Inference on High Dimensional Selective Labeling Models

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

[1]Department of Economics, Boston College
[2]Deptartment of Economics, Harvard University
[3]Department of Economics, Louisiana State University

October 24, 2024

Abstract

A class of simultaneous equation models arise in the many domains where observed binary outcomes are themselves a consequence of the existing choices of of one of the agents in the model. These models are gaining increasing interest in the computer science and machine learning literatures where they refer the potentially endogenous sample selection as the *selective labels* problem. Empirical settings for such models arise in fields as diverse as criminal justice, health care, and insurance. For important recent work in this area, see for example Lakkaraju et al. (2017), Kleinberg et al. (2018), and Coston, Rambachan, and Chouldechova (2021) where the authors focus on judicial bail decisions, and where one observes the outcome of whether a defendant filed to return for their court appearance only if the judge in the case decides to release the defendant on bail. Identifying and estimating such models can be computationally challenging for two reasons. One is the nonconcavity of the bivariate likelihood function, and the other is the large number of covariates in each equation. Despite these challenges, in this paper we propose a novel distribution free estimation procedure that is computationally friendly in many covariates settings. The new method combines the semiparametric batched gradient descent algorithm introduced in Khan, Lan, Tamer, and Yao (2022) with a novel sorting algorithms incorporated to control for selection bias. Asymptotic properties of the new procedure are established under increasing dimension conditions in both equations, and its finite sample properties are explored through a simulation study and an application using similar judicial bail data.

Key Words: Selective Label Models, Semiparametric Batched Gradient Descent, Selection Bias.

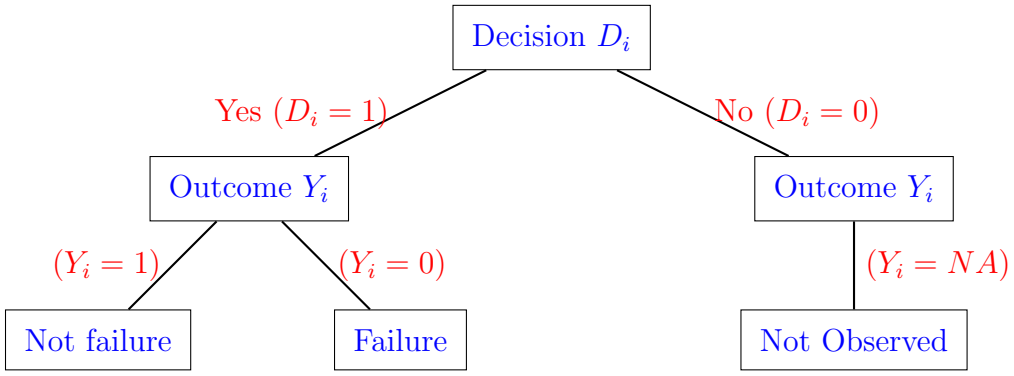
JEL Codes: C14, C31, C35, C63.

1. Introduction

This paper addresses the challenge of inference in large-dimensional selective labeling models. These models emerge in numerous domains where the observed binary outcomes result from the choices made by one of the agents within the system. Recently, they have garnered significant attention in the fields of computer science and machine learning, where this issue is known as the "selective labels problem", referring to potentially endogenous sample selection (in the language of econometricians). Applications of these models span a wide range of areas, including criminal justice, healthcare, and insurance. For important recent work in this area, see for example Lakkaraju et al. (2017), Kleinberg et al. (2018) and Coston et al. (2021). These authors focus on judicial bail decisions, where one observes the outcome of whether a defendant filed to return for their court appearance only if the judge in the case decides to release the defendant on bail. Letting d_i denote the binary decision to grant bail, and y_i denote the binary outcome of the defendant returning for court appearance, they consider a model of the form

$$Y_i = \begin{cases} 0 \text{ or } 1, & \text{if } D_i = 1 \\ \text{not observed (NA)}, & \text{otherwise} \end{cases} \tag{1.1}$$

This process and the ensuing model can be best explained with the diagram below. The top node indicates the decision made by the agent (judge in our criminology example) which corresponds to a *yes* ($D_i = 1$) or *no* ($D_i = 0$) on individual i. The other observed dependent variable, corresponding to the two nodes beneath the top one, is denoted by Y_i, where $Y_i \in \{0, 1, NA\}$ and denotes the resulting outcome (return to court in our example). The selective labels problem occurs because the observation of outcome Y_i is constrained by the decision D_i made by the judge:



Of course controlling for selection bias has a rich history in the econometrics literature, but usually for models where the outcome variable after selection is continuous. Seminal work

in the parametric literature include Heckman (1974) and in the semiparametric literature, see Ahn and Powell (1993a), Das, Newey, and Vella (2003) and Newey (2009).

With the availability of regressors for each of the equations in the binary outcome our econometric model is of the form:

$$D_i = I\left(z_{0,i} + Z_i'\delta_0 - U_i > 0\right) \tag{1.2}$$

$$Y_i = D_i \cdot I\left(x_{0,i} + X_i'\beta_0 - V_i > 0\right) \tag{1.3}$$

The above system of equations is of a similar structure to that used in the classical sample selection model introduced in Heckman (1974). $(z_{0,i}, Z_i)$ and $(x_{0,i}, X_i)$ denote vectors of observed regressors in selection and outcome equations, respectively. D_i and Y_i denote observed *binary* outcomes. δ_0 and β_0 are vectors of unknown regression coefficients, that are the same dimensions as Z_i, X_i respectively and the main parameters of interest. This model is different from the standard selection model, which is for the case where the outcome equation is linear.

Some Extension of Model

It is possible that the model above can accommodate endogeneity in some of the regressors. We recall that in the standard selection model with linear outcome equation, Ahn and Powell (1993a) identify and estimate the coefficients of endogenous regressors using an instrumental variable approach. However there methods is not applicable to the selective labeling model where the outcome equation is binary.

In fact even in the absence of selection bias, identifying regression coefficients of endogenous regressors in binary outcome models is complicated and requires assumptions stronger than the standard instrumental variable assumptions. Blundell and Powell (2004) adopt a control function approach that among other conditions, requires modeling the first stage equation relating endogenous regressors to instruments. Furthermore, they require the endogenous regressor(s) be continuously distributed with large support and thus rule out examples in the treatment effects literature. Abrevaya et al. (2010) consider a binary outcome model with a binary endogenous variable and show how to infer the sign of its coefficient but not its magnitude. Vytlacil and Yildiz (2007) attain point identification of both the regression coefficient and the average treatment effect but require a monotonicity condition and a large regressor support condition. Shaikh and Vytlacil (2011) maintain monotonicity but relax large support conditions to partially identify the average treatment effect. Chen et al. (2024) relax the monotonicity condition and consider a wider class of nonlinear models

with discrete endogenous regressors, also focusing on reduced form parameters such as the average treatment effect. Khan et al. (2023a) impose a factor structure on unobservables in a triangular system of binary equations and show ho this aids in point identifying coefficients of endogenous regressors.

None of the papers in the above expansive list consider models with sample selection as is the case in our selective labeling model. For the model where the outcome equation without selection is binary, recent work is in Abrevaya et al. (2010), who considered identification, estimation and inference of the unknown parameters. However, the estimation approach taken in that paper was based on rank regression methods, analogous to that used in Han (1987). Consequently the objective functions involved are non-smooth and nonconvex, making its implementation very difficult, even more so in large dimensional models which is what this paper is about.

The structure of the rest of the paper is organized as follows. In the next section we define our new (algorithmic) estimation procedures for the unknown regression coefficients. Both are designed to be computationally efficient, and suitable to implement for models where the dimensions of z_i and/or x_i are large. Section 3 explores the asymptotic properties of the new methods, establishing their limiting distribution theory for models of both fixed and increasing dimension. The latter class is gaining widespread and growing interest in the big data and machine learning literatures, but has yet to be studied for this selective labeling model. Section 4 explores the finite sample properties of our procedures by means of a simulation study. Section 7 concludes by summarizing our results and suggesting areas for future work. An Appendix collects tabular results from the simulation study and all the proofs of the main theorems.

2. Algorithmic Estimation Procedures

This section proposes two novel computationally friendly algorithms for estimating β_0. Our proposed methods are to first estimate parameter vector δ_0 in the selection equation, and with that, use matching as in Ahn and Powell (1993a) or series expansion as in Das et al. (2003) and Newey (2009) to estimate the selection correction function, and finally estimate β_0. We will not use rank estimation in either step because the dimension of Z_i and X_i potentially can be large. Instead, we adopt iteration-based methods which feature simple implementation and fast computation speed.

2.1. Estimating First Step

We first introduce the algorithm for the first-step estimator of δ_0. Let $\Phi_\delta(\cdot)$ be a $(q+1)$-dimensional vector of basis functions. The algorithm for estimating δ_0 is described as follows.

Algorithm 0 for estimating δ_0:

1. Start with $k = 0$ and $\widehat{\delta}^0, \widehat{\pi}^0, \widehat{F}_U^0$, where $\widehat{\delta}^0$ is the initial guess of the parameter vector in the selection equation, $\widehat{\pi}^0$ is the initial guess of sieve coefficient, and $\widehat{F}_U^0$ is the initial guess of distribution function of U_i.

2. With $\widehat{\delta}^k$, update the sieve coefficient to $\widehat{\pi}^{k+1}$ using the following

$$\widehat{\pi}^{k+1} = \left(\sum_{i=1}^{n} \Phi_\delta(z_{0,i} + Z_i'\widehat{\delta}^k)' \Phi_\delta(z_{0,i} + Z_i'\widehat{\delta}^k) \right)^{-1} \left(\sum_{i=1}^{n} \Phi_\delta(z_{0,i} + Z_i'\widehat{\delta}^k)' D_i \right)$$

3. With $\widehat{\pi}^{k+1}$, update $\widehat{F}_U^k$ to $\widehat{F}_U^{k+1}$ as $\widehat{F}_U^{k+1}(u) = \Psi(u)'\widehat{\pi}^{k+1}$.

4. With $\widehat{F}_U^{k+1}$, update $\widehat{\delta}^k$ to $\widehat{\delta}^{k+1}$ using

$$\widehat{\delta}^{k+1} = \widehat{\delta}^k - \frac{\gamma_k}{n} \sum_{i=1}^{n} \left(\widehat{F}_U^{k+1}\left(z_{0,i} + Z_i'\widehat{\delta}^k \right) - D_i \right) Z_i$$

   where $\gamma_k > 0$ is learning rate.

5. Set $k = k + 1$ and go back to Step 2, until some terminating conditions are satisfied.

Above is the sieve-based gradient descent estimator (SBGD) proposed by Khan, Lan, Tamer, and Yao (2022). Under some regularity conditions, Khan et al. (2022) show that for k sufficiently large, $\widehat{\delta}_k$ is consistent and asymptotically normally distributed under increasing dimensions.

2.2. Estimating Second Step

Denote the first-step estimator as $\widehat{\delta}$. With $\widehat{\delta}$ in hand, we now consider estimating β_0. We will do something similar to SBGD, but essentially control for selection bias. To provide some intuition, suppose that we know the joint CDF of U_i and V_i denoted as $F(u, v)$, then the probability of $V_i < v$ conditional on $U_i < u$ is given by $P(V_i < v | U_i < u) = F(u, v)/F_U(u) \equiv$

$G(u,v)$, where $F_U(u)$ denotes the marginal distribution of U_i. So $P(Y_i = 1|D_i = 1) = G(z_{0,i} + Z_i'\delta_0, x_{0,i} + X_i'\beta_0)$. If we also knew δ_0, batch gradient descent using the loss function in Khan et al. (2022) leads to the following iterative algorithm for estimating β_0

$$\widehat{\beta}^{k+1} = \widehat{\beta}^k - \frac{\gamma_k}{S_n}\sum_{i=1}^{n} D_i\left(G(z_{0,i} + Z_i'\delta_0, x_{0,i} + X_i'\widehat{\beta}^k) - Y_i\right)X_i, \tag{2.1}$$

where $S_n = \sum_{i=1}^n D_i$. Obviously, in each round of the above update, the conditional probability conditioned on event $D_i = 1$ effectively controls for the selection bias. However, since both $F(u,v)$ and δ_0 are unknown, the above algorithm is infeasible. The latter issue can be resolved by plugging in our first-step estimator $\widehat{\delta}$, while the former remains unsolved. An intuitive solution to the above problem is to obtain an estimator for the conditional probability $G(u,v)$ and then plug such estimators into the above update. We propose two methods to estimate such conditional probability, one being local in nature and the other global.

The first local estimator is closer to that in Ahn and Powell (1993a) in that it uses matching to control for selection bias. To provide some intuition, suppose that the first-step estimator $\widehat{\delta}$ is consistent (see following for technical conditions) and $\widehat{\beta}^k$, the starting point in the k-th iteration, is close to β_0. In this case,

$$E\left(Y_j|\, Z_{e,j}, X_{e,j}, D_j = 1\right) \approx G(z_{0,j} + Z_j'\widehat{\delta}, x_{0,j} + X_j'\widehat{\beta}^k).$$

So long as G is smooth enough and $(z_{0,j} + Z_j'\widehat{\delta}, x_{0,j} + X_j'\widehat{\beta}^k)$ is close enough to $(z_{0,i} + Z_i'\widehat{\delta}, x_{0,i} + X_i'\widehat{\beta}^k)$, Y_j can be used as a (noisy) replacement for $G(z_{0,i} + Z_i'\widehat{\delta}, x_{0,i} + X_i'\widehat{\beta}^k)$. Then the conditional probability for the Y_i can be constructed as a weighted combination of Y_j's, where decreasing weight is assign to each Y_j as the distance between $(z_{0,j} + Z_j'\widehat{\delta}, x_{0,j} + X_j'\widehat{\beta}^k)$ and $(z_{0,i} + Z_i'\widehat{\delta}, x_{0,i} + X_i'\widehat{\beta}^k)$ increases. The above weighting scheme is essentially in line with Ahn and Powell (1993a).

To improve computational efficiency of our algorithm, we propose nearest-neighbour-type[1] weighting scheme. Define

$$d_{ij}^k = \|(z_{0,j} + Z_j'\widehat{\delta}, x_{0,j} + X_j'\widehat{\beta}^k) - (z_{0,i} + Z_i'\widehat{\delta}, x_{0,i} + X_i'\widehat{\beta}^k)\|.$$

For any i with $D_i = 1$, rearrange the indices of Y_j with $j \neq i$ and $D_j = 1$,

[1]Kernel-based weights are also easy to construct, which can be similarly done as in Khan et al. (2022). However, constructing the weights involves $O(n^2)$ computational burdens in each round, which may cause heavy computation burdens, see Yao (2024).

as $\nu^k(i,1),\cdots,\nu^k(i,S_n-1)$, such that $d^k_{i,\nu^k(i,1)} \leq \cdots \leq d^k_{i,\nu^k(i,S_n-1)}$. Then the weights based on m-nearest neighbour is given by

$$W^k_{ij} = \begin{cases} 1/m & \text{if } j = \nu^k(i,1),\cdots,\nu^k(i,m), \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Constructing weights based on m nearest neighbour has computational complexity of order $O(mn\log(n))$, which is much faster than constructing kernel-based weights as long as m is small. In applications, m can be chosen as small as 1, as proposed in Yatchew (1997).

The algorithm for estimating β_0 based on matching is provided as folows.

