# KERNEL REGRESSION UTILIZING EXTERNAL INFORMATION AS CONSTRAINTS

BY CHI-SHIAN DAI[1] AND JUN SHAO[*]

[1]*Department of Statistics, University of Wisconsin-Madison cdai39@wisc.edu*

[*]*School of Statistics, East China Normal University and Department of Statistics, University of Wisconsin-Madison shao@stat.wisc.edu*

With advanced technologies in data collection and storage, data analysis in modern scientific research and practice has shifted from analyzing a single dataset to coupling several datasets. Article [3] proposes an approach by formulating some constraints to link a main "internal" dataset and an additional "external" dataset. In this article, we consider nonparametric kernel regression in an internal dataset analysis utilizing constraints for auxiliary information from an external dataset with either summary statistics or individual-level data. We show that the proposed constrained kernel regression estimator is asymptotically normal and is better than the standard kernel regression without using external information in terms of the asymptotic mean integrated square error. Furthermore, we consider the situation where internal and external data have different populations and propose some adjustments to address the difference. Simulation results are obtained to confirm our theory and to quantify the improvements from utilizing external data.

**1. Introduction.** With advanced technologies in data collection and storage, in many modern statistical analyses we have not only primary individual-level data carefully collected from a population of interest but also information from some external datasets, which typically have very large sizes but often contain relatively crude information such as summary statistics or partial individual-level data, due to practical and ethical reasons. Sources of external datasets include, for example, population-based census, administrative datasets, and databases from past investigations. In what follows, the primary individual-level data are referred to as the internal data. Since the internal dataset is obtained to address specific scientific questions, it may contain more measured covariates from each sampled subject and, consequently, its size is much smaller than those of the external datasets due to cost considerations. Thus, there is a growing need for internal data analysis utilizing summary or individual-level information from external datasets. This line of research fits into a more general framework of data integration [13, 18, 20, 24, 33, 34, 35] and is different from the traditional meta-analysis in which the analysis is based on multiple datasets with summary statistics, without an internal individual-level dataset possibly containing more covariates.

In this paper, we study regression between a univariate response variable $Y$ and a covariate vector $\boldsymbol{U}$, based on an internal individual-level dataset in which both $Y$ and $\boldsymbol{U}$ are measured, and an external dataset with summary statistics or individual-level data on $Y$ and $\boldsymbol{X}$, where $\boldsymbol{X}$ is a part of the vector $\boldsymbol{U}$, i.e., $\boldsymbol{U} = (\boldsymbol{X}, \boldsymbol{Z})$, with $\boldsymbol{Z}$ being the part of $\boldsymbol{U}$ not measured in the external dataset due to the high cost of measuring $\boldsymbol{Z}$ or the progress of new technology and/or new scientific relevance for measuring $\boldsymbol{Z}$.

[*]The research was partially supported by the National Natural Science Foundation of China Grant 11831008 and the U.S. National Science Foundation Grant DMS-1914411.

*MSC2020 subject classifications:* Primary 62G08, 62G22; secondary 62P10.

*Keywords and phrases:* Adjustment for heterogeneous populations, constraints, data integration, external data, two-step kernel regression, summary statistics.

Under the same setting and a parametric model between the response $Y$ and covariate vector $\boldsymbol{U}$, [3] proposes a constrained maximum likelihood estimation by utilizing the summary information from an external dataset in the form of constraints added to the observed likelihood for internal data. Other parametric or semiparametric approaches on using information from external datasets can be found, for example, in [2, 4, 6, 13, 14, 23, 26, 31].

We focus on nonparametric kernel regression [1, 29, 30], a well-established approach that does not require any assumption on the regression function between $Y$ and $\boldsymbol{U}$, except for some smoothness conditions. Because of the well-known curse of dimensionality for nonparametric kernel-type methods, we focus on low dimensional covariate $\boldsymbol{U}$. A discussion of handling large dimensional $\boldsymbol{U}$ is given in Section 6.

To make use of summary or individual-level information from the external dataset, we propose a two-step constrained kernel (CK) regression method. In the first step, we apply a constrained optimization procedure to obtain fitted regression value $\widehat{\mu}_i$ at each observed $\boldsymbol{U}_i$ in the internal dataset with sample size $n$, $i = 1, ..., n$, subject to some constraints constructed using summary or individual-level information from the external dataset. As a prediction, $\widehat{\mu}_i$ is usually better than the fitted value at $\boldsymbol{U}_i$ from the standard kernel regression, as it utilizes external information. In the second step, we apply the standard kernel regression treating $\widehat{\mu}_i$'s as the observed $Y$-values to obtain the entire estimated regression function.

To measure the performance of nonparametric regression methods, [9] proposes an asymptotic mean integrated square error (AMISE). In terms of AMISE, we conduct both theoretical and empirical studies on the performance of the proposed CK. The results show that when the sample size of external dataset is at least comparable with the sample size of internal dataset, the CK improves the standard kernel method without using external information. Moreover, the improvement can be substantial.

We organize this paper as follows. Section 2 describes the methodology and establishes the asymptotic normality of CK estimator and its superiority over the standard kernel estimator in AMISE. In Section 3, we consider the same problem when individual-level external data are available. While in Sections 2-3, the assumption that internal and external data share the same population is imposed, in Section 4, we study extensions to situations where this assumption does not hold. Section 5 presents some simulation results as complements to asymptotic results. Section 6 contains some discussions. The proofs of main theorems are given in Section 7 and some technical details are in the Appendix as Supplementary Material.

## 2. The Use of External Summary Statistics.

2.1. *Methodology.* The internal dataset contains individual-level observations $(Y_i, \boldsymbol{U}_i)$, $i = 1, ..., n$, independent and identically distributed (iid) from the population of $(Y, \boldsymbol{U})$, where $Y$ is a univariate response of interest, $\boldsymbol{U}$ is a $p$-dimensional vector of continuous covariates associated with $Y$, $n$ is the sample size of internal dataset, and $p$ is a fixed integer smaller than $n$ and does not vary with $n$. We are interested in the estimation of regression function

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

the conditional expectation of $Y$ given $\boldsymbol{U} = \boldsymbol{u}$, for any $\boldsymbol{u} \in \mathbb{U}$, the range of $\boldsymbol{U}$.

Let $\kappa(\boldsymbol{u})$ be a given kernel function on $\mathbb{R}^p$, where $\mathbb{R}^d$ denotes the $d$-dimensional Euclidean space throughout the paper. We assume that $\boldsymbol{U}$ is standardized so that the same bandwidth $b > 0$ is used for every component of $\boldsymbol{U}$ in kernel regression. The standard kernel regression estimator of $\mu(\boldsymbol{u})$ in (1) for any fixed $\boldsymbol{u} \in \mathbb{U}$ based on the internal dataset is

$$\begin{aligned} \widehat{\mu}_K(\boldsymbol{u}) &= \arg\min_{\mu} \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)(Y_i - \mu)^2 \\ &= \sum_{i=1}^{n} Y_i \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \end{aligned} \tag{2}$$

where $\kappa_b(\boldsymbol{a}) = b^{-p}\kappa(\boldsymbol{a}/b)$, $\boldsymbol{a} \in \mathbb{R}^p$.

The external dataset is another iid sample of size $m$ from the population of $(Y, \boldsymbol{X})$ (the same population for internal data; extensions are considered in Section 4), independent of the internal sample, where $\boldsymbol{X}$ is a $q$-dimensional sub-vector of $\boldsymbol{U}$, $q \leq p$. In this section, we consider the scenario where only some summary statistics are available from the external dataset. Specifically, the external dataset provides the vector $\widehat{\boldsymbol{\beta}}_g$ of least squares estimate of $\boldsymbol{\beta}$ based on external data under a working model $\mathrm{E}(Y|\boldsymbol{X}) = \boldsymbol{\beta}^\top \boldsymbol{g}(\boldsymbol{X})$ (not necessarily correct), where $\boldsymbol{g}$ is a known function from $\mathbb{R}^q$ to $\mathbb{R}^k$ with a fixed $k$ and $\boldsymbol{a}^\top$ denotes the transpose of vector $\boldsymbol{a}$ throughout the paper.

Regardless of whether the working model is correct or not, the asymptotic limit of $\widehat{\boldsymbol{\beta}}_g$ is $\boldsymbol{\beta}_g = \boldsymbol{\Sigma}_g^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})Y\}$ under some moment conditions, where $\boldsymbol{\Sigma}_g = \mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top\}$ is assumed to be finite and positive definite. From $\mathrm{E}(Y|\boldsymbol{X}) = \mathrm{E}\{\mathrm{E}(Y|\boldsymbol{U})|\boldsymbol{X}\} = \mathrm{E}\{\mu(\boldsymbol{U})|\boldsymbol{X}\}$, we obtain that

$$\begin{aligned} \mathrm{E}\{\boldsymbol{\beta}_g^\top \boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top\} &= \mathrm{E}\{Y\boldsymbol{g}(\boldsymbol{X})^\top\}\boldsymbol{\Sigma}_g^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top\} \\ &= \mathrm{E}\{\mathrm{E}(Y|\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top\} \\ &= \mathrm{E}[\mathrm{E}\{\mu(\boldsymbol{U})|\boldsymbol{X}\}\boldsymbol{g}(\boldsymbol{X})^\top] \\ &= \mathrm{E}\{\mu(\boldsymbol{U})\boldsymbol{g}(\boldsymbol{X})^\top\}. \end{aligned} \tag{3}$$

Hence, the summary information from external dataset can be utilized through the constraint

$$\mathrm{E}[\{\boldsymbol{\beta}_g^\top \boldsymbol{g}(\boldsymbol{X}) - \mu(\boldsymbol{U})\}\boldsymbol{g}(\boldsymbol{X})^\top] = 0. \tag{4}$$

We propose a two-step procedure. In the first step, we make use of (4) and the external information to obtain predicted values $\widehat{\mu}_1, ..., \widehat{\mu}_n$ of $\mu(\boldsymbol{U}_1), ..., \mu(\boldsymbol{U}_n)$, respectively, to improve $\widehat{\mu}_K(\boldsymbol{U}_1), ..., \widehat{\mu}_K(\boldsymbol{U}_n)$ from the standard kernel regression. To achieve this, we estimate $\boldsymbol{\mu} = (\mu(\boldsymbol{U}_1), ..., \mu(\boldsymbol{U}_n))^\top$ by the $n$-dimensional vector $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, ..., \widehat{\mu}_n)^\top$ that is the solution to the following constrained minimization,

$$\widehat{\boldsymbol{\mu}} = \arg \min_{(\mu_1,...,\mu_n)^\top \in \mathbb{R}^n} \sum_{i=1}^n \sum_{j=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)(Y_j - \mu_i)^2 \Big/ \sum_{k=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k) \tag{5}$$

$$\text{subject to} \quad \sum_{i=1}^n \{\widehat{\boldsymbol{\beta}}_g^\top \boldsymbol{g}(\boldsymbol{X}_i) - \mu_i\}\boldsymbol{g}(\boldsymbol{X}_i)^\top = 0, \tag{6}$$

where the constraint in (6) is an empirical analog of (4) for the estimation of $\boldsymbol{\mu}$ based on internal data and $l$ in (5) is a bandwidth that may be different from $b$ in (2). A discussion about the selection of bandwidths is given in Section 2.3.

To motivate the objective function in (5) being minimized, note that

$$\sum_{j=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)\{Y_j - \mu(\boldsymbol{U}_i)\}^2 \Big/ \sum_{k=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k) \approx \mathrm{E}[\{Y - \mu(\boldsymbol{U})\}^2 | \boldsymbol{U} = \boldsymbol{U}_i]$$

for each $i$ and, hence, the objective function in (5) divided by $n$ approximates

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}[\{Y - \mu(\boldsymbol{U})\}^2 | \boldsymbol{U} = \boldsymbol{U}_i] \approx \mathrm{E}[\{Y - \mu(\boldsymbol{U})\}^2].$$

To derive an explicit form of $\widehat{\boldsymbol{\mu}}$ in (5), let $\boldsymbol{G}$ be the $n \times n$ matrix whose $i$th row is $\boldsymbol{g}(\boldsymbol{X}_i)^\top$ and let $\widehat{\boldsymbol{h}}$ and $\widehat{\boldsymbol{\mu}}_K$ be the $n$-dimensional vectors whose $i$th components are $\widehat{\boldsymbol{\beta}}_g^\top \boldsymbol{g}(\boldsymbol{X}_i)$ and $\widehat{\mu}_K(\boldsymbol{U}_i)$, respectively, with $\widehat{\mu}_K$ being defined by (2). Then solving (5)-(6) is the same as solving

$$\widehat{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{\nu} \in \mathbb{R}^n} (\boldsymbol{\nu}^\top \boldsymbol{\nu} - 2\boldsymbol{\nu}^\top \widehat{\boldsymbol{\mu}}_K) \quad \text{subject to} \quad \boldsymbol{G}^\top(\boldsymbol{\nu} - \widehat{\boldsymbol{h}}) = 0.$$

From the Lagrange multiplier $L(\boldsymbol{\nu}, \boldsymbol{\lambda}) = \boldsymbol{\nu}^\top \boldsymbol{\nu} - 2\boldsymbol{\nu}^\top \widehat{\boldsymbol{\mu}}_K + 2\boldsymbol{\lambda}^\top \boldsymbol{G}^\top (\boldsymbol{\nu} - \widehat{\boldsymbol{h}})$ and $\nabla_{\boldsymbol{\nu}} L(\boldsymbol{\nu}, \boldsymbol{\lambda}) = 2\boldsymbol{\nu} - 2\widehat{\boldsymbol{\mu}}_K + 2\boldsymbol{G}\boldsymbol{\lambda}$, we obtain that $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_K - \boldsymbol{G}\boldsymbol{\lambda}$. From the constraint, $\boldsymbol{G}^\top \widehat{\boldsymbol{h}} = \boldsymbol{G}^\top \widehat{\boldsymbol{\mu}} = \boldsymbol{G}^\top \widehat{\boldsymbol{\mu}}_K - \boldsymbol{G}^\top \boldsymbol{G}\boldsymbol{\lambda}$. Solving for $\boldsymbol{\lambda}$, we obtain that $\boldsymbol{\lambda} = (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top \widehat{\boldsymbol{\mu}}_K - (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top \widehat{\boldsymbol{h}}$. Hence, $\widehat{\boldsymbol{\mu}}$ has an explicit form

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_K + \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top (\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\mu}}_K). \tag{7}$$

This estimator adds an adjustment term to $\widehat{\boldsymbol{\mu}}_K$, the estimator in (2) from the standard kernel regression. The adjustment involves the difference $\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\mu}}_K$ and the projection matrix $\boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top$. Since the additional information from the external dataset is used in constraint (6), $\widehat{\boldsymbol{\mu}}$ in (7) is expected to be better than $\widehat{\boldsymbol{\mu}}_K$ that does not use external information. Proposition 2.1 in Section 2.2 quantifies this improvement.

To obtain an improved estimator of the entire regression function $\mu(\boldsymbol{u})$ defined by (1), not just the function $\mu(\boldsymbol{u})$ at $\boldsymbol{U}_1, ..., \boldsymbol{U}_n$, we propose a second step to apply the standard kernel regression with responses $Y_1, ..., Y_n$ replaced by $\widehat{\mu}_1, ..., \widehat{\mu}_n$. Specifically, our proposed estimator of $\mu(\boldsymbol{u})$ is

$$\widehat{\mu}_{CK}(\boldsymbol{u}) = \sum_{i=1}^n \widehat{\mu}_i \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^n \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i), \tag{8}$$

where $b$ is the same bandwidth in (2).

If we apply kernel regression with $\widehat{\mu}_1, ..., \widehat{\mu}_n$ in (8) replaced respectively by $\widehat{\mu}_1^{(0)}, ..., \widehat{\mu}_n^{(0)}$ defined as the solution to minimization in (5) without applying the constraint $\boldsymbol{G}^\top (\boldsymbol{\nu} - \widehat{\boldsymbol{h}}) = 0$, i.e., $\widehat{\mu}_i^{(0)}$ is the estimator (2) at $\boldsymbol{u} = \boldsymbol{U}_i$ but with $b$ replaced by $l$, then we obtain another estimator

$$\widehat{\mu}_{DK}(\boldsymbol{u}) = \sum_{i=1}^n \widehat{\mu}_i^{(0)} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^n \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i). \tag{9}$$

It does not use external information, but is obtained by applying kernel regression twice, and will be referred to as double kernel regression estimator. Intuitively, $\widehat{\mu}_{DK}$ should not be better or worse than the standard $\widehat{\mu}_K$, as no additional information is utilized in (9). In the asymptotic theory presented next, we show that $\widehat{\mu}_{DK}$ is asymptotically equivalent to the standard kernel regression using a kernel different from $\kappa$.

