# Kernel regression utilizing heterogeneous datasets 

Chi-Shian Dai \& Jun Shao

To cite this article: Chi-Shian Dai \& Jun Shao (2023): Kernel regression utilizing heterogeneous datasets, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2023.2202579

To link to this article: https://doi.org/10.1080/24754269.2023.2202579

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor \& Francis Group

Published online: 28 Apr 2023.


Submit your article to this journal

```\(\square\)
```



Article views: 124


View related articles


View Crossmark data $\checkmark$

# Kernel regression utilizing heterogeneous datasets 

Chi-Shian Dai ${ }^{\text {a }}$ and Jun Shao ${ }^{\text {a,b }}$<br>${ }^{\text {a }}$ Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA; ${ }^{\text {b }}$ School of Statistics, East China Normal University, Shanghai, People's Republic of China


#### Abstract

Data analysis in modern scientific research and practice has shifted from analysing a single dataset to coupling several datasets. We propose and study a kernel regression method that can handle the challenge of heterogeneous populations. It greatly extends the constrained kernel regression [Dai, C.-S., \& Shao, J. (2023). Kernel regression utilizing external information as constraints. Statistica Sinica, 33, in press] that requires a homogeneous population of different datasets. The asymptotic normality of proposed estimators is established under some conditions and simulation results are presented to confirm our theory and to quantify the improvements from datasets with heterogeneous populations.


## ARTICLE HISTORY

Received 5 December 2022
Revised 5 April 2023
Accepted 8 April 2023

## KEYWORDS

Conditional expectation; constraints; data coupling and integration; external data; heterogeneous populations; kernel estimation

## 1. Introduction

With advanced technologies in data collection and storage, in modern statistical analyses we have not only a primary random sample from a population of interest, which results in a dataset referred to as the internal dataset, but also some independent external datasets from sources such as past investigations and publicly available datasets. In this paper, we consider nonparametric kernel regression (Bierens, 1987; Wand \& Jones, 1994, December; Wasserman, 2006) between a univariate response $Y$ and a covariate vector $\boldsymbol{U}$ from a sampled subject, using the internal dataset with the help from independent external datasets. Specifically, we consider kernel estimation of the conditional expectation (regression function) of $Y$ given $\boldsymbol{U}=\boldsymbol{u}$ under an internal data population,

$$
\begin{equation*}
\mu_{1}(\boldsymbol{u})=E(Y \mid \boldsymbol{U}=\boldsymbol{u}, D=1) \tag{1}
\end{equation*}
$$

where $D=1$ indicates internal population and $\boldsymbol{u}$ is a fixed point in $\mathscr{U}$, the range of $\boldsymbol{U}$. The indicator $D$ can be either random or deterministic. The subscript 1 in $\mu_{1}(\boldsymbol{u})$ emphasizes that it is for internal data population $(D=1)$, which may be different from $\mu(\boldsymbol{u})=E(Y \mid \boldsymbol{U}=\boldsymbol{u})$, a mixture of quantities from the internal and external data populations.

When external datasets also have measurements $Y$ and $\boldsymbol{U}$, we may simply combine the internal and external datasets when the populations for internal and external data are identical (homogeneous). However, heterogeneity typically exists among populations for different datasets, especially when there are multiple external datasets collected in different ways and/or different time periods. In Section 2, we propose a method to handle heterogeneity among different populations and derive a kernel regression more efficient than the one using internal data alone. The result is also a crucial building block for the more complicated case in Section 3 where external datasets contain fewer measured covariates as described next.

In applications, it often occurs that an external dataset has measured $Y$ and $\boldsymbol{X}$ from each subject, where $\boldsymbol{X}$ is a part of the vector $\boldsymbol{U}$, i.e., some components of $\boldsymbol{U}$ are not measured due to high measurement cost or the progress of technology and/or scientific relevance. With some unmeasured components of $\boldsymbol{U}$, the external dataset cannot be directly used to estimate $\mu_{1}(\boldsymbol{u})$ in (1), since conditioning on the entire $\boldsymbol{U}$ is involved. To solve this problem, Dai and Shao (2023) proposes a two-step kernel regression using external information as a constraint to improve kernel regression based on internal data alone, following the idea of using constraints in Chatterjee et al. (2016) and H. Zhang et al. (2020). However, these three cited papers mainly assume that the internal and external datasets share the same population, which may be unrealistic. The challenge in dealing with the heterogeneity among different populations is similar to the difficulty in handling nonignorable missing data if unmeasured components of $\boldsymbol{U}$ is treated as missing data, although in missing data problems we usually want to estimate $\mu(\boldsymbol{u})=E(Y \mid \boldsymbol{U}=\boldsymbol{u}) \neq$ $\mu_{1}(\boldsymbol{u})$ in (1).

In Section 3, we develop a methodology to handle population heterogeneity for internal and external datasets, which extends the procedure in Dai and Shao (2023) to heterogeneous populations and greatly widens its application scope.

Under each scenario, we derive asymptotic normality in Section 4 for the proposed kernel estimators and obtain explicitly the asymptotic variances, which is important for large sample inference. Some simulation results are presented in Section 5 to compare finite sample performance of several estimators. Discussions on extensions and handling high dimension covariates are given in Section 6. All technical details are in the Appendix.

Our research fits into a general framework of data integration (Kim et al., 2021; Lohr \& Raghunathan, 2017; Merkouris, 2004; Rao, 2021; Yang \& Kim, 2020; Y. Zhang et al., 2017).

## 2. Efficient kernel estimation by combining datasets

The internal dataset contains observations $\left(Y_{i}, \boldsymbol{U}_{i}\right), i=1, \ldots, n$, independent and identically distributed (iid) from $\mathcal{P}_{1}$, the internal population of $(Y, \boldsymbol{U})$, where $Y$ is the response and $\boldsymbol{U}$ is a $p$-dimensional covariate vector associated with $Y$. We are interested in the estimation of conditional expectation $\mu_{1}(\boldsymbol{u})$ in (1). The standard kernel regression estimator of $\mu_{1}(\boldsymbol{u})$ based on the internal dataset alone is

$$
\begin{equation*}
\widehat{\mu}_{1}(\boldsymbol{u})=\sum_{i=1}^{n} Y_{i} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) / \sum_{i=1}^{n} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \tag{2}
\end{equation*}
$$

where $\kappa_{b}(\boldsymbol{a})=b^{-p_{\kappa}}(\boldsymbol{a} / b), \kappa(\cdot)$ is a given kernel function on $\mathscr{U}$ (the range of $\boldsymbol{u}$ ), and $b>0$ is a bandwidth depending on $n$. We assume that $\boldsymbol{U}$ is standardized so that the same bandwidth $b$ is used for every component of $\boldsymbol{U}$ in kernel regression. Because of the well-known curse of dimensionality for kernel-type methods, we focus on a low dimension $p$ not varying with $n$. A discussion of handling a large dimensional $\boldsymbol{U}$ is given in Section 6.

We consider the case with one external dataset, independent of the internal dataset. Extension to multiple external datasets is straightforward and discussed in Section 6.

In this section we consider the situation where the external dataset contains iid observations $\left(Y_{i}, \boldsymbol{U}_{i}\right), i=n+$ $1, \ldots, N$, from $\mathcal{P}_{0}$, the external population of $(Y, \boldsymbol{U})$.

### 2.1. Combing data from homogeneous populations

If we assume that the two populations $\mathcal{P}_{1}$ and $\mathcal{P}_{0}$ are identical, then we can simply combine two datasets to obtain the kernel estimator

$$
\begin{equation*}
\widehat{\mu}_{1}^{E 1}(\boldsymbol{u})=\sum_{i=1}^{N} Y_{i} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) / \sum_{i=1}^{N} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right), \tag{3}
\end{equation*}
$$

which is obviously more efficient than $\widehat{\mu}_{1}(\boldsymbol{u})$ in (2) as the sample size is increased to $N>n$. The estimator $\widehat{\mu}_{1}^{E 1}(\boldsymbol{u})$ in (3), however, is not correct (i.e., it is biased) when populations $\mathcal{P}_{1}$ and $\mathcal{P}_{0}$ are different, because $E(Y \mid \boldsymbol{U}=\boldsymbol{u}$, $D=0$ ) for external population may be different from $\mu_{1}(\boldsymbol{u})=E(Y \mid \boldsymbol{U}=\boldsymbol{u}, D=1)$ for internal population.

### 2.2. Combing data from heterogeneous populations

We now derive a kernel estimator using two datasets and is asymptotically correct regardless of whether $\mathcal{P}_{1}$ and $\mathcal{P}_{0}$ are the same or not. Let $f(y \mid \boldsymbol{u}, D)$ be the conditional density of $Y$ given $\boldsymbol{U}=\boldsymbol{u}$ and $D=1$ or 0 (for internal or external population). Then

$$
\begin{equation*}
\mu_{1}(\boldsymbol{x})=E(Y \mid \boldsymbol{U}=\boldsymbol{u}, D=1)=E\left\{\left.Y \frac{f(Y \mid \boldsymbol{u}, D=1)}{f(Y \mid \boldsymbol{u}, D=0)} \right\rvert\, \boldsymbol{U}=\boldsymbol{u}, D=0\right\} \tag{4}
\end{equation*}
$$

The ratio $f(Y \mid \boldsymbol{u}, D=1) / f(Y \mid \boldsymbol{u}, D=0)$ links internal and external populations so that we can overcome the difficulty in utilizing the external data under heterogeneous populations.

If we can construct an estimator $\widehat{f}(y \mid \boldsymbol{u}, D)$ of $f(y \mid \boldsymbol{u}, D)$ for every $y$, $\boldsymbol{u}$, and $D=0$ or 1 , then we can modify the estimator in (3) by replacing every $Y_{i}$ with $i>n$ by constructed response $\widehat{Y}_{i}=Y_{i} \widehat{f}\left(Y_{i} \mid \boldsymbol{U}_{i}, D=1\right) / \widehat{f}\left(Y_{i} \mid \boldsymbol{U}_{i}, D=0\right)$.

The resulting kernel estimator is

$$
\begin{equation*}
\widehat{\mu}_{1}^{E 2}(\boldsymbol{u})=\left\{\sum_{i=1}^{n} Y_{i} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)+\sum_{i=n+1}^{N} \widehat{Y}_{i} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)\right\} / \sum_{i=1}^{N} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) . \tag{5}
\end{equation*}
$$

Note that we use internal data $\left(Y_{i}, \boldsymbol{U}_{i}\right), i=1, \ldots, n$, to obtain estimator $\widehat{f}\left(Y_{i} \mid \boldsymbol{U}_{i}, D=1\right)$ and external data $\left(Y_{i}, \boldsymbol{U}_{i}\right)$, $i=n+1, \ldots, N$, to construct estimator $\widehat{f}\left(Y_{i} \mid \boldsymbol{U}_{i}, D=0\right)$. Applying kernel estimation, we obtain that

$$
\begin{gather*}
\widehat{f}(y \mid \boldsymbol{U}=\boldsymbol{u},=1)=\sum_{i=1}^{n} \tilde{\kappa}_{\tilde{b}}\left(y-Y_{i}, \boldsymbol{u}-\boldsymbol{U}_{i}\right) / \sum_{i=1}^{n} \bar{\kappa}_{\bar{b}}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right), \\
\widehat{f}(y \mid \boldsymbol{U}=\boldsymbol{u}, D=0)=\sum_{i=n+1}^{N} \tilde{\kappa}_{\tilde{b}}\left(y-Y_{i}, \boldsymbol{u}-\boldsymbol{U}_{i}\right) / \sum_{i=n+1}^{N} \bar{\kappa}_{\bar{b}}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right), \tag{6}
\end{gather*}
$$

where $\tilde{\kappa}$ and $\bar{\kappa}$ are kernels with dimensions $p+1$ and $p$ and bandwidths $\tilde{b}$ and $\bar{b}$, respectively. The estimator in (5) is asymptotically valid under some regularity conditions for kernel and bandwidth, summarized in Theorem 4.1 of Section 4.

### 2.3. Combing data from heterogeneous populations with additional information

If additional information exists, then the approach in Section 2.2 can be improved. Assume that the internal and external datasets are formed according to a random binary indicator $D$ such that $\left(Y_{i}, \boldsymbol{U}_{i}, D_{i}\right), i=1, \ldots, N$, are iid distributed as $(Y, \boldsymbol{U}, D)$, where $Y_{i}$ and $\boldsymbol{U}_{i}$ are observed internal data when $D_{i}=1, Y_{i}$ and $\boldsymbol{U}_{i}$ are observed external data when $D_{i}=0$, and $N$ is still the known total sample size for internal and external data. In this situation, the internal and external sample sizes are $n=\sum_{i=1}^{N} D_{i}$ and $N-n$, respectively, both of which are random. In most applications, the assumption of random $D$ is not substantial. From the identity

$$
\begin{equation*}
\frac{f(Y \mid \boldsymbol{u}, D=1)}{f(Y \mid \boldsymbol{u}, D=0)}=\frac{P(D=1 \mid \boldsymbol{U}=\boldsymbol{u}, Y)}{P(D=0 \mid \boldsymbol{U}=\boldsymbol{u}, Y)} \frac{P(D=0 \mid \boldsymbol{U}=\boldsymbol{u})}{P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})}, \tag{7}
\end{equation*}
$$

we just need to estimate $P(D=1 \mid \boldsymbol{U}=\boldsymbol{u}, Y)$ and $P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})$ for every $\boldsymbol{u}$, constructed using for example the nonparametric estimators in Fan et al. (1998) for binary response. For each estimator, both internal and external data on $(Y, \boldsymbol{U})$ and the indicator $D$ are used.

A further improvement can be made if the following semi-parametric model holds,

$$
\begin{equation*}
\frac{P(D=0 \mid \boldsymbol{U}, Y)}{P(D=1 \mid \boldsymbol{U}, Y)}=\exp \{\alpha(\boldsymbol{U})+\gamma Y\} \tag{8}
\end{equation*}
$$

where $\alpha(\cdot)$ is an unspecified unknown function and $\gamma$ is an unknown parameter. From (7)-(8),

$$
\begin{equation*}
\frac{f(Y \mid \boldsymbol{u}, D=1)}{f(Y \mid \boldsymbol{u}, D=0)}=\mathrm{e}^{-\gamma Y} E\left(\mathrm{e}^{\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right) \tag{9}
\end{equation*}
$$

If $\gamma=0$, then $f(Y \mid \boldsymbol{u}, D=1)=f(Y \mid \boldsymbol{u}, D=0)$ and the estimator $\widehat{\mu}_{1}^{E 1}(\boldsymbol{u})$ in (3) is correct. Under (9) with $\gamma \neq 0$, we just need to derive an estimator $\widehat{\gamma}$ of $\gamma$ and apply kernel estimation to estimate $E\left(\mathrm{e}^{\widehat{\gamma} Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right)$ as a function of $\boldsymbol{u}$. Note that we do not need to estimate the unspecified function $\alpha(\cdot)$ in (8), which is a nice feature of semi-parametric model (8).

We now derive an estimator $\widehat{\gamma}$. Applying (7)-(8) to (4), we obtain that

$$
\begin{aligned}
\mu_{1}(\boldsymbol{u}) & =E\left\{\left.Y \frac{P(D=1 \mid \boldsymbol{U}=\boldsymbol{u}, Y)}{P(D=0 \mid \boldsymbol{U}=\boldsymbol{u}, Y)} \right\rvert\, \boldsymbol{U}=\boldsymbol{u}, D=0\right\} \frac{P(D=0 \mid \boldsymbol{U}=\boldsymbol{u})}{P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})} \\
& =E\left(Y \mathrm{e}^{-\alpha(\boldsymbol{u})-\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right) \frac{E\{P(D=0 \mid \boldsymbol{U}=\boldsymbol{u}, Y) \mid \boldsymbol{U}=\boldsymbol{u}\}}{P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})}
\end{aligned}
$$

$$
\begin{aligned}
& =\mathrm{e}^{-\alpha(\boldsymbol{u})} E\left(Y \mathrm{e}^{-\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right) \frac{E\left\{\mathrm{e}^{\alpha(\boldsymbol{u})+\gamma Y} P(D=1 \mid \boldsymbol{U}=\boldsymbol{u}, Y) \mid \boldsymbol{U}=\boldsymbol{u}\right\}}{P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})} \\
& =E\left(Y \mathrm{e}^{-\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right) \frac{E\left\{\mathrm{e}^{\gamma Y} E(D \mid \boldsymbol{U}=\boldsymbol{u}, Y) \mid \boldsymbol{U}=\boldsymbol{u}\right\}}{P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})} \\
& =E\left(Y \mathrm{e}^{-\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right) \frac{E\left(\mathrm{e}^{\gamma Y} D \mid \boldsymbol{U}=\boldsymbol{u}\right)}{P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})} \\
& =E\left(Y \mathrm{e}^{-\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right) E\left(\mathrm{e}^{\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right)
\end{aligned}
$$

where the second and third equalities follow from (8) and the last equation follows from

$$
\begin{aligned}
E\left(\mathrm{e}^{\gamma Y} D \mid \boldsymbol{U}=\boldsymbol{u}\right)= & E\left(\mathrm{e}^{\gamma Y} D \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right) P(D=1 \mid \boldsymbol{U}=\boldsymbol{u}) \\
& +E\left(\mathrm{e}^{\gamma Y} D \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right) P(D=0 \mid \boldsymbol{U}=\boldsymbol{u}) \\
= & E\left(\mathrm{e}^{\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right) P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})
\end{aligned}
$$

as $E\left(\mathrm{e}^{\gamma Y} D \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right)=0$. For every real number $t$, define

$$
h(\boldsymbol{u}, t)=E\left(Y \mathrm{e}^{-t Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right) E\left(\mathrm{e}^{t Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right)
$$

