Kernel Regression Utilizing Heterogeneous Datasets

Chi-Shian Dai[1] and Jun Shao[2] [1]

[1] *Department of Statistics, University of Wisconsin-Madison*

[2] *School of Statistics, East China Normal University*

Abstract: Data analysis in modern scientific research and practice has shifted from analyzing a single dataset to coupling several datasets. We propose and study a kernel regression method that can handle the challenge of heterogeneous populations and greatly extends the constrained kernel regression (Dai and Shao, 2022) for homogeneous populations of different datasets. 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.

Key words and phrases: 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 administrative datasets. In this paper, we consider nonparametric kernel regression (Bierens, 1987; Wand and Jones, 1994; 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 internal data population,

$$\mu_1(\boldsymbol{u}) = E(Y \mid \boldsymbol{U} = \boldsymbol{u}, D = 1), \tag{1}$$

where $D = 1$ indicates internal population and $\boldsymbol{u}$ is a fixed point in $\mathscr{U}$, the range of $\boldsymbol{U}$. 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 the external dataset 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 2.1, 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 building block for the more complicated

case in Section 2.2 where the external dataset contains fewer measured covariates.

In applications, it is often the case that the external dataset has measured Y and $\boldsymbol{X}$ from each subject, where $\boldsymbol{X}$ is a part of the vector $\boldsymbol{U}$ with some components of $\boldsymbol{U}$ not measured due to the high measurement cost or the progress of new technology and/or new scientific relevance for measuring some components of $\boldsymbol{U}$. With some unmeasured components of $\boldsymbol{U}$, the external datasets 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 (2022) proposes a two-step kernel regression using external information as a constraint to improve kernel regression based on internal data along, following the idea of using constraints in Chatterjee et al. (2016) and 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 that in handling nonignorable missing data problems 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 2.2, we develop a methodology to handle population heterogeneity for internal and external datasets, which extends the procedure in Dai and Shao (2022) to heterogeneous populations and greatly widens its application scope.

Under each scenario, we derive asymptotic normality in Section 3 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 4 to compare finite sample performance of several estimators. Discussions on extensions and handling high dimension covariates are given in Section 5. All technical details are in the Appendix.

Our research fits into a general framework of data integration (Kim et al., 2021; Lohr and Raghunathan, 2017; Merkouris, 2004; Rao, 2021; Yang and Kim, 2020; Zhang et al., 2017; Zieschang, 1990).

2 Methodology

The internal dataset contains observations $(Y_i, \boldsymbol{U}_i)$, $i = 1, ..., n$, 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

$$\widehat{\mu}_1(\boldsymbol{u}) = \sum_{i=1}^{n} Y_i \, \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i), \tag{2}$$

where $\kappa_b(\boldsymbol{a}) = b^{-p}\kappa(\boldsymbol{a}/b)$, $\boldsymbol{a} \in \mathscr{U}$, $\kappa(\boldsymbol{u})$ is a given kernel function, 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 5.

We consider the case with one external dataset, independent of the internal dataset. Extension to multiple external datasets is straightforward and discussed in Section 5.

2.1 The Case of No Unmeasured Covariates

In this subsection we consider the situation where the external dataset contains iid observations $(Y_i, \boldsymbol{U}_i)$, $i = n + 1, ..., N$, from $\mathcal{P}_0$, the external population of $(Y, \boldsymbol{U})$.

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

$$\widehat{\mu}_1^{E1}(\boldsymbol{u}) = \sum_{i=1}^{N} Y_i \, \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^{N} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i), \tag{3}$$

which is obviously more efficient than $\widehat{\mu}_1(\boldsymbol{u})$ in (2) as the sample size is increased to $N > n$. The estimator $\widehat{\mu}_1^{E1}(\boldsymbol{u})$ 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.

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 $f(y|\boldsymbol{u}, D)$ be the conditional density of Y given $\boldsymbol{U} = \boldsymbol{u}$ and $D = 1$ or 0 (for internal or external population). Then

$$\mu_1(\boldsymbol{x}) = E(Y|\boldsymbol{U} = \boldsymbol{u}, D = 1) = E\left\{ Y \frac{f(Y|\boldsymbol{u}, D = 1)}{f(Y|\boldsymbol{u}, D = 0)} \,\middle|\, \boldsymbol{U} = \boldsymbol{u}, D = 0 \right\}. \tag{4}$$

The ratio $f(Y|\boldsymbol{u}, D = 1)/f(Y|\boldsymbol{u}, D = 0)$ 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 $\widehat{f}(y|\boldsymbol{u}, D)$ of $f(y|\boldsymbol{u}, D)$ 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 $\widehat{Y}_i = Y_i \widehat{f}(Y_i|\boldsymbol{U}_i, D = 1)/\widehat{f}(Y_i|\boldsymbol{U}_i, D = 0)$. The resulting kernel estimator is

$$\widehat{\mu}_1^{E2}(\boldsymbol{u}) = \left\{ \sum_{i=1}^{n} Y_i \, \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) + \sum_{i=n+1}^{N} \widehat{Y}_i \, \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \right\} \Big/ \sum_{i=1}^{N} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i). \tag{5}$$

Note that we use internal data $(Y_i, \boldsymbol{U}_i)$, $i = 1, ..., n$, to obtain estimator $\widehat{f}(Y_i|\boldsymbol{U}_i, D = 1)$ and external data $(Y_i, \boldsymbol{U}_i)$, $i = n+1, ..., N$, to construct estimator $\widehat{f}(Y_i|\boldsymbol{U}_i, D = 0)$. Applying kernel estimation, we obtain that

$$\widehat{f}(y|\boldsymbol{U}=\boldsymbol{u},D=1)=\sum_{i=1}^{n}\tilde{\kappa}_{\tilde{b}}(y-Y_i,\boldsymbol{u}-\boldsymbol{U}_i)\Big/\sum_{i=1}^{n}\bar{\kappa}_{\bar{b}}(\boldsymbol{u}-\boldsymbol{U}_i),$$
$$\widehat{f}(y|\boldsymbol{U}=\boldsymbol{u},D=0)=\sum_{i=n+1}^{N}\tilde{\kappa}_{\tilde{b}}(y-Y_i,\boldsymbol{u}-\boldsymbol{U}_i)\Big/\sum_{i=n+1}^{N}\bar{\kappa}_{\bar{b}}(\boldsymbol{u}-\boldsymbol{U}_i), \tag{6}$$

where $\tilde{\kappa}$ and $\bar{\kappa}$ are kernels with 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 1 of Section 3.

If additional information exists, then this approach can be improved. Assume that the internal and external datasets are formed according to a random binary indicator D such that $(Y_i,\boldsymbol{U}_i,D_i)$, $i=1,...,N$, 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, this assumption is not substantial. From the identity

$$\frac{f(Y|\boldsymbol{u},D=1)}{f(Y|\boldsymbol{u},D=0)}=\frac{P(D=1|\boldsymbol{U}=\boldsymbol{u},Y)}{P(D=0|\boldsymbol{U}=\boldsymbol{u},Y)}\frac{P(D=0|\boldsymbol{U}=\boldsymbol{u})}{P(D=1|\boldsymbol{U}=\boldsymbol{u})}. \tag{7}$$

we just need to estimate $P(D=1|\boldsymbol{U}=\boldsymbol{u},Y)$ and $P(D=1|\boldsymbol{U}=\boldsymbol{u})$ 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 improved can be made if the following semi-parametric model holds,

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

where $\alpha(\cdot)$ is an unspecified unknown function and γ is an unknown parameter. It follows from (7)-(8) that

$$\frac{f(Y|\boldsymbol{u}, D=1)}{f(Y|\boldsymbol{u}, D=0)} = e^{-\gamma Y} E(e^{\gamma Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=1). \tag{9}$$

If $\gamma = 0$, then $f(Y|\boldsymbol{u}, D=1) = f(Y|\boldsymbol{u}, D=0)$ and the estimator in (3) is correct. Under (9) with $\gamma \neq 0$, we just need to derive an estimator $\widehat{\gamma}$ of γ and apply kernel estimation to estimate $E(e^{\widehat{\gamma} Y} \mid \boldsymbol{U}=\boldsymbol{u}, D=1)$ 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

$$\begin{aligned}
\mu_1(\boldsymbol{u}) &= E\left\{ Y \frac{P(D=1|\boldsymbol{U}=\boldsymbol{u}, Y)}{P(D=0|\boldsymbol{U}=\boldsymbol{u}, Y)} \middle| \boldsymbol{U}=\boldsymbol{u}, D=0 \right\} \frac{P(D=0|\boldsymbol{U}=\boldsymbol{u})}{P(D=1|\boldsymbol{U}=\boldsymbol{u})} \\
&= E\left(Y e^{-\alpha(\boldsymbol{u})-\gamma Y} | \boldsymbol{U}=\boldsymbol{u}, D=0 \right) \frac{E\{P(D=0|\boldsymbol{U}=\boldsymbol{u}, Y)|\boldsymbol{U}=\boldsymbol{u}\}}{P(D=1|\boldsymbol{U}=\boldsymbol{u})} \\
&= e^{-\alpha(\boldsymbol{u})} E\left(Y e^{-\gamma Y} | \boldsymbol{U}=\boldsymbol{u}, D=0 \right) \frac{E\{e^{\alpha(\boldsymbol{u})+\gamma Y} P(D=1|\boldsymbol{U}=\boldsymbol{u}, Y)|\boldsymbol{U}=\boldsymbol{u}\}}{P(D=1|\boldsymbol{U}=\boldsymbol{u})} \\
&= E\left(Y e^{-\gamma Y} | \boldsymbol{U}=\boldsymbol{u}, D=0 \right) \frac{E\{e^{\gamma Y} E(D|\boldsymbol{U}=\boldsymbol{u}, Y)|\boldsymbol{U}=\boldsymbol{u}\}}{P(D=1|\boldsymbol{U}=\boldsymbol{u})} \\
&= E\left(Y e^{-\gamma Y} | \boldsymbol{U}=\boldsymbol{u}, D=0 \right) \frac{E(e^{\gamma Y} D|\boldsymbol{U}=\boldsymbol{u})}{P(D=1|\boldsymbol{U}=\boldsymbol{u})} \\
&= E\left(Y e^{-\gamma Y} | \boldsymbol{U}=\boldsymbol{u}, D=0 \right) E\left(e^{\gamma Y} | \boldsymbol{U}=\boldsymbol{u}, D=1 \right),
\end{aligned}$$

where the second and third equalities follow from (8).