Algorithm 1 for estimating β_0:

1. Start with $k=0$, first-step estimator $\widehat{\delta}$, initial guess of weights $\{W^0_{ij}\}^n_{i,j=1}$ and initial guess $\widehat{\beta}_0$.
2. With $\widehat{\beta}_k$, update the weights $\{W^k_{ij}\}^n_{i,j=1}$ to $\{W^{k+1}_{ij}\}^n_{i,j=1}$ using (2.2).
3. With $\{W^{k+1}_{ij}\}^n_{i,j=1}$, update $\widehat{\beta}^k$ to $\widehat{\beta}^{k+1}$ using

$$\widehat{\beta}^{k+1} = \widehat{\beta}^k - \frac{\gamma_k}{S_n}\sum_{i=1}^n\sum_{j=1}^n W^{k+1}_{ij} D_i D_j \left(Y_j - Y_i\right) X_i,$$

   where $\gamma_k > 0$ is the learning rate.
4. Set $k = k+1$ and go back to Step 2 until some terminating conditions are satisfied.

We next propose an algorithmic approach which controls for selection with by nonparametrically estimating the selection correction function globally based on a series approximation. This was done for the standard selection model with linear outcome equation in Das, Newey, and Vella (2003) and Newey (2009). Let $\Phi(\cdot,\cdot)$ be a $(q+1)^2$-dimensional vector of basis functions. To provide some intuition to our algorithm, note that we have the following representation of D_iY_i,

$$D_iY_i = D_i\Phi\left(z_{0,i} + Z_i'\widehat{\delta}, x_{0,i} + X_i'\widehat{\beta}^k\right)'\Pi_q + \mathcal{E}_{q,k,i}, \tag{2.3}$$

where Π_q is the unknown true sieve parameter[2], and $\mathcal{E}_{q,k,i}$ is the error that can be decomposed

[2]For any sequence of sieve functions $\{\phi_{s,t}(u,v)\}^\infty_{s,t=0}$ that is complete in $C(R^2)$ space, for any func-

as follows

$$\begin{aligned}\mathcal{E}_{q,k,i} = & \underbrace{D_i\left(G\left(z_{0,i}+Z_i'\widehat{\delta}, x_{0,i}+X_i'\widehat{\beta}^k\right)-\Phi_q\left(z_{0,i}+Z_i'\widehat{\delta}, x_{0,i}+X_i'\widehat{\beta}^k\right)'\Pi_q\right)}_{\text{Error due to truncation of the sieve space}} \\ & + \underbrace{D_i\left(G\left(z_{0,i}+Z_i'\delta_0, x_{0,i}+X_i'\widehat{\beta}^k\right)-G\left(z_{0,i}+Z_i'\widehat{\delta}, x_{0,i}+X_i'\widehat{\beta}^k\right)\right)}_{\text{Error due to first-step estimation}} \\ & + \underbrace{D_i\left(G\left(z_{0,i}+Z_i'\delta_0, x_{0,i}+X_i'\beta_0\right)-G\left(z_{0,i}+Z_i'\delta_0, x_{0,i}+X_i'\widehat{\beta}^k\right)\right)}_{\text{Error due to second-step estimation of the } k\text{-th round}} \\ & + \underbrace{D_i\left(Y_i-G\left(z_{0,i}+Z_i'\delta_0, x_{0,i}+X_i'\beta_0\right)\right)}_{\text{Sampling randomness}}.\end{aligned}$$

When $G(u,v)$ is smooth enough, the first term on the right side of the above equation will be small as long as q is large. Moreover, suppose again that the first-step estimator $\widehat{\delta}$ is consistent and $\widehat{\beta}^k$ is close to β_0, then both second and third terms are small. Finally, the expectation of the last term conditioned on $Z_{e,i}$, $X_{e,i}$ and $D_i = 1$ is zero. The above discussion indicates that the unknown sieve parameter Π_q can be estimated by an OLS-type estimator

$$\begin{aligned}\widehat{\Pi}_q^k = & \left[\sum_{i=1}^n D_i\Phi\left(z_{0,i}+Z_i'\widehat{\delta}, x_{0,i}+X_i'\widehat{\beta}_k\right)\Phi\left(z_{0,i}+Z_i'\widehat{\delta}, x_{0,i}+X_i'\widehat{\beta}_k\right)'\right]^{-1} \times \\ & \left[\sum_{i=1}^n D_iY_i\Phi\left(z_{0,i}+Z_i'\widehat{\delta}, x_{0,i}+X_i'\widehat{\beta}_k\right)\right], \qquad (2.4)\end{aligned}$$

and the unknown conditional probability function $G(u,v)$ is estimated by $\widehat{G}^k(u,v) = \Phi(u,v)'\widehat{\Pi}^k$. Given the estimator of $G(u,v)$, we can plug it back to (2.1), and conduct the update. The algorithm is detailed as follows.

Algorithm 2 for estimating β_0:

1. Start with $k = 0$, first-step estimator $\widehat{\delta}$, initial guess $\widehat{\beta}^0$, $\widehat{\Pi}_q^0$, $\widehat{G}_n^0(u,v)$. The second is the guess of sieve coefficient, and the third is the guess of the conditional probability.
2. With $\widehat{\beta}^k$, update $\widehat{\Pi}^k$ to $\widehat{\Pi}^{k+1}$ using (2.4).
3. With $\widehat{\Pi}_q^{k+1}$, update $\widehat{G}^k$ to $\widehat{G}^{k+1}$ using $\widehat{G}^{k+1}(u,v) = \Phi(u,v)'\widehat{\Pi}_q^{k+1}$.

tion $G(u,v) \in C(R^2)$, there exists a sequence of sieve coefficients $\{\pi_{s,t}\}_{s,t=0}^{\infty}$ such that $G(u,v) = \sum_{s,t=0}^{\infty}\pi_{s,t}\phi_{s,t}(u,v)$. Then Π_q is the vector of first $(q+1)^2$ function-specific sieve coefficients. See Chen (2007) for more detailed discussion.

4. With $\widehat{G}^{k+1}$, update $\widehat{\beta}^k$ to $\widehat{\beta}^{k+1}$ using

$$\widehat{\beta}^{k+1} = \widehat{\beta}^k - \frac{\gamma_k}{S_n} \sum_{i=1}^{n} D_i \left(\widehat{G}^{k+1} \left(z_{0,i} + Z_i' \widehat{\delta}, x_{0,i} + X_i' \widehat{\beta}^k \right) - Y_i \right) X_i,$$

where $\gamma_k > 0$ is the learning rate.

5. Set $k = k + 1$ and go back to Step 2 until some terminating conditions are satisfied.

Comparisons between two algorithms. When using nearest-neighbour-based weights, the matching-based algorithm features fast computation. The only tuning parameter is m, the number of neighbours. When we use the nearest neighbour to perform the update, the algorithm is completely tuning free and is the most computationally efficient. However, the resulting estimator is more volatile and has larger variance. Compared with the matching-based algorithm, the sieve-based estimator has smaller variance, but it takes longer time to compute. The most time-consuming part is the calculation of the sieve coefficient $\widehat{\xi}_{k+1}$. Given tuning parameter Q, we use a total of $(Q+1)^2$ basis functions, so the computational complexity of calculating $\widehat{\xi}_{k+1}$ will be at least $O(Q^6)$, which is large even we choose $Q = 10$.

Choice of terminating conditions. For the sieve-based approach, we follow Khan et al. (2022) and terminate the update if $\max_j \left| \widehat{\beta}_{k+1,j} - \widehat{\beta}_{k,j} \right| < \varrho$, where ϱ is some small positive constant, say 10^{-5}. For the matching-based approach, it generally does not converge so we consider an alternative terminating condition. In particular, let T be a positive integer. We terminate the update if for a successive of T rounds of updates, $\max_{1 \le m \le k} \beta_{k,j}$ and $\min_{1 \le m \le k} \beta_{k,j}$ do not change for all j.

3. Asymptotic Analysis

This section studies the statistical properties of the iteration-based estimators proposed in the previous section. To ease notation, denote $Z_{e,i} = (z_{0,i}, Z_i')' \in \mathcal{Z}_e \subseteq R^{pz+1}$, $X_{e,i} = (x_{0,i}, X_i')' \in \mathcal{X}_e \subseteq R^{px+1}$, $\delta_0 \in \mathscr{D} \subseteq R^{pz}$, $\beta_0 \in \mathscr{B} \subseteq R^{px}$. Let $F(u,v)$ be the joint CDF of (U_i, V_i). Denote the marginal CDF of U_i and V_i as $F_U(u) = \lim_{v \to \infty} F(u,v)$ and $F_V(v) = \lim_{u \to \infty} F(u,v)$. Finally, denote $\mathbb{Z} = z_0 + Z'\delta_0$, $\mathbb{Z}_{0,i} = z_{0,i} + Z_i'\delta_0$, $\widehat{\mathbb{Z}}_0 = z_0 + Z'\widehat{\delta}$, $\widehat{\mathbb{Z}}_{0,i} = z_{0,i} + Z_i'\widehat{\delta}$, $\mathbb{X} = x_0 + X'\beta_0$, $\mathbb{X}_i = x_{0,i} + X_i'\beta_0$, and $\mathbb{X}_i^k = x_{0,i} + X_i'\widehat{\beta}^k$

We impose the following conditions on the data generating process and the data set we observe.

Condition 1. *$Z_{e,i}, X_{e,i}, U_i, V_i, D_i, Y_i$ are iid over i and satisfy (1.2) and (1.3). U_i and V_i are jointly independent of $Z_{e,i}$ and $X_{e,i}$. Finally, we observe the data set $\mathcal{S}_n = \{Z_{e,i}, X_{e,i}, D_i, Y_i\}_{i=1}^n$.*

We next make assumption on the first-step estimator. In particular, we assume that $\widehat{\delta}$ has the following asymptotically linear representation.

Condition 2. *The first-step estimator $\widehat{\delta}$ guarantees the following representation,*

$$\left\| \sqrt{n}\left(\widehat{\delta} - \delta_0\right) - \Psi_\delta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\delta\left(Z_{e,i}\right)\left(D_i - F_U\left(\mathbb{Z}_{0,i}\right)\right) \right\| = o_p\left(1\right),$$

where Ψ_δ is a $(p_Z+1)\times(p_Z+1)$ nonrandom invertible matrix and $\psi_\delta(\cdot)$ is a $(p_Z+1)\times 1$ nonrandom function, whose arguments can change with q. Finally, there holds $\|\widehat{\delta} - \delta_0\| = O_p(\sqrt{p_Z/n})$.

Remark 1. *Condition 2 simply repeats the results of Theorem 8 in Khan et al (2023); more primitive conditions that guarantee such condition can be found therein. According to Theorem 8 of Khan et al. (2022), we have that $\psi_\delta(Z_e) = Z - E_{\widetilde{Z}_e}\left(\widetilde{Z}\middle|\, \widetilde{z}_0 + \widetilde{Z}'\delta_0 = \mathbb{Z}_0\right)$ and $\Psi_\delta = E\left(F_U'\left(\mathbb{Z}_0\right)\psi_\delta\left(Z_e\right)Z'\right)$, where $\widetilde{Z}_e$ is a independent copy of Z_e, and $E_{\widetilde{Z}_e}$ is expectation with respect to $\widetilde{Z}_e$.*

The next condition is imposed on the data generating process.

Condition 3. *For all n, there hold*

(i) $\mathcal{Z}_e = [0,1]^{p_Z+1}$ and $\mathcal{X}_e = [0,1]^{p_X+1}$;

(ii) There exists some constant $C_0 > 0$ such that $\mathscr{D} \subseteq [-C_0, C_0]^{p_Z}$ and $\mathscr{B} \subseteq [-C_0, C_0]^{p_X}$;

(iii) There exists a positive constant $C_G > 0$ such that $\|\nabla_u G\|_\infty, \|\nabla_v G\|_\infty$, and $\|\nabla_{vv} G\|_\infty$ are upper bounded by C_G;

(iv) There exists some constant $C_D > 0$ such that $P(D_i = 1) = E(F_U(z_{0,i} + Z_i'\delta_0)) \geq C_D$.

3.1. The Matching-Based Estimator

In this paper we will specifically focus on the matching-based algorithm with $m = 1$.

3.2. The Sieve-Based Estimator

This section studies the statistical properties of the KBGD algorithm proposed in the previous section. Define We further introduce some technical conditions.

Condition 4. *The vector of basis functions Φ_q satisfies*

(i) Let Φ_j denote the j-th argument of Φ. Then for all $1 \leq j \leq (q+1)^2$, $\|\Phi_j\|_\infty \leq C_{\Phi,q}$, $\|\nabla_u \Phi_j\|_\infty \leq C_{\Phi,1,q}$, $\|\nabla_v \Phi_j\|_\infty \leq C_{\Phi,1,q}$, $\|\nabla_{uu} \Phi_j\|_\infty \leq C_{\Phi,2,q}$, $\|\nabla_{uv} \Phi_j\|_\infty \leq C_{\Phi,2,q}$, and $\|\nabla_{vv} \Phi_j\|_\infty \leq C_{\Phi,2,q}$, where $C_{\Phi,q}, C_{\Phi,1,q}$ and $C_{\Phi,2,q}$ are all positive constants that depend on q only;

(ii) Define $\Gamma_q(\beta) = E[\Phi_q(\mathbb{Z}_{0,i}, x_{0,i} + X_i'\beta)\Phi_q(\mathbb{Z}_{0,i}, x_{0,i} + X_i'\beta)' | D_i = 1]$. There exist $0 < \underline{\lambda}_\Phi \leq \overline{\lambda}_\Phi < \infty$ such that $\underline{\lambda}_\Phi \leq \inf_{\beta \in \mathscr{B}} \underline{\lambda}(\Gamma_q(\beta)) \leq \sup_{\beta \in \mathscr{B}} \overline{\lambda}(\Gamma_q(\beta)) \leq \overline{\lambda}_\Phi$ for all q;

(iii) $\|G(u,v) - \Phi(u,v)'\Pi_q\|_\infty \leq \mathscr{R}_q$;

Condition 5. *$\mathscr{X}(\upsilon_Z, \upsilon_X, \beta) = E(X | \mathbb{Z}_0 = \upsilon_Z, x_0 + X'\beta = \upsilon_X, D = 1)$ exists for all $\upsilon_Z, \upsilon_X, \beta \in \mathscr{B}$. Moreover, there hold*

(i) Let $\mathscr{X}_j(\upsilon_Z, \upsilon_X, \beta)$ denote the j-th argument of $\mathscr{X}(\upsilon_Z, \upsilon_X, \beta)$. There exists a positive constant C_X such that for all $1 \leq j \leq p_X$, $\|\nabla_{\upsilon_Z} \mathscr{X}_j(\upsilon_Z, \upsilon_X, \beta)\|_\infty \leq C_X$, $\|\nabla_{\upsilon_X} \mathscr{X}_j(\upsilon_Z, \upsilon_X, \beta)\|_\infty \leq C_X$, and $\|\nabla_\beta \mathscr{X}_j(\upsilon_Z, \upsilon_X, \beta)\|_\infty \leq C_X p_X^{1/2}$;

(ii) $\sup_{\upsilon_Z, \upsilon_X, \beta} \|\mathscr{X}(\upsilon_Z, \upsilon_X, \beta) - \Pi_q^X(\beta)\Phi(\upsilon_Z, \upsilon_X)\| \leq \mathscr{R}_{X,q}$, where $\Pi_q^X(\beta)$ is the unknown sieve coefficient matrix.