2.2. *Asymptotic Theory.* We now establish some asymptotic results (as the sample size $n$ of the internal dataset increases to infinity), which enables us to compare three estimators in (2), (8), and (9).

The first result, shown in the Appendix (Supplementary Material), establishes that, with probability tending to 1 as $n \to \infty$, $\|\widehat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}\|^2$ is larger than $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$, where $\|\boldsymbol{a}\|^2 = \boldsymbol{a}^\top \boldsymbol{a}$ for vector $\boldsymbol{a}$ throughout this paper.

PROPOSITION 2.1. *Assume the conditions in Theorem 2.1. Then,*

$$\frac{\|\widehat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}\|^2 - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2}{nb^4} \xrightarrow{p} \mathrm{E}\{A(\boldsymbol{U})\boldsymbol{g}(\boldsymbol{X})^\top\} \boldsymbol{\Sigma}_g^{-1} \mathrm{E}\{A(\boldsymbol{U})\boldsymbol{g}(\boldsymbol{X})\},$$

*where $\xrightarrow{p}$ denotes convergence in probability as $n \to \infty$ and $A(\boldsymbol{u})$ is defined in (11).*

Our main result is the asymptotic normality of $\widehat{\mu}_{CK}(\boldsymbol{u})$ and $\widehat{\mu}_{DK}(\boldsymbol{u})$ in (8)-(9) for a fixed $\boldsymbol{u}$, under regularity conditions (A1)-(A5). The proof of Theorem 2.1 is deferred to Section 7 with some technical details given in the Appendix as Supplementary Material.

THEOREM 2.1. *Assume the following conditions.*

(A1) *The response $Y$ has a finite $\mathrm{E}|Y|^s$ with $s > 2 + p/2$ and $\boldsymbol{\Sigma}_g = \mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top\}$ is positive definite. The covariate vector $\boldsymbol{U}$ has a compact support $\mathbb{U} \subset \mathbb{R}^p$. The density of $\boldsymbol{U}$ is bounded away from infinity and zero on $\mathbb{U}$, and has bounded second-order derivatives.*

(A2) *Functions $\mu(\boldsymbol{u}) = \mathrm{E}(Y|\boldsymbol{U} = \boldsymbol{u})$, $\sigma^2(\boldsymbol{u}) = \mathrm{E}[\{Y - \mu(\boldsymbol{U})\}^2|\boldsymbol{U} = \boldsymbol{u}]$, and $\boldsymbol{g}(\boldsymbol{x})$ are Lipschitz continuous; $\mu(\boldsymbol{u})$ has bounded third-order derivatives; and $\mathrm{E}(|Y|^s|\boldsymbol{U} = \boldsymbol{u})$ is bounded.*

(A3) *The kernel $\kappa$ is a positive, bounded, and Lipschitz continuous density with mean zero and finite sixth moments.*

(A4) *The bandwidths $b$ in (2) and $l$ in (5) satisfy $b \to 0$, $l \to 0$, $l/b \to r \in (0, \infty)$, and $nb^{4+p} \to c \in [0, \infty)$, as the internal sample size $n \to \infty$.*

(A5) *The external sample size $m$ satisfies $n = O(m)$, i.e., $n/m$ is bounded by a fixed constant.*

*Then, for any $\boldsymbol{u} \in \mathbb{U}$,*

$$\sqrt{nb^p}\{\widehat{\mu}_t(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \xrightarrow{d} N\big(B_t(\boldsymbol{u}), V_t(\boldsymbol{u})\big), \tag{10}$$

*where $\xrightarrow{d}$ denotes convergence in distribution as $n \to \infty$, $t = DK$ or $CK$,*

$$\begin{aligned}
B_{DK}(\boldsymbol{u}) &= c^{1/2}(1 + r^2)A(\boldsymbol{u}), \\
B_{CK}(\boldsymbol{u}) &= c^{1/2}[(1 + r^2)A(\boldsymbol{u}) - r^2\boldsymbol{g}(\boldsymbol{X})^\top\boldsymbol{\Sigma}_g^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}], \\
A(\boldsymbol{u}) &= \int \kappa(\boldsymbol{v})\left\{\tfrac{1}{2}\boldsymbol{v}^\top\nabla^2\mu(\boldsymbol{u})\boldsymbol{v} + \boldsymbol{v}^\top\nabla \log f_U(\boldsymbol{u})\nabla\mu(\boldsymbol{u})^\top\boldsymbol{v}\right\}d\boldsymbol{v}, \\
V_{CK}(\boldsymbol{u}) &= V_{DK}(\boldsymbol{u}) = \frac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})}\int\left\{\int \kappa(\boldsymbol{v} - r\boldsymbol{w})\kappa(\boldsymbol{w})d\boldsymbol{w}\right\}^2 d\boldsymbol{v},
\end{aligned} \tag{11}$$

*$\boldsymbol{\Sigma}_g = \mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top\}$, and $f_U$ is the density of $\boldsymbol{U}$.*

REMARK 2.1. *(A1) is stronger than the usual condition in the theory of kernel regression, which only requires $s > 2$ and the density $f_U$ is positive on $\mathbb{U}$. We need this stronger assumption to control the efficiency of $n$ estimates in the first step of our procedure.*

REMARK 2.2. *Under (A5), $m$ is at least comparable with $n$, and $m \to \infty$ as $n \to \infty$. Theorem 2.1 indicates that the use of external information does not improve the convergence rate $1/\sqrt{nb^p}$ in estimating $\mu(\boldsymbol{u})$, regardless of whether external information is used and what $m$ is. This is because only the internal dataset with size $n$ has information on $\boldsymbol{Z}$.*

Although the convergence rate and limiting variance $V_t(\boldsymbol{u})$ of $\widehat{\mu}_t(\boldsymbol{u})$ are the same for $t = CK$ and $DK$, the result in Theorem 2.1 indicates that the use of external information affects the asymptotic bias $B_t(\boldsymbol{u})$. Let $D_g(\boldsymbol{X}) = \boldsymbol{g}(\boldsymbol{X})^\top\boldsymbol{\Sigma}_g^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}$. Then

$$\mathrm{E}[\{A(\boldsymbol{U}) - D_g(\boldsymbol{X})\}D_g(\boldsymbol{X})] = 0$$

and, consequently,

$$\begin{aligned}
\mathrm{E}\{B_{CK}(\boldsymbol{U})\}^2 &= c\,\mathrm{E}[A(\boldsymbol{U}) + r^2\{A(\boldsymbol{U}) - D_g(\boldsymbol{X})\}]^2 \\
&= c\,\mathrm{E}\{A(\boldsymbol{U})\}^2 + cr^2(2 + r^2)\mathrm{E}\{A(\boldsymbol{U}) - D_g(\boldsymbol{X})\}^2 \\
&\le c\,\mathrm{E}\{A(\boldsymbol{U})\}^2 + cr^2(2 + r^2)\mathrm{E}\{A(\boldsymbol{U})\}^2 \\
&= \mathrm{E}\{B_{DK}(\boldsymbol{U})\}^2
\end{aligned}$$

with equality holds if and only if $c=0$ or $D_g(\boldsymbol{X})=0$ (i.e., $\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}=0$). Therefore, if $c>0$ and $\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}\neq 0$, $\widehat{\mu}_{CK}(\boldsymbol{u})$ is asymptotically less biased than $\widehat{\mu}_{DK}(\boldsymbol{u})$ and consequently is better than $\widehat{\mu}_{DK}(\boldsymbol{u})$ in terms of the asymptotic mean integrated square error (AMISE) defined as

$$\mathrm{AMISE}(\widehat{\mu}_t)=\mathrm{E}[\{B_t(\boldsymbol{U})\}^2+V_t(\boldsymbol{U})],\quad t=CK \text{ or } DK,$$

a globe accuracy measure considered in the literature (see, e.g., [9]). A precise comparison between $\widehat{\mu}_{CK}$ and $\widehat{\mu}_{DK}$ can be made as follows:

$$\frac{\mathrm{AMISE}(\widehat{\mu}_{DK})-\mathrm{AMISE}(\widehat{\mu}_{CK})}{\mathrm{AMISE}(\widehat{\mu}_{DK})}=\frac{cr^2(2+r^2)\mathrm{E}[\boldsymbol{g}(\boldsymbol{X})^\top\boldsymbol{\Sigma}_g^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}]^2}{c(1+r^2)^2\mathrm{E}\{A(\boldsymbol{U})\}^2+\mathrm{E}\{V_{DK}(\boldsymbol{U})\}}$$

This ratio measures the efficiency improvement due to external information and is always between 0 and 1. For any fixed $r$, the efficiency improvement is an increasing function of $c$ and a decreasing function of $\mathrm{E}\{V_{DK}(\boldsymbol{U})\}$. Since $c$ is the limit of $nb^{4+p}$ according to (A3), and in applications, the bandwidth $b$ is often chosen to minimize variability, the value $c$ is often related with the variance $\mathrm{E}\{V_{DK}(\boldsymbol{U})\}$. For example, when $\sigma^2(\boldsymbol{u})=\sigma^2$ does not depend on $\boldsymbol{u}$, Theorem 4.2 in [7] shows that the optimal bandwidth is the one with $c=c_0\sigma^2$ for a constant $c_0>0$, in which case the efficiency improvement is

$$\frac{r^2(2+r^2)\mathrm{E}[\boldsymbol{g}(\boldsymbol{X})^\top\boldsymbol{\Sigma}_g^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}]^2}{(1+r^2)^2\mathrm{E}\{A(\boldsymbol{U})\}^2+c_0\int\left\{\int\kappa(\boldsymbol{v}-r\boldsymbol{w})\kappa(\boldsymbol{w})d\boldsymbol{w}\right\}^2d\boldsymbol{v}}.$$

As $r\to\infty$, it becomes $\mathrm{E}[\boldsymbol{g}(\boldsymbol{X})^\top\boldsymbol{\Sigma}_g^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}]^2/\mathrm{E}\{A(\boldsymbol{U})\}]^2$.

From the theory of standard kernel regression (e.g., [22]), under (A1)-(A4), the kernel estimator $\widehat{\mu}_K(\boldsymbol{u})$ in (2) also satisfies (10) with $t=K$,

$$B_K(\boldsymbol{u})=c^{1/2}A(\boldsymbol{u})\quad\text{and}\quad V_K(\boldsymbol{u})=\frac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})}\int\{\kappa(\boldsymbol{v})\}^2d\boldsymbol{v}.$$

If we replace the kernel $\kappa(\boldsymbol{u})$ in $\widehat{\mu}_K(\boldsymbol{u})$ with the convolution kernel $\int\kappa(\boldsymbol{v}-r\boldsymbol{w})\kappa(\boldsymbol{w})d\boldsymbol{w}$ and denote its asymptotic bias and variance by $\widetilde{B}_K(\boldsymbol{u})$ and $\widetilde{V}_K(\boldsymbol{u})$, respectively, then immediately we see that $\widetilde{V}_K(\boldsymbol{u})=V_{DK}(\boldsymbol{u})$. Also, replacing the kernel $\kappa$ in $A(\boldsymbol{u})$ with the convolution kernel and denoting $\nabla^2\mu(\boldsymbol{u})/2+\nabla\log f_U(\boldsymbol{u})\nabla\mu(\boldsymbol{u})^\top$ by $\boldsymbol{D}(\boldsymbol{u})$, we obtain that

$$\begin{aligned}\widetilde{B}_K(\boldsymbol{u})&=\sqrt{c}\int\left\{\int\kappa(\boldsymbol{v}-r\boldsymbol{w})\kappa(\boldsymbol{w})d\boldsymbol{w}\right\}\boldsymbol{v}^\top\boldsymbol{D}(\boldsymbol{u})\boldsymbol{v}d\boldsymbol{v}\\&=\sqrt{c}\int\left\{\int\kappa(\boldsymbol{z})(\boldsymbol{z}+r\boldsymbol{w})^\top\boldsymbol{D}(\boldsymbol{u})(\boldsymbol{z}+r\boldsymbol{w})d\boldsymbol{z}\right\}\kappa(\boldsymbol{w})d\boldsymbol{w}\\&=\sqrt{c}(1+r^2)A(\boldsymbol{u})\\&=B_{DK}(\boldsymbol{u}),\end{aligned}$$

where the second equality holds by switching the order of integration and changing $\boldsymbol{z}=\boldsymbol{v}-r\boldsymbol{w}$ and the third equality holds because $\int\kappa(\boldsymbol{v})d\boldsymbol{v}=0$ and $A(\boldsymbol{u})=\int\kappa(\boldsymbol{v})\boldsymbol{v}^\top\boldsymbol{D}(\boldsymbol{u})\boldsymbol{v}d\boldsymbol{v}$. In other words, $\widehat{\mu}_{DK}(\boldsymbol{u})$ is asymptotically equivalent to the standard kernel regression estimator with the convolution kernel, in terms of the AMISE.

A comparison between $\widehat{\mu}_{CK}$ in (8) and $\widehat{\mu}_K$ in (2) is given in the following theorem whose proof is deferred to Section 7.

THEOREM 2.2. *Under the conditions in Theorem 2.1 and an additional condition that* $\int\nabla^2\kappa(\boldsymbol{u})\kappa(\boldsymbol{u})d\boldsymbol{u}$ *being strictly negative definite,* $\mathrm{AMISE}(\widehat{\mu}_{CK})<\mathrm{AMISE}(\widehat{\mu}_K)$ *for* $c$ *and* $r$ *in a neighborhood of 0.*

EXAMPLE 2.1 (Gaussian kernels). The Gaussian kernel $\kappa(\boldsymbol{u}) = (2\pi)^{-p/2}e^{-\|\boldsymbol{u}\|^2/2}$ is the density of $N(0, \boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the identity matrix of order $p$. The Gaussian kernel satisfies the condition on $\kappa$ in Theorem 2.2, as

$$\int \nabla^2\kappa(\boldsymbol{u})\kappa(\boldsymbol{u})d\boldsymbol{u} = \frac{1}{(2\pi)^p}\int \left(\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{I}_p\right) e^{-\|\boldsymbol{u}\|^2} d\boldsymbol{u} = -\frac{1}{2^{1+p/2}(2\pi)^{p/2}}\boldsymbol{I}_p$$

is strictly negative definite. In this example, the convolution kernel $\int \kappa(\boldsymbol{v} - r\boldsymbol{w})\kappa(\boldsymbol{w})d\boldsymbol{w}$ is the density of $N(0, (1 + r^2)\boldsymbol{I}_p)$.

2.3. *Bandwidth Selection.* (A4) in Theorem 2.1 provides the rates of the bandwidths $l$ and $b$ for $\widehat{\mu}_{CK}$. In an application, we need to choose $l$ and $b$ with a given sample size $n$. We apply the following $K$-fold cross-validation (CV) as described in [11]. Let $\mathcal{G}_1, ..., \mathcal{G}_K$ be a random partition of the internal dataset with approximately equal sizes $n_1, ..., n_K$, and let $\widehat{\mu}_{CK}^{(-k)}(\boldsymbol{u})$ be the kernel regression estimator (8) with bandwidths $l$ and $b$ but without using data $\{(Y_i, \boldsymbol{U}_i), i \in \mathcal{G}_k\}$, $k = 1, ..., K$. Then, over a reasonable range, we select $(l, b)$ that minimizes

$$\mathrm{CV}(l, b) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in\mathcal{G}_k}\{\widehat{\mu}_{CK}^{(-k)}(\boldsymbol{U}_i) - Y_i\}^2.$$

**3. The Use of External Individual-Level Data.** In some applications, the individual-level data, $\{Y_{n+j}, \boldsymbol{X}_{n+j}, j = 1, ..., m\}$, are available from the external dataset that is an iid sample, independent of the internal data. Again, only $\boldsymbol{X}$, part of the entire covariate vector $\boldsymbol{U}$, is measured in the external dataset. Although we consider the frame work of internal and external datasets, our result also covers the scenario where we have a single dataset with first $n$ observations $(Y_1, \boldsymbol{U}_1), ..., (Y_n, \boldsymbol{U}_n)$ and another $m$ observations $(Y_{n+1}, \boldsymbol{X}_{n+1})$, ...,$(Y_{n+m}, \boldsymbol{X}_{n+m})$ without $\boldsymbol{Z}$-values due to the fact that the measurement of $\boldsymbol{Z}$ is difficult or expensive, or $\boldsymbol{Z}$-values for all $j > n$ are missing.

