Estimating High Dimensional Monotone Index Models by Iterative Convex Optimization*

Shakeeb Khan[1], Xiaoying Lan[1], Elie Tamer[2], and Qingsong Yao[3]

[1]Dept. of Economics, Boston College

[2]Dept. of Economics, Harvard University

[3]Dept. of Economics, Louisiana State University

First Version: 3/2023; This Version: April 23, 2024

Abstract

In this paper we propose new approaches to estimating large dimensional monotone index models. This class of models has been popular in the applied and theoretical econometrics literatures as it includes discrete choice, nonparametric transformation, and duration models. A main advantage of our approach is computational. For instance, rank estimation procedures such as those proposed in Han (1987) and Cavanagh and Sherman (1998) that optimize a nonsmooth, non convex objective function are difficult to use with more than a few regressors and so limits their use in with economic data sets. For such monotone index models with increasing dimension, we propose to use a new class of estimators based on batched gradient descent (BGD) involving nonparametric methods such as kernel estimation or sieve estimation, and study their asymptotic properties. The BGD algorithm uses an iterative procedure where the key step exploits a strictly convex objective function, resulting in computational advantages. A contribution of our approach is that our model is large dimensional and semiparametric and so does not require the use of parametric distributional assumptions.

Key Words Monotone Index models, Convex Optimization, Kernel and Sieve Estimation.

*We are grateful to the Editor, the Associate Editor, two anonymous referees, and conference participants at the BC/BU 2020 Econometrics Workshop, the 2019 Midwestern Econometrics Study group, the 2020 12th World Congress of the Econometric Society, the 2021 NASM of the Econometric Society, 2022 CIRAQ Econometrics conference, 2022 ISNPS conference, 2022 Advanced Methods Conference at TSE, and seminar participants from Georgetown, UC Berkeley, UC Louvain, UC Riverside, University of Bristol, UVA, University of Warwick and Yale for helpful comments.

1 Introduction

Monotone index models have received a great deal of attention in both the theoretical and applied econometrics literature, as many economic variables of interest are of a limited or qualitative nature. A leading special case in this class is the binary choice model which is usually represented by some variation of the following threshold crossing model:

$$y_i = \mathbf{1}\left(X_{0,i}\beta^\star + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^\star - u_i \geq 0\right) \tag{1}$$

where $\mathbf{1}(\cdot)$ is usual indicator function, y_i is the observed response variable, taking the values 0 or 1 and $\mathbf{X}_{e,i} = \left(X_{0,i}, \mathbf{X}_i^{\mathrm{T}}\right)^{\mathrm{T}}$ is an observed $p+1$ dimensional vector of covariates which effect the behavior of y_i. Both the scalar disturbance term u_i with distribution function denoted by $G(\cdot)$ and the $(p+1)$-dimensional true parameter vector $\boldsymbol{\beta}_e^\star = \left(\beta^\star, \boldsymbol{\beta}^{\star\mathrm{T}}\right)^{\mathrm{T}}$ are unobserved, the latter often being the parameter estimated from a random sample $(\mathbf{X}_{e,i}, y_i), i = 1, 2, \cdots, n$[1].

The disturbance term u_i is restricted in ways that ensure identification of $\boldsymbol{\beta}_e^\star$. Parametric restrictions specify the distribution of u_i up to a finite dimensional parameter and assume that u_i distributed independently of the covariates $\mathbf{X}_i$. Under such a restriction, $\boldsymbol{\beta}_e^\star$ can be estimated (up to scale) using maximum likelihood or nonlinear least squares. Estimators that are robust to these parametric distributional assumptions have been proposed and analyzed resulting in a variety of estimation procedures for $\boldsymbol{\beta}_e^\star$.

An important class of semiparametric restrictions used in the literature is based on independence/index restrictions. Estimation procedures under this restriction include those proposed by Han (1987), Ichimura (1993), Klein and Spady (1993). These cover but are not limited to the above binary response model. This class of index models have a robustness advantage over parametric approaches, but estimators within this class are difficult to compute[2] due to nonconvexity and in some cases also nonsmoothness of their respective objective functions[3]. Furthermore the difficulty increases with the dimension of $\mathbf{X}_i$. Recent work which is motivated by computational concerns is Ahn et al. (2018). However, the employs a two step procedure involving a fully nonparametric estimator in the first stage which makes it difficult to handle covariates with large dimensions.

[1] To clarify discussion, throughout the paper we will only focus on model (1). But we note that all the proposed algorithms and their theoretical properties in this paper apply directly to the most general class of monotone index models without any modifications.

[2] Other estimation of index models includes Stoker (1986) and Powell et al. (1989). While these are relatively easy to compute, such derivative based estimators cannot be applied unless all components of $\mathbf{X}_{e,i}$ are continuously distributed.

[3] Generally, finding a a local optimum is generally NP-hard, let alone the global optimum (Murty and Kabadi, 1987).

A related drawback of all these procedures is that they are designed to estimate parameters in models of a small and *fixed* dimension. A relatively recent and thriving literature in econometrics and machine learning is recognizing the many advantages of allowing for large dimensional models or models with a large set of controls. This class is a special case of models that consider the situation when the dimension of $\mathbf{X}_i$ is large, and this is now often modeled with its dimension increasing with the sample size. Due primarily to its empirical relevance, there has been a burgeoning literature on estimation and inference on certain econometric and statistics models with a large number of regressors or a large number of moment conditions. For a survey of examples in economics and finance, see Fan et al. (2020). Recent papers include Newey and Windmeijer (2009), Chernozhukov et al. (2017),Belloni et al. (2018), Cattaneo et al. (2018a), Cattaneo et al. (2018b),

Related to our work is the recent literature on estimating large dimensional binary choice or monotone index models such as Sur and Candès (2019) and Fan et al. (2020). Sur and Candès (2019) considers inference in a large dimensional logit model, relying on the logistic distribution of the disturbance term where it is shown that χ^2 asymptotic approximations of the LR statistic are suspect when the dimension of x is large. Fan et al. (2020) on the other hand estimates parameters by optimizing the objective function introduced in Han (1987), but with the number parameters increasing with the sample size. Optimizing these rank based objective functions is unfortunately hard even with recent developments in algorithms and search methods for optimizing non smooth and/or non convex objective functions. See for example important recent work based on mixed integer programming (MIP) as in, e.g. Fan et al. (2020) and Shin and Todorov (2021).

Therefore, in light of the drawbacks in the existing literature, this paper proposes a new estimation procedure that is amenable to easier computation. Specifically we aim to construct a computationally feasible estimator for a semiparametric binary choice and monotone index models with *increasing* dimension based on a convex objective function and then establish its asymptotic properties. As we will discuss in detail in the next section, our algorithm uses an iterative estimator based on a batched gradient descent (BGD) method, and we show how to use nonparametric methods to approximate the distribution in each stage of the iteration. One is kernel regression, and the other is the method of sieves[4].

[4] See Chen (2007) who pioneered the use of sieve methods in econometrics.

1.1 Notations

Throughout the rest of this paper, we will be using the following notations. For any real sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = o\left(b_n\right)$ if $\limsup_{n\to\infty} |a_n/b_n| = 0$, $a_n = O\left(b_n\right)$ if $\limsup_{n\to\infty} |a_n/b_n| < \infty$, and $a_n \sim b_n$ if both $a_n = O\left(b_n\right)$ and $b_n = O\left(a_n\right)$. For any random sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = O_p\left(b_n\right)$ if for any $0 < \tau < 1$ there are N and $C > 0$ such that $P\left\{|a_n/b_n| > C\right\} < \tau$ holds for all $n \geq N$, we write $a_n = o_p\left(b_n\right)$ if for any $C > 0$, $\lim_{n\to\infty} P\left\{|a_n/b_n| > C\right\} \to 0$. For any Borel sets $A \subseteq \mathbb{R}^k$, denote its Lebesgue measure as $m\left(A\right)$. For any symmetric matrix A, we write $A \succ 0$ if A is positive definite, and $A \succeq 0$ if A is positive semi-definite. For any symmetric matrices A and B, we write $A \succ B$ if $A - B \succ 0$ and $A \succeq B$ if $A - B \succeq 0$. For any matrix A, we denote $\sigma\left(A\right)$ as its singular value, and denote $\overline{\sigma}\left(A\right)$ and $\underline{\sigma}\left(A\right)$ as its largest and smallest singular value. For any symmetric matrix A, we denote $\lambda\left(A\right)$ as its eigenvalue, and denote $\overline{\lambda}\left(A\right)$ and $\underline{\lambda}\left(A\right)$ as its largest and smallest eigenvalue. For any vector $\mathbf{x} = \left(x_1, \cdots, x_p\right)^{\mathrm{T}}$, we denote its Euclidean norm as $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{p} x_i^2}$. For any matrices $A = \left(a_{ij}\right)_{n\times m}$, we denote $\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}$. Note that when A is positive semi-definite, there holds $\|A\mathbf{x}\| \leq \overline{\lambda}\left(A\right) \cdot \|\mathbf{x}\|$; for general square matrix A, there holds $\|A\mathbf{x}\| \leq \overline{\sigma}\left(A\right) \cdot \|\mathbf{x}\|$. Finally, for any function $f\left(\mathbf{x}\right)$ with domain D, define $\|f\|_{\infty} = \sup_{\mathbf{x}\in D} f\left(\mathbf{x}\right)$.

2 The BGD Estimator

To provide intuition for our semiparametric estimators that we introduce later, we start here by considering a simplified version of the model where the cumulative distribution function $G\left(\cdot\right)$ is completely known. Under such setup, we explore the *batch gradient descent estimator* (BGD estimator) of $\boldsymbol{\beta}_e^{\star}$ when its dimensionality p may increase, which is also important on its own right. Throughout the following analysis we assume that the data set satisfies the following assumption.

Assumption 1. *An i.i.d. data set $\mathscr{D}_n = \left\{\left(\mathbf{X}_{e,i}, y_i\right)\right\}_{i=1}^{n}$ of sample size n is observed, where y_i is generated*[5] *by $y_i = \mathbf{1}\left(X_{0,i}\beta_0^{\star} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star} - u_i \geq 0\right)$ with unobserved shock u_i that is independent of $\mathbf{X}_{e,i}$ and has CDF $G\left(\cdot\right)$.*

Given any loss function $\ell_G\left(\boldsymbol{\beta}_e, \mathbf{X}_e, y\right)$ that depends on G and is differentiable with respect

[5] Here we are decomposing the vector $\mathbf{X}_{e,i}$ into a scalar component $X_{0,i}$ and the vector $\mathbf{X}_i$, and decomposing the vector of parameters $\boldsymbol{\beta}_e^{\star}$ into the scalar term $\beta_0^{\star}$ and the vector $\boldsymbol{\beta}^{\star}$. As we will see this is done for notational convenience when imposing scale normalizations.

to $\boldsymbol{\beta}_e \in \mathcal{B}_e$, the BGD estimator of $\boldsymbol{\beta}_e^\star$ is constructed based on the following iteration,

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n} \sum_{i=1}^{n} \partial \ell_G \left(\boldsymbol{\beta}_{e,k}, \mathbf{X}_{e,i}, y_i \right) / \partial \boldsymbol{\beta}_e, \tag{2}$$

where $\delta_k > 0$ is the learning rate. Note that $n^{-1} \sum_{i=1}^{n} \partial \ell_G \left(\boldsymbol{\beta}_e, \mathbf{X}_{e,i}, y_i \right) / \partial \boldsymbol{\beta}_e$ constitutes a sample analogue of the derivative $\partial \mathbb{E} \left[\ell_G \left(\boldsymbol{\beta}_e, \mathbf{X}_e, y \right) \right] / \partial \boldsymbol{\beta}_e$. Unlike the stochastic gradient descent (SGD) algorithm, in the BGD algorithm, in each round of update we evaluate the derivative of the loss function over all data points. This increases the computational burden but provides a more accurate estimator for the derivative of the expected loss function. Given the initial guess of the parameter, $\boldsymbol{\beta}_{e,1}$, we iterate based on (2) until some terminating conditions are satisfied.

In this paper, we consider the following loss function

$$\ell_G \left(\boldsymbol{\beta}_e, \mathbf{X}_e, y \right) = \int_{-A}^{\mathbf{X}_e^{\mathrm{T}} \boldsymbol{\beta}_e} G(z)\, dz - y \mathbf{X}_e^{\mathrm{T}} \boldsymbol{\beta}_e, \tag{3}$$

for some sufficiently large positive constant A. A similar loss function to (3) was also considered in Agarwal et al. (2014). This loss function is attractive primarily because it is convex and hence easy to optimize. Moreover, under some mild conditions, we can show that

$$\frac{\partial \mathbb{E} \left(\ell_G \left(\boldsymbol{\beta}_e^\star, \mathbf{X}_e, y \right) \right)}{\partial \boldsymbol{\beta}_e} = \mathbb{E} \left\{ \left(G \left(\mathbf{X}_e^{\mathrm{T}} \boldsymbol{\beta}_e^\star \right) - \mathbb{E} \left(y | \, \mathbf{X}_e \right) \right) \mathbf{X}_e \right\} = 0,$$

and

$$\frac{\partial^2 \mathbb{E} \left(\ell_G \left(\boldsymbol{\beta}_e, \mathbf{X}_e, y \right) \right)}{\partial \boldsymbol{\beta}_e \partial \boldsymbol{\beta}_e^{\mathrm{T}}} = \mathbb{E} \left\{ G' \left(\mathbf{X}_e^{\mathrm{T}} \boldsymbol{\beta}_e \right) \mathbf{X}_e \mathbf{X}_e^{\mathrm{T}} \right\} \succ 0, \forall \boldsymbol{\beta}_e \in \mathcal{B}_e.$$

So $\boldsymbol{\beta}_e^\star$ uniquely minimizes $\mathbb{E} \ell_G \left(\boldsymbol{\beta}_e, \mathbf{X}_e, y \right)$ over $\mathcal{B}_e$.