Condition 6. *Let*

$$\Psi(\beta) = \int_0^1 E\left(\nabla_v G\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i} + \tau X_i'\Delta\beta\right)\left(X_i - \mathscr{X}\left(\mathbb{Z}_{0,i}, x_{0,i} + X_i'\beta, \beta\right)\right)X_i' \middle| D_i = 1\right) d\tau.$$

There hold $\sup_{\beta \in \mathscr{B}} \overline{\lambda}(\Psi(\beta) + \Psi'(\beta)) \leq \overline{\lambda}_\Psi < \infty$ and $\inf_{\beta \in \mathscr{B}} \underline{\lambda}(\Psi(\beta) + \Psi'(\beta)) \geq \underline{\lambda}_\Psi > 0$.

Define

$$\begin{aligned}\Xi_{1,n} = {} & n^{-1/2} p_X^{1/2} q^4 C_{\Phi,q}^3 (p_Z C_{\Phi,u,q} + C_{\Phi,q}) + \mathscr{R}_q \left(p_X^{1/2} q^2 C_{\Phi,q}^2 + p_X\right) \\ & + (\log(n)/n)^{-1/2} \left(p_X q^4 C_{\Phi,q}^3 + p_X^{3/2}\right) + \mathscr{R}_q^X.\end{aligned}$$

Under the above conditions, we have the following result, whose proof is in the Appendix, Section B.

Theorem 1. *Let Conditions 1-6 hold and* $\Xi_{1,n} \to 0$. *Then there exists some* $0 < \overline{\gamma} < \infty$ *such that if we choose* $\gamma_k = \gamma$ *for all* $k \geq 1$, *where* $0 < \gamma < \overline{\gamma}$, *there holds*

$$\sup_{k \geq k(n,\gamma)} \|\beta_k - \beta_0\| = O_p(\Xi_{1,n}),$$

where $k(n,\gamma)$ *is a threshold that depends on* n *and* γ.

Theorem 2. *Let all the conditions in Theorem 1, then we have that*

$$\begin{aligned}\Delta\beta_{k+1} = &\left(I_{p_X} - \gamma\Psi(\beta_0)\right)\Delta\beta_k + \frac{\gamma}{n}\sum_{i=1}^{n} D_i\left(X_i - X^E(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}, \beta_0)\right)\varepsilon_i \\ &- \frac{\gamma}{n}\sum_{i=1}^{n}\Sigma_{X,Z}\Psi_\delta^{-1}\psi_\delta(\mathbb{Z}_{0,i})\left(D_i - F_U(\mathbb{Z}_{0,i})\right) + W_{n,k},\end{aligned}$$

where $\Sigma_{X,Z} = E\left(\nabla_u G(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i})\left(X_i - X^E(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}, \beta_0)\right) Z_i' | D_i = 1\right)$ *and* $\sup_{k \geq k(n,\gamma)} \|W_{n,k}\| = O_p(\Xi_{2,n})$. *We have that*

$$\begin{aligned}\sup_{k \geq k(n,\gamma) - \log\left(\Xi_{2,n}^{-1}\right)/\log(1-\gamma\underline{\lambda}_\Psi/8)} \Bigg\| &\sqrt{n}\Delta\beta_k - \Psi^{-1}(\beta^\star)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\left(X_i - X^E(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}, \beta_0)\right)\varepsilon_i \\ &+\Psi^{-1}(\beta^\star)\Sigma_{X,Z}\Psi_\delta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_\delta(\mathbb{Z}_{0,i})\left(D_i - F_U(\mathbb{Z}_{0,i})\right)\Bigg\| = o_p(1).\end{aligned}$$

Theorem 2 directly leads to the following corollary.

Corollary 1. *Let all the conditions in Theorem 2 hold. Define* $\widehat{\beta} = \beta_k$ *for any* $k \geq k(n,\gamma) - \log\left(\Xi_{2,n}^{-1}\right)/\log(1 - \gamma\underline{\lambda}_\Psi/8)$. *For any* $p_X \times 1$ *vector* ω, *suppose that* $\omega'\Psi^{-1}(\beta^\star)\left(X - X^E(\mathbb{Z}_0, \mathbb{X}_0, \beta_0)\right) \to_{a.} \mathbb{X}(\omega)$ *and* $\omega'\Psi^{-1}(\beta^\star)\Sigma_{X,Z}\Psi_\delta^{-1}\psi_\delta(\mathbb{Z}_0) \to_{a.s.} \mathbb{Z}(\omega)$ *as* $n \to \infty$, *where* $\mathbb{X}(\omega)$ *and* $\mathbb{Z}(\omega)$ *are random variables with bounded second moment. Then we have that*

$$\begin{aligned}\sqrt{n}\omega'\Delta\widehat{\beta} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\mathbb{X}(\omega)\varepsilon_i - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{Z}(\omega)\left(D_i - F_U(\mathbb{Z}_{0,i})\right) + o_p(1) \\ &\to_d N\Big(0, E\left(G(\mathbb{Z}_0, \mathbb{X}_{0,i})\left(1 - G(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i})\right)F_U(\mathbb{Z}_{0,i})\mathbb{X}_i(\omega)\mathbb{X}_i(\omega)'\right) \\ &\quad + E\left(F_U(\mathbb{Z}_{0,i})\left(1 - F_U(\mathbb{Z}_{0,i})\right)\mathbb{Z}_i(\omega)\mathbb{Z}_i(\omega)'\right)\Big).\end{aligned}$$

4. Monte Carlo Simulations

This section conducts some simulation experiments to evaluate the performance of the proposed estimators. We consider four competing methods.

The first method is parametric estimation using nonlinear least squares, which is similar to Heckman's two-step estimation but accounts for the binary response in the second stage. We assume that U_i and V_i in (1.2) and (1.3) have zero mean and unit variance, and are jointly normally distributed with covariance ρ. Then we can jointly estimate (rescaled) δ_0, β_0 together with ρ. In particular, let $F_1(\cdot)$ and $F_2(\cdot,\cdot,\rho)$ denote the CDF's of univariate standard normal distribution and bivariate normal distribution with zero mean, unit variance, and correlation ρ. Also define $\bar{Z}_{e,i} = (1, z_{0,i}, Z_i')'$, $\bar{X}_{e,i} = (1, x_{0,i}, X_i')'$, $\bar{\delta} = (c_{\delta,0}, c_{\delta,1}, \delta')'$, $\bar{\beta} = (c_{\beta,0}, c_{\beta,1}, \beta')'$. In the first step, we minimize the following loss function

$$L_{1,n}(\bar{\delta}) = \frac{1}{n}\sum_{i=1}^{n}(D_i - F_1(\bar{Z}_{e,i}'\bar{\delta}))^2$$

and obtain the minimizer $\widehat{\bar{\delta}}$. Then in the second step, we minimize the following loss function

$$L_{2,n}(\bar{\beta},\rho) = \frac{1}{n}\sum_{i=1}^{n} D_i\left(Y_i - \frac{F_2(\bar{Z}_{e,i}'\widehat{\bar{\delta}}, \bar{X}_{e,i}'\bar{\beta},\rho)}{F_1(\bar{Z}_{e,i}'\widehat{\bar{\delta}})}\right)^2$$

and obtain the minimizer $\widehat{\bar{\beta}}$. Then the two-step NLS estimators for δ_0 and β_0 are given by $\widehat{c}_{\delta,1}^{-1}\widehat{\delta}$ and $\widehat{c}_{\beta,1}^{-1}\widehat{\beta}$.

The second method is parametric maximum likelihood estimation. Like in the first method, we also assume that U_i and V_i in (1.2) and (1.3) are jointly normally distributed, then the log-likelihood function is then given by

$$\begin{aligned} L_{3,n}\left(\bar{\delta},\bar{\beta},\rho\right) = \frac{1}{n}\sum_{i=1}^{n}\Big((1-D_i)\log(1-F_1(\bar{Z}_{e,i}'\bar{\delta})) + D_iY_i\log\left(F_2(\bar{Z}_{e,i}'\bar{\delta},\bar{X}_{e,i}'\bar{\beta},\rho)\right) \\ +D_i(1-Y_i)\log\left(F_1(\bar{Z}_{e,i}'\bar{\delta}) - F_2(\bar{Z}_{e,i}'\bar{\delta},\bar{X}_{e,i}'\bar{\beta},\rho)\right)\Big). \end{aligned}$$

Suppose the MLE estimators are given by $\widehat{\bar{\delta}}$ and $\widehat{\bar{\beta}}$, then the MLE estimators for δ_0 and β_0 are given by $\widehat{c}_{\delta,1}^{-1}\widehat{\delta}$ and $\widehat{c}_{\beta,1}^{-1}\widehat{\beta}$.

The third method is semiparametric estimation based on matching. In particular, we first obtain the estimator of δ_0 in the first step, then we conduct Algorithm 1. To improve the

computational efficiency, in both first and second step estimation, we use nearest neighbour matching with $m = 1$.

The fourth method is semiparametric estimation based on series approximation. For the sieve functions, we consider the Legendre polynomials used in Khan et al. (2022), and use tensor products of one-variate sieve functions as the sieve functions for bivariate functions.

We report the bias, the root mean squared error, and the running time of all methods. Tables of results are reported in the appendix, Section A.

5. Empirical Illustrations

In this section we further illustrate the finite sample properties of our proposed procedures through two empirical illustrations- one which controls for selection in a binarized female labor supply equation and the other which studies the likelihood of rearrested after controlling for the bias induced by the sample only being of those released after the first arrest.

5.1. Female Labor Supply

In this subsection we revisit the well known Mroz (1987) labor supply data set. This data set was also used in Ahn and Powell (1993b), Newey et al. (1990) and Khan and Nekipelov (2024) to compare parametric and semiparametric methods. However, in those papers the focus was the sample selection model whereas here we estimate a selective labeling model.

In the original Mroz (1987) study, the sample consists of measurements on the characteristics of 753 married women, with 428 being employed and 325 unemployed. The dependent variable in the outcome equation, the annual hours of work, is specified to depend upon six regressors: the logarithm of the wage rate, household income less the woman's labor income, indicators for young and older children in the household, the woman's age and years of education. Mroz's study also used the square of experience and various interaction terms as instrumental variables for the wage rate, and were also included in his Probit analysis of employment status, resulting in 18 parameters to be estimated in the first equation. Ahn and Powell (1993b) use the same conditioning variables in the first equation but only the original 10 variables in their first stage kernel regression to attain estimators of the slope coefficients in the outcome (hours worked) equation.

Here, notationally, the first-stage estimates the model with a binary outcome D_i indicating

TABLE 1. ESTIMATION RESULTS FOR MROZ (1987)'S DATA SET

	Selection Equation			Outcome Equation		
	2Step NLS	Joint MLE	2Step SBGD	2Step NLS	Joint MLE	2Step SBGD
Num. of Kids < 6 years	$-0.43^{\dagger}$	$-0.43^{\dagger}$	$-0.44^{\dagger}$	0.55	0.60	0.97
	(0.07)	(0.18)	(0.10)	(0.47)	(1.87)	(0.76)
Num. of Kids 6-18 years	0.03	0.04	0.04	-1.00	-1.00	-1.00
	(0.07)	(0.26)	(0.06)	(0.00)	(0.00)	(0.00)
Age	6.32	$-2.07^{\dagger}$	0.13	0.28	0.33	$0.74^{\dagger}$
	(6.83)	(0.30)	(6.60)	(0.50)	(0.57)	(0.45)
Education	1.86	$-1.25^{\dagger}$	0.24	$-1.58^{\dagger}$	$-1.34^{\dagger}$	$-1.87^{\dagger}$
	(4.90)	(0.20)	(4.98)	(0.51)	(0.76)	(0.58)
Log wage				$0.85^{\dagger}$	0.48	0.85
				(0.43)	(1.74)	(0.61)
Mother's education	0.05	0.03	0.03			
	(0.07)	(0.16)	(0.06)			
Father's education	-0.02	0.01	-0.01			
	(0.07)	(0.13)	(0.07)			
Unemployment Rate	-0.06	-0.06	-0.05			
	(0.05)	(0.24)	(0.05)			
SMSA dummy	-0.00	-0.01	-0.01			
	(0.06)	(0.16)	(0.05)			
Experience	1.00	1.00	1.00			
	(0.00)	(0.00)	(0.00)			
Non-wife income	$-0.17^{\dagger}$	$-0.15^{\dagger}$	$-0.15^{\dagger}$	$0.69^{\dagger}$	0.28	$0.53^{\dagger}$
	(0.06)	(0.17)	(0.07)	(0.38)	(1.37)	(0.11)
Experience squared	$-0.48^{\dagger}$	$-0.43^{\dagger}$	$-0.48^{\dagger}$			
	(0.16)	(0.15)	(0.09)			

Note: Standard deviation is provided in the parenthesis. † indicates statistical significance at 10%.

the status of female labor participation. When $D_i = 1$, we can further observe the hours worked, denoted by Y_i^*. To apply our estimator to a selective labeling model, we binarize Y_i^* as

$$Y_i = I(Y_i^* > \text{med}(Y_i^*)),$$

where $\text{med}(Y_i^*)$ is the observed median of Y_i^* for $D_i = 1$. We use Y_i as our second-stage binary outcome, indicating whether a women in the labor market work more compared to the average level. We follow Newey et al. (1990) and consider 18 covariates in the first-stage

model and 6 covariates in the second-stage model, see Mroz (1987) and Newey et al. (1990) for more details.

Similar to the simulation section, we consider three competing methods including two-step NLS, MLE, and two-step SBGD. When conducting semiparametric estimation we normalize the coefficient of experience to 1 in the first-step equation because, intuitively, an individual with more working experience is more likely to be in the job market. For the outcome equation, we normalize the coefficient of the number of kids between 6 and 18 years to -1 because we expect that women with more number of young children will spend more time with family so are less likely to work more than average. When estimating the model, we standardize all the regressors so the standardized regressors have zero mean and unit variance. For the first-step SBGD estimation, we use the same sieve functions as in Khan et al. (2022), and for the second-step SBGD estimation, we use tensor products of the above sieve functions. The order of sieve functions is determined by an AIC-type criterion function[3]. For both steps of estimation, we use MLE estimators as the initial guess. Finally, we choose $\gamma_k = 1$ for both steps, and terminate the iteration when the maximum change across all arguments of the parameters after an iteration is less than 10^{-6} or the number of iterations exceeds 10^6.

We report the estimation results of both steps using different methods in Table 1. We can see in Table 1 that estimation results of two-step NLS and two-step SBGD are similar in terms of statistical significance. In particular, coefficients of number of kids under 6 years old, non-wife income and squared experience are all estimated to be negative and significant at 10%. Indeed, for such three coefficients, using different specifications or estimation methods leads to almost identical results. We also find that when using joint MLE estimation, apart from the above three coefficients, the coefficients of age and education are also significant at 10%. But when we compare them with the coefficients estimated based on two-step SBGD, we find that age and education do not have significant impacts on the job market participation. Moreover, the estimation results of the two coefficients differ in both sign and scale. When we further look at the second-stage estimation results reported in the right three columns in Table 1, we find that, interestingly, parametric and semiparametric estimators differ systematically for some coefficients. For example, under both two-step NLS and joint MLE, the coefficient of age is estimated to be insignificant at 10%, but under semiparametric estimation, the impacts of age become statistically significant. The significance mainly

[3]For selection equation, the criterion is given by $\log\left(\widehat{\sigma}_q^2\right) + 2q/n$, where n is the sample size, q is the number of sieve functions, and $\widehat{\sigma}_q^2 = \sum_{i=1}^n (D_i - \widehat{D}_{q,i})^2/n$, where $\widehat{D}_{q,i}$ is the predicted $E(D_i|Z_{e,i})$. For the outcome equation, the criterion is given by $\log\left(\widehat{\sigma}_q^2\right) + 2q/\sum_{i=1}^n D_i$, where q is the number of sieve functions, and $\widehat{\sigma}_q^2 = \sum_{i=1}^n D_i \cdot (Y_i - \widehat{Y}_{q,i})^2/\sum_{i=1}^n D_i$, where $\widehat{Y}_{q,i}$ is the predicted $E(Y_i|Z_{e,i}, X_{e,i}, D_i = 1)$.

comes from the increase of the scale of the estimated coefficient: compared with two-step NLS or joint MLE, the estimated coefficient of age under two-step SBGD almost triples. This highlights the difference between semiparametric and parametric estimation, which also indicates potential model misspecification. Our semiparametric estimator indicates departure from bivariate normality of in this data set.