Like in Section 2, in this section we still assume the same population for internal and external data. Under this assumption, if $p = q$ (i.e., $\boldsymbol{U} = \boldsymbol{X}$ and $\boldsymbol{Z}$ is degenerated), then we can simply combine the internal and external datasets into a single dataset of size $N = n+m$. Therefore, we focus on the scenario of $q < p$ in this section.

With external individual data and $q < p$, results obtained in Section 2 still hold with $\widehat{\boldsymbol{\beta}}_g$ calculated based on external individual data and any $\boldsymbol{g}$. The difference is that, with available external individual data, we are able to choose different $\boldsymbol{g}$'s to construct $\widehat{\boldsymbol{\beta}}_g$'s in constraint (6). Further, with available external individual data we can allow heterogeneity in populations of internal and external data, which is considered in Section 4.

We now consider the choice of $\boldsymbol{g}$ in constraint (4) or (6). Even if we can find a correct working model $\mathrm{E}(Y|\boldsymbol{X}) = \boldsymbol{\beta}^\top\boldsymbol{g}(\boldsymbol{X})$ with a specific function $\boldsymbol{g}$, it does not necessarily lead to a CK estimator having the smallest AMISE among different $\boldsymbol{g}$'s, as our next result shows.

Note that $\boldsymbol{g}$ does not affect the asymptotic variance $V_{CK}$, but it affects the asymptotic bias $B_{CK}$ through the following identity established in Section 2.2,

$$\mathrm{E}\{B_{CK}(\boldsymbol{U})\}^2 = c\,\mathrm{E}\{A(\boldsymbol{U})\}^2 + cr^2(2 + r^2)\mathrm{E}\{A(\boldsymbol{U}) - D_g(\boldsymbol{X})\}^2,$$

where $D_g(\boldsymbol{X}) = \boldsymbol{g}(\boldsymbol{X})^\top\boldsymbol{\Sigma}_g^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}$ and the function $A(\boldsymbol{u})$ is defined by (11). From this identity, the best $\boldsymbol{g}$ is the one minimizing $\mathrm{E}\{A(\boldsymbol{U}) - D_g(\boldsymbol{X})\}^2$. From the theory of conditional expectation, we should choose $D_g(\boldsymbol{X}) = \mathrm{E}\{A(\boldsymbol{U})|\boldsymbol{X}\}$, i.e., $\boldsymbol{g}$ should be the one-dimensional function $g^*(\boldsymbol{X}) = \mathrm{E}\{A(\boldsymbol{U})|\boldsymbol{X}\}$.

Unfortunately, the best function $g^*(\boldsymbol{x}) = \mathrm{E}\{A(\boldsymbol{U})|\boldsymbol{X} = \boldsymbol{x}\}$ is typically unknown. In the following, we propose an estimator of $g^*(\boldsymbol{X})$ and study the asymptotic property of $\widehat{\mu}_{CK}$ with the estimated function $g^*$.

First, we construct an estimator $\widehat{A}(\boldsymbol{u})$ of $A(\boldsymbol{u})$ in (11). Suppose that the kernel $\kappa$ has the property that, for any components $u_k$ and $u_j$ of $\boldsymbol{u}$, $\int u_k u_j \kappa(\boldsymbol{u})d\boldsymbol{u} = 0$ when $k \neq j$, and $\int u_k^2 \kappa(\boldsymbol{u})d\boldsymbol{u} = 1$. Then the function $A(\boldsymbol{u})$ in (11) has the form

$$A(\boldsymbol{u}) = \frac{1}{2}\sum_{k=1}^{p}\left\{\nabla_{kk}^2\mu(\boldsymbol{u}) + \frac{2\nabla_k\mu(\boldsymbol{u})\nabla_k f_U(\boldsymbol{u})}{f_U(\boldsymbol{u})}\right\}$$

$$= \frac{1}{2}\sum_{k=1}^{p}\left\{\frac{\nu_k(\boldsymbol{u})}{f_U(\boldsymbol{u})} - \frac{\nu_0(\boldsymbol{u})\nabla_{kk}^2 f_U(\boldsymbol{u})}{f_U^2(\boldsymbol{u})}\right\},$$

where $\nu_k(\boldsymbol{u}) = \nabla_{kk}^2\{\mu(\boldsymbol{u})f_U(\boldsymbol{u})\}$, $\nu_0(\boldsymbol{u}) = \mu(\boldsymbol{u})f_U(\boldsymbol{u})$, $\nabla_k$ denotes the $k$th component of $\nabla$, and $\nabla_{kk}^2$ denotes the $k$th diagonal element of $\nabla^2$. We then obtain an estimator $\widehat{A}(\boldsymbol{u})$ by estimating $f_U(\boldsymbol{u})$, $\nu_0(\boldsymbol{u})$, $\nu_k(\boldsymbol{u})$, and $\nabla_{kk}^2 f_U(\boldsymbol{u})$, $k = 1, \ldots, p$, with

$$\widehat{f}_U(\boldsymbol{u}) = \frac{1}{n\lambda_1^p}\sum_{i=1}^{n}\widetilde{\kappa}\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{\lambda_1}\right) \tag{12}$$

$$\widehat{\nu}_0(\boldsymbol{u}) = \frac{1}{n\lambda_1^p}\sum_{i=1}^{n}\widetilde{\kappa}\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{\lambda_1}\right)Y_i$$

$$\widehat{\nu}_k(\boldsymbol{u}) = \frac{1}{n\lambda_1^{p+2}}\sum_{i=1}^{n}\nabla_{kk}^2\widetilde{\kappa}\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{\lambda_1}\right)Y_i$$

$$\nabla_{kk}^2\widehat{f}_U(\boldsymbol{u}) = \frac{1}{n\lambda_1^{p+2}}\sum_{i=1}^{n}\nabla_{kk}^2\widetilde{\kappa}\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{\lambda_1}\right),$$

respectively, where $\widetilde{\kappa}$ is a twice differentiable kernel and $\lambda_1$ is a bandwidth. Then, we apply kernel regression to estimate $g^*(\boldsymbol{x})$ by

$$\widehat{g}^*(\boldsymbol{x}) = \sum_{j=1}^{n}\kappa_{\lambda_2}(\boldsymbol{x}-\boldsymbol{X}_j)\widehat{A}(\boldsymbol{U}_j)\Big/\sum_{j=1}^{n}\kappa_{\lambda_2}(\boldsymbol{x}-\boldsymbol{X}_j) \tag{13}$$

and use this $\widehat{g}^*$ as the function $\boldsymbol{g}$ in (6) to obtain $\widehat{\mu}_{CK}$ in (8), where $\lambda_2$ is another bandwidth.

We establish the following result for $\widehat{\mu}_{CK}$ based on $\widehat{g}^*$. Its proof is deferred to Section 7.

THEOREM 3.1. *Assume the conditions in Theorem 2.1, $q < p$, and the following additional conditions.*

(B1) *The kernel $\kappa$ in (A3) satisfies $\int u_k^2\kappa(\boldsymbol{u})d\boldsymbol{u} = 1$ and $\int u_k u_j\kappa(\boldsymbol{u})d\boldsymbol{u} = 0$ when $k \neq j$. The kernel $\widetilde{\kappa}$ in (12) is zero outside a compact set and has second-order derivatives that are Lipschitz continuous. Further, $\int\widetilde{\kappa}(\boldsymbol{u})d\boldsymbol{u} = 1$ and there is an integer $\varsigma > 2 + 8/p$ such that for all $j < \varsigma$, $\int(\sum_{k=1}^{p}u_k)^j\widetilde{\kappa}(\boldsymbol{u})d\boldsymbol{u} = 0$.*

(B2) *The bandwidth $\lambda_1$ in (12) has order $n^{-1/(p+4+2\varsigma)}$ and $\lambda_2$ in (13) has order $n^{-1/(q+4)}$.*

(B3) *For some constant $s > \max\{2 + p/2, 4\}$, $\mathrm{E}|Y|^s < \infty$ and $\mathrm{E}(|Y|^s|\boldsymbol{X} = \boldsymbol{x})f_U(\boldsymbol{x})$ is bounded. Further, $\mu(\boldsymbol{u})\,f_U(\boldsymbol{u})$ and $f_U(\boldsymbol{u})$ are continuously differentiable up to order $2+\varsigma$.*

*Then, the result in Theorem 2.1 with $\boldsymbol{g} = g^*$ holds for $\widehat{\mu}_{CK}$ using the estimated $\widehat{g}^*$ in (13).*

Although $g^*$ is theoretically the best choice for constraint (4), the estimator $\widehat{g}^*$ is complicated as it involves the estimation of a second-order gradient. Furthermore, the estimation of $g^*$ has to use $\boldsymbol{U}$-data from the internal dataset with a sample size that may be much smaller than the size of the external dataset. Thus, the estimator $\widehat{\mu}_{CK}$ using $\widehat{g}^*$ may not perform well for finite sample size $n$ although it is asymptotically optimal. We recommend the simple alternative $\boldsymbol{g}(\boldsymbol{x}) = (1, \boldsymbol{x})^\top$, which is asymptotically justified by Theorem 2.1 and performs well in the simulation study presented in Section 5.

**4. Different Internal and External Populations.** In this section, we consider extensions of our results in Sections 2-3 to situations where the populations for internal and external data are different. Recall that there are response $Y$ and $\boldsymbol{U}=(\boldsymbol{X},\boldsymbol{Z})$ in the internal dataset and only $Y$ and $\boldsymbol{X}$ in the external dataset. Let $R$ be the indicator for internal and external data. What we observe are iid $(Y_i,\boldsymbol{U}_i,R_i)$, $i=1,...,N$, where $N$ is the total sample size of internal and external data, $(Y_i,\boldsymbol{U}_i)$ is observed if $R_i=1$ (internal data), and only $(Y_i,\boldsymbol{X}_i)$ is observed if $R_i=0$ (external data). Our interest is still to estimate the regression function for internal data population, i.e.,

$$\mu_1(\boldsymbol{u})=\mathrm{E}(Y\mid \boldsymbol{U}=\boldsymbol{u},R=1), \tag{14}$$

which reduces to $\mu(\boldsymbol{u})$ in (1) when internal and external populations are the same.

If the internal and external datasets are considered to be one single dataset and $q<p$ (i.e., $\boldsymbol{Z}$ is not degenerated), then $R$ is the indicator of whether $\boldsymbol{Z}$ is observed or missing. However, there are two important differences between the missing data problem and our problem of utilizing external data in analyzing internal data:

(i) The estimand (quantity to be estimated) is different, i.e., in the missing data problem, the estimand is usually $\mu(\boldsymbol{u})$ in (1), rather than $\mu_1(\boldsymbol{u})$ in (14).
(ii) The missing data problem is only defined if $q<p$. When $p=q$, i.e., $\boldsymbol{U}=\boldsymbol{X}$, the internal and external populations can be different and, hence, utilizing external data is still needed.

Thus, unlike Section 3, in this section we consider both scenarios of $q<p$ and $p=q$.

4.1. *Robustness of Results in Sections 2-3.* With iid $(Y_i,\boldsymbol{U}_i,R_i)$'s, the results in Sections 2-3 hold with $n=\sum_{i=1}^{N}R_i$, when internal and external populations are the same, i.e.,

$$(Y,\boldsymbol{X},\boldsymbol{Z})\perp R, \tag{15}$$

where $A\perp B$ denotes that $A$ and $B$ are independent. This is "missing completely at random" if an unmeasured $\boldsymbol{Z}$ is treated as a missing value.

Are the results in Sections 2-3 robust against the violation of (15)? The answer depends on how much the external population deviates from the internal population.

With $R$ indicating the internal and external datasets, constraint (4) should be replaced by

$$\mathrm{E}[\{\boldsymbol{\beta}_g^{\top}\boldsymbol{g}(\boldsymbol{X})-\mu_1(\boldsymbol{U})\}\boldsymbol{g}(\boldsymbol{X})^{\top}|R=1]=0, \tag{16}$$

where

$$\boldsymbol{\beta}_g=[\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^{\top}|R=0\}]^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})Y|R=0\}, \tag{17}$$

because constraint (16) is used in the estimation of $\mu_1(\boldsymbol{u})$ in (14) with internal data (conditioning on $R=1$), whereas $\boldsymbol{\beta}_g$ in (17) is the limit of estimator $\widehat{\boldsymbol{\beta}}_g$ based on external data (conditioning on $R=0$).

When $q<p$, assume that

$$(Y,\boldsymbol{X})\perp R, \tag{18}$$

which is weaker than (15) and allows the internal and external populations to be different. Similar to the derivation (3) that leads to constraint (4), under (18) $\boldsymbol{\beta}_g$ in (17) equals $[\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^{\top}|R=1\}]^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})Y|R=1\}$ and, therefore,

$$\begin{aligned}
\mathrm{E}\{\boldsymbol{\beta}_g^{\top}\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^{\top}|R=1\}&=\mathrm{E}\{Y\boldsymbol{g}(\boldsymbol{X})^{\top}|R=1\}\\
&=\mathrm{E}[\mathrm{E}\{Y\boldsymbol{g}(\boldsymbol{X})^{\top}|\boldsymbol{X},R=1\}|R=1]\\
&=\mathrm{E}[\mathrm{E}\{Y|\boldsymbol{X},R=1\}\boldsymbol{g}(\boldsymbol{X})^{\top}|R=1]\\
&=\mathrm{E}[\mathrm{E}\{\mu_1(\boldsymbol{U})|\boldsymbol{X},R=1\}\boldsymbol{g}(\boldsymbol{X})^{\top}|R=1]\\
&=\mathrm{E}\{\mu_1(\boldsymbol{U})\boldsymbol{g}(\boldsymbol{X})^{\top}|R=1\}
\end{aligned}$$

i.e., (16) holds. Consequently, all results in Sections 2-3 still hold under condition (18) when $q < p$, regardless of whether we have summary-level or individual-level external data. We do not need to consider the case of $p = q$ as (15) and (18) are the same when $p = q$.

Next, we consider two extensions by replacing (18) with even weaker assumptions.

4.2. *Extension under* $\mathrm{E}(Y|\mathbf{X}, R = 1) = \mathrm{E}(Y|\mathbf{X}, R = 0)$. Without (18), constraint (16) may not be satisfied and thus the results in Sections 2-3 may not hold. In this subsection, we consider an extension under the assumption

$$h_0(\boldsymbol{x}) = h_1(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x}, \tag{19}$$

where $h_j(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}, R = j)$, $j = 0, 1$. Note that (19) is weaker than

$$Y \perp R \,|\, \boldsymbol{X}, \tag{20}$$

since (19) involves only first order conditional moments, and (20) is similar to the unconfoundedness assumption in [25] and is weaker than (18) which is weaker than the missing completely at random condition (15).

Under assumption (19),

$$\mathrm{E}[\{\mu_1(\boldsymbol{U}) - h(\boldsymbol{X})\}\boldsymbol{g}(\boldsymbol{X})^\top | R = 1] = 0 \tag{21}$$

holds, where $h = h_1 = h_0$; thus, we can replace constraint (16) by (21). Conditioning on $R = 1$ in (21) means that constraint (21) is used with internal data, which is the same as that in constraint (16). Also, under (19) we do not need to consider the case of $p = q$, as (19) and $p = q$ imply $\mu(\boldsymbol{U}) = \mu_1(\boldsymbol{U})$ so that we can combine the internal and external datasets to estimate $\mu(\boldsymbol{U}) = \mu_1(\boldsymbol{U})$.

