Statistica Sinica

# Kernel Regression Utilizing External Information as Constraints

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

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

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

*Abstract:* 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 Chatterjee et al. (2016) proposes a parametric likelihood approach for analyzing a main "internal" dataset using constraints formulated with information from 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 summary statistics. Under some conditions, 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. Simulation results are obtained to confirm our theory and to quantify the improvements from utilizing external data. An example of application is also included for illustration.

*Key words and phrases:* Asymptotic mean integrated square error, constraints, data integration, external summary statistics, two-step kernel regression.

## 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 independent external datasets, which typically have very large sizes but often contain relatively crude information such as summary statistics, 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 information from external datasets. This line of research fits into a more general framework of data integration Kim et al. (2021); Lohr and Raghunathan (2017); Merkouris (2004); Rao (2021); Yang and Kim (2020); Zhang et al. (2017); Zieschang (1990) 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 based 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 $\boldsymbol{Z}$.

Under the same setting and a parametric model between the response $Y$ and covariate vector $\boldsymbol{U}$, Chatterjee et al. (2016) 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 (Breslow and Holubkov, 1997; Chen and Chen, 2000; Deville and Särndal, 1992; Kim et al., 2021; Lawless et al., 1999; Qin et al., 2015; Scott and Wild, 1997; Wu and Sitter, 2001).

We focus on nonparametric kernel regression Bierens (1987); Wand and Jones (1994); Wasserman (2006), 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 kernel-type methods, we focus on a low dimensional covariate $\boldsymbol{U}$. A discussion of handling a large dimensional $\boldsymbol{U}$ is given in Section 5.

To make use of summary 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 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, Fan and Gijbels (1992) 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, under some conditions 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. We start with the situation where internal and external data share the same population and then study robustness and some extensions to heterogeneous populations. Section 3 presents some simulation results and Section 4 contains an example. The paper ends with some discussions in Section 5. All technical details are in the Appendix as Supplementary Material.

## 2. Methodology and Theory

### 2.1 Two-step constrained kernel estimation

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

$$\mu(\boldsymbol{u}) = \mathrm{E}(Y \mid \boldsymbol{U} = \boldsymbol{u}), \tag{2.1}$$

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 (2.1) for any fixed $\boldsymbol{u} \in \mathbb{U}$ based on the internal dataset is

$$\begin{aligned} \widehat{\mu}_K(\boldsymbol{u}) &= \arg\min_{\mu} \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)(Y_i - \mu)^2 \\ &= \sum_{i=1}^{n} Y_i \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \end{aligned} \tag{2.2}$$

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 $(Y, \boldsymbol{X})$, independent of the internal sample, where $\boldsymbol{X}$ is a $q$-dimensional sub-vector of $\boldsymbol{U}$, $q \le p$. We consider the scenario where only some summary statistics are available from the external dataset. Specifically, the external dataset provides a 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 $\boldsymbol{a}^\top$ denotes the transpose of vector $\boldsymbol{a}$ throughout the paper, and $\boldsymbol{g}$ is a function from $\mathbb{R}^q$ to $\mathbb{R}^k$ with a fixed $k$. The form of $\boldsymbol{g}$ is known and given as part of the external information. For example, $\boldsymbol{g}(\boldsymbol{X}) = (1, \boldsymbol{X}^\top)^\top$.

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

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

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

$$\mathrm{E}[\{\boldsymbol{\beta}_g^\top \boldsymbol{g}(\boldsymbol{X}) - \mu(\boldsymbol{U})\}\boldsymbol{g}(\boldsymbol{X})^\top] = 0. \tag{2.3}$$

In (2.3), the external information $\boldsymbol{\beta}_g^\top \boldsymbol{g}(\boldsymbol{X})$ can be viewed as a projection of $\mu(\boldsymbol{U})$ into the linear space of $\boldsymbol{g}(\boldsymbol{X})$. Since $\mu(\boldsymbol{U})$ is directly involved in constraint (2.3), this constraint is particularly useful for kernel regression. It is different from the constraint in Chatterjee et al. (2016) that is useful for parametric likelihood analysis with internal data but not for kernel regression.