Its estimator by kernel regression is

$$
\begin{equation*}
\widehat{h}(\boldsymbol{u}, t)=\frac{\sum_{i=1}^{N}\left(1-D_{i}\right) \check{\kappa}_{\check{b}}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) Y_{i} \mathrm{e}^{-t Y_{i}}}{\sum_{i=1}^{N}\left(1-D_{i}\right) \check{\kappa}_{\check{b}}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)} \frac{\sum_{i=1}^{N} D_{i} \check{\kappa}_{\check{b}}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \mathrm{e}^{t Y_{i}}}{\sum_{i=1}^{N} D_{i} \check{\kappa}_{\breve{b}}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)} \tag{10}
\end{equation*}
$$

where $\check{\kappa}$ is a kernel and $\check{b}$ is a bandwidth. Then, we estimate $\gamma$ by

$$
\begin{equation*}
\widehat{\gamma}=\arg \min _{t} \frac{1}{N} \sum_{i=1}^{N} D_{i}\left\{Y_{i}-\widehat{h}\left(\boldsymbol{U}_{i}, t\right)\right\}^{2} \tag{11}
\end{equation*}
$$

motivated by the fact that the objective function for minimization in (11) approximates $E\left[D\{Y-h(\boldsymbol{U}, t)\}^{2} \mid D=1\right]$ and, for any $t$,

$$
E\left[D\{Y-h(\boldsymbol{U}, \gamma)\}^{2} \mid D=1\right] \leq E\left[D\{Y-h(\boldsymbol{U}, t)\}^{2} \mid D=1\right]
$$

because $h(\boldsymbol{u}, \gamma)=\mu_{1}(\boldsymbol{u})$.
Once $\widehat{\gamma}$ is obtained, our estimator of $\mu_{1}(\boldsymbol{u})$ is

$$
\begin{equation*}
\widehat{\mu}_{1}^{E 3}(\boldsymbol{u})=\left\{\sum_{i=1}^{N} D_{i} Y_{i} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)+\sum_{i=1}^{N}\left(1-D_{i}\right) \widehat{Y}_{i} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)\right\} / \sum_{i=1}^{N} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \tag{12}
\end{equation*}
$$

with

$$
\widehat{Y}_{i}=Y_{i} \mathrm{e}^{-\widehat{\gamma} Y_{i}} \sum_{j=1}^{n} \mathrm{e}^{\widehat{\gamma} Y_{j}} \check{\kappa}_{\breve{b}}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right) / \sum_{j=1}^{n} \check{\kappa}_{\check{b}}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right),
$$

in view of (9).
In applications, we need to choose bandwidths with given sample sizes $n$ and $N-n$. We can apply the $k$-fold crossvalidation as described in Györfi et al. (2002). Requirements on the rates of bandwidths are described in theorems in Section 3.

## 3. Constrained kernel regression with unmeasured covariates

We still consider the case with one external dataset, independent of the internal dataset. In this section, the external dataset contains iid observations $\left(Y_{i}, \boldsymbol{X}_{i}\right), i=n+1, \ldots, N$, from the external population $\mathcal{P}_{0}$, where $\boldsymbol{X}$ is a $q$-dimensional sub-vector of $\boldsymbol{U}$ with $q<p$.

Since the external dataset has only $\boldsymbol{X}$, not the entire $\boldsymbol{U}$, we cannot apply the method in Section 2 when $q<p$. Instead, we consider kernel regression using external information in a constraint. First, we consider the estimation
of the $n$-dimensional vector $\boldsymbol{\mu}_{1}=\left(\mu_{1}\left(\boldsymbol{U}_{1}\right), \ldots, \mu_{1}\left(\boldsymbol{U}_{n}\right)\right)^{\top}$, where $\boldsymbol{A}^{\top}$ denotes the transpose of vector or matrix $\boldsymbol{A}$ throughout. Note that the standard kernel regression (2) estimates $\boldsymbol{\mu}_{1}$ as

$$
\widehat{\boldsymbol{\mu}}_{1}=\left(\sum_{i=1}^{n} Y_{i} \kappa_{b}\left(\boldsymbol{U}_{1}-\boldsymbol{U}_{i}\right) / \sum_{i=1}^{n} \kappa_{b}\left(\boldsymbol{U}_{1}-\boldsymbol{U}_{i}\right), \ldots, \sum_{i=1}^{n} Y_{i} \kappa_{b}\left(\boldsymbol{U}_{n}-\boldsymbol{U}_{i}\right) / \sum_{i=1}^{n} \kappa_{b}\left(\boldsymbol{U}_{n}-\boldsymbol{U}_{i}\right)\right)^{\top}
$$

Taking partial derivatives with respect to $\mu_{i}$ 's, we obtain that

$$
\begin{equation*}
\widehat{\boldsymbol{\mu}}_{1}=\arg \min _{\mu_{1}, \ldots, \mu_{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_{b}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right)\left(Y_{j}-\mu_{i}\right)^{2} / \sum_{k=1}^{n} \kappa_{b}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{k}\right) . \tag{13}
\end{equation*}
$$

We improve $\widehat{\boldsymbol{\mu}}_{1}$ by the following constrained minimization,

$$
\begin{align*}
\widehat{\boldsymbol{\mu}}_{1}^{C j}= & \arg \min _{\mu_{1}, \ldots, \mu_{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_{l}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right)\left(Y_{j}-\mu_{i}\right)^{2} / \sum_{k=1}^{n} \kappa_{l}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{k}\right)  \tag{14}\\
& \text { subject to } \sum_{i=1}^{n}\left\{\mu_{i}-\widehat{h}_{1}^{E j}\left(\boldsymbol{X}_{i}\right)\right\} \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)^{\top}=0, \tag{15}
\end{align*}
$$

where $\boldsymbol{g}(\boldsymbol{x})^{\top}=\left(1, \boldsymbol{x}^{\top}\right), l$ in (14) is a bandwidth that may be different from $b$ in (2) or (13), and $\widehat{h}_{1}^{E j}(\boldsymbol{x})$ is the kernel estimator of $h_{1}(\boldsymbol{x})=E(Y \mid \boldsymbol{X}=\boldsymbol{x}, D=1)$ using the $j$ th of the three methods described in Section $2, j=1,2,3$. Specifically, $\widehat{h}_{1}^{E 1}(\boldsymbol{x})$ is given by (3), $\widehat{h}_{1}^{E 2}(\boldsymbol{x})$ is given by (5), and $\widehat{h}_{1}^{E 3}(\boldsymbol{x})$ is given by (12), with $\boldsymbol{u}$ and $\boldsymbol{U}$ replaced by $\boldsymbol{x}$ and $\boldsymbol{X}$, respectively, and kernels and bandwidths suitably adjusted as dimensions of $\boldsymbol{U}$ and $\boldsymbol{X}$ are different. Note that $\widehat{h}_{1}^{E j}$ can be computed as both internal and external datasets have measured $\boldsymbol{X}_{i}$ 's.

It turns out that $\widehat{\boldsymbol{\mu}}_{1}^{C j}$ in (14) has an explicit form $\widehat{\boldsymbol{\mu}}_{1}^{C j}=\widehat{\boldsymbol{\mu}}_{1}+\boldsymbol{G}\left(\boldsymbol{G}^{\top} \boldsymbol{G}\right)^{-1} \boldsymbol{G}^{\top}\left(\widehat{\boldsymbol{h}}_{1}^{E j}-\widehat{\boldsymbol{\mu}}_{1}\right)$, where $\boldsymbol{G}$ is the $n \times n$ matrix whose $i$ th row is $\boldsymbol{g}\left(\boldsymbol{X}_{i}\right)^{\top}$ and $\widehat{\boldsymbol{h}}_{1}^{E j}$ is the $n$-dimensional vector whose $i$ th component is $\widehat{h}_{1}^{E j}\left(\boldsymbol{X}_{i}\right)$. Constraint (15) is an empirical analog of the theoretical constraint

$$
E\left[\left\{\mu_{1}(\boldsymbol{U})-h_{1}(\boldsymbol{X})\right\} \boldsymbol{g}(\boldsymbol{X})^{\top} \mid D=1\right]=0
$$

(based on internal data), as $E\{E(Y \mid \boldsymbol{U}, D=1) \mid \boldsymbol{X}, D=1\}=E(Y \mid \boldsymbol{X}, D=1)=h_{1}(\boldsymbol{X})$. Thus, if $\widehat{h}_{1}^{E j}(\cdot)$ is a good estimator of $h_{1}(\cdot)$, then $\widehat{\boldsymbol{\mu}}_{1}^{C j}$ in (14) is more accurate than the unconstrained $\widehat{\boldsymbol{\mu}}_{1}$ in (13).

To obtain an improved estimator of the entire regression function $\mu_{1}(\cdot)$ in (1), not just the function at $\boldsymbol{u}=\boldsymbol{U}_{i}$, $i=1, \ldots, n$, we apply the standard kernel regression with response vector $\left(Y_{1}, \ldots, Y_{n}\right)^{\top}$ replaced by $\widehat{\boldsymbol{\mu}}_{1}^{C j}$ in (14), which results in the following three estimators of $\mu_{1}(\boldsymbol{u})$ :

$$
\begin{equation*}
\widehat{\mu}_{1}^{C j}(\boldsymbol{u})=\sum_{i=1}^{n} \widehat{\mu}_{i}^{C j} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) / \sum_{i=1}^{n} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \quad j=1,2,3, \tag{16}
\end{equation*}
$$

where $\widehat{\mu}_{i}^{C j}$ is the $i$ th component of $\widehat{\boldsymbol{\mu}}_{1}^{C j}$ in (14) and $b$ is the same bandwidth in (2). The first estimator $\widehat{\mu}_{i}^{C 1}$ is simple, but can be incorrect when populations $\mathcal{P}_{1}$ and $\mathcal{P}_{0}$ are different. The asymptotic validity of $\widehat{\boldsymbol{\mu}}_{1}^{C 2}$ and $\widehat{\boldsymbol{\mu}}_{1}^{C 3}$ are established in the next section.

## 4. Asymptotic normality

We now establish the asymptotic normality of $\widehat{\mu}_{1}^{E j}(\boldsymbol{u})$ and $\widehat{\mu}_{1}^{C j}(\boldsymbol{u})$ for a fixed $\boldsymbol{u}$, as the sample size of the internal dataset increases to infinity. All technical proofs are given in the Appendix.

The first result is about $\widehat{\mu}_{1}^{E 2}(\boldsymbol{u})$ in (5). The result is also applicable to $\widehat{\mu}_{1}^{E 1}(\boldsymbol{u})$ in (3) with an added condition that $\mathcal{P}_{1}=\mathcal{P}_{0}$.

Theorem 4.1: Assume the following conditions.
(B1) The densities $f_{1}(\boldsymbol{u})$ and $f_{0}(\boldsymbol{u})$ for $\boldsymbol{U}$, respectively under internal and external populations have continuous and bounded first- and second-order partial derivatives.
(B2) $\mu_{1}^{2}(\boldsymbol{u}) f_{k}(\boldsymbol{u}), \sigma_{k}^{2}(\boldsymbol{u}) f_{k}(\boldsymbol{u})$, and the first- and second-order partial derivatives of $\mu_{1}(\boldsymbol{u}) f_{k}(\boldsymbol{u})$ are continuous and bounded, where $\sigma_{1}^{2}(\boldsymbol{u})=E\left[\left\{Y-\mu_{1}(\boldsymbol{U})\right\}^{2} \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right], \sigma_{0}^{2}(\boldsymbol{u})=E\left[\left\{\tilde{Y}-\mu_{1}(\boldsymbol{U})\right\}^{2} \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right]$, and
$\tilde{Y}=Y f(Y \mid \boldsymbol{U}, D=1) / f(Y \mid \boldsymbol{U}, D=0)$. Also, $E\left(|Y|^{s} \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right) f_{1}(\boldsymbol{u})$ and $E\left(|\tilde{Y}|^{s} \mid \boldsymbol{U}=\boldsymbol{u}, D=0\right) f_{0}(\boldsymbol{u})$ are bounded for a constant $s>2$.
(B3) The kernel $\kappa$ is second order, i.e., $\int \boldsymbol{u} \kappa(\boldsymbol{u}) \mathrm{d} \boldsymbol{u}=0$ and $0<\int \boldsymbol{u}^{\top} \boldsymbol{u} \kappa(\boldsymbol{u}) \mathrm{d} \boldsymbol{u}<\infty$.
(B4) The bandwidth $b$ satisfies $b \rightarrow 0$ and $(a+1) n b^{p+4} \rightarrow c \in[0, \infty)$, where $a=\lim _{n \rightarrow \infty}(N-n) / n$ (assumed to exist without loss of generality).
(B5) The kernels $\tilde{\kappa}$ and $\bar{\kappa}$ in (6) have bounded supports and orders $\tilde{m}>2+2 / p$ and $\bar{m}>2$, respectively, as defined by Bierens (1987), $f(y, \boldsymbol{u} \mid D=1), f(y, \boldsymbol{u} \mid D=0)$ are $\tilde{m} t h$-order continuously differentiable with bounded partial derivatives, and $f_{1}(\boldsymbol{u})$ and $f_{0}(\boldsymbol{u})$ are $\bar{m} t h$-order continuously differentiable with bounded partial derivatives. Functions $f(y, \boldsymbol{u} \mid D=0)$ and $f_{1}(\boldsymbol{u})$ are bounded away from zero. The bandwidths $\tilde{b}$ and $\bar{b}$ satisfy $n \tilde{b}^{p+1} / \log (n) \rightarrow \infty$ and $n \bar{b}^{p} / \log (n) \rightarrow \infty$.

Then, for any fixed $\boldsymbol{u}$ with $f_{0}(\boldsymbol{u})>0$ and $f_{1}(\boldsymbol{u})>0$ and $\widehat{\mu}_{1}^{E 2}$ in (5),

$$
\begin{equation*}
\sqrt{n b^{p}}\left\{\widehat{\mu}_{1}^{E 2}(\boldsymbol{u})-\mu_{1}(\boldsymbol{u})\right\} \xrightarrow{d} N\left(B_{a}(\boldsymbol{u}), V_{a}(\boldsymbol{u})\right), \tag{17}
\end{equation*}
$$

where $\xrightarrow{d}$ denotes convergence in distribution as $n \rightarrow \infty$,

$$
\begin{aligned}
& B_{a}(\boldsymbol{u})=\frac{c^{1 / 2}\left\{f_{1}(\boldsymbol{u}) A_{1}(\boldsymbol{u})+a f_{0}(\boldsymbol{u}) A_{0}(\boldsymbol{u})\right\}}{(a+1)^{1 / 2}\left\{f_{1}(\boldsymbol{u})+a f_{0}(\boldsymbol{u})\right\}}, \\
& A_{1}(\boldsymbol{u})=\int \kappa(\boldsymbol{v})\left\{\frac{1}{2} \boldsymbol{v}^{\top} \nabla^{2} \mu_{1}(\boldsymbol{u}) \boldsymbol{v}+\boldsymbol{v}^{\top} \nabla \log f_{1}(\boldsymbol{u}) \nabla \mu_{1}(\boldsymbol{u})^{\top} \boldsymbol{v}\right\} \mathrm{d} \boldsymbol{v}, \\
& A_{0}(\boldsymbol{u})=\int \kappa(\boldsymbol{v})\left\{\frac{1}{2} \boldsymbol{v}^{\top} \nabla^{2} \mu_{1}(\boldsymbol{u}) \boldsymbol{v}+\boldsymbol{v}^{\top} \nabla \log f_{0}(\boldsymbol{u}) \nabla \mu_{1}(\boldsymbol{u})^{\top} \boldsymbol{v}\right\} \mathrm{d} \boldsymbol{v}, \\
& V_{a}(\boldsymbol{u})=\frac{f_{1}(\boldsymbol{u}) \sigma_{1}^{2}(\boldsymbol{u})+a f_{0}(\boldsymbol{u}) \sigma_{0}^{2}(\boldsymbol{u})}{\left\{f_{1}(\boldsymbol{u})+a f_{0}(\boldsymbol{u})\right\}^{2}} \int \kappa(\boldsymbol{v})^{2} \mathrm{~d} \boldsymbol{v} .
\end{aligned}
$$

Conditions (B1)-(B4) are typically assumed for kernel estimation (Bierens, 1987). Condition (B5) is a sufficient condition for

$$
\begin{equation*}
\max _{i=n+1, \ldots, N}\left|\frac{\widehat{f}\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=1\right)}{\widehat{f}\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=0\right)}-\frac{f\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=1\right)}{f\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=0\right)}\right|=\frac{o_{p}(1)}{\sqrt{n b^{p}}} \tag{18}
\end{equation*}
$$

(Lemma 8.10 in Newey \& McFadden, 1994), where $o_{p}(1)$ denotes a term tending to 0 in probability. Result (18) implies that the estimation of ratio $f(Y \mid \boldsymbol{U}, D=1) / f(Y \mid \boldsymbol{U}, D=0)$ does not affect the asymptotic distribution of $\widehat{\mu}_{1}^{E 2}(\boldsymbol{u})$ in (5).

Note that both the squared bias $B_{a}^{2}(\boldsymbol{u})$ and variance $V_{a}(\boldsymbol{u})$ in (17) are decreasing in the limit $a=\lim _{n \rightarrow \infty}$ $(N-n) / n$, a quantity reflecting how many external data we have. In the extreme case of $a=0$, i.e., the size of the external dataset is negligible compared with the size of the internal dataset, result (17) reduces to the wellknown asymptotic normality for the standard kernel estimator $\widehat{\mu}_{1}(\boldsymbol{u})$ in (2) (Bierens, 1987). In the other extreme case of $a=\infty$, on the other hand, $B_{a}(\boldsymbol{u})=V_{a}(\boldsymbol{u})=0$ and, hence, $\widehat{\mu}_{1}^{E 2}(\boldsymbol{u})$ has a convergence rate tending to 0 faster than $1 / \sqrt{n b^{p}}$, the convergence rate of the standard kernel estimator $\widehat{\mu}_{1}(\boldsymbol{u})$.