Based on loss function (3), the BGD estimator is constructed using the following iterative procedure

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n} \sum_{i=1}^{n} \left(G \left(\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_{e,k} \right) - y_i \right) \mathbf{X}_{e,i}. \tag{4}$$

We summarize our algorithm as follows in algorithm 1.

Remark 1. *The key to the above approach is the construction of a convex objective function that facilitates computation even with high dimensions. This transformed convex objective works for any monotone model. In particular, for any model of the form* $y_i = G(\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_e^\star) + \varepsilon_i$ *with* $\mathbb{E} \left(\varepsilon_i | \mathbf{X}_{e,i} \right) = 0$ *and monotone* $G(\cdot)$*, a similar convex criterion as in (3) can be used for inference on* $\boldsymbol{\beta}_e^\star$*.*

We now describe the asymptotic properties of $\boldsymbol{\beta}_{e,k}$ based on (4). We first make the

Algorithm 1: The BGD Estimator

```
input : Data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^n$, sequence of learning rate $\{\delta_k\}_{k=1}^\infty$, initial guess
        $\boldsymbol{\beta}_{e,1}$, CDF $G(\cdot)$, and terminating condition $\mathcal{T}$
output: The BGD estimator $\widehat{\boldsymbol{\beta}}_e$
$k \leftarrow 1$;
while The terminating condition $\mathcal{T}$ is not satisfied do
    $\boldsymbol{\beta}_{e,k+1} \leftarrow \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n} \sum_{i=1}^n \left(G\left(\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_{e,k}\right) - y_i\right) \mathbf{X}_{e,i}$;
    $k \leftarrow k + 1$;
$\widehat{\boldsymbol{\beta}}_e \leftarrow \boldsymbol{\beta}_{e,k}$;
```

following assumption.

Assumption 2. *(i) $\mathcal{X}_e = [0,1]^{p+1}$; (ii) $\mathcal{B}_e$ is convex, and there exists some constant $B_0 > 0$ such that for any $\boldsymbol{\beta}_e \in \mathcal{B}_e$, $|\beta_j| \leq B_0$ for any $0 \leq j \leq p$; (iii) there exists integer υ_G such that G has up to υ_G-th bounded derivatives; (iv) Define $M_n(\boldsymbol{\beta}_e) = \frac{1}{n}\sum_{i=1}^n G'\left(\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_e\right) \mathbf{X}_{e,i} \mathbf{X}_{e,i}^{\mathrm{T}}$ and $M(\boldsymbol{\beta}_e) = \mathbb{E}[M_n(\boldsymbol{\beta}_e)]$. For any $\boldsymbol{\beta}_e \in \mathcal{B}_e$, there holds $0 < \underline{\lambda}_e \leq \underline{\lambda}(M(\boldsymbol{\beta}_e)) \leq \overline{\lambda}(M(\boldsymbol{\beta}_e)) \leq \overline{\lambda}_e < \infty$.*

Remark 2. *Assumption 2(i) and (ii) facilitate the theoretical analysis of our estimator. Note that when the space of the $\mathbf{X}_e$ is compact[6], we can always redefine the true parameter vector $\boldsymbol{\beta}_e^\star$ so that Assumption 2(i) holds. For Assumption 2(ii), note that to ensure that $\boldsymbol{\beta}_{e,k}$ falls into a compact set for each k, some form of truncation on $\boldsymbol{\beta}_{e,k+1}$ in (4) is needed. While according to our results below, as long as $\mathcal{B}_e$ is sufficiently large, it can be shown that $\boldsymbol{\beta}_{e,k}$ will fall into $\mathcal{B}_e$ for all k with probability going to 1. We then assume that $\boldsymbol{\beta}_{e,k} \in \mathcal{B}_e$ for all k. Assumption 2(iii) imposes some smoothness conditions on G, where the requirement on υ_G will be stated below. Such smoothness assumption can be easily satisfied by many commonly-used distributions such as normal, Logistic, or t distribution. Assumption 2(iv) requires that the eigenvalue of $M(\boldsymbol{\beta}_e)$ is bounded below and above uniformly over $\mathcal{B}_e$.*

For any $\boldsymbol{\beta}_e \in \mathcal{B}_e$, define $\Delta\boldsymbol{\beta}_e = \boldsymbol{\beta}_e - \boldsymbol{\beta}_e^\star$. Also define $\varepsilon_i = y_i - G\left(\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_e^\star\right)$, where $\mathbb{E}\left[\varepsilon_i | \mathbf{X}_{e,i}\right] = 0$. When Assumption 1 and Assumption 2 hold, we have the following result.

Theorem 1. *Let Assumption 1 and Assumption 2 hold. In addition, let $\upsilon_G = 3$, $p^5 (\log p)^2 n^{-1} \to 0$, and the learning rate δ_k is chosen such that $\delta_k = \delta \leq 2/\left(3\overline{\lambda}_e\right)$. Let $\boldsymbol{\beta}_e$ be updated based on algorithm 1. We have that*

[6] This condition is similarly imposed in Ichimura (1993). Note also that in Klein and Spady (1993), Condition C.4a requires that, in the single-index case, $\mathbf{X}_e^{\mathrm{T}} \boldsymbol{\beta}_e^\star$ is bounded. Since there is no further sparsity restrictions on $\boldsymbol{\beta}^\star$, this condition is also similar to our Assumption 2(i).

(i) Define

$$k_{1,n}^{BGD}=\frac{\log\left\|\Delta\boldsymbol{\beta}_{e,1}\right\|+\frac{1}{2}\log\left(n/\left(p\log p\right)\right)}{-\log\left(1-\underline{\lambda}_{e}\delta/2\right)};$$

we have that

$$\sup_{k\geq k_{1,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k}\right\|=O_{p}\left(\sqrt{p\left(\log p\right)/n}\right);$$

(ii) Define $k_{2,n}^{BGD}$ such that $\left(1-\underline{\lambda}_{e}\delta\right)^{k_{2,n}^{BGD}}\sqrt{p\log p}\rightarrow0$, we have that

$$\sup_{k\geq k_{2,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k+k_{1,n}^{BGD}}-M^{-1}\left(\boldsymbol{\beta}_{e}^{\star}\right)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i}\mathbf{X}_{e,i}\right\|=o_{p}\left(1/\sqrt{n}\right);$$

(iii) For any $k\geq k_{1,n}^{BGD}+k_{2,n}^{BGD}+1$, define $\widehat{\boldsymbol{\beta}}_{e}=\widehat{\boldsymbol{\beta}}_{k}$. Also define

$$\Sigma_{1}^{\star}=M^{-1}\left(\boldsymbol{\beta}_{e}^{\star}\right)\mathbb{E}\left[G_{i}^{\star}\left(1-G_{i}^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]M^{-1}\left(\boldsymbol{\beta}_{e}^{\star}\right),$$

and

$$\widehat{\Sigma}_{1,n}=M_{n}^{-1}\left(\widehat{\boldsymbol{\beta}}_{e}\right)\left\{\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_{i}\left(1-\widehat{G}_{i}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right\}M_{n}^{-1}\left(\widehat{\boldsymbol{\beta}}_{e}\right),$$

where $G_{i}^{\star}=G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e}^{\star}\right)$ and $\widehat{G}_{i}=G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{e}\right)$. Suppose further that $\mathbb{E}\left(\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right)$ has uniformly (with respect to p) upper bounded eigenvalues. Then we have that

$$\left\|\widehat{\Sigma}_{1,n}-\Sigma_{1}^{\star}\right\|\rightarrow_{p}0.$$

(iv) For any $p+1$ vector ρ such that $\lim_{n\rightarrow\infty}\|\rho\|<\infty$, $\lim_{n\rightarrow\infty}\rho^{\mathrm{T}}\Sigma_{1}^{\star}\rho=\sigma^{2}\left(\rho\right)$, and that $\rho^{\mathrm{T}}M^{-1}\left(\boldsymbol{\beta}_{e}^{\star}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_{i}\mathbf{X}_{e,i}\rightarrow_{d}N\left(0,\sigma^{2}\left(\rho\right)\right)$, we have that

$$\rho^{\mathrm{T}}\Delta\widehat{\boldsymbol{\beta}}_{e}/\sqrt{\widehat{\sigma}^{2}\left(\rho\right)/n}\rightarrow_{d}N\left(0,1\right),$$

where $\widehat{\sigma}^{2}\left(\rho\right)=\rho^{\mathrm{T}}\widehat{\Sigma}_{1,n}\rho$.

Proof of Theorem 1. See Section B of Supplementary Material. □

Note that unlike methods based on stochastic gradient descent (e.g., Toulis and Airoldi (2017)) where the learning rate is required to decline with the sample size, when conducting BGD, we can choose a constant learning rate throughout the whole iterations. However, any decreasing learning learning rate satisfying $\sum_{k=1}^{\infty}\delta_{k}=\infty$ will lead to the same estimator under our setup.

When p is fixed, Theorem 1(i) implies that $\sup_{k\geq k_{1,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k}\right\|=O_{p}\left(1/\sqrt{n}\right)$, and Theorem 1(ii) implies that for k sufficiently large, the BGD estimator is an asymptotically linear estimator, so we have $\sqrt{n}\Delta\boldsymbol{\beta}_{e,k+k_{1,n}^{BGD}}\rightarrow_{d}N\left(0,\Sigma_{1}^{\star}\right)$ by the central limit theorem. The asymptotic variance can be estimated based on Theorem 1(iii). When p is diverging, the inference on $\boldsymbol{\beta}_e^\star$ based on the BGD estimator is given by Theorem 1(iv). Note that for any given vector ρ, we require that $\frac{1}{\sqrt{n}}\rho^{\mathrm{T}}M^{-1}\left(\boldsymbol{\beta}_e^\star\right)\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}$ is asymptotically normally distributed. An alternative approach is to apply the high-dimensional central limit theorem to $\frac{1}{n}\sum_{i=1}^{n}M^{-1}\left(\boldsymbol{\beta}_e^\star\right)\mathbf{X}_{e,i}\varepsilon_i$ (e.g., Chernozhukov et al., 2017).

The number of iterations required to obtain $1/\sqrt{n}$ consistency, $k_{1,n}^{BGD}$, is determined by many factors including the sample size n, the distance between the true parameter and the initial guess $||\Delta\boldsymbol{\beta}_{e,1}||$, as well as the lower bound of the eigenvalues of $M\left(\boldsymbol{\beta}_e\right)$. In general, $k_{1,n}^{BGD}$ is of order $O\left(\log n\right)$, but in practice when we apply the above algorithm, the specific number of iteration is difficult to determine. For detailed discussion of the number of iterations, see Remark 6 at the end of Section 4.

3 Semiparametric BGD Estimation

In the previous section, we focused on iterative estimators based on the BGD algorithm under the parametric setup. It is interesting to see whether this iterative convex estimation approach can handle the case where no parametric form is assumed on the distribution of u. Clearly, when the distribution of u is not known, the algorithm used above is not feasible. We propose instead a modified algorithm that can handle this semiparametric case.

To ensure identification we normalize $\beta_0^\star$ to be 1, so our estimation target is $\boldsymbol{\beta}^\star$. To simplify our notation, we denote the space of $\mathbf{X}$ as $\mathcal{X}$, and the corresponding parameter space of $\boldsymbol{\beta}$ as $\mathcal{B}$. Suppose that an initial guess for $\boldsymbol{\beta}^\star$ is given by $\boldsymbol{\beta}_1$. In the k-th round of iteration, to update $\boldsymbol{\beta}_k$ based on the BGD algorithm, we require the knowledge of G as in Section 2, which is infeasible when G is unknown. A natural idea is that we can construct an estimator for G based on $\boldsymbol{\beta}_k$. More intuitively, suppose for a moment that in the k-th round of iteration, $\boldsymbol{\beta}_k$ is close to the unknown true parameter $\boldsymbol{\beta}^\star$, then we have that $G\left(z\right)=\mathbb{E}\left[y|\,X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^\star=z\right]\approx\mathbb{E}\left[y|\,X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_k=z\right]$ for any $z\in R$. This motivates replacing the unknown $G(\cdot)$ in iteration k with a nonparametric estimator based on an empirical analogue of $\mathbb{E}\left[y|\,X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_k=z\right]$. We consider kernel estimation and the method of sieves to obtain such estimators.

3.1 The Kernel-BGD or KBGD Estimator

In this section we consider using kernel techniques to estimate $G(\cdot)$. The Nadaraya-Watson kernel estimator of $G(\cdot)$ at iteration k is of the form

$$\widehat{G}\left(z\,\middle|\,\boldsymbol{\beta}_k\right) = \frac{\sum_{j=1}^{n} K_{h_n}\left(z - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right) y_j}{\sum_{j=1}^{n} K_{h_n}\left(z - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right)}, z \in R, \tag{5}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is some kernel function, and h_n is some bandwidth parameter depending on n. Given the estimated CDF $\widehat{G}\left(\cdot\,|\,\boldsymbol{\beta}_k\right)$, we can update the parameter as if it were the true CDF $G(\cdot)$. In particular, $\boldsymbol{\beta}_k$ is updated as

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\,\middle|\,\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i. \tag{6}$$

As before, we keep updating $\boldsymbol{\beta}_k$ based on (5) and (6), until some terminating conditions are reached. The resulting estimator is labeled as the *kernel-based batch gradient descent estimator* (KBGD estimator). We summarize our algorithm as follows in algorithm 2.