5.2. Juvenile Defendants in Criminal Courts

In this section we apply our estimation procedures to a sample of juvenile defendants drawn from the State Court Processing Statistics (SCPS) for 1998. SCPS 1998 tracked felony cases filed in May 1998 until final disposition or until one year had elapsed from the date of filing. The SCPS presents data on felony cases filed in approximately 40 of the nation's 75 most populous counties in 1998. These 75 counties account for more than a third of the United States population and approximately half of all reported crimes. Data were collected on arrest charges, demographic characteristics, criminal history, pretrial release and detention, adjudication, and sentencing. Within each sampled site, data were gathered on each juvenile felony case. Cases were tracked through adjudication or for up to one year. Further details of this data set can be found in USDOJ (2003), Griffin, Torbet, and Szymanski (1998).

To implement our procedure we estimate a selective labeling model. The first (selected) equation is modeled as a semiparametric binary response model, where the dependent binary variable was RELEASE, indicating whether of not a defendant was released before trial, and the regressors were type of charge[4], RELADULT, a binary variable indicating previous relationship with the adult court at time of arrest, RELJUV, a binary variable indicating previous relationship with the juvenile court at time of arrest, age at time of arrest, sex, race, prior arrest as a juvenile, prior arrest as an adult.

For the binary outcome equation, the dependent variable is the binary variable REARREST indicating whether or not the defendant was rearrested before the trial date. The regressors were the same as selection equation expect that we include a new regressor HOMECOF indicating whether the defendant was confined to home, and exclude the variable ADPRIOR, so we have exclusion.

[4]We classfiy are charge types into five categories. ChargeType 1 refers to violent offenses, including murder, rape, robbery, assault, other violent offenses. ChargeType 2 refers to drug offenses, including drug trafficking and other drug offenses. ChargeType 3 refers to property offenses, including burglary, theft, motor vehicle theft, fraud, forgery, and other property offenses. ChargeType 4 refers to public order offenses, including weapons, and other public order offenses. ChargeType 5 refers other offenses, including driving-related offenses, other felony, and misdemeanor. We drop observations with ChargeType 5 due to too few observations. Then we drop variable ChargeType4 to ensure identification.

When estimating semiparametric selection and outcome equations, we need to respectively normalize one covariate so that its coefficient is 1. When estimating the selection equation, we normalize the coefficient of negative ChargType1 to be 1. This is motivated by the intuition that when the defendant is involved in violent offenses such as murder or rape, the defendant is less likely to be get released. When estimating the outcome equation, we normalize negative HOMECOF. This is also intuitive since, if the defendant is released but is confined to home, then the defendant is less likely to conduct crimes and is less likely to get rearrested before the trail date. The estimation results are reported in Table 2. As was the case in the previous empirical example, these differences indicate misspecification in the parametric estimators of regression coefficents and indicates the bivariate normality assumption is not reflected in this data set.

Table 2. Estimation Results of Juvenile Defendants in Criminal Courts

	Selection Equation			Outcome Equation		
	2Step NLS	Joint MLE	2Step SBGD	2Step NLS	Joint MLE	2Step SBGD
Reladult	$-0.45^{\dagger}$	-0.42	$-0.45^{\dagger}$			
	(0.17)	(0.32)	(0.25)			
Reljuv	$-0.48^{\dagger}$	$-0.48^{\dagger}$	$-0.51^{\dagger}$			
	(0.15)	(0.10)	(0.24)			
Sex	-0.17	-0.16	-0.15	$2.08^{\dagger}$	0.21	$0.33^{\dagger}$
	(0.13)	(0.13)	(0.14)	(1.24)	(0.13)	(0.05)
Juvprior	$-0.39^{\dagger}$	$-0.37^{\dagger}$	$-0.38^{\dagger}$	1.00	1.00	1.00
	(0.15)	(0.14)	(0.20)	(0.00)	(0.00)	(0.00)
Adprior	-0.21	-0.20	-0.22	0.24	0.18	$0.59^{\dagger}$
	(0.17)	(0.15)	(0.19)	(0.18)	(0.12)	(0.31)
Age	$0.81^{\dagger}$	$0.76^{\dagger}$	$0.82^{\dagger}$	0.34	0.40	0.14
	(0.14)	(0.18)	(0.33)	(0.23)	(0.12)	(0.31)
Black	$1.08^{\dagger}$	$1.09^{\dagger}$	$1.28^{\dagger}$	$2.52^{\dagger}$	$7.27^{\dagger}$	$1.13^{\dagger}$
	(0.37)	(0.24)	(0.66)	(0.13)	(0.67)	(0.64)
White	$1.23^{\dagger}$	$1.27^{\dagger}$	$1.42^{\dagger}$	1.59	$6.29^{\dagger}$	0.31
	(0.37)	(0.20)	(0.69)	(1.20)	(0.68)	(0.64)
Charge Type 1	-1.00	-1.00	-1.00	0.43	$0.36^{\dagger}$	$0.86^{\dagger}$
	(0.00)	(0.00)	(0.00)	(0.39)	(0.21)	(0.27)
Charge Type 2	$0.52^{\dagger}$	$0.52^{\dagger}$	0.52	-0.18	0.02	-0.07
	(0.31)	(0.27)	(0.48)	(0.30)	(0.23)	(0.17)
Charge Type 3	0.23	0.22	0.22	0.13	0.19	0.05
	(0.29)	(0.19)	(0.36)	(0.28)	(0.21)	(0.19)

Note: Standard deviation is provided in the parenthesis. † indicates statistical significance at 10%.

6. Extensions (Preliminary)

In this section we discuss extensions of the methods introduced tin the previous section to estimate regression coefficients in different models. As was the case with the semiparametric selective labeling model we considered previously, the motivation of the estimators proposed in this section are computational ease when there are a moderate or large number of regressors. One is for a semiparametric multinomial choice model and the the other for

semiparametric panel data selective labeling models.

6.1. Multinomial Choice Models

We illustrate here how our proposed method can be used to estimate the standard multinomial response model where the dependent variable takes one of $J+1$ mutually exclusive and exhaustive alternatives numbered from 0 to J. Following the notation similar to that used in Khan et al. (2021), for individual i, alternative $\jmath$ is assumed to have an unobservable indirect utility $y^*_{i\jmath}$. The alternative with the highest indirect utility is assumed chosen. Thus the observed choice $y_{i\jmath}$ can be defined as

$$y_{i\jmath} = \mathbf{1}[y^*_{i\jmath} > y^*_{ik}, \forall k \neq \jmath]$$

with the convention that $y_{i\jmath} = 0$ indicates that the choice of alternative $\jmath$ is not made by individual i. As is standard in the literature, an assumption of joint continuity of the indirect utilities rules out ties (with probability one). In addition, we maintain the familiar linear form for indirect utilities[5]

$$\begin{aligned} y^*_{i0} &= 0, \\ y^*_{i\jmath} &= x'_{i\jmath}\beta_0 - \epsilon_{i\jmath}, \ \jmath = 1, ..., J, \end{aligned} \tag{6.1}$$

where β_0 is a p-dimensional vector of unknown preference parameters of interest whose first component is normalized to have absolute value 1 (scale normalization). Note that for alternative $\jmath = 0$, the standard (location) normalization $y^*_{i0} = 0$ is imposed. The vector $\epsilon_i \equiv (\epsilon_{i1}, ..., \epsilon_{iJ})'$ of unobserved error terms, attained by stacking all the scalar idiosyncratic errors $\epsilon_{i\jmath}$, is assumed to be jointly continuously distributed and independent of the $p \times J$-dimensional vector of regressors $x_i \equiv (x'_{i1}, ..., x'_{iJ})'$[6]. We stress that expression (6.1) is rather general. By properly re-organizing $x_{i\jmath}$'s and β_0, (6.1) can accommodate both alternative-specific and individual-specific covariates[7]

Previous semiparametric contributions to estimating this model include Lee (1995), who

[5]Our method can be applied to more general models with indirect utilities $y^*_{i\jmath} = u_\jmath(x'_{i\jmath}\beta_0, -\epsilon_{i\jmath})$, $\jmath = 1, 2$, where $u_\jmath(\cdot, \cdot)$'s are unknown (to econometrician) $\mathbb{R}^2 \mapsto \mathbb{R}$ functions strictly increasing in each of their arguments. It will be clear that our rank procedure does not rely on the additive separability of the regressors and error terms.

[6]We impose the independence restriction here to simplify exposition. As explained in Khan et al. (2021), this matching-based approach allows ϵ_i to be correlated with individual-specific regressors.

[7]Note the identification of models with both alternative-specific and individual-specific regressors will need to take two steps, of which the first step only identifies the coefficients on alternative-specific regressors.

proposes a profile likelihood approach, extending the results in Klein and Spady (1993) for the binary response model. Ahn et al. (2017) propose a two-step estimator that requires nonparametric methods but show the second step is of closed-form. Shi et al. (2018) also propose a two-step estimator in panel setups exploiting a cyclic monotonicity condition, which also requires a high dimensional nonparametric first stage, but whose second stage is not closed-form as Ahn et al. (2017) is.

Khan et al. (2021) proposed a local rank procedure. We define it here for the special case where $J = 2$, in which case the model is

$$\begin{aligned} y^*_{i0} &= 0, \\ y^*_{ij} &= x'_{ij}\beta_0 - \epsilon_{ij}, \;\; j = 1, 2. \end{aligned}$$

One way to estimate β_0 for this model proposed in Khan et al. (2021) was a weighted rank type estimator:[8] rank correlation estimator, analogous to the maximum rank correlation (MRC) estimator proposed in Han (1987), defined as the maximizer, over the parameter space $\mathcal{B}$, of the objective function

$$G_{1n}(b) = \frac{1}{n(n-1)} \sum_{i \neq \ell} \mathbf{1}[x_{i2} = x_{\ell 2}](y_{i1} - y_{\ell 1}) \cdot sgn((x_{i1} - x_{\ell 1})'b), \tag{6.2}$$

where above we denote pairs of individuals by i, ℓ and recall the second subscript denotes the choice.

The motivation for that estimator were robustness properties, notably when many of the regressors were discrete. Note the matching of regressors is analogous to how we controlled for selection in the earlier part of the paper for the selective labelling model. We also note the nonsmoothness and nonconvexity of the objective function make implementation difficult, especially when there are a moderately large number of regressors.

For the model at hand we propose the following algorithm, which is analogous to the matching algorithm earlier in the paper (Algorithm 1) and keep notation as close as possible to that used there. A sieve based algorithm could also be considered be we omit that here.

Define $d^k_{i\ell} = \|(x_{i2} - x_{\ell 2})'\hat{\beta}_k\|$. Rearrange the indices of $y_{\ell 2}$ with $\ell \neq i$ as $\nu^k_{i,1}, \nu^k_{i,2}, \cdots, \nu^k_{i,n}$, such that $d^k_{i,\nu^k_{i,1}} \leq d^k_{i,\nu^k_{i,2}} \cdots \leq d^k_{i,\nu^k_{i,n}}$. Then the weights based on m-nearest neighbour is given by

[8] Here the weights correspond to binary, "exact" matches of each component of the vector x_2. For continuously distributed regressors they were replaced with kernel weights in Khan et al. (2021).

$$W_{i\ell}^k = \begin{cases} 1/m & \text{if } \ell = \nu_{i,1}^k, \nu_{i,2}^k, \cdots, \nu_{i,m}^k, \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

Algorithm 3 for estimating β_0:

1. Start with $k = 0$, weights $\{W_{i\ell}^k\}_{i,\ell=1}^n$ and initial guess $\widehat{\beta}_0$.
2. With $\widehat{\beta}_k$, update the weights $\{W_{i\ell}^k\}_{i,\ell=1}^n$ as $\{W_{i\ell}^{k+1}\}_{i,\ell=1}^n$ using (6.3).
3. With $\{W_{i\ell}^k\}_{i,\ell=1}^n$, update $\widehat{\beta}^k$ to $\widehat{\beta}^{k+1}$ using

$$\widehat{\beta}^{k+1} = \widehat{\beta}^k - \frac{\gamma_k}{n} \sum_{i=1}^n \sum_{\ell=1}^n W_{i\ell}^k \left(y_{i1} - y_{\ell 1}\right) x_i,$$

   where $\gamma_k > 0$ is the learning rate.
4. Set $k = k + 1$ and go back to Step 2 until some terminating conditions are satisfied.

6.2. Selective Label Models with Endogenous Treatment

In recent seminal work, Lee (2009) considers identifying treatment parameters in treatment effect models with attrition but does not allow for explanatory variables. Extending our selective labeling model to allow for an endogenous treatment variable, denoted by T_i which could be continuous or discrete, could be expressed as

$$\begin{aligned} D_i &= I[z_{0,i} + Z_i'\delta_0 + T_i\beta_{01} > U_i] \\ Y_i &= D_i \cdot I[x_{0,i} + X_i'\delta_0 + T_i\beta_{02} > V_i] \\ T_i &= 1[W_i\gamma_0 > \nu_i] \end{aligned}$$

The key here is that (U_i, V_i, ν_i) are mutually correlated. so treatment is endogenous. A parameter of interest is β_{02} which relates to the effect of treatment on outcome. In a model where for $D_i = 1$ Y_i was linear and there were no regressors besides T_i, Lee (2009) considered identification. In a model where there was no selection/attrition issue so D_i was identical to 1, a identification and estimation was considered in Vytlacil and Yildiz (2007), Abrevaya et al. (2010), Shaikh and Vytlacil (2011). Vytlacil and Yildiz (2007) attained point identification under a monotonicity condition as well as support conditions on exogenous covariates

effecting Y_i. None of methods proposed in these papers are applicable to the model above where there is endogenous treatment, attrition and selective labeling.

6.3. Panel Data Models

A second area where the models we considered and estimated can be extended to are those for data sets where we observed multiple observations for each agent. In these settings we are able to control for unobserved heterogeneity in ways we could not before, so consequently panel data models are very useful in applied research.

Not only do they allow researchers to study the intertemporal behaviour of individuals, they also enable them to control for the presence of unobserved permanent individual heterogeneity. To date there exists a large body of literature on panel data models with unobserved individual effects that enter additively in the (possibly latent) regression model. Considerable advances in the panel data literature have been made in the direction of dynamic linear and nonlinear models that allow for the presence of lags of the dependent variable. These are reviewed for example Arellano and Honoré (2001), who also describe results for dynamic non-linear panel data models. An important setting involves sample selection models- see Kyriazidou (1997), Kyriazidou (2001). However, little is known about settings in a selection panel data model when the outcome variable is binary, as was the case in the cross sectional models considered at the outset of this paper.