The next result is about $\widehat{\mu}_{1}^{C 2}(\boldsymbol{u})$ in (16) as described in Section 3.
Theorem 4.2: Assume (B1)-(B5) with $\boldsymbol{U}$ and $p$ replaced by $\boldsymbol{X}$ and $q$, respectively, and the following conditions, where $f_{k}(\boldsymbol{u})$ and $\sigma_{k}^{2}(\boldsymbol{u}), k=0,1$, are defined in (B1)-(B2).
(C1) The range $\mathscr{U}$ of $\boldsymbol{U}$ is a compact set in the p-dimensional Euclidean space and $f_{1}(\boldsymbol{u})$ is bounded away from infinity and zero on $\mathscr{U} ; f_{1}(\boldsymbol{u})$ and $f_{0}(\boldsymbol{u})$ have continuous and bounded first- and second-order partial derivatives.
(C2) Functions $\mu_{1}(\boldsymbol{u})=E(Y \mid \boldsymbol{U}=\boldsymbol{u})$ and $\sigma_{1}^{2}(\boldsymbol{u})$ are Lipschitz continuous; $\mu_{1}(\boldsymbol{u})$ has bounded third-order partial derivatives; $h_{1}(\boldsymbol{x})=E(Y \mid \boldsymbol{X}=\boldsymbol{x}, D=1)$ has bounded first- and second-order partial derivatives; and $E\left(|Y|^{s} \mid \boldsymbol{U}=\boldsymbol{u}, D=1\right)$ is bounded with $s>2+p / 2$.
(C3) All kernel functions are positive, bounded, and Lipschitz continuous with mean zero and finite sixth moments.
(C4) $a=\lim _{n \rightarrow \infty}(N-n) / n>0$ and the bandwidths $b$ in (2) and lin (14) satisfy $b \rightarrow 0, l \rightarrow 0, l / b \rightarrow r \in(0, \infty)$, $n b^{p} \rightarrow \infty$, and $n b^{4+p} \rightarrow c \in[0, \infty)$, as $n \rightarrow \infty$.
(C5) The densities $f_{X 0}(\boldsymbol{x})$ and $f_{X 1}(\boldsymbol{x})$ for $\boldsymbol{X}$, respectively under internal and external populations are bounded away from zero. There exists a constant $s>4$ such that $E\left(|Y|^{s} \mid D=1\right)$ and $E\left(|\tilde{Y}|^{s} \mid D=0\right)$ are finite, $E\left(|Y|^{s} \mid\right.$
$\boldsymbol{X}=\boldsymbol{x}, D=1) f_{X 1}(\boldsymbol{x})$ and $E\left(|\tilde{Y}|^{s} \mid \boldsymbol{X}=\boldsymbol{x}, D=0\right) f_{X 0}(\boldsymbol{x})$ are bounded, and the bandwidth $b_{h}$ for $\widehat{h}_{1}$ satisfies $n^{1-2 / s} b_{h}^{q} / \log (n) \rightarrow \infty$.

Then, for any fixed $\boldsymbol{u} \in \mathscr{U}$ and $\widehat{\mu}_{1}^{C 2}(\boldsymbol{u})$ in (16),

$$
\begin{equation*}
\sqrt{n b^{p}}\left\{\widehat{\mu}_{1}^{C 2}(\boldsymbol{u})-\mu_{1}(\boldsymbol{u})\right\} \xrightarrow{d} N\left(B_{r}(\boldsymbol{u}), V_{r}(\boldsymbol{u})\right), \tag{19}
\end{equation*}
$$

where

$$
\begin{aligned}
& B_{r}(\boldsymbol{u})=c^{1 / 2}\left[\left(1+r^{2}\right) A_{1}(\boldsymbol{u})-r^{2} \boldsymbol{g}(\boldsymbol{x})^{\top} \boldsymbol{\Sigma}_{g}^{-1} E\left\{\boldsymbol{g}(\boldsymbol{X}) A_{1}(\boldsymbol{U}) \mid D=1\right\}\right] \\
& A_{1}(\boldsymbol{u})=\int \kappa(\boldsymbol{v})\left\{\frac{1}{2} \boldsymbol{v}^{\top} \nabla^{2} \mu_{1}(\boldsymbol{u}) \boldsymbol{v}+\boldsymbol{v}^{\top} \nabla \log f_{1}(\boldsymbol{u}) \nabla \mu_{1}(\boldsymbol{u})^{\top} \boldsymbol{v}\right\} \mathrm{d} \boldsymbol{v} \\
& V_{r}(\boldsymbol{u})=\frac{\sigma_{1}^{2}(\boldsymbol{u})}{f_{1}(\boldsymbol{u})} \int\left\{\int \kappa(\boldsymbol{v}-r \boldsymbol{w}) \kappa(\boldsymbol{w}) d \boldsymbol{w}\right\}^{2} \mathrm{~d} \boldsymbol{v}
\end{aligned}
$$

and $\boldsymbol{\Sigma}=E\left\{\boldsymbol{g}(\boldsymbol{X}) \boldsymbol{g}(\boldsymbol{X})^{\top} \mid D=1\right\}$ is assumed to be positive definite without loss of generality.
The next result is about $\widehat{\gamma}$ in (11).
Theorem 4.3: Suppose that (8) holds for binary random $D$ indicating internal and external data. Assume also the following conditions.
(D1) The kernel $\check{\kappa}$ in (10) is Lipschitz continuous, satisfies $\int \check{\kappa}(\boldsymbol{u}) \mathrm{d} \boldsymbol{u}=1$, has a bounded support, and has order $d>\max \{(p+4) / 2, p\}$.
(D2) The bandwidth $\check{b}$ in (10) satisfies $N \check{b}^{2 q} /(\log N)^{2} \rightarrow \infty$ and $N \check{b}^{2 d} \rightarrow 0$ as the total sample size of internal and external datasets $N \rightarrow \infty$, where $d$ is given in (D1).
(D3) $\gamma$ in (8) is an interior point of a compact domain $\Gamma$ and it is the unique solution to $h_{1}(\cdot)=h(\cdot, t), t \in \Gamma$. For any $\boldsymbol{u}, h(\boldsymbol{u}, t)$ is second-order continuously differentiable in $t$, and $h, \nabla_{t} h, \nabla_{t}^{2} h$ are bounded over $t$ and $\boldsymbol{u}$. As $t \rightarrow \gamma, h(\cdot, t), \nabla_{t} h(\cdot, t)$, and $\nabla_{t}^{2} h(\cdot, t)$ convergence uniformly.
(D4) $\sup _{t \in \Gamma} E\left\|\boldsymbol{W}_{t}\right\|^{4}<\infty$ and $\sup _{t \in \Gamma} E\left[\left\|\boldsymbol{W}_{t}\right\|^{4} \mid \boldsymbol{U}\right] f_{U}(\boldsymbol{U})$ is bounded, where $\|\boldsymbol{a}\|^{2}=\boldsymbol{a}^{\top} \boldsymbol{a}, \boldsymbol{W}_{t}=\left(D e^{t Y},(1-D)\right.$ $\left.Y e^{-t Y}, D,(1-D), D Y e^{t Y},(1-D) Y^{2} \mathrm{e}^{-t Y}, D Y^{2} \mathrm{e}^{t Y},(1-D) Y^{3} \mathrm{e}^{-t Y}\right)^{\top}$, and $f_{U}$ is the density of $\boldsymbol{U}$. Furthermore, there is a function $\tau(Y, D)$ with $E\{\tau(Y, D)\}<\infty$ such that $\left\|\boldsymbol{W}_{t}-\boldsymbol{W}_{t^{\prime}}\right\|<\tau(Y, D)\left|t-t^{\prime}\right|$.
(D5) The function $\boldsymbol{\omega}_{t}(\boldsymbol{u})=E\left(\boldsymbol{W}_{t} \mid \boldsymbol{U}=\boldsymbol{u}\right) f_{U}(\boldsymbol{u})$ is bounded away from zero, and it is dth-order continuously differentiable with bounded partial derivatives on an open set containing the support of $\boldsymbol{U}$. There is a functional $G(Y, D, \omega)$ linear in $\boldsymbol{\omega}$ such that $|G(Y, D, \omega)| \leq \iota(Y, D)\|\boldsymbol{\omega}\|_{\infty}$ and, for small enough $\| \boldsymbol{\omega}$ $\boldsymbol{\omega}_{\gamma}\left\|_{\infty},\left|\psi(Y, D, \boldsymbol{\omega})-\psi\left(Y, D, \omega_{\gamma}\right)-G\left(Y, D, \boldsymbol{\omega}-\boldsymbol{\omega}_{\gamma}\right)\right| \leq \iota(Y, D)\right\| \boldsymbol{\omega}-\boldsymbol{\omega}_{\gamma} \|_{\infty}^{2}$, where $\iota(Y, D)$ is a function with $E\{\iota(Y, D)\}<\infty, \psi(Y, D, \boldsymbol{\omega})=-2 D\left(Y-\frac{\omega_{1} \omega_{2}}{\omega_{3} \omega_{4}}\right)\left(\frac{\omega_{2} \omega_{5}-\omega_{1} \omega_{6}}{\omega_{3} \omega_{4}}\right)$, $\omega_{j}$ is the jth component of $\boldsymbol{\omega},\|\boldsymbol{\omega}\|_{\infty}=$ $\sup _{\boldsymbol{x} \in \mathscr{U}}\|\boldsymbol{\omega}(\boldsymbol{u})\|,\left\|\boldsymbol{\omega}-\boldsymbol{\omega}_{\gamma}\right\|_{\infty}=\sup _{\boldsymbol{x} \in \mathscr{U}}\left\|\boldsymbol{\omega}(\boldsymbol{u})-\boldsymbol{\omega}_{\gamma}(\boldsymbol{u})\right\|$, and $\mathscr{U}$ is the range of $\boldsymbol{U}$. Also, there exists an almost everywhere continuous 8-dimensional function $\boldsymbol{v}(\boldsymbol{U})$ with $\int\|\boldsymbol{v}(\boldsymbol{u})\| \mathrm{d} \boldsymbol{u}<\infty$ and $E\left\{\sup _{\|\boldsymbol{\delta}\| \leq \epsilon} \| \boldsymbol{v}(\boldsymbol{U}+\right.$ $\left.\boldsymbol{\delta}) \|^{4}\right\}<\infty$ for some $\epsilon>0$ such that $E\{G(Y, D, \boldsymbol{\omega})\}=\int \boldsymbol{v}(\boldsymbol{u})^{\top} \boldsymbol{\omega}(\boldsymbol{u}) \mathrm{d} \boldsymbol{u}$ for all $\|\boldsymbol{\omega}\|_{\infty}<\infty$.

Then, as the total sample size of internal and external datasets $N \rightarrow \infty$,

$$
\begin{equation*}
\sqrt{N}(\widehat{\gamma}-\gamma) \xrightarrow{d} N\left(0, \sigma_{\gamma}^{2}\right), \tag{20}
\end{equation*}
$$

where $\sigma_{\gamma}^{2}=\left[2 E\left\{D \nabla_{\gamma} h(\boldsymbol{U}, \gamma)\right\}^{2}\right]^{-1} \operatorname{Var}\left[\psi\left(Y, D, \boldsymbol{\omega}_{\gamma}\right)+\boldsymbol{v}(\boldsymbol{U})^{\top} \boldsymbol{W}_{\gamma}-E\left\{\boldsymbol{v}(\boldsymbol{U})^{\top} \boldsymbol{W}_{\gamma}\right\}\right]$.
Conditions (D1)-(D5) are technical assumptions discussed in Lemmas 8.11 and 8.12 in Newey and McFadden (1994). As discussed by Newey and McFadden (1994), the condition that $\check{\kappa}$ has a bounded support can be relaxed, as it is imposed for a simple proof.

Combining Theorems 4.1-4.3, we obtain the following result for $\widehat{\mu}_{1}^{E 3}(\boldsymbol{u})$ in (12) or $\widehat{\mu}_{1}^{C 3}(\boldsymbol{u})$ in (16).
Corollary 4.1: Suppose that (8) holds for the binary random D indicating internal and external data.
(i) Under (B1)-(B4) and (D1)-(D5), result (17) holds with $\widehat{\mu}_{1}^{E 2}(\boldsymbol{u})$ replaced by $\widehat{\mu}_{1}^{E 3}(\boldsymbol{u})$.
(ii) Under (C1)-(C4) and (D1)-(D5) with $\boldsymbol{U}$ and $p$ replaced by $\boldsymbol{X}$ and q, respectively, result (19) holds with $\widehat{\mu}_{1}^{C 2}(\boldsymbol{u})$ replaced by $\widehat{\mu}_{1}^{C 3}(\boldsymbol{u})$.

## 5. Simulation results

### 5.1. The performance of $\widehat{\mu}_{1}^{C j}$ given by (16)

We first present simulation results to examine and compare the performance of the standard kernel estimator $\widehat{\mu}_{1}$ in (2) without using external information and our proposed estimator (16) with three variations, $\widehat{\mu}_{1}^{C 1}$, $\widehat{\mu}_{1}^{C 2}$, and $\widehat{\mu}_{1}^{C 3}$, as described in the end of Section 3. We consider $\boldsymbol{U}=(X, Z)^{\top}$ with univariate covariates $X$ and $Z$, where $Z$ is unmeasured in the external dataset ( $p=2$ and $q=1$ ). The covariates are generated in two ways:
(i) normal covariates: $(X, Z)^{\top}$ is bivariate normal with means 0 , variances 1 , and correlation 0.5 ;
(ii) bounded covariates: $X=B W_{1}+(1-B) W_{2}$ and $Z=B W_{1}+(1-B) W_{3}$, where $W_{1}, W_{2}$, and $W_{3}$ are identically distributed as uniform on $[-1,1], B$ is uniform on $[0,1]$, and $W_{1}, W_{2}, W_{3}$, and $B$ are independent.

Conditioned on $(X, Z)^{\top}$, the response $Y$ is normal with mean $\mu(X, Z)$ and variance 1 , where $\mu(X, Z)$ follows one of the following four models:
(M1) $\mu(X, Z)=X / 2-Z^{2} / 4$;
(M2) $\mu(X, Z)=\cos (2 X) / 2+\sin (Z)$;
(M3) $\mu(X, Z)=\cos (2 X Z) / 2+\sin (Z)$;
(M4) $\mu(X, Z)=X / 2-Z^{2} / 4+\cos (X Z) / 4$

Note that all four models are nonlinear in $(X, Z)^{\top}$; M1-M2 are additive models, while M3-M4 are non-additive.
A total of $N=1,200$ data are generated from the population of $(Y, X, Z)$ as previously described. A data point is treated as internal or external according to a random binary $D$ with conditional probability $P(D=1 \mid$ $Y, X, Z)=1 / \exp \left(\gamma_{0}+2|X|+\gamma Y\right)$, where $\gamma=0$ or $1 / 2$, and $\gamma_{0}=1$ or -1.5 . Under the setting $\gamma_{0}=1$ or -1.5 , the unconditional $P(D=1) \approx n / N$ is around $13 \%$ or $50 \%$.

The simulation studies performance of kernel estimators in terms of mean integrated square error (MISE). The following measure is calculated by simulation with $S$ replications:

$$
\begin{equation*}
\operatorname{MISE}\left(\widehat{\mu}_{1}^{*}\right)=\frac{1}{S} \sum_{s=1}^{S} \frac{1}{T} \sum_{t=1}^{T}\left\{\widehat{\mu}_{1}^{*}\left(\boldsymbol{U}_{s, t}\right)-\mu_{1}\left(\boldsymbol{U}_{s, t}\right)\right\}^{2} \tag{21}
\end{equation*}
$$

where $\left\{\boldsymbol{U}_{s, t}: t=1, \ldots, T\right\}$ are test data for each simulation replication $s$, the simulation is repeated independently for $s=1, \ldots, S$, and $\widehat{\mu}_{1}^{*}$ is one of $\widehat{\mu}_{1}, \widehat{\mu}_{1}^{C 1}, \widehat{\mu}_{1}^{C 2}$, and $\widehat{\mu}_{1}^{C 3}$, independent of test data. We consider two ways of generating test data $\boldsymbol{U}_{s, t}$ 's. The first one is to use $T=121$ fixed grid points on $[-1,1] \times[-1,1]$ with equal space. The second one is to take a random sample of $T=121$ without replacement from the covariate $\boldsymbol{U}$ 's of the internal dataset, for each fixed $s=1, \ldots, S$ and independently across $s$.

To show the benefit of using external information, we calculate the improvement in efficiency defined as follows:

$$
\begin{equation*}
\operatorname{IMP}=1-\frac{\min \left\{\operatorname{MISE}\left(\widehat{\mu}_{1}^{*}\right)\right\}}{\operatorname{MISE}\left(\widehat{\mu}_{1}\right)} \tag{22}
\end{equation*}
$$

where the minimum is over $\widehat{\mu}_{1}^{*}=$ one of $\widehat{\mu}_{1}, \widehat{\mu}_{1}^{C 1}, \widehat{\mu}_{1}^{C 2}$, and $\widehat{\mu}_{1}^{C 3}$.
In all cases, we use the Gaussian kernel. The bandwidths $b$ and $l$ affect the performance of kernel methods. We consider two types of bandwidths in the simulation. The first one is 'the best bandwidth'; for each method, we evaluate MISE in a pool of bandwidths and display the one that has the minimal MISE. This shows the best we can achieve in terms of bandwidth, but it cannot be used in applications. The second one is to select bandwidth from a pool of bandwidths via 10 -fold cross validation (Györfi et al., 2002), which produces a decent bandwidth that can be applied to real data.