In order to use constraint (21) with $q < p$ and external summary statistic such as $\widehat{\boldsymbol{\beta}}_g$, the results in Section 2 hold if the working model $h(\boldsymbol{X}) = E(Y|\boldsymbol{X}) = \boldsymbol{\beta}^\top \boldsymbol{g}(\boldsymbol{X})$ is correct. Otherwise, we must have external individual-level data to correctly estimate $h = h_1 = h_0$ under (19), which is a price to pay for weakening assumption (18) to (19). Instead of requiring a correct model for $h$, we consider the standard nonparametric kernel estimation of $h$, using external individual-level data on $(Y, \boldsymbol{X})$. Under (19), we can add the internal data on $(Y, \boldsymbol{X})$ in this kernel estimation.

Let $\widehat{h}$ be the resulting kernel estimator. We then extend our CK estimator $\widehat{\mu}_{CK}$ in (8) by replacing constraint (6) with

$$\sum_{i=1}^{n} \{\mu_i - \widehat{h}(\boldsymbol{X}_i)\}\boldsymbol{g}(\boldsymbol{X}_i)^\top = 0, \tag{22}$$

which is an empirical analog of constraint (21).

THEOREM 4.1. *Assume (19), (A1)-(A5) in Theorem 2.1, $q < p$, and the following additional conditions for the kernel estimator $\widehat{h}$:*

(A1′) *The function $h$ in (21) has bounded second-order derivatives.*
(A2′) *The kernel used in the estimation of $h$ satisfies condition (A3) with a bandwidth of the order $m^{-1/(4+q)}$ as the external sample size $m \to \infty$.*

*Then, as $n \to \infty$, result (10) in Theorem 2.1 holds with $t = CK$, $\mu(\cdot)$ replaced by $\mu_1(\cdot)$ in (14), and any fixed $\boldsymbol{g}$ in constraint (22).*

Comparing Theorem 4.1 with Theorem 2.1, we find that weakening assumption (18) to (19) does not affect the AMISE of $\widehat{\mu}_{CK}$. Also, the argument in choosing $\boldsymbol{g}$ still holds (Theorem 3.1).

4.3. *Extension under* $\mathrm{E}(Y|\mathbf{X}, R=1) \neq \mathrm{E}(Y|\mathbf{X}, R=0)$. When (19) does not hold, we need to link internal and external data so that $\mathrm{E}(Y|\boldsymbol{X}=\boldsymbol{x}, R=1)$ can be estimated utilizing the external data. The discussion in this subsection is for the case of either $q<p$ or $p=q$, assuming that we have external individual-level data.

Let $f(y|\boldsymbol{x}; R=j)$ be the conditional density of $Y$ given $\boldsymbol{X}=\boldsymbol{x}$ and $R=j$, $j=0,1$. Then

$$h_1(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X}=\boldsymbol{x}, R=1) = \mathrm{E}\left\{ Y \frac{f(Y|\boldsymbol{x}, R=1)}{f(Y|\boldsymbol{x}, R=0)} \middle| \boldsymbol{X}=\boldsymbol{x}, R=0 \right\}. \tag{23}$$

If we can construct an estimator $\widehat{f}(Y|\boldsymbol{x}, R=j)$ of $f(Y|\boldsymbol{x}, R=j)$ for every $\boldsymbol{x}$ and $j=0$ or 1 (preferably a nonparametric estimator), then $h_1(\boldsymbol{x})$ in (23) can be estimated by kernel with $Y\widehat{f}(Y|\boldsymbol{X}, R=1)/\widehat{f}(Y|\boldsymbol{X}, R=0)$ as "response". From the Bayes formula,

$$\frac{f(Y|\boldsymbol{x}, R=1)}{f(Y|\boldsymbol{x}, R=0)} = \frac{\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x}, Y)}{\mathrm{P}(R=0|\boldsymbol{X}=\boldsymbol{x}, Y)} \frac{\mathrm{P}(R=0|\boldsymbol{X}=\boldsymbol{x})}{\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x})}. \tag{24}$$

Thus, we can replace $\widehat{f}(Y|\boldsymbol{X}=\boldsymbol{x}, R=1)/\widehat{f}(Y|\boldsymbol{X}=\boldsymbol{x}, R=0)$ by the product of estimators of $\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x}, Y)/\mathrm{P}(R=0|\boldsymbol{X}=\boldsymbol{x}, Y)$ and $\mathrm{P}(R=0|\boldsymbol{X}=\boldsymbol{x})/\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x})$ for every $\boldsymbol{x}$, constructed using for example the nonparametric estimator in [8]. Both internal and external individual-level data on $(Y, \boldsymbol{X})$ have to be used to estimate $h_1$ under this approach. We denote this nonparametric estimator by $\widehat{h}_1$. The extension of CK estimator $\widehat{\mu}_{CK}$ in (8) to the case of unequal $h_1$ and $h_0$ is to replace constraint (22) with

$$\sum_{i=1}^{n} \{\mu_i - \widehat{h}_1(\boldsymbol{X}_i)\} \boldsymbol{g}(\boldsymbol{X}_i)^\top = 0. \tag{25}$$

Although this approach does not require any extra condition when quantities in (24) are estimated nonparametrically, its performance with not very large dataset may not be satisfactory, because kernel estimation is applied twice in estimating $h_1$ through (23)-(24). If additional information exists for quantities in (24), parametrically or semiparametrically, then this approach can be improved. For example, we consider a semi-parametric model on the following ratio that appears on the right hand side of (24),

$$\frac{\mathrm{P}(R=0 \mid \boldsymbol{X}, Y)}{\mathrm{P}(R=1 \mid \boldsymbol{X}, Y)} = \exp\{a(\boldsymbol{X}) + \gamma Y\}, \tag{26}$$

where $a(\cdot)$ is an unspecified unknown function and $\gamma$ is an unknown parameter. If $\gamma=0$, then (20) holds and thus (19) holds. If $\gamma \neq 0$, then (19) does not hold in general. Applying (26) to (23)-(24), we obtain that

$$\begin{aligned}
h_1(\boldsymbol{x}) &= \mathrm{E}\left\{ Y \frac{\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x}, Y)}{\mathrm{P}(R=0|\boldsymbol{X}=\boldsymbol{x}, Y)} \middle| \boldsymbol{X}=\boldsymbol{x}, R=0 \right\} \frac{\mathrm{P}(R=0|\boldsymbol{X}=\boldsymbol{x})}{\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x})} \\
&= \mathrm{E}\left( Y e^{-a(\boldsymbol{x})-\gamma Y} | \boldsymbol{X}=\boldsymbol{x}, R=0 \right) \frac{\mathrm{E}\{\mathrm{P}(R=0|\boldsymbol{X}=\boldsymbol{x}, Y)|\boldsymbol{X}=\boldsymbol{x}\}}{\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x})} \\
&= e^{-a(\boldsymbol{x})} \mathrm{E}\left( Y e^{-\gamma Y} | \boldsymbol{X}=\boldsymbol{x}, R=0 \right) \frac{\mathrm{E}\{e^{a(\boldsymbol{x})+\gamma Y} \mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x}, Y)|\boldsymbol{X}=\boldsymbol{x}\}}{\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x})} \\
&= \mathrm{E}\left( Y e^{-\gamma Y} | \boldsymbol{X}=\boldsymbol{x}, R=0 \right) \frac{\mathrm{E}\{e^{\gamma Y} \mathrm{E}(R|\boldsymbol{X}=\boldsymbol{x}, Y)|\boldsymbol{X}=\boldsymbol{x}\}}{\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x})} \\
&= \mathrm{E}\left( Y e^{-\gamma Y} | \boldsymbol{X}=\boldsymbol{x}, R=0 \right) \frac{\mathrm{E}(e^{\gamma Y} R|\boldsymbol{X}=\boldsymbol{x})}{\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x})} \\
&= \mathrm{E}\left( Y e^{-\gamma Y} | \boldsymbol{X}=\boldsymbol{x}, R=0 \right) \mathrm{E}\left( e^{\gamma Y} | \boldsymbol{X}=\boldsymbol{x}, R=1 \right),
\end{aligned}$$

where the first and second equalities follow from (26) and the last equality follows from

$$\begin{aligned}\mathrm{E}(e^{\gamma Y}R|\boldsymbol{X}=\boldsymbol{x}) &= \mathrm{E}\left(e^{\gamma Y}R|\boldsymbol{X}=\boldsymbol{x},R=1\right)\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x})\\ &\quad+\mathrm{E}\left(e^{\gamma Y}R|\boldsymbol{X}=\boldsymbol{x},R=0\right)\mathrm{P}(R=0|\boldsymbol{X}=\boldsymbol{x})\\ &= \mathrm{E}\left(e^{\gamma Y}|\boldsymbol{X}=\boldsymbol{x},R=1\right)\mathrm{P}(R=1|\boldsymbol{X}=\boldsymbol{x}).\end{aligned}$$

Note that the function $a(\boldsymbol{x})$ in (26) is not involved in $h_1(\boldsymbol{x})$ and hence we do not need to estimate $a(\boldsymbol{x})$, a nice feature of semiparametric model (26).

If $\gamma$ is known, then we can estimate $h_1$ by estimating $\mathrm{E}\left(Ye^{-\gamma Y}|\boldsymbol{X}=\boldsymbol{x},R=0\right)$ with external data and estimating $\mathrm{E}\left(e^{\gamma Y}|\boldsymbol{X}=\boldsymbol{x},R=1\right)$ with internal data. For the realistic situation of unknown $\gamma$, we propose the following approach. For every real number $t$, define

$$h(\boldsymbol{x},t)=\mathrm{E}(Ye^{-tY}|\boldsymbol{X}=\boldsymbol{x},R=0)\mathrm{E}(e^{tY}|\boldsymbol{X}=\boldsymbol{x},R=1).$$

Its estimator by kernel regression is

$$\widehat{h}(\boldsymbol{x},t)=\frac{\sum_{i=1}^{N}(1-R_i)\check{\kappa}_\ell(\boldsymbol{x}-\boldsymbol{X}_i)Y_ie^{-tY_i}}{\sum_{i=1}^{N}(1-R_i)\check{\kappa}_\ell(\boldsymbol{x}-\boldsymbol{X}_i)}\,\frac{\sum_{i=1}^{N}R_i\check{\kappa}_\ell(\boldsymbol{x}-\boldsymbol{X}_i)e^{tY_i}}{\sum_{i=1}^{N}R_i\check{\kappa}_\ell(\boldsymbol{x}-\boldsymbol{X}_i)} \tag{27}$$

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

$$\widehat{\gamma}=\arg\min_t\frac{1}{N}\sum_{i=1}^{N}R_i\{Y_i-\widehat{h}(\boldsymbol{X}_i,t)\}^2, \tag{28}$$

motivated by the fact that the objective function for minimization in (28) approximates $\mathrm{E}[R\{Y-h(\boldsymbol{X},t)\}^2|R=1]$ and, for any $t$,

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

because $h(\boldsymbol{x},\gamma)=h_1(\boldsymbol{x})=\mathrm{E}(Y|\boldsymbol{X}=\boldsymbol{x},R=1)$ from the previous derivation.

Once $\gamma$ in (26) is estimated by $\widehat{\gamma}$ in (28), we define CK estimator $\widehat{\mu}_{CK}$ in (8) with constraint

$$\sum_{i=1}^{n}\{\mu_i-\widehat{h}(\boldsymbol{X}_i,\widehat{\gamma})\}\boldsymbol{g}(\boldsymbol{X}_i)^\top=0. \tag{29}$$

The following theorem shows the asymptotic normality of $\widehat{\gamma}$ and $\widehat{\mu}_{CK}$.

THEOREM 4.2. *Assume following conditions.*

(C1) *The kernel $\check{\kappa}$ in (27) is Lipschitz continuous; $\int\check{\kappa}(\boldsymbol{u})d\boldsymbol{u}=1$; $\check{\kappa}$ has a bounded support; and there is an integer $d>\max\{q(p+4)/(2p),q\}$ such that for all $j<d$, $\int(\sum_{k=1}^{q}x_k)^j\check{\kappa}(\boldsymbol{x})d\boldsymbol{x}=0$.*

(C2) *The bandwidth $\ell$ in (27) satisfies $N\ell^{2q}/(\log N)^2\to\infty$ and $N\ell^{2d}\to0$ as the total sample size of internal and external datasets $N\to\infty$, where $d$ is given in* (C1).

(C3) *$\gamma$ in (26) 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 $\boldsymbol{x}$, $h(\boldsymbol{x},t)$ is second-order continuously differentiable in $t$ and $h$, $\nabla_th$, $\nabla_t^2h$ are bounded over $t$ and $\boldsymbol{x}$. As $t\to\gamma$, $h(\cdot,t)$, $\nabla_th(\cdot,t)$, and $\nabla_t^2h(\cdot,t)$ convergence uniformly.*

(C4) *$\sup_{t\in\Gamma}\mathrm{E}\|\boldsymbol{W}_t\|^4<\infty$ and $\sup_{t\in\Gamma}\mathrm{E}[\|\boldsymbol{W}_t\|^4|\boldsymbol{X}]f_X(\boldsymbol{X})$ is bounded, where $f_X$ is the density of $\boldsymbol{X}$ and $\boldsymbol{W}_t=(Re^{tY},(1-R)Ye^{-tY},R,(1-R),RYe^{tY},(1-R)Y^2e^{-tY},RY^2e^{tY},(1-R)Y^3e^{-tY})^\top$. Furthermore, there is a function $\tau(Y,R)$ with $\mathrm{E}\{\tau(Y,R)\}<\infty$ such that $\|\boldsymbol{W}_t-\boldsymbol{W}_{t'}\|<\tau(Y,R)|t-t'|$.*

(C5) *The function* $\boldsymbol{\omega}_t(\boldsymbol{x}) = \mathrm{E}(\boldsymbol{W}_t|\boldsymbol{X}=\boldsymbol{x})f_X(\boldsymbol{x})$ *is bounded away from zero, and it is* $d$*th-order continuously differentiable with bounded derivatives on an open set containing the support of* $\boldsymbol{X}$.

(C6) *There is a functional* $G(Y,R,\boldsymbol{\omega})$ *linear in* $\boldsymbol{\omega}$ *such that* $|G(Y,R,\boldsymbol{\omega})| \le \iota(Y,R)\|\boldsymbol{\omega}\|_\infty$ *and, for small enough* $\|\boldsymbol{\omega}-\boldsymbol{\omega}_\gamma\|_\infty$, $|\psi(Y,R,\boldsymbol{\omega}) - \psi(Y,R,\boldsymbol{\omega}_\gamma) - G(Y,R,\boldsymbol{\omega}-\boldsymbol{\omega}_\gamma)| \le \iota(Y,R)\|\boldsymbol{\omega}-\boldsymbol{\omega}_\gamma\|_\infty^2$, *where* $\iota(Y,R)$ *is a function with* $\mathrm{E}\{\iota(Y,R)\} < \infty$, $\psi(Y,R,\boldsymbol{\omega}) = -2R\left(Y - \frac{\omega_1\omega_2}{\omega_3\omega_4}\right)\left(\frac{\omega_2\omega_5-\omega_1\omega_6}{\omega_3\omega_4}\right)$, $\omega_j$ *is the* $j$*th component of* $\boldsymbol{\omega}$, $\|\boldsymbol{\omega}\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{X}}\|\boldsymbol{\omega}(\boldsymbol{x})\|$, $\|\boldsymbol{\omega}-\boldsymbol{\omega}_\gamma\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{X}}\|\boldsymbol{\omega}(\boldsymbol{x})-\boldsymbol{\omega}_\gamma(\boldsymbol{x})\|$, *and* $\mathbb{X}$ *is the range of* $\boldsymbol{X}$. *Also, there exists an almost everywhere continuous function* $\boldsymbol{\nu}(\boldsymbol{X}) \in \mathbb{R}^8$ *with* $\int\|\boldsymbol{\nu}(\boldsymbol{x})\|d\boldsymbol{x} < \infty$ *and* $\mathrm{E}\{\sup_{\|\boldsymbol{\delta}\|\le\epsilon}\|\boldsymbol{\nu}(\boldsymbol{X}+\boldsymbol{\delta})\|^4\} < \infty$ *for some* $\epsilon > 0$ *such that* $\mathrm{E}\{G(Y,R,\boldsymbol{\omega})\} = \int\boldsymbol{\nu}(\boldsymbol{x})^\top\boldsymbol{\omega}(\boldsymbol{x})d\boldsymbol{x}$ *for all* $\|\boldsymbol{\omega}\|_\infty < \infty$.