We express the panel data selective labeling model as

$$\begin{aligned} d_{it} &= I[\alpha_{1i} + w'_{it}\delta_0 + \eta_{it} > 0] && (6.4)\\ y_{it} &= d_{it} \cdot I[\alpha_{2i} + x'_{it}\beta_0 + \epsilon_{it} > 0] && (6.5)\\ i &= 1, 2, ...n, t = 1, 2....T \end{aligned}$$

and for its dynamic variant as

$$\begin{aligned} d_{it} &= I[\alpha_{1i} + w'_{it}\delta_0 + \gamma_0 d_{i,t-1} + \eta_{it} > 0] && (6.6)\\ y_{it} &= d_{it} \cdot I[\alpha_{2i} + x'_{it}\beta_0 + \theta_0 y_{i,t-1} + \epsilon_{it} > 0] && (6.7)\\ i &= 1, 2, ...n, t = 1, 2....T \end{aligned}$$

where $d_{it}, w_{it}, y_{it}, x_{it}$ are observed variables, α_{1i}, α_{2i} are unobserved, denoting individual specific heterogeneity, η_{it}, ϵ_{it} are also unobserved, denoting idiosyncratic shocks.

We note the dynamic variant included lagged dependent variables as explanatory variables.

For work in other dynamic nonlinear panel data models see, for example, Honoré and Kyriazidou (2000) and Khan et al. (2023b) (binary), Hu (2002) (censored), Khan et al. (2016) (Roy), Kyriazidou (2001) (selection). We would characterize the model here in equations (6.6), (6.7) with lagged dependent variables, as a dynamic selective labeling model.

There is much recent interest in dynamic binary choice panel since Honoré and Kyriazidou (2000), but little work for system of (static or dynamic) binary equations like the ones above. Our work here would propose similar algorithmic procedures to estimate δ_0, β_0 in situations where, as in the cross sectional setting, they are of moderate or large dimension.

7. Conclusions

This paper considers estimation and inference for large dimensional semiparametric selective labeling models. Statistically these models have a similar structure to sample selection models with binary, as opposed to linear outcome equations in the second stage. It is this binary/binary structure which makes computation of the model particularly difficult when compared to the standard selection model, especially for large dimensional (i.e many regressor) models.

To address this problem we propose novel algorithmic procedures which are computationally fast, and derive their asymptotic properties even for the case where the dimension increases with the sample size. We demonstrate the finite sample properties of our proposed procedures by a simulation study.

Our work here motivates areas for future research. For example to further ease implementation, a bivariate penalization scheme would be useful for model selection in this settings, and its asymptotic validity would need to be proven. Furthermore, the usefulness of our methods in other empirical settings in economics, biostatistics and medicine would be worthy of exploration.

References

Abrevaya, J., J. Hausman, and S. Khan (2010): "Testing for Causal Effects in a Generalized Regression Model with Endogenous Regressors," *Econometrica*, 78, 2043–2061.

Ahn, H. and J. Powell (1993a): "Semiparametric Estimation of Censored Selection Models," *Journal of Econometrics*, 58, 3–29.

——— (1993b): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3–29.

Ahn, H., J. Powell, H. Ichimura, and P. Ruud (2017): "Simple Estimators for Invertible Index Models," *Journal of Business Economics and Statistics*, 36, 1–10.

Arellano, M. and B. Honoré (2001): "Panel Data Models: Some Recent Developments," *Handbook of econometrics. Volume 5*, 3229–96.

Blundell, R. and J. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 56, 37–58.

Chen, S., S. Khan, and X. Tang (2024): "Endogeneity in Weakly Separable Models witgout Endogeneity," *Journal of Econometrics*, 238, 1–14.

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

Coston, A., A. Rambachan, and A. Chouldechova (2021): "Charecterizing Fairness Over the Set of Good Models Under Selective Labels," *Proceedings of the 38th International Conference on Machine Learning*, 139, 2144–2155.

Das, M., W. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

Griffin, P., P. Torbet, and L. Szymanski (1998): "Trying Juveniles as Adults in Criminal Court: An Analysis of State Transfer Provisions." Tech. rep., U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention.

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

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694.

Honoré, B. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

Hu, L. (2002): "Estimation of a Censored Dynamic Panel Model," *Econometrica*, 70, 2499–2517.

Khan, S., X. Lan, E. Tamer, and Q. Yao (2022): "Estimating High Dimensional Monotone Index Models By Iterative Convex Optimization," Boston College Working Paper.

Khan, S., A. Maurel, and Y. Zhang (2023a): "Informational Content of Factor Structures in Simultaneous Binary Response Models," *Advances in Econometrics*, 45, 385–410.

Khan, S. and D. Nekipelov (2024): "On Uniform Inference in Nonlinear Models with Endogeneity," *Journal of Econometrics*, 240, 105261.

Khan, S., F. Ouyang, and E. Tamer (2021): "Inference on Semiparametric Multinomial Response Models," *Quantitative Economics*, 12, 743–777.

Khan, S., M. Ponomareva, and E. Tamer (2016): "Identification of dynamic binary response models," *Journal of Econometrics*, 194, 57–75.

——— (2023b): "Identification of dynamic binary response models," *Journal of Econometrics*, 237, 105515.

Klein, R. and R. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018): "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133, 237–293.

Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364.

——— (2001): "Estimation of Dynamic Panel Data Sample Selection Models," *Review of Economic Studies*, 68, 543–572.

Lakkaraju, H., J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan (2017): "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables," KDD Research Paper.

Lee, D. (2009): "Training, Wages and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Lee, L.-F. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381–428.

Mroz, T. (1987): "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," *Econometrica*, 55, 765–799.

Newey, W., J. Powell, and J. Walker (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review Papers and Proceedings*, 80, 324–328.

Newey, W. K. (2009): "Two-step Series Estimation of Sample Selection Models," *Econometrics Journal*, 12, 217–229.

Shaikh, A. and E. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables," *Econometrica*, 79, 949–955.

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

USDOJ (2003): "Juvenile Defendants in Criminal Courts (JDCC): Survey of 40 Counties in the United States," *Inter-university Consortium for Political and Social Research*.

Vytlacil, E. and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757–779.

Yatchew, A. (1997): "An elementary estimator of the partial linear model," *Economics letters*, 57, 135–143.

A. Monte Carlo Results

Table 3. Finite Sample Performance
$B-\delta, R-\delta, B-\beta, R-\beta$ denote aggregate (across components) bias and RMSE of estimators for δ, β. η, ϵ have marginal cauchy distributions, implying misspecification of MLE. Time denotes computational time in seconds.

$p = 10$										
	$n = 100000$					$n = 200000$				
Method	B-δ	R-δ	B-β	R-β	Time	B-δ	R-δ	B-β	R-β	Time
MLE	1.1562	0.7345	1.1592	0.7208	149.64	1.1477	0.7210	1.1406	0.7045	286.83
M-GD	0.0624	0.1539	0.8273	0.4774	1144.5	0.0344	0.0989	0.6358	0.3520	3163.6
S-GD	0.0336	0.1456	0.0209	0.1786	1936.1	0.0308	0.0956	0.0072	0.1310	3845.5
$p = 50$										
	$n = 100000$					$n = 200000$				
Method	B-δ	R-δ	B-β	R-β	Time	B-δ	R-δ	B-β	R-β	Time
MLE	1.1789	0.7357	1.2034	0.7455	881.57	1.1711	0.7301	1.1927	0.7242	1690.6
M-GD	0.0357	0.1740	0.1305	0.3402	1777.1	0.0266	0.1233	0.1609	0.2508	4634.4
S-GD	0.0394	0.1712	0.0800	0.2346	2099.3	0.0461	0.1226	0.0601	0.1669	4161.2

TABLE 4. FINITE SAMPLE PERFORMANCE OF KERNEL-BASED ESTIMATORS: $\eta_i \sim$ CAUCHY, $\varepsilon_i \sim 0.5\eta_i + \sqrt{0.75}$CAUCHY, $p = 10$

					$n = 100000$						
	Method	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
	MLE	0.0034	0.0228	0.0158	0.0622	0.2147	0.6721	0.0107	0.0237	0.0201	0.1106
Bias	M-GD	0.0053	0.0033	0.0013	0.0034	0.0065	0.0148	0.0017	0.0040	0.0061	0.0159
	S-GD	0.0034	0.0015	0.0002	0.0015	0.0030	0.0081	0.0010	0.0025	0.0031	0.0093
	MLE	0.0432	0.0337	0.0282	0.0689	0.2203	0.6771	0.0240	0.0365	0.0512	0.1394
RMSE	M-GD	0.0374	0.0234	0.0218	0.0293	0.0488	0.0905	0.0184	0.0251	0.0437	0.0834
	S-GD	0.0363	0.0218	0.0210	0.0282	0.0469	0.0856	0.0180	0.0242	0.0413	0.0777
	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	MLE	0.0067	0.0188	0.0157	0.0572	0.1987	0.6327	0.0111	0.0153	0.0049	0.1981
Bias	M-GD	0.0487	0.0473	0.0244	0.0493	0.1000	0.1990	0.0212	0.0488	0.0926	0.1959
	S-GD	0.0023	0.0004	0.0014	0.0014	0.0027	0.0013	0.0011	0.0030	0.0014	0.0058
	MLE	0.0579	0.0375	0.0346	0.0685	0.2078	0.6411	0.0300	0.0379	0.0554	0.2221
RMSE	M-GD	0.0882	0.0717	0.0455	0.0751	0.1458	0.2890	0.0397	0.0712	0.1357	0.2783
	S-GD	0.0497	0.0298	0.0278	0.0352	0.0555	0.1036	0.0235	0.0313	0.0503	0.0931
					$n = 200000$						
	Method	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
	MLE	0.0027	0.0231	0.0166	0.0603	0.2146	0.6678	0.0109	0.0247	0.0200	0.1070
Bias	M-GD	0.0027	0.0022	0.0021	0.0008	0.0037	0.0072	0.0010	0.0023	0.0047	0.0076
	S-GD	0.0022	0.0019	0.0017	0.0006	0.0034	0.0068	0.0009	0.0021	0.0041	0.0070
	MLE	0.0299	0.0288	0.0235	0.0640	0.2172	0.6703	0.0180	0.0308	0.0363	0.1202
RMSE	M-GD	0.0259	0.0163	0.0154	0.0199	0.0319	0.0572	0.0121	0.0162	0.0278	0.0523
	S-GD	0.0254	0.0155	0.0151	0.0195	0.0309	0.0555	0.0120	0.0161	0.0264	0.0502
	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	MLE	0.0013	0.0200	0.0142	0.0559	0.1957	0.6311	0.0100	0.0184	0.0031	0.1908
Bias	M-GD	0.0400	0.0373	0.0194	0.0371	0.0732	0.1551	0.0171	0.0372	0.0716	0.1479
	S-GD	0.0016	0.0010	0.0008	0.0008	0.0009	0.0000	0.0004	0.0007	0.0001	0.0007
	MLE	0.0377	0.0301	0.0264	0.0622	0.2006	0.6357	0.0226	0.0316	0.0414	0.2045
RMSE	M-GD	0.0644	0.0529	0.0333	0.0558	0.1056	0.2145	0.0298	0.0532	0.1003	0.2044
	S-GD	0.0328	0.0210	0.0199	0.0251	0.0426	0.0769	0.0180	0.0223	0.0368	0.0687

TABLE 5. FINITE SAMPLE PERFORMANCE OF KERNEL-BASED ESTIMATORS: $\eta_i \sim$ CAUCHY, $\varepsilon_i \sim 0.5\eta_i + \sqrt{0.75}$CAUCHY, $p = 50$

					$n = 100000$						
	Method	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
	MLE	0.0073	0.0218	0.0164	0.0595	0.2131	0.6662	0.0120	0.0264	0.0231	0.1011
Bias	M-GD	0.0007	0.0006	0.0009	0.0015	0.0005	0.0014	0.0007	0.0004	0.0015	0.0014
	S-GD	0.0004	0.0006	0.0014	0.0006	0.0026	0.0038	0.0000	0.0005	0.0005	0.0031
	MLE	0.0458	0.0313	0.0278	0.0665	0.2181	0.6708	0.0232	0.0364	0.0488	0.1278
RMSE	M-GD	0.0376	0.0216	0.0209	0.0288	0.0455	0.0794	0.0176	0.0255	0.0384	0.0710
	S-GD	0.0379	0.0207	0.0209	0.0283	0.0436	0.0781	0.0174	0.0219	0.0377	0.0699
	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	MLE	0.0006	0.0198	0.0154	0.0550	0.1984	0.6393	0.0081	0.0175	0.0070	0.1977
Bias	M-GD	0.0074	0.0048	0.0043	0.0038	0.0096	0.0235	0.0012	0.0045	0.0108	0.0215
	S-GD	0.0033	0.0014	0.0019	0.0006	0.0043	0.0138	0.0010	0.0022	0.0044	0.0095
	MLE	0.0598	0.0372	0.0363	0.0671	0.2073	0.6485	0.0291	0.0398	0.0588	0.2228
RMSE	M-GD	0.0641	0.0459	0.0370	0.0524	0.0909	0.1811	0.0322	0.0470	0.0845	0.1659
	S-GD	0.0512	0.0290	0.0295	0.0361	0.0590	0.1118	0.0241	0.0315	0.0522	0.0958
					$n = 200000$						
	Method	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
	MLE	0.0023	0.0230	0.0161	0.0614	0.2163	0.6711	0.0112	0.0244	0.0204	0.1082
Bias	M-GD	0.0022	0.0006	0.0002	0.0001	0.0015	0.0021	0.0001	0.0014	0.0009	0.0030
	S-GD	0.0031	0.0016	0.0008	0.0012	0.0038	0.0073	0.0005	0.0024	0.0033	0.0076
	MLE	0.0295	0.0285	0.0225	0.0649	0.2185	0.6737	0.0187	0.0309	0.0374	0.1227
RMSE	M-GD	0.0255	0.0152	0.0144	0.0193	0.0308	0.0575	0.0130	0.0161	0.0272	0.0516
	S-GD	0.0253	0.0146	0.0143	0.0192	0.0306	0.0577	0.0128	0.0164	0.0270	0.0515
	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	MLE	0.0036	0.0221	0.0147	0.0567	0.1992	0.6397	0.0091	0.0166	0.0074	0.1988
Bias	M-GD	0.0120	0.0083	0.0044	0.0072	0.0143	0.0324	0.0046	0.0084	0.0159	0.0294
	S-GD	0.0053	0.0023	0.0006	0.0017	0.0032	0.0101	0.0008	0.0020	0.0049	0.0073
	MLE	0.0397	0.0312	0.0259	0.0628	0.2039	0.6441	0.0210	0.0299	0.0412	0.2114
RMSE	M-GD	0.0482	0.0343	0.0243	0.0380	0.0687	0.1349	0.0229	0.0350	0.0635	0.1249
	S-GD	0.0344	0.0205	0.0193	0.0253	0.0423	0.0801	0.0160	0.0277	0.0368	0.0699

TABLE 6. FINITE SAMPLE PERFORMANCE: $\eta_i \sim N(0,1), \varepsilon_i \sim 0.5\eta_i + \sqrt{0.75}N(0,1)$ $B-\delta, R-\delta, B-\beta, R-\beta$ DENOTE AGGREGATE (ACROSS COMPONENTS) BIAS AND RMSE OF ESTIMATORS FOR δ, β. TIME DENOTES COMPUTATIONAL TIME IN SECONDS.

