times m_2\left(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i)\right)K_{h_j}(X_{ij} - x)/V\big(m_1(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))\big), \quad j = 1, 2, ..., d, \\
S_m(\boldsymbol{\zeta}, \mathbf{m}; u) &= \frac{1}{n}\sum_{i=1}^{n}[Y_i - m_1(u) - m_2(u) \times (\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i) - u)] \times K_{h_m}\left(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i) - u\right)\frac{W_i(\boldsymbol{\zeta}; u)}{V\big(m_1(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))\big)},
\end{aligned}$$

where $W_i(\boldsymbol{\zeta}; u) = [1, \mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i) - u]'$ and $\Upsilon_{ij}(x)$ is defined in Section 2. Then the deterministic terms of $U_j(\boldsymbol{\beta}, \mathbf{m}; x)$ and $S_m(\boldsymbol{\zeta}, \mathbf{m}; u)$ are given respectively by $\mathbf{u}_j(\boldsymbol{\zeta}, \mathbf{m}; x)$ and $\mathbf{s}_m(\boldsymbol{\zeta}, m_1; u)$, which are defined in Section 3.

Define $\mathbf{U}(\boldsymbol{\beta}, \mathbf{m}; \mathbf{x}, u) = \big(\{U_j(\boldsymbol{\beta}, \mathbf{m}; x_j)\}_{j=1}^d, S_m(\boldsymbol{\zeta}, \mathbf{m}; u)\big)'$. Then, the proposed iterative algorithms and the model (1.2) show that $\mathbf{U}([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}}; \mathbf{x}, u) = 0$ and $\mathbf{u}(\mathbf{g}, [m, m_2]; \mathbf{x}, u) = 0$ for any bounded function m_2.

Define $\mathcal{B}_n^d = \{f : \|f\|_\infty \leq C, \|f(z_1) - f(z_2)\| \leq c\|z_1 - z_2\| + b_n,\ z_1, z_2 \in [-1, 1]^d\}$,

$$\mathcal{B}_n^1 = \{f : \|f\|_\infty \leq C, \|f(z_1) - f(z_2)\| \leq c\|z_1 - z_2\| + b_n^1,\ z_1, z_2 \in [-1, 1]\},$$

for some constants $C > 0$ and $c > 0$, where $b_n = \max_{k=1,\cdots,d}\{h_k + (nh_k)^{-1/2}(\log n)^{1/2}\}$, $b_n^1 = \{h_{d+1} + (nh_{d+1})^{-1/2}(\log n)^{1/2}\}$ with $h_{d+1} = h_m$.

To show the uniform consistency of $\widehat{\mathbf{g}}$ and $\widehat{m}$, it suffices to prove the following conclusions:

(i) For any continuous function vectors $\boldsymbol{\zeta}$, m_1 and bounded functions γ, m_2,

$$\sup_{(\mathbf{x},u)\in[-1,1]^{d+1}} \|\mathbf{U}(\boldsymbol{\beta},\mathbf{m};\mathbf{x},u)-\mathbf{u}(\boldsymbol{\zeta},\mathbf{m};\mathbf{x},u)(1,0)'\| = o_p(1).$$

(ii) $\sup_{(\mathbf{x},u)\in[-1,1]^{d+1}} \|\mathbf{U}(\boldsymbol{\beta},\mathbf{m};\mathbf{x},u)-\mathbf{u}(\boldsymbol{\zeta},\mathbf{m};\mathbf{x},u)(1,0)'\| = o_p(1)$ uniformly holds over $\boldsymbol{\zeta}\in\mathcal{B}_n^d, m_1\in\mathcal{B}_n^1$ and bounded γ, m_2.

(iii) $P\{\widehat{\mathbf{g}}\in\mathcal{B}_n^d, \widehat{m}\in\mathcal{B}_n^1\}\to 1$.

Once (i)-(iii) are established, applying the Arzela-Ascoli theorem in $\mathcal{B}_n^k$ $(k=1,\, d)$ for the estimators $\{\widehat{\mathbf{g}},\widehat{m}\}$, we can show that for any subsequence of $\{\widehat{\mathbf{g}},\widehat{m}\}$, there exists a convergence subsequence $\{\widehat{\mathbf{g}},\widehat{m}\}_{nk}$, such that uniformly over $(\mathbf{x},u)\in[-1,1]^{d+1}$, $\{\widehat{\mathbf{g}},\widehat{m}\}_{nk}\to\{\mathbf{g}^*,m^*\}$ in probability, and it is easily seen that $\mathbf{g}^*\in\mathcal{C}_d$ and $m^*\in\mathcal{C}_1$, where $\mathcal{C}_1$ is the continuous function class. Note that

$$\begin{aligned}\mathbf{u}(\mathbf{g}^*,[m^*,\widehat{m}];\mathbf{x},u)(1,0)' \;&=\; \mathbf{u}(\mathbf{g}^*,[m^*,\widehat{m}];\mathbf{x},u)(1,0)' - \mathbf{u}(\{\widehat{\mathbf{g}},\widehat{\mathbf{m}}\}_{nk};\mathbf{x},u)(1,0)'\\ &\quad + \mathbf{u}(\{\widehat{\mathbf{g}},\widehat{\mathbf{m}}\}_{nk};\mathbf{x},u)(1,0)' - \mathbf{U}(\{[\widehat{\mathbf{g}},\widehat{\dot{\mathbf{g}}}],\widehat{\mathbf{m}}\}_{nk};\mathbf{x},u).\end{aligned}$$

It follows from (ii) and (iii) that $\mathbf{u}(\mathbf{g}^*,[m^*,\widehat{m}];\mathbf{x},u)=0$ over $(\mathbf{x},u)\in[-1,1]^{d+1}$. Since $\mathbf{u}(\boldsymbol{\zeta},[m_1,\widehat{m}];\mathbf{x},u)=0$ has a unique root at $[\mathbf{g},m]$ by condition (A5), we conclude that $[\mathbf{g},m]=[\mathbf{g}^*,m^*]$, which ensures the uniform consistency of $\widehat{\mathbf{g}}$ and $\widehat{m}$. This completes the proof of Theorem 1.