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

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

*where* $\sigma_\gamma^2 = [2E\{R\nabla_\gamma h(\boldsymbol{X},\gamma)\}^2]^{-1}\mathrm{Var}[\psi(Y,R,\boldsymbol{\omega}_\gamma) + \boldsymbol{\nu}(\boldsymbol{X})^\top\boldsymbol{W}_\gamma - \mathrm{E}\{\boldsymbol{\nu}(\boldsymbol{X})^\top\boldsymbol{W}_\gamma\}]$. *If, in addition, (A1)-(A5) in Theorem 2.1 holds, then result (10) in Theorem 2.1 holds for* $\widehat{\mu}_{CK}$ *using constraint (29), with* $\mu(\cdot)$ *replaced by* $\mu_1(\cdot)$ *in (14).*

Conditions (C1)-(C6) are technique assumptions discussed in [21]; in particular, Lemmas 8.11 and 8.12 in [21]. As discussed by [21], the condition that $\check{\kappa}$ has a bounded support can be relaxed, as it is imposed for a simple proof.

A result similar to Theorem 4.2 can be established for $\widehat{\mu}_{CK}$ using constraint (25), with conditions given in [8] for the estimation of $h_1$.

**5. Simulation Results.** In this section, we present simulation results to examine the performance of our proposed CK estimator and compare it with the methods ignoring external data: the standard kernel estimator (2) and double kernel estimator (9).

We consider univariate covariates $\boldsymbol{X} = X$ and $\boldsymbol{Z} = Z$ ($p=2$ and $q=1$) in two cases:

(i) normal covariates in which $(X,Z)$ is bivariate normal with means 0, variances 1, and correlation 0.5;

(ii) uniform covariates in which $X$ and $Z$ are uniform on $[-1,1]$ with correlation 0.5; specifically, $X = BW_1 + (1-B)W_2$ 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 $(X,Z)$, the response $Y$ is normal with mean $\mu(X,Z)$ and variance 1, where $\mu(X,Z)$ follows 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 $(X,Z)$; M1-M2 are additive models, while M3-M4 are non-additive.

The internal and external data are generated according to the following two settings:

S1. The internal and external datasets are independently sampled from the same population of $(Y,X,Z)$ as previously described with sizes $n = 200$ and $m = 1,000$, respectively.

S2. A total of $N = 1,200$ data are generated from the population of $(Y,X,Z)$ as previously described; given $(Y,X,Z)$, a binary $R$ is generated according to (26) with $a(X) = 1 + 2|X|$ and $\gamma = 0$ or 1/2; internal and external data are indicted by $R=1$ and $R=0$, respectively. Under this setting, the unconditional $\mathrm{P}(R=1)$ is between 10% and 15%.

Note that S1 is for the situation considered in Sections 2-3, whereas $\gamma = 0$ and 1/2 in S2 are for the scenarios in Sections 4.2 and 4.3, respectively.

The following measure is calculated by simulation with $S$ replications to evaluate the performance of each estimator:

$$\text{MISE} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{T}\sum_{t=1}^{T}\{\widehat{\mu}_1^{(s)}(\boldsymbol{U}_{s,t}) - \mu_1(\boldsymbol{U}_{s,t})\}^2, \tag{30}$$

where $\{\boldsymbol{U}_{s,t} : t = 1, ..., T\}$ are test data for each simulation replication $s$, the simulation is repeated independently for $s = 1, ..., S$, $\mu_1$ is defined by (14), and $\widehat{\mu}_1^{(s)}$ is an estimator of $\mu_1$ 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 $s$; hence, the simulated $nb^p \times$MISE approximates AMISE. Under S1 and S2 with $\gamma = 0$, $\mu_1(X, Z) = \mu(X, Z)$. Under S2 with $\gamma \neq 0$, $\mu_1$ is different from $\mu$ and does not have a close form. In the simulation, $\mu_1$ is approximated by Monte Carlo with replication $10^5$.

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

$$\text{IMP} = 1 - \frac{\min\{\text{MISE}(\widehat{\mu}_{CK}) \text{ over all CK methods}\}}{\min\{\text{MISE}(\widehat{\mu}_K), \text{MISE}(\widehat{\mu}_{DK})\}}. \tag{31}$$

In all cases, we use the Gaussian kernel as introduced in Example 2.1. The bandwidths 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 CV (Section 2.3), which produces a decent bandwidth that can be applied to real data. For the CK method, we only tune the bandwidths $b$ and $l$ in (8), not the bandwidths for $\widehat{h}$ and $\widehat{g}^*$ to save computation time.

To see the effects of different functions $\boldsymbol{g}$ in constraint (6) for the CK method, under setting S1 we consider five different choices of $\boldsymbol{g}$, i.e., $\boldsymbol{g}(X) = 1$, $(1, X)^\top$, $(1, \widehat{h}(X))^\top$, $(1, X, \widehat{h}(X))^\top$, and $\widehat{g}^*$, where $\widehat{g}^*$ is given by (13) and $\widehat{h}$ is a kernel estimator of $h(x) = \text{E}(Y|X = x)$.

The simulated MISE defined in (30) based on $S = 200$ replications is presented in Table 1 for setting S1. Note that, for the case where $\boldsymbol{g}(X) = 1$ or $(1, X)^\top$, the results in Table 1 for the CK estimator are applicable to both scenarios of external summary statistics and external individual-level data.

From Table 2, we can see that the proposed CK estimator may be substantially better (in terms of MISE) than the standard kernel or double kernel estimator without using external information. The improvement in efficiency IMP defined in (31) is often over 10% and can be as high as 49%. The bandwidths selected by CV work well although they may not achieve the best efficiency gain. The three choices of $g$ functions in constraint (6), i.e., $\boldsymbol{g}(X) = (1, X)^\top$, $(1, \widehat{h}(X))^\top$, and $(1, X, \widehat{h}(X))^\top$, work well and have comparable performances, but do not show definite superiority of one over the other. Thus, $\boldsymbol{g}(X) = (1, X)^\top$ is recommended for its simplicity. However, the choice $\widehat{g}^*$ in (13) does not perform well, although Theorem 3.1 shows that $\widehat{g}^*$ has the optimal performance asymptotically. This indicates the slow convergence of $\widehat{g}^*$ due to the estimation of second-order derivatives. Note that $\widehat{g}^*$ is constructed using internal data with size $n = 200$ under setting S1.

Under setting S2, our main interest is to evaluate the performance of CK estimator with a fixed choice $\boldsymbol{g}(X) = (1, X)^\top$ under different scenarios in which the internal and external populations are different as described in Section 4. Other than the standard kernel and double kernel without using external data, we study four CK estimators discussed in Sections 2-4: $\widehat{\mu}_{CK}$ with constraint (6), which is incorrect since (18) does not hold; $\widehat{\mu}_{CK}$ with constraint (22), which is correct when $\gamma = 0$ but incorrect when $\gamma \neq 0$; $\widehat{\mu}_{CK}$ with constraint (25); and $\widehat{\mu}_{CK}$ with constraint (29). The last two estimators are asymptotically valid under S1-S2.

The simulated MISE based on $S = 200$ replications is shown in Table 2 for the case of $\gamma = 0$ and in Table 3 for the case of $\gamma = 1/2$.

Consider first the results in Table 2. Since $\gamma = 0$, the CK estimators using three different constraints, (22), (25), and (29), are all correct and more efficient than estimators without using external information. The estimator using constraint (22) is the best among the three, as it uses the information that $\gamma = 0$. The CK estimator using constraint (6) is biased as (18) does not hold and its performance depends on the magnitude of bias; in some cases it can be much worse than the others and in other cases it is as good as the CK estimator using constraint (22).

Next, the results in Table 3 for setting S2 with $\gamma = 1/2$ indicate that the CK estimator using correct constraint (25) or (29) is better than the CK estimator using incorrect constraint (6) or (22), or the estimator without using external information. Since the estimator with constraint (29) uses more information, it is in general better than the one with constraint (25). With an incorrect constraint, the CK estimator can be much worse than that without using external information.

Finally, in S2 with both $\gamma = 0$ and $\gamma = 1/2$, the estimator $\widehat{\gamma}$ defined by (28) performs well.

Overall, the simulation results support our asymptotic theory and show that the CK estimator is better than the kernel estimators without using external information.

**6. Empirical Application.** In this section, we apply the proposed algorithm to real data. University of Queensland Vital Signs Dataset (UQVSD) [17] recorded vital signals of intensive care patients with several sensors, including ECG, Pulse oximetry, Capnography, Gas analysis, and more. It has rich information; however, the sample size is only 32. Furthermore, we utilize an external data set, Medical Information Mart for Intensive Care III (MIMIC-III) [12], which is a freely available digital health record database with a vast sample size (54060) integrated rich clinical information of patients needing critical care. Since both data sets study intensive care units, we consider these two data sets to come from the same distribution.

This empirical application aims to predict systolic ($Y$) blood pressure, which is a critical biomarker to measure patients' health conditions. Since UQVSD and MIMIC-III share three covariates, HR, RR, and SpO2 (See table 5) and the primary data set UQVSD only contains 32 samples, we consider a linear combination of these three variables as a single index covariate $X$. The coefficient of this linear combination is the first eigenvector of a well-known sufficient dimension reduction algorithm SAVE [28], from which the first eigenvector provides more than 94 % variability. As a result, the external data set (MIMIC-III) provides information of linear regression of systolic ($Y$) on the single index covariate ($X$).

In the primary data UQVSD, we consider additional covariates collected via a sensor-Gas analysis, which contains the information of in/et O2 and in/et CO2. Figure 1-4 show the fitted result of systolic ($Y$) on the single index covariate ($X$) and one of these four additional covariates ($Z$) via our proposed algorithm CK (8), and other two methods DK (9) and KR (2). One can see that CK provides a clean pattern indicating patients at risk of hypertension while DK, KR, provides a vague prediction. All the bandwidths are selected via five-fold cross-validation.

TABLE 1

| | Variables | MIMIC-III | UQVSD | Description |
|---|---|---|---|---|
| $Y$ | Sys | $\surd$ | $\surd$ | Systolic blood pressure |
| $X$ | HR | $\surd$ | $\surd$ | Heart Rate |
| | RR | $\surd$ | $\surd$ | Respiratory rate |
| | SpO2 | $\surd$ | $\surd$ | Blood oxygen saturation |
| $Z$ | inO2 | | $\surd$ | Inspired O2 concentration |
| | etO2 | | $\surd$ | End tidal O2 concentration |
| | inCO2 | | $\surd$ | Inspired CO2 concentration |
| | etCO2 | | $\surd$ | End tidal CO2 concentration |

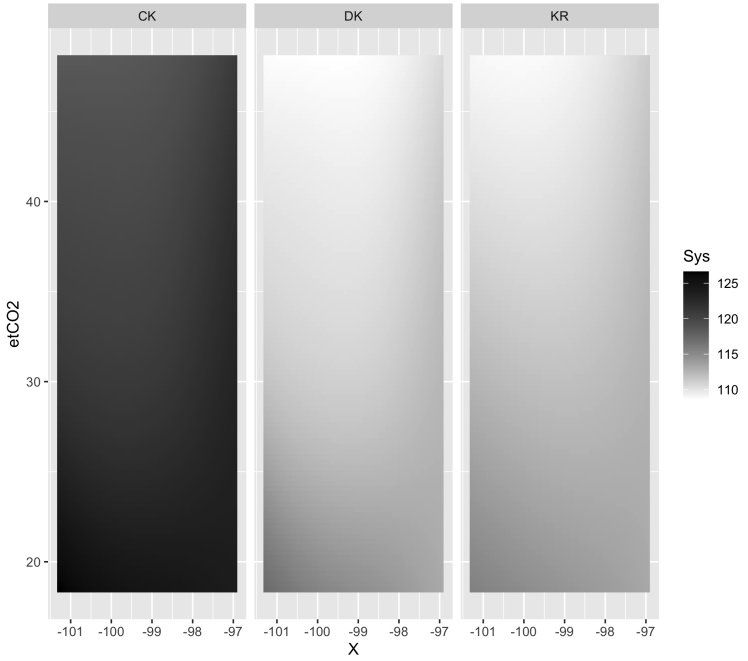


FIG 1.

**7. Discussion.** Curse of dimensionality is a well-known problem for nonparametric methods. Thus, the proposed CK method in Sections 2-4 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 (see [15]), i.e., $\mu(\boldsymbol{U})$ in (1) is assumed to be

$$\mu(\boldsymbol{U}) = \mu(\boldsymbol{\eta}^\top \boldsymbol{U}), \tag{32}$$

where $\boldsymbol{\eta}$ is an unknown $p$-dimensional vector. The well-known SIR technique (see [15]) can be applied to obtain a consistent and asymptotically normal estimator $\widehat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ in (32). 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 (32) [5, 16, 19, 32].

We turn to the dimension of $\boldsymbol{X}$ in the external dataset. In the situation where (15) or (18) holds, constraint (6) can be used and the least square type estimator $\widehat{\boldsymbol{\beta}}_g$ is not seriously affected by the dimension of $\boldsymbol{X}$ unless the dimension of $\boldsymbol{X}$ is ultra-high in the sense that the dimension of $\boldsymbol{X}$ over the size of external dataset does not tend to 0. If the dimension of $\boldsymbol{X}$ is ultra-high or (18) does not hold so that a kernel estimator of $h_1$ is used as described in

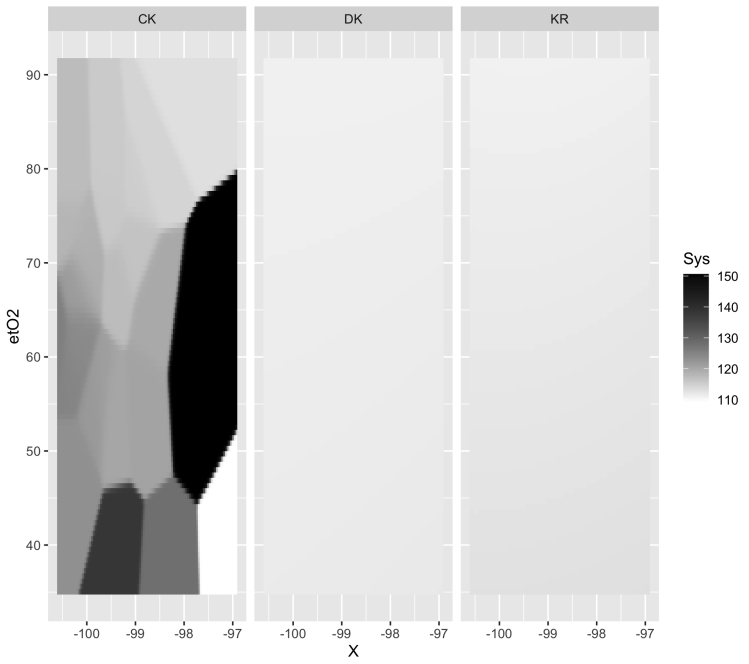


FIG 2.

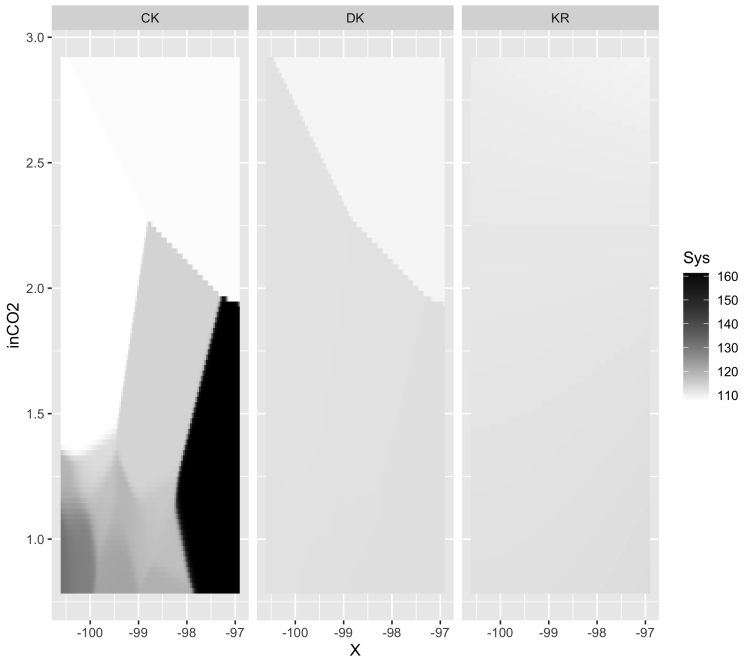


FIG 3.