$p = 10$										
	$n = 100000$					$n = 200000$				
Method	B-δ	R-δ	B-β	R-β	Time	B-δ	R-δ	B-β	R-β	Time
MLE	0.0058	0.0753	0.0152	0.0859	138.550	0.0095	0.0522	0.0030	0.0628	277.94
M-GD	0.0158	0.0769	0.1173	0.1272	1460.0	0.0171	0.0536	0.0875	0.0945	4067.1
S-GD	0.0202	0.0778	0.0250	0.0922	1864.7	0.0257	0.0551	0.0054	0.0659	3693.2
$p = 50$										
	$n = 100000$					$n = 200000$				
Method	B-δ	R-δ	B-β	R-β	Time	B-δ	R-δ	B-β	R-β	Time
MLE	0.0198	0.0945	0.0240	0.1112	797.77	0.0141	0.0670	0.0188	0.0799	1547.6
M-GD	0.0183	0.0965	0.0312	0.1360	1817.8	0.0131	0.0682	0.0215	0.0972	4735.5
S-GD	0.0261	0.0973	0.0230	0.1172	1981.3	0.0192	0.0683	0.0283	0.0844	3879.5

TABLE 7. FINITE SAMPLE PERFORMANCE OF KERNEL-BASED ESTIMATORS:
$\eta_i \sim N(0,1), \varepsilon_i \sim 0.5\eta_i + \sqrt{0.75}N(0,1),\ p = 10$

						$n = 100000$					
	Method	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
	MLE	0.0002	0.0004	0.0003	0.0006	0.0008	0.0004	0.0001	0.0007	0.0007	0.0017
Bias	M-GD	0.0009	0.0010	0.0001	0.0011	0.0019	0.0027	0.0004	0.0013	0.0021	0.0041
	S-GD	0.0011	0.0013	0.0001	0.0015	0.0027	0.0037	0.0005	0.0016	0.0025	0.0053
	MLE	0.0212	0.0124	0.0119	0.0155	0.0250	0.0424	0.0103	0.0128	0.0213	0.0393
RMSE	M-GD	0.0216	0.0126	0.0122	0.0157	0.0256	0.0433	0.0106	0.0130	0.0217	0.0402
	S-GD	0.0216	0.0126	0.0121	0.0157	0.0256	0.0442	0.0105	0.0132	0.0218	0.0408
	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	MLE	0.0021	0.0001	0.0001	0.0007	0.0020	0.0030	0.0008	0.0002	0.0021	0.0042
Bias	M-GD	0.0063	0.0076	0.0044	0.0059	0.0139	0.0289	0.0044	0.0067	0.0130	0.0263
	S-GD	0.0025	0.0007	0.0001	0.0012	0.0026	0.0069	0.0006	0.0008	0.0032	0.0064
	MLE	0.0258	0.0131	0.0139	0.0176	0.0270	0.0483	0.0121	0.0150	0.0235	0.0454
RMSE	M-GD	0.0351	0.0214	0.0188	0.0237	0.0387	0.0736	0.0172	0.0213	0.0366	0.0677
	S-GD	0.0259	0.0140	0.0142	0.0181	0.0287	0.0534	0.0123	0.0158	0.0253	0.0489
						$n = 200000$					
	Method	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
	MLE	0.0015	0.0004	0.0007	0.0001	0.0014	0.0015	0.0000	0.0002	0.0016	0.0021
Bias	M-GD	0.0020	0.0008	0.0010	0.0004	0.0024	0.0034	0.0002	0.0008	0.0025	0.0037
	S-GD	0.0024	0.0013	0.0012	0.0008	0.0035	0.0054	0.0005	0.0013	0.0036	0.0059
	MLE	0.0148	0.0081	0.0077	0.0103	0.0170	0.0295	0.0071	0.0089	0.0145	0.0279
RMSE	M-GD	0.0153	0.0085	0.0079	0.0104	0.0175	0.0303	0.0072	0.0090	0.0149	0.0286
	S-GD	0.0153	0.0084	0.0079	0.0104	0.0178	0.0315	0.0072	0.0091	0.0153	0.0298
	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	MLE	0.0000	0.0001	0.0006	0.0000	0.0007	0.0002	0.0002	0.0002	0.0008	0.0002
Bias	M-GD	0.0066	0.0052	0.0027	0.0043	0.0109	0.0207	0.0024	0.0052	0.0104	0.0190
	S-GD	0.0000	0.0002	0.0005	0.0002	0.0007	0.0021	0.0001	0.0001	0.0004	0.0012
	MLE	0.0182	0.0100	0.0094	0.0126	0.0202	0.0353	0.0087	0.0113	0.0173	0.0333
RMSE	M-GD	0.0252	0.0157	0.0130	0.0178	0.0294	0.0541	0.0116	0.0166	0.0271	0.0512
	S-GD	0.0187	0.0102	0.0095	0.0130	0.0210	0.0376	0.0089	0.0116	0.0182	0.0351

TABLE 8. FINITE SAMPLE PERFORMANCE OF KERNEL-BASED ESTIMATORS:
$\eta_i \sim N(0,1), \varepsilon_i \sim 0.5\eta_i + \sqrt{0.75}N(0,1),\ p = 50$

						$n = 100000$					
	Method	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
	MLE	0.0007	0.0004	0.0005	0.0009	0.0002	0.0013	0.0002	0.0004	0.0006	0.0006
Bias	M-GD	0.0003	0.0003	0.0006	0.0006	0.0003	0.0000	0.0003	0.0002	0.0001	0.0005
	S-GD	0.0003	0.0004	0.0008	0.0000	0.0018	0.0023	0.0007	0.0005	0.0012	0.0031
	MLE	0.0204	0.0124	0.0116	0.0149	0.0237	0.0422	0.0104	0.0129	0.0211	0.0400
RMSE	M-GD	0.0205	0.0130	0.0118	0.0150	0.0239	0.0432	0.0106	0.0133	0.0215	0.0404
	S-GD	0.0205	0.0128	0.0118	0.0152	0.0242	0.0447	0.0106	0.0133	0.0217	0.0416
	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	MLE	0.0008	0.0002	0.0008	0.0008	0.0003	0.0016	0.0003	0.0010	0.0008	0.0017
Bias	M-GD	0.0005	0.0008	0.0014	0.0001	0.0008	0.0020	0.0004	0.0015	0.0015	0.0022
	S-GD	0.0012	0.0001	0.0008	0.0011	0.0006	0.0017	0.0004	0.0005	0.0000	0.0000
	MLE	0.0252	0.0140	0.0132	0.0174	0.0289	0.0481	0.0117	0.0146	0.0237	0.0444
RMSE	M-GD	0.0327	0.0178	0.0167	0.0211	0.0347	0.0583	0.0144	0.0180	0.0285	0.0556
	S-GD	0.0260	0.0150	0.0134	0.0178	0.0307	0.0536	0.0118	0.0154	0.0253	0.0486
						$n = 200000$					
	Method	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
	MLE	0.0005	0.0001	0.0001	0.0002	0.0017	0.0020	0.0000	0.0000	0.0007	0.0004
Bias	M-GD	0.0003	0.0002	0.0001	0.0002	0.0010	0.0009	0.0002	0.0003	0.0000	0.0006
	S-GD	0.0005	0.0009	0.0004	0.0009	0.0005	0.0017	0.0004	0.0010	0.0012	0.0034
	MLE	0.0137	0.0087	0.0079	0.0107	0.0167	0.0306	0.0070	0.0091	0.0148	0.0284
RMSE	M-GD	0.0139	0.0090	0.0080	0.0109	0.0168	0.0308	0.0072	0.0094	0.0149	0.0287
	S-GD	0.0139	0.0089	0.0080	0.0109	0.0169	0.0313	0.0071	0.0093	0.0151	0.0294
	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	MLE	0.0004	0.0005	0.0003	0.0003	0.0014	0.0023	0.0003	0.0010	0.0005	0.0013
Bias	M-GD	0.0008	0.0006	0.0003	0.0001	0.0013	0.0021	0.0002	0.0003	0.0001	0.0006
	S-GD	0.0007	0.0010	0.0005	0.0005	0.0021	0.0057	0.0005	0.0015	0.0012	0.0035
	MLE	0.0165	0.0097	0.0100	0.0126	0.0195	0.0366	0.0085	0.0108	0.0173	0.0331
RMSE	M-GD	0.0207	0.0125	0.0123	0.0153	0.0238	0.0438	0.0101	0.0130	0.0215	0.0400
	S-GD	0.0171	0.0101	0.0103	0.0130	0.0206	0.0407	0.0086	0.0114	0.0186	0.0360

B. Proofs of Main Theorems

In the following, we use $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) if there exists some constant C such that $a_n \leq b_n$ ($a_n \geq Cb_n$) for all sufficiently large n.

Proof of Theorem 1. Denote $\widehat{\mathbb{Z}}_i = z_{0,i} + Z_i'\widehat{\delta}$, $\mathbb{Z}_{0,i} = z_{0,i} + Z_i'\delta_0$, $\mathbb{X}_i^k = x_{0,i} + X_i'\widehat{\beta}^k$, and

$\mathbb{X}_{0,i} = x_{0,i} + X_i'\beta_0$. Also denote $\varepsilon_i = Y_i - G\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right)$. We have that

$$\begin{aligned}
\widehat{\beta}^{k+1} &= \widehat{\beta}^{k} - \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i\left(\Phi_q\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k\right)'\widehat{\Pi}_q^k - Y_i\right)X_i \\
&= \widehat{\beta}^{k} - \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i X_i \Phi_q\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k\right)'\left(\widehat{\Pi}_q^k - \Pi_q\right) - \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i X_i\left(\Phi_q\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k\right)'\Pi_q - G\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k\right)\right) \\
&\quad - \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i X_i\left(G\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k\right) - G\left(\mathbb{Z}_{0,i}, \mathbb{X}_i^k\right)\right) - \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i X_i\left(G\left(\mathbb{Z}_{0,i}, \mathbb{X}_i^k\right) - G\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right)\right) \\
&\quad - \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i X_i \varepsilon_i.
\end{aligned}$$

Now we derive an expression for $\widehat{\Pi}_q^k - \Pi_q$. Denote

$$\widehat{\Gamma}_{n,q}\left(\delta, \beta\right) = \frac{1}{S_n}\sum_{i=1}^{n} D_i \Phi_q\left(z_{0,i} + Z_i'\delta, x_{0,i} + X_i'\beta\right)\Phi_q\left(z_{0,i} + Z_i'\delta, x_{0,i} + X_i'\beta\right)'.$$

Then

$$\begin{aligned}
\widehat{\Pi}_q^k - \Pi_q &= \widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta}, \widehat{\beta}^k\right)\frac{1}{S_n}\sum_{j=1}^{n} D_j \Phi_q\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right) D_j Y_j - \Pi_q \\
&= \widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta}, \widehat{\beta}^k\right)\frac{1}{S_n}\sum_{j=1}^{n} D_j \Phi_q\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right)\left(G\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right) - G\left(\mathbb{Z}_{0,j}, \mathbb{X}_j^k\right)\right) \\
&\quad + \widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta}, \widehat{\beta}^k\right)\frac{1}{S_n}\sum_{j=1}^{n} D_j \Phi_q\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right)\left(G\left(\mathbb{Z}_{0,j}, \mathbb{X}_j^k\right) - G\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right)\right) \\
&\quad + \widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta}, \widehat{\beta}^k\right)\frac{1}{S_n}\sum_{j=1}^{n} D_j \Phi_q\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right)\left(G\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right) - D_j\Phi_q\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right)'\Pi_q\right) \\
&\quad + \widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta}, \widehat{\beta}^k\right)\frac{1}{S_n}\sum_{j=1}^{n} D_j \Phi_q\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right)\varepsilon_j.
\end{aligned}$$

Define $\widehat{X}_n^E\left(\nu_Z, \nu_X, \beta\right) = \frac{1}{S_n}\sum_{i=1}^{n} D_i X_i \Phi_q\left(\widehat{\mathbb{Z}}_{0,i}, x_{0,i} + X_i'\beta\right)'\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta}, \beta\right)\Phi_q\left(\nu_Z, \nu_X\right)$. Taking the expression of $\widehat{\Pi}_q^k - \Pi_q$ into the expression of $\widehat{\beta}^{k+1}$, we have that

$$\widehat{\beta}^{k+1} = \widehat{\beta}^{k} - \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i\left(X_i - \widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k, \beta^k\right)\right)\left(\Phi_q\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k\right)'\Pi_q - G\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k\right)\right)$$
$$- \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i\left(X_i - \widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k, \beta^k\right)\right)\left(G\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k\right) - G\left(\mathbb{Z}_{0,i}, \mathbb{X}_i^k\right)\right)$$
$$- \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i\left(X_i - \widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k, \beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i}, \mathbb{X}_i^k\right) - G\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right)\right)$$
$$- \frac{\gamma}{S_n}\sum_{i=1}^{n} D_i\left(X_i - \widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k, \beta^k\right)\right)\varepsilon_i.$$

The following lemmas are necessary for analyzing the above algorithm.

Lemma 1. *Let Conditions 1-5 hold, then*

$$\sup_{\beta\in\mathscr{B}}\left\|\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta\right) - \Gamma_q\left(\beta\right)\right\| = O_p\left(n^{-1/2}p_Z q^2 C_{\Phi,q}C_{\Phi,1,q} + n^{-1/2}C_{\Phi,q}^2 + \left(\log\left(n\right)/n\right)^{-1/2}p_X^{1/2}q^2C_{\Phi,q}\right).$$

If further $\Xi_{1,n}\to 0$, *then* $P\left(\sup_{\beta\in\mathscr{B}}\overline{\lambda}\left(\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)\right)\leq 1.5\underline{\lambda}_{\Phi}^{-1}\right)\to 1$.