Proof of (i). For convenience, we only give the proof of $\|U_j(\boldsymbol{\beta},\mathbf{m};x)-\mathbf{u}_j(\boldsymbol{\zeta},\mathbf{m};x)(1,0)'\|$. The similar arguments lead to the conclusions about $S_m(\boldsymbol{\zeta},\mathbf{m};u)$ and $S_\ell(\boldsymbol{\zeta},m_1,\boldsymbol{\sigma};z)$. To estimate $U_j(\beta,\mathbf{m};x)-\mathbf{u}_j(\boldsymbol{\zeta},\mathbf{m};x)(1,0)'$, we consider the following decomposition,

$$\begin{aligned}&U_j(\boldsymbol{\beta},\mathbf{m};x)-\mathbf{u}_j(\boldsymbol{\zeta},\mathbf{m};x)(1,0)'\\ &\quad=\{U_j(\boldsymbol{\beta},\mathbf{m};x)-U_j([\boldsymbol{\zeta},0],\mathbf{m};x)\}+\{U_j([\boldsymbol{\zeta},0],\mathbf{m};x)-\widetilde{\mathbf{u}}_j(\boldsymbol{\zeta},\mathbf{m};x)\}\\ &\qquad\qquad+[\widetilde{\mathbf{u}}_j(\boldsymbol{\zeta},\mathbf{m};x)-\mathbf{u}_j(\boldsymbol{\zeta},\mathbf{m};x)(1,0)']\\ &\quad\equiv I_1+I_2+I_3,\end{aligned}$$

where $\widetilde{\mathbf{u}}_j(\boldsymbol{\zeta},\mathbf{m};x)$ is the mean of $U_j([\boldsymbol{\zeta},0],\mathbf{m};x)$.

First consider I_1. Denote the modulus of continuity of f by $\mathbf{w}_f(h)$. Observing that

$$\|I_1\| \leq \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_{m_1}[\gamma_j(x)(X_{ij}-x)]\|\Upsilon_{ij}(x)\| \times \left|m_2\left(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i)\right)\right| \frac{K_{h_j}(X_{ij}-x)}{V\big(m_1(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))\big)}.$$

For any given $i \in \{1, 2, ..., n\}$ and any bounded function γ_j, note that

$$\begin{aligned}
&\mathbf{w}_{m_1}[\gamma_j(x)(X_{ij}-x)]\|\Upsilon_{ij}(x)\| \times \left|m_2\left(\mathbf{e}'\zeta(\mathbf{X}_i)\right)\right| \frac{K_{h_j}(X_{ij}-x)}{V\big(m_1(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))\big)} \\
&\leq \mathbf{w}_{m_1}(C[X_{ij}-x])\|\Upsilon_{ij}(x)\| \times \left|m_2\left(\mathbf{e}'\zeta(\mathbf{X}_i)\right)\right| \frac{K_{h_j}(X_{ij}-x)}{V\big(m_1(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))\big)},
\end{aligned}$$

and it is easy to check that

$$\int_{-1}^{1} \mathbf{w}_{m_1}(u_j-x)K_{h_j}(u_j-x)f_j(u_j)du_j = O_p(\mathbf{w}_{m_1}(h_j)) \quad \text{for all}\, x \in [-1,1].$$

For any bounded functions $\boldsymbol{\beta}$ and $\mathbf{m}$, this follows from Condition A3 that

$$E[\|I_1\|] \leq O_p(\mathbf{w}_{m_1}(h_j)).$$

Hence Lemma 1 shows that, for any given continuous functions $\boldsymbol{\beta}$ and $\mathbf{m}$,

$$\sup_{x\in[-1,1]} \|I_1\| = O_p(\mathbf{w}_{m_1}(h_j)) + \mathcal{O}_p((\log n)^{1/2}(nh_j)^{-1/2}) \to 0 \quad \text{as } n \to \infty. \tag{7.1}$$

To estimate I_2, it suffices to verify the conditions given in Lemma 2. Condition (A1) means that K_h lies in a Sobolev space denoted by H^2 with the property: $\|f\|_\infty \leq c\|f\|_{\mathrm{H}^2}$, for any $f \in \mathrm{H}^2$. Then using Condition A3, we have

$$\sup_{x\in[-1,1]} \|I_2\| \leq O_p((\log n)^{1/2}(nh_j)^{-1/2}). \tag{7.2}$$

Next consider I_3. By replacing $\zeta_j(x)$ with $\zeta_j(X_{ij})$ in $\widetilde{\mathbf{u}}_j(\zeta, \mathbf{m}; x)$, a difference controlled by $\mathbf{w}_{(m_1\circ\zeta_j)}(h_j)$ is caused for all $x \in [-1,1]$. Moreover, we note that

$$\begin{aligned}
&\int_{[-1,1]^d} \left[m\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right) - m_1\left(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i)\right)\right] \times m_2\left(\mathbf{e}'\zeta(\mathbf{X}_i)\right) K_{h_j}(X_{ij}-x)/V\big(m_1(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))\big)dF(\mathbf{X}_i) \\
&\quad = \int_{-1}^{1} H(u_j)K_{h_j}(u_j-x)f_j(u_j)du_j \to \mathbf{u}_j(\boldsymbol{\zeta}, \mathbf{m}, \sigma_1; x) \;\text{as}\, h_j \to 0,
\end{aligned}$$

where F is the joint distribution function of $\mathbf{X}_i$,

$H(u_j) = E\left[[m\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right) - m_1\left\{\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i)\right\}] \times m_2(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))/V\big(m_1(\mathbf{e}'\boldsymbol{\zeta}(\mathbf{X}_i))\big)|X_{ij} = u_j\right].$