I change the first and second to the second and third – Chi

For every real number t, define

$$h(\boldsymbol{u}, t) = E(Y e^{-tY} | \boldsymbol{U}=\boldsymbol{u}, D=0) E(e^{tY} | \boldsymbol{U}=\boldsymbol{u}, D=1).$$

Its estimator by kernel regression is

$$\widehat{h}(\boldsymbol{u},t)=\frac{\displaystyle\sum_{i=1}^{N}(1-D_i)\check{\kappa}_{\check{b}}(\boldsymbol{u}-\boldsymbol{U}_i)Y_ie^{-tY_i}}{\displaystyle\sum_{i=1}^{N}(1-D_i)\check{\kappa}_{\check{b}}(\boldsymbol{u}-\boldsymbol{U}_i)}\;\frac{\displaystyle\sum_{i=1}^{N}D_i\check{\kappa}_{\check{b}}(\boldsymbol{u}-\boldsymbol{U}_i)e^{tY_i}}{\displaystyle\sum_{i=1}^{N}D_i\check{\kappa}_{\check{b}}(\boldsymbol{u}-\boldsymbol{U}_i)} \tag{10}$$

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

$$\widehat{\gamma}=\arg\min_t\frac{1}{N}\sum_{i=1}^{N}D_i\{Y_i-\widehat{h}(\boldsymbol{U}_i,t)\}^2, \tag{11}$$

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

$$E[D\{Y-h(\boldsymbol{U},\gamma)\}^2|D=1]\le E[D\{Y-h(\boldsymbol{U},t)\}^2|D=1]$$

because $h(\boldsymbol{u},\gamma)=\mu_1(\boldsymbol{u})$.

Once $\widehat{\gamma}$ is obtained, our estimator of $\mu_1(\boldsymbol{u})$ is

$$\widehat{\mu}_1^{E3}(\boldsymbol{u})=\left\{\sum_{i=1}^{N}D_iY_i\,\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)+\sum_{i=1}^{N}(1-D_i)\widehat{Y}_i\,\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\right\}\Big/\sum_{i=1}^{N}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i) \tag{12}$$

with

$$\widehat{Y}_i=Y_i\,e^{-\widehat{\gamma}Y_i}\sum_{j=1}^{n}e^{\widehat{\gamma}Y_j}\,\check{\kappa}_{\check{b}}(\boldsymbol{U}_i-\boldsymbol{U}_j)\Big/\sum_{j=1}^{n}\check{\kappa}_{\check{b}}(\boldsymbol{U}_i-\boldsymbol{U}_j),$$

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 cross-validation as described in Györfi et al. (2002). Requirements on the rates of bandwidths are described in theorems in Section 3.

2.2 Constrained Kernel Regression with Unmeasured Covariates

We still consider the case with one external dataset, independent of the internal dataset. In this subsection, the external dataset contains iid observations $(Y_i,\boldsymbol{X}_i)$, $i=n+1,...,N$, 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 2.1 when $q < p$. Instead, we consider kernel regression using external information in a constraint. First, we consider the estimation of the n-dimensional vector $\boldsymbol{\mu}_1 = \big(\mu_1(\boldsymbol{U}_1), ..., \mu_1(\boldsymbol{U}_n)\big)^\top$, where $\boldsymbol{A}^\top$ denotes the transpose of vector or matrix $\boldsymbol{A}$ throughout. Note that the standard kernel regression (2) actually estimates $\boldsymbol{\mu}_1$ by

$$\widehat{\boldsymbol{\mu}}_1 = \arg\min_{\mu_1,...,\mu_n} \sum_{i=1}^n \sum_{j=1}^n \kappa_b(\boldsymbol{U}_i - \boldsymbol{U}_j)(Y_j - \mu_i)^2 \Big/ \sum_{k=1}^n \kappa_b(\boldsymbol{U}_i - \boldsymbol{U}_k). \tag{13}$$

We improve $\widehat{\boldsymbol{\mu}}_1$ by the following constrained minimization,

$$\widehat{\boldsymbol{\mu}}_1^C = \arg\min_{\mu_1,...,\mu_n} \sum_{i=1}^n \sum_{j=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)(Y_j - \mu_i)^2 \Big/ \sum_{k=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k) \tag{14}$$

$$\text{subject to} \quad \sum_{i=1}^n \{\mu_i - \widehat{h}_1(\boldsymbol{X}_i)\}\boldsymbol{g}(\boldsymbol{X}_i)^\top = 0, \tag{15}$$

where $\boldsymbol{g}(\boldsymbol{x})^\top = (1, \boldsymbol{x}^\top)$, l in (14) is a bandwidth that may be different from b in (2) or (13), and $\widehat{h}_1(\boldsymbol{x})$ is the kernel estimator of $h_1(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}, D = 1)$ using one of the methods described in Section 2.1 with $\boldsymbol{U}$ replaced by $\boldsymbol{X}$. Note that this can be done since both internal and external datasets have measured $\boldsymbol{X}_i$'s. More details of $\widehat{h}_1(\boldsymbol{x})$ are given in the end of this subsection.

Constraint (15) is an empirical analog (based on internal data) of

$$E\left[\left\{\mu_1(\boldsymbol{U}) - h_1(\boldsymbol{X})\right\}\boldsymbol{g}(\boldsymbol{X})^\top \,\middle|\, D = 1\right] = 0,$$

as $E\left\{E(Y \mid \boldsymbol{U}, D = 1) \mid \boldsymbol{X}, D = 1\right\} = E(Y \mid \boldsymbol{X}, D = 1) = h_1(\boldsymbol{X})$. Thus, if $\widehat{h}_1(\cdot)$ is a good estimator of $h_1(\cdot)$, then $\widehat{\boldsymbol{\mu}}_1^C$ in (14) is more accurate than the unconstrained $\widehat{\boldsymbol{\mu}}_1$ in (13).

It turns out that $\widehat{\boldsymbol{\mu}}_1^C$ in (14) has an explicit form $\widehat{\boldsymbol{\mu}}_1^C = \widehat{\boldsymbol{\mu}}_1 + \boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{G})^{-1}\boldsymbol{G}^\top(\widehat{\boldsymbol{h}}_1 - \widehat{\boldsymbol{\mu}}_1)$,

where $\boldsymbol{G}$ is the $n \times n$ matrix whose ith row is $\boldsymbol{g}(\boldsymbol{X}_i)^\top$ and $\widehat{\boldsymbol{h}}_1$ is the n-dimensional vector whose ith component is $\widehat{h}_1(\boldsymbol{X}_i)$.

To obtain an improved estimator of the entire regression function $\mu_1(\cdot)$ in (1), not just the function at $\boldsymbol{u} = \boldsymbol{U}_i$, $i = 1, ..., n$, we apply the standard kernel regression with response vector $(Y_1, ..., Y_n)^\top$ replaced by $\widehat{\boldsymbol{\mu}}_1^C$ in (14), which results in an estimator of $\mu_1(\boldsymbol{u})$ for any $\boldsymbol{u} \in \mathscr{U}$ as

$$\widehat{\mu}_1^C(\boldsymbol{u}) = \sum_{i=1}^{n} \widehat{\mu}_i \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i), \tag{16}$$

where $\widehat{\mu}_i$ is the ith component of $\widehat{\boldsymbol{\mu}}_1^C$ and b is the same bandwidth in (2).

With $\boldsymbol{U}$ replaced by $\boldsymbol{X}$, there are essentially three methods described in Section 2.1 for constructing the kernel estimator $\widehat{h}_1(\boldsymbol{x})$ of $h_1(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}, D = 1)$. The first one is the simplest estimator (3) (with $\boldsymbol{U}$ replaced by $\boldsymbol{X}$), which may be incorrect unless populations $\mathcal{P}_1$ and $\mathcal{P}_0$ are identical. We still denote the resulting estimator in (16) as $\widehat{\mu}_1^{C1}(\boldsymbol{u})$.

The second one is (5) with $\boldsymbol{U}$ replaced by $\boldsymbol{X}$ and $\widehat{f}(Y_i|\boldsymbol{X}_i, D = 1)/\widehat{f}(Y_i|\boldsymbol{X}_i, D = 0)$ constructed either by the ratio of two kernel estimators given by (6) or kernel estimators for binary responses through (7), with $\boldsymbol{U}$ replaced by $\boldsymbol{X}$. The resulting estimator in (16) is denoted as $\widehat{\mu}_1^{C2}(\boldsymbol{u})$.

The third and last one uses $\widehat{h}(\boldsymbol{X}_i, \widehat{\gamma})$ for $\widehat{h}_1(\boldsymbol{X}_i)$ in constraint (15), where $\widehat{\gamma}$ and $h(\boldsymbol{X}_i, t)$ are defined by (11) and (10), respectively, with $\boldsymbol{U}$ replaced by $\boldsymbol{X}$. The resulting estimator in (16) is denoted as $\widehat{\mu}_1^{C3}(\boldsymbol{u})$.