Algorithm 2: The KBGD Estimator

```
input  : Data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^{n}$, sequence of learning rate $\{\delta_k\}_{k=1}^{\infty}$, initial guess
         $\boldsymbol{\beta}_1$, kernel function $K$, bandwidth $h_n$, and terminating condition $\mathcal{T}$
output: The KBGD estimator $\widehat{\boldsymbol{\beta}}$

$k \leftarrow 1$;
while The terminating condition $\mathcal{T}$ is not satisfied do
    for $i \leftarrow 1$ to $n$ do
        $\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k\right) \leftarrow \frac{\sum_{j=1}^{n} K_{h_n}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right) y_j}{\sum_{j=1}^{n} K_{h_n}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right)}$;
    $\boldsymbol{\beta}_{k+1} \leftarrow \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_{e,i}$;
    $k \leftarrow k + 1$;
$\widehat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}_k$;
```

For any fixed z and $\boldsymbol{\beta}$, under mild conditions there holds $\widehat{G}\left(z\,|\,\boldsymbol{\beta}\right) \rightarrow_p \mathbb{E}\left(y\,|\,X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} = z\right)$. Denote such limit as $L(z, \boldsymbol{\beta})$. Obviously, $L(z, \boldsymbol{\beta}^{\star}) = G(z)$ holds for any $z \in \mathbb{R}$. Before we move to a formal description of the statistical properties of the KBGD estimator, we first provide some further discussion on $L(z, \boldsymbol{\beta})$. For simplicity, in the following we only focus on the case where all the covariates are continuous. We leave further discussion of the case where some covariates are discrete to Remark 5. Note that when there are discrete covariates, our algorithm can be directly applied without any modification, although some further assumptions might be required.

Denote the joint density of $\mathbf{X}_e$ and $\mathbf{X}$ as $f_e(\mathbf{X}_e) = f_e(X_0, \mathbf{X})$ and $f(\mathbf{X}) = \int f_e(X_0, \mathbf{X})\, dX_0$, respectively. Denote $z(\mathbf{X}_e, \boldsymbol{\beta}) = X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$. Also denote $f_{\mathbf{X},z}(\mathbf{X}, z | \boldsymbol{\beta})$ as the joint density of $\mathbf{X}$ and $z(\mathbf{X}_e, \boldsymbol{\beta})$ given $\boldsymbol{\beta}$. Note that for any $\mathbf{x}$ and z,

$$P\left[\mathbf{X} \le \mathbf{x}, z(\mathbf{X}_e, \boldsymbol{\beta}) \le z\right] = \int_{\widetilde{\mathbf{X}} \le \mathbf{x}} \left[\int_{\widetilde{X}_0 \le z - \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}} f_e\left(\widetilde{X}_0, \widetilde{\mathbf{X}}\right) d\widetilde{X}_0\right] d\widetilde{\mathbf{X}},$$

so the joint density of $\mathbf{X}$ and $z(\mathbf{X}_e, \boldsymbol{\beta})$ given $\boldsymbol{\beta}$ is given by $f_{\mathbf{X},z}(\mathbf{X}, z | \boldsymbol{\beta}) = f_e\left(z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X}\right)$, and the marginal density of $z(\mathbf{X}_e, \boldsymbol{\beta})$ is given by

$$f_z(z | \boldsymbol{\beta}) = \int_{\mathcal{X}} f_{\mathbf{X},z}(\mathbf{X}, z | \boldsymbol{\beta})\, d\mathbf{X} = \int_{\mathcal{X}} f_e\left(z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X}\right) d\mathbf{X}.$$

Define $f_{\mathbf{X}|z}(\mathbf{X} | z, \boldsymbol{\beta}) = f_{\mathbf{X},z}(\mathbf{X}, z | \boldsymbol{\beta}) / f_z(z | \boldsymbol{\beta})$ as the conditional density of $\mathbf{X}$ given z and $\boldsymbol{\beta}$, we have that

$$L(z, \boldsymbol{\beta}) = \mathbb{E}\left(G\left(z - \mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right) \middle| z(\mathbf{X}_e, \boldsymbol{\beta}) = z\right) = \int_{\mathcal{X}} G\left(z - \mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right) f_{\mathbf{X}|z}(\mathbf{X} | z, \boldsymbol{\beta})\, d\mathbf{X}, \quad (7)$$

where recall that $\Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}^{\star}$.

Based on the above notations, we formally study the asymptotic properties of the KBGD estimator under increasing dimensions. We first introduce some further assumptions.

Assumption 3. *The kernel function $K(\cdot)$ satisfies: (i) K is bounded and twice continuously differentiable with bounded first and second derivatives, and the second derivative satisfies Lipschitz condition on the whole real line; (ii) $\int K(s)\, ds = 1$; (iii) there exists positive integer υ_K such that $\int s^{\upsilon} K(s)\, du = 0$ for $1 \le \upsilon \le \upsilon_K - 1$; (iv) $K(s) = 0$ for $|s| > 1$.*

Assumption 4. *(i) There exists some constant $\zeta > 1$ such that $\zeta^{-1} \le f_e(\mathbf{X}_e) \le \zeta$ holds for all $\mathbf{X}_e \in \mathcal{X}_e$; (ii) there exists positive integer υ_f such that $f_e(\mathbf{X}_e)$ has bounded up to υ_f-th derivatives.*

Remark 3. *Assumption 4(i) together with Assumption 2(i) is a commonly-used assumption in the machine learning literature (e.g., Wager and Athey, 2018) and allows us to construct a subset of $\mathcal{X}_e$ such that $f_z(z(\mathbf{X}_e, \boldsymbol{\beta}) | \boldsymbol{\beta})$ is uniformly lower bounded from zero.*

The following lemma will be useful in the proof of our theorem.

Lemma 1. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 3, and Assumption 4 hold with $\upsilon_G = 3$, $\upsilon_K = 2$, and $\upsilon_f = 3$. Suppose moreover that K is chosen such that $K \ge 0$. Define $\psi(n, p, h) = h^{-1}\sqrt{\log(pnh^{-1})/n + h^2}$. We have that*

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left\| \frac{1}{n} \sum_{i=1}^{n} \widehat{G}\left(z(\mathbf{X}_{e,i}, \boldsymbol{\beta}) | \boldsymbol{\beta}\right) \mathbf{X}_i - \mathbb{E}\left[L\left(z(\mathbf{X}_{e,i}, \boldsymbol{\beta}), \boldsymbol{\beta}\right) \mathbf{X}_i\right] \right\| = O_p\left(p^{\frac{2p+1}{2(p+1)}} \psi^{\frac{1}{p+1}}(n, p, h_n)\right).$$

Proof of Lemma 1. See Section A of Supplementary Material. □

Lemma 1 implies that $\frac{1}{n}\sum_{i=1}^{n}\widehat{G}\left(Z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\,\boldsymbol{\beta}\right)\mathbf{X}_i$ will be close to $\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]$ uniformly with respect to $\boldsymbol{\beta}$ as n increases if p increases mildly. Note that such uniform convergence results are free of trimming; we do not need to trim $\mathbf{X}_{e,i}$ even when the density of $z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)$ is small. So even when $\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\,\boldsymbol{\beta}\right)$ is a poor estimator for $L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)$ for some $\mathbf{X}_{e,i}$ and $\boldsymbol{\beta}$, our results are still valid. While on the same time, the cost of not conducting any trimming is that our guaranteed convergence rate depends heavily on the dimensionality. To see this, note that to ensure that our estimator is consistent, p must satisfy $p^{\frac{2p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right)\to 0$. Suppose that $p/n\to 0$ and we choose $h_n=\left(\left(\log n\right)/n\right)^{1/6}$, we have that $\psi\left(n,p,h_n\right)\sim\left(\left(\log n\right)/n\right)^{1/3}$. This implies that when p is fixed, the convergence rate in Lemma 1 is $\left(\left(\log n\right)/n\right)^{1/3(p+1)}$. When p increases with n, the dimension p should satisfy $p\log p=O\left(\log n\right)$, implying that p is allowed to increase only mildly with n. Such restriction on p basically comes from the fact that as $\mathbf{X}_{e,i}$ moves towards the boundary of $\mathcal{X}_e$, the density of random variable $z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)$ decreases faster towards zero given a larger p, which makes the convergence rate sensitive to the increase of p.

Given Lemma 1, we are ready to study the statistical properties of the KBGD estimator. For notational simplicity, in the following we denote $z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}_k\right)$ and $z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}^\star\right)$ as $z_{i,k}$ and $z_i^\star$, respectively. We have that under all the conditions as imposed in Lemma 1, there holds

$$\boldsymbol{\beta}_{k+1}=\boldsymbol{\beta}_k-\delta_k\mathbb{E}\left[\left(L\left(z_{i,k},\boldsymbol{\beta}_k\right)-G\left(z_i^\star\right)\right)\cdot\mathbf{X}_i\right]+\delta_k\cdot\left(\text{small order terms}\right). \tag{8}$$

Note that $z_{i,k}=z_i^\star+\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}_k$ and $L\left(z_{i,k},\boldsymbol{\beta}_k\right)=\int_{\mathcal{X}}G\left(z_{i,k}-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}_k\right)f_{\mathbf{X}|z}\left(\mathbf{X}|\,z_{i,k},\boldsymbol{\beta}_k\right)d\mathbf{X}$, so $\left(L\left(z_{i,k},\boldsymbol{\beta}_k\right)-G\left(z_i^\star\right)\right)\cdot\mathbf{X}_i$ equals to

$$\begin{aligned}&\left\{\int_{\mathcal{X}}\left[G\left(z_i^\star+\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}_k-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}_k\right)-G\left(z_i^\star\right)\right]f_{\mathbf{X}|z}\left(\mathbf{X}|\,z_{i,k},\boldsymbol{\beta}_k\right)d\mathbf{X}\right\}\cdot\mathbf{X}_i\\&=\int_0^1\int_{\mathcal{X}}\left[G'\left(z_i^\star+t\left(\mathbf{X}_i-\mathbf{X}\right)^{\mathrm{T}}\Delta\boldsymbol{\beta}_k\right)f_{\mathbf{X}|z}\left(\mathbf{X}|\,z_{i,k},\boldsymbol{\beta}_k\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}-\mathbf{X}_i\mathbf{X}^{\mathrm{T}}\right)\right]\Delta\boldsymbol{\beta}_k d\mathbf{X}dt,\end{aligned} \tag{9}$$

where the integration is understood to be element-wise. (There were some typos so I made some revisions for the following notations.) To further simplify our notation, define

$$W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta},t\right)=G'\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^\star\right)+t\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)f_{\mathbf{X}|z}\left(\widetilde{\mathbf{X}}\middle|\,z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right),$$

$$V\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta},t\right)=\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}-\mathbf{X}\widetilde{\mathbf{X}}^{\mathrm{T}}\right)W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta},t\right),$$

and

$$\mathit{\Lambda}\left(\boldsymbol{\beta},t\right)=\mathbb{E}\left[\int_{\mathcal{X}}V\left(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta},t\right)d\mathbf{X}\right],$$

we have that

$$\mathbb{E}\left[\left(L\left(z_{i,k},\boldsymbol{\beta}_k\right)-G\left(z_i^{\star}\right)\right)\cdot\mathbf{X}_i\right]=\int_0^1 \mathit{\Lambda}\left(\boldsymbol{\beta}_k,t\right)\Delta\boldsymbol{\beta}_k dt,$$

which indicates that

$$\Delta\boldsymbol{\beta}_{k+1}=\left\{\int_0^1\left(I_p-\delta_k\mathit{\Lambda}\left(\boldsymbol{\beta}_k,t\right)\right)dt\right\}\Delta\boldsymbol{\beta}_k+\delta_k\cdot\left(\text{small order terms}\right).$$

To ensure that with probability going to 1 the above iteration shrinks $\|\Delta\boldsymbol{\beta}_k\|$, we make the following assumption.

Assumption 5. *There hold*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B},t\in[0,1]}\overline{\lambda}\left(\mathit{\Lambda}\left(\boldsymbol{\beta},t\right)+\mathit{\Lambda}^{\mathrm{T}}\left(\boldsymbol{\beta},t\right)\right)\leq\overline{\lambda}_{\mathit{\Lambda}}<\infty,$$

and

$$\inf_{\boldsymbol{\beta}\in\mathcal{B},t\in[0,1]}\underline{\lambda}\left(\mathit{\Lambda}\left(\boldsymbol{\beta},t\right)+\mathit{\Lambda}^{\mathrm{T}}\left(\boldsymbol{\beta},t\right)\right)\geq\underline{\lambda}_{\mathit{\Lambda}}>0.$$

Remark 4. *Here we provide an illustrative example where Assumption 5 holds. The technical details for establishing this can be found in Section B of Supplementary Material. Consider the case where u is uniformly distributed over $[-T_1,T_2]$ for both T_1 and T_2 large and $\mathbf{X}_e\sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma^2&0\\0&\Sigma_{\mathbf{X}}\end{pmatrix}\right)$. Then for T_1 and T_2 sufficiently large, Assumption 5 holds. More generally, if the random error u is uniformly distributed over a sufficiently large support, and has sufficiently small and exponentially decreasing probabilities outside the large support, Assumption 5 will also hold.*

Based on the above assumptions, we have the following result.

Theorem 2. *Suppose that all the assumptions in Lemma 1 and Assumption 5 hold with $\upsilon_G=3$, $\upsilon_K=2$, and $\upsilon_f=3$, $\delta_k=\delta<\min\left\{1/\left(2\underline{\lambda}_{\mathit{\Lambda}}\right),1/\left(2p^2\left\|G'\right\|_{\infty}^2\right)\right\}$, and that $\boldsymbol{\beta}_k$ is updated based on algorithm 2. Define*

$$k_{1,n}^{KBGD}=\frac{\log\left(\left\|\Delta\boldsymbol{\beta}_1\right\|\right)-\log\left(p^{\frac{2p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right)\right)}{-\log\left(1-\delta\underline{\lambda}_{\mathit{\Lambda}}/4\right)}.$$

Then we have that

$$\sup_{k\geq k_{1,n}^{KBGD}+1}\left\|\Delta\boldsymbol{\beta}_k\right\|=O_p\left(p^{\frac{2p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right)\right).$$

In particular, if h_n is chosen such that $h_n=\left(\left(\log n\right)/n\right)^{1/6}$, then

$$\sup_{k\geq k_{1,n}^{KBGD}+1}\left\|\Delta\boldsymbol{\beta}_k\right\|=O_p\left(p^{\frac{2p+1}{2(p+1)}}\left(\frac{\log n}{n}\right)^{\frac{1}{3p+3}}\right).$$

Proof of Theorem 2. See See Section B of Supplementary Material. □

Theorem 2 implies that the iterative estimator based on (5) and (6) is consistent under increasing dimensions, no matter whether the starting point is close to the unknown true parameter or not. However, the convergence speed heavily depends on the dimensionality of the problem, p, even when p is fixed. This is not ideal under our single-index setup but is not surprising since our algorithm does not involve any trimming procedure.