By Lemma 1, it can be shown that

$$\sup_{x\in[-1,1]} \|I_3\| = O_p(\mathbf{w}_{(m_1\circ\zeta_j)}(h_j)). \tag{7.3}$$

Thus we complete the proof of (i) by combining (7.1), (7.2) with (7.3).

Proof of (ii). Noting that $\mathbf{x}$ and u are bounded, the arguments used to prove (ii) is essentially the same with those in Chen, et.al (2009). In fact, uniform laws of large numbers for infinite space plays an important role in proving this part.

Proof of (iii). Because the proofs involving these two components are the same with each other, we only give the proof for $\widehat{\mathbf{g}} \in \mathcal{B}_n^d$.

Given any $x_1, x_2 \in [-1,1]$ with $|x_1 - x_2| \le h_j$, denoting the first component of U_j by U_{j1}. Since $U_{j1}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{m}}; x_1) = 0$ and $U_{j1}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{m}}; x_2) = 0$. By the Taylor expansion and Condition A1, it follows that

$$\begin{aligned}
&U_{j1}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{m}}; x_1) - U_{j1}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{m}}; x_2) = \frac{1}{n}\sum_{i=1}^{n}\Big[Y_i - \widehat{m}\Big\{\mathbf{e}'_{-j}\widehat{\mathbf{g}}(\mathbf{X}_i) + \widehat{g}_j(x_1) + \widehat{\dot{g}}_j(x_1)(X_{ij} - x_1)\Big\}\Big] \\
&\times\frac{\widehat{\dot{m}}\left(\mathbf{e}'\widehat{\mathbf{g}}(\mathbf{X}_i)\right)}{V\big(\widehat{m}(\mathbf{e}'\widehat{\mathbf{g}}(\mathbf{X}_i))\big)}[K_{h_j}(X_{ij} - x_1) - [K_{h_j}(X_{ij} - x_2)] + \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\widehat{\dot{m}}\left(\mathbf{e}'\widehat{\mathbf{g}}(\mathbf{X}_i)\right)}{V\big(\widehat{m}(\mathbf{e}'\widehat{\mathbf{g}}(\mathbf{X}_i))\big)}K_{h_j}(X_{ij} - x_2)\Big] \\
&\times\dot{\widehat{m}}\Big(\mathbf{e}'_{-j}\widehat{\mathbf{g}}(\mathbf{X}_i) + \widehat{g}_j(x_1) + \widehat{\dot{g}}_j(x_1)(X_{ij} - x_1)\Big)\{\widehat{g}_j(x_2) - \widehat{g}_j(x_1) - \widehat{\dot{g}}_j(x_1)(x_2 - x_1) \\
&\quad+(\widehat{\dot{g}}_j(x_2) - \widehat{\dot{g}}_j(x_1))(X_{ij} - x_1)\} + \mathcal{O}\left\{(\widehat{g}_j(x_2) - \widehat{g}_j(x_1))^2 + (x_2 - x_1)^2 + b_n^2 + b_n(\widehat{g}_j(x_2) - \widehat{g}_j(x_1))\right\} \\
&\equiv I \times (x_2 - x_1) + II \times (\widehat{g}_j(x_2) - \widehat{g}_j(x_1)) + II \times \widehat{\dot{g}}_j(x_1)(x_2 - x_1) + III \times (\widehat{\dot{g}}_j(x_2) - \widehat{\dot{g}}_j(x_1)) \\
&\qquad+O\left\{(\widehat{g}_j(x_2) - \widehat{g}_j(x_1))^2 + (x_2 - x_1)^2 + b_n^2 + b_n(\widehat{g}_j(x_2) - \widehat{g}_j(x_1))\right\}
\end{aligned}$$

By the similar discussion in Cai, Fan and Yao (2000), we have $I = \mathcal{O}_p(1)$, $II = \mathcal{O}_p(1)$ and $III = \mathcal{O}_p(b_n)$. Note that $\widehat{\dot{g}}_j$ is bounded, (iii) is held immediately.

Proof of Theorem 2.