3 Theory

We now establish the asymptotic normality of $\widehat{\mu}_1^E(\boldsymbol{u})$ and $\widehat{\mu}_1^C(\boldsymbol{u})$ 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 $\widehat{\mu}_1^{E2}(\boldsymbol{u})$ in (5). The result is also applicable to $\widehat{\mu}_1^{E1}(\boldsymbol{u})$ in (3) with an added condition that $\mathcal{P}_1 = \mathcal{P}_0$.

Theorem 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 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 derivatives of $\mu_1(\boldsymbol{u})f_k(\boldsymbol{u})$ are continuous and bounded, where $\sigma_1^2(\boldsymbol{u}) = E[\{Y - \mu_1(\boldsymbol{U})\}^2 \mid \boldsymbol{U} = \boldsymbol{u}, D = 1]$, $\sigma_0^2(\boldsymbol{u}) = E[\{\tilde{Y} - \mu_1(\boldsymbol{U})\}^2 \mid \boldsymbol{U} = \boldsymbol{u}, D = 0]$, and $\tilde{Y} = Yf(Y|\boldsymbol{U}, D = 1)/f(Y|\boldsymbol{U}, D = 0)$. Also, $E(|Y|^s|\boldsymbol{U} = \boldsymbol{u}, D = 1)f_1(\boldsymbol{u})$ and $E(|\tilde{Y}|^s|\boldsymbol{U} = \boldsymbol{u}, D = 0)f_0(\boldsymbol{u})$ are bounded for a constant $s > 2$.*

(B3) *The kernel κ is second order, i.e., $\int \boldsymbol{u}\,\kappa(\boldsymbol{u})d\boldsymbol{u} = 0$ and $0 < \int \boldsymbol{u}^\top \boldsymbol{u}\,\kappa(\boldsymbol{u})d\boldsymbol{u} < \infty$.*

(B4) *The bandwidth b satisfies $b \to 0$ and $(a+1)nb^{p+4} \to c \in [0, \infty)$, where $a = \lim_{n\to\infty}(N-n)/n$ (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), $f(y, \boldsymbol{u}|D = 1)$, $f(y, \boldsymbol{u}|D = 0)$ are $\tilde{m}$th order continuously differentiable with bounded derivatives, and $f_1(\boldsymbol{u})$ and $f_0(\boldsymbol{u})$ are $\bar{m}$th order continuously differentiable with bounded derivatives. Functions $f(y, \boldsymbol{u}|D = 0)$ and $f_1(\boldsymbol{u})$ are bounded away from zero. The bandwidths $\tilde{b}$ and $\bar{b}$ satisfy $n\tilde{b}^{p+1}/\log(n) \to \infty$ and $n\bar{b}^p/\log(n) \to \infty$.*

Then, for any fixed $\boldsymbol{u}$ with $f_0(\boldsymbol{u}) > 0$ and $f_1(\boldsymbol{u}) > 0$ and $\widehat{\mu}_1^{E2}$ in (5),

$$\sqrt{nb^p}\{\widehat{\mu}_1^{E2}(\boldsymbol{u}) - \mu_1(\boldsymbol{u})\} \xrightarrow{d} N\big(B_a(\boldsymbol{u}), V_a(\boldsymbol{u})\big), \tag{17}$$

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

$$\begin{aligned}
B_a(\boldsymbol{u}) &= \frac{c^{1/2}\{f_1(\boldsymbol{u})A_1(\boldsymbol{u}) + af_0(\boldsymbol{u})A_0(\boldsymbol{u})\}}{(a+1)^{1/2}\{f_1(\boldsymbol{u}) + af_0(\boldsymbol{u})\}},\\
A_1(\boldsymbol{u}) &= \int \kappa(\boldsymbol{v}) \left\{ \tfrac{1}{2}\boldsymbol{v}^\top \nabla^2 \mu_1(\boldsymbol{u})\boldsymbol{v} + \boldsymbol{v}^\top \nabla \log f_1(\boldsymbol{u}) \nabla \mu_1(\boldsymbol{u})^\top \boldsymbol{v} \right\} d\boldsymbol{v},\\
A_0(\boldsymbol{u}) &= \int \kappa(\boldsymbol{v}) \left\{ \tfrac{1}{2}\boldsymbol{v}^\top \nabla^2 \mu_1(\boldsymbol{u})\boldsymbol{v} + \boldsymbol{v}^\top \nabla \log f_0(\boldsymbol{u}) \nabla \mu_1(\boldsymbol{u})^\top \boldsymbol{v} \right\} d\boldsymbol{v},\\
V_a(\boldsymbol{u}) &= \frac{f_1(\boldsymbol{u})\sigma_1^2(\boldsymbol{u}) + af_0(\boldsymbol{u})\sigma_0^2(\boldsymbol{u})}{\{f_1(\boldsymbol{u}) + af_0(\boldsymbol{u})\}^2} \int \kappa(\boldsymbol{v})^2 d\boldsymbol{v}.
\end{aligned}$$

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

$$\max_{i=n+1,\ldots,N} \left| \frac{\widehat{f}(Y_i|\boldsymbol{U} = \boldsymbol{U}_i, D = 1)}{\widehat{f}(Y_i|\boldsymbol{U} = \boldsymbol{U}_i, D = 0)} - \frac{f(Y_i|\boldsymbol{U} = \boldsymbol{U}_i, D = 1)}{f(Y_i|\boldsymbol{U} = \boldsymbol{U}_i, D = 0)} \right| = \frac{o_p(1)}{\sqrt{nb^p}} \tag{18}$$

(Lemma 8.10 in Newey and McFadden (1994)), where $o_p(1)$ denotes a term tending to 0 in probability. Result (18) implies that the estimation of ratio $f(Y|\boldsymbol{U}, D = 1)/f(Y|\boldsymbol{U}, D = 0)$ does not affect the asymptotic distribution of $\widehat{\mu}_1^{E2}(\boldsymbol{u})$ in (5).

Note that both the bias $B_a(\boldsymbol{u})$ and variance $V_a(\boldsymbol{u})$ in (17) are decreasing in the limit $a = \lim_{n\to\infty}(N - n)/n$, 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 well-known 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, $\widehat{\mu}_1^{E2}(\boldsymbol{u})$ has a convergence rate tending to 0 faster than $1/\sqrt{nb^p}$, the convergence rate of the standard kernel estimator $\widehat{\mu}_1(\boldsymbol{u})$.

The next result is about $\widehat{\mu}_1^{C2}(\boldsymbol{u})$ in (16) as described in the end of Section 2.2.

Theorem 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 derivatives.*

(C2) Functions $\mu_1(\boldsymbol{u}) = E(Y|\boldsymbol{U} = \boldsymbol{u})$ *and* $\sigma_1^2(\boldsymbol{u})$ *are Lipschitz continuous;* $\mu_1(\boldsymbol{u})$ *has bounded third-order derivatives;* $h_1(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}, D = 1)$ *has bounded first and second order derivatives; and* $E(|Y|^s|\boldsymbol{U} = \boldsymbol{u}, D = 1)$ *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) $a = \lim_{n\to\infty}(N - n)/n > 0$ *and the bandwidths* b *in (2) and* l *in (14) satisfy* $b \to 0$*,* $l \to 0$*,* $l/b \to r \in (0, \infty)$*,* $nb^p \to \infty$*, and* $nb^{4+p} \to c \in [0, \infty)$*, as* $n \to \infty$*.*

(C5) The densities $f_{X0}(\boldsymbol{x})$ *and* $f_{X1}(\boldsymbol{x})$ *for* $\boldsymbol{X}$ *respectively under internal and external populations are bounded away from zero. There exists a constant* $s > 4$ *such that* $E(|Y|^s \mid D = 1)$ *and* $E(|\tilde{Y}|^s \mid D = 0)$ *are finite,* $E(|Y|^s \mid \boldsymbol{X} = \boldsymbol{x}, D = 1)f_{X1}(\boldsymbol{x})$ *and* $E(|\tilde{Y}|^s \mid \boldsymbol{X} = \boldsymbol{x}, D = 0)f_{X0}(\boldsymbol{x})$ *are bounded, and the bandwidth* b_h *for* $\widehat{h}_1$ *satisfies* $n^{1-2/s}b_h^q/\log(n) \to \infty$*. (Should we use something different for b?)*

Use b_h?? – Chi

Then, for any fixed $\boldsymbol{u} \in \mathscr{U}$ *and* $\widehat{\mu}_1^{C2}(\boldsymbol{u})$ *in (16),*

$$\sqrt{nb^p}\{\widehat{\mu}_1^{C2}(\boldsymbol{u}) - \mu_1(\boldsymbol{u})\} \xrightarrow{d} N\big(B_r(\boldsymbol{u}), V_r(\boldsymbol{u})\big), \tag{19}$$

where

$$\begin{aligned}
B_r(\boldsymbol{u}) &= c^{1/2}[(1+r^2)A_1(\boldsymbol{u}) - r^2\boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} E\{\boldsymbol{g}(\boldsymbol{X})A_1(\boldsymbol{U})|D=1\}],\\
A_1(\boldsymbol{u}) &= \int \kappa(\boldsymbol{v}) \left\{\tfrac{1}{2}\boldsymbol{v}^\top \nabla^2 \mu_1(\boldsymbol{u})\boldsymbol{v} + \boldsymbol{v}^\top \nabla \log f_1(\boldsymbol{u}) \nabla \mu_1(\boldsymbol{u})^\top \boldsymbol{v}\right\} d\boldsymbol{v},\\
V_r(\boldsymbol{u}) &= \frac{\sigma_1^2(\boldsymbol{u})}{f_1(\boldsymbol{u})} \int \left\{\int \kappa(\boldsymbol{v} - r\boldsymbol{w})\kappa(\boldsymbol{w}) d\boldsymbol{w}\right\}^2 d\boldsymbol{v},
\end{aligned}$$

and $\boldsymbol{\Sigma} = E\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top \mid D=1\}$ *is assumed to be positive definite without loss of generality.*

The next result is about $\widehat{\gamma}$ in (11).