We proceed to establish the asymptotic normality of the KBGD estimator. Due to technical difficulties, throughout the following analysis in this section we only consider the case where p is fixed. As we can see in Theorem 2, even in the case of fixed dimensionality, the guaranteed convergence rate of the KBGD estimator based on (5) and (6) is at best $((\log n)/n)^{\frac{1}{3p+3}}$, which still depends on p. To obtain asymptotic normality, we need to slightly modify our algorithm. In particular, we introduce trimming to our algorithm. When updating the parameter, we only use observations that fall into a pre-selected region as did in Ichimura (1993). In particular, the algorithm is modified as,

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n} \mathbf{1}_i^{\phi} \cdot \left(\widehat{G}\left(z_{i,k}\middle|\,\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i, \tag{10}$$

where $\widehat{G}\left(z_{i,k}\middle|\,\boldsymbol{\beta}_k\right) = \widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}_k\right)\middle|\,\boldsymbol{\beta}_k\right)$ is defined in (5), $\mathbf{1}_i^{\phi} = \mathbf{1}\left(\mathbf{X}_{e,i} \in \mathcal{X}_e^{\phi}\right)$, and $\mathcal{X}_e^{\phi}$ is a subset of $\mathcal{X}_e$ given by

$$\mathcal{X}_e^{\phi} = \left\{\mathbf{X}_e \in \mathcal{X}_e : \frac{\phi}{(1+B_0)p} \le X_j \le 1 - \frac{\phi}{(1+B_0)p}, 1 \le j \le p\right\} \tag{11}$$

for some $\phi > 0$ whose choice will be provided later.

Different from (6), the update of $\boldsymbol{\beta}_k$ based on (10) uses only a subset of the whole sample for which the covariate vector $\mathbf{X}_{e,i}$ falls into $\mathcal{X}_e^{\phi}$. The reason why we choose the trimming set as in (11) is that, as we show in Section A of Supplementary Material, for any $0 < \phi < 1$, there holds

$$\inf_{(\mathbf{X}_e,\boldsymbol{\beta}) \in \mathcal{X}_e^{\phi} \times \mathcal{B}} f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\middle|\,\boldsymbol{\beta}\right) \ge \zeta\left(\frac{\phi}{2(1+B_0)p}\right)^p$$

for some constant $C > 0$ that depends on ϕ. When p and ϕ are both fixed, $f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\middle|\,\boldsymbol{\beta}\right)$ is uniformly lower bounded from zero for any combination $(\mathbf{X}_e,\boldsymbol{\beta}) \in \mathcal{X}_e^{\phi} \times \mathcal{B}$, so the uniform estimation accuracy of $L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)$ with respect to $\mathbf{X}_{e,i}$ and $\boldsymbol{\beta}$ will be improved. Note that trimming will cause some efficiency loss by throwing away some observations, but such loss can be controlled to be small if we choose ϕ to be close to zero. We also point that trimming is only applied to the update of the parameter; when nonparametrically estimating G, we still use all the data points.

To simplify our following notation, given the trimming parameter ϕ, we denote $\mathbf{1}^{\phi}\cdot\mathbf{X}$ as $\mathbf{X}^{\phi}$. We also define

$$\Lambda_{\phi}\left(\boldsymbol{\beta},t\right)=\mathbb{E}\left(\mathbf{1}_{i}^{\phi}\cdot\int_{\mathcal{X}}V\left(\mathbf{X}_{e,i},\mathbf{X}_{e},\boldsymbol{\beta},t\right)d\mathbf{X}\right).$$

The following theorem provides a counterpart to the results in Theorem 2.

Theorem 3. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 3, Assumption 4, and Assumption 5 hold with $\upsilon_G=3$, $\upsilon_K=2$, and $\upsilon_f=3$, $\phi<\underline{\lambda}_{\Lambda}/\left(24p^2\left\|G'\right\|_{\infty}\zeta\right)$, $\delta_k=\delta<\min\left\{1/\left(2\underline{\lambda}_{\Lambda}\right),1/\left(2p^2\left\|G'\right\|_{\infty}^2\right)\right\}$, and that $\boldsymbol{\beta}$ is updated under (5) and (10), which is the trimmed version of algorithm 2. Define*

$$\widetilde{k}_{1,n}^{KBGD}=\frac{\log\left(\left\|\Delta\boldsymbol{\beta}_1\right\|\right)-\log\left(\psi\left(n,p,h_n\right)\right)}{-\log\left(1-\delta\underline{\lambda}_{\Lambda}/8\right)},$$

then there holds

$$\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+1}\left\|\Delta\boldsymbol{\beta}_k\right\|=O_p\left(\psi\left(n,p,h_n\right)\right).$$

Proof of Theorem 3. See Section B of Supplementary Material. □

Note that when p is fixed, $\psi\left(n,p,h_n\right)$ no longer depends on p asymptotically. The improvement over the convergence rate basically comes from the improvement of the uniform convergence rate of the kernel estimator due to trimming. Also note that under trimming, the minimum number of iteration in Theorem 3, $\widetilde{k}_{1,n}^{KBGD}$, is of order $\log n$ as long as $nh_n\to\infty$. This implies that under trimming, a faster convergence rate is guaranteed with the minimum number of iterations being of the same magnitude as that of the estimator without trimming.

We now proceed to establish the asymptotic normality of $\boldsymbol{\beta}_k$. Define

$$\boldsymbol{\xi}_n^{\phi}=\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_i^{\star}\middle|\boldsymbol{\beta}^{\star}\right)-y_i\right)\mathbf{X}_i^{\phi}.$$

We note that

$$\begin{aligned}\Delta\boldsymbol{\beta}_{k+1}&=\Delta\boldsymbol{\beta}_k-\frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right)-y_i\right)\mathbf{X}_i^{\phi},\\&=\Delta\boldsymbol{\beta}_k-\frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right)-\widehat{G}\left(z_i^{\star}\middle|\boldsymbol{\beta}^{\star}\right)\right)\mathbf{X}_i^{\phi}-\delta_k\boldsymbol{\xi}_n^{\phi}\\&=\int_0^1\left\{I_p-\frac{\delta_k}{n}\sum_{i=1}^{n}\mathbf{X}_i^{\phi}\left.\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k}\right\}dt\Delta\boldsymbol{\beta}_k-\delta_k\boldsymbol{\xi}_n^{\phi},\end{aligned}\qquad(12)$$

where the integration is understood to be element-wise. Define $\Lambda_{\phi}^{\star}=\Lambda_{\phi}\left(\boldsymbol{\beta}^{\star},0\right)$. To understand the properties of the above algorithm, we need the following lemmas.

Lemma 2. *Suppose that all the and assumptions and assumptions in Theorem 3 hold. For any sequence of subset $\{\mathcal{B}_n\}_{n=1}^{\infty}$ with $\mathcal{B}_n \subseteq \mathcal{B}$, we have that*

$$\sup_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_n, t \in [0,1]} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^{\phi} \left. \frac{\partial \widehat{G}\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right) \middle| \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\widetilde{\boldsymbol{\beta}}} - \Lambda_{\phi}^{\star} \right\|$$
$$= O_p\left(h_n^{-2}\sqrt{\left(\log\left(nh_n^{-1}\right)\right)/n} + h_n^3 + \sup_{\boldsymbol{\beta} \in \mathcal{B}_n} \|\Delta \boldsymbol{\beta}\| \right).$$

Proof of Lemma 2. See Section A of Supplementary Material. □

Lemma 3. *Suppose that all the assumptions and conditions in Theorem 3 hold. If h_n is chosen such that $nh_n^6 \to 0$, we have that $\sqrt{n}\boldsymbol{\xi}_n^{\phi} \to_d N\left(0, \Sigma_{\boldsymbol{\xi}}^{\phi}\right)$, where*

$$\Sigma_{\boldsymbol{\xi}}^{\phi} = \mathbb{E}\left[\left(1 - G\left(z_i^{\star}\right)\right) G\left(z_i^{\star}\right) \left(\mathbf{X}_i^{\phi} - \mathbb{E}(\mathbf{X}^{\phi}|z_i^{\star})\right)\left(\mathbf{X}_i^{\phi} - \mathbb{E}(\mathbf{X}^{\phi}|z_i^{\star})\right)^{\mathrm{T}} \right].$$

Proof of Lemma 3. See Section A of Supplementary Material. □

Now we are in a position to illustrate the results of the asymptotic normality of our KBGD estimator.

Theorem 4. *Suppose that all the assumptions and conditions in Theorem 3 hold. Suppose moreover that h_n is chosen such that $nh_n^6 \to 0$ and $nh_n^4/\left(\log n\right)^2 \to \infty$, and that $\boldsymbol{\beta}$ is updated under (5) and (10). Then*

(i) There holds

$$\sup_{k \geq \widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + 1} \|\Delta \boldsymbol{\beta}_k\| = O_p\left(n^{-1/2}\right),$$

where $k_{2,n}^{KBGD}$ is given by

$$k_{2,n}^{KBGD} = \frac{\log\left(n^{1/2}\right) + \log\left(\psi\left(n, p, h_n\right)\right)}{-\log\left(1 - \delta\lambda_{\Lambda}/16\right)};$$

(ii) Define $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_k$ for any $k - \widetilde{k}_{1,n}^{KBGD} - k_{2,n}^{KGBD} \to \infty$, we have that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\star}\right) \to N\left(0, \Sigma_{\boldsymbol{\beta}}^{\phi}\right),$$

where $\Sigma_{\boldsymbol{\beta}}^{\phi} = \Lambda_{\phi}^{\star -1} \Sigma_{\boldsymbol{\xi}}^{\phi} \left(\Lambda_{\phi}^{\star -1}\right)^{\mathrm{T}}$.

Proof of Theorem 4. See Section B of Supplementary Material. □

We introduce the estimator for the variance matrix, based on which the confidence interval of $\boldsymbol{\beta}^{\star}$ can be then constructed.

Theorem 5. *Suppose that all the assumptions and conditions in Theorem 4 hold. Let $\widehat{\boldsymbol{\beta}}$ be defined as in Theorem 4. Define $\widehat{\Lambda}_\phi = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^\phi \partial \widehat{G}\left(z\left(\mathbf{X}_{e,i}, \widehat{\boldsymbol{\beta}}\right)\middle| \widehat{\boldsymbol{\beta}}\right)/\partial \boldsymbol{\beta}^{\mathrm{T}}$. Moreover, define*

$$\widehat{\Sigma}_{\boldsymbol{\xi}}^\phi = \frac{1}{n}\sum_{i=1}^n \left(\widetilde{G}_i\left(1-\widetilde{G}_i\right)\left(\mathbf{X}_i^\phi - \widehat{\mathbb{E}}\left(\mathbf{X}^\phi\middle|\, \widehat{z}_i\right)\right)\left(\mathbf{X}_i^\phi - \widehat{\mathbb{E}}\left(\mathbf{X}^\phi\middle|\, \widehat{z}_i\right)\right)^{\mathrm{T}}\right),$$

where

$$\widetilde{G}_i = \frac{\sum_{j=1}^n \widetilde{K}_{\widetilde{h}_n}\left(\widehat{z}_i - \widehat{z}_j\right) y_j}{\sum_{j=1}^n \widetilde{K}_{\widetilde{h}_n}\left(\widehat{z}_i - \widehat{z}_j\right)},\ \widehat{\mathbb{E}}\left(\mathbf{X}^\phi\middle|\, \widehat{z}_i\right) = \frac{\sum_{j=1}^n \widetilde{K}_{\widetilde{h}_n}\left(\widehat{z}_i - \widehat{z}_j\right)\mathbf{X}_j^\phi}{\sum_{j=1}^n \widetilde{K}_{\widetilde{h}_n}\left(\widehat{z}_i - \widehat{z}_j\right)}, \widehat{z}_i = X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}},$$

$\widetilde{K} \geq 0$ is any second-order kernel function and $\widetilde{h}_n$ is the bandwidth parameter satisfying $\widetilde{h}_n \to 0$ and $\widetilde{h}_n^2\sqrt{n/\log(n)} \to \infty$. Then we have that

$$\left\|\widehat{\Lambda}_\phi^{-1}\widehat{\Sigma}_{\boldsymbol{\xi}}^\phi\left(\widehat{\Lambda}_\phi^{-1}\right)^{\mathrm{T}} - \Sigma_{\boldsymbol{\beta}}^\phi\right\| \to_p 0.$$

Proof of Theorem 5. See Section B of Supplementary Material. □

We provide some remarks for the KBGD estimators.