The simulated MISE values based on $S=200$ replications are shown in Tables 1-4.
Consider first the results in Tables $1-2$. Since $\gamma=0$, all three estimators, $\widehat{\mu}_{1}^{C 1}, \widehat{\mu}_{1}^{C 2}$, and $\widehat{\mu}_{1}^{C 3}$, are correct and more efficient than the standard estimator $\widehat{\mu}_{1}$ in (2) without using external information. The estimator $\widehat{\mu}_{1}^{C 1}$ is the best, as it uses the correct information that populations are homogeneous $(\gamma=0)$ and is simpler than $\widehat{\mu}_{1}^{C 2}$ and $\widehat{\mu}_{1}^{C 3}$.

Next, the results in Tables 3-4 for $\gamma=1 / 2$ indicate that the estimator $\widehat{\mu}_{1}^{C 2}$ or $\widehat{\mu}_{1}^{C 3}$ using a correct constraint is better than the estimator $\widehat{\mu}_{1}^{C 1}$ using an incorrect constraint or the estimator $\widehat{\mu}_{1}$ without using external information. Since $\widehat{\mu}_{1}^{C 3}$ uses more information, it is in general better than $\widehat{\mu}_{1}^{C 2}$. Furthermore, with an incorrect constraint, $\widehat{\mu}_{1}^{C 1}$ can be much worse than $\widehat{\mu}_{1}$ without using external information.

Table 1. Simulated MISE (21) and IMP (22) when the external dataset contains only $X$, with $S=200$ under $\gamma=0, n / N \approx 13 \%$.

| Covariate | Model | Test data | $b, 1$ | Estimator |  |  |  | IMP \% | Mean of $\hat{\gamma}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{C 1}$ | $\widehat{\mu}_{1}^{C 2}$ | $\widehat{\mu}_{1}^{\text {C3 }}$ |  |  |
| Normal | M1 | Sample | Best | 0.040 | 0.017 | 0.020 | 0.022 | 56.96 | -0.003 |
|  |  |  | CV | 0.057 | 0.043 | 0.042 | 0.051 | 27.16 | -0.007 |
|  |  | Grid | Best | 0.026 | 0.009 | 0.013 | 0.014 | 66.49 | -0.012 |
|  |  |  | CV | 0.037 | 0.022 | 0.022 | 0.028 | 40.53 | -0.003 |
|  | M2 | Sample | Best | 0.042 | 0.021 | 0.027 | 0.026 | 50.58 | -0.042 |
|  |  |  | CV | 0.074 | 0.056 | 0.061 | 0.060 | 23.52 | -0.039 |
|  |  | Grid | Best | 0.030 | 0.015 | 0.025 | 0.023 | 50.36 | -0.026 |
|  |  |  | CV | 0.051 | 0.033 | 0.040 | 0.036 | 35.62 | -0.035 |
|  | M3 | Sample | Best | 0.039 | 0.020 | 0.025 | 0.023 | 48.55 | -0.023 |
|  |  |  | CV | 0.063 | 0.050 | 0.053 | 0.055 | 21.16 | -0.020 |
|  |  | Grid | Best | 0.028 | 0.014 | 0.020 | 0.021 | 51.73 | -0.010 |
|  |  |  | CV | 0.046 | 0.029 | 0.033 | 0.033 | 36.33 | -0.024 |
|  | M4 | Sample | Best | 0.041 | 0.018 | 0.020 | 0.023 | 56.28 | -0.007 |
|  |  |  | CV | 0.064 | 0.050 | 0.047 | 0.056 | 25.83 | -0.017 |
|  |  | Grid | Best | 0.026 | 0.009 | 0.014 | 0.015 | 65.54 | -0.006 |
|  |  |  | CV | 0.036 | 0.021 | 0.020 | 0.026 | 44.18 | -0.007 |
| Bounded | M1 | Sample | Best | 0.010 | 0.002 | 0.006 | 0.007 | 79.58 | -0.007 |
|  |  |  | CV | 0.016 | 0.006 | 0.007 | 0.010 | 63.65 | 0.001 |
|  |  | Grid | Best | 0.013 | 0.002 | 0.006 | 0.007 | 82.93 | -0.000 |
|  |  |  | CV | 0.041 | 0.012 | 0.013 | 0.017 | 70.86 | 0.009 |
|  | M2 | Sample |  | 0.011 |  | 0.007 | 0.008 | 74.88 | -0.031 |
|  |  |  | CV | 0.037 | 0.008 | 0.011 | 0.013 | 77.46 | -0.018 |
|  |  | Grid | Best | 0.016 | 0.004 | 0.009 | 0.010 | 74.00 | -0.023 |
|  |  |  | CV | 0.086 | 0.018 | 0.019 | 0.021 | 79.20 | -0.019 |
|  | M3 | Sample | Best | 0.011 | 0.003 | 0.006 | 0.007 | 73.85 | -0.003 |
|  |  |  | CV | 0.037 | 0.009 | 0.011 | 0.014 | 76.48 | -0.008 |
|  |  | Grid | Best | 0.016 | 0.004 | 0.008 | 0.009 | 73.94 | -0.004 |
|  |  |  | CV | 0.083 | 0.018 | 0.020 | 0.023 | 78.46 | 0.004 |
|  | M4 | Sample | Best | 0.009 | 0.002 | 0.005 | 0.006 | 77.43 | -0.007 |
|  |  |  | CV | 0.018 | 0.006 | 0.008 | 0.011 | 68.73 | 0.002 |
|  |  | Grid | Best | 0.013 | 0.002 | 0.006 | 0.007 | 82.23 | -0.007 |
|  |  |  | CV | 0.036 | 0.010 | 0.011 | 0.015 | 71.59 | 0.003 |

Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.005 and 0.006 .

Table 2. Simulated MISE (21) and IMP (22) when the external dataset contains only $X$, with $S=200$ under $\gamma=0, n / N \approx 50 \%$.

| Covariate | Model | Test data | $b, 1$ | Estimator |  |  |  | IMP \% | Mean of $\hat{\gamma}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{C 1}$ | $\widehat{\mu}_{1}^{C 2}$ | $\widehat{\mu}_{1}^{\text {c3 }}$ |  |  |
| Normal | M1 | Sample | Best | 0.012 | 0.007 | 0.006 | 0.007 | 44.75 | -0.017 |
|  |  |  | CV | 0.033 | 0.027 | 0.021 | 0.028 | 36.37 | -0.005 |
|  |  | Grid | Best | 0.006 | 0.004 | 0.005 | 0.005 | 34.29 | -0.009 |
|  |  |  | CV | 0.017 | 0.012 | 0.008 | 0.012 | 53.01 | -0.007 |
|  | M2 | Sample | Best | 0.013 | 0.008 | 0.010 | 0.009 | 36.08 | -0.057 |
|  |  |  | CV | 0.052 | 0.027 | 0.027 | 0.026 | 48.95 | -0.062 |
|  |  | Grid | Best | 0.010 | 0.005 | 0.007 | 0.006 | 45.83 | -0.058 |
|  |  |  | CV | 0.036 | 0.016 | 0.018 | 0.016 | 57.27 | -0.063 |
|  | M3 | Sample | Best | 0.014 | 0.009 | 0.010 | 0.010 | 37.36 | -0.032 |
|  |  |  | CV | 0.050 | 0.027 | 0.027 | 0.027 | 46.48 | -0.041 |
|  |  | Grid | Best | 0.008 | 0.005 | 0.006 | 0.005 | 43.13 | -0.027 |
|  |  |  | CV | 0.030 | 0.013 | 0.015 | 0.014 | 55.91 | -0.031 |
|  | M4 | Sample | Best | 0.014 | 0.009 | 0.008 | 0.009 | 44.00 | -0.020 |
|  |  |  | CV | 0.041 | 0.033 | 0.026 | 0.033 | 36.63 | -0.014 |
|  |  | Grid | Best | 0.006 | 0.004 | 0.005 | 0.005 | 37.33 | -0.008 |
|  |  |  | CV | 0.017 | 0.012 | 0.008 | 0.013 | 53.05 | -0.023 |
| Bounded | M1 | Sample | Best | 0.002 | 0.001 | 0.001 | 0.002 | 40.67 | -0.008 |
|  |  |  | CV | 0.009 | 0.005 | 0.004 | 0.005 | 52.83 | 0.009 |
|  |  | Grid | Best | 0.004 | 0.002 | 0.002 | 0.002 | 55.17 | -0.010 |
|  |  |  | CV | 0.021 | 0.008 | 0.007 | 0.008 | 67.74 | -0.000 |
|  | M2 | Sample | Best | 0.004 | 0.002 | 0.002 | 0.002 | 58.55 | -0.035 |
|  |  |  | CV | 0.027 | 0.007 | 0.006 | 0.007 | 77.51 | -0.039 |
|  |  | Grid | Best | 0.006 | 0.002 | 0.003 | 0.003 | 64.61 | -0.033 |
|  |  |  | CV | 0.056 | 0.011 | 0.011 | 0.011 | 80.27 | -0.036 |
|  | M3 | Sample | Best | 0.004 | 0.002 | 0.002 | 0.002 | 60.82 | -0.006 |
|  |  |  | CV | 0.024 | 0.006 | 0.006 | 0.006 | 76.60 | -0.008 |
|  |  | Grid | Best | 0.006 | 0.002 | 0.002 | 0.003 | 66.02 | -0.011 |
|  |  |  | CV | 0.050 | 0.011 | 0.011 | 0.011 | 78.56 | -0.010 |
|  | M4 | Sample | Best | 0.002 | 0.001 | 0.001 | 0.002 | 43.16 | -0.011 |
|  |  |  | CV | 0.010 | 0.005 | 0.005 | 0.005 | 54.67 | -0.003 |
|  |  | Grid | Best | 0.004 | 0.002 | 0.002 | 0.002 | 53.45 | -0.007 |
|  |  |  | CV | 0.024 | 0.009 | 0.008 | 0.009 | 68.53 | -0.002 |

Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.004 and 0.005 .

Table 3. Simulated MISE (21) and IMP (22) when the external dataset contains only $X$, with $S=200$ under $\gamma=0.5, n / N \approx 13 \%$.

| Covariate | Model | Test data | b, I | Estimator |  |  |  | IMP \% | Mean of $\hat{\gamma}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{C 1}$ | $\widehat{\mu}_{1}^{C 2}$ | $\widehat{\mu}_{1}^{\text {C3 }}$ |  |  |
| Normal | M1 | Sample | Best | 0.042 | 0.181 | 0.033 | 0.028 | 32.15 | 0.449 |
|  |  |  | CV | 0.055 | 0.193 | 0.061 | 0.055 | -0.91 | 0.449 |
|  |  | Grid | Best | 0.022 | 0.164 | 0.019 | 0.013 | 40.36 | 0.453 |
|  |  |  | CV | 0.030 | 0.193 | 0.048 | 0.030 | 2.13 | 0.435 |
|  | M2 | Sample | Best | 0.041 | 0.295 | 0.047 | 0.026 | 37.01 | 0.429 |
|  |  |  | CV | 0.077 | 0.324 | 0.103 | 0.066 | 15.26 | 0.424 |
|  |  | Grid | Best | 0.030 | 0.279 | 0.040 | 0.022 | 27.11 | 0.428 |
|  |  |  | CV | 0.059 | 0.315 | 0.087 | 0.043 | 28.20 | 0.421 |
|  | M3 | Sample | Best | 0.048 | 0.335 | 0.047 | 0.028 | 40.81 | 0.452 |
|  |  |  | CV | 0.076 | 0.370 | 0.104 | 0.066 | 12.50 | 0.448 |
|  |  | Grid | Best | 0.031 | 0.300 | 0.035 | 0.021 | 31.59 | 0.441 |
|  |  |  | CV | 0.053 | 0.355 | 0.082 | 0.041 | 23.25 | 0.447 |
|  | M4 | Sample | Best | 0.048 | 0.193 | 0.035 | 0.030 | 37.27 | 0.452 |
|  |  |  | CV | 0.067 | 0.205 | 0.067 | 0.061 | 8.73 | 0.433 |
|  |  | Grid | Best | 0.023 | 0.177 | 0.023 | 0.014 | 39.42 | 0.437 |
|  |  |  | CV | 0.036 | 0.204 | 0.049 | 0.031 | 15.80 | 0.452 |
| Bounded | M1 | Sample | Best | 0.009 | 0.146 | 0.011 | 0.007 | 23.90 | 0.482 |
|  |  |  | CV | 0.017 | 0.150 | 0.026 | 0.011 | 35.66 | 0.483 |
|  |  | Grid | Best | 0.013 | 0.152 | 0.016 | 0.007 | 44.41 | 0.470 |
|  |  |  | CV | 0.037 | 0.159 | 0.030 | 0.017 | 52.24 | 0.485 |
|  | M2 | Sample | Best | 0.012 | 0.196 | 0.018 | 0.008 | 33.06 | 0.468 |
|  |  |  | CV | 0.044 | 0.206 | 0.036 | 0.017 | 60.76 | 0.468 |
|  |  | Grid | Best | 0.016 | 0.205 | 0.022 | 0.010 | 39.95 | 0.463 |
|  |  |  | CV | 0.088 | 0.214 | 0.043 | 0.024 | 72.45 | 0.469 |
|  | M3 | Sample | Best | 0.011 | 0.209 | 0.016 | 0.007 | 34.46 | 0.479 |
|  |  |  | CV | 0.040 | 0.216 | 0.034 | 0.016 | 60.92 | 0.475 |
|  |  | Grid | Best | 0.018 | 0.218 | 0.019 | 0.011 | 39.73 | 0.486 |
|  |  |  | CV | 0.089 | 0.228 | 0.042 | 0.026 | 70.77 | 0.483 |
|  | M4 | Sample | Best | 0.010 | 0.152 | 0.012 | 0.007 | 28.69 | 0.484 |
|  |  |  | CV | 0.019 | 0.157 | 0.025 | 0.012 | 36.41 | 0.485 |
|  |  | Grid | Best | 0.014 | 0.164 | 0.015 | 0.007 | 46.99 | 0.485 |
|  |  |  | CV | 0.040 | 0.170 | 0.029 | 0.019 | 52.63 | 0.492 |

Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.005 and 0.006 .

Table 4. Simulated MISE (21) and IMP (22) when the external dataset contains only $X$, with $S=200$ under $\gamma=0.5, n / N \approx 50 \%$.

| Covariate | Model | Test data | $b, 1$ | Estimator |  |  |  | IMP \% | Mean of $\hat{\gamma}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{C 1}$ | $\widehat{\mu}_{1}^{C 2}$ | $\widehat{\mu}_{1}^{\text {c3 }}$ |  |  |
| Normal | M1 | Sample | Best | 0.013 | 0.035 | 0.008 | 0.009 | 39.89 | 0.431 |
|  |  |  | CV | 0.037 | 0.055 | 0.027 | 0.034 | 25.63 | 0.434 |
|  |  | Grid | Best | 0.007 | 0.032 | 0.005 | 0.004 | 34.86 | 0.439 |
|  |  |  | CV | 0.014 | 0.046 | 0.014 | 0.012 | 14.98 | 0.437 |
|  | M2 | Sample | Best | 0.015 | 0.061 | 0.014 | 0.010 | 33.07 | 0.408 |
|  |  |  | CV | 0.063 | 0.086 | 0.041 | 0.030 | 52.47 | 0.406 |
|  |  | Grid | Best | 0.009 | 0.061 | 0.011 | 0.007 | 28.26 | 0.406 |
|  |  |  | CV | 0.043 | 0.078 | 0.032 | 0.018 | 59.17 | 0.400 |
|  | M3 | Sample | Best | 0.015 | 0.070 | 0.014 | 0.010 | 33.24 | 0.441 |
|  |  |  | CV | 0.056 | 0.093 | 0.037 | 0.029 | 48.27 | 0.440 |
|  |  | Grid | Best | 0.008 | 0.070 | 0.011 | 0.006 | 30.34 | 0.439 |
|  |  |  | CV | 0.034 | 0.095 | 0.032 | 0.017 | 51.53 | 0.438 |
|  | M4 | Sample | Best | 0.015 | 0.041 | 0.010 | 0.010 | 35.19 | 0.423 |
|  |  |  | CV | 0.048 | 0.063 | 0.033 | 0.040 | 31.52 | 0.414 |
|  |  | Grid | Best | 0.006 | 0.033 | 0.005 | 0.005 | 21.67 | 0.422 |
|  |  |  | CV | 0.017 | 0.054 | 0.016 | 0.014 | 17.89 | 0.423 |
| Bounded | M1 | Sample | Best | 0.003 | 0.022 | 0.003 | 0.002 | 20.26 | 0.477 |
|  |  |  | CV | 0.010 | 0.027 | 0.008 | 0.006 | 39.55 | 0.474 |
|  |  | Grid | Best | 0.004 | 0.024 | 0.003 | 0.002 | 37.63 | 0.480 |
|  |  |  | CV | 0.022 | 0.029 | 0.009 | 0.009 | 57.30 | 0.480 |
|  | M2 | Sample | Best | 0.004 | 0.034 | 0.004 | 0.002 | 45.43 | 0.462 |
|  |  |  | CV | 0.030 | 0.042 | 0.011 | 0.008 | 72.70 | 0.462 |
|  |  | Grid | Best | 0.007 | 0.037 | 0.006 | 0.003 | 56.85 | 0.457 |
|  |  |  | CV | 0.057 | 0.041 | 0.014 | 0.012 | 78.52 | 0.468 |
|  | M3 | Sample | Best | 0.004 | 0.037 | 0.004 | 0.002 | 47.13 | 0.474 |
|  |  |  | CV | 0.026 | 0.045 | 0.010 | 0.007 | 73.40 | 0.487 |
|  |  | Grid | Best | 0.006 | 0.043 | 0.005 | 0.002 | 56.60 | 0.476 |
|  |  |  | CV | 0.050 | 0.046 | 0.014 | 0.011 | 77.10 | 0.483 |
|  | M4 | Sample | Best | 0.003 | 0.024 | 0.002 | 0.002 | 26.50 | 0.483 |
|  |  |  | CV | 0.010 | 0.030 | 0.007 | 0.006 | 43.47 | 0.475 |
|  |  | Grid | Best | 0.004 | 0.027 | 0.003 | 0.002 | 45.69 | 0.480 |
|  |  |  | CV | 0.023 | 0.032 | 0.009 | 0.009 | 61.42 | 0.485 |

Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.004 and 0.005 .