Theorem 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 $\int \check{\kappa}(\boldsymbol{u})d\boldsymbol{u} = 1$, has a bounded support, and has order $d > \max\{(p+4)/2, p\}$.

(D2) The bandwidth ℓ in (10) satisfies $N\ell^{2q}/(\log N)^2 \to \infty$ and $N\ell^{2d} \to 0$ as the total sample size of internal and external datasets $N \to \infty$, where d is given in (D1).

(D3) γ in (8) is an interior point of a compact domain Γ 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 $t \to \gamma$, $h(\cdot, t)$, $\nabla_t h(\cdot, t)$, and $\nabla_t^2 h(\cdot, t)$ convergence uniformly.

(D4) $\sup_{t\in\Gamma} E\|\boldsymbol{W}_t\|^4 < \infty$ and $\sup_{t\in\Gamma} E[\|\boldsymbol{W}_t\|^4|\boldsymbol{U}]f_U(\boldsymbol{U})$ is bounded, where $\|\boldsymbol{a}\|^2 = \boldsymbol{a}^\top\boldsymbol{a}$, $\boldsymbol{W}_t = (De^{tY}, (1-D)Ye^{-tY}, D, (1-D), DYe^{tY}, (1-D)Y^2e^{-tY}, DY^2e^{tY}, (1-D)Y^3e^{-tY})^\top$,

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 $\|\boldsymbol{W}_t - \boldsymbol{W}_{t'}\| < \tau(Y,D)|t-t'|$.

(D5) The function $\boldsymbol{\omega}_t(\boldsymbol{u}) = E(\boldsymbol{W}_t|\boldsymbol{U}=\boldsymbol{u})f_U(\boldsymbol{u})$ is bounded away from zero, and it is dth-order continuously differentiable with bounded derivatives on an open set containing the support of $\boldsymbol{U}$. There is a functional $G(Y,D,\boldsymbol{\omega})$ linear in $\boldsymbol{\omega}$ such that $|G(Y,D,\boldsymbol{\omega})| \leq \iota(Y,D)\|\boldsymbol{\omega}\|_\infty$ and, for small enough $\|\boldsymbol{\omega}-\boldsymbol{\omega}_\gamma\|_\infty$, $|\psi(Y,D,\boldsymbol{\omega})-\psi(Y,D,\boldsymbol{\omega}_\gamma)-G(Y,D,\boldsymbol{\omega}-\boldsymbol{\omega}_\gamma)| \leq \iota(Y,D)\|\boldsymbol{\omega}-\boldsymbol{\omega}_\gamma\|_\infty^2$, where $\iota(Y,D)$ is a function with $E\{\iota(Y,D)\} < \infty$, $\psi(Y,D,\boldsymbol{\omega}) = -2D\left(Y - \frac{\omega_1\omega_2}{\omega_3\omega_4}\right)\left(\frac{\omega_2\omega_5-\omega_1\omega_6}{\omega_3\omega_4}\right)$, ω_j is the jth component of $\boldsymbol{\omega}$, $\|\boldsymbol{\omega}\|_\infty = \sup_{\boldsymbol{x}\in\mathscr{U}}\|\boldsymbol{\omega}(\boldsymbol{u})\|$, $\|\boldsymbol{\omega}-\boldsymbol{\omega}_\gamma\|_\infty = \sup_{\boldsymbol{x}\in\mathscr{U}}\|\boldsymbol{\omega}(\boldsymbol{u})-\boldsymbol{\omega}_\gamma(\boldsymbol{u})\|$, and $\mathscr{U}$ is the range of $\boldsymbol{U}$. Also, there exists an almost everywhere continuous 8-dimensional function $\boldsymbol{\nu}(\boldsymbol{U})$ with $\int\|\boldsymbol{\nu}(\boldsymbol{u})\|d\boldsymbol{u} < \infty$ and $E\{\sup_{\|\boldsymbol{\delta}\|\leq\epsilon}\|\boldsymbol{\nu}(\boldsymbol{U}+\boldsymbol{\delta})\|^4\} < \infty$ for some $\epsilon > 0$ such that $E\{G(Y,D,\boldsymbol{\omega})\} = \int\boldsymbol{\nu}(\boldsymbol{u})^\top\boldsymbol{\omega}(\boldsymbol{u})d\boldsymbol{u}$ for all $\|\boldsymbol{\omega}\|_\infty < \infty$.

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

$$\sqrt{N}(\widehat{\gamma}-\gamma) \xrightarrow{d} N(0,\sigma_\gamma^2), \tag{20}$$

where $\sigma_\gamma^2 = [2E\{D\nabla_\gamma h(\boldsymbol{U},\gamma)\}^2]^{-1}\mathrm{Var}[\psi(Y,D,\boldsymbol{\omega}_\gamma) + \boldsymbol{\nu}(\boldsymbol{U})^\top\boldsymbol{W}_\gamma - E\{\boldsymbol{\nu}(\boldsymbol{U})^\top\boldsymbol{W}_\gamma\}]$.

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 1-3, we obtain the following result for $\widehat{\mu}_1^{E3}(\boldsymbol{u})$ in (12) or $\widehat{\mu}_1^{C3}(\boldsymbol{u})$ in (16).

Corollary 1. *Suppose that (8) holds for the binary random D indicating internal and ex-*

ternal data.

(i) Under (B1)-(B4) and (D1)-(D5), result (17) holds with $\widehat{\mu}_1^{E2}(\boldsymbol{u})$ *replaced by* $\widehat{\mu}_1^{E3}(\boldsymbol{u})$ *in (12).*

(ii) Under (C1)-(C4) and (D1)-(D5) with $\boldsymbol{U}$ *and* p *replaced by* $\boldsymbol{X}$ *and* q*, respectively, result (19) holds with* $\widehat{\mu}_1^{C2}(\boldsymbol{u})$ *replaced by* $\widehat{\mu}_1^{C3}(\boldsymbol{u})$*.*

4 Simulation Results

In this section, we present simulation results to examine and compare the performance of the standard kernel estimator (2) without using external information and our proposed estimator (16) with three variations, $\widehat{\mu}_1^{C1}$, $\widehat{\mu}_1^{C2}$, and $\widehat{\mu}_1^{C3}$, as described in the end of Section 2.2.

We consider $\boldsymbol{U} = (X, Z)^\top$ with univariate covariate $\boldsymbol{X} = X$ and unmeasured Z ($p = 2$ and $q = 1$) in two cases:

(i) normal covariates: $(X, Z)^\top$ is bivariate normal with means 0, variances 1, and correlation 0.5;

(ii) bounded covariates: $X = BW_1 + (1 - B)W_2$ and $Z = BW_1 + (1 - B)W_3$, 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. $\mu(X, Z) = \cos(2X)/2 + \sin(Z)$;

M3. $\mu(X, Z) = \cos(2XZ)/2 + \sin(Z)$;

M4. $\mu(X, Z) = X/2 - Z^2/4 + \cos(XZ)/4$.

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 $P(D = 1 \mid Y, X, Z) = 1/\exp(1 + 2|X| + \gamma Y)$, where $\gamma = 0$ or $1/2$. Under this setting, the unconditional $P(D = 1)$ is between 10% and 15%.

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:

$$\text{MISE} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{T}\sum_{t=1}^{T}\{\widehat{\mu}_1^*(\boldsymbol{U}_{s,t}) - \mu_1(\boldsymbol{U}_{s,t})\}^2, \tag{21}$$

where $\{\boldsymbol{U}_{s,t} : t = 1, ..., T\}$ are test data for each simulation replication s, the simulation is repeated independently for $s = 1, ..., S$, and $\widehat{\mu}_1^*(\cdot)$ is an estimator of $\mu_1(\cdot)$ using a method described previously based on internal and external data, 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 $s = 1, ..., S$ and independently across s.

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

$$\text{IMP} = 1 - \frac{\min\{\text{MISE}(\widehat{\mu}_1^{Cj}) \text{ over } j = 1, 2, 3\}}{\text{MISE}(\widehat{\mu}_1)}. \tag{22}$$

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 based on $S = 200$ replications is shown in Table 1 for the case of $\gamma = 0$ and in Table 2 for the case of $\gamma = 1/2$.

Consider first the results in Table 1. Since $\gamma = 0$, all three estimators, $\widehat{\mu}_1^{C1}$, $\widehat{\mu}_1^{C2}$, and $\widehat{\mu}_1^{C3}$, are correct and more efficient than the standard estimator $\widehat{\mu}_1$ in (2) without using external information. The estimator $\widehat{\mu}_1^{C1}$ is the best, as it uses the correct information that populations are homogeneous ($\gamma = 0$).

Next, the results in Table 2 for $\gamma = 1/2$ indicate that the estimator $\widehat{\mu}_1^{C2}$ or $\widehat{\mu}_1^{C3}$ using a correct constraint is better than the estimator $\widehat{\mu}_1^{C1}$ using an incorrect constraint or the estimator $\widehat{\mu}_1$ without using external information. Since $\widehat{\mu}_1^{C3}$ uses more information, it is in general better than $\widehat{\mu}_1^{C2}$. With an incorrect constraint, $\widehat{\mu}_1^{C1}$ can be much worse than $\widehat{\mu}_1$ without using external information.

5 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 (Li, 1991), i.e., $\mu_1(\boldsymbol{U})$ in (1) is assumed to be

$$\mu_1(\boldsymbol{U}) = \mu_1(\boldsymbol{\eta}^\top \boldsymbol{U}), \tag{23}$$

where $\boldsymbol{\eta}$ is an unknown p-dimensional vector. The well-known SIR technique (Li, 1991) can be applied to obtain a consistent and asymptotically normal estimator $\widehat{\boldsymbol{\eta}}$ 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 and Weisberg, 1991; Li and Wang, 2007; Shao et al., 2007; Xia et al., 2002; Ma and Zhu, 2012).