Proof. Recall that $\Gamma_q\left(\beta\right) = E\left[\Phi_q\left(\mathbb{Z}_{0,i}, x_{0,i}+X_i'\beta\right)\Phi_q\left(\mathbb{Z}_{0,i}, x_{0,i}+X_i'\beta\right)'|D_i=1\right]$. To show the first result, we first note that

$$\sup_{\beta\in\mathscr{B}}\left\|\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta\right) - \widehat{\Gamma}_{n,q}\left(\delta_0,\beta\right)\right\| = \sup_{\beta\in\mathscr{B}}\left\|S_n^{-1}\sum_{i=1}^{n} D_i\Phi_q\left(\widehat{\mathbb{Z}}_i, x_{0,i}+X_i'\beta\right)\Phi_q\left(\widehat{\mathbb{Z}}_i, x_{0,i}+X_i'\beta\right)'\right.$$
$$\left. - S_n^{-1}\sum_{i=1}^{n} D_i\Phi_q\left(\mathbb{Z}_{0,i}, x_{0,i}+X_i'\beta\right)\Phi_q\left(\mathbb{Z}_{0,i}, x_{0,i}+X_i'\beta\right)'\right\|$$
$$\lesssim q^2 C_{\Phi,q}C_{\Phi,1,q}\max_{1\leq i\leq n}\left|\widehat{\mathbb{Z}}_i - \mathbb{Z}_{0,i}\right|.$$

According to Condition 2, $\max_{1\leq i\leq n}\left|\widehat{\mathbb{Z}}_i - \mathbb{Z}_{0,i}\right| \lesssim p_Z^{1/2}\left\|\widehat{\delta}-\delta_0\right\| = O_p\left(p_Z n^{-1/2}\right)$, so

$$\sup_{\beta\in\mathscr{B}}\left\|\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta\right) - \widehat{\Gamma}_{n,q}\left(\delta_0,\beta\right)\right\| = O_p\left(n^{-1/2}p_Z q^2 C_{\Phi,q}C_{\Phi,1,q}\right).$$

Also note that $\|\widehat{\Gamma}_{n,q}\left(\delta_0,\beta\right) - P_D^{-1}\frac{1}{n}\sum_{i=1}^{n} D_i\Phi_q\left(\mathbb{Z}_{0,i}, x_{0,i}+X_i'\beta\right)\Phi_q\left(\mathbb{Z}_{0,i}, x_{0,i}+X_i'\beta\right)'\|$ is uni-

formly of order $O_p(n^{-1/2}C_{\Phi,q}^2)$, it then remains to bound the following distance

$$\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\right.$$
$$\left.-\ E\left(D_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\right)\right\|.$$

Note that each argument of $D_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'$ is bounded by $C_{\Phi,q}^2$, and each argument of $\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta_1\right)\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'$ has partial derivative with respect to β that is bounded (in norm) by $C_{\Phi,q}C_{\Phi,v,q}p_X^{1/2}$. So using Lemma A1 of Khan et al (2023), we have that

$$\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\right.$$
$$\left.-\ E\left(D_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\right)\right\|=O_p\left(\left(\log\left(n\right)/n\right)^{-1/2}p_X^{1/2}q^2C_{\Phi,q}\right).$$

This shows the first result. Since $\Xi_{1,n}\to 0$, we have that $\sup_{\beta\in\mathscr{B}}\left\|\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta\right)-\Gamma_q\left(\beta\right)\right\|\to_p 0$. Using the fact that

$$\sup_{\beta\in\mathscr{B}}\left|\underline{\lambda}\left(\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta\right)\right)-\underline{\lambda}\left(\Gamma_q\left(\beta\right)\right)\right|\leq\sup_{\beta\in\mathscr{B}}\left\|\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta\right)-\Gamma_q\left(\beta\right)\right\|\to_p 0,$$

we have that $\underline{\lambda}\left(\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta\right)\right)\geq 2/3\inf_{\beta\in\mathscr{B}}\underline{\lambda}\left(\Gamma_q\left(\beta\right)\right)\geq 2/3\underline{\lambda}_\Phi$ with probability going to 1. Then $\overline{\lambda}\left(\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)\right)\leq 1.5\underline{\lambda}_\Phi^{-1}$ for all $\beta\in\mathscr{B}$ with probability going to 1. This proves the results. □

Lemma 2. *Let Conditions 1-5 hold and $\Xi_{1,n}\to 0$, then*

$$\sup_{Z_e\in\mathcal{Z}_e,X_e\in\mathcal{X}_e,\beta\in\mathscr{B}}\left\|\widehat{X}_n^E\left(\widehat{\mathbb{Z}},x_0+X'\beta,\beta\right)-X^E\left(\mathbb{Z}_0,x_0+X'\beta,\beta\right)\right\|$$
$$=O_p\left(\mathscr{R}_q^X+n^{-1/2}p_Zp_X^{1/2}q^4C_{\Phi,q}^3C_{\Phi,1,q}+n^{-1/2}p_X^{1/2}q^2C_{\Phi,q}^4+\left(\log\left(n\right)/n\right)^{-1/2}p_Xq^4C_{\Phi,q}^3\right).$$

Proof. Note that

$$\sup_{\beta\in\mathscr{B}}\left\|X^E\left(\mathbb{Z}_0,x_0+X'\beta,\beta\right)-\Pi_q^X\left(\beta\right)\Phi\left(\mathbb{Z}_0,x_0+X'\beta\right)\right\|\leq\mathscr{R}_q^X$$

according to Condition 5(ii), and that the following

$$\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_iX_i\Phi_q\left(\widehat{\mathbb{Z}}_i,x_{0,i}+X_i'\beta\right)'\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)\Phi_q\left(\widehat{\mathbb{Z}},x_0+X'\beta\right)-\right.$$
$$\left.\frac{1}{nP_D}\sum_{i=1}^{n}D_iX_i\Phi_q\left(\widehat{\mathbb{Z}}_i,x_{0,i}+X_i'\beta\right)'\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)\Phi_q\left(\widehat{\mathbb{Z}},x_0+X'\beta\right)\right\|$$

is uniformly of order $O_p(n^{-1/2}p_X^{-1/2}q^2C_{\Phi,q}^2)$, so we only need to bound the following

$$\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_iX_i\Phi_q\left(\widehat{\mathbb{Z}}_i,x_{0,i}+X_i'\beta\right)'\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)\Phi_q\left(\widehat{\mathbb{Z}},x_0+X'\beta\right)\right.$$
$$\left.-E\left(D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\Gamma_q^{-1}\left(\beta\right)\Phi_q\left(\mathbb{Z}_0,x_0+X'\beta\right)\right)\right\|.$$

Note that the above is bounded by

$$\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_iX_i\left(\Phi_q\left(\widehat{\mathbb{Z}}_i,x_{0,i}+X_i'\beta\right)-\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)\right)'\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)\Phi_q\left(\widehat{\mathbb{Z}},x_0+X'\beta\right)\right\|\ (i)$$
$$+\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\left(\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)-\Gamma_q^{-1}\left(\beta\right)\right)\Phi_q\left(\widehat{\mathbb{Z}},x_0+X'\beta\right)\right\|\ (ii)$$
$$+\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\Gamma_q^{-1}\left(\beta\right)\left(\Phi_q\left(\widehat{\mathbb{Z}},x_0+X'\beta\right)-\Phi_q\left(\mathbb{Z}_0,x_0+X'\beta\right)\right)\right\|\ (iii)$$
$$+\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\Gamma_q^{-1}\left(\beta\right)\Phi_q\left(\mathbb{Z}_0,x_0+X'\beta\right)\right.$$
$$\left.-E\left(D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\right)\Gamma_q^{-1}\left(\beta\right)\Phi_q\left(\mathbb{Z}_0,x_0+X'\beta\right)\right\|\ (iv)$$

According to the proof of Lemma 1, we know that $\sup_{\beta\in\mathscr{B}}\overline{\lambda}\left(\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)\right)\leq 1.5\underline{\lambda}_{\Phi}^{-1}$ with probability going to 1. So (i) and (iii) are both of order $O_p\left(n^{-1/2}p_Zp_X^{1/2}q^2C_{\Phi,q}C_{\Phi,1,q}\right)$. Term (ii) is bounded by $p_X^{1/2}q^2C_{\Phi,q}^2\sup_{\beta\in\mathscr{B}}\left\|\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta\right)-\Gamma_q^{-1}\left(\beta\right)\right\|$, which according to the proof of Lemma 1, is of order $O_p\left(n^{-1/2}p_Zp_X^{1/2}q^4C_{\Phi,q}^3C_{\Phi,u,q}+n^{-1/2}p_X^{1/2}q^2C_{\Phi,q}^4+\left(\log\left(n\right)/n\right)^{1/2}p_Xq^4C_{\Phi,q}^3\right)$. For the last term (iv), we know that each argument of $D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'$ is bounded by $C_{\Phi,q}$ and each argument of $D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta_1\right)'-D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta_2\right)'$ is bounded

by $C_{\Phi,1,q}p_X^{1/2}\left\|\beta_1-\beta_2\right\|$. Using Lemma A1 of Khan et al (2024), we have that

$$\begin{aligned}(iv) &\lesssim qC_{\Phi,q}\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'-E\left(D_iX_i\Phi_q\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)'\right)\right\| \\ &= O_p\left(\left(\log\left(n\right)/n\right)^{1/2}p_Xq^2C_{\Phi,q}^2\right).\end{aligned}$$

This proves the results. □

Lemma 3. *Let Conditions 1-5 hold and* $\Xi_{1,n}\to 0$*, then*

$$\sup_{k\geq 1}\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_i\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(\Phi_q\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k\right)'\Pi_q-G\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k\right)\right)\right\|=O_p\left(p_X^{1/2}\mathscr{R}_q\right),$$

$$\sup_{k\geq 1}\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_i\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)\right)\right\|=O_p\left(n^{-1/2}p_Zp_X^{1/2}\right),$$

and

$$\begin{aligned}&\sup_{k\geq 1}\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_i\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right. \\ &\quad\left.-P_D^{-1}\int_0^1E\left(D_i\nabla_vG\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}+\tau X_i'\Delta\beta^k\right)\left(X_i-P_D^{-1}X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta^k\right)\right)X_i'\right)\Delta\beta^k d\tau\right\|=O_p\left(\Xi_{1,n}\right).\end{aligned}$$

Proof. The first two results are obvious if we note that

$$\sup_{\beta\in\mathscr{B}}\left\|\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,x_0+X'\beta,\beta\right)-X^E\left(\mathbb{Z}_0,x_0+X'\beta,\beta\right)\right\|=o_p\left(1\right)$$

and $X^E\left(\mathbb{Z}_0,x_0+X'\beta,\beta\right)$ uniformly bounded. So we will only look at the third term. Note that

$$\begin{aligned}&\sup_{k\geq 1}\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_i\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right. \\ &\quad\left.-\frac{1}{S_n}\sum_{i=1}^{n}D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right\|=O_p\left(\Xi_{1,n}\right).\end{aligned}$$

Then we will bound

$$\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_i\left(X_i-P_D^{-1}X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta,\beta\right)\right)\left(G\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right.$$
$$\left.-P_D^{-1}E\left(D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right)\right\|.$$

Since $n/S_n - P_D^{-1} = O_p(n^{-1/2})$, we only need to look at the following distance

$$\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta,\beta\right)\right)\left(G\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right.$$
$$\left.-E\left(D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right)\right\|.$$

Note that each argument of $D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta,\beta\right)\right)\left(G\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)$ is upper bounded, and moreover, the gradient of each argument with respect to β is bounded by $p_X^{1/2}$ up to some scale in norm. So using Khan et al (2024)'s Lemma A1, we have that

$$\begin{aligned}&\sup_{\beta\in\mathscr{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta,\beta\right)\right)\left(G\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right.\\&\quad\left.-E\left(D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right)\right\|\\&=O_p\left(p_X\left(\left(\log n\right)/n\right)^{1/2}\right).\end{aligned}$$

We finish the proof by noting that, using Fubini's theorem, we have

$$\begin{aligned}&E\left(D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right)\\&=E\left(D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta^k,\beta^k\right)\right)\int_0^1\nabla G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}+X_i'\Delta\beta^k\right)X_i'\Delta\beta^k d\tau\right)\\&=\int_0^1 E\left(D_i\nabla G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}+X_i'\Delta\beta^k\right)\left(X_i-P_D^{-1}X^E\left(\mathbb{Z}_{0,i},x_{0,i}+X_i'\beta^k,\beta^k\right)\right)X_i'\Delta\beta^k\right)d\tau.\end{aligned}$$

□

Lemma 4. *Let Conditions 1 and 3(i) hold, then* $\|S_n^{-1}\sum_{i=1}^n D_iX_i\varepsilon_i\| = O_p\left(n^{-1/2}p_X^{1/2}\right)$.

Proof. Note that $E\left(\varepsilon_i|\,X_i,Z_i,D_i=1\right)=0$ and $E\left\|n^{-1}\sum_{i=1}^n D_iX_i\varepsilon_i\right\|^2=\sum_{j=1}^{p_X}E\left(\frac{1}{n}\sum_{i=1}^n D_ix_{ij}\varepsilon_i\right)^2\leq p_X/n$, which proves the result. □

Lemma 5. *Let Conditions 1-6 hold, we have that*

$$\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{S_n}\sum_{i=1}^{n}\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)\right)-\Sigma_{X,Z}\Delta\widehat{\delta}\right\|=O_p\left(n^{-1/2}p_Zp_X\Xi_{1,n}\right.$$

Proof. Note that

$$\begin{aligned}
&\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{S_n}\sum_{i=1}^{n}\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)\right)-\Sigma_{X,Z}\Delta\widehat{\delta}\right\| \\
&\lesssim\left|\frac{n}{S_n}-P_D^{-1}\right|\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)\right)\right\|(i) \\
&+\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\left(G\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)\right)\right\|(ii) \\
&+\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(X_i-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\left(\nabla_u G\left(\widetilde{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-\nabla_u G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)Z_i'\Delta\widehat{\delta}\right\|(iii) \\
&+\left\|\frac{1}{n}\sum_{i=1}^{n}\left(X_i-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\nabla_u G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)Z_i'-E\left(\left(X_i-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\nabla_u G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)Z_i'\right.\right.
\end{aligned}$$

Obviously, $(i)\lesssim\left|n/S_n-P_D^{-1}\right|p_X^{1/2}\sup_{k\geq k(n,\gamma),1\leq i\leq n}\left|\widehat{\mathbb{Z}}_i-\mathbb{Z}_{0,i}\right|=O_p\left(n^{-1}p_Z p_X^{1/2}\right)$. For (ii), we have

$$(ii)\leq\sup_{k\geq k(n,\gamma),1\leq i\leq n}\left\|\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right\|\sup_{1\leq i\leq n}\left|\widehat{\mathbb{Z}}_i-\mathbb{Z}_{0,i}\right|.$$

Obviously,

$$\begin{aligned}
&\sup_{k\geq k(n,\gamma),1\leq i\leq n}\left\|\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right\| \\
&\leq\sup_{k\geq k(n,\gamma),1\leq i\leq n}\left\|\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k,\beta^k\right)\right\| \\
&+\sup_{k\geq k(n,\gamma),1\leq i\leq n}\left\|X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k,\beta^k\right)-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta^k\right)\right\| \\
&+\sup_{k\geq k(n,\gamma),1\leq i\leq n}\left\|X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta^k\right)-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k,\beta_0\right)\right\|=O_p\left(p_X\Xi_{1,n}\right).
\end{aligned}$$

So (ii) is of order $O_p\left(n^{-1/2}p_Z p_X\Xi_{1,n}\right)$. (iii), is bounded by $p_Z^{1/2}p_X^{1/2}\max_{1\leq i\leq n}\left|\nabla_u G\left(\widetilde{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-\nabla_u G\left(\mathbb{Z}_{0,i}\right.\right.$

Since

$$\begin{aligned}
&\sup_{k\geq k(n,\gamma),1\leq i\leq n}\left|\nabla_u G\left(\widetilde{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-\nabla_u G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right| \\
&\leq \sup_{k\geq k(n,\gamma),1\leq i\leq n}\left|\nabla_u G\left(\widetilde{\mathbb{Z}}_i,\mathbb{X}_i^k\right)-\nabla_u G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)\right| \\
&+ \sup_{k\geq k(n,\gamma),1\leq i\leq n}\left|\nabla_u G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)-\nabla_u G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right| \\
&\lesssim \left|\widetilde{\mathbb{Z}}_i-\mathbb{Z}_{0,i}\right|+\left|\mathbb{X}_i^k-\mathbb{X}_{0,i}\right| = O_p\left(p_X^{1/2}\Xi_{1,n}\right).
\end{aligned}$$

This implies that (iii) is of order $O_p\left(n^{-1/2}p_Z p_X \Xi_{1,n}\right)$. Finally, the last term (iv) is obviously of order $O_p\left(n^{-1}p_Z p_X^{1/2}\right)$. Together we have shown the result. □

Lemma 6. *Let Conditions 1-6 hold, we have that*

$$\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{S_n}\sum_{i=1}^n\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)-\Psi\left(\beta_0\right)\Delta\beta^k\right\| = O_p\left(p_X^{3/2}\Xi_{1,n}^2\right).$$