For convenience of notation, denote $H_d = diag(h_1, \cdots, h_d)$,

$$a_n = \max_{1\le k\le d+1}\{h_k^2 + (nh_k)^{-1/2}(\log n)^{1/2}\},\ c_n = \sup_{\mathbf{x}\in[-1,1]^d}\|\widehat{\mathbf{g}}(\mathbf{x}) - \mathbf{g}(\mathbf{x})\|,$$
$$d_n = \sup_{\mathbf{x}\in[-1,1]^d}\|H_d\widehat{\dot{\mathbf{g}}}(\mathbf{x}) - H_d\dot{\mathbf{g}}(\mathbf{x})\|,\ e_n = \sup_{u\in[-1,1]}|\widehat{m}(u) - m(u)|,$$
$$\pi_n = \sup_{u\in[-1,1]}|h_m\widehat{\dot{m}}(u) - h_m\dot{m}(u)|, \quad \mu_2 = \int_0^1 x^2K(x)dx, \quad \mathbb{Q} = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix}.$$

First, we claim that uniformly over $x \in [-1,1]$, we have

$$U_j([\hat{\mathbf{g}}, \hat{\dot{\mathbf{g}}}], [\hat{m}, \hat{\dot{m}}]; x) - U_j([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; x) = (1,0)'[\mathrm{H}_{\mathbf{g}j}(\widehat{\mathbf{g}} - \mathbf{g})(x)] + \mathrm{H}_{\mathbf{m}j}(\widehat{m} - m)(x)$$
$$+O_p\big(c_n(a_n + c_n + d_n + e_n) + d_n(a_n + d_n + e_n + \pi_n)\big), \tag{7.4}$$

where $\mathrm{H}_{\mathbf{g}j}$ is an integral-type map from $\mathcal{C}_d$ to $\mathcal{C}_1$ and $\mathrm{H}_{\mathbf{m}j}$ is an integral operator on $\mathcal{C}_1$, both are defined in Appendix A.

To prove (7.4), we write,

$$U_j([\hat{\mathbf{g}}, \hat{\dot{\mathbf{g}}}], [\hat{m}, \hat{\dot{m}}]; x) - U_j([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; x) \equiv J_1 + J_2$$

where

$$J_1 = U_j([\hat{\mathbf{g}}, \hat{\dot{\mathbf{g}}}], [\hat{m}, \hat{\dot{m}}]; x) - U_j([\mathbf{g}, \dot{\mathbf{g}}], [\hat{m}, \hat{\dot{m}}]; x),$$
$$J_2 = U_j([\mathbf{g}, \dot{\mathbf{g}}], [\hat{m}, \hat{\dot{m}}]; x) - U_j([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; x).$$

Similar to the proof of Theorem 1, we can show that $\hat{\dot{g}}_j$ and $\hat{\dot{m}}_j$ are both bounded with high probability, furthermore $\|\hat{\dot{g}}_j - \dot{g}_j\| \to 0$ and $\|\hat{\dot{m}}_j - \dot{m}_j\| \to 0$. Thus by uninform the laws of large numbers and Taylor expansion of $m(\cdot)$, we conclude that

$$J_1 = (1,0)'\big(\mathrm{H}_{\mathbf{g}j}(\widehat{\mathbf{g}} - \mathbf{g})\big)(x) + O_p(a_n(c_n + d_n + \pi_n) + c_ne_n + c_n^2 + d_n(d_n + c_n + \pi_n + e_n)).$$

Similarly, we can obtain that

$$J_2 = (1,0)'\mathrm{H}_{\mathbf{m}j}(\hat{m} - m)(x) + O_p(a_ne_n + a_n\pi_n + e_n\pi_n).$$

Consequently, this together with J_1 and J_2 yields the conclusion of (7.4).

Next we consider $S_m(\widehat{\mathbf{g}}, [\widehat{m}, \widehat{\dot{m}}]; u) - S_m(\mathbf{g}, [m, \dot{m}]; u)$, which can be decomposed into the following aspects: $J_3 + J_4$, where

$$J_3 := S_m(\widehat{\mathbf{g}}, [\widehat{m}, \widehat{\dot{m}}]; u) - S_m(\widehat{\mathbf{g}}, [m, \dot{m}]; u)$$

and

$$J_4 := S_m(\widehat{\mathbf{g}}, [m, \dot{m}]; u) - S_m(\mathbf{g}, [m, \dot{m}]; u).$$

By a simple calculation, we easily obtain that

$$\begin{aligned} J_3 &= \frac{1}{n}\sum_{i=1}^{n} \big[\hat{m}(u) - m(u) + (\hat{\dot{m}} - \dot{m})(u)(\mathbf{e}'\hat{\mathbf{g}}(\mathbf{X}_i) - u)\big] K_{h_m}(\mathbf{e}'\hat{\mathbf{g}}(\mathbf{X}_i) - u) \frac{W_i(\hat{\mathbf{g}}; u)}{V\big(\widehat{m}(\mathbf{e}'\hat{\mathbf{g}}(\mathbf{X}_i))\big)} \\ &= \mathbb{Q}\big[(\hat{m} - m)(u), h_m(\hat{\dot{m}} - \dot{m})(u)\big] f_{\hat{\mathbf{g}}}(u)/V(m(u)) + O_p(e_n^2 + e_n\pi_n + a_n e_n + a_n\pi_n) \\ &= \mathbb{Q}\big[(\hat{m} - m)(u), h_m(\hat{\dot{m}} - \dot{m})(u)\big] f_{\mathbf{g}}(u)/V(m(u)) + O_p(e_n^2 + e_n\pi_n + d_n c_n + d_n\pi_n + a_n e_n) \end{aligned}$$

where the second equality is derived by repeating the process for bounding J_1 as above, and the last one is derived by taking the Taylor expansion of $f_{\hat{\mathbf{g}}}(u)$ at $\mathbf{e}'\mathbf{g}$. Similarly, we also have that

$$J_4 = (1,0)'E[m(\mathbf{e}'\mathbf{g}(X)) - m(\mathbf{e}'\hat{\mathbf{g}}(X))\big|\, \mathbf{e}'\mathbf{g}(X) = u] f_{\mathbf{g}}(u)/V(m(u)) + O_p(d_n^2 + a_n d_n + a_n c_n + c_n d_n).$$