Sections 4.2-4.3, then we may consider the following approach. Instead of using constraint

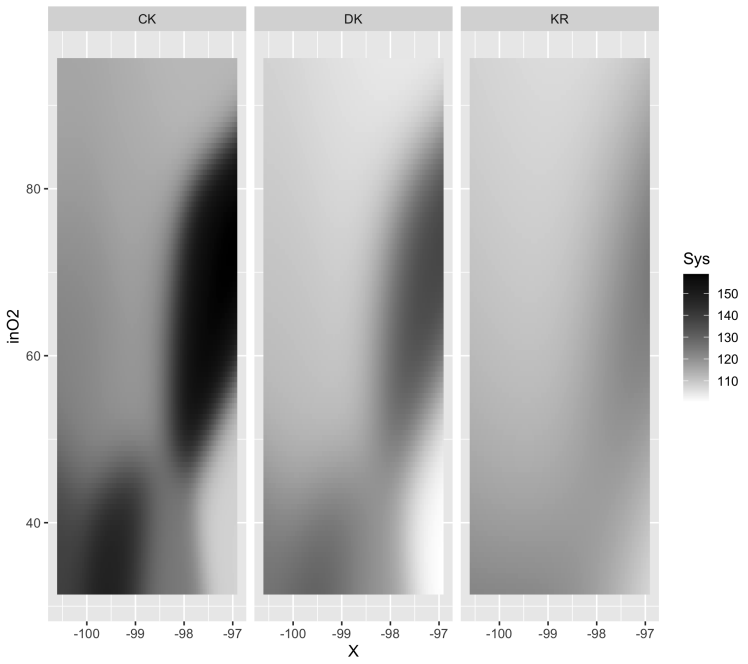


FIG 4.

(6), (22), or (25), we use component-wise constraints

$$\sum_{i=1}^{n}\{\mu_i - \widehat{h}^{(k)}(X_i^{(k)})\}g_k(X_i^{(k)})^\top = 0, \qquad k = 1, ..., q, \tag{33}$$

where $X_i^{(k)}$ is the $k$th component of $\boldsymbol{X}_i$, $g_k(X^{(k)})$ is a function of $X^{(k)}$, and $\widehat{h}^{(k)}(X_i^{(k)})$ equals $\widehat{\boldsymbol{\beta}}_{g_k}^\top g_k(X^{(k)})$ when (6) is used, equals a kernel estimator of $\mathrm{E}(Y|X^{(k)})$ when (22) is used, or equals a kernel estimator through (23)-(24) when (25) is used. More constraints are involved in (33), but estimation only involves one dimensional $X^{(k)}$, $k = 1, ..., q$.

The kernel $\kappa$ we adopted in (2), (5), and (8) is called the second order kernel so that the convergence rate of $\widehat{\mu}_{CK}(\boldsymbol{u}) - \mu(\boldsymbol{u})$ is $n^{-2/(4+p)}$. A $d$th order kernel with $d \geq 2$ as defined by [1] may be used to achieve convergence rate $n^{-d/(2d+p)}$. Alternatively, we may also apply other nonparametric smoothing techniques such as the local polynomial [10] to achieve convergence rate $n^{-d/(2d+p)}$, $d \geq 2$.

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 (33) by estimating $\widehat{h}^{(k)}$ 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 4.

Finally, in Section 2, we only consider the summary-level external information from a linear regression. However, we can generalize it to any generalized estimating equation (GEE), such as logistic regression. Assume that the summary-level estimate $\widehat{\boldsymbol{\beta}}$ is solved from the following GEE.

$$\sum_{i=n+1}^{N} \boldsymbol{H}(\widehat{\boldsymbol{\beta}}, Y_i, \boldsymbol{X}_i) = 0,$$

where $\boldsymbol{H}$ is a known $k$ dimensional functions. For example, $\boldsymbol{H}(\boldsymbol{\beta}, Y, \boldsymbol{X}) = (Y - \boldsymbol{\beta}^\top \boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top$ is a special case for a linear regression estimate. From an analogy of (3), the following equations are valid constraints for GEE types of summary-level information.

$$\sum_{i=1}^{n} \boldsymbol{H}(\widehat{\boldsymbol{\beta}}, \mu_i, \boldsymbol{X}_i) = 0.$$

**8. Proofs of Theorems.** This section contains the proofs of Theorems, with some technical details and all lemmas used in the proofs given in the Appendix as Supplementary Material.

8.1. *Proof of Theorem 2.1.* Consider the following decomposition:

$$\sqrt{nb^p}\{\widehat{\mu}_{CK}(\boldsymbol{u}) - \mu(\boldsymbol{u})\} = T_1 + \cdots + T_6, \tag{34}$$

where

$$\begin{aligned}
T_1 &= n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top (\boldsymbol{I}_n - \boldsymbol{P}) \boldsymbol{B}_l^{-1} \boldsymbol{\Delta}_l \boldsymbol{\epsilon} / \widehat{f}_b(\boldsymbol{u}), \\
T_2 &= n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \{\boldsymbol{\mu} - \mu(\boldsymbol{u}) \mathbf{1}_n\} / \widehat{f}_b(\boldsymbol{u}), \\
T_3 &= n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top (\boldsymbol{B}_l^{-1} \boldsymbol{\Delta}_l \boldsymbol{\mu} - \boldsymbol{\mu}) / \widehat{f}_b(\boldsymbol{u}), \\
T_4 &= -n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \boldsymbol{P} (\boldsymbol{B}_l^{-1} \boldsymbol{\Delta}_l \boldsymbol{\mu} - \boldsymbol{\mu}) / \widehat{f}_b(\boldsymbol{u}), \\
T_5 &= n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \boldsymbol{P} (\widehat{\boldsymbol{h}} - \boldsymbol{h}) / \widehat{f}_b(\boldsymbol{u}), \\
T_6 &= n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \boldsymbol{P} (\boldsymbol{h} - \boldsymbol{\mu}) / \widehat{f}_b(\boldsymbol{u}),
\end{aligned}$$

$\widehat{f}_b(\boldsymbol{u}) = \sum_{i=1}^n \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)/n$, $\boldsymbol{\delta}_b(\boldsymbol{u}) = (\kappa_b(\boldsymbol{u} - \boldsymbol{U}_1), \ldots, \kappa_b(\boldsymbol{u} - \boldsymbol{U}_n))^\top$, $\boldsymbol{I}_n$ is the identity matrix of order $n$, $\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top$ with $\boldsymbol{G}$ defined in (7), $\boldsymbol{B}_l$ is the $n \times n$ diagonal matrix whose $i$th 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 $\epsilon_i = Y_i - \mu(\boldsymbol{U}_i)$, $\boldsymbol{\mu} = (\mu(\boldsymbol{U}_1), ..., \mu(\boldsymbol{U}_n))^\top$, $\mathbf{1}_n$ is the $n$-vector with all components being 1, $\boldsymbol{h} = \boldsymbol{G}\boldsymbol{\beta}_g$ with $\boldsymbol{\beta}_g$ defined in (3), and $\widehat{\boldsymbol{h}} = \boldsymbol{G}\widehat{\boldsymbol{\beta}}_g$ with $\widehat{\boldsymbol{\beta}}_g$ defined in (6).

We first show that $T_1$ in (34) is asymptotically normal with mean 0 and variance $V_{CK}(\boldsymbol{u})$ defined in (11). Define

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

Consider a further decomposition of $T_1$:

$$T_1 = \sqrt{n}\, V + T_{11} + T_{12} + T_{13}, \tag{35}$$

where

$$\begin{aligned}
V &= \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} S(\boldsymbol{U}_i, \epsilon_i, \boldsymbol{U}_j, \epsilon_j), \\
T_{11} &= \frac{b^{p/2}}{n^{3/2}} \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\kappa_l(0)\epsilon_i}{f_U(\boldsymbol{u}) f_U(\boldsymbol{U}_i)}, \\
T_{12} &= \frac{b^{p/2}}{n^{3/2}} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{f_U(\boldsymbol{u}) f_U(\boldsymbol{U}_i)} \left\{ \frac{f_U(\boldsymbol{u}) f_U(\boldsymbol{U}_i)}{\widehat{f}_b(\boldsymbol{u}) \widehat{f}_l(\boldsymbol{U}_i)} - 1 \right\} \epsilon_j, \\
T_{13} &= -n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \boldsymbol{P} \boldsymbol{B}_l^{-1} \boldsymbol{\Delta}_l \boldsymbol{\epsilon} / \widehat{f}_b(\boldsymbol{u}).
\end{aligned}$$

The decomposition (35) follows from

$$\sqrt{n}\,V + T_{11} = \frac{b^{p/2}}{n^{3/2}} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{f_U(\boldsymbol{u}) f_U(\boldsymbol{U}_i)} \epsilon_j$$

and

$$\begin{aligned}\sqrt{n}\,V + T_{11} + T_{12} &= \frac{b^{p/2}}{n^{3/2}} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\widehat{f}_l(\boldsymbol{u}) \widehat{f}_l(\boldsymbol{U}_i)} \epsilon_j \\ &= n^{-1/2} b^{p/2} \boldsymbol{\delta}_b(\boldsymbol{u})^\top \boldsymbol{B}_l^{-1} \boldsymbol{\Delta}_l \boldsymbol{\epsilon} / \widehat{f}_b(\boldsymbol{u}).\end{aligned}$$

Note that $V$ is a V-statistic with

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

which has variance

$$\begin{aligned}\mathrm{Var}\{S_1(\boldsymbol{U}_1, \epsilon_1)\} &= \frac{b^{p/2}}{4 f_U^2(\boldsymbol{u})} \int f_U(\boldsymbol{u}_1)\sigma^2(\boldsymbol{u}_1) \left\{ \int \kappa_l(\boldsymbol{u}_2 - \boldsymbol{u}_1)\kappa_b(\boldsymbol{u} - \boldsymbol{u}_2) d\boldsymbol{u}_2 \right\}^2 d\boldsymbol{u}_1 \\ &= \frac{b^{p/2}}{4 f_U^2(\boldsymbol{u})} \int f_U(\boldsymbol{u}_1)\sigma^2(\boldsymbol{u}_1) \left\{ \int \kappa_l(\boldsymbol{v})\kappa_b(\boldsymbol{u} - \boldsymbol{u}_1 - l\boldsymbol{\nu}) d\boldsymbol{\nu} \right\}^2 d\boldsymbol{u}_1 \\ &= \frac{1}{4 f_U^2(\boldsymbol{u})} \int f_U(\boldsymbol{u} - b\boldsymbol{w})\sigma^2(\boldsymbol{u} - b\boldsymbol{w}) \left\{ \int \kappa(\boldsymbol{v})\kappa\left(\boldsymbol{w} - \boldsymbol{\nu}\frac{l}{b}\right) d\boldsymbol{\nu} \right\}^2 d\boldsymbol{w},\end{aligned}$$

where $\sigma^2(\cdot)$ is given in condition (A2), 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 $f_U(\cdot)$ and $\sigma^2(\cdot)$, $\mathrm{Var}\{S_1(\boldsymbol{u}_1, \epsilon_1)\}$ converges to $V_{CK}(\boldsymbol{u})$. Therefore, by the theory for asymptotic normality of V-statistics (e.g., Theorem 3.16 in [27]),

$$\sqrt{n}\,V \xrightarrow{d} N\big(0, V_{CK}(\boldsymbol{u})\big).$$

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

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

where $O_p(a_n)$ denotes a term bounded by $a_n$ in probability and $o_p(1)$ denotes a term $\xrightarrow{p} 0$. This proves that $T_{11} = o_p(1)$. Note that $\mathrm{E}(T_{12} \mid \boldsymbol{U}_1, ..., \boldsymbol{U}_n) = 0$ and, from

$$\begin{aligned}&\left| \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{f_U(\boldsymbol{u}) f_U(\boldsymbol{U}_i)} - \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\widehat{f}_b(\boldsymbol{u}) \widehat{f}_l(\boldsymbol{U}_i)} \right| \\ \le\ &\max\left\{ \frac{1}{f_U(\boldsymbol{u})}, \frac{1}{\widehat{f}_b(\boldsymbol{u})} \right\} \max_{i=1,...,n} \left| \frac{f_U(\boldsymbol{U}_i)}{\widehat{f}_l(\boldsymbol{U}_i)} - 1 \right| \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{f_U(\boldsymbol{u}) f_U(\boldsymbol{U}_i)},\end{aligned}$$

$\mathrm{Var}(T_{12} \mid \boldsymbol{U}_1, ..., \boldsymbol{U}_n)$ is bounded by

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

Therefore, under the assumed condition that $f_U$ is bounded away from zero, (S5) in Lemma 3 implies $T_{12} = o_p(1)$. Define

$$W_j(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\boldsymbol{g}(\boldsymbol{X}_i)^\top}{\widehat{f}_b(\boldsymbol{u})}(\boldsymbol{G}^\top\boldsymbol{G})^{-1}\sum_{i=1}^n \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)\boldsymbol{g}(\boldsymbol{X}_i)}{\widehat{f}_l(\boldsymbol{U}_i)}.$$

Then

$$T_{13} = \frac{b^{p/2}}{n^{1/2}}\sum_{j=1}^n W_j(\boldsymbol{u})\epsilon_j.$$

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

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

Under the assumed condition that $f_U$ is bounded away from zero, (S5) and (S7) in Lemma 3 imply

$$\max_{j=1,\ldots,n}\left|W_j(\boldsymbol{u}) - \boldsymbol{g}(\boldsymbol{u})^\top\boldsymbol{\Sigma}_g^{-1}\boldsymbol{g}(\boldsymbol{X}_j)\right| = o_p(1).$$

As a result, $\mathrm{Var}(T_{13} \mid \boldsymbol{U}_1, ..., \boldsymbol{U}_n) = O_p(b^p) = o_p(1)$. This concludes that $T_{13} = o_p(1)$. Consequently, by (35), $T_1$ has the same asymptotic distribution as $\sqrt{n}V$, the claimed result.

It remains to show that $T_2 + \cdots + T_6$ in (34) converges to $B_{CK}(\boldsymbol{u})$ in probability. From Lemma 4 and (A4),

$$T_2 = \sqrt{nb^p}b^2A(\boldsymbol{u})\{1 + o_p(1)\} = \sqrt{c}A(\boldsymbol{u})\{1 + o_p(1)\}.$$

Note that

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

where the second equality follows from (S5) in Lemma 3, Lemma 4 and (A4), and the last equality follows from Lemma 2 and continuity of $A(\cdot)$.