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

$$\sum_{i=1}^{n} \{\mu_i - \widehat{h}_1^{(k)}(X_i^{(k)})\} \boldsymbol{g}_k(X_i^{(k)})^\top = 0, \qquad k = 1, ..., q, \tag{24}$$

where $X_i^{(k)}$ is the kth component of $\boldsymbol{X}_i$, $\boldsymbol{g}_k(X^{(k)}) = (1, X^{(k)})^\top$, and $\widehat{h}_1^{(k)}(X_i^{(k)})$ is an estimator of $h_1^{(k)}(X^{(k)}) = E(Y \mid X^{(k)}, D = 1)$ using methods described in Section 2. More constraints are involved in (24), but estimation only involves one dimensional $X^{(k)}$, $k = 1, ..., q$.

The kernel κ 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 mth order kernel with $m > 2$ as defined by Bierens (1987) may be used to achieve convergence rate $n^{-m/(2m+p)}$. Alternatively, we may apply other nonparametric smoothing techniques such as the local polynomial (Fan et al., 1997) to achieve convergence rate $n^{-m/(2m+p)}$ 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.

Appendix

Proof of Theorem 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 $\widehat{\mu}_1(\boldsymbol{u}) = \sum_{i=1}^n \kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)Y_i / \sum_{i=1}^n \kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)$, $\widehat{\mu}_0(\boldsymbol{u}) = \sum_{i=n+1}^N \kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\tilde{Y}_i / \sum_{i=n+1}^N \kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)$, and $\widehat{p}(\boldsymbol{u}) = \sum_{i=1}^n \kappa_b(\boldsymbol{u}-\boldsymbol{U}_i) / \sum_{i=1}^N \kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)$. Under (B3)-(B4), Theorem 2 in Nadaraya (1964) shows that $\widehat{p}(\boldsymbol{u})$ converges to $P(D=1|\boldsymbol{U}=\boldsymbol{u})$ in probability. Under (B1)-(B4), $\sqrt{nb^p}\{\widehat{\mu}_1(\boldsymbol{u}) - \mu_1(\boldsymbol{u})\} \xrightarrow{d} N\big(B_1(\boldsymbol{u}), V_1(\boldsymbol{u})\big)$, $B_1(\boldsymbol{u}) = c^{1/2}A_1(\boldsymbol{u})$, $V_1(\boldsymbol{u}) = \frac{\sigma_1^2(\boldsymbol{u})}{f_1(\boldsymbol{u})}\int \kappa(\boldsymbol{v})^2 d\boldsymbol{v}$, and $\sqrt{n/(N-n)}\sqrt{(N-n)b^p}\{\widehat{\mu}_0(\boldsymbol{u}) - \mu_1(\boldsymbol{u})\} \xrightarrow{d} N\big(B_0(\boldsymbol{u}), V_0(\boldsymbol{u})\big)$, $B_0(\boldsymbol{u}) = c^{1/2}A_0(\boldsymbol{u})$, $V_0(\boldsymbol{u}) = \frac{\sigma_0^2(\boldsymbol{u})}{a f_0(\boldsymbol{u})}\int \kappa(\boldsymbol{v})^2 d\boldsymbol{v}$. 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 $|\widehat{\mu}_1^{E2}(\boldsymbol{u}) - \tilde{\mu}_1(\boldsymbol{u})|$ is bounded by

$$\{1-\widehat{p}(\boldsymbol{u})\} \max_{i>n} \left| \frac{\widehat{f}(Y_i|\boldsymbol{U}=\boldsymbol{U}_i, D=1)}{\widehat{f}(Y_i|\boldsymbol{U}=\boldsymbol{U}_i, D=0)} - \frac{f(Y_i|\boldsymbol{U}=\boldsymbol{U}_i, D=1)}{f(Y_i|\boldsymbol{U}=\boldsymbol{U}_i, D=0)} \right| \frac{\sum_{i=n+1}^N |Y_i| \kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)}{\sum_{i=n+1}^N \kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)}, \tag{25}$$

which is $o_p(1/\sqrt{nb^p})$ by result (18) under condition (B5).

Proof of Theorem 2. Write

$$\sqrt{nb^p}\{\widehat{\mu}_1^{C2}(\boldsymbol{u}) - \mu_1(\boldsymbol{u})\} = T_1 + \cdots + T_6, \tag{26}$$

where $T_1 = n^{-1/2}b^{p/2}\boldsymbol{\delta}_b(\boldsymbol{u})^\top(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{B}_l^{-1}\boldsymbol{\Delta}_l\boldsymbol{\epsilon}/\widehat{f}_b(\boldsymbol{u})$, $T_2 = n^{-1/2}b^{p/2}\boldsymbol{\delta}_b(\boldsymbol{u})^\top\{\boldsymbol{\mu}_1 - \mu_1(\boldsymbol{u})\mathbf{1}_n\}/\widehat{f}_b(\boldsymbol{u})$,

$T_3 = n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top (\boldsymbol{B}_l^{-1} \boldsymbol{\Delta}_l \boldsymbol{\mu}_1 - \boldsymbol{\mu}_1)/\widehat{f}_b(\boldsymbol{u})$, $T_4 = -n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \boldsymbol{P} (\boldsymbol{B}_l^{-1} \boldsymbol{\Delta}_l \boldsymbol{\mu}_1 - \boldsymbol{\mu}_1)/\widehat{f}_b(\boldsymbol{u})$, $T_5 = n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \boldsymbol{P} (\widehat{\boldsymbol{h}}_1 - \boldsymbol{h}_1)/\widehat{f}_b(\boldsymbol{u})$, $T_6 = n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \boldsymbol{P} (\boldsymbol{h}_1 - \boldsymbol{\mu}_1)/\widehat{f}_b(\boldsymbol{u})$, $\widehat{f}_b(\boldsymbol{u}) = \sum_{i=1}^n \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)/n$, $\boldsymbol{\delta}_b(\boldsymbol{u}) = (\kappa_b(\boldsymbol{u} - \boldsymbol{U}_1), \ldots, \kappa_b(\boldsymbol{u} - \boldsymbol{U}_n))^\top$, $\boldsymbol{I}_n$ 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 ith diagonal element is $\widehat{f}_l(\boldsymbol{U}_i)$, $\boldsymbol{\Delta}_l$ is the $n \times n$ matrix whose (i,j)th entry is $\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)/n$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$ with $\epsilon_i = Y_i - \mu_1(\boldsymbol{U}_i)$, $\boldsymbol{h}_1$ is the n-dimensional vector whose ith component is $h_1(\boldsymbol{X}_i)$, $\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top$, and $\boldsymbol{G}$, $\widehat{\boldsymbol{h}}_1$, and $\boldsymbol{\mu}_1$ are defined in Section 2.

We first show that T_1 in (26) is asymptotically normal with mean 0 and variance $V_r(\boldsymbol{u})$ defined in Theorem 2. Consider a further decomposition $T_1 = \sqrt{n}\, V + T_{11} + T_{12} + T_{13}$, where

$$V = \frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n S(\boldsymbol{U}_i, \epsilon_i, \boldsymbol{U}_j, \epsilon_j)$$

is a V-statistic with

$$S(\boldsymbol{U}_i, \epsilon_i, \boldsymbol{U}_j, \epsilon_j) = \frac{b^{p/2}}{2 f_1(\boldsymbol{u})} \left\{ \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j) \epsilon_j}{f_1(\boldsymbol{U}_i)} + \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_j) \kappa_l(\boldsymbol{U}_j - \boldsymbol{U}_i) \epsilon_i}{f_1(\boldsymbol{U}_j)} \right\},$$

$$T_{11} = \frac{b^{p/2}}{n^{3/2}} \sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \kappa_l(0) \epsilon_i}{f_1(\boldsymbol{u}) f_1(\boldsymbol{U}_i)},$$

$$T_{12} = \frac{b^{p/2}}{n^{3/2}} \sum_{j=1}^n \sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{f_1(\boldsymbol{u}) f_1(\boldsymbol{U}_i)} \left\{ \frac{f_1(\boldsymbol{u}) f_1(\boldsymbol{U}_i)}{\widehat{f}_b(\boldsymbol{u}) \widehat{f}_l(\boldsymbol{U}_i)} - 1 \right\} \epsilon_j,$$

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

$$S_1(\boldsymbol{U}_1, \epsilon_1) = E\{S(\boldsymbol{U}_1, \epsilon_1, \boldsymbol{U}_2, \epsilon_2) \mid \boldsymbol{U}_1, \epsilon_1\} = \frac{b^{p/2}}{2 f_1(\boldsymbol{u})} \left\{ \int \kappa_l(\boldsymbol{u}_2 - \boldsymbol{U}_1) \kappa_b(\boldsymbol{u} - \boldsymbol{u}_2) d\boldsymbol{u}_2 \right\} \epsilon_1$$

having variance

$$\begin{aligned} \mathrm{Var}\{S_1(\boldsymbol{U}_1, \epsilon_1)\} &= \frac{b^{p/2}}{4 f_1^2(\boldsymbol{u})} \int f_1(\boldsymbol{u}_1) \sigma_1^2(\boldsymbol{u}_1) \left\{ \int \kappa_l(\boldsymbol{u}_2 - \boldsymbol{u}_1) \kappa_b(\boldsymbol{u} - \boldsymbol{u}_2) d\boldsymbol{u}_2 \right\}^2 d\boldsymbol{u}_1 \\ &= \frac{b^{p/2}}{4 f_1^2(\boldsymbol{u})} \int f_1(\boldsymbol{u}_1) \sigma_1^2(\boldsymbol{u}_1) \left\{ \int \kappa_l(\boldsymbol{v}) \kappa_b(\boldsymbol{u} - \boldsymbol{u}_1 - l\boldsymbol{\nu}) d\boldsymbol{\nu} \right\}^2 d\boldsymbol{u}_1 \end{aligned}$$

$$= \frac{1}{4f_1^2(\boldsymbol{u})} \int f_1(\boldsymbol{u} - b\boldsymbol{w})\sigma_1^2(\boldsymbol{u} - b\boldsymbol{w}) \left\{ \int \kappa(\boldsymbol{v})\kappa\left(\boldsymbol{w} - \boldsymbol{\nu}\frac{l}{b}\right) d\boldsymbol{\nu} \right\}^2 d\boldsymbol{w},$$

where $\sigma_1^2(\cdot)$ is given in condition (C2), the second and third equalities follow from changing variables $\boldsymbol{u}_2 - \boldsymbol{u}_1 = l\boldsymbol{\nu}$ and $\boldsymbol{u} - \boldsymbol{u}_1 = b\boldsymbol{w}$, respectively. From the continuity of $f_1(\cdot)$ and $\sigma_1^2(\cdot)$, $\mathrm{Var}\{S_1(\boldsymbol{u}_1, \epsilon_1)\}$ converges to $V_r(\boldsymbol{u})$. Therefore, by the theory for asymptotic normality of V-statistics (e.g., Theorem 3.16 in Shao (2003)), $\sqrt{n}\, V \xrightarrow{d} N\big(0, V_r(\boldsymbol{u})\big)$.