Hence, combining with J_3 and J_4, we have that

$$\begin{aligned} &S_m(\widehat{\mathbf{g}}, [\widehat{m}, \widehat{\dot{m}}]; u) - S_m(\mathbf{g}, [m, \dot{m}]; u) = \mathbb{Q}\big[(\hat{m} - m)(u), h_m(\hat{\dot{m}} - \dot{m})(u)\big] f_{\mathbf{g}}(u)/V(m(u)) \\ &+(1,0)'\big(\mathbf{H}_{mg}\big)(\hat{\mathbf{g}} - \mathbf{g})(u) + O_p(d_n^2 + d_n c_n + d_n\pi_n + e_n^2 + e_n\pi_n + a_n e_n). \qquad (7.5) \end{aligned}$$

On the other hand, using Condition (A1), Lemma 1 implies that

$$\sup_{(\mathbf{x},u)\in[-1,1]^{d+1}} \|\mathbf{U}([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; \mathbf{x}, u)\| = O_p(a_n).$$

Note that $\mathbf{U}([\hat{\mathbf{g}}, \hat{\dot{\mathbf{g}}}], [\hat{m}, \hat{\dot{m}}]; \mathbf{x}, u) = 0$, and it follows from the second component in (7.4) and (7.5) that $d_n = O_p(a_n + c_n(b_n + c_n + e_n + d_n))$ and $\pi_n = O_p(a_n + e_n(b_n + c_n + e_n))$. This further implies that

$$d_n = O_p(a_n + c_n(b_n + c_n + e_n)). \qquad (7.6)$$

Let U_{j1} be the first component of U_j, and denote $\mathbf{U}^1 = (U_{11}, ..., U_{d1})'$. Combining (7.4) with (7.6) we have

$$\mathrm{H}_{\mathbf{g}}(\widehat{\mathbf{g}} - \mathbf{g})(\mathbf{x}) + \mathrm{H}_{\mathbf{m}}(\widehat{m} - m)(\mathbf{x}) = -\mathbf{U}^1([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; \mathbf{x}) + O_p(c_n b_n + c_n e_n + e_n a_n) \quad (7.7)$$

By the same way as above, let S_{m1} be the first components of S_m, and it follows from (7.5) that

$$\big(\hat{m} - m\big)(u) + \big(\mathbf{H}_{mg}\big)(\widehat{\mathbf{g}} - \mathbf{g})(u) = -S_{m1}(\mathbf{g}, [m, \dot{m}]; u) + O_p(a_n b_n + a_n e_n + c_n e_n). \quad (7.8)$$

Consequently, this together with (7.7) and (7.8) implies that

$$\begin{aligned}[\mathrm{H}_{\mathbf{g}} - \mathrm{H}_{\mathbf{m}} \circ \mathrm{H}_{mg}](\widehat{\mathbf{g}} - \mathbf{g})(\mathbf{x}) &= -\mathbf{U}^1([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; \mathbf{x}) \\ &+ \mathrm{H}_{\mathbf{m}}(S_{m1}(\mathbf{g}, [m, \dot{m}]))(\mathbf{x}) + O_p(a_n b_n + a_n e_n + c_n e_n), \quad (7.9)\end{aligned}$$

where we used the fact that $\mathrm{H}_{\mathbf{m}}$ is a bounded operator on $\mathcal{C}_1$. Following Condition A7, $(\mathrm{H}_{\mathbf{g}} - \mathrm{H}_{\mathbf{m}} \circ \mathrm{H}_{mg})^{-1}$ exists and is bounded on $\mathcal{C}_d$, the supremum norm of the left-side hand of (7.9) is equivalent to c_n.

Besides, Lemma 1 can show that $\big\|\mathbf{U}^1([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; \cdot) - \mathrm{H}_{\mathbf{m}}(S_{m1}(\mathbf{g}, [m, \dot{m}]))\big\| = O_p(a_n)$. By (7.9), we have $c_n = O_p(a_n + a_n e_n)$. Similarly, it can be shown that $e_n = O_p(a_n + a_n c_n)$ and $\nu_n = \mathcal{O}(a_n)$. Hence we have

$$c_n = O_p(a_n) = O_p(e_n). \quad (7.10)$$

Then by Condition A7, we note that Ψ is linear and so Ψ^{-1} is linear and bounded on $\mathcal{C}_d \times \mathcal{C}_1 \times \mathcal{C}_1$. Combining (7.7), (7.8) and (7.10), we have

$$\Psi \begin{pmatrix} \widehat{\mathbf{g}} - \mathbf{g} \\ \widehat{m} - m \end{pmatrix} (\mathbf{x}, u) = \begin{pmatrix} -\mathbf{U}^1([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; \mathbf{x}) \\ -S_{m1}(\mathbf{g}, [m, \dot{m}]; u) \end{pmatrix} + O_p(a_n b_n). \quad (7.11)$$

On the other hand, note that $U_{j1}([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}], \sigma_\ell; ; x)$ can be expressed as

$$U_{j1}([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}]; x) = V_{n,j}(x) + B_{n,j}(x)$$

where

$$V_{n,j}(x) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - m\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right)\right]\dot{m}\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right)\frac{K_{h_j}(X_{ij}-x)}{V\big(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))\big)}$$
$$B_{n,j}(x) = \frac{1}{n}\sum_{i=1}^{n}[\dot{m}\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right)]^2\ddot{g}_j(x)(X_{ij}-x)^2\frac{K_{h_j}(X_{ij}-x)}{V\big(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))\big)} + o_p(h_j^2).$$