### 5.2. The performance of $\widehat{\mu}_{1}^{E j}$ given by (3), (5), or (12)

Under the same simulation setting as described in Section 5.1 but with covariate $Z$ measured in both internal and external datasets, we compare the performance of three estimators, $\widehat{\mu}_{1}^{E 1}, \widehat{\mu}_{1}^{E 2}$, and $\widehat{\mu}_{1}^{E 3}$ given by (3), (5), and (12), respectively, with the standard kernel estimator $\widehat{\mu}_{1}$ in (2) without using external information. The mean integrated squared error (MISE) and improvement (IMP) are calculated using formulas (21) and (22), respectively, with $\widehat{\mu}_{1}^{*}=$ one of $\widehat{\mu}_{1}, \widehat{\mu}_{1}^{E 1}, \widehat{\mu}_{1}^{E 2}$, and $\widehat{\mu}_{1}^{E 3}$.

Tables 5-5 present the simulation results. The relative performance of $\widehat{\mu}_{1}^{E 1}, \widehat{\mu}_{1}^{E 2}, \widehat{\mu}_{1}^{E 3}$, and $\widehat{\mu}_{1}$ follows the same pattern as $\widehat{\mu}_{1}^{C 1}, \widehat{\mu}_{1}^{C 2}, \widehat{\mu}_{1}^{C 3}$, and $\widehat{\mu}_{1}$ in Section 5.1.

The only difference between the results here and those in Section 5.1 is that the use of more external data (a smaller $n / N$ ) results in a better performance of $\widehat{\mu}_{1}^{E 2}$ or $\widehat{\mu}_{1}^{E 3}$ (or $\widehat{\mu}_{1}^{E 1}$ when it is correct). This is actually consistent with our theoretical result Theorem 4.1 in Section 4, which shows that both the squared bias $B_{a}^{2}(\boldsymbol{u})$ and variance $V_{a}(\boldsymbol{u})$ in (17) are decreasing in the limit $a=\lim _{n \rightarrow \infty}(N-n) / n$. On the other hand, the simulation results in Section 5.1 and Theorem 4.2 in Section 4 do not show a clear indication of using more external data produces better estimators. The main reason for this is that, when $Z$ is not observed in the external dataset, the estimator $\widehat{\mu}_{1}^{C j}$ relies more on internal data to recover the loss of $Z$ from external dataset in a complicated way.

### 5.3. The performance of $\widehat{\mu}_{1}^{C j}$ given by (16) with $q=2$

We re-consider the simulation in Section 5.1 but with the dimension of $\boldsymbol{X}$ to be $q=2$, i.e., $\boldsymbol{U}=\left(X_{1}, X_{2}, Z\right)^{\top}$. We only consider normally distributed covariates with means 0 , variances 1 , and the correlations in $\left(X_{1}, Z\right),\left(X_{2}, Z\right)$, and ( $X_{1}, X_{2}$ ) being $0.5,0.5$, and 0.25 , respectively. Given $\boldsymbol{U}$, the response variable $Y$ is normally distributed with mean $\mu\left(X_{1}, X_{2}, Z\right)=X_{1} / 2+X_{2} / 4-Z^{2} / 4$ and variance 1 . Moreover, $P(D=1 \mid Y, \boldsymbol{X}, Z)=1 / \exp \left(\gamma_{0}+2\left|X_{1}\right|+\right.$ $\gamma Y$ ), while the remaining settings are the same as in Section 5.1. In calculating MISE (21), we only a random $\boldsymbol{U}_{s, t}$ with $T=121$, not fixed grid points. Also, we consider only evaluate the performance of estimators $\widehat{\mu}_{1}^{C j}$, since estimators $\widehat{\mu}_{1}^{E j}$ are simpler.

Table 5. Simulated $\operatorname{MISE}$ (21) and $\operatorname{IMP}(22)$ when the external dataset contains both $X$ and $Z$, with $S=200$ under $\gamma=0, n / N \approx 13 \%$.

| Covariate | Model | Test data | $b, 1$ | Estimator |  |  |  | IMP \% | Mean of $\hat{\gamma}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{E 1}$ | $\widehat{\mu}_{1}^{E 2}$ | $\widehat{\mu}_{1}^{E 3}$ |  |  |
| Normal | M1 | Sample | Best | 0.055 | 0.020 | 0.035 | 0.025 | 64.46 | 0.020 |
|  |  |  | CV | 0.072 | 0.032 | 0.052 | 0.036 | 55.51 | 0.015 |
|  |  | Grid | Best | 0.047 | 0.010 | 0.027 | 0.017 | 78.44 | 0.013 |
|  |  |  | CV | 0.060 | 0.019 | 0.034 | 0.024 | 68.49 | 0.014 |
|  | M2 | Sample | Best | 0.066 | 0.026 | 0.044 | 0.032 | 60.99 | -0.054 |
|  |  |  | CV | 0.088 | 0.035 | 0.052 | 0.040 | 60.73 | -0.058 |
|  |  | Grid | Best | 0.072 | 0.019 | 0.032 | 0.025 | 74.04 | -0.051 |
|  |  |  | CV | 0.091 | 0.027 | 0.041 | 0.034 | 70.63 | -0.058 |
|  | M3 | Sample | Best | 0.065 | 0.027 | 0.047 | 0.033 | 58.73 | -0.014 |
|  |  |  | CV | 0.078 | 0.035 | 0.053 | 0.041 | 54.98 | -0.009 |
|  |  | Grid | Best | 0.055 | 0.016 | 0.031 | 0.023 | 70.04 | -0.004 |
|  |  |  | CV | 0.067 | 0.026 | 0.038 | 0.033 | 60.79 | 0.002 |
|  | M4 | Sample | Best | 0.055 | 0.020 | 0.037 | 0.025 | 63.83 | 0.018 |
|  |  |  | CV | 0.075 | 0.034 | 0.053 | 0.040 | 54.39 | 0.012 |
|  |  | Grid | Best | 0.045 | 0.010 | 0.026 | 0.019 | 77.21 | 0.015 |
|  |  |  | CV | 0.057 | 0.017 | 0.031 | 0.023 | 71.13 | 0.012 |
| Bounded | M1 | Sample | Best | 0.016 | 0.005 | 0.009 | 0.009 | 69.52 | 0.003 |
|  |  |  | CV | 0.023 | 0.011 | 0.015 | 0.015 | 53.87 | -0.003 |
|  |  | Grid | Best | 0.036 | 0.017 | 0.023 | 0.022 | 54.15 | 0.009 |
|  |  |  | CV | 0.075 | 0.043 | 0.050 | 0.050 | 42.12 | 0.005 |
|  | M2 | Sample | Best | 0.026 | 0.006 | 0.011 | 0.010 | 78.31 | -0.024 |
|  |  |  | CV | 0.030 | 0.015 | 0.020 | 0.018 | 49.71 | -0.027 |
|  |  | Grid | Best | 0.084 | 0.018 | 0.028 | 0.025 | 77.95 | -0.021 |
|  |  |  | CV | 0.100 | 0.055 | 0.064 | 0.059 | 44.75 | -0.016 |
|  | M3 | Sample | Best | 0.025 | 0.006 | 0.011 | 0.011 | 78.40 | 0.000 |
|  |  |  | CV | 0.037 | 0.014 | 0.022 | 0.019 | 63.38 | 0.010 |
|  |  | Grid | Best | 0.073 | 0.017 | 0.024 | 0.023 | 76.96 | 0.006 |
|  |  |  | CV | 0.107 | 0.048 | 0.060 | 0.061 | 55.27 | 0.004 |
|  | M4 | Sample | Best | 0.016 | 0.005 | 0.009 | 0.010 | 69.74 | 0.002 |
|  |  |  | CV | 0.027 | 0.011 | 0.017 | 0.015 | 59.30 | 0.001 |
|  |  | Grid | Best | 0.038 | 0.016 | 0.021 | 0.021 | 59.45 | 0.011 |
|  |  |  | CV | 0.063 | 0.042 | 0.048 | 0.044 | 33.36 | 0.011 |

Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.005 and 0.007 .

Table 6. Simulated MISE (21) and IMP (22) when the external dataset contains both $X$ and $Z$, with $S=200$ under $\gamma=0, n / N \approx 50 \%$.

| Covariate | Model | Test data | $b, 1$ | Estimator |  |  |  | IMP \% | Mean of $\hat{\gamma}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{E 1}$ | $\widehat{\mu}_{1}^{E 2}$ | $\widehat{\mu}_{1}^{E 3}$ |  |  |
| Normal | M1 | Sample | Best | 0.027 | 0.017 | 0.026 | 0.017 | 36.71 | 0.024 |
|  |  |  | CV | 0.033 | 0.020 | 0.024 | 0.020 | 39.59 | 0.023 |
|  |  | Grid | Best | 0.017 | 0.007 | 0.013 | 0.008 | 57.00 | 0.021 |
|  |  |  | CV | 0.019 | 0.010 | 0.014 | 0.011 | 46.34 | 0.023 |
|  | M2 | Sample | Best | 0.041 | 0.022 | 0.025 | 0.023 | 46.91 | -0.089 |
|  |  |  | CV | 0.043 | 0.023 | 0.025 | 0.024 | 46.51 | -0.093 |
|  |  | Grid | Best | 0.022 | 0.013 | 0.016 | 0.014 | 42.47 | -0.095 |
|  |  |  | CV | 0.027 | 0.013 | 0.016 | 0.015 | 49.76 | -0.090 |
|  | M3 | Sample | Best | 0.039 | 0.022 | 0.025 | 0.023 | 43.14 | -0.034 |
|  |  |  | CV | 0.042 | 0.022 | 0.025 | 0.023 | 48.01 | -0.026 |
|  |  | Grid | Best | 0.023 | 0.013 | 0.016 | 0.014 | 43.63 | -0.032 |
|  |  |  | CV | 0.024 | 0.013 | 0.015 | 0.014 | 44.94 | -0.028 |
|  | M4 | Sample | Best | 0.029 | 0.017 | 0.026 | 0.018 | 40.57 | 0.014 |
|  |  |  | CV | 0.039 | 0.021 | 0.026 | 0.022 | 45.54 | 0.013 |
|  |  | Grid | Best | 0.018 | 0.008 | 0.013 | 0.009 | 54.08 | 0.005 |
|  |  |  | CV | 0.020 | 0.011 | 0.014 | 0.011 | 46.79 | 0.015 |
| Bounded | M1 | Sample | Best | 0.006 | 0.004 | 0.005 | 0.005 | 31.24 | 0.007 |
|  |  |  | CV | 0.008 | 0.005 | 0.006 | 0.006 | 28.27 | 0.008 |
|  |  | Grid | Best | 0.019 | 0.012 | 0.012 | 0.012 | 39.18 | -0.003 |
|  |  |  | CV | 0.025 | 0.017 | 0.018 | 0.017 | 30.05 | 0.004 |
|  | M2 | Sample | Best | 0.007 | 0.005 | 0.006 | 0.006 | 32.97 | -0.035 |
|  |  |  | CV | 0.012 | 0.012 | 0.013 | 0.012 | 2.38 | -0.040 |
|  |  | Grid | Best | 0.021 | 0.013 | 0.015 | 0.014 | 38.04 | -0.036 |
|  |  |  | CV | 0.042 | 0.036 | 0.040 | 0.036 | 14.28 | -0.030 |
|  | M3 | Sample | Best | 0.007 | 0.005 | 0.006 | 0.005 | 31.15 | -0.006 |
|  |  |  | CV | 0.010 | 0.010 | 0.011 | 0.010 | 6.03 | -0.010 |
|  |  | Grid | Best | 0.022 | 0.013 | 0.015 | 0.015 | 37.83 | -0.007 |
|  |  |  | CV | 0.032 | 0.027 | 0.030 | 0.029 | 13.81 | -0.002 |
|  | M4 | Sample | Best | 0.007 | 0.004 | 0.005 | 0.005 | 32.99 | -0.004 |
|  |  |  | CV | 0.008 | 0.006 | 0.007 | 0.006 | 26.81 | -0.004 |
|  |  | Grid | Best | 0.019 | 0.012 | 0.012 | 0.013 | 36.25 | 0.007 |
|  |  |  | CV | 0.025 | 0.017 | 0.019 | 0.018 | 32.72 | 0.001 |

Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.004 and 0.006 .

Table 7. Simulated MISE (21) and IMP (22) when the external dataset contains both $X$ and $Z$, with $S=200$ under $\gamma=0.5, n / N \approx$ 13\%.

| Covariate | Model | Test data | $b, 1$ | Estimator |  |  |  | IMP \% | Mean of $\hat{\gamma}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{E 1}$ | $\widehat{\mu}_{1}^{E 2}$ | $\widehat{\mu}_{1}^{E 3}$ |  |  |
|  | M1 | Sample | Best | 0.055 | 0.180 | 0.060 | 0.034 | 37.57 | 0.436 |
|  |  |  | CV | 0.064 | 0.184 | 0.064 | 0.039 | 38.64 | 0.434 |
|  |  | Grid | Best | 0.040 | 0.177 | 0.035 | 0.018 | 54.10 | 0.431 |
|  |  |  | CV | 0.053 | 0.196 | 0.045 | 0.022 | 59.51 | 0.432 |
|  | M2 | Sample | Best | 0.076 | 0.170 | 0.064 | 0.041 | 46.31 | 0.380 |
|  |  |  | CV | 0.092 | 0.177 | 0.071 | 0.045 | 51.02 | 0.377 |
|  |  | Grid | Best | 0.071 | 0.179 | 0.050 | 0.032 | 54.78 | 0.386 |
|  |  |  | CV | 0.082 | 0.190 | 0.056 | 0.038 | 53.31 | 0.385 |
|  | M3 | Sample | Best | 0.072 | 0.188 | 0.064 | 0.043 | 39.63 | 0.425 |
|  |  |  | CV | 0.091 | 0.192 | 0.067 | 0.049 | 45.96 | 0.421 |
|  |  | Grid | Best | 0.053 | 0.191 | 0.042 | 0.033 | 38.08 | 0.436 |
|  |  |  | CV | 0.074 | 0.202 | 0.048 | 0.035 | 52.56 | 0.425 |
|  | M4 | Sample | Best | 0.060 | 0.187 | 0.066 | 0.033 | 44.90 | 0.435 |
|  |  |  | CV | 0.070 | 0.197 | 0.070 | 0.040 | 42.84 | 0.427 |
|  |  | Grid | Best | 0.041 | 0.180 | 0.037 | 0.018 | 57.34 | 0.421 |
|  |  |  | CV | 0.059 | 0.197 | 0.041 | 0.021 | 64.72 | 0.430 |
| Bounded | M1 | Sample | Best | 0.016 | 0.155 | 0.019 | 0.010 | 39.49 | 0.478 |
|  |  |  | CV | 0.022 | 0.160 | 0.022 | 0.013 | 41.00 | 0.494 |
|  |  | Grid | Best | 0.037 | 0.193 | 0.036 | 0.021 | 44.14 | 0.478 |
|  |  |  | CV | 0.061 | 0.212 | 0.047 | 0.037 | 38.34 | 0.489 |
|  | M2 | Sample |  |  | 0.162 | 0.022 | 0.015 | 48.90 | 0.460 |
|  |  |  | CV | 0.036 | 0.164 | 0.031 | 0.022 | 37.67 | 0.464 |
|  |  | Grid | Best | 0.087 | 0.202 | 0.040 | 0.049 | 53.27 | 0.473 |
|  |  |  | CV | 0.105 | 0.228 | 0.086 | 0.083 | 20.55 | 0.471 |
|  | M3 | Sample | Best | 0.029 | 0.171 | 0.020 | 0.015 | 48.81 | 0.485 |
|  |  |  | CV | 0.039 | 0.174 | 0.029 | 0.022 | 43.49 | 0.492 |
|  |  | Grid | Best | 0.076 | 0.205 | 0.036 | 0.045 | 52.75 | 0.490 |
|  |  |  | CV | 0.107 | 0.240 | 0.076 | 0.073 | 31.97 | 0.492 |
|  | M4 | Sample | Best | 0.016 | 0.163 | 0.019 | 0.009 | 39.27 | 0.479 |
|  |  |  | CV | 0.025 | 0.169 | 0.025 | 0.015 | 37.49 | 0.479 |
|  |  | Grid | Best | 0.039 | 0.193 | 0.034 | 0.023 | 41.60 | 0.484 |
|  |  |  | CV | 0.072 | 0.230 | 0.062 | 0.052 | 27.74 | 0.483 |

[^0]Table 8. Simulated MISE (21) and IMP (22) when the external dataset contains both $X$ and $Z$, with $S=200$ under $\gamma=0.5, n / N \approx$ 50\%.