The term $T_4$ divided by $-n^{-1/2}b^{p/2}$ is equal to

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\boldsymbol{g}(\boldsymbol{X}_i)^\top}{\widehat{f}_b(\boldsymbol{u})}(\boldsymbol{G}^\top\boldsymbol{G})^{-1}\sum_{j=1}^n \frac{\boldsymbol{g}(\boldsymbol{X}_j)}{n\widehat{f}(\boldsymbol{U}_j)}\sum_{i=1}^n \kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)\{\mu(\boldsymbol{U}_i) - \mu(\boldsymbol{U}_j)\} \\
&= \left\{\boldsymbol{g}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_g^{-1}\frac{1}{n}\sum_{j=1}^n \frac{\boldsymbol{g}(\boldsymbol{X}_j)}{n\widehat{f}(\boldsymbol{U}_j)}\sum_{i=1}^n \kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)\{\mu(\boldsymbol{U}_i) - \mu(\boldsymbol{U}_j)\}\right\}\{1 + o_p(1)\}
\end{aligned}$$

$$= \left\{ \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} \frac{l^{2/p}}{n} \sum_{j=1}^{n} \boldsymbol{g}(\boldsymbol{X}_j) A(\boldsymbol{U}_j) \right\} \{1 + o_p(1)\}$$

$$= l^{2/p} \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} \mathrm{E}\{\boldsymbol{g}(\boldsymbol{X}) A(\boldsymbol{U})\} \{1 + o_p(1)\},$$

where the first equality follows from (S5)-(S6) in the Appendix and the law of large numbers, the second equality follows from Lemma 4, and the last equality follows from the law of large numbers. Hence,

$$T_4 = -\sqrt{c} r^2 g(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} \mathrm{E}\{\boldsymbol{g}(\boldsymbol{X}) A(\boldsymbol{U})\} \{1 + o_p(1)\}.$$

Similarly, $T_5$ divided by $n^{-1/2} b^{p/2}$ is equal to

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \boldsymbol{g}(\boldsymbol{X}_i)^\top}{\widehat{f}_b(\boldsymbol{u})} (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i) \{\widehat{h}(\boldsymbol{X}_i) - h(\boldsymbol{X}_i)\}$$

$$= \left[ \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i) \{\widehat{h}(\boldsymbol{X}_i) - h(\boldsymbol{X}_i)\} \right] \{1 + o_p(1)\}$$

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

where $\widehat{h}(\boldsymbol{X}_i)$ and $h(\boldsymbol{X}_i)$ are the $i$th components of $\widehat{\boldsymbol{h}}$ and $\boldsymbol{h}$, respectively, the first equality follows from (S5)-(S6), and the inequality holds for a non-random constant $M$. Since $\widehat{h}(\boldsymbol{X}_j) - h(\boldsymbol{X}_j) = \boldsymbol{g}(\boldsymbol{X}_j)^\top (\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g)$,

$$\max_{j=1,\ldots,n} |\widehat{h}(\boldsymbol{X}_j) - h(\boldsymbol{X}_j)| \leq \max_{j=1,\ldots,n} \|\boldsymbol{g}(\boldsymbol{X}_j)\| \|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\| = O_p(1/\sqrt{m}),$$

following from $\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\| = O_p(1/\sqrt{m})$ and the boundedness of $\boldsymbol{g}(\boldsymbol{X})$. By assumption (A5), $O_p(1/\sqrt{m}) = o_p(n^{-2/(p+4)})$, where $o_p(a_n) = a_n o_p(1)$. Consequently, by (A4),

$$T_5 = o_p(1).$$

From (S5)-(S6) in Lemma 3 and the Central Limit Theorem,

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

$$= O_p(b^{p/2}) = o_p(1).$$

Combining these results, we obtain that $T_2 + \cdots + T_6 = B_{CK}(\boldsymbol{u}) + o_p(1)$. This establishes the result for $\widehat{\mu}_{CK}$.

For $\widehat{\mu}_{DK}$, the result follows from the fact that $\widehat{\mu}_{DK}$ equals to $\widehat{\mu}_{CK}$ with $\boldsymbol{g} = 0$.

As a technical note, we show that a large enough $s$ ensures that Lemma 2 and (A4) hold. The condition $s > 2 + p/2$ as stated in (A1) is sufficient for

$$n^{1-2/s-\theta} b^p \to \infty, \qquad \text{where } \theta > 0 \text{ and } b = O_p(n^{-1/(p+4)}).$$

8.2. *Proof of Theorem 2.2.* Let $c \geq 0$ and $r > 0$ be the constants in (A4) and $\varphi_c(r)$ denote the AMISE of $\widehat{\mu}_{CK}$ for particular $c$ and $r$. By (11)-(12),

$$\varphi_c(r) = c\tau_1 + c\tau_2 r^2 (2 + r^2) + \tau_3 \int \left\{ \int \kappa(\boldsymbol{w} - r\boldsymbol{v}) \kappa(\boldsymbol{v}) d\boldsymbol{v} \right\}^2 d\boldsymbol{w},$$

where $\tau_1 = \mathrm{E}\{A(\boldsymbol{U})\}^2$, $\tau_2 = \mathrm{E}\{A(\boldsymbol{U}) - \boldsymbol{g}(\boldsymbol{X})^\top \boldsymbol{\Sigma}_g^{-1} \mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}\}^2$, and $\tau_3 = \mathrm{E}\{\sigma^2(\boldsymbol{U})/f_U(\boldsymbol{U})\}$. The first and second order derivatives of this function with respect to $r$ are

$$\varphi_c'(r) = 4c\tau_2(r + r^3) + \tau_3 \int \left\{ \int \kappa(\boldsymbol{w} - \boldsymbol{v}r)\kappa(\boldsymbol{v})d\boldsymbol{v} \right\} \left\{ \int -\nabla\kappa(\boldsymbol{w} - \boldsymbol{v}r)^\top \boldsymbol{v}\kappa(\boldsymbol{v})d\boldsymbol{v} \right\} d\boldsymbol{w},$$

and

$$\begin{aligned}\varphi_c''(r) = {} & 4c\tau_2(1 + 3r^2) + \tau_3 \int \left\{ \int \nabla\kappa(\boldsymbol{w} - \boldsymbol{v}r)^\top \boldsymbol{v}\kappa(\boldsymbol{v})d\boldsymbol{v} \right\}^2 d\boldsymbol{w} \\ & + \tau_3 \int \left\{ \int \kappa(\boldsymbol{w} - \boldsymbol{v}r)\kappa(\boldsymbol{v})d\boldsymbol{v} \right\} \left\{ \int \boldsymbol{v}^\top \nabla^2\kappa(\boldsymbol{w} - \boldsymbol{v}r)\boldsymbol{v}\kappa(\boldsymbol{v})d\boldsymbol{v} \right\} d\boldsymbol{w}.\end{aligned}$$

Since the mean of $\kappa(\cdot)$ is 0,

$$\varphi_c'(0) = 0 \quad \text{and} \quad \varphi_c''(0) = 4c\tau_2 + \tau_3 \int \boldsymbol{v}^\top \left\{ \int \nabla^2\kappa(\boldsymbol{w})\kappa(\boldsymbol{w})d\boldsymbol{w} \right\} \boldsymbol{v}\kappa(\boldsymbol{v})d\boldsymbol{v}.$$

Under the assumed condition that $\int \nabla^2\kappa(\boldsymbol{w})\kappa(\boldsymbol{w})d\boldsymbol{w}$ is negative definite, there exists a constant $c^* > 0$ such that for all $c < c^*$, $\varphi_c''(0) < 0$. This means that for $c < c^*$, $r = 0$ is a local maximum of $\varphi_c(r)$. Consequently, for all $c$ and $r$ in a neighborhood of 0,

$$\mathrm{AMISE}(\widehat{\mu}_{CK})(r) < \mathrm{AMISE}(\widehat{\mu}_K) = \varphi_c(0).$$

8.3. *Proof of Theorem 3.1.* Let $T_{13}$, $T_4$, $T_5$ and $T_6$ be the quantities defined in the proof of Theorem 2.1 but with $\boldsymbol{P}$ replaced by $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{G}}(\widehat{\boldsymbol{G}}^\top\widehat{\boldsymbol{G}})^{-1}\widehat{\boldsymbol{G}}^\top$, where $\widehat{\boldsymbol{G}} = (\widehat{g}^*(\boldsymbol{X}_1), ..., \widehat{g}^*(\boldsymbol{X}_n))^\top$. Following the proof of Theorem 2.1 and using the same notation, we still have decompositions (34) and (35), where $V$, $T_{11}$, $T_{12}$, $T_2$, and $T_3$ are unchanged. Thus, we just need to show that the results for $T_{13}$, $T_4$, $T_5$ and $T_6$ in the proof of Theorem 2.1 still hold for the modified $T_{13}$, $T_4$, $T_5$ and $T_6$. From Lemma 5, we can show that

$$\widehat{\boldsymbol{G}}^\top\widehat{\boldsymbol{G}}/n \xrightarrow{p} \boldsymbol{\Sigma}_g, \tag{36}$$

$$\frac{1}{n}\sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\widehat{g}^*(\boldsymbol{X}_i)}{\widehat{f}_b(\boldsymbol{u})} \xrightarrow{p} g^*(\boldsymbol{x}), \tag{37}$$

$$\frac{1}{n}\sum_{j=1}^n \widehat{g}^*(\boldsymbol{X}_j)A(\boldsymbol{U}_j) \xrightarrow{p} \mathrm{E}\{g^*(\boldsymbol{X})A(\boldsymbol{U})\}, \tag{38}$$

$$\max_{j=1,...,n} |\widehat{g}^*(\boldsymbol{X}_j)| = O_p(1), \tag{39}$$

$$m\left|\Omega^{-1} - \widehat{\Omega}^{-1}\right| = o_p(n^{-2/(p+4)}), \tag{40}$$

where

$$\Omega = \sum_{i=n+1}^{n+m} \{g^*(\boldsymbol{X}_i)\}^2 \qquad \text{and} \qquad \widehat{\Omega} = \sum_{i=n+1}^{n+m} \{\widehat{g}^*(\boldsymbol{X}_i)\}^2.$$

From (36)-(38), $T_4$ converge in probability to the same limits as those in the proof of Theorem 2.1. From (36), (37), and (39), $T_5 = o_p(1)$ follows from $|\widehat{\beta}_{g^*} - \widehat{\beta}_{\widehat{g}^*}| = o_p(n^{-2/(p+4)})$, where $\widehat{\beta}_{g^*} = \Omega^{-1}\sum_{i=n+1}^{n+m} g^*(\boldsymbol{X}_i)Y_i$ and $\widehat{\beta}_{\widehat{g}^*} = \widehat{\Omega}^{-1}\sum_{i=n+1}^{n+m} \widehat{g}^*(\boldsymbol{X}_i)Y_i$. Note that

$$|\widehat{\beta}_{g^*} - \widehat{\beta}_{\widehat{g}^*}| \le m\,\Omega^{-1}\left(\frac{1}{m}\sum_{i=n+1}^{n+m} |Y_i|\right) \sup_{\boldsymbol{x}\in\mathbb{X}} |\widehat{g}^*(\boldsymbol{x}) - g^*(\boldsymbol{x})|$$

$$+\ m\left|\Omega^{-1}-\widehat{\Omega}^{-1}\right|\left\{\frac{1}{m}\sum_{i=n+1}^{n+m}g^*(\boldsymbol{X}_i)Y_i\right\}$$
$$=o_p(n^{-2/(p+4)}),$$

where $\mathbb{X}$ is the range of $\boldsymbol{X}$ and the equality follows from (40), law of large numbers, and Lemma 5. This proves that $T_5=o_p(1)$. Next,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\{\widehat{g}^*(\boldsymbol{X}_i)-g^*(\boldsymbol{X}_i)\}\{h(\boldsymbol{X}_i)-\mu(\boldsymbol{U}_i)\}\right|$$
$$\le\left\{\frac{1}{n}\sum_{i=1}^{n}|h(\boldsymbol{X}_i)-\mu(\boldsymbol{U}_i)|\right\}\sup_{\boldsymbol{x}\in\mathbb{X}}|\widehat{g}^*(\boldsymbol{x})-g^*(\boldsymbol{x})|$$
$$=o_p(n^{-2/(p+4)}),$$

where the equality follows from Lemma 5. This result together with (36) and (37) shows that $T_6=o_p(1)$. Similarly,

$$\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\widehat{g}^*(\boldsymbol{X}_i)-g^*(\boldsymbol{X}_i)\}\kappa_l(\boldsymbol{U}_i-\boldsymbol{U}_j)\epsilon_j\right|$$
$$\le\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_l(\boldsymbol{U}_i-\boldsymbol{U}_j)|\epsilon_j|\right\}\sup_{\boldsymbol{x}\in\mathbb{X}}|\widehat{g}^*(\boldsymbol{x})-g^*(\boldsymbol{x})|$$
$$=o_p(n^{-2/(p+4)}).$$

This result together with (36) and (37) show that $T_{13}=o_p(1)$. The proof is completed.

8.4. *Proof of Theorem 4.1.* The proof follows the same argument in proving Theorem 2.1 with $h(\boldsymbol{X})=\mathrm{E}(Y|\boldsymbol{X})$ and $\widehat{h}$ being its kernel estimator defined in (22). We only need to show that $\max_{j=1,\ldots,n}|\widehat{h}(\boldsymbol{X}_j)-h(\boldsymbol{X}_j)|=o_p(n^{-2/(p+4)})$. From (A1′)-(A2′) and (S3) in Lemma 2, $\sup_{\boldsymbol{x}\in\mathbb{X}}|\widehat{h}(\boldsymbol{x})-h(\boldsymbol{x})|=O_p\big(\log(n+m)(n+m)^{-2/(q+4)}\big)$. Then the result follows from (A5) and $q<p$.

8.5. *Proof of Theorem 4.2.* The main technical detail is in showing that $\sqrt{N}(\widehat{\gamma}-\gamma)$ is asymptotically normal with mean 0 and a finite variance, as $N\to\infty$. This is provided in Section S2 of the Appendix. The result implies that $\widehat{\gamma}-\gamma=O_p(1/\sqrt{N})$. Lemma 8.10 in [21] shows that $\sup_{\boldsymbol{x}\in\mathbb{X}}|\widehat{h}(\boldsymbol{x},\gamma)-h_1(\boldsymbol{x})|=O_p\big((\log N)^{1/2}(N\ell^q)^{-1/2}+\ell^d\big)$, which is $o_p(N^{-2/(p+4)})$ under the assumed conditions $d>\max\{q(p+4)/(2p),q\}$ and $N\ell^{2d}\to 0$. Since $\widehat{\gamma}-\gamma$ converges faster than $\sup_{\boldsymbol{x}\in\mathbb{X}}|\widehat{h}(\boldsymbol{x},\gamma)-h_1(\boldsymbol{x})|$, $\sup_{\boldsymbol{x}\in\mathbb{X}}|\widehat{h}(\boldsymbol{x},\widehat{\gamma})-h_1(\boldsymbol{x})|=o_p(N^{-2/(p+4)})$. Then, the rest of proof follows the proof of Theorem 4.1.

**Supplementary Material.** The supplementary material contains the Appendix for all technical lemmas and proofs.

## REFERENCES

[1] BIERENS, H. J. (1987). Kernel estimators of regression functions. In *Advances in Econometrics: Fifth World Congress* **1** 99–144.

[2] BRESLOW, N. E. and HOLUBKOV, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 447–461.
[3] CHATTERJEE, N., CHEN, Y.-H., MAAS, P. and 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** 107–117.
[4] CHEN, Y.-H. and CHEN, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62** 449–460.
[5] COOK, R. D. and WEISBERG, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association* **86** 328–332.
[6] DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association* **87** 376–382.
[7] EUBANK, R. L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd ed. CRC Press.
[8] FAN, J., FARMEN, M. and GIJBELS, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 591–608.
[9] FAN, J. and GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics* **20** 2008–2036.
[10] FAN, J., GASSER, T., GIJBELS, I., BROCKMANN, M. and ENGEL, J. (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics* **49** 79–99.
[11] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.
[12] JOHNSON, A. E. W., POLLARD, T. J., SHEN, L., LEHMAN, L.-W. H., FENG, M., GHASSEMI, M., MOODY, B., SZOLOVITS, P., ANTHONY CELI, L. and MARK, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data* **3** 160035.
[13] KIM, H. J., WANG, Z. and KIM, J. K. (2021). Survey data integration for regression analysis using model calibration. *arXiv 2107.06448*.
[14] LAWLESS, J., KALBFLEISCH, J. and WILD, C. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** 413–438.
[15] LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327.
[16] LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102** 997-1008.
[17] LIU, D., GÖRGES, M. and JENKINS, S. A. (2012). University of Queensland Vital Signs Dataset: Development of an Accessible Repository of Anesthesia Patient Monitoring Data for Research. *Anesthesia & Analgesia* **114**.
[18] LOHR, S. L. and RAGHUNATHAN, T. E. (2017). Combining survey data with other data sources. *Statistical Science* **32** 293–312.
[19] MA, Y. and ZHU, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107** 168–179.
[20] MERKOURIS, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association* **99** 1131–1139.
[21] NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4** 2111–2245.
[22] OPSOMER, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis* **73** 166–179.
[23] QIN, J., ZHANG, H., LI, P., ALBANES, D. and YU, K. (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* **102** 169–180.
[24] RAO, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B* **83** 242–272.
[25] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.
[26] SCOTT, A. J. and WILD, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84** 57–71.
[27] SHAO, J. (2003). *Mathematical Statistics*, 2nd ed. Springer, New York.
[28] SHAO, Y., COOK, R. D. and WEISBERG, S. (2007). Marginal Tests with Sliced Average Variance Estimation. *Biometrika* **94** 285–296.
[29] WAND, M. P. and JONES, M. C. (1994). *Kernel Smoothing*. *Chapman & Hall/CRC Monographs on Statistics & Applied Probability* **60**. Chapman & Hall, Boca Raton, FL, U.S.