Lemma 1 is applied again to show that

$$B_{n,j}(x) = \mu_2\ddot{g}_j(x)E\left[[\dot{m}\left(\mathbf{e}'\mathbf{g}(X)\right)]^2/V\big(m(\mathbf{e}'\mathbf{g}(X))\big)|X_{\cdot j}=x\right]f_j(x)h_j^2 + o_p(h_j^2).$$

Similarly, we can obtain

$$S_{m1}(\mathbf{g},[m,\dot{m}];u) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))]\frac{K_{h_m}\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)-u\right)}{V\big(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))\big)} + B_{d+1}(u)h_m^2 + o_p(h_m^2),$$

where $B_{d+1}(\cdot)$ is defined in Appendix A. Denote $V_{n,d+1}(u)$ by

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))]\frac{K_{h_m}\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)-u\right)}{V\big(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))\big)}$$

and we write $\mathrm{A}_n(\mathbf{x}) = -(V_{n,1}(x_1), V_{n,2}(x_2), ..., V_{n,d}(x_d), V_{n,d+1}(u))$. Then it follows from (7.11) that

$$\Psi\left(\begin{array}{c}\widehat{\mathbf{g}}-\mathbf{g}\\ \widehat{m}-m\end{array}\right)(\mathbf{x},u) = \mathrm{A}_n(\mathbf{x},u) + H^2\mathrm{B}(\mathbf{x},u) + O_p(a_nb_n) + o_p(\sum_{k=1}^{d+1}h_k^2),$$

where $H = diag(h_1,\cdots,h_{d+1})$ and $\mathrm{B}(\mathbf{x},u)$ is defined in Appendix A. The classical central limit theorem implies that $(nH)^{1/2}\mathrm{A}_n(\mathbf{x},u)$ is asymptotically normal with zero mean and finite covariance matrix $\mathbf{M}$. Thus the proof of Theorem 2 is completed.

Proof of Theorem 3.

First, we derive the asymptotic variance of $\sum_{j=1}^{d}\int_{-1}^{1}\widehat{g}_j\psi_j(x)dx+\int_{-1}^{1}\widehat{m}(u)\psi_m(u)du$. Conditioned on h_j and h_m, from (7.11), we have

$$\sum_{j=1}^{d}\int_{-1}^{1}[\widehat{g}_j(x)-g_j(x)]\psi_j(x)dx+\int_{-1}^{1}(\widehat{m}-m)(u)\psi_m(u)du$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-m\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right)\right]\dot{m}\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right)\sum_{j=1}^{d}\frac{?\phi_j(X_{ij})}{\mathrm{H}_{\mathbf{g}j}(\mathbf{e}_j)(X_{ij})}+\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i-m\left(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)\right)}{V(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)))}?\phi_m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i))$$
$$+O_p(a_nb_n). \tag{7.12}$$

where we used the fact that $\int_{-1}^{1}xK(x)dx=0$. Using the classical central limit theorem, we get

$$\sum_{j=1}^{d}\int_{-1}^{1}[\widehat{g}_j(x)-g_j(x)]\psi_j(x)dx+\int_{-1}^{1}(\widehat{m}-m)(u)\psi_m(u)du\to N(0,\sigma_v^2), \tag{7.13}$$

where σ_v^2 is defined in Section 3. In particular, the quantity (7.12) shows that the variance of $\sum_{j=1}^{d}\int_{-1}^{1}\widehat{g}_j(x)\psi_j(x)dx+\int_{-1}^{1}\widehat{m}(u)\psi_m(u)du$ equals to

$$\bar{\sigma}^2=E\left[\Big(\dot{m}\left(\mathbf{e}'\mathbf{g}(X)\right)\mathbf{e}'\widetilde{\boldsymbol{\phi}}(X)+\phi_m(\mathbf{e}'\mathbf{g}(X))\Big)^2\right]\Big/\Big(n\sum_{j=1}^{d+1}h_j\Big). \tag{7.14}$$

Next show the asymptotic efficiency of $\sum_{j=1}^{d}\int_{-1}^{1}\widehat{g}_j\psi_j(x)dx+\int_{-1}^{1}\widehat{m}(u)\psi_m(u)du$. Consider the following parametric submodel with unknown parametric β,

$$(\mathbf{g}(\mathbf{x},\beta),m(u,\beta))=\big(\mathbf{g}(\mathbf{x}),m(u)\big)+\beta(\widetilde{\boldsymbol{\phi}}(\mathbf{x}),(V\big(m(g(\mathbf{x}))\big)\phi_m)(u)\big).$$