| Covariate | Model | Test data | $b, 1$ | Estimator |  |  |  | IMP \% |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{E 1}$ | $\widehat{\mu}_{1}^{E 2}$ | $\widehat{\mu}_{1}^{E 3}$ |  | Mean of $\hat{\gamma}$ |
| Normal | M1 | Sample | Best | 0.027 | 0.061 | 0.032 | 0.019 | 29.74 | 0.430 |
|  |  |  | CV | 0.034 | 0.061 | 0.033 | 0.022 | 36.98 | 0.422 |
|  |  | Grid | Best | 0.015 | 0.041 | 0.015 | 0.008 | 47.14 | 0.423 |
|  |  |  | CV | 0.018 | 0.049 | 0.018 | 0.010 | 45.47 | 0.417 |
|  | M2 | Sample | Best | 0.043 | 0.060 | 0.033 | 0.027 | 36.46 | 0.353 |
|  |  |  | CV | 0.044 | 0.062 | 0.035 | 0.030 | 33.60 | 0.360 |
|  |  | Grid | Best | 0.023 | 0.053 | 0.022 | 0.016 | 27.23 | 0.367 |
|  |  |  | CV | 0.024 | 0.052 | 0.022 | 0.016 | 31.83 | 0.353 |
|  | M3 | Sample | Best | 0.043 | 0.068 | 0.031 | 0.029 | 32.86 | 0.417 |
|  |  |  | CV | 0.042 | 0.067 | 0.030 | 0.029 | 30.68 | 0.424 |
|  |  | Grid | Best | 0.022 | 0.058 | 0.018 | 0.015 | 30.01 | 0.416 |
|  |  |  | CV | 0.023 | 0.059 | 0.019 | 0.017 | 27.17 | 0.417 |
|  | M4 | Sample | Best | 0.033 | 0.062 | 0.034 | 0.022 | 34.75 | 0.413 |
|  |  |  | CV | 0.039 | 0.064 | 0.033 | 0.023 | 40.31 | 0.414 |
|  |  | Grid | Best | 0.017 | 0.046 | 0.017 | 0.010 | 42.46 | 0.409 |
|  |  |  | CV | 0.020 | 0.056 | 0.018 | 0.011 | 44.86 | 0.421 |
| Bounded | M1 | Sample | Best | 0.006 | 0.028 | 0.007 | 0.006 | 8.96 | 0.481 |
|  |  |  | CV | 0.007 | 0.029 | 0.008 | 0.006 | 13.22 | 0.482 |
|  |  | Grid | Best | 0.018 | 0.047 | 0.016 | 0.017 | 13.08 | 0.482 |
|  |  |  | CV | 0.023 | 0.054 | 0.023 | 0.021 | 8.56 | 0.478 |
|  | M2 | Sample | Best | 0.008 | 0.033 | 0.008 | 0.006 | 20.59 | 0.465 |
|  |  |  | CV | 0.015 | 0.040 | 0.017 | 0.013 | 14.82 | 0.465 |
|  |  | Grid | Best | 0.023 | 0.059 | 0.021 | 0.020 | 12.38 | 0.460 |
|  |  |  | CV | 0.045 | 0.081 | 0.050 | 0.048 | -6.84 | 0.473 |
|  | M3 | Sample | Best | 0.008 | 0.037 | 0.008 | 0.006 | 16.16 | 0.493 |
|  |  |  | CV | 0.010 | 0.040 | 0.012 | 0.009 | 2.05 | 0.497 |
|  |  | Grid | Best | 0.022 | 0.060 | 0.018 | 0.020 | 17.75 | 0.495 |
|  |  |  | CV | 0.033 | 0.078 | 0.041 | 0.040 | -21.18 | 0.498 |
|  | M4 | Sample | Best | 0.007 | 0.032 | 0.007 | 0.005 | 21.96 | 0.476 |
|  |  |  | CV | 0.008 | 0.032 | 0.008 | 0.006 | 21.82 | 0.489 |
|  |  | Grid | Best | 0.019 | 0.052 | 0.016 | 0.019 | 14.56 | 0.481 |
|  |  |  | CV | 0.023 | 0.059 | 0.022 | 0.023 | 2.61 | 0.485 |

Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.004 and 0.006 .

Table 9. Simulated MISE (21) and IMP (22) when the external dataset contains only normally distributed $\left(X_{1}, X_{2}\right)$, with $S=200$.

| Test data | $\gamma$ | $n / N$ | $b, 1$ | Estimator |  |  |  | IMP \% | Mean of $\hat{\gamma}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | $\widehat{\mu}_{1}$ | $\widehat{\mu}_{1}^{C 1}$ | $\widehat{\mu}_{1}^{C 2}$ | $\widehat{\mu}_{1}^{\text {C3 }}$ |  |  |
| Sample | 0 | 13\% | Best | 0.090 | 0.046 | 0.054 | 0.051 | 49.496 | -0.006 |
|  |  |  | CV | 0.080 | 0.056 | 0.063 | 0.062 | 29.115 | 0.007 |
|  |  | 50\% | Best | 0.027 | 0.020 | 0.023 | 0.020 | 25.701 | -0.006 |
|  |  |  | CV | 0.031 | 0.025 | 0.027 | 0.026 | 20.040 | -0.008 |
|  | 0.5 | 13\% | Best | 0.085 | 0.198 | 0.067 | 0.055 | 35.178 | 0.421 |
|  |  |  | CV | 0.070 | 0.190 | 0.072 | 0.062 | 11.429 | 0.406 |
|  |  | 50\% | Best | 0.025 | 0.049 | 0.025 | 0.021 | 17.000 | 0.389 |
|  |  |  | CV | 0.031 | 0.052 | 0.031 | 0.026 | 15.798 | 0.391 |

Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.004 and 0.006 .

The results are shown in Table 9. Compared with results in Tables $1-4$ for the case of $q=1$, the MISEs in this case are larger due to the fact of having more covariates $(q=2)$. But the relative performances of estimators are the same as those shown in Tables 1-4.

## 6. Discussion

Curse of dimensionality is a well-known problem for nonparametric methods. Thus, the proposed method in Section 2 is intended for low dimensional covariate $\boldsymbol{U}$, i.e., $p$ is small. If $p$ is not small, then we should reduce the dimension of $\boldsymbol{U}$ prior to applying the CK, or any kernel methods. For example, consider a single index model assumption (K.-C. Li, 1991), i.e., $\mu_{1}(\boldsymbol{U})$ in (1) is assumed to be

$$
\begin{equation*}
\mu_{1}(\boldsymbol{U})=\mu_{1}\left(\boldsymbol{\eta}^{\top} \boldsymbol{U}\right) \tag{23}
\end{equation*}
$$

where $\eta$ is an unknown $p$-dimensional vector. The well-known SIR technique (K.-C. Li, 1991) can be applied to obtain a consistent and asymptotically normal estimator $\widehat{\eta}$ of $\boldsymbol{\eta}$ in (23). Once $\boldsymbol{\eta}$ is replaced by $\widehat{\boldsymbol{\eta}}$, the kernel method
can be applied with $\boldsymbol{U}$ replaced by the one-dimensional 'covariate' $\widehat{\boldsymbol{\eta}}^{\top} \boldsymbol{U}$. We can also apply other dimension reduction techniques developed under assumptions weaker than (23) (Cook \& Weisberg, 1991; B. Li \& Wang, 2007; Ma \& Zhu, 2012; Y. Shao et al., 2007; Xia et al., 2002).

We turn to the dimension of $\boldsymbol{X}$ in the external dataset. When the dimension of $\boldsymbol{X}$ is high, we may consider the following approach. Instead of using constraint (15), we use component-wise constraints

$$
\begin{equation*}
\sum_{i=1}^{n}\left\{\mu_{i}-\widehat{h}_{1}^{(k)}\left(X_{i}^{(k)}\right)\right\} \boldsymbol{g}_{k}\left(X_{i}^{(k)}\right)^{\top}=0, \quad k=1, \ldots, q \tag{24}
\end{equation*}
$$

where $X_{i}^{(k)}$ is the $k$ th component of $\boldsymbol{X}_{i}, \boldsymbol{g}_{k}\left(X^{(k)}\right)=\left(1, X^{(k)}\right)^{\top}$, and $\widehat{h}_{1}^{(k)}\left(X_{i}^{(k)}\right)$ is an estimator of $h_{1}^{(k)}\left(X^{(k)}\right)=E(Y \mid$ $X^{(k)}, D=1$ ) using methods described in Section 2. More constraints are involved in (24), but estimation only involves one dimensional $X^{(k)}, k=1, \ldots, q$.

The kernel $\kappa$ we adopted in (2) and (16) is the second-order kernel so that the convergence rate of $\widehat{\mu}_{1}^{E}(\boldsymbol{u})-\mu_{1}(\boldsymbol{u})$ is $n^{-2 /(4+p)}$. An $m$ th-order kernel with $m>2$ as defined by Bierens (1987) may be used to achieve convergence rate $n^{-m /(2 m+p)}$. Alternatively, we may apply other nonparametric smoothing techniques such as the local polynomial (Fan et al., 1997) to achieve convergence rate $n^{-m /(2 m+p)}$ with $m \geq 2$.

Our results can be extended to the scenarios where several external datasets are available. Since each external source may provide different covariate variables, we may need to apply component-wise constraints (24) by estimating $\widehat{h}_{1}^{(k)}$ via combining all the external sources that collects covariate $X^{(k)}$. If populations of external datasets are different, then we may have to apply a combination of the methods described in Section 2.

## Acknowledgments

The authors would like to thank two anonymous referees for helpful comments and suggestions.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Jun Shao's research was partially supported by the National Natural Science Foundation of China [11831008] and the U.S. National Science Foundation [DMS-1914411].

## References

Bierens, H. J. (1987). Kernel estimators of regression functions. In Advances in Econometrics: Fifth World Congress (Vol. 1, pp. 99-144). Cambridge University Press.
Chatterjee, N., Chen, Y. H., Maas, P., \& Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. Journal of the American Statistical Association, 111(513), 107-117. https://doi.org/10.1080/01621459.2015.1123157
Cook, R. D., \& Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. Journal of the American Statistical Association, 86(414), 328-332. https://doi.org/10.2307/2290564
Dai, C.-S., \& Shao, J. (2023). Kernel regression utilizing external information as constraints. Statistica Sinica, 33, in press. https://doi.org/10.5705/ss.202021.0446
Fan, J., Farmen, M., \& Gijbels, I. (1998). Local maximum likelihood estimation and inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(3), 591-608. https://doi.org/10.1111/1467-9868.00142
Fan, J., Gasser, T., Gijbels, I., Brockmann, M., \& Engel, J. (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficiency. Annals of the Institute of Statistical Mathematics, 49(1), 79-99. https://doi.org/10.1023/A:1003162622169
Györfi, L., Kohler, M., Krzyżak, A., \& Walk, H. (2002). A distribution-free theory of nonparametric regression. Springer.
Kim, H. J., Wang, Z., \& Kim, J. K. (2021). Survey data integration for regression analysis using model calibration. arXiv 2107.06448.

Li, B., \& Wang, S. (2007). On directional regression for dimension reduction. Journal of the American Statistical Association, 102(479), 997-1008. https://doi.org/10.1198/016214507000000536
Li, K.-C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414), 316-327. https://doi.org/10.1080/01621459.1991.10475035
Lohr, S. L., \& Raghunathan, T. E. (2017). Combining survey data with other data sources. Statistical Science, 32(2), 293-312. https://doi.org/10.1214/16-STS584
Ma, Y., \& Zhu, L. (2012). A semiparametric approach to dimension reduction. Journal of the American Statistical Association, 107(497), 168-179. https://doi.org/10.1080/01621459.2011.646925
Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. Journal of the American Statistical Association, 99(468), 1131-1139. https://doi.org/10.1198/016214504000000601

Nadaraya, E. A. (1964). On estimating regression. Theory of Probability \& Its Applications, 9(1), 141-142. https://doi.org/10.1137/ 1109020
Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. Econometric Theory, 10(2), 1-21. https://doi.org/10.1017/S0266466600008409.
Newey, W. K., \& McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of Econometrics, 4, $2111-2245$. https://doi.org/10.1016/S1573-4412(05)80005-4
Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. Sankhya B, 83(1), 242-272. https://doi.org/10.1007/s13571-020-00227-w
Shao, J. (2003). Mathematical statistics. 2nd ed., Springer.
Shao, Y., Cook, R. D., \& Weisberg, S. (2007). Marginal tests with sliced average variance estimation. Biometrika, 94(2), $285-296$. https://doi.org/10.1093/biomet/asm021
Wand, M. P., \& Jones, M. C. (1994, December). Kernel smoothing. Number 60 in Chapman \& Hall/CRC Monographs on Statistics \& Applied Probability, Boca Raton.
Wasserman, L. (2006). All of nonparametric statistics. Springer.
Xia, Y., Tong, H., Li, W. K., \& Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 64(3), 363-410. https://doi.org/10.1111/1467-9868.03411
Yang, S., \& Kim, J. K. (2020). Statistical data integration in survey sampling: a review. Japanese Journal of Statistics and Data Science, 3(2), 625-650. https://doi.org/10.1007/s42081-020-00093-w
Zhang, H., Deng, L., Schiffman, M., Qin, J., \& Yu, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. Biometrika, 107(3), 689-703. https://doi.org/10.1093/biomet/asaa014
Zhang, Y., Ouyang, Z., \& Zhao, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. The Annals of Applied Statistics, 11(1), 161-184. https://doi.org/10.1214/16-AOAS998

## Appendix

Proof of Theorem 4.1.: Let $\tilde{\mu}_{1}(\boldsymbol{u})=\widehat{p}(\boldsymbol{u}) \widehat{\mu}_{1}(\boldsymbol{u})+\{1-\widehat{p}(\boldsymbol{u})\} \widehat{\mu}_{0}(\boldsymbol{u})$, where $\widehat{\mu}_{1}(\boldsymbol{u})=\sum_{i=1}^{n} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) Y_{i} / \sum_{i=1}^{n} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)$, $\widehat{\mu}_{0}(\boldsymbol{u})=\sum_{i=n+1}^{N} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \tilde{Y}_{i} / \sum_{i=n+1}^{N} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)$, and $\widehat{p}(\boldsymbol{u})=\sum_{i=1}^{n} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) / \sum_{i=1}^{N} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)$. Under (B3)-(B4), Theorem 2 in Nadaraya (1964) shows that $\widehat{p}(\boldsymbol{u})$ converges to $P(D=1 \mid \boldsymbol{U}=\boldsymbol{u})$ in probability. Under (B1)-(B4), $\sqrt{n b^{p}}\left\{\widehat{\mu}_{1}(\boldsymbol{u})-\right.$ $\left.\mu_{1}(\boldsymbol{u})\right\} \xrightarrow{d} N\left(B_{1}(\boldsymbol{u}), V_{1}(\boldsymbol{u})\right), B_{1}(\boldsymbol{u})=c^{1 / 2} A_{1}(\boldsymbol{u}), V_{1}(\boldsymbol{u})=\frac{\sigma_{1}^{2}(\boldsymbol{u})}{f_{1}(\boldsymbol{u})} \int \kappa(\boldsymbol{v})^{2} \mathrm{~d} \boldsymbol{v}$, and $\sqrt{n /(N-n)} \sqrt{(N-n) b^{p}}\left\{\widehat{\mu}_{0}(\boldsymbol{u})-\mu_{1}(\boldsymbol{u})\right\} \xrightarrow{d}$ $N\left(B_{0}(\boldsymbol{u}), V_{0}(\boldsymbol{u})\right), B_{0}(\boldsymbol{u})=c^{1 / 2} A_{0}(\boldsymbol{u}), V_{0}(\boldsymbol{u})=\frac{\sigma_{0}^{2}(\boldsymbol{u})}{a f_{0}(\boldsymbol{u})} \int \kappa(\boldsymbol{v})^{2} \mathrm{~d} \boldsymbol{v}$. Then (17) holds for $\tilde{\mu}_{1}(\boldsymbol{u})$, by Slutsky's theorem, the independence between $\widehat{\mu}_{1}$ and $\widehat{\mu}_{0}$, and the definition of $a$. The desired result (17) follows from the fact that $\left|\widehat{\mu}_{1}^{E 2}(\boldsymbol{u})-\tilde{\mu}_{1}(\boldsymbol{u})\right|$ is bounded by

$$
\begin{equation*}
\{1-\widehat{p}(\boldsymbol{u})\} \max _{i>n}\left|\frac{\widehat{f}\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=1\right)}{\widehat{f}\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=0\right)}-\frac{f\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=1\right)}{f\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=0\right)}\right| \frac{\sum_{i=n+1}^{N}\left|Y_{i}\right| \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)}{\sum_{i=n+1}^{N} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)} \tag{A1}
\end{equation*}
$$

which is $o_{p}\left(1 / \sqrt{n b^{p}}\right)$ by result (18) under condition (B5).