Conditioned on $\boldsymbol{U}_1, ..., \boldsymbol{U}_n$, T_{11} has mean 0 and variance

$$\begin{aligned}\mathrm{Var}(T_{11}|\boldsymbol{U}_1, ..., \boldsymbol{U}_n) &= \frac{b^p}{4f_1^2(\boldsymbol{u})n^3} \sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)^2 \kappa_l(0)^2 \sigma_1^2(\boldsymbol{U}_i)}{f_1(\boldsymbol{U}_i)} \\ &\leq \frac{\sup_{\boldsymbol{u}\in\mathscr{U}} \kappa(\boldsymbol{u})^3}{4f_1^2(\boldsymbol{u})n^3 b^{2p}} \sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\sigma_1^2(\boldsymbol{U}_i)}{f_1(\boldsymbol{U}_i)} = o_p(1).\end{aligned}$$

This proves that $T_{11} = o_p(1)$. Note that $E(T_{12} \mid \boldsymbol{U}_1, ..., \boldsymbol{U}_n) = 0$ and $\mathrm{Var}(T_{12} \mid \boldsymbol{U}_1, ..., \boldsymbol{U}_n)$ is bounded by

$$\max\left\{ \frac{1}{f_1^2(\boldsymbol{u})}, \frac{1}{\widehat{f}_b^2(\boldsymbol{u})} \right\} \max_{i=1,...,n} \left| \frac{f_1(\boldsymbol{U}_i)}{\widehat{f}_l(\boldsymbol{U}_i)} - 1 \right|^2 \mathrm{Var}(\sqrt{n}V + T_{11}|\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n).$$

Therefore, under the assumed condition that f_1 is bounded away from zero, Lemma 3 in Dai and Shao (2022) implies $T_{12} = o_p(1)$. Note that

$$T_{13} = \frac{b^{p/2}}{n^{1/2}} \sum_{j=1}^n W_j(\boldsymbol{u})\epsilon_j, \;\; W_j(\boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\boldsymbol{g}(\boldsymbol{X}_i)^\top}{\widehat{f}_b(\boldsymbol{u})} (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \sum_{i=1}^n \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)\boldsymbol{g}(\boldsymbol{X}_i)}{\widehat{f}_l(\boldsymbol{U}_i)}.$$

Conditioned on $\boldsymbol{U}_1, ..., \boldsymbol{U}_n$, T_{13} has mean 0 and variance

$$\mathrm{Var}(T_{13} \mid \boldsymbol{U}_1, ..., \boldsymbol{U}_n) = \frac{b^p}{n} \sum_{j=1}^n W_j^2(\boldsymbol{u})\sigma_1^2(\boldsymbol{U}_j) = O_p(b^p) = o_p(1),$$

because, under the assumed condition that f_1 is bounded away from zero, Lemma 3 in Dai

and Shao (2022) implies $\max_{j=1,\ldots,n} |W_j(\boldsymbol{u}) - \boldsymbol{g}(\boldsymbol{u})^\top \boldsymbol{\Sigma}_g^{-1} \boldsymbol{g}(\boldsymbol{X}_j)| = o_p(1)$. 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 (2022) and (C4), $T_2 = \sqrt{c} A_1(\boldsymbol{u})\{1 + o_p(1)\}$. Note that

$$\begin{aligned} T_3 &= \frac{\sqrt{nb^p} l^2}{n \widehat{f}_b(\boldsymbol{u})} \sum_{j=1}^n \kappa_b(\boldsymbol{u} - \boldsymbol{U}_j) \left[\frac{1}{nl^2 \widehat{f}_b(\boldsymbol{U}_j)} \sum_{i=1}^n \kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)\{\mu_1(\boldsymbol{U}_i) - \mu_1(\boldsymbol{U}_j)\} \right] \\ &= \left\{ \frac{\sqrt{c} r^2}{n \widehat{f}_b(\boldsymbol{u})} \sum_{j=1}^n \kappa_b(\boldsymbol{u} - \boldsymbol{U}_j) A_1(\boldsymbol{U}_j) \right\} \{1 + o_p(1)\} = \sqrt{c} r^2 A_1(\boldsymbol{u})\{1 + o_p(1)\}, \end{aligned}$$

where the second equality follows from (A4) and Lemmas 3-4 in Dai and Shao (2022), and the last equality follows from Lemma 2 in Dai and Shao (2022) and continuity of $A_1(\cdot)$. Also,

$$\begin{aligned} -\frac{n^{1/2} T_4}{b^{p/2}} =& \frac{1}{n} \sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \boldsymbol{g}(\boldsymbol{X}_i)^\top}{\widehat{f}_b(\boldsymbol{u})} (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \sum_{j=1}^n \frac{\boldsymbol{g}(\boldsymbol{X}_j)}{n \widehat{f}_b(\boldsymbol{U}_j)} \sum_{i=1}^n \kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)\{\mu_1(\boldsymbol{U}_i) - \mu_1(\boldsymbol{U}_j)\} \\ =& \left\{ \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} \frac{1}{n} \sum_{j=1}^n \frac{\boldsymbol{g}(\boldsymbol{X}_j)}{n \widehat{f}_b(\boldsymbol{U}_j)} \sum_{i=1}^n \kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)\{\mu_1(\boldsymbol{U}_i) - \mu_1(\boldsymbol{U}_j)\} \right\} \{1 + o_p(1)\} \\ =& \left\{ \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} \frac{l^{2/p}}{n} \sum_{j=1}^n \boldsymbol{g}(\boldsymbol{X}_j) A_1(\boldsymbol{U}_j) \right\} \{1 + o_p(1)\} \\ =& \sqrt{c} r^2 \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} E\{\boldsymbol{g}(\boldsymbol{X}) A_1(\boldsymbol{U})\}\{1 + o_p(1)\}, \end{aligned}$$

where the first equality follows from Lemma 3 in Dai and Shao (2022) and the law of large numbers, the second equality follows from Lemma 4 in Dai and Shao (2022), and the last equality follows from the law of large numbers. Similarly, is equal to

$$\begin{aligned} \frac{n^{1/2} T_5}{b^{p/2}} =& \frac{1}{n} \sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \boldsymbol{g}(\boldsymbol{X}_i)^\top}{\widehat{f}_b(\boldsymbol{u})} (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \sum_{i=1}^n \boldsymbol{g}(\boldsymbol{X}_i)\{\widehat{h}_1(\boldsymbol{X}_i) - h_1(\boldsymbol{X}_i)\} \\ =& \left[\boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} \frac{1}{n} \sum_{i=1}^n \boldsymbol{g}(\boldsymbol{X}_i)\{\widehat{h}_1(\boldsymbol{X}_i) - h_1(\boldsymbol{X}_i)\} \right] \{1 + o_p(1)\} \end{aligned}$$

$$\leq \{1 + o_p(1)\}O_p(1) \max_{j=1,\ldots,n} |\widehat{h}_1(\boldsymbol{X}_j) - h_1(\boldsymbol{X}_j)|$$

the equality follows from Lemma 3 in Dai and Shao (2022). Under (B1)-(B5) with $\boldsymbol{U}$ and p replaced by $\boldsymbol{X}$ and q, respectively, and (C5), Lemma 8.10 in Newey and McFadden (1994) implies that

$$\max |\widehat{h}_1(\boldsymbol{X}_i) - h_1(\boldsymbol{X}_1)| = O_p(\sqrt{\log(n)}n^{-2/(q+4)}), \tag{27}$$

which is $o_p(1/\sqrt{nb^p}) = o_p(n^{-2/(p+4)})$ and, hence, $T_5 = o_p(1)$. From Lemma 3 in Dai and Shao (2022) and the Central Limit Theorem,

$$T_6 = \frac{b^{p/2}}{n^{1/2}} \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\boldsymbol{g}(\boldsymbol{X}_i)^\top}{\widehat{f}_b(\boldsymbol{u})} (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i)\{h_1(\boldsymbol{X}_i) - \mu_1(\boldsymbol{U}_i)\} = O_p(b^{p/2}) = o_p(1).$$

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 3.