Obviously, $\beta_0=0$ is the true value of β. Based on the definition of the quasi-likelihood (2.1), the score of this parametric submodel at β_0 is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[Y_i-m\left(\mathbf{e}'\mathbf{g}(X_i)\right)\right]\dot{m}\left(\mathbf{e}'\mathbf{g}(X_i)\right)\mathbf{e}'\widetilde{\boldsymbol{\phi}}(X_i)+\frac{Y_i-m\left(\mathbf{e}'\mathbf{g}(X_i)\right)}{V(m(\mathbf{e}'\mathbf{g}(\mathbf{X}_i)))}\phi_m(\mathbf{e}'\mathbf{g}(X_i)), \tag{7.15}$$

whose variance is $\bar{\sigma}^2$. Thus, the maximum likelihood estimator of β, denoted by $\widetilde{\beta}$, satisfies

$$\sqrt{n}(\widetilde{\beta}-\beta_0)\to N(0,(\bar{\sigma}^2)^{-1}).$$

For any vector functions $\boldsymbol{\psi}(\mathbf{x}, u) = (\{\psi_j(x)\}_{j=1}^d, \psi_m(u))$, we observe that

$$\begin{aligned}\int_{[\mathbf{x},u]\in[-1,1]^{d+1}} &[(\mathbf{g}(\mathbf{x}, \widetilde{\beta}), m(u, \widetilde{\beta})) - (\mathbf{g}(\mathbf{x}, \beta_0), m(u, \beta_0))]\boldsymbol{\psi}(\mathbf{x}, u)' d\mathbf{x} du \\ &= (\widetilde{\beta} - \beta_0) \int_{[\mathbf{x},u]\in[-1,1]^{d+1}} [\widetilde{\boldsymbol{\phi}}(\mathbf{x}), \phi_m(u)]\boldsymbol{\psi}(\mathbf{x}, u)' d\mathbf{x} du. \qquad (7.16)\end{aligned}$$

Moreover, we observe that

$$\int_{[\mathbf{x},u]\in[-1,1]^{d+1}} [\widetilde{\boldsymbol{\phi}}(\mathbf{x}), (\sigma^2\phi_m)(u)]\boldsymbol{\psi}(\mathbf{x}, u)' d\mathbf{x} du = \bar{\sigma}^2.$$

Then it follows from (7.16) that

$$\sqrt{n} \int_{[\mathbf{x},u]\in[-1,1]^{d+1}} [(\mathbf{g}(\mathbf{x}, \widetilde{\beta}), m(u, \widetilde{\beta})) - (\mathbf{g}(\mathbf{x}, \beta_0), m(u, \beta_0))]\boldsymbol{\psi}(\mathbf{x}, u)' d\mathbf{x} du \to N(0, \bar{\sigma}^2).$$

This together with (7.14) shows that the asymptotic variance of $\sum_{j=1}^d \int_{-1}^1 \widehat{g}_j(x)\psi_j(x)dx + \int_{-1}^1 \widehat{m}(u)\psi_m(u)du$ is the same as that of $\int_{[\mathbf{x},u]\in[-1,1]^{d+1}} (\mathbf{g}(\mathbf{x}, \widetilde{\beta}), m(u, \widetilde{\beta}))\boldsymbol{\psi}(\mathbf{x}, u)' d\mathbf{x} du$. In other words, $\sum_{j=1}^d \int_{-1}^1 \widehat{g}_j(x)\psi_j(x)dx + \int_{-1}^1 \widehat{m}(u)\psi_m(u)du$ is asymptotically efficient for the estimation of $\sum_{j=1}^d \int_{-1}^1 g_j\psi_j(x)dx + \int_{-1}^1 m(u)\psi_m(u)du$. Thus we complete the proof of Theorem 3.

Convergence of the iterative algorithm

Let $\mathbf{u}(\boldsymbol{\beta}, \mathbf{m}; \mathbf{x}, u)$ be defined as above. In the proof of Theorem 1, we have shown that, there exists a solution of $\mathbf{U}(\boldsymbol{\beta}, \mathbf{m}; \mathbf{x}, u) = 0$ denoted by $([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})$, such that $([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})$ converges to $([\mathbf{g}, \dot{\mathbf{g}}], [m, \dot{m}])$ in the sup-norm on $\mathbf{x} \in [-1, 1]$. We claim that, on a neighbor of $([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})$, the proposed iterative algorithms converges to $([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})$ as the number of iterative steps increase. It can be checked from (2.8) and (2.9) that, the solution of $([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})(\mathbf{x})$ is determined by the solutions of $([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})(\mathbf{x}_1)$,...,$([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})(\mathbf{x}_n)$. Thus, it suffices to consider the solutions of $([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})(\mathbf{x}_1)$, ..., $([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})(\mathbf{x}_n)$. In particular, for some $r \in \mathbb{N}^+$, provided that $\|([\mathbf{g}^{(r)}, \dot{\mathbf{g}}^{(r)}], \mathbf{m}^{(r)}) - ([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})\| = \tilde{a}_n$ on all the sample points, we need to show that $\|([\mathbf{g}^{(r+1)}, \dot{\mathbf{g}}^{(r+1)}], \mathbf{m}^{(r+1)}) - ([\widehat{\mathbf{g}}, \widehat{\dot{\mathbf{g}}}], \widehat{\mathbf{m}})\| \le \theta\tilde{a}_n + \tilde{b}_n$ with some $\theta \in (0, 1)$, where $\tilde{b}_n$ is referred to as statistical error independent of our iterative algorithm. That is, this implies that the proposed iteramtive algorithm locally converges with linear order.