Remark 5. *Our previous discussion has be restricted to the case where all the covariates are continuously distributed, while our algorithm can be directly applied to the case where there are discrete covariates without any modifications. In contrast to the average derivative approach (Stoker, 1986; Powell et al., 1989) that uses the differentiation with respect to covariates, the KBGD estimator performs differentiation with respect to the parameters, so it does not impose requirements on the continuity of the covariates. It should be noted that we do require at least one continuous covariate to guarantee identification of the parameters. For simplicity, we recommend normalizing a continuous covariate as X_0. Finally, we point out that stronger assumption should be imposed to make our results valid when there are discrete covariates. In particular, suppose that $\mathbf{X}_e = \left(\mathbf{X}_c^{\mathrm{T}}, \mathbf{X}_d^{\mathrm{T}}\right)^{\mathrm{T}}$, where $\mathbf{X}_c$ is the collection of all the continuous covariates, whereas $\mathbf{X}_d$ is the collection of all the discrete covariates. Also denote the density function of $\mathbf{X}_c$ conditional on $\mathbf{X}_d$ as $f_{\mathbf{X}_c|\mathbf{X}_d}\left(\mathbf{X}_c|\mathbf{X}_d\right)$. Then we require that all the conditions imposed on the $f_e\left(\mathbf{X}_e\right)$ hold for $f_{\mathbf{X}_c|\mathbf{X}_d}\left(\mathbf{X}_c|\mathbf{X}_d\right)$ for any realizations of $\mathbf{X}_d$.*

Remark 6. *We finally provide some remarks on the implementation of our KBGD estimator. The KBGD estimator might be sensitive to the data magnitude. So when implementing such an estimator, we recommend first standardizing the data so that each covariate has zero mean and unit variance. Note that when constructing the KBGD estimator, we normalize the coefficient of $X_{0,i}$ to 1, indicating that the coefficients of $\mathbf{X}_{e,i}$ can not all be zeros. So we* *need to test whether at least one covariate affects the conditional probability of $y_i = 1$. One option is to run a Logit or Probit regression and test whether all the coefficients are equal to zero. Moreover, $X_{0,i}$ must have positive impacts over the conditional probability.*

When applying our algorithm, it is also crucial to choose the tuning parameters including the constant learning rate δ, bandwidth of kernel estimator h_n, and terminating conditions of the algorithm. Theorem 4 implies that the learning rate δ can be chosen as a constant throughout all rounds of iterations but is required to be smaller than $1/\left(2\underline{\lambda}_{\Lambda}\right)$ and $1/\left(4p^2\left\|G'\right\|_{\infty}\right)$, neither of which is known. So we recommend setting the constant δ to be 1 in the first place. If the iteration based on such learning rate does not converge (for example, diverges to infinity or oscillates), we suggest choosing a smaller δ (say, half of the magnitude as the one we chose before) and perform the iteration from the starting point again.

For the choice of the bandwidth h_n, when using a fourth-order kernel function, Theorem 4 requires that h_n is chosen such that $nh_n^6 \to 0$ and $nh_n^4/\left(\log n\right)^2 \to \infty$. As a rule of thumb, we recommend choosing $h_n = C \cdot n^{-1/5}$. For the choice of the constant C, we can choose $C = C_k = std\left(z_{i,k}\right)$ for the k-th round of iteration and $C = std\left(\widehat{z}_i\right)$ when estimating the variance $\Sigma_{\boldsymbol{\beta}}^{\phi}$.

We finally discuss the terminating conditions. As we show in Theorem 4, to obtain $1/\sqrt{n}$-consistency and asymptotic normality, the iteration number is required to be only of order $\log\left(n\right)$. However, such rule can not be directly applied to determine the number of iterations since the initial distance $\left\|\Delta\boldsymbol{\beta}_1\right\|$ as well as the lower bound on the eigenvalues $\underline{\lambda}_{\Lambda}$ are both unknown. We recommend the following terminating condition (NOTE: I multiply the tolerance with δ. One referee points out that if δ is small, then $\max_{1\leq j\leq p}|\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}|$ will be small. So we need to adjust it by multiplying δ.)

$$\max_{1\leq j\leq p}|\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}| < \delta \cdot \varrho$$

for some predetermined tolerance ϱ. During the simulation, we choose $\varrho = 10^{-5}$. Note that in many cases, $\max_{1\leq j\leq p}|\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}|$ may not be monotonically decreasing with k; in some extreme cases, $\max_{1\leq j\leq p}|\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}|$ may even be oscillating and does not shrink to zero. On these conditions, we recommend stopping iteration when the maximum distance achieves its minimum value.

3.2 The SBGD Estimator

In this section, we consider an alternative nonparametric approximation for the unknown CDF based on the method of sieves. Given a set of basis functions $\{r_j\left(z\right)\}_{j=0}^{\infty}$ that is complete in $C(R)$ space, any smooth CDF G can be represented by $G(z) = \sum_{j=0}^{\infty} \pi_j^{\star} r_j(z)$ for any $z \in R$, where $\{\pi_j^{\star}\}_{j=0}^{\infty}$ is the unknown coefficients of the basis functions. In practice, to make the algorithm tractable, we truncate the sequence of basis functions and only use the first $q+1$ ones for approximation, where q increases with sample size n at some rate. To approximate G, it then remains to provide an estimator for the unknown coefficients of the basis functions $\{\pi_j^{\star}\}_{j=0}^{q}$. Our estimation procedure for $\{\pi_j^{\star}\}_{j=0}^{q}$ shares similar intuition as the one that motivates the Nadaraya-Watson kernel estimator in the previous section. In particular, suppose for a moment that in the k-th round of update, we start with $\boldsymbol{\beta}_k$, which is close to the unknown true parameter $\boldsymbol{\beta}^{\star}$. In this case, define $\boldsymbol{r}_q(z) = (r_0(z), \cdots, r_q(z))^{\mathrm{T}}$ and $\boldsymbol{\pi}_q^{\star} = \left(\pi_1^{\star}, \cdots, \pi_q^{\star}\right)^{\mathrm{T}}$, we have that

$$y_i \approx G(z_{i,k}) + \varepsilon_i \approx \boldsymbol{r}_q^{\mathrm{T}}(z_{i,k})\, \boldsymbol{\pi}_q^{\star} + \varepsilon_i,$$

where recall that $z_{i,k} = X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k$. The above relationship motivates the following linear projection type estimator for the sieve coefficients

$$\widehat{\boldsymbol{\pi}}_{q,n,k} = \left(\sum_{i=1}^{n} \boldsymbol{r}_q(z_{i,k})\, \boldsymbol{r}_q^{\mathrm{T}}(z_{i,k})\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{r}_q(z_{i,k})\, y_i\right). \tag{13}$$

Given the sieve coefficient estimator $\widehat{\boldsymbol{\pi}}_{q,n,k}$, the unknown CDF G in the k-th round of update is approximated by

$$\widehat{G}(z|\, \boldsymbol{\beta}_k) = \boldsymbol{r}_q^{\mathrm{T}}(z)\, \widehat{\boldsymbol{\pi}}_{n,q,k}, \ -\infty < z < \infty. \tag{14}$$

Based on the estimated CDF $\widehat{G}(z|\, \boldsymbol{\beta}_k)$, the update of the parameter can be carried out based on (6). We iterate sequentially based on (13), (14) and (6) until some terminating conditions are satisfied. The resulting estimator is then labeled as the *sieve-based batch gradient descent estimator* (SBGD estimator). We summarize our algorithm as follows in algorithm 3.

Remark 7. *In the above SBGD procedure, we update the sieve parameters via a linear projection into the basis vector* $\mathbf{r}_q$*. An alternative procedure can be based on the flexible Logit regression proposed by Hirano, Imbens, and Ridder (2003). The advantage of using flexible Logit regression is that the estimated CDF* $\widehat{G}(z|\, \boldsymbol{\beta}_k)$ *always falls between 0 and 1 for all* z*, which makes the update more stable. A disadvantage of such update is that the flexible Logit regression is based on MLE, which does not allow for an analytical solution. Using numerical optimization to solve for the sieve coefficients in each round of update will increase the computational burdens.*

Remark 8. *Compared with the KBGD algorithm, the SBGD procedure has at least two advantages. On the one side, the sieve-based approximation for the unknown CDF is global*

Algorithm 3: The SBGD Estimator

```
input  : Data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^{n}$, sequence of learning rate $\{\delta_k\}_{k=1}^{\infty}$, initial guess
         $\boldsymbol{\beta}_1$, the order of sieves $q$, sieve functions $\boldsymbol{r}(z) = (r_0(z), \cdots, r_q(z))^{\mathrm{T}}$, and
         terminating condition $\mathcal{T}$
output: The SBGD estimator $\widehat{\boldsymbol{\beta}}$

k ← 1;
while The terminating condition $\mathcal{T}$ is not satisfied do
    $\widehat{\boldsymbol{\pi}}_{q,n,k} \leftarrow$
      $\left(\sum_{i=1}^{n} \boldsymbol{r}_q\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{r}_q\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)y_i\right)$;
    for $i \leftarrow 1$ to $n$ do
        $\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k\right) \leftarrow \boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\widehat{\boldsymbol{\pi}}_{q,n,k}$;
    $\boldsymbol{\beta}_{k+1} \leftarrow \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_{e,i}$;
    $k \leftarrow k + 1$;
$\widehat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}_k$;
```

and guarantees uniform approximation error rate. This allows us to update the parameter without performing any form of trimming as we did for the KBGD estimator. Moreover, we can then develop the asymptotic distribution of the SBGD estimator for the case of increasing dimensionality. On the other hand, the KBGD procedure relies on the kernel estimation of CDF G at n data points, whose computational complexity of each update is of order $O(n^2)$. While the most time-consuming part of the SBGD procedure is the OLS procedure (13), with computational complexity of order $O(nq^2 + q^3)$. When $q/\sqrt{n} \to 0$, the computational burden of SBGD estimator will be substantially lower than that of KBGD estimator.

Define $R_q(z) = G(z) - \boldsymbol{r}_q^{\mathrm{T}}(z)\boldsymbol{\pi}_q^{\star}$, $\Gamma_{q,n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, $\Gamma_{q,n,k} = \Gamma_{q,n}(\boldsymbol{\beta}_k)$, and $\mathfrak{X}_{q,n}(z, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q,n}^{-1}(\boldsymbol{\beta})\boldsymbol{r}_q(z)\mathbf{X}_i$. Through tedious algebra shown in the Supplementary Material to this paper, we can show that the SBGD procedure has the following representation,

$$\begin{aligned}
\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k &- \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,n}(z_{i,k}, \boldsymbol{\beta}_k)\right)\left(G(z_{i,k}) - G(z_i^{\star})\right) \\
&- \frac{\delta_k}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}(z_{i,k})\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_q(z_{j,k})R_q(z_{j,k}) + \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_q(z_{j,k})\varepsilon_j\right) \\
&+ \frac{\delta_k}{n}\sum_{i=1}^{n}\left(R_q(z_{i,k})\mathbf{X}_i + \varepsilon_i\mathbf{X}_i\right), \qquad (15)
\end{aligned}$$

where recall that $z_{i,k} = X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k$. To study the properties of the above procedure, we introduce some additional assumptions.

Assumption 6. *(i) There holds* $\max_{0\le j\le q}\left\|r_j\right\|_\infty \le D_{q,0}$*,* $\max_{0\le j\le q}\left\|r'_j\right\|_\infty \le D_{q,1}$*, and* $\max_{0\le j\le q}\left\|r''_j\right\|_\infty \le D_{q,2}$*; (ii) Define* $\Gamma_q\left(\boldsymbol{\beta}\right) = \mathbb{E}\left(\boldsymbol{r}_q\left(X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{r}_q^{\mathrm{T}}\left(X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\right)$*, there hold* $\inf_{\boldsymbol{\beta}\in\mathcal{B}}\underline{\lambda}\left(\Gamma_q\left(\boldsymbol{\beta}\right)\right) \ge \underline{\lambda}_\Gamma > 0$ *and* $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\lambda}\left(\Gamma_q\left(\boldsymbol{\beta}\right)\right) \le \overline{\lambda}_\Gamma < \infty$ *for all q; (iii) There hold* $\sup_{z\in R}\left|G\left(z\right) - \boldsymbol{r}^{\mathrm{T}}\left(z\right)\boldsymbol{\pi}_q^\star\right| \le \mathcal{E}_{q,0}$ *and* $\sup_{z\in R}\left|G'\left(z\right) - \left(\boldsymbol{r}'\left(z\right)\right)^{\mathrm{T}}\boldsymbol{\pi}_q^\star\right| \le \mathcal{E}_{q,1}$*, where* $\boldsymbol{r}'(z) = \left(r'_0(z), \cdots, r'_q(z)\right)^{\mathrm{T}}$.

For any $-\infty < z < \infty$, define the population counterpart of $\mathfrak{X}_{q,n}\left(z, \boldsymbol{\beta}\right)$ as

$$\mathfrak{X}_q\left(z, \boldsymbol{\beta}\right) = \mathbb{E}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}\right)\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_q\left(z\right)\mathbf{X}\right).$$

To gain some insights into $\mathfrak{X}_q(z, \boldsymbol{\beta})$, we note that $\mathfrak{X}_q(z, \boldsymbol{\beta}) = \boldsymbol{r}_q^{\mathrm{T}}\left(z\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}\right)\mathbb{E}\left(\boldsymbol{r}_q\left(z(\mathbf{X}_e, \boldsymbol{\beta})\mathbf{X}\right)\right)$. Take X_1, the first argument of $\mathbf{X}$, as an example, under mild conditions we have that $\mathbb{E}\left(X_1 \middle| z(\mathbf{X}_e, \boldsymbol{\beta} = z)\right) = \sum_{j=0}^q \pi^\star_{X_1,j} r_q(z) + \mathcal{E}_{X_1,q}$, where $\boldsymbol{\pi}^\star_{X_1,q} = \left(\pi^\star_{X_1,0}, \cdots, \pi^\star_{X_1,q}\right)^{\mathrm{T}}$ is the first $(q+1)$ sieve parameter, and $\mathcal{E}_{X_1,q}$ is the approximation error which shrinks to 0 uniformly with respect to z as q increases. Then $\Gamma_q^{-1}\left(\boldsymbol{\beta}\right)\mathbb{E}\left(\boldsymbol{r}_q\left(z(\mathbf{X}_e, \boldsymbol{\beta})X_j\right)\right)$ equals to $\boldsymbol{\pi}^\star_{X_1,q}$ plus some approximation error that shrinks to zero as q increases. In this case, $\boldsymbol{r}_q^{\mathrm{T}}\left(z\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}\right)\mathbb{E}\left(\boldsymbol{r}_q\left(z(\mathbf{X}_e, \boldsymbol{\beta})X_1\right)\right)$ equals to $\mathbb{E}\left(X_1 \middle| z(\mathbf{X}_e, \boldsymbol{\beta} = z)\right)$ up to some approximation error, and as a result, $\lim_{q\to\infty}\mathfrak{X}_q(z, \boldsymbol{\beta}) = \mathbb{E}\left(\mathbf{X} \middle| z(\mathbf{X}_e, \boldsymbol{\beta} = z)\right)$.