Proof. Note that similar to the proof of the above lemma, we have that

$$\begin{aligned}
&\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{S_n}\sum_{i=1}^n\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)-\Psi\left(\beta_0\right)\Delta\beta^k\right\| \\
&\lesssim \left|\frac{n}{S_n}-P_D^{-1}\right|\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{n}\sum_{i=1}^n\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\left(G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right\| (i) \\
&+\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{n}\sum_{i=1}^n\left(\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\left(G\left(\mathbb{Z}_{0,i},\mathbb{X}_i^k\right)-G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)\right\| (ii) \\
&+\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{n}\sum_{i=1}^n\left(X_i-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\left(\nabla_v G\left(\mathbb{Z}_{0,i},\widetilde{\mathbb{X}}_i\right)-\nabla_v G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\right)X_i'\Delta\beta^k\right\| (iii) \\
&+\left\|\frac{1}{n}\sum_{i=1}^n\left(X_i-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\nabla_v G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)X_i'-E\left(\left(X_i-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\nabla_v G\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)X_i'\right.\right.
\end{aligned}$$

Obviously, term (i) is of order $O_p\left(n^{-1/2}p_X\Xi_{1,n}\right)$, term (ii) is of order $O_p\left(p_X^{1/2}\Xi_{1,n}^2\right)$, term (iii) is of order $O_p\left(p_X^{3/2}\Xi_{1,n}^2\right)$, and term (iv) is of order $O_p\left(n^{-1/2}p_X\Xi_{1,n}\right)$. This proves the result. □

Lemma 7. *Let Conditions 1-6 hold, we have that*

$$\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_i\left(X_i-\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)\right)\varepsilon_i-\frac{1}{nP_D}\sum_{i=1}^{n}D_i\left(X_i-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\varepsilon_i\right\|$$
$$=O_p\left(p_Xq^3C_{\Phi,q}C_{\Phi,2,q}\Xi_{1,n}^2+p_X^2q^6C_{\Phi,q}^2C_{\Phi,2,q}^2\Xi_{1,n}^4\right).$$

Proof. We first show the results for

$$\sup_{k\geq k(n,\gamma)}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right)-X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\varepsilon_i\right\|$$

Using Taylor expansion, we have that

$$\Phi_q\left(\widehat{\mathbb{Z}}_j,\mathbb{X}_j^k\right)=\Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)+\underbrace{\nabla_u\Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)Z_j'\Delta\widehat{\delta}}_{\Phi(1,j,n)}+\underbrace{\nabla_v\Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)X_j'\Delta\beta^k}_{\Phi(2,j,n,k)}+\mathcal{E}_{\Phi,q,j}^k,$$

where $\sup_{j,k}\left\|\Theta_{\Phi,q,j}^k\right\|=O_p\left(p_Xq C_{\Phi,2,q}\Xi_{1,n}^2\right)$, and that

$$\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta^k\right)=\Gamma_q\left(\beta_0\right)+\underbrace{\Gamma_q\left(\beta_0\right)-S_n^{-1}\sum_{i=1}^{n}D_i\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)'}_{\Gamma(1,n)}$$
$$+\underbrace{S_n^{-1}\sum_{i=1}^{n}D_i\nabla_u\left[\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)'\right]Z_j'\Delta\widehat{\delta}}_{\Gamma(2,n)}$$
$$+\underbrace{S_n^{-1}\sum_{i=1}^{n}D_i\nabla_v\left[\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)'\right]X_j'\Delta\beta^k}_{\Gamma(3,n,k)}+\Theta_{\Gamma,q}^k,$$

where $\left\|\Gamma\left(1,n\right)\right\|=O_p\left(n^{-1/2}qC_{\Phi,q}^2\right)$, $\left\|\Gamma\left(2,n\right)\right\|=O_p\left(n^{-1/2}p_Zq^2C_{\Phi,1,q}^2\right)$, $\sup_k\left\|\Gamma\left(3,n,k\right)\right\|=O_p\left(p_X^{1/2}q^2C_{\Phi,1,q}^2\Xi_{1,n}\right)$, and $\sup_k\left\|\Theta_{\Gamma,q}^k\right\|=O_p\left(p_Xq^2C_{\Phi,2,q}\Xi_{1,n}^2\right)$. Since $\Xi_{2,n}\to 0$, all of the above terms are of $o_p\left(1\right)$. Then note that

$$\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta^k\right)-\Gamma_q^{-1}\left(\beta_0\right)=\Gamma_q^{-1}\left(\beta_0\right)\left(\Gamma_q\left(\beta_0\right)-\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta^k\right)\right)\left(\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta^k\right)-\Gamma_q^{-1}\left(\beta_0\right)\right)$$
$$+\Gamma_q^{-1}\left(\beta_0\right)\left(\Gamma_q\left(\beta_0\right)-\widehat{\Gamma}_{n,q}\left(\widehat{\delta},\beta^k\right)\right)\Gamma_q^{-1}\left(\beta_0\right).$$

Define

$$\mathcal{E}_{\Gamma,q,n}=\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta},\beta^k\right)-\Gamma_q^{-1}\left(\beta_0\right)-\Gamma_q^{-1}\left(\beta_0\right)\left(\Gamma\left(1,n\right)+\Gamma\left(2,n\right)+\Gamma\left(3,n,k\right)\right)\Gamma_q^{-1}\left(\beta_0\right)$$

we have $\sup_k \|\mathcal{E}_{\Gamma,q,n}\| = O_p\left(p_X q^4 C_{\Phi,1,q}^4 \Xi_{1,n}^2 + p_X^2 q^4 C_{\Phi,2,q}^2 \Xi_{1,n}^4\right)$.

Using the expansion of $\Phi_q\left(\widehat{\mathbb{Z}}_j, \mathbb{X}_j^k\right)$ and $\widehat{\Gamma}_{n,q}^{-1}\left(\widehat{\delta}, \beta^k\right)$, we have that

$$\begin{aligned}
\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i, \mathbb{X}_i^k, \beta^k\right) &= S_n^{-1}\sum_{j=1}^n D_j X_j \left(\Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right) + \Phi\left(1,j,n\right) + \Phi\left(2,j,n,k\right) + \mathcal{E}_{\Phi,n,j}\right) \\
&\times \left(\widehat{\Gamma}_{n,q}^{-1}\left(\delta_0,\beta_0\right) + \Gamma_q^{-1}\left(\beta_0\right)\left(\Gamma\left(1,n\right) + \Gamma\left(2,n\right) + \Gamma\left(3,n,k\right)\right)\Gamma_q^{-1}\left(\beta_0\right) + \mathcal{E}_{\Gamma,n}\right) \\
&\times \left(\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right) + \Phi\left(1,i,n\right) + \Phi\left(2,i,n,k\right) + \mathcal{E}_{\Phi,n,i}\right)' \\
&= S_n^{-1}\sum_{j=1}^n D_j X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)\Gamma_q^{-1}\left(\beta_0\right)\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right) \\
&+ S_n^{-1}\sum_{j=1}^n D_j X_j \left(\Phi\left(1,j,n\right) + \Phi\left(2,j,n,k\right)\right)\Gamma_q^{-1}\left(\beta_0\right)\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right) \\
&+ S_n^{-1}\sum_{j=1}^n D_j X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)\left(\Gamma_q^{-1}\left(\beta_0\right)\left(\Gamma\left(1,n\right) + \Gamma\left(2,n\right) + \Gamma\left(3,n,k\right)\right)\right)\Gamma_q^{-1}\left(\beta_0\right)\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right) \\
&+ S_n^{-1}\sum_{j=1}^n D_j X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)\Gamma_q^{-1}\left(\beta_0\right)\left(\Phi\left(1,i,n\right) + \Phi\left(2,i,n,k\right)\right) + \mathcal{E}_{X,n,i,k},
\end{aligned}$$

where

$$\sup_{k,i}\|\mathcal{E}_{X,n,i,k}\| = O_p\left(p_X q^3 C_{\Phi,q} C_{\Phi,2,q}\Xi_{1,n}^2 + p_X^2 q^6 C_{\Phi,q}^2 C_{\Phi,2,q}^2 \Xi_{1,n}^4\right).$$

Also note that

$$\begin{aligned}
&S_n^{-1}\sum_{j=1}^n D_j X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)'\Gamma_q^{-1}\left(\beta_0\right)\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right) - X^E\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}, \beta_0\right) \\
&= \left(S_n^{-1}\sum_{j=1}^n D_j X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)'\Gamma_q^{-1}\left(\beta_0\right) - \Pi_q\right)\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right) + \left(X^E\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}, \beta_0\right) - \Pi_q\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right)\right) \\
&= \left(S_n^{-1}\sum_{j=1}^n D_j X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)' - \frac{1}{P_D}E\left(D_j X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)'\right)\right)\Gamma_q^{-1}\left(\beta_0\right)\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right) \\
&+ \left(E\left(X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)' \middle| D_j = 1\right)\Gamma_q^{-1}\left(\beta_0\right) - \Pi_q\right)\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right) \\
&+ \left(X^E\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}, \beta_0\right) - \Pi_q\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right)\right)
\end{aligned}$$

Obviously,

$$\left\|\left(E\left(X_j \Phi_q\left(\mathbb{Z}_{0,j}, \mathbb{X}_{0,j}\right)' \middle| D_j = 1\right)\Gamma_q^{-1}\left(\beta_0\right) - \Pi_q\right)\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right)\right\| \leq q C_{\Phi,q}\mathscr{R}_q^X,$$

$$\left\|X^E\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}, \beta_0\right) - \Pi_q\Phi_q\left(\mathbb{Z}_{0,i}, \mathbb{X}_{0,i}\right)\right\| \leq \mathscr{R}_q^X.$$

Moreover,

$$S_n^{-1}\sum_{j=1}^{n} D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)' - \frac{1}{P_D} E\left(D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)'\right)$$
$$= \left(\frac{1}{S_n/n} - \frac{1}{P_D}\right)\frac{1}{n}\sum_{j=1}^{n} D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)' + \frac{1}{P_D}\left(\frac{1}{n}\sum_{j=1}^{n} D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)' - E\left(D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right.\right.\right.$$

Then

$$\left\|\left(\frac{1}{S_n/n} - \frac{1}{P_D}\right)\frac{1}{n}\sum_{j=1}^{n} D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)' \Gamma_q^{-1}\left(\beta_0\right)\frac{1}{n}\sum_{i=1}^{n}\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\varepsilon_i\right\| = O_p\left(n^{-1}p_X^{1/2}q^2 C_{\Phi,q}^2\right)$$

and

$$\left\|\frac{1}{P_D}\left(\frac{1}{n}\sum_{j=1}^{n} D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)' - E\left(D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)'\right)\right)\Gamma_q^{-1}\left(\beta_0\right)\frac{1}{n}\sum_{i=1}^{n}\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\varepsilon_i\right\|$$
$$= O_p\left(n^{-1}p_X^{1/2}q^2 C_{\Phi,q}^2\right).$$

Now we are ready to derive the result. Note that

$$\left\|\frac{1}{S_n}\sum_{i=1}^{n} D_i\left(\widehat{X}_n^E\left(\widehat{\mathbb{Z}}_i,\mathbb{X}_i^k,\beta^k\right) - X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\varepsilon_i\right\|$$
$$\leq \left\|\frac{1}{S_n}\sum_{i=1}^{n} D_i\left(\frac{1}{S_n}\sum_{i=1}^{n} D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)'\Gamma_q^{-1}\left(\beta_0\right)\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right) - X^E\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i},\beta_0\right)\right)\varepsilon_i\right\|$$
$$+ \left\|\frac{1}{S_n}\sum_{i=1}^{n} D_j X_j\left(\Phi\left(1,j,n\right) + \Phi\left(2,j,n\right)\right)\Gamma_q^{-1}\left(\beta_0\right)\frac{1}{S_n}\sum_{i=1}^{n} D_i\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\varepsilon_i\right\|$$
$$+ \left\|\frac{1}{S_n}\sum_{i=1}^{n} D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)\left(\Gamma_q^{-1}\left(\beta_0\right)\left(\Gamma\left(1,n\right) + \Gamma\left(2,n,k\right)\right)\right)\Gamma_q^{-1}\left(\beta_0\right)\frac{1}{S_n}\sum_{i=1}^{n} D_i\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\varepsilon_i\right\|$$
$$+ \left\|\frac{1}{S_n}\sum_{j=1}^{n} D_j X_j \Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)\Gamma_q^{-1}\left(\beta_0\right)\frac{1}{S_n}\sum_{i=1}^{n} D_i\left(\Phi\left(1,i,n\right) + \Phi\left(2,i,n\right)\right)\varepsilon_i\right\|^2 + E\left\|\frac{1}{S_n}\sum_{i=1}^{n} D_i\mathcal{E}_{X,n,i,k}\varepsilon_i\right\|$$

The first term is obviously $O_p\left(n^{-1}p_X^{1/2}q^2C_{\Phi,q}^2 + qC_{\Phi,q}\mathscr{R}_q^X\right)$. We will look at the remaining

terms one by one. First of all,

$$E\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_jX_j\left(\Phi\left(1,j,n\right)+\Phi\left(2,j,n\right)\right)\Gamma_q^{-1}\left(\beta_0\right)\frac{1}{S_n}\sum_{i=1}^{n}D_i\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\varepsilon_i\right\|^2$$
$$\lesssim E\left[\frac{\sum_{i=1}^{n}D_iq^2C_{\Phi,q}^2}{S_n^2}\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_jX_j\left(\Phi\left(1,j,n\right)+\Phi\left(2,j,n\right)\right)\right\|^2\right]\lesssim E\left[\frac{\left(\sum_{i=1}^{n}D_i\right)p_Xq^4C_{\Phi,q}^2\left(p_Z^2n^{-1}+p_X\right.\right.}{S_n^2}$$
$$=O\left(n^{-1}p_Xq^4C_{\Phi,q}^2\left(p_Z^2n^{-1}+p_X\Xi_{1,n}^2\right)\right),$$

where the last result comes from the fact that $\sum_{i=1}^{n}D_i/S_n^2$ is bounded by 1 and $S_n/n\to P_D>0$. Similarly,

$$E\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_jX_j\Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)\left(\Gamma_q^{-1}\left(\beta_0\right)\left(\Gamma\left(1,n\right)+\Gamma\left(2,n\right)+\Gamma\left(3,n\right)\right)\right)\Gamma_q^{-1}\left(\beta_0\right)\frac{1}{S_n}\sum_{i=1}^{n}D_i\Phi_q\left(\mathbb{Z}_{0,i},\mathbb{X}_{0,i}\right)\varepsilon\right.$$
$$=O\left(\frac{\sum_{i=1}^{n}D_iq^2C_{\Phi,q}^2C_{\Phi,1,q}^4p_X^2q^6\Xi_{1,n}^2}{S_n^2}\right),$$

$$E\left\|\frac{1}{S_n}\sum_{j=1}^{n}D_jX_j\Phi_q\left(\mathbb{Z}_{0,j},\mathbb{X}_{0,j}\right)\Gamma_q^{-1}\left(\beta_0\right)\frac{1}{S_n}\sum_{i=1}^{n}D_i\left(\Phi\left(1,i,n\right)+\Phi\left(2,i,n\right)\right)\varepsilon_i\right\|^2$$
$$=O_p\left(n^{-1}p_Xq^4C_{\Phi,q}^2\left(p_Z^2n^{-1}+p_X\Xi_{1,n}^2\right)\right).$$

Finally,

$$E\left\|\frac{1}{S_n}\sum_{i=1}^{n}D_i\mathcal{E}_{X,n,i,k}\varepsilon_i\right\|^2\leq\sup_{k,i}\left\|\mathcal{E}_{X,n,i,k}\right\|^2=O_p\left(p_X^2q^6C_{\Phi,q}^2C_{\Phi,2,q}^2\Xi_{1,n}^4+p_X^4q^{12}C_{\Phi,q}^4C_{\Phi,2,q}^4\Xi_{1,n}^8\right).$$

□