To this end, we write $\Pi_{kj}^{(r)} := \sum_{i=1}^{n} \rho^{(2r)}(\mathbf{X}_i)\Upsilon_{ij}(x)\Upsilon_{ij}(x)' K_{h_j}(X_{ij}-X_{kj})/V(\mu_i^{(r)})$ for the ease of symbols. For any X_{kj} with $k=1,...,n$ and $j=1,...,d$, we obtain from (2.8) that

$$\begin{pmatrix} \big(g_j^{(r+1)}-\hat{g}_j\big)(X_{kj}) \\ \big(\dot{g}_j^{(r+1)}-\hat{\dot{g}}_j\big)(X_{kj}) \end{pmatrix} := \big(\Pi_{kj}^{(r)}\big)^{-1} \times \Theta_j^{(r)}, \tag{7.17}$$

where

$$\begin{aligned}
\Theta_j^{(r)} &= \sum_{i=1}^{n}\Big(Y_i-\mu_i^{(r)}+\rho^{(r)}(\mathbf{X}_i)g_j^{(r)}(X_{ij})\Big)\Big(\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\Big)-\Pi_j^{(r)}\begin{pmatrix}\hat{g}_j(X_{kj})\\ \hat{\dot{g}}_j(X_{kj})\end{pmatrix}\\
&= \sum_{i=1}^{n}\Big(Y_i-\mu_i^{(r)}+\rho^{(r)}(\mathbf{X}_i)g_j^{(r)}(X_{ij})-\rho^{(r)}(\mathbf{X}_i)\hat{g}_j(X_{kj})-\rho^{(r)}(\mathbf{X}_i)(X_{ij}-X_{kj})\hat{\dot{g}}_j(X_{kj})\Big)\\
&\quad \Big(\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\Big)\\
&= \sum_{i=1}^{n}\Big(Y_i-m\big(\sum_{j=1}^{d}g(X_{ij})\big)+\hat{m}\big(\sum_{j=1}^{d}\hat{g}(X_{ij})\big)-\mu_i^{(r)}+m\big(\sum_{j=1}^{d}g(X_{ij})\big)-\hat{m}\big(\sum_{j=1}^{d}\hat{g}(X_{ij})\big)\\
&\qquad +\rho^{(r)}(\mathbf{X}_i)\big[g_j^{(r)}(X_{ij})-\hat{g}_j(X_{kj})\big]\Big)\Big(\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\Big)\\
&\qquad -\sum_{i=1}^{n}\rho^{(r)}(\mathbf{X}_i)(X_{ij}-X_{kj})\hat{\dot{g}}_j(X_{kj})\Big(\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\Big)\\
&=: \Omega_1+\Omega_2+\Omega_3+\Omega_4
\end{aligned}$$

where

$$\begin{aligned}
\Omega_1 &= \sum_{i=1}^{n}\Big(\hat{m}\big(\sum_{j=1}^{d}\hat{g}(X_{ij})\big)-\mu_i^{(r)}+\rho^{(r)}(\mathbf{X}_i)\big[g_j^{(r)}(X_{ij})-\hat{g}_j(X_{kj})\big]\Big)\Big(\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\Big)\\
\Omega_2 &= \sum_{i=1}^{n}\Big(m\big(\sum_{j=1}^{d}g(X_{ij})\big)-\hat{m}\big(\sum_{j=1}^{d}\hat{g}(X_{ij})\big)\Big)\Big(\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\Big)\\
\Omega_3 &= \sum_{i=1}^{n}\Big(Y_i-m\big(\sum_{j=1}^{d}g(X_{ij})\big)\Big)\Big(\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\Big)\\
\Omega_4 &= \sum_{i=1}^{n}\rho^{(r)}(\mathbf{X}_i)(X_{kj}-X_{ij})\hat{\dot{g}}_j(X_{kj})\Big(\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\Big).
\end{aligned}$$

We call Ω_1 as algorithmic error, since it is induced by the r-th step iteration. Obviously, Ω_2 is referred to as estimation error, and Ω_3 is mainly determined by the noise error of our model. Ω_4 reflects the error induced by the local linear approximation to nonparametric functions.

From the decomposition of $\Theta_j^{(r)}$ as above, we see that Ω_1 is the most important term for justifying the local convergence of our proposed algorithm. Furthermore, Ω_1 can be bounded by the following two terms:

$$\left\|\left(\Pi_{kj}^{(r)}\right)^{-1}\Omega_1\right\|_\infty \leq \sum_{i=1}^{n}\left((\|\dot{\hat{m}}\|_\infty + |\rho^{(r)}(\mathbf{X}_i)|)\left|g_j^{(r)}(X_{ij}) - \hat{g}_j(X_{kj})\right|\right)\Pi_{ijk}^{(r)} + \sum_{i=1}^{n}\|\hat{m} - m^{(r)}\|_\infty\Pi_{ijk}^{(r)}$$

where $\Pi_{ijk}^{(r)} = \left\|\left(\Pi_{kj}^{(r)}\right)^{-1}\Upsilon_{ij}(x)\rho^{(r)}(\mathbf{X}_i)\frac{K_{h_j}(X_{ij}-X_{kj})}{V(\mu_i^{(r)})}\right\|_\infty$. It remains to prove that $\sum_{i=1}^{n}\left((\|\dot{\hat{m}}\|_\infty + |\rho^{(r)}(\mathbf{X}_i)|)\right)\Pi_{ijk}^{(r)}$ as well as $\sum_{i=1}^{n}\Pi_{ijk}^{(r)}$ is strictly less than 1. Since $m^{(r)}$ and $g_j^{(r)}$ $(j = 1, ..., d)$ are close enough to the true functions (m, g_j), $\rho^{(r)}$ and $\mu^{(r)}$ approximate well to ρ and μ respectively. If $\|\rho\|_\infty$ is strictly less than one, we can justify the above conclusion following the definition of $\left(\Pi_{kj}^{(r)}\right)^{-1}$. In addition, we observe that the other terms in $\Theta_j^{(r)}$ except Ω_1 primarily reflect statistical error and the errors induced by iterative algorithm can be negligible. Thus, we complete the proof of the algorithm convergence together with Theorem 1, which generates the parameter $\tilde{b}_n$.