Proof of Theorem 4.2.: Write

$$
\begin{equation*}
\sqrt{n b^{p}}\left\{\widehat{\mu}_{1}^{C 2}(\boldsymbol{u})-\mu_{1}(\boldsymbol{u})\right\}=T_{1}+\cdots+T_{6} \tag{A2}
\end{equation*}
$$

where $T_{1}=n^{-1 / 2} b^{p / 2} \boldsymbol{\delta}_{b}(\boldsymbol{u})^{\top}\left(\boldsymbol{I}_{n}-\boldsymbol{P}\right) \boldsymbol{B}_{l}^{-1} \boldsymbol{\Delta}_{l} \boldsymbol{\epsilon} / \widehat{f}_{b}(\boldsymbol{u}), T_{2}=n^{-1 / 2} b^{p / 2} \boldsymbol{\delta}_{b}(\boldsymbol{u})^{\top}\left\{\boldsymbol{\mu}_{1}-\mu_{1}(\boldsymbol{u}) \mathbf{1}_{n}\right\} / \widehat{f}_{b}(\boldsymbol{u}), T_{3}=n^{-1 / 2} b^{p / 2} \boldsymbol{\delta}_{b}(\boldsymbol{u})^{\top}\left(\boldsymbol{B}_{l}^{-1}\right.$ $\left.\boldsymbol{\Delta}_{l} \boldsymbol{\mu}_{1}-\boldsymbol{\mu}_{1}\right) / \widehat{f_{b}}(\boldsymbol{u}), T_{4}=-n^{-1 / 2} b^{p / 2} \boldsymbol{\delta}_{b}(\boldsymbol{u})^{\top} \boldsymbol{P}\left(\boldsymbol{B}_{l}^{-1} \boldsymbol{\Delta}_{l} \boldsymbol{\mu}_{1}-\boldsymbol{\mu}_{1}\right) / \widehat{f_{b}}(\boldsymbol{u}), T_{5}=n^{-1 / 2} b^{p / 2} \boldsymbol{\delta}_{b}(\boldsymbol{u})^{\top} \boldsymbol{P}\left(\widehat{\boldsymbol{h}}_{1}-\boldsymbol{h}_{1}\right) / \widehat{f}_{b}(\boldsymbol{u}), T_{6}=n^{-1 / 2} b^{p / 2}$ $\boldsymbol{\delta}_{b}(\boldsymbol{u})^{\top} \boldsymbol{P}\left(\boldsymbol{h}_{1}-\boldsymbol{\mu}_{1}\right) / \widehat{f}_{b}(\boldsymbol{u}), \widehat{f}_{b}(\boldsymbol{u})=\sum_{i=1}^{n} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) / n, \boldsymbol{\delta}_{b}(\boldsymbol{u})=\left(\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{1}\right), \ldots, \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{n}\right)\right)^{\top}, \boldsymbol{I}_{n}$ is the identity matrix of order $n, \mathbf{1}_{n}$ is the $n$-vector with all components being $1, \boldsymbol{B}_{l}$ is the $n \times n$ diagonal matrix whose $i$ th diagonal element is $\widehat{f}_{l}\left(\boldsymbol{U}_{i}\right)$, $\boldsymbol{\Delta}_{l}$ is the $n \times n$ matrix whose $(i, j)$ th entry is $\kappa_{l}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right) / n, \boldsymbol{\epsilon}=\left(\epsilon_{1}, \ldots, \epsilon_{n}\right)^{\top}$ with $\epsilon_{i}=Y_{i}-\mu_{1}\left(\boldsymbol{U}_{i}\right)$, $\boldsymbol{h}_{1}$ is the $n$-dimensional vector whose $i$ th component is $h_{1}\left(\boldsymbol{X}_{i}\right), \boldsymbol{P}=\boldsymbol{G}\left(\boldsymbol{G}^{\top} \boldsymbol{G}\right)^{-1} \boldsymbol{G}^{\top}$, and $\boldsymbol{G}, \widehat{\boldsymbol{h}}_{1}$, and $\boldsymbol{\mu}_{1}$ are defined in Section 2.

We first show that $T_{1}$ in (A2) is asymptotically normal with mean 0 and variance $V_{r}(\boldsymbol{u})$ defined in Theorem 4.2. Consider a further decomposition $T_{1}=\sqrt{n} V+T_{11}+T_{12}+T_{13}$, where

$$
V=\frac{1}{n^{2}} \sum_{j=1}^{n} \sum_{i=1}^{n} S\left(\boldsymbol{U}_{i}, \epsilon_{i}, \boldsymbol{U}_{j}, \epsilon_{j}\right)
$$

is a V-statistic with

$$
\begin{aligned}
S\left(\boldsymbol{U}_{i}, \epsilon_{i}, \boldsymbol{U}_{j}, \epsilon_{j}\right) & =\frac{b^{p / 2}}{2 f_{1}(\boldsymbol{u})}\left\{\frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \kappa_{l}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right) \epsilon_{j}}{f_{1}\left(\boldsymbol{U}_{i}\right)}+\frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{j}\right) \kappa_{l}\left(\boldsymbol{U}_{j}-\boldsymbol{U}_{i}\right) \epsilon_{i}}{f_{1}\left(\boldsymbol{U}_{j}\right)}\right\}, \\
T_{11} & =\frac{b^{p / 2}}{n^{3 / 2}} \sum_{i=1}^{n} \frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \kappa_{l}(0) \epsilon_{i}}{f_{1}(\boldsymbol{u}) f_{1}\left(\boldsymbol{U}_{i}\right)} \\
T_{12} & =\frac{b^{p / 2}}{n^{3 / 2}} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \kappa_{l}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right)}{f_{1}(\boldsymbol{u}) f_{1}\left(\boldsymbol{U}_{i}\right)}\left\{\frac{f_{1}(\boldsymbol{u}) f_{1}\left(\boldsymbol{U}_{i}\right)}{\left.\widehat{f_{b}(\boldsymbol{u}) \widehat{f}_{l}\left(\boldsymbol{U}_{i}\right)}-1\right\} \epsilon_{j}}\right.
\end{aligned}
$$

and $T_{13}=-n^{-1 / 2} b^{p / 2} \boldsymbol{\delta}_{b}(\boldsymbol{u})^{\top} \boldsymbol{P} \boldsymbol{B}_{l}^{-1} \boldsymbol{\Delta}_{l} \boldsymbol{\epsilon} / \widehat{f}_{b}(\boldsymbol{u})$. Note that

$$
S_{1}\left(\boldsymbol{U}_{1}, \epsilon_{1}\right)=E\left\{S\left(\boldsymbol{U}_{1}, \epsilon_{1}, \boldsymbol{U}_{2}, \epsilon_{2}\right) \mid \boldsymbol{U}_{1}, \epsilon_{1}\right\}=\frac{b^{p / 2}}{2 f_{1}(\boldsymbol{u})}\left\{\int \kappa_{l}\left(\boldsymbol{u}_{2}-\boldsymbol{U}_{1}\right) \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{u}_{2}\right) \mathrm{d} \boldsymbol{u}_{2}\right\} \epsilon_{1}
$$

having variance

$$
\begin{aligned}
\operatorname{Var}\left\{S_{1}\left(\boldsymbol{U}_{1}, \epsilon_{1}\right)\right\} & =\frac{b^{p / 2}}{4 f_{1}^{2}(\boldsymbol{u})} \int f_{1}\left(\boldsymbol{u}_{1}\right) \sigma_{1}^{2}\left(\boldsymbol{u}_{1}\right)\left\{\int \kappa_{l}\left(\boldsymbol{u}_{2}-\boldsymbol{u}_{1}\right) \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{u}_{2}\right) \mathrm{d} \boldsymbol{u}_{2}\right\}^{2} d \boldsymbol{u}_{1} \\
& =\frac{b^{p / 2}}{4 f_{1}^{2}(\boldsymbol{u})} \int f_{1}\left(\boldsymbol{u}_{1}\right) \sigma_{1}^{2}\left(\boldsymbol{u}_{1}\right)\left\{\int \kappa_{l}(\boldsymbol{v}) \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{u}_{1}-l \boldsymbol{v}\right) \mathrm{d} \boldsymbol{v}\right\}^{2} \mathrm{~d} \boldsymbol{u}_{1} \\
& =\frac{1}{4 f_{1}^{2}(\boldsymbol{u})} \int f_{1}(\boldsymbol{u}-b \boldsymbol{w}) \sigma_{1}^{2}(\boldsymbol{u}-b \boldsymbol{w})\left\{\int \kappa(\boldsymbol{v}) \kappa\left(\boldsymbol{w}-\boldsymbol{v} \frac{l}{b}\right) \mathrm{d} \boldsymbol{v}\right\}^{2} \mathrm{~d} \boldsymbol{w},
\end{aligned}
$$

where $\sigma_{1}^{2}(\cdot)$ is given in condition (C2), the second and third equalities follow from changing variables $\boldsymbol{u}_{2}-\boldsymbol{u}_{1}=\boldsymbol{l} \boldsymbol{v}$ and $\boldsymbol{u}-$ $\boldsymbol{u}_{1}=b \boldsymbol{w}$, respectively. From the continuity of $f_{1}(\cdot)$ and $\sigma_{1}^{2}(\cdot), \operatorname{Var}\left\{S_{1}\left(\boldsymbol{u}_{1}, \epsilon_{1}\right)\right\}$ converges to $V_{r}(\boldsymbol{u})$. Therefore, by the theory for asymptotic normality of V-statistics (e.g., Theorem 3.16 in J. Shao, 2003), $\sqrt{n} V \xrightarrow{d} N\left(0, V_{r}(\boldsymbol{u})\right)$.

Conditioned on $\boldsymbol{U}_{1}, \ldots, \boldsymbol{U}_{n}, T_{11}$ has mean 0 and variance

$$
\begin{aligned}
\operatorname{Var}\left(T_{11} \mid \boldsymbol{U}_{1}, \ldots, \boldsymbol{U}_{n}\right) & =\frac{b^{p}}{4 f_{1}^{2}(\boldsymbol{u}) n^{3}} \sum_{i=1}^{n} \frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)^{2} \kappa_{l}(0)^{2} \sigma_{1}^{2}\left(\boldsymbol{U}_{i}\right)}{f_{1}\left(\boldsymbol{U}_{i}\right)} \\
& \leq \frac{\sup _{\boldsymbol{u} \in \mathscr{U}} \kappa(\boldsymbol{u})^{3}}{4 f_{1}^{2}(\boldsymbol{u}) n^{3} b^{2 p}} \sum_{i=1}^{n} \frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \sigma_{1}^{2}\left(\boldsymbol{U}_{i}\right)}{f_{1}\left(\boldsymbol{U}_{i}\right)}=o_{p}(1) .
\end{aligned}
$$

This proves that $T_{11}=o_{p}(1)$. Note that $E\left(T_{12} \mid \boldsymbol{U}_{1}, \ldots, \boldsymbol{U}_{n}\right)=0$ and $\operatorname{Var}\left(T_{12} \mid \boldsymbol{U}_{1}, \ldots, \boldsymbol{U}_{n}\right)$ is bounded by

$$
\max \left\{\frac{1}{f_{1}^{2}(\boldsymbol{u})}, \frac{1}{\widehat{f_{b}^{2}(\boldsymbol{u})}}\right\} \max _{i=1, \ldots, n}\left|\frac{f_{1}\left(\boldsymbol{U}_{i}\right)}{\widehat{f}_{l}\left(\boldsymbol{U}_{i}\right)}-1\right|^{2} \operatorname{Var}\left(\sqrt{n} V+T_{11} \mid \boldsymbol{U}_{1}, \ldots, \boldsymbol{U}_{n}\right) .
$$

Therefore, under the assumed condition that $f_{1}$ is bounded away from zero, Lemma 3 in Dai and Shao (2023) implies $T_{12}=$ $o_{p}(1)$. Note that

$$
T_{13}=\frac{b^{p / 2}}{n^{1 / 2}} \sum_{j=1}^{n} W_{j}(\boldsymbol{u}) \epsilon_{j}, \quad W_{j}(\boldsymbol{u})=\frac{1}{n} \sum_{i=1}^{n} \frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)^{\top}}{\widehat{f}_{b}(\boldsymbol{u})}\left(\boldsymbol{G}^{\top} \boldsymbol{G}\right)^{-1} \sum_{i=1}^{n} \frac{\kappa_{l}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right) \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)}{\widehat{f}_{l}\left(\boldsymbol{U}_{i}\right)} .
$$

Conditioned on $\boldsymbol{U}_{1}, \ldots, \boldsymbol{U}_{n}, T_{13}$ has mean 0 and variance

$$
\operatorname{Var}\left(T_{13} \mid \boldsymbol{U}_{1}, \ldots, \boldsymbol{U}_{n}\right)=\frac{b^{p}}{n} \sum_{j=1}^{n} W_{j}^{2}(\boldsymbol{u}) \sigma_{1}^{2}\left(\boldsymbol{U}_{j}\right)=O_{p}\left(b^{p}\right)=o_{p}(1),
$$

because, under the assumed condition that $f_{1}$ is bounded away from zero, Lemma 3 in Dai and Shao (2023) implies $\max _{j=1, \ldots, n}\left|W_{j}(\boldsymbol{u})-\boldsymbol{g}(\boldsymbol{u})^{\top} \boldsymbol{\Sigma}_{g}^{-1} \boldsymbol{g}\left(\boldsymbol{X}_{j}\right)\right|=o_{p}(1)$. Thus, $T_{13}=o_{p}(1)$. Consequently, $T_{1}$ has the same asymptotic distribution as $\sqrt{n} V$, the claimed result.

From Lemma 4 in Dai and Shao (2023) and (C4), $T_{2}=\sqrt{c} A_{1}(\boldsymbol{u})\left\{1+o_{p}(1)\right\}$. Note that

$$
\begin{aligned}
T_{3} & =\frac{\sqrt{n b^{p}}{ }^{2}}{n \widehat{f}_{b}(\boldsymbol{u})} \sum_{j=1}^{n} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{j}\right)\left[\frac{1}{n l^{2} \widehat{f}_{b}\left(\boldsymbol{U}_{j}\right)} \sum_{i=1}^{n} \kappa_{l}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)\left\{\mu_{1}\left(\boldsymbol{U}_{i}\right)-\mu_{1}\left(\boldsymbol{U}_{j}\right)\right\}\right] \\
& =\left\{\frac{\sqrt{c} r^{2}}{n \widehat{f}_{b}(\boldsymbol{u})} \sum_{j=1}^{n} \kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{j}\right) A_{1}\left(\boldsymbol{U}_{j}\right)\right\}\left\{1+o_{p}(1)\right\}=\sqrt{c} r^{2} A_{1}(\boldsymbol{u})\left\{1+o_{p}(1)\right\},
\end{aligned}
$$

where the second equality follows from (A4) and Lemmas 3-4 in Dai and Shao (2023), and the last equality follows from Lemma 2 in Dai and Shao (2023) and continuity of $A_{1}(\cdot)$. Also,

$$
\begin{aligned}
-\frac{n^{1 / 2} T_{4}}{b^{p / 2}} & =\frac{1}{n} \sum_{i=1}^{n} \frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)^{\top}}{\widehat{f}_{b}(\boldsymbol{u})}\left(\boldsymbol{G}^{\top} \boldsymbol{G}\right)^{-1} \sum_{j=1}^{n} \frac{\boldsymbol{g}\left(\boldsymbol{X}_{j}\right)}{n \widehat{f}_{b}\left(\boldsymbol{U}_{j}\right)} \sum_{i=1}^{n} \kappa_{l}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)\left\{\mu_{1}\left(\boldsymbol{U}_{i}\right)-\mu_{1}\left(\boldsymbol{U}_{j}\right)\right\} \\
& =\left\{\boldsymbol{g}(\boldsymbol{x})^{\top} \boldsymbol{\Sigma}_{g}^{-1} \frac{1}{n} \sum_{j=1}^{n} \frac{\boldsymbol{g}\left(\boldsymbol{X}_{j}\right)}{\left.\widehat{n \widehat{f}_{b}\left(\boldsymbol{U}_{j}\right)} \sum_{i=1}^{n} \kappa_{l}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right)\left\{\mu_{1}\left(\boldsymbol{U}_{i}\right)-\mu_{1}\left(\boldsymbol{U}_{j}\right)\right\}\right\}\left\{1+o_{p}(1)\right\}}\right. \\
& =\left\{\boldsymbol{g}(\boldsymbol{x})^{\top} \boldsymbol{\Sigma}_{g}^{-1} \frac{l^{2 / p}}{n} \sum_{j=1}^{n} \boldsymbol{g}\left(\boldsymbol{X}_{j}\right) A_{1}\left(\boldsymbol{U}_{j}\right)\right\}\left\{1+o_{p}(1)\right\} \\
& =\sqrt{c} r^{2} \boldsymbol{g}(\boldsymbol{x})^{\top} \boldsymbol{\Sigma}_{g}^{-1} E\left\{\boldsymbol{g}(\boldsymbol{X}) A_{1}(\boldsymbol{U})\right\}\left\{1+o_{p}(1)\right\},
\end{aligned}
$$

where the first equality follows from Lemma 3 in Dai and Shao (2023) and the law of large numbers, the second equality follows from Lemma 4 in Dai and Shao (2023), and the last equality follows from the law of large numbers. Similarly,

$$
\begin{aligned}
\frac{n^{1 / 2} T_{5}}{b^{p / 2}} & =\frac{1}{n} \sum_{i=1}^{n} \frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)^{\top}}{\widehat{f}_{b}(\boldsymbol{u})}\left(\boldsymbol{G}^{\top} \boldsymbol{G}\right)^{-1} \sum_{i=1}^{n} \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)\left\{\widehat{h}_{1}\left(\boldsymbol{X}_{i}\right)-h_{1}\left(\boldsymbol{X}_{i}\right)\right\} \\
& =\left[\boldsymbol{g}(\boldsymbol{x})^{\top} \boldsymbol{\Sigma}_{g}^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)\left\{\widehat{h}_{1}\left(\boldsymbol{X}_{i}\right)-h_{1}\left(\boldsymbol{X}_{i}\right)\right\}\right]\left\{1+o_{p}(1)\right\} \\
& \leq\left\{1+o_{p}(1)\right\} O_{p}(1) \max _{j=1, \ldots, n}\left|\widehat{h}_{1}\left(\boldsymbol{X}_{j}\right)-h_{1}\left(\boldsymbol{X}_{j}\right)\right|,
\end{aligned}
$$

where the second equality follows from Lemma 3 in Dai and Shao (2023). Under (B1)-(B5) with $\boldsymbol{U}$ and $p$ replaced by $\boldsymbol{X}$ and $q$, and (C5), Lemma 8.10 in Newey and McFadden (1994) implies that

$$
\begin{equation*}
\max \left|\widehat{h}_{1}\left(\boldsymbol{X}_{i}\right)-h_{1}\left(\boldsymbol{X}_{1}\right)\right|=O_{p}\left(\sqrt{\log (n)} n^{-2 /(q+4)}\right) \tag{A3}
\end{equation*}
$$

which is $o_{p}\left(1 / \sqrt{n b^{p}}\right)=o_{p}\left(n^{-2 /(p+4)}\right)$ and, hence, $T_{5}=o_{p}(1)$. From Lemma 3 in Dai and Shao (2023) and the Central Limit Theorem,

$$
T_{6}=\frac{b^{p / 2}}{n^{1 / 2}} \sum_{i=1}^{n} \frac{\kappa_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)^{\top}}{\widehat{f}_{b}(\boldsymbol{u})}\left(\boldsymbol{G}^{\top} \boldsymbol{G}\right)^{-1} \sum_{i=1}^{n} \boldsymbol{g}\left(\boldsymbol{X}_{i}\right)\left\{h_{1}\left(\boldsymbol{X}_{i}\right)-\mu_{1}\left(\boldsymbol{U}_{i}\right)\right\}=O_{p}\left(b^{p / 2}\right)=o_{p}(1) .
$$