We have the following lemma.

Lemma 4. *Define* $\chi_{1,n} = \sqrt{pq^2 D_{q,0}^4 \log\left(pqD_{q,0}D_{q,1}n\right)/n}$*, and* $\chi_{2,n} = \sqrt{p}qD_{q,0}^2\left(\chi_{1,n} + \mathcal{E}_{q,0}\right)$*. Suppose that Assumption 1, Assumption 2(i)-(iii), and Assumption 6 hold, and moreover,* $\upsilon_G \ge 1$ *and the combination of p, q and υ_G guarantees that $\chi_{1,n} \to 0$ as $n \to \infty$. Then if $\boldsymbol{\beta}_k$ is updated based on (13), (14) and (6), the following holds,*

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \delta_k\mathbb{E}\left[\left(\mathbf{X} - \mathfrak{X}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}_k\right), \boldsymbol{\beta}_k\right)\right)\left(G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}_k\right)\right) - G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right)\right)\right)\right] + \delta_k\mathfrak{R}_{n,k},$$

where $\sup_{k\ge1}\left\|\mathfrak{R}_{n,k}\right\| = O_p\left(\chi_{2,n}\right)$.

Proof of Lemma 4. See Section A of Supplementary Material. □

Lemma 4 provides a parallel result to (8). In particular, define

$$\Psi_q\left(\boldsymbol{\beta}, t\right) = \mathbb{E}\left[G'\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right) + t\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathfrak{X}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right)\mathbf{X}^{\mathrm{T}}\right)\right],$$

under all the conditions imposed in Lemma 4, we have that

$$\Delta\boldsymbol{\beta}_{k+1} = \left\{\int_0^1\left(I_p - \delta_k\Psi_q\left(\boldsymbol{\beta}_k, t\right)\right)dt\right\}\Delta\boldsymbol{\beta}_k + \delta_k\mathfrak{R}_{n,k}. \tag{16}$$

Obviously, (16) is also a parallel result to (9). As a result, to ensure that (16) constitutes a contraction for $\left\|\Delta\boldsymbol{\beta}_k\right\|$, we impose the following assumption.

Assumption 7. *For any $q \geq 0$, there hold*

$$\inf_{0 \leq t \leq 1, \boldsymbol{\beta} \in \mathcal{B}} \underline{\lambda} \left(\Psi_q \left(\boldsymbol{\beta}, t \right) + \Psi_q^{\mathrm{T}} \left(\boldsymbol{\beta}, t \right) \right) \geq \underline{\lambda}_{\Psi} > 0,$$

$$\sup_{0 \leq t \leq 1, \boldsymbol{\beta} \in \mathcal{B}} \underline{\lambda} \left(\Psi_q \left(\boldsymbol{\beta}, t \right) + \Psi_q^{\mathrm{T}} \left(\boldsymbol{\beta}, t \right) \right) \leq \overline{\lambda}_{\Psi} < \infty.$$

Remark 9. *Assumption 7 is similar to Assumption 5. Indeed, since we already pointed out that $\lim_{q \to \infty} \mathfrak{X}_q(z, \boldsymbol{\beta}) = \mathbb{E} \left(\mathbf{X} | \, z(\mathbf{X}_e, \boldsymbol{\beta} = z) \right)$, then*

$$\lim_{q \to \infty} \Psi_q \left(\boldsymbol{\beta}, t \right) = \mathbb{E} \left[G' \left(z \left(\mathbf{X}_e, \boldsymbol{\beta}^{\star} \right) + t \mathbf{X}^{\mathrm{T}} \Delta \boldsymbol{\beta} \right) \left(\mathbf{X} \mathbf{X}^{\mathrm{T}} - \mathbb{E} \left(\mathbf{X} | \, z(\mathbf{X}_e, \boldsymbol{\beta}) = z \right) \mathbf{X}^{\mathrm{T}} \right) \right].$$

We can then verify that under the data generating process in Remark 4, Assumption 7 will also hold asymptotically.

Based on the above assumptions, we have the following result.

Theorem 6. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 6 and Assumption 7 hold, and moreover, $\upsilon_G \geq 1$ and the combination of p, q and υ_G guarantees that $\chi_{1,n} \to 0$ as $n \to \infty$. Suppose moreover that the learning rate is chosen such that $\delta_k = \delta$ with $0 < \delta < \min \left\{ 1/\left(2 \underline{\lambda}_{\Psi} \right), \underline{\lambda}_{\Psi} / (2 \left\| G' \right\|_{\infty}^2 p^2 \{ 1 + \underline{\lambda}_{\Gamma}^{-1} (q+1) D_{q,0}^2 \}^2) \right\}$, and that $\boldsymbol{\beta}$ is updated based on algorithm 3. Define*

$$k_{1,n}^{SBGD} = \frac{\log \left(\left\| \Delta \boldsymbol{\beta}_1 \right\| \right) - \log \left(\chi_{2,n} \right)}{- \log \left(1 - \underline{\lambda}_{\Psi} \delta / 4 \right)},$$

then we have that

$$\sup_{k \geq k_{1,n}^{SBGD} + 1} \left\| \Delta \boldsymbol{\beta}_k \right\| = O_p \left(\chi_{2,n} \right).$$

Proof of Theorem 6. See Section B of Supplementary Material. □

Based on the consistency results in Theorem 6, we are ready to establish the asymptotic normality of our SBGD estimator. Different from the KBGD estimator studied in the previous section, now we allow the dimensionality p to diverge with n at some rate.

Apply the mean value theorem to (15), we have that

$$\begin{aligned} \Delta \boldsymbol{\beta}_{k+1} = & \left\{ I_p - \delta_k \int_0^1 \frac{1}{n} \sum_{i=1}^n G' \left(z_i^{\star} + t \mathbf{X}_i^{\mathrm{T}} \Delta \boldsymbol{\beta}_k \right) \left(\mathbf{X}_i \mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n} \left(z_{i,k}, \boldsymbol{\beta}_k \right) \mathbf{X}_i^{\mathrm{T}} \right) dt \right\} \Delta \boldsymbol{\beta}_k \\ & - \frac{\delta_k}{n} \sum_{i=1}^n \mathbf{X}_i \boldsymbol{r}_q^{\mathrm{T}} \left(z_{i,k} \right) \Gamma_{q,n,k}^{-1} \left(\frac{1}{n} \sum_{j=1}^n \boldsymbol{r}_q \left(z_{j,k} \right) R_q \left(z_{j,k} \right) + \frac{1}{n} \sum_{i=1}^n \boldsymbol{r}_q \left(z_{j,k} \right) \varepsilon_j \right) \\ & + \frac{\delta_k}{n} \sum_{i=1}^n \left(R_q \left(z_{i,k} \right) \mathbf{X}_i + \varepsilon_i \mathbf{X}_i \right). \end{aligned}$$

Define $\Psi_q^\star = \mathbb{E}\left[G'\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right)\right)\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathfrak{X}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right), \boldsymbol{\beta}^\star\right)\mathbf{X}^{\mathrm{T}}\right)\right]$ and $\mathfrak{V}_q = \mathbb{E}\left(\mathbf{X}_i \boldsymbol{r}_q^{\mathrm{T}}\left(z_i^\star\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^\star\right)\right)$. Similar to Lemma 2 and Lemma 3, we provide two additional lemmas that are useful to understand the above dynamics.

Lemma 5. *Suppose that Assumption 1, Assumption 2(i)-(iii), and Assumption 6 hold, $\upsilon_G \geq 2$ and the combination of p, q and υ_G guarantees that $\chi_{1,n} \to 0$ as $n \to \infty$. Then for any sequence $\{\mathcal{B}_n\}_{n=1}^\infty$ with $\mathcal{B}_n \subseteq \mathcal{B}$ we have that*

$$\sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^n G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right) - \Psi_q^\star\right\|$$
$$= O_p\left(pqD_{q,0}^2\chi_{1,n} + \sqrt{p^3}q^2D_{q,0}^3D_{q,1}\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\|\Delta\boldsymbol{\beta}\|\right).$$

Proof of Lemma 5. See Section A of Supplementary Material. □

Lemma 6. *Suppose that Assumption 1, Assumption 2(i)-(iii), and Assumption 6 hold, $\upsilon_G \geq 2$ and the combination of p, q and υ_G guarantees that $\chi_{1,n} \to 0$ as $n \to \infty$. Define $\boldsymbol{r}_{q,i,k} = \boldsymbol{r}_q\left(z_{i,k}\right)$, and $R_{q,i,k} = R_q\left(z_{i,k}\right)$. Also define $\chi_{3,n} = \sqrt{p^2qD_{q,1}^2\log\left(pqD_{q,2}n\right)/n}$, then we have that*

$$\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^n\boldsymbol{r}_{q,j,k}R_{q,j,k} + \frac{1}{n}\sum_{j=1}^n\boldsymbol{r}_{q,j,k}\varepsilon_j\right) + \right.$$
$$\left.\frac{1}{n}\sum_{i=1}^n R_q\left(z_{i,k}\right)\mathbf{X}_i - \frac{1}{n}\sum_{i=1}^n\mathfrak{X}_q\left(z_i^\star, \boldsymbol{\beta}^\star\right)\varepsilon_j\right\| = O_p\left(\chi_{4,n}\right),$$

where $\chi_{4,n} = \sqrt{p}qD_{q,0}^2\mathcal{E}_{q,0} + \sqrt{pq}D_{q,0}\chi_{2,n}\chi_{3,n} + \chi_{2,n}\sqrt{p^2q^4D_{q,0}^6D_{q,1}^2/n}$.

Proof of Lemma 6. See Section A of Supplementary Material. □

Based on the above two lemmas, we are now ready to study the asymptotic distribution of the SBGD estimator.

Theorem 7. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 6 and Assumption 7 hold, $\upsilon_G \geq 2$, the combination of p, q and υ_G guarantees that $\chi_{1,n} \to 0$ as $n \to \infty$, and that $\boldsymbol{\beta}$ is updated based on algorithm 3. We have that*

(i) There holds

$$\Delta\boldsymbol{\beta}_{k+1} = \left(I_p - \delta\Psi_q^\star\right)\Delta\boldsymbol{\beta}_k + \frac{\delta}{n}\sum_{i=1}^n\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^\star, \boldsymbol{\beta}^\star\right)\right)\varepsilon_i + \widetilde{\mathfrak{R}}_{n,k},$$

where $\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\widetilde{\mathfrak{R}}_{n,k}\right\| = O_p\left(\chi_{5,n}\right)$ with $\chi_{5,n} = pqD_{q,0}^2\chi_{1,n}\chi_{2,n} + \sqrt{p^3}q^2D_{q,0}^3D_{q,1}\chi_{2,n}^2 + \chi_{4,n}$;

(ii) Define $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_{k+k_{1,n}^{SBGD}+k_{2,n}^{SBGD}+1}$ *with*

$$k_{2,n}^{SBGD} = \frac{-\log \chi_{2,n} + \log \sqrt{n}}{-\log\left(1 - \underline{\lambda}_{\Psi}\delta/4\right)},$$

and any $k \geq 1$. *If the combination of* p, q *and* υ_G *further guarantees that* $\sqrt{n}\chi_{5,n} \to 0$ *as* $n \to \infty$, *we have that*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\star}\right) = \Psi_q^{\star-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^{\star}, \boldsymbol{\beta}^{\star}\right)\right)\varepsilon_i + o_p\left(n^{-\frac{1}{2}}\right).$$

Then for any $p \times 1$ *vector* ρ *such that* $\|\rho\| < \infty$ *and* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho^{\mathrm{T}}\Psi_q^{\star-1}\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^{\star}, \boldsymbol{\beta}^{\star}\right)\right)\varepsilon_i \to_d N\left(0, \sigma_S^2(\rho)\right)$ *with*

$$\sigma_S^2(\rho) = \lim_{n\to\infty}\rho^{\mathrm{T}}\Psi_q^{\star-1}\mathbb{E}\left\{G\left(z_i^{\star}\right)\left(1 - G\left(z_i^{\star}\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^{\star}, \boldsymbol{\beta}^{\star}\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^{\star}, \boldsymbol{\beta}^{\star}\right)\right)^{\mathrm{T}}\right\}\left(\Psi_q^{\star-1}\right)^{\mathrm{T}}\rho,$$

there holds

$$\sqrt{n}\rho^{\mathrm{T}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\star}\right) \to_d N\left(0, \sigma_S^2(\rho)\right).$$

Proof of Theorem 7. See Section B of Supplementary Material. □

Similar to Fan et al. (2020), Theorem 7 demonstrates the asymptotic normality of the SBGD estimator under high dimensionality. Based on such asymptotic distribution, we finally provide the estimator for the asymptotic variance so that inference on the unknown parameter can be conducted.