I change $\boldsymbol{X}$ to $\boldsymbol{U}$. – Chi

Define

$$\widehat{\omega}_{t1} = \frac{1}{N}\sum_{i=1}^{N} R_i \check{\kappa}_b(\boldsymbol{u} - \boldsymbol{U}_i)e^{tY_i}, \qquad \widehat{\omega}_{t2} = \frac{1}{N}\sum_{i=1}^{N} (1 - R_i)\check{\kappa}_b(\boldsymbol{u} - \boldsymbol{U}_i)Y_i e^{-tY_i},$$
$$\widehat{\omega}_{t3} = \frac{1}{N}\sum_{i=1}^{N} R_i \check{\kappa}_b(\boldsymbol{u} - \boldsymbol{U}_i), \qquad \widehat{\omega}_{t4} = \frac{1}{N}\sum_{i=1}^{N} (1 - R_i)\check{\kappa}_b(\boldsymbol{u} - \boldsymbol{U}_i),$$
$$\widehat{\omega}_{t5} = \frac{1}{N}\sum_{i=1}^{N} R_i \check{\kappa}_b(\boldsymbol{u} - \boldsymbol{U}_i)Y_i e^{tY_i}, \qquad \widehat{\omega}_{t6} = \frac{1}{N}\sum_{i=1}^{N} (1 - R_i)\check{\kappa}_b(\boldsymbol{u} - \boldsymbol{U}_i)Y_i^2 e^{-tY_i},$$
$$\widehat{\omega}_{t7} = \frac{1}{N}\sum_{i=1}^{N} R_i \check{\kappa}_b(\boldsymbol{u} - \boldsymbol{U}_i)Y_i^2 e^{tY_i}, \qquad \widehat{\omega}_{t8} = \frac{1}{N}\sum_{i=1}^{N} (1 - R_i)\check{\kappa}_b(\boldsymbol{u} - \boldsymbol{U}_i)Y_i^3 e^{-tY_i}.$$

Then, $\widehat{h}(\boldsymbol{u},t)=\widehat{\omega}_{t1}\widehat{\omega}_{t2}/\widehat{\omega}_{t3}\widehat{\omega}_{t4}$, $\nabla_t\widehat{h}(\boldsymbol{u},t)=(\widehat{\omega}_{t2}\widehat{\omega}_{t5}-\widehat{\omega}_{t1}\widehat{\omega}_{t6})/\widehat{\omega}_{t3}\widehat{\omega}_{t4}$, and $\nabla_t^2\widehat{h}(\boldsymbol{u},t)=(\widehat{\omega}_{t1}\widehat{\omega}_{t8}-2\widehat{\omega}_{t5}\widehat{\omega}_{t6}+\widehat{\omega}_{t2}\widehat{\omega}_{t7})/\widehat{\omega}_{t3}\widehat{\omega}_{t4}$. Let $L(t)=E[R\{Y-h(\boldsymbol{U},t)\}^2]$, $\widehat{L}_n(t)=N^{-1}\sum_{i=1}^N R_i\{Y_i-\widehat{h}(\boldsymbol{U}_i,t)\}^2$, and $L_n(t)=N^{-1}\sum_{i=1}^N R_i\{Y_i-h(\boldsymbol{U}_i,t)\}^2$. Taking derivatives with respect to t, we obtain

$$\nabla_t\widehat{L}_n(t)=\frac{1}{N}\sum_{i=1}^N -2R_i\{Y_i-\widehat{h}(\boldsymbol{U}_i,t)\}\nabla_t\widehat{h}(\boldsymbol{U}_i,t)=\frac{1}{N}\sum_{i=1}^N\psi\{Y_i,R_i,\widehat{\boldsymbol{\omega}}_t(\boldsymbol{U}_i)\},$$

$$\nabla_t L_n(t)=\frac{1}{N}\sum_{i=1}^N -2R_i\{Y_i-h(\boldsymbol{U}_i,t)\}\nabla_t h(\boldsymbol{U}_i,t)=\frac{1}{N}\sum_{i=1}^N\psi\{Y_i,R_i,\boldsymbol{\omega}_t(\boldsymbol{U}_i)\},$$

and

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

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

Step 1: Since γ is the unique minimizer of $L(t)$, from Theorem 2.1 in Newey and McFadden (1994), it suffices to prove that $\sup_{t\in\Gamma}|\nabla_t\widehat{L}_n(t)-\nabla_t L(t)|\xrightarrow{p}0$. Note that

$$\begin{aligned}\sup_{t\in\Gamma}|\nabla_t\widehat{L}_n(t)-\nabla_t L(t)|\leq&\sup_{t\in\Gamma}|\nabla_t L_n(t)-\nabla_t L(t)|+\sup_{t\in\Gamma}|\nabla_t\widehat{L}_n(t)-\nabla_t L_n(t)|\\ \leq&\sup_{t\in\Gamma}|\nabla_t L_n(t)-\nabla_t L(t)|\\ &+\frac{2}{n}\sum_{i=1}^n R_i|Y_i|\left\{\sup_{t\in\Gamma,\boldsymbol{x}\in\mathscr{U}}|\nabla_t\widehat{h}(\boldsymbol{u},t)-\nabla_t h(\boldsymbol{u},t)|\right.\\ &\left.+\sup_{t\in\Gamma,\boldsymbol{u}\in\mathscr{U}}|\widehat{h}(\boldsymbol{u},t)\nabla_t\widehat{h}(\boldsymbol{u},t)-h(\boldsymbol{u},t)\nabla_t h(\boldsymbol{u},t)|\right\}\end{aligned}$$

From (D3), $|2R\{Y-h(\boldsymbol{u},t)\}\nabla_t h(\boldsymbol{U},t)|$ is bounded by $c|Y|$ for a constant c and hence Lemma 2.4 in Newey and McFadden (1994) implies that $\sup_{t\in\Gamma}|\nabla_t L_n(t)-\nabla_t L(t)|=o_p(1)$. Based on Lemma B.3 in Newey (1994), conditions (D1)-(D4) imply that $\sup_{\boldsymbol{u}\in\mathscr{U}}|\widehat{\boldsymbol{\omega}}_t(\boldsymbol{u})-\boldsymbol{\omega}_t(\boldsymbol{u})|\to 0$ 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 $\sup_{t\in\Gamma,\boldsymbol{u}\in\mathscr{U}}|\widehat{\boldsymbol{\omega}}_t(\boldsymbol{u})-\boldsymbol{\omega}_t(\boldsymbol{u})|\xrightarrow{p}0$. 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 $\sup_{t\in\Gamma,\boldsymbol{u}\in\mathscr{U}}|\nabla_t\widehat{h}(\boldsymbol{u},t)-\nabla_t h(\boldsymbol{u},t)|\xrightarrow{p}0$. 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 $\sqrt{N}\nabla_t\widehat{L}_n(\gamma)\xrightarrow{d}N(0,\sigma_L^2)$ with $\sigma_L^2=\mathrm{Var}\{m(Y,R,\boldsymbol{U},\boldsymbol{\omega}_\gamma)+\tau(Y,R,\boldsymbol{U},\gamma)\}$.

Step 3: Note that $\nabla_t^2 L_n(t)=N^{-1}\sum_{i=1}^N -2R_i\{Y_i-h(\boldsymbol{U}_i,t)\}\nabla_t^2 h(\boldsymbol{U}_i,t)+2R_i\{\nabla_t h(\boldsymbol{U}_i,t)\}^2$ and $\sup_{|t-\gamma|\le|\widehat{\gamma}-\gamma|}|\nabla_t^2\widehat{L}_n(t)-\nabla_t^2 L(\gamma)|\le A_1+A_2+A_3$, where $A_1=|\nabla_t^2 L_n(\gamma)-\nabla_t^2 L(\gamma)|$, $A_2=\sup_{t\in\Gamma}|\nabla_t^2\widehat{L}_n(t)-\nabla_t^2 L_n(t)|$, and the last term $A_3=\sup_{|t-\gamma|\le|\widehat{\gamma}-\gamma|}|\nabla_t^2 L_n(t)-\nabla_t^2 L_n(\gamma)|$. 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

$$\begin{aligned}|\nabla_t^2 L_n(t)-\nabla_t^2 L_n(\gamma)|\le{}&\frac{2}{N}\sum_{i=1}^N|\{\nabla_t h(\boldsymbol{U}_i,t)\}^2-\{\nabla_t h(\boldsymbol{U}_i,\gamma)\}^2|\\&+\frac{2}{N}\sum_{i=1}^N|Y_i||\nabla_t^2 h(\boldsymbol{U}_i,t)-\nabla_t^2 h(\boldsymbol{U}_i,\gamma)|\\&+\frac{2}{N}\sum_{i=1}^N|h(\boldsymbol{U}_i,t)\nabla_t^2 h(\boldsymbol{U}_i,t)-h(\boldsymbol{U}_i,\gamma)\nabla_t^2 h(\boldsymbol{U}_i,\gamma)|.\end{aligned}$$

Under (D3), $h(\cdot,t)$, $\nabla h(\cdot,t)$, and $\nabla h(\cdot,t)$ converge uniformly for all $\boldsymbol{x}$ as $t\to\gamma$ and, thus, the $A_3=o_p(1)$ because $\widehat{\gamma}\xrightarrow{p}\gamma$. This shows that $\sup_{|t-\gamma|\le|\widehat{\gamma}-\gamma|}|\nabla_t^2\widehat{L}_n(t)-\nabla_t^2 L(\gamma)|\xrightarrow{p}0$.

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, $\sqrt{N}(\widehat{\gamma}-\gamma)\xrightarrow{d}N(0,[2E\{R\nabla_\gamma h(\boldsymbol{U},\gamma)\}^2]^{-1}\sigma_L^2)$. This completes the proof of (20).