We propose a two-step procedure. In the first step, we make use of (2.3) 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,

$$\widehat{\boldsymbol{\mu}} = \arg \min_{(\mu_1,...,\mu_n)^\top \in \mathbb{R}^n} \sum_{i=1}^n \sum_{j=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)(Y_j - \mu_i)^2 \Big/ \sum_{k=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k) \tag{2.4}$$

$$\text{subject to} \quad \sum_{i=1}^n \{\widehat{\boldsymbol{\beta}}_g^\top \boldsymbol{g}(\boldsymbol{X}_i) - \mu_i\}\boldsymbol{g}(\boldsymbol{X}_i)^\top = 0, \tag{2.5}$$

where the constraint in (2.5) is an empirical analog of (2.3) for the estimation of $\boldsymbol{\mu}$ based on internal data and $l$ in (2.4) is a bandwidth that may be different from $b$ in (2.2). A discussion about the selection of bandwidths is given in Section 2.3.

To motivate the objective function in (2.4) 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 (2.4) 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 (2.4), 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 (2.2). Then solving (2.4)-(2.5) 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

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_K + \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top (\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\mu}}_K). \tag{2.6}$$

This estimator adds an adjustment term to $\widehat{\boldsymbol{\mu}}_K$, the estimator in (2.2) 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 (2.5), $\widehat{\boldsymbol{\mu}}$ in (2.6) is expected to be better than $\widehat{\boldsymbol{\mu}}_K$ that does not use external information, when the sample size of the external dataset is at least comparable with the sample size of the internal dataset. Proposition 1 in Section 2.2 quantifies this improvement.

To obtain an improved estimator of the entire regression function $\mu(\boldsymbol{u})$ defined by (2.1), 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 constrained kernel estimator of $\mu(\boldsymbol{u})$ is

$$\widehat{\mu}_{CK}(\boldsymbol{u}) = \sum_{i=1}^{n} \widehat{\mu}_i \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i), \tag{2.7}$$

where $b$ is the same bandwidth in (2.2).

### 2.2 Asymptotic theory

We now establish the asymptotic normality of $\widehat{\mu}_{CK}(\boldsymbol{u})$ in (2.7) for a fixed $\boldsymbol{u}$, as the sample size $n$ of the internal dataset increases to infinity. All technical proofs in this section are given in the Appendix (Supplementary Material).

**Theorem 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 (2.2) and $l$ in (2.4) satisfy $b \to 0$, $l \to 0$, $l/b \to r \in (0, \infty)$, $nb^p \to \infty$, and $nb^{4+p} \to c \in [0, \infty)$, as 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 fixed $\boldsymbol{u} \in \mathbb{U}$,*

$$\sqrt{nb^p}\{\widehat{\mu}_{CK}(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \xrightarrow{d} N\big(B_{CK}(\boldsymbol{u}), V_{CK}(\boldsymbol{u})\big),$$

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

$$\begin{aligned} B_{CK}(\boldsymbol{u}) &= c^{1/2}[(1+r^2)A(\boldsymbol{u}) - r^2 \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} \mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}], \\ A(\boldsymbol{u}) &= \int \kappa(\boldsymbol{v}) \left\{ \tfrac{1}{2} \boldsymbol{v}^\top \nabla^2 \mu(\boldsymbol{u}) \boldsymbol{v} + \boldsymbol{v}^\top \nabla \log f_U(\boldsymbol{u}) \nabla \mu(\boldsymbol{u})^\top \boldsymbol{v} \right\} d\boldsymbol{v}, \\ V_{CK}(\boldsymbol{u}) &= \frac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})} \int \left\{ \int \kappa(\boldsymbol{v} - r\boldsymbol{w}) \kappa(\boldsymbol{w}) d\boldsymbol{w} \right\}^2 d\boldsymbol{v}, \end{aligned} \tag{2.8}$$

*and $f_U$ is the density of $\boldsymbol{U}$.*

(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 $\mathbb{U}$. It is a sufficient condition in our proof of the efficiency of $\widehat{\boldsymbol{\mu}}$ in (2.6) in the first step.