Combining these results, we obtain that $T_{2}+\cdots+T_{6}=B_{r}(\boldsymbol{u})+o_{p}(1)$. This completes the proof.
Proof of Theorem 4.3.: Define

$$
\begin{aligned}
& \widehat{\omega}_{t 1}=\frac{1}{N} \sum_{i=1}^{N} R_{i} \check{\kappa}_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) \mathrm{e}^{t Y_{i}}, \quad \widehat{\omega}_{t 2}=\frac{1}{N} \sum_{i=1}^{N}\left(1-R_{i}\right) \check{\kappa}_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) Y_{i} \mathrm{e}^{-t Y_{i}}, \\
& \widehat{\omega}_{t 3}=\frac{1}{N} \sum_{i=1}^{N} R_{i} \check{\kappa}_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right), \quad \widehat{\omega}_{t 4}=\frac{1}{N} \sum_{i=1}^{N}\left(1-R_{i}\right) \check{\kappa}_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right), \\
& \widehat{\omega}_{t 5}=\frac{1}{N} \sum_{i=1}^{N} R_{i} \check{\kappa}_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) Y_{i} \mathrm{e}^{t Y_{i}}, \quad \widehat{\omega}_{t 6}=\frac{1}{N} \sum_{i=1}^{N}\left(1-R_{i}\right) \check{\kappa}_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) Y_{i}^{2} \mathrm{e}^{-t Y_{i}}, \\
& \widehat{\omega}_{t 7}=\frac{1}{N} \sum_{i=1}^{N} R_{i} \check{\kappa}_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) Y_{i}^{2} \mathrm{e}^{t Y_{i}}, \quad \widehat{\omega}_{t 8}=\frac{1}{N} \sum_{i=1}^{N}\left(1-R_{i}\right) \check{\kappa}_{b}\left(\boldsymbol{u}-\boldsymbol{U}_{i}\right) Y_{i}^{3} \mathrm{e}^{-t Y_{i}} .
\end{aligned}
$$

Then, $\widehat{h}(\boldsymbol{u}, t)=\widehat{\omega}_{t 1} \widehat{\omega}_{t 2} / \widehat{\omega}_{t 3} \widehat{\omega}_{t 4}, \nabla_{t} \widehat{h}(\boldsymbol{u}, t)=\left(\widehat{\omega}_{t 2} \widehat{\omega}_{t 5}-\widehat{\omega}_{t 1} \widehat{\omega}_{t 6}\right) / \widehat{\omega}_{t 3} \widehat{\omega}_{t 4}$, and $\nabla_{t}^{2} \widehat{h}(\boldsymbol{u}, t)=\left(\widehat{\omega}_{t 1} \widehat{\omega}_{t 8}-2 \widehat{\omega}_{t 5} \widehat{\omega}_{t 6}+\widehat{\omega}_{t 2} \widehat{\omega}_{t 7}\right) / \widehat{\omega}_{t 3} \widehat{\omega}_{t 4}$. Let $L(t)=E\left[R\{Y-h(\boldsymbol{U}, t)\}^{2}\right], \widehat{L}_{n}(t)=N^{-1} \sum_{i=1}^{N} R_{i}\left\{Y_{i}-\widehat{h}\left(\boldsymbol{U}_{i}, t\right)\right\}^{2}$, and $L_{n}(t)=N^{-1} \sum_{i=1}^{N} R_{i}\left\{Y_{i}-h\left(\boldsymbol{U}_{i}, t\right)\right\}^{2}$. Taking derivatives with respect to $t$, we obtain

$$
\begin{aligned}
& \nabla_{t} \widehat{L}_{n}(t)=\frac{1}{N} \sum_{i=1}^{N}-2 R_{i}\left\{Y_{i}-\widehat{h}\left(\boldsymbol{U}_{i}, t\right)\right\} \nabla_{t} \widehat{h}\left(\boldsymbol{U}_{i}, t\right)=\frac{1}{N} \sum_{i=1}^{N} \psi\left\{Y_{i}, R_{i}, \widehat{\boldsymbol{\omega}}_{t}\left(\boldsymbol{U}_{i}\right)\right\}, \\
& \nabla_{t} L_{n}(t)=\frac{1}{N} \sum_{i=1}^{N}-2 R_{i}\left\{Y_{i}-h\left(\boldsymbol{U}_{i}, t\right)\right\} \nabla_{t} h\left(\boldsymbol{U}_{i}, t\right)=\frac{1}{N} \sum_{i=1}^{N} \psi\left\{Y_{i}, R_{i}, \boldsymbol{\omega}_{t}\left(\boldsymbol{U}_{i}\right)\right\},
\end{aligned}
$$

and

$$
\nabla_{t} L(t)=-2 E\left[R\{Y-h(\boldsymbol{u}, t)\} \nabla_{t} h(\boldsymbol{u}, t)\right]=\mathrm{E}\left[\psi\left\{Y, R, \boldsymbol{\omega}_{t}(\boldsymbol{U})\right\}\right],
$$

where $\psi$ is given in (D5). Note that $\nabla_{t} L(\gamma)=0$ and $\nabla_{t}^{2} L(\gamma)=2 E\left[\left\{\nabla_{t} h(\boldsymbol{U}, \gamma)\right\}^{2} R\right]=v_{\gamma} \geq 0$. We establish the asymptotic normality of $\widehat{\gamma}$ in the following four steps.

Step 1: Since $\gamma$ is the unique minimizer of $L(t)$, from Theorem 2.1 in Newey and McFadden (1994), it suffices to prove that $\sup _{t \in \Gamma}\left|\nabla_{t} \widehat{L}_{n}(t)-\nabla_{t} L(t)\right| \xrightarrow{p} 0$. Note that

$$
\begin{aligned}
\sup _{t \in \Gamma}\left|\nabla_{t} \widehat{L}_{n}(t)-\nabla_{t} L(t)\right| \leq & \sup _{t \in \Gamma}\left|\nabla_{t} L_{n}(t)-\nabla_{t} L(t)\right|+\sup _{t \in \Gamma}\left|\nabla_{t} \widehat{L}_{n}(t)-\nabla_{t} L_{n}(t)\right| \\
\leq & \sup _{t \in \Gamma}\left|\nabla_{t} L_{n}(t)-\nabla_{t} L(t)\right| \\
& +\frac{2}{n} \sum_{i=1}^{n} R_{i}\left|Y_{i}\right|\left\{\sup _{t \in \Gamma, \boldsymbol{x} \in \mathscr{U}}\left|\nabla_{t} \widehat{h}(\boldsymbol{u}, t)-\nabla_{t} h(\boldsymbol{u}, t)\right|\right. \\
& \left.+\sup _{t \in \Gamma, \boldsymbol{u} \in \mathscr{U}}\left|\widehat{h}(\boldsymbol{u}, t) \nabla_{t} \widehat{h}(\boldsymbol{u}, t)-h(\boldsymbol{u}, t) \nabla_{t} h(\boldsymbol{u}, t)\right|\right\}
\end{aligned}
$$

From (D3), $\left|2 R\{Y-h(\boldsymbol{u}, t)\} \nabla_{t} h(\boldsymbol{U}, t)\right|$ is bounded by $c|Y|$ for a constant $c$ and hence Lemma 2.4 in Newey and McFadden (1994) implies that $\sup _{t \in \Gamma}\left|\nabla_{t} L_{n}(t)-\nabla_{t} L(t)\right|=o_{p}(1)$. Based on Lemma B. 3 in Newey (1994), conditions (D1)-(D4) imply that
$\sup _{\boldsymbol{u} \in \mathscr{U}}\left|\widehat{\boldsymbol{\omega}}_{t}(\boldsymbol{u})-\boldsymbol{\omega}_{t}(\boldsymbol{u})\right| \rightarrow 0$ for all $t \in \Gamma$. As a result, by a similar argument of the proof of Lemma B. 3 in Newey (1994), we obtain that $\sup _{t \in \Gamma, \boldsymbol{u} \in \mathscr{U}}\left|\widehat{\boldsymbol{\omega}}_{t}(\boldsymbol{u})-\boldsymbol{\omega}_{t}(\boldsymbol{u})\right| \xrightarrow{p} 0$. Since $\boldsymbol{\omega}_{t}$ is bounded away from zero and $h(\cdot, t)$ and $\nabla_{t} h(\cdot, t)$ are Lipschitz continuous functions with respect to $\boldsymbol{\omega}_{t}, \sup _{t \in \Gamma, \boldsymbol{u} \in \mathscr{U}}|\widehat{h}(\boldsymbol{u}, t)-h(\boldsymbol{u}, t)| \xrightarrow{p} 0$ and $\sup _{t \in \Gamma, \boldsymbol{u} \in \mathscr{U}}\left|\nabla_{t} \widehat{h}(\boldsymbol{u}, t)-\nabla_{t} h(\boldsymbol{u}, t)\right| \xrightarrow{p} 0$. These results together with the previous inequality implies that $\widehat{\gamma} \xrightarrow{p} \gamma$.

Step 2: Conditions (D1)-(D5) ensure that Lemma 8.11 in Newey and McFadden (1994) holds and hence $\sqrt{N} \nabla_{t} \widehat{L}_{n}(\gamma) \xrightarrow{d}$ $N\left(0, \sigma_{L}^{2}\right)$ with $\sigma_{L}^{2}=\operatorname{Var}\left\{m\left(Y, R, \boldsymbol{U}, \boldsymbol{\omega}_{\gamma}\right)+\tau(Y, R, \boldsymbol{U}, \gamma)\right\}$.

Step 3: Note that $\nabla_{t}^{2} L_{n}(t)=N^{-1} \sum_{i=1}^{N}-2 R_{i}\left\{Y_{i}-h\left(\boldsymbol{U}_{i}, t\right)\right\} \nabla_{t}^{2} h\left(\boldsymbol{U}_{i}, t\right)+2 R_{i}\left\{\nabla_{t} h\left(\boldsymbol{U}_{i}, t\right)\right\}^{2}$ and $\sup _{|t-\gamma| \leq|\hat{\gamma}-\gamma|} \mid \nabla_{t}^{2} \widehat{L}_{n}(t)-$ $\nabla_{t}^{2} L(\gamma) \mid \leq A_{1}+A_{2}+A_{3}$, where $A_{1}=\left|\nabla_{t}^{2} L_{n}(\gamma)-\nabla_{t}^{2} L(\gamma)\right|, A_{2}=\sup _{t \in \Gamma}\left|\nabla_{t}^{2} \widehat{L}_{n}(t)-\nabla_{t}^{2} L_{n}(t)\right|$, and the last term $A_{3}=$ $\sup _{|t-\gamma| \leq|\widehat{\gamma}-\gamma|}\left|\nabla_{t}^{2} L_{n}(t)-\nabla_{t}^{2} L_{n}(\gamma)\right|$. The law of large numbers guarantees that $A_{1}=o_{p}(1)$. A similar argument in Step 1 shows that $A_{2}=o_{p}(1)$. For $A_{3}$, we have

$$
\begin{aligned}
\left|\nabla_{t}^{2} L_{n}(t)-\nabla_{t}^{2} L_{n}(\gamma)\right| \leq & \frac{2}{N} \sum_{i=1}^{N}\left|\left\{\nabla_{t} h\left(\boldsymbol{U}_{i}, t\right)\right\}^{2}-\left\{\nabla_{t} h\left(\boldsymbol{U}_{i}, \gamma\right)\right\}^{2}\right| \\
& +\frac{2}{N} \sum_{i=1}^{N}\left|Y_{i}\right|\left|\nabla_{t}^{2} h\left(\boldsymbol{U}_{i}, t\right)-\nabla_{t}^{2} h\left(\boldsymbol{U}_{i}, \gamma\right)\right| \\
& +\frac{2}{N} \sum_{i=1}^{N}\left|h\left(\boldsymbol{U}_{i}, t\right) \nabla_{t}^{2} h\left(\boldsymbol{U}_{i}, t\right)-h\left(\boldsymbol{U}_{i}, \gamma\right) \nabla_{t}^{2} h\left(\boldsymbol{U}_{i}, \gamma\right)\right| .
\end{aligned}
$$

Under (D3), $h(\cdot, t), \nabla h(\cdot, t)$, and $\nabla h(\cdot, t)$ converge uniformly for all $x$ as $t \rightarrow \gamma$ and, thus, the $A_{3}=o_{p}(1)$ because $\widehat{\gamma} \xrightarrow{p}$. This shows that $\sup _{|t-\gamma| \leq|\widehat{\gamma}-\gamma|}\left|\nabla_{t}^{2} \widehat{L}_{n}(t)-\nabla_{t}^{2} L(\gamma)\right| \xrightarrow{p} 0$.

Step 4: By Taylor's expansion, $\widehat{L}_{n}(\widehat{\gamma})-\widehat{L}_{n}(\gamma)=0-\widehat{L}_{n}(\gamma)=\nabla_{t} \widehat{L}_{n}(\xi)(\widehat{\gamma}-\gamma)$ for some $\xi \in(\gamma, \widehat{\gamma})$. From the results in Steps $1-3, \sqrt{N}(\widehat{\gamma}-\gamma) \xrightarrow{d} N\left(0,\left[2 E\left\{R \nabla_{\gamma} h(\boldsymbol{U}, \gamma)\right\}^{2}\right]^{-1} \sigma_{L}^{2}\right)$. This completes the proof of (20).

Proof of Corollary 4.1.: (i) From Theorem 4.3, (20) shows that $\hat{\gamma}-\gamma=O_{p}(1 / \sqrt{N})$. Furthermore, Lemma 8.10 in Newey and McFadden (1994) shows that

$$
\begin{equation*}
\max _{i}\left|\frac{\mathrm{e}^{-\gamma Y_{i}} \sum_{j=1}^{n} \mathrm{e}^{\gamma Y_{j}} \check{\kappa}_{\breve{b}}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right)}{\sum_{j=1}^{n} \check{\kappa}_{b}\left(\boldsymbol{U}_{i}-\boldsymbol{U}_{j}\right)}-\frac{f\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=1\right)}{f\left(Y_{i} \mid \boldsymbol{U}=\boldsymbol{U}_{i}, D=0\right)}\right|=O_{p}\left(\sqrt{\frac{\log N}{N \check{b}^{p}}}+\check{b}^{d}\right), \tag{A4}
\end{equation*}
$$

which is $o_{p}\left(N^{-2 /(p+4)}\right)=o_{p}(1) / \sqrt{n b^{p}}$ under the assumed condition $d>\max \{(p+4) / 2, p\}$ and $N \check{b}^{2 d} \rightarrow 0$. Since $\widehat{\gamma}-$ $\gamma$ converges faster than (A4), (18) holds. As a result, (17) holds with $\mu_{1}^{E 2}(\boldsymbol{u})$ replaced by $\mu_{1}^{E 3}(\boldsymbol{u})$ under (B1)-(B4) and (D1)-(D5).
(ii) Under (D1)-(D5) with $\boldsymbol{U}$ replaced by $\boldsymbol{X}$ and $p$ replaced by $q$, Lemma 8.10 in Newey and McFadden (1994) implies that

$$
\sup _{x \in \mathbb{X}}\left|\widehat{h}(\boldsymbol{x}, \gamma)-h_{1}(\boldsymbol{x})\right|=O_{p}\left((\log N)^{1 / 2}\left(N \check{b}^{q}\right)^{-1 / 2}+\check{b}^{d}\right)=o_{p}\left(n^{-2 /(p+4)}\right) .
$$

From the asymptotic normality of $\widehat{\gamma}, \widehat{\gamma}-\gamma=O_{p}(1 / \sqrt{N})$, which converges to 0 faster than $\sup _{\boldsymbol{x} \in \mathscr{X}}\left|\widehat{h}(\boldsymbol{x}, \gamma)-h_{1}(\boldsymbol{x})\right| \rightarrow 0$. Hence (A3) holds while $\widehat{h}_{1}$ is estimated by $\widehat{h}(\boldsymbol{X}, \widehat{\gamma})$. Then, the rest of proof of the second claims follows the argument in the proof of Theorem 4.2.


[^0]:    Note: Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.005 and 0.007 .