[30] WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York.
[31] WU, C. and SITTER, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* **96** 185–193.
[32] XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2002). An Adaptive Estimation of Dimension Reduction Space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64** 363–410.
[33] YANG, S. and KIM, J. K. (2020). Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science* **3** 625–650.
[34] ZHANG, Y., OUYANG, Z. and ZHAO, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics* **11** 161.
[35] ZIESCHANG, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association* **85** 986–1001.

TABLE 2
*Simulated MISE (30) and IMP (31) with $S = 200$ under setting S1*

| Covariate | Model | Test data | $b, l$ | Estimator (2) | Estimator (9) | CK estimator (8) with constraint (6) and $\boldsymbol{g} =$ 1 | $(1, X)$ | $(1, \widehat{h})$ | $(1, X, \widehat{h})$ | $\widehat{g}^*$ | IMP % |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Normal | M1 | Sample | Best | 0.055 | 0.055 | 0.052 | 0.041 | 0.047 | 0.058 | 0.056 | 24.79 |
| | | | CV | 0.065 | 0.065 | 0.063 | 0.054 | 0.053 | 0.068 | 0.066 | 18.00 |
| | | Grid | Best | 0.029 | 0.029 | 0.026 | 0.019 | 0.017 | 0.020 | 0.029 | 40.84 |
| | | | CV | 0.029 | 0.029 | 0.027 | 0.023 | 0.021 | 0.024 | 0.030 | 28.03 |
| | M2 | Sample | Best | 0.080 | 0.081 | 0.081 | 0.079 | 0.076 | 0.072 | 0.082 | 10.65 |
| | | | CV | 0.084 | 0.086 | 0.084 | 0.083 | 0.080 | 0.076 | 0.087 | 10.25 |
| | | Grid | Best | 0.059 | 0.060 | 0.057 | 0.055 | 0.054 | 0.050 | 0.060 | 15.72 |
| | | | CV | 0.066 | 0.067 | 0.067 | 0.065 | 0.068 | 0.063 | 0.068 | 4.33 |
| | M3 | Sample | Best | 0.089 | 0.089 | 0.089 | 0.087 | 0.092 | 0.091 | 0.090 | 2.30 |
| | | | CV | 0.098 | 0.099 | 0.096 | 0.095 | 0.099 | 0.099 | 0.100 | 3.36 |
| | | Grid | Best | 0.057 | 0.057 | 0.054 | 0.053 | 0.055 | 0.053 | 0.058 | 7.52 |
| | | | CV | 0.061 | 0.062 | 0.062 | 0.061 | 0.066 | 0.064 | 0.063 | 0.22 |
| | M4 | Sample | Best | 0.061 | 0.061 | 0.058 | 0.051 | 0.054 | 0.061 | 0.062 | 17.06 |
| | | | CV | 0.071 | 0.072 | 0.070 | 0.063 | 0.060 | 0.071 | 0.073 | 15.64 |
| | | Grid | Best | 0.030 | 0.031 | 0.027 | 0.024 | 0.020 | 0.021 | 0.031 | 35.40 |
| | | | CV | 0.031 | 0.031 | 0.030 | 0.027 | 0.023 | 0.026 | 0.031 | 25.86 |
| Uniform | M1 | Sample | Best | 0.020 | 0.021 | 0.018 | 0.011 | 0.010 | 0.012 | 0.020 | 49.13 |
| | | | CV | 0.028 | 0.029 | 0.025 | 0.017 | 0.016 | 0.018 | 0.027 | 42.62 |
| | | Grid | Best | 0.043 | 0.045 | 0.042 | 0.030 | 0.028 | 0.032 | 0.044 | 36.43 |
| | | | CV | 0.060 | 0.058 | 0.055 | 0.041 | 0.039 | 0.045 | 0.055 | 32.73 |
| | M2 | Sample | Best | 0.039 | 0.039 | 0.036 | 0.035 | 0.033 | 0.030 | 0.040 | 23.19 |
| | | | CV | 0.048 | 0.048 | 0.044 | 0.042 | 0.040 | 0.037 | 0.048 | 24.17 |
| | | Grid | Best | 0.097 | 0.097 | 0.093 | 0.090 | 0.090 | 0.078 | 0.098 | 18.91 |
| | | | CV | 0.114 | 0.115 | 0.111 | 0.108 | 0.110 | 0.095 | 0.116 | 17.18 |
| | M3 | Sample | Best | 0.033 | 0.033 | 0.031 | 0.030 | 0.029 | 0.029 | 0.034 | 13.78 |
| | | | CV | 0.042 | 0.043 | 0.039 | 0.037 | 0.036 | 0.036 | 0.043 | 15.94 |
| | | Grid | Best | 0.083 | 0.083 | 0.080 | 0.077 | 0.079 | 0.073 | 0.084 | 12.61 |
| | | | CV | 0.103 | 0.104 | 0.101 | 0.097 | 0.102 | 0.093 | 0.104 | 9.30 |
| | M4 | Sample | Best | 0.021 | 0.022 | 0.018 | 0.012 | 0.011 | 0.012 | 0.021 | 47.77 |
| | | | CV | 0.029 | 0.030 | 0.026 | 0.018 | 0.017 | 0.018 | 0.027 | 43.01 |
| | | Grid | Best | 0.046 | 0.049 | 0.045 | 0.033 | 0.030 | 0.035 | 0.047 | 34.47 |
| | | | CV | 0.063 | 0.062 | 0.057 | 0.044 | 0.042 | 0.046 | 0.058 | 32.09 |

Red indicates the best one among all methods.

TABLE 3
*Simulated MISE (30) and IMP (31) with $S = 200$ under setting S2 and $\gamma = 0$*

| Covariate | Model | Test data | $b, l$ | Estimator (2) | Estimator (9) | CK estimator (8) with constraint (6) | (22) | (25) | (29) | IMP % | $\widehat{\gamma}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Normal | M1 | Sample | Best | 0.057 | 0.057 | 0.040 | 0.038 | 0.051 | 0.044 | 33.26 | -0.003 |
| | | | CV | 0.073 | 0.064 | 0.048 | 0.046 | 0.060 | 0.055 | 36.66 | -0.006 |
| | | Grid | Best | 0.054 | 0.056 | 0.027 | 0.019 | 0.038 | 0.030 | 64.86 | -0.012 |
| | | | CV | 0.063 | 0.068 | 0.042 | 0.036 | 0.054 | 0.047 | 42.10 | -0.013 |
| | M2 | Sample | Best | 0.080 | 0.080 | 0.130 | 0.073 | 0.081 | 0.078 | 8.75 | -0.033 |
| | | | CV | 0.091 | 0.091 | 0.143 | 0.083 | 0.093 | 0.089 | 8.79 | -0.033 |
| | | Grid | Best | 0.093 | 0.093 | 0.086 | 0.085 | 0.095 | 0.090 | 8.60 | -0.037 |
| | | | CV | 0.110 | 0.112 | 0.103 | 0.101 | 0.115 | 0.108 | 8.18 | -0.043 |
| | M3 | Sample | Best | 0.077 | 0.077 | 0.092 | 0.072 | 0.078 | 0.076 | 6.49 | -0.016 |
| | | | CV | 0.087 | 0.086 | 0.102 | 0.080 | 0.088 | 0.085 | 6.97 | -0.018 |
| | | Grid | Best | 0.067 | 0.067 | 0.066 | 0.059 | 0.067 | 0.062 | 11.94 | 0.001 |
| | | | CV | 0.087 | 0.090 | 0.088 | 0.081 | 0.091 | 0.087 | 6.89 | -0.006 |
| | M4 | Sample | Best | 0.061 | 0.061 | 0.053 | 0.040 | 0.054 | 0.047 | 33.83 | -0.004 |
| | | | CV | 0.076 | 0.067 | 0.062 | 0.051 | 0.064 | 0.059 | 32.94 | -0.010 |
| | | Grid | Best | 0.059 | 0.060 | 0.045 | 0.022 | 0.045 | 0.033 | 62.17 | -0.000 |
| | | | CV | 0.064 | 0.069 | 0.061 | 0.040 | 0.060 | 0.054 | 36.77 | -0.014 |
| Uniform | M1 | Sample | Best | 0.022 | 0.024 | 0.004 | 0.004 | 0.014 | 0.014 | 80.43 | -0.003 |
| | | | CV | 0.029 | 0.024 | 0.005 | 0.005 | 0.014 | 0.014 | 80.96 | -0.007 |
| | | Grid | Best | 0.059 | 0.060 | 0.017 | 0.019 | 0.034 | 0.040 | 69.79 | 0.008 |
| | | | CV | 0.090 | 0.093 | 0.042 | 0.042 | 0.061 | 0.065 | 53.57 | -0.014 |
| | M2 | Sample | Best | 0.039 | 0.039 | 0.035 | 0.032 | 0.041 | 0.037 | 17.94 | -0.014 |
| | | | CV | 0.044 | 0.045 | 0.043 | 0.039 | 0.047 | 0.044 | 11.36 | -0.016 |
| | | Grid | Best | 0.124 | 0.124 | 0.103 | 0.119 | 0.126 | 0.120 | 16.93 | -0.005 |
| | | | CV | 0.158 | 0.166 | 0.126 | 0.148 | 0.170 | 0.157 | 20.25 | -0.032 |
| | M3 | Sample | Best | 0.033 | 0.033 | 0.028 | 0.027 | 0.035 | 0.034 | 19.54 | -0.006 |
| | | | CV | 0.039 | 0.041 | 0.035 | 0.033 | 0.042 | 0.041 | 13.63 | -0.009 |
| | | Grid | Best | 0.103 | 0.103 | 0.094 | 0.097 | 0.103 | 0.101 | 9.34 | 0.009 |
| | | | CV | 0.130 | 0.140 | 0.118 | 0.125 | 0.144 | 0.139 | 9.81 | -0.017 |
| | M4 | Sample | Best | 0.023 | 0.023 | 0.005 | 0.005 | 0.015 | 0.014 | 77.03 | -0.003 |
| | | | CV | 0.029 | 0.025 | 0.006 | 0.006 | 0.015 | 0.015 | 78.59 | -0.007 |
| | | Grid | Best | 0.063 | 0.064 | 0.023 | 0.025 | 0.040 | 0.044 | 62.57 | 0.007 |
| | | | CV | 0.096 | 0.097 | 0.046 | 0.049 | 0.066 | 0.072 | 52.23 | -0.015 |

Red indicates the best one among all methods.
For CK estimator under all constraints, $\boldsymbol{g}(X) = (1, X)$.
Result for $\widehat{\gamma}$ is the simulation mean with simulation standard deviation between 0.005 and 0.006.

TABLE 4
*Simulated MISE (30) and IMP (31) with $S = 200$ under setting S2 and $\gamma = 1/2$*

| Covariate | Model | Test data | $b, l$ | Estimator (2) | Estimator (9) | CK estimator (8) with constraint (6) | (22) | (25) | (29) | IMP % | $\widehat{\gamma}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Normal | M1 | Sample | Best | 0.054 | 0.054 | 0.237 | 0.318 | 0.048 | 0.040 | 25.26 | 0.453 |
| | | | CV | 0.068 | 0.062 | 0.257 | 0.341 | 0.057 | 0.051 | 17.08 | 0.458 |
| | | Grid | Best | 0.049 | 0.049 | 0.126 | 0.190 | 0.035 | 0.028 | 42.18 | 0.449 |
| | | | CV | 0.056 | 0.057 | 0.187 | 0.259 | 0.047 | 0.040 | 28.00 | 0.441 |
| | M2 | Sample | Best | 0.082 | 0.082 | 0.247 | 0.509 | 0.083 | 0.081 | 1.07 | 0.426 |
| | | | CV | 0.093 | 0.093 | 0.268 | 0.528 | 0.095 | 0.091 | 1.12 | 0.429 |
| | | Grid | Best | 0.089 | 0.089 | 0.293 | 0.588 | 0.089 | 0.085 | 4.21 | 0.419 |
| | | | CV | 0.099 | 0.101 | 0.305 | 0.608 | 0.103 | 0.098 | 1.31 | 0.426 |
| | M3 | Sample | Best | 0.084 | 0.083 | 0.357 | 0.560 | 0.085 | 0.084 | -0.50 | 0.449 |
| | | | CV | 0.097 | 0.099 | 0.368 | 0.565 | 0.101 | 0.095 | 1.99 | 0.442 |
| | | Grid | Best | 0.070 | 0.070 | 0.312 | 0.513 | 0.069 | 0.067 | 4.42 | 0.452 |
| | | | CV | 0.082 | 0.086 | 0.342 | 0.551 | 0.087 | 0.081 | 0.38 | 0.456 |
| | M4 | Sample | Best | 0.063 | 0.063 | 0.213 | 0.335 | 0.056 | 0.051 | 18.90 | 0.442 |
| | | | CV | 0.078 | 0.071 | 0.230 | 0.358 | 0.064 | 0.058 | 18.90 | 0.432 |
| | | Grid | Best | 0.053 | 0.053 | 0.095 | 0.189 | 0.040 | 0.033 | 37.11 | 0.439 |
| | | | CV | 0.062 | 0.066 | 0.178 | 0.290 | 0.060 | 0.053 | 14.89 | 0.440 |
| Uniform | M1 | Sample | Best | 0.021 | 0.024 | 0.221 | 0.242 | 0.015 | 0.013 | 36.29 | 0.480 |
| | | | CV | 0.028 | 0.025 | 0.219 | 0.241 | 0.015 | 0.014 | 42.94 | 0.475 |
| | | Grid | Best | 0.057 | 0.058 | 0.301 | 0.326 | 0.035 | 0.039 | 31.95 | 0.483 |
| | | | CV | 0.083 | 0.087 | 0.322 | 0.346 | 0.061 | 0.064 | 22.69 | 0.487 |
| | M2 | Sample | Best | 0.040 | 0.040 | 0.265 | 0.338 | 0.042 | 0.040 | 0.76 | 0.478 |
| | | | CV | 0.049 | 0.049 | 0.275 | 0.350 | 0.051 | 0.050 | -2.07 | 0.469 |
| | | Grid | Best | 0.127 | 0.127 | 0.459 | 0.555 | 0.125 | 0.122 | 4.15 | 0.463 |
| | | | CV | 0.155 | 0.161 | 0.489 | 0.591 | 0.163 | 0.154 | 0.59 | 0.463 |
| | M3 | Sample | Best | 0.033 | 0.033 | 0.287 | 0.328 | 0.035 | 0.034 | -4.03 | 0.475 |
| | | | CV | 0.039 | 0.039 | 0.301 | 0.343 | 0.041 | 0.041 | -6.37 | 0.486 |
| | | Grid | Best | 0.105 | 0.105 | 0.433 | 0.484 | 0.102 | 0.101 | 4.04 | 0.483 |
| | | | CV | 0.126 | 0.126 | 0.466 | 0.519 | 0.130 | 0.126 | -0.30 | 0.482 |
| | M4 | Sample | Best | 0.021 | 0.024 | 0.220 | 0.243 | 0.016 | 0.014 | 31.91 | 0.476 |
| | | | CV | 0.030 | 0.027 | 0.217 | 0.238 | 0.017 | 0.015 | 42.21 | 0.477 |
| | | Grid | Best | 0.064 | 0.065 | 0.315 | 0.344 | 0.043 | 0.044 | 30.67 | 0.486 |
| | | | CV | 0.086 | 0.091 | 0.343 | 0.370 | 0.068 | 0.068 | 20.44 | 0.489 |

Red indicates the best one among all methods.
For CK estimator under all constraints, $\boldsymbol{g}(X) = (1, X)$.
Result for $\widehat{\gamma}$ is the simulation mean with simulation standard deviation between 0.005 and 0.006.