From the theory of standard kernel regression (Opsomer, 2000), under (A1)-(A4), the kernel estimator $\widehat{\mu}_K(\boldsymbol{u})$ in (2.2) also satisfies

$$\begin{aligned} &\sqrt{nb^p}\{\widehat{\mu}_K(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \xrightarrow{d} N\big(B_K(\boldsymbol{u}) V_K(\boldsymbol{u})\big), \\ &B_K(\boldsymbol{u}) = c^{1/2} A(\boldsymbol{u}), \quad V_K(\boldsymbol{u}) = \tfrac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})} \textstyle\int \{\kappa(\boldsymbol{v})\}^2 d\boldsymbol{v}. \end{aligned} \tag{2.9}$$

Theorem 1 and (2.9) indicate that the use of external information does not improve the convergence rate $1/\sqrt{nb^p}$ in estimating $\mu(\boldsymbol{u})$, regardless of what $m$ is, due to the facts that

(i) the summary information from external data is not in the form of kernel regression and (ii) the estimation of $\mu(\boldsymbol{u})$ involves $\boldsymbol{Z} = \boldsymbol{z}$ which is not in the external dataset.

The use of external information does affect the asymptotic bias and variance in kernel estimation of $\mu(\boldsymbol{u})$. We now compare asymptotic performances of the proposed estimator (2.7) and the standard kernel estimator (2.2) without using external information, although they have the same convergence rate.

Our first result is about a comparison of predicting $\boldsymbol{\mu} = (\mu(\boldsymbol{U}_1), ..., \mu(\boldsymbol{U}_n))^\top$. For the standard kernel (2.2), $\boldsymbol{\mu}$ is predicted as $\widehat{\boldsymbol{\mu}}_K = (\widehat{\mu}_K(\boldsymbol{U}_1), ..., \widehat{\mu}_K(\boldsymbol{U}_n))^\top$; for the proposed estimator (2.7), $\boldsymbol{\mu}$ is predicted as $\widehat{\boldsymbol{\mu}}$ in (2.6). The following result shows that, with probability tending to 1 as $n \to \infty$, $\|\widehat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}\|^2 \geq \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$, where $\|\boldsymbol{a}\|^2 = \boldsymbol{a}^\top \boldsymbol{a}$ for vector $\boldsymbol{a}$.

**Proposition 1.** *Under the conditions in Theorem 1 and $nb^4 \to \infty$,*

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

*where $\xrightarrow{p}$ denotes convergence in probability as $n \to \infty$ and $A(\boldsymbol{u})$ is defined in (2.8).*

This result shows the usefulness of constraint (2.3) from external information. Even in the situation where there is no covariate in the external dataset, i.e., $\boldsymbol{g} \equiv 1$ and $\boldsymbol{\beta}_g = E(Y)$, constraint (2.3) is still useful because it becomes $E(Y) = E\{\mu(\boldsymbol{U})\}$ with $E(Y)$ estimated by external information $\widehat{\boldsymbol{\beta}}_g$ = the sample mean of $Y$'s in the external dataset to help the estimation of $\mu$ using internal data.

For any kernel estimator $\widehat{\mu}(\boldsymbol{u})$ satisfying $\sqrt{nb^p}\{\widehat{\mu}(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \xrightarrow{d} N\big(B(\boldsymbol{u}), V(\boldsymbol{u})\big)$, we consider the asymptotic mean integrated square error (AMISE), a globe accuracy measure

often used in the literature (Fan and Gijbels, 1992):

$$\mathrm{AMISE}(\widehat{\mu}) = \mathrm{E}[\{B(\boldsymbol{U})\}^2 + V(\boldsymbol{U})].$$

We now compare the proposed $\widehat{\mu}_{CK}$ in (2.7) with the standard $\widehat{\mu}_K$ in (2.2) in terms of AMISE. From (2.8) and (2.9),

$$\mathrm{E}\{V_K(\boldsymbol{U}) - V_{CK}(\boldsymbol{U})\} = \{\rho(0) - \rho(r)\}\,\mathrm{E}\{\sigma^2(\boldsymbol{U})/f_U(\boldsymbol{U})\},$$

where $r$ is given in (A4) and

$$\rho(r) = \int \left\{ \int \kappa(\boldsymbol{w} - r\boldsymbol{v})\kappa(\boldsymbol{v})d\boldsymbol{v} \right\}^2 d\boldsymbol{w}. \tag{2.10}$$