Theorem 8. *Suppose that all the conditions listed in Theorem 7 hold and* $pq^2D_{q,0}^4\mathcal{E}_{q,1} \to 0$ *as* $n \to 0$. *Let* $\widehat{\boldsymbol{\beta}}$ *be as defined as in Theorem 7. Define* $\widehat{\boldsymbol{r}}_{q,i} = \boldsymbol{r}_q\left(z\left(\mathbf{X}_{e,i}, \widehat{\boldsymbol{\beta}}\right)\right)$, $\widehat{\boldsymbol{r}}'_{q,i} = \boldsymbol{r}'_q\left(z\left(\mathbf{X}_{e,i}, \widehat{\boldsymbol{\beta}}\right)\right)$, $\widehat{\boldsymbol{\pi}}_q = \left(\sum_{i=1}^{n}\widehat{\boldsymbol{r}}_{q,i}\widehat{\boldsymbol{r}}_{q,i}^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\widehat{\boldsymbol{r}}_{q,i}y_i\right)$, $\widehat{G}_i = \widehat{\boldsymbol{r}}_{q,i}^{\mathrm{T}}\widehat{\boldsymbol{\pi}}$, $\widehat{G}'_i = \widehat{\boldsymbol{r}}_{q,i}^{\prime\mathrm{T}}\widehat{\boldsymbol{\pi}}_q$, $\widehat{\Psi}_{q,i}^{\star} = \frac{1}{n}\sum_{i=1}^{n}\widehat{G}'_i \cdot \left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(\widehat{z}_i, \widehat{\boldsymbol{\beta}}\right)\mathbf{X}_i^{\mathrm{T}}\right)$, $\widehat{\mathfrak{X}}_{q,i} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j\widehat{\boldsymbol{r}}_{q,j}^{\mathrm{T}}\Gamma_{q,n}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{r}}_{q,i}$, *and*

$$\widehat{\sigma}_S^2(\rho) = \rho^{\mathrm{T}}\widehat{\Psi}_q^{\star-1}\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{G}_i\left(1 - \widehat{G}_i\right)\left(\mathbf{X}_i - \widehat{\mathfrak{X}}_{q,i}\right)\left(\mathbf{X}_i - \widehat{\mathfrak{X}}_{q,i}\right)^{\mathrm{T}}\right\}\left(\widehat{\Psi}_q^{\star-1}\right)^{\mathrm{T}}\rho,$$

If $\sqrt{p^6q^{10}D_{q,0}^{18}D_{q,1}^2/n} + p^2q^4D_{q,0}^7D_{q,1}\mathcal{E}_{q,0} + pq^3D_{q,0}^6\mathcal{E}_{q,1} \to 0$, *then for any* $p \times 1$ *vector* ρ *such that* $\|\rho\| < \infty$, *there holds*

$$\left|\widehat{\sigma}_S^2(\rho) - \sigma_S^2(\rho)\right| \to_p 0.$$

Proof of Theorem 8. See Section B of Supplementary Material. □

We conclude this section with some guidance of implementation of the SBGD estimator.

Remark 10. *For SBGD estimators, the choice of the constant learning rate δ and the stopping rule are the same as those of KBGD estimators, as was discussed in Remark 6. For the choice of sieve functions, we can use polynomial series for the case where the error term u_i has bounded support and Hermite polynomials for the case where u_i has unbounded support. Note that when using polynomial series $\{1, z, z^2, \cdots, z^q\}$, the correlation between the sieve functions increases as the approximation order q increases, which may lead to a violation of Assumption 6(ii). To improve the finite sample performance of our method, we recommend using Chebyshev or Legendre polynomials. Moreover, in the case where u_i has unbounded support, following Bierens (2014), we recommend first conducting the following transformation $G(z) = \widetilde{G}(T(z))$, where $T : R \mapsto [-1, 1]$ is a differentiable function, and then using standard Chebyshev or Legendre polynomials to approximate $\widetilde{G}$. For example, in our following simulations and empirical applications, we use $T(z) = 2\pi^{-1} \arctan(z)$ and Legendre polynomials. For the uniform error bound of truncated Legendre polynomials, see Wang and Xiang (2012). Finally, for the choice of the order of the sieve functions, we recommend choosing q that minimizes some pre-specified loss function, say, $\sum_{i=1}^{n}(y_i - \widehat{G}_i)^2$, where $\widehat{G}_i$ is defined in Theorem 8.*

4 Monte Carlo Experiments

This section conducts Monte Carlo simulations to study the performance of our KBGD and SBGD estimators. We focus on two aspects of our estimators. First, we study the finite-sample properties of the KBGD estimator, including the bias and the root mean squared error (RMSE). Let the j-th argument of the true parameter be $\beta_j^\star$, and the simulation is repeated R times, where its estimator in the r-th round of simulation is $\widehat{\beta}_j^r$, then the bias and RMSE are respectively given by Bias $= |\frac{1}{R}\sum_{r=1}^{R}(\widehat{\beta}_j^r - \beta_j^\star)|$ and RMSE $= \sqrt{\sum_{r=1}^{R}(\widehat{\beta}_j^r - \beta_j^\star)^2/R}$. We also investigate whether the confidence interval based on the asymptotic distribution has good coverage rate. We consider nominal coverage rate $\alpha = 0.95$, so the confidence interval for $\beta_j^\star$ in the r-th round of repetition is given by $CI_j^r = [\widehat{\beta}_j^r - 1.96 \cdot \widehat{\text{std}}_j^r, \widehat{\beta}_j^r + 1.96 \cdot \widehat{\text{std}}_j^r]$, where $\widehat{\text{std}}_j^r$ is the estimated standard deviation of $\widehat{\beta}_j^r$. The actual coverage rate is then given by $CR = \frac{1}{R}\sum_{r=1}^{R}\mathbf{1}(\beta_j^\star \in CI_j^r)$.

Second, we are also interested in how sensitive our estimators are to the initial guess of the true parameter. In each repetition of our simulation, we consider three different initial guesses: the true parameter vector, the parameter vector estimated based on the Logit regression, and the parameter with all elements being zeros. If the estimation results starting from different initial guesses are close or even identical to each other, the estimation methods are insensitive to the initial guesses and thus are robust in terms of computation. Denote $\widehat{\boldsymbol{\beta}}^r_{\mathrm{T}}$, $\widehat{\boldsymbol{\beta}}^r_{\mathrm{L}}$, and $\widehat{\boldsymbol{\beta}}^r_{\mathrm{Z}}$ as the estimators with starting points being true parameter, Logit estimator, and vector of zeros (origin point). We use $S_{\mathrm{L}} = \sqrt{\frac{1}{R}\sum_{i=1}^{n}||\widehat{\boldsymbol{\beta}}^r_{\mathrm{L}} - \widehat{\boldsymbol{\beta}}^r_{\mathrm{T}}||^2/||\boldsymbol{\beta}^{\star}||^2}$ and $S_{\mathrm{Z}} = \sqrt{\frac{1}{R}\sum_{i=1}^{n}||\widehat{\boldsymbol{\beta}}^r_{\mathrm{Z}} - \widehat{\boldsymbol{\beta}}^r_{\mathrm{T}}||^2/||\boldsymbol{\beta}^{\star}||^2}$ as the measurement of the sensitivity, where $\beta^{\star}$ is the true parameter vector. To compare the performance of our method with the existing estimators, we also consider Ichimura's semiparametric least squares (SLS) estimator (Ichimura, 1993), Klein and Spady's semiparametric maximum likelihood (SMLE) estimator (Klein and Spady, 1993), and maximum rank correlation (MRC) estimator (Han, 1987; Fan et al., 2020).

Table 1: Finite Sample Performance: $u \sim \mathit{Cauchy}$

			β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
$n = 2500$	Bias	KBGD	0.0134	0.0123	0.0108	0.0109	0.0225	0.0323	0.0053	0.0120	0.0222	0.0388
		SBGD	0.0146	0.0128	0.0115	0.0113	0.0246	0.0327	0.00059	0.0121	0.0243	0.0396
	RMSE	KBGD	0.2493	0.1464	0.1450	0.1866	0.3076	0.5529	0.1146	0.1611	0.2642	0.5068
		SBGD	0.2508	0.1433	0.1459	0.1852	0.3123	0.5409	0.1162	0.1575	0.2618	0.4895
	CR	KBGD	0.9550	0.9610	0.9500	0.9510	0.9490	0.9540	0.9640	0.9580	0.9660	0.9630
		SBGD	0.9520	0.9480	0.9340	0.9380	0.9280	0.9460	0.9430	0.9440	0.9460	0.9550
$n = 5000$	Bias	KBGD	0.0087	0.0025	0.0052	0.0071	0.0130	0.0063	0.0021	0.0080	0.0130	0.0141
		SBGD	0.0076	0.0015	0.0046	0.0062	0.0108	0.0023	0.0014	0.0068	0.0109	0.0103
	RMSE	KBGD	0.1692	0.0960	0.0986	0.1263	0.2019	0.3626	0.0863	0.1034	0.1756	0.3307
		SBGD	0.1679	0.0942	0.0978	0.1248	0.1984	0.3530	0.0857	0.1023	0.1711	0.3211
	CR	KBGD	0.9600	0.9590	0.9530	0.9530	0.9460	0.9500	0.9400	0.9720	0.9640	0.9480
		SBGD	0.9530	0.9540	0.9430	0.9570	0.9540	0.9510	0.9380	0.9550	0.9510	0.9450

Table 2: Finite Sample Performance: $u \sim t(4)$

			β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
$n = 2500$	Bias	KBGD	0.0024	0.0032	0.0042	0.0014	0.0042	0.0163	0.0021	0.0033	0.0043	0.0135
		SBGD	0.0017	0.0040	0.0041	0.0024	0.0056	0.0195	0.0019	0.0036	0.0054	0.0161
	RMSE	KBGD	0.1509	0.0861	0.0831	0.1112	0.1708	0.3083	0.0748	0.0936	0.1539	0.2874
		SBGD	0.1519	0.0868	0.0844	0.1123	0.1718	0.3105	0.0760	0.0949	0.1551	0.2907
	CR	KBGD	0.9540	0.9540	0.9500	0.9480	0.9440	0.9520	0.9440	0.9450	0.9490	0.9370
		SBGD	0.9430	0.9460	0.9440	0.9390	0.9510	0.9430	0.9350	0.9460	0.9490	0.9350
$n = 5000$	Bias	KBGD	0.0021	0.0002	0.0012	0.0020	0.0001	0.0019	0.0040	0.0008	0.0019	0.0015
		SBGD	0.0020	0.0003	0.0013	0.0023	0.0009	0.0000	0.0042	0.0011	0.0027	0.0032
	RMSE	KBGD	0.1077	0.0606	0.0560	0.0777	0.1264	0.2157	0.0509	0.0658	0.1077	0.2052
		SBGD	0.1081	0.0609	0.0560	0.0779	0.1270	0.2180	0.0511	0.0663	0.1084	0.2070
	CR	KBGD	0.9440	0.9490	0.9610	0.9350	0.9390	0.9550	0.9540	0.9540	0.9500	0.9420
		SBGD	0.9420	0.9530	0.9580	0.9250	0.9390	0.9560	0.9520	0.9510	0.9480	0.9430

Table 3: Finite Sample Performance: $u \sim \chi^2(3) - 3$

			β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
$n = 2500$	Bias	KBGD	0.0104	0.0116	0.0075	0.0202	0.0220	0.0570	0.0034	0.0124	0.0246	0.0548
		SBGD	0.0114	0.0125	0.0073	0.0207	0.0226	0.0599	0.0038	0.0127	0.0253	0.0574
	RMSE	KBGD	0.2216	0.1297	0.1258	0.1660	0.2574	0.4578	0.1042	0.1328	0.2223	0.4228
		SBGD	0.2220	0.1315	0.1261	0.1663	0.2594	0.4637	0.1042	0.1346	0.2250	0.4274
	CR	KBGD	0.9650	0.9440	0.9520	0.9510	0.9500	0.9520	0.9580	0.9510	0.9600	0.9550
		SBGD	0.9570	0.9450	0.9370	0.9500	0.9430	0.9520	0.9360	0.9460	0.9540	0.9540
$n = 5000$	Bias	KBGD	0.0045	0.0059	0.0032	0.0046	0.0102	0.0151	0.0009	0.0071	0.0147	0.0249
		SBGD	0.0049	0.0062	0.0028	0.0047	0.0101	0.0155	0.0008	0.0071	0.0149	0.0256
	RMSE	KBGD	0.1531	0.0889	0.0868	0.1129	0.1769	0.3109	0.0732	0.0916	0.1479	0.2904
		SBGD	0.1531	0.0887	0.0868	0.1131	0.1765	0.3104	0.0734	0.0915	0.1480	0.2901
	CR	KBGD	0.9650	0.9510	0.9410	0.9470	0.9490	0.9450	0.9540	0.9650	0.9560	0.9450
		SBGD	0.9660	0.9510	0.9420	0.9460	0.9510	0.9480	0.9500	0.9680	0.9590	0.9430

Table 4: Finite Sample Performance: $u \sim N(0,1)$

			β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
$n = 2500$	Bias	KBGD	0.0034	0.0046	0.0024	0.0046	0.0150	0.0220	0.0013	0.0065	0.0106	0.0192
		SBGD	0.0048	0.0050	0.0022	0.0048	0.0153	0.0222	0.0013	0.0064	0.0111	0.0201
	RMSE	KBGD	0.1345	0.0783	0.0722	0.0941	0.1576	0.2870	0.0626	0.0844	0.1407	0.2639
		SBGD	0.1366	0.0794	0.0730	0.0947	0.1598	0.2903	0.0634	0.0852	0.1427	0.2680
	CR	KBGD	0.9500	0.9490	0.9540	0.9520	0.9510	0.9530	0.9590	0.9490	0.9390	0.9520
		SBGD	0.9490	0.9350	0.9300	0.9460	0.9370	0.9420	0.9450	0.9410	0.9430	0.9380
$n = 5000$	Bias	KBGD	0.0030	0.0033	0.0006	0.0033	0.0053	0.0121	0.0033	0.0003	0.0029	0.0051
		SBGD	0.0035	0.0035	0.0006	0.0036	0.0055	0.0129	0.0036	0.0005	0.0033	0.0062
	RMSE	KBGD	0.0964	0.0538	0.0527	0.0706	0.1083	0.1912	0.0452	0.0584	0.0949	0.1727
		SBGD	0.0967	0.0543	0.0534	0.0706	0.1088	0.1928	0.0456	0.0587	0.0954	0.1740
	CR	KBGD	0.9390	0.9490	0.9500	0.9470	0.9430	0.9450	0.9450	0.9390	0.9460	0.9570
		SBGD	0.9390	0.9450	0.9420	0.9420	0.9360	0.9480	0.9500	0.9440	0.9450	0.9590

When studying finite-sample performance, we consider data generating process

$$y_i = \mathbf{1}(X_{0,i} + \beta_1^\star X_{1,i} \cdots + \beta_{10}^\star X_{10,i} - u_i > 0), i = 1, 2, \cdots, n,$$

where data are i.i.d over i, and $X_{0,i}, X_{1,i}, \cdots, X_{10,i}, u_i$ are also independent. We set $\boldsymbol{\beta}^\star = (1, 1, 0.5, 1, 2, 4, -0.5, -1, -2, -4)^{\mathrm{T}}$, $X_{0,i} \sim N(0,1)$, $X_{1,i} \sim \text{Bernoulli}\,(1/2)$, $X_{2,i} \sim \text{Poisson}\,(2)$, $X_{j,i} \sim (\chi^2(1) - 1)/\sqrt{2}$ for $3 \leq j \leq 10$. We consider 4 setups for the random error u_i: $u_i \sim \textit{Cauchy}$, $u_i \sim t(4)$, $u_i \sim \chi^2(3) - 3$ and $u_i \sim N(0,1)$. We consider two sample sizes $n = 2500$ and 5000. Finally, we repeat the simulation 1000 times.