Proof of Corollary 1. (i) From Theorem 3, (20) shows that $\widehat{\gamma}-\gamma=O_p(1/\sqrt{N})$. Further-

more, Lemma 8.10 in Newey and McFadden (1994) shows that

$$\max_i \left| \frac{e^{-\gamma Y_i} \sum_{j=1}^n e^{\gamma Y_j} \check{\kappa}_\ell(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_{j=1}^n \check{\kappa}_\ell(\boldsymbol{U}_i - \boldsymbol{U}_j)} - \frac{f(Y_i|\boldsymbol{U} = \boldsymbol{U}_i, D = 1)}{f(Y_i|\boldsymbol{U} = \boldsymbol{U}_i, D = 0)} \right| = O_p\left(\sqrt{\frac{\log N}{N\ell^p}} + \ell^d \right), \quad (28)$$

which is $o_p(N^{-2/(p+4)}) = o_p(1)/\sqrt{nb^p}$ under the assumed condition $d > \max\{(p+4)/2, p\}$ and $N\ell^{2d} \to 0$. Since $\widehat{\gamma} - \gamma$ converges faster than (28), (18) holds. As a result, (17) holds with $\mu_1^{E2}(\boldsymbol{u})$ replaced by $\mu_1^{E3}(\boldsymbol{u})$ 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

$$\sup_{\boldsymbol{x} \in \mathbb{X}} |\widehat{h}(\boldsymbol{x}, \gamma) - h_1(\boldsymbol{x})| = O_p\big((\log N)^{1/2}(N\ell^q)^{-1/2} + \ell^d\big) = o_p(n^{-2/(p+4)}).$$

From the asymptotic normality of $\widehat{\gamma}$, $\widehat{\gamma} - \gamma = O_p(1/\sqrt{N})$, which converges to 0 faster than $\sup_{\boldsymbol{x} \in \mathscr{X}} |\widehat{h}(\boldsymbol{x}, \gamma) - h_1(\boldsymbol{x})| \to 0$. Hence (27) 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 2.

Acknowledgements

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*, Volume 1, pp. 99–144.

Chatterjee, N., Y. H. Chen, P. Maas, and R. J. Carroll (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.

Cook, R. D. and S. Weisberg (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association 86*(414), 328–332.

Dai, C.-S. and J. Shao (2022). Kernel regression utilizing external information as constraints. *Statistica Sinica 33*, under revision.

Fan, J., M. Farmen, and I. Gijbels (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(3), 591–608.

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, and J. Engel (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics 49*(1), 79–99.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

Kim, H. J., Z. Wang, and J. K. Kim (2021). Survey data integration for regression analysis using model calibration. *arXiv 2107.06448*.

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*(479), 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 316–327.

Lohr, S. L. and T. E. Raghunathan (2017). Combining survey data with other data sources. *Statistical Science 32*(2), 293–312.

Ma, Y. and L. Zhu (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association 107*(497), 168–179.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association 99*(468), 1131–1139.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications 9*(1), 141–142.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory 10*, 233–253.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B 83*(1), 242–272.

Shao, J. (2003). *Mathematical Statistics* (2nd ed.). Springer, New York.

Shao, Y., R. D. Cook, and S. Weisberg (2007). Marginal tests with sliced average variance estimation. *Biometrika 94*(2), 285–296.

Wand, M. P. and M. C. Jones (1994, December). *Kernel Smoothing*. Number 60 in Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton, FL, U.S.: Chapman & Hall.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, New York.

Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64*(3), 363–410.

Yang, S. and J. K. Kim (2020). Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science 3*(2), 625–650.

Zhang, H., L. Deng, M. Schiffman, J. Qin, and K. Yu (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika 107*(3), 689–703.

Zhang, Y., Z. Ouyang, and H. Zhao (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics 11*(1), 161–184.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association 85*(412), 986–1001.

Table 1: Simulated MISE (21) and IMP (22) with $S = 200$ under $\gamma = 0$

				Estimator					
Covariate	Model	Test data	b, l	$\widehat{\mu}_1$	$\widehat{\mu}_1^{C1}$	$\widehat{\mu}_1^{C2}$	$\widehat{\mu}_1^{C3}$	IMP %	Mean of $\widehat{\gamma}$
Normal	M1	Sample	Best	0.057	0.038	0.051	0.044	33.26	-0.003
			CV	0.073	0.046	0.060	0.055	36.66	-0.006
		Grid	Best	0.054	0.019	0.038	0.030	64.86	-0.012
			CV	0.063	0.036	0.054	0.047	42.10	-0.013
	M2	Sample	Best	0.080	0.073	0.081	0.078	8.75	-0.033
			CV	0.091	0.083	0.093	0.089	8.79	-0.033
		Grid	Best	0.093	0.085	0.095	0.090	8.60	-0.037
			CV	0.110	0.101	0.115	0.108	8.18	-0.043
	M3	Sample	Best	0.077	0.072	0.078	0.076	6.49	-0.016
			CV	0.087	0.080	0.088	0.085	6.97	-0.018
		Grid	Best	0.067	0.059	0.067	0.062	11.94	0.001
			CV	0.087	0.081	0.091	0.087	6.89	-0.006
	M4	Sample	Best	0.061	0.040	0.054	0.047	33.83	-0.004
			CV	0.076	0.051	0.064	0.059	32.94	-0.010
		Grid	Best	0.059	0.022	0.045	0.033	62.17	-0.000
			CV	0.064	0.040	0.060	0.054	36.77	-0.014
Bounded	M1	Sample	Best	0.022	0.004	0.014	0.014	80.43	-0.003
			CV	0.029	0.005	0.014	0.014	80.96	-0.007
		Grid	Best	0.059	0.019	0.034	0.040	69.79	0.008
			CV	0.090	0.042	0.061	0.065	53.57	-0.014
	M2	Sample	Best	0.039	0.032	0.041	0.037	17.94	-0.014
			CV	0.044	0.039	0.047	0.044	11.36	-0.016
		Grid	Best	0.124	0.119	0.126	0.120	16.93	-0.005
			CV	0.158	0.148	0.170	0.157	20.25	-0.032
	M3	Sample	Best	0.033	0.027	0.035	0.034	19.54	-0.006
			CV	0.039	0.033	0.042	0.041	13.63	-0.009
		Grid	Best	0.103	0.097	0.103	0.101	9.34	0.009
			CV	0.130	0.125	0.144	0.139	9.81	-0.017
	M4	Sample	Best	0.023	0.005	0.015	0.014	77.03	-0.003
			CV	0.029	0.006	0.015	0.015	78.59	-0.007
		Grid	Best	0.063	0.025	0.040	0.044	62.57	0.007
			CV	0.096	0.049	0.066	0.072	52.23	-0.015

$\widehat{\mu}_1$: the standard kernel estimator given by (2).
$\widehat{\mu}_1^{C1}$, $\widehat{\mu}_1^{C2}$, and $\widehat{\mu}_1^{C3}$: the estimators given by (16) as described in the end of Section 2.2.
Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.005 and 0.006.

Table 2: Simulated MISE (21) and IMP (22) with $S = 200$ under $\gamma = 1/2$

				Estimator					
Covariate	Model	Test data	b, l	$\widehat{\mu}_1$	$\widehat{\mu}_1^{C1}$	$\widehat{\mu}_1^{C2}$	$\widehat{\mu}_1^{C3}$	IMP %	Mean of $\widehat{\gamma}$
Normal	M1	Sample	Best	0.054	0.318	0.048	0.040	25.26	0.453
			CV	0.068	0.341	0.057	0.051	17.08	0.458
		Grid	Best	0.049	0.190	0.035	0.028	42.18	0.449
			CV	0.056	0.259	0.047	0.040	28.00	0.441
	M2	Sample	Best	0.082	0.509	0.083	0.081	1.07	0.426
			CV	0.093	0.528	0.095	0.091	1.12	0.429
		Grid	Best	0.089	0.588	0.089	0.085	4.21	0.419
			CV	0.099	0.608	0.103	0.098	1.31	0.426
	M3	Sample	Best	0.084	0.560	0.085	0.084	-0.50	0.449
			CV	0.097	0.565	0.101	0.095	1.99	0.442
		Grid	Best	0.070	0.513	0.069	0.067	4.42	0.452
			CV	0.082	0.551	0.087	0.081	0.38	0.456
	M4	Sample	Best	0.063	0.335	0.056	0.051	18.90	0.442
			CV	0.078	0.358	0.064	0.058	18.90	0.432
		Grid	Best	0.053	0.189	0.040	0.033	37.11	0.439
			CV	0.062	0.290	0.060	0.053	14.89	0.440
Bounded	M1	Sample	Best	0.021	0.242	0.015	0.013	36.29	0.480
			CV	0.028	0.241	0.015	0.014	42.94	0.475
		Grid	Best	0.057	0.326	0.035	0.039	31.95	0.483
			CV	0.083	0.346	0.061	0.064	22.69	0.487
	M2	Sample	Best	0.040	0.338	0.042	0.040	0.76	0.478
			CV	0.049	0.350	0.051	0.050	-2.07	0.469
		Grid	Best	0.127	0.555	0.125	0.122	4.15	0.463
			CV	0.155	0.591	0.163	0.154	0.59	0.463
	M3	Sample	Best	0.033	0.328	0.035	0.034	-4.03	0.475
			CV	0.039	0.343	0.041	0.041	-6.37	0.486
		Grid	Best	0.105	0.484	0.102	0.101	4.04	0.483
			CV	0.126	0.519	0.130	0.126	-0.30	0.482
	M4	Sample	Best	0.021	0.243	0.016	0.014	31.91	0.476
			CV	0.030	0.238	0.017	0.015	42.21	0.477
		Grid	Best	0.064	0.344	0.043	0.044	30.67	0.486
			CV	0.086	0.370	0.068	0.068	20.44	0.489

$\widehat{\mu}_1$: the standard kernel estimator given by (2).
$\widehat{\mu}_1^{C1}$, $\widehat{\mu}_1^{C2}$, and $\widehat{\mu}_1^{C3}$: the estimators given by (16) as described in the end of Section 2.2.
Simulation standard deviations of $\widehat{\gamma}$ for all cases are between 0.005 and 0.006.