Under mild conditions (e.g., Example 1 and Proposition 2), $\rho(0) - \rho(r) \geq 0$ and, hence, using external information reduces variability in kernel estimation. On the other hand, if we define $A_g(\boldsymbol{X}) = \boldsymbol{g}(\boldsymbol{X})^\top \boldsymbol{\Sigma}_g^{-1} \mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})A(\boldsymbol{U})\}$, then $\mathrm{E}[\{A(\boldsymbol{U}) - A_g(\boldsymbol{X})\}A_g(\boldsymbol{X})] = 0$ and, consequently,

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

where $c$ and $r$ are given in (A4). This indicates that the expected squared asymptotic bias of $\widehat{\mu}_{CK}$ is larger than that of $\widehat{\mu}_K$ and the difference is measured by $\mathrm{E}\{A(\boldsymbol{U}) - A_g(\boldsymbol{X})\}^2$, i.e., how good is $A_g(\boldsymbol{X})$ as an approximation to $A(\boldsymbol{U})$ using external information. If external information is very useful so that $\mathrm{E}\{A(\boldsymbol{U}) - A_g(\boldsymbol{X})\}^2$ is close to 0, then $\mathrm{E}\{B_{CK}(\boldsymbol{U})\}^2$ is close to $\mathrm{E}\{B_K(\boldsymbol{U})\}^2$.

Combining the results for expected asymptotic variance and squared asymptotic bias, we conclude that, in terms of AMISE, the proposed $\widehat{\mu}_{CK}$ is better than the standard $\widehat{\mu}_K$ if and only if (see the proof of Proposition 2 in the Appendix)

$$c < \tau \frac{\rho(0) - \rho(r)}{r^2(2 + r^2)}, \qquad \tau = \frac{\mathrm{E}\{\sigma^2(\boldsymbol{U})/f_U(\boldsymbol{U})\}}{\mathrm{E}\{A(\boldsymbol{U}) - A_g(\boldsymbol{X})\}^2}. \tag{2.11}$$

The value of $\tau$ in (2.11) can be viewed as a bias-variance trade-off in using external information. In application, the bandwidth $b$ (and thus its limit $c = \lim_{n\to\infty} nb^{4+p}$) is often chosen to be related with the variability. For example, when $\sigma^2(\boldsymbol{u}) = \sigma^2$ does not depend on $\boldsymbol{u}$, Theorem 4.2 in Eubank (1999) shows that the optimal bandwidth is the one with $c = c_0\sigma^2$ for a constant $c_0 > 0$. Thus, if external information is useful and $\tau$ is large, then a $c$ satisfying the inequality in (2.11) can be achieved and $\widehat{\mu}_{CK}$ is better than $\widehat{\mu}_K$ in items of AMISE. On the other hand, if $\tau$ is small, we may not be able to choose a $c$ satisfying the inequality in (2.11) to achieve a meaningful/reasonable improvement.

**Example 1** (Gaussian kernels)**.** The Gaussian kernel $\kappa(\boldsymbol{u}) = (2\pi)^{-p/2}e^{-\|\boldsymbol{u}\|^2/2}$ is the density of $p$-dimensional normal distribution $N(0, \boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the identity matrix of order $p$. For this kernel, $\int \kappa(\boldsymbol{w} - r\boldsymbol{v})\kappa(\boldsymbol{v})d\boldsymbol{v}$ is the density of $N(0, (1 + r^2)\boldsymbol{I}_p)$ and, thus, the function in (2.10) is

$$\rho(r) = \int \left[\{2\pi(1 + r^2)\}^{-p/2}e^{-\|\boldsymbol{w}\|^2/2}\right]^2 d\boldsymbol{w} = (2\sqrt{\pi})^{-p}\{(1 + r^2)\}^{-p/2}.$$

Hence, $\rho(0) - \rho(r) = \{1 - (1 + r^2)^{-p/2}\}/(2\sqrt{\pi})^p > 0$ for any $r > 0$ and, in terms of AMISE, $\widehat{\mu}_{CK}$ is better than $\widehat{\mu}_K$ if and only if

$$c < \tau \frac{