Regarding the implementation of KBGD and SBGD algorithms, the learning rate is chosen as $\delta = 1$, the starting point is the Logit estimator, and the stopping rule is either

$\max_{1\leq j\leq p}|\widehat{\beta}_{j,k+1}-\widehat{\beta}_{j,k}| < 10^{-5}$ or $k \geq 20000$. for KBGD estimator, we use fourth-order Epanechnikov kernel to construct the Nadaraya-Watson estimator. In each iteration, the bandwidth h_n is chosen as $h_n = \sigma_{\widehat{z}} \cdot n^{-1/5}$, where n is sample size, $\sigma_{\widehat{z}}$ is the standard deviation of $z_{i,k}$, and $z_{i,k} = X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k$. For the SBGD estimator, we use transformed Legendre polynomials according to the discussion in Remark 10. For the choice of the order of the sieve functions q, we also use the procedure proposed in Remark 10, where the smallest and the largest order are 9 and 25, respectively.

Table 1 through Table 4 report the finite-sample properties of our estimators. It can be seen that our estimators perform well in finite sample cases. Both estimators have small bias which is close to zero, and the RMSE decrease with the increase of sample size at roughly $1/\sqrt{n}$ rate. Moreover, the confidence interval constructed based on the asymptotic variance and normal approximation has actual coverage rate that is quite close to the nominal rate 0.95.

Table 5: Sensitivity to Initial Points: $u \sim Cauchy$

		Sensitivity		Running Time (Seconds)		
	Method	S_L	S_Z	True	Logit	Zeros
$n = 2500$	KBGD	0.0024	0.0023	161.19	122.15	208.63
	SBGD	0.0028	0.0023	1.2784	0.8431	1.2784
	SLS	0.1516	505.26	70.132	64.268	106.62
	SMLE	0.1555	960.61	524.01	522.01	284.76
	MRC	0.1213	0.5207	13.085	11.710	24.112
$n = 5000$	KBGD	0.0026	0.0025	529.91	499.48	788.40
	SBGD	0.0025	0.0024	1.7869	1.5380	2.4520
	SLS	0.1033	1157.6	248.89	240.39	366.79
	SMLE	0.1073	2806.7	2285.3	1345.0	945.60
	MRC	0.0757	0.4289	46.215	44.501	116.79

We study the sensitivity of different methods to the initial points. The setups of the data generating process and the implementation of KBGD and SBGD algorithms are the same as before[7], expect that now we consider different starting points. When implementing the SLS and SMLE estimators, the bandwidth is chosen as $h_n = n^{-1/5}$ and $h_n = n^{-1/6.02}$, respectively. The optimization is conducted using MATLAB code `fminsearch`, where the optimization setup is given by `optimset(’MaxFunEvals’, 1e5, ’MaxIter’, 1e5, ’display’, ’off’, ’TolFun’, 1e-6, ’TolX’, 1e-6, ’Display’, ’off’)`. Table 5 through Table 8 report the simulation results.

[7] We also fix the order of sieve functions at $q = 11$ for comparisons.

Table 6: Sensitivity to Initial Points: $u \sim t(4)$

		Sensitivity		Running Time (Seconds)		
	Method	S_L	S_Z	True	Logit	Zeros
$n = 2500$	KBGD	0.0026	0.0021	109.25	33.136	184.70
	SBGD	0.0023	0.0022	0.6700	0.5013	1.0117
	SLS	0.0879	399.90	63.759	60.848	95.000
	SMLE	0.0901	559.56	95.108	100.18	178.05
	MRC	0.0689	0.5052	9.2122	7.8178	21.141
$n = 5000$	KBGD	0.0022	0.0025	405.38	67.322	689.25
	SBGD	0.0021	0.0024	1.3466	0.8639	2.1190
	SLS	0.0637	1125.2	244.08	247.96	360.41
	SMLE	0.0631	1857.7	240.17	231.56	725.61
	MRC	0.0508	0.4157	44.072	38.570	119.16

Table 7: Sensitivity to Initial Points: $u \sim \chi^2(3) - 3$

		Sensitivity		Running Time (Seconds)		
	Method	S_L	S_Z	True	Logit	Zeros
$n = 2500$	KBGD	0.0030	0.0024	143.91	73.499	201.50
	SBGD	0.0025	0.0023	0.8701	0.7267	1.1815
	SLS	0.1305	239.37	58.384	62.397	106.08
	SMLE	0.1303	647.64	62.495	64.875	208.23
	MRC	0.1080	0.5103	9.8474	8.3594	21.378
$n = 5000$	KBGD	0.0032	0.0024	445.74	240.86	756.32
	SBGD	0.0027	0.0024	1.4611	1.1934	2.2780
	SLS	0.0953	1910.6	247.92	243.61	361.14
	SMLE	0.0931	2397.5	245.75	238.29	918.13
	MRC	0.0767	0.4130	46.638	40.899	117.53

We can see that for both KBGD and SBGD estimators, S_{L} and S_{Z} are both close to zero, indicating that the resulting estimators starting from Logit estimator or zeros are almost identical to the ones starting from the unknown true parameter. Such a result demonstrates that our algorithms are quite robust to different initial guesses and do not suffer from the issue of local minimum. On the contrary, we can see that SLS, SMLE, and MRC are all sensitive to the initial guess. Under the above methods, the estimators starting from parametric Logit regression differ significantly from those starting from the unknown true parameter, and such difference even explodes for SLS and SMLE when we consider estimators starting from the origin point. The above results highlight the numerical robustness of our estimators. Finally, for computation time, we can see that KBGD algorithm is relatively time-consuming while SBGD algorithm is extremely time-efficient compared with existing methods.

Table 8: Sensitivity to Initial Points: $u \sim N(0,1)$

		Sensitivity		Running Time (Seconds)		
	Method	S_L	S_Z	True	Logit	Zeros
$n = 2500$	KBGD	0.0025	0.0022	110.99	25.494	178.22
	SBGD	0.0021	0.0023	0.6477	0.4543	0.9677
	SLS	0.0758	383.14	62.927	61.633	92.688
	SMLE	0.0757	515.67	64.967	65.545	223.319
	MRC	0.0598	0.5114	8.8695	7.5822	21.086
$n = 5000$	KBGD	0.0023	0.0024	353.21	52.811	694.78
	SBGD	0.0023	0.0026	1.1284	0.7802	2.0912
	SLS	0.0497	1256.8	237.60	240.89	379.09
	SMLE	0.0480	1826.7	255.10	264.30	1198.6
	MRC	0.0373	0.4199	43.467	37.779	122.74

5 Empirical Application

Table 9: Estimation Results

	(I)	(II)	(III)
Estd. Coefficients	1.0695***	0.9350***	0.9486***
	(0.0914)	(0.1580)	(0.1437)
Num. of Obs.	21805	21805	21805
Estimation Methods	Logit	KBGD	SBGD
Running Time	0.0440	714.73	97.550
Num. of Iterations	–	1617	1741

Note: Running time are all in seconds. For Logit regression, we report the coefficient of education divided by that of total asset. *** indicates significance at 1% level. For the KBGD estimator, in each round of update, the calculation of kernel estimators is distributed over 6 cores using parallel computation. For SBGD estimator, the smallest and largest orders of sieves are 9 and 31, respectively.

As a demonstrative example, this section applies our new algorithms to study how education background affects household risk aversion. In the existing researches, it's extensively documented that risk aversion is significantly correlated with the level of education, although the directions of correlation are mixed, see Outreville (2015) for a comprehensive review. In this study, we investigate how the household-level educational background affects household's risk preferences as well as its investing behaviors.

We use the national survey data from 2019 China Household Financial Survey Project (CHFS) (Gan et al., 2014), which provides household-level information over demographics, asset and debt, income and consumption, social security and insurance, and various household's subjective preferences. The dependent variable we are interested in is the degree of risk aversion of the household. In particular, y_i is constructed to take value of 0 if the i-th household is completely against any form of risks and thus is described as being extremely risk averse; it takes value of 1 if the family is willing to bear some form of risks when making investments. We study how the conditional probability of $y_i = 1$ is affected by a set of factors based on the binary choice model. The key factor that we are particularly interested in is educational backgrounds, which is defined as the average years of education across all the members in the household. We also consider a set of control variables including gender, ethnicity, health conditions, marital status, geographic region of residence, economic knowledge and household totoal asset, whose impacts on the risk aversion are of interest on their own right.

Before estimation, we standardize all covariates so that the resulting variables have zero mean and unity variance. When conducting semiparametric estimation, we normalize the coefficient of total asset to 1. This is because, intuitively, household with more total asset will be more tolerant to risks, indicating that total asset has positive impacts over the conditional probability of $y_i = 1$[8]. To provide a comparison to the semiparametric estimation results, we first conduct parametric Logit regression and report the normalized coefficients in regression (I) in Table 9. We then conduct KBGD and SBGD estimation and report the estimated coefficients of education in (II) and (III). The implementation details of the algorithms are the same as those in the previous section.

As we can see from Table 9, no matter which estimation methods we use, the coefficient of educational background is estimated to be positive with significance at 1% level. This implies that, holding other conditions fixed, increase in the average years of education of the households leads to increase of household's willingness to bear risks. Comparing the semiparametric estimation results with that of Logit regression, we find that the KBGD and SBGD estimators are close to each other, which are both smaller than that of Logit regression, indicating that parametric estimation might suffer from model misspecification and lead to an overestimation of the impacts of education on risk aversion. We finally compare the computation time of each method. We can see that both KBGD and SBGD estimators take much longer to converge compared with the parametric estimation. Comparatively, the SBGD algorithm is significantly faster than the KBGD algorithm. This implies that SBGD algorithm might be more advantageous when there are large number of data points.

[8] Preliminary Probit and Logit estimation also support such argument.

6 Conclusions

In this paper, we proposed new estimation procedures for binary choice and monotonic index models with increasing dimensions. Existing semiparametric estimation procedures for this model cannot be implemented in practice when the number of regressors is large, and thus are particularly unsuitable for big data models such as those considered in much of the machine learning literature. In contrast, our algorithm-based procedures can be used for many regressor models as it involves convex optimization at each iteration of the procedure. We show this iterative procedure also has desirable asymptotic properties when the number of regressors increases with the sample size in ways that are standard in big data literature.

References

Alekh Agarwal, Sham Kakade, Nikos Karampatziakis, Le Song, and Gregory Valiant. Least squares revisited: Scalable approaches for multi-class prediction. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2014.

Hyungtaik Ahn, Hidehiko Ichimura, James L Powell, and Paul A Ruud. Simple estimators for invertible index models. *Journal of Business & Economic Statistics*, 36(1):1–10, 2018.

A. Belloni, V. Chernozhukov, D. Chetverikov, and Y. Wei. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Annals of Statistics*, 46:3643–3675, 2018.

Herman J Bierens. Consistency and asymptotic normality of sieve ml estimators under low-level conditions. *Econometric Theory*, 30(5):1021–1076, 2014.

M.D. Cattaneo, M. Jansson, and W.K. Newey. Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34:277–301, 2018a.

M.D. Cattaneo, M. Jansson, and W.K. Newey. Inference in linear regression models with many covariates and heteroskedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361, 2018b.

Christopher Cavanagh and Robert P Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351–382, 1998.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632, 2007.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

Yanqin Fan, Fang Han, Wei Li, and Xiao-Hua Zhou. On rank estimators in increasing dimensions. *Journal of Econometrics*, 214(2):379–412, 2020.

Li Gan, Zhichao Yin, Nan Jia, Shu Xu, Shuang Ma, Lu Zheng, et al. Data you need to know about china. *Springer Berlin Heidelberg. https://doi*, 10:978–3, 2014.

Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

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

Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

W.K. Newey and F. Windmeijer. Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719, 2009.

J François Outreville. The relationship between relative risk aversion and the level of education: A survey and implications for the demand for life insurance. *Journal of economic surveys*, 29(1):97–111, 2015.

James L Powell, James H Stock, and Thomas M Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

Y. Shin and Z. Todorov. Exact computation of the maximum rank correlation estimator. *Forthcoming, Econometrics Journal*, 2021.

Thomas M Stoker. Consistent estimation of scaled coefficients. *Econometrica: Journal of the Econometric Society*, pages 1461–1481, 1986.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116 (29):14516–14525, 2019.

Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Haiyong Wang and Shuhuang Xiang. On the convergence rates of legendre approximation. *Mathematics of computation*, 81(278):861–877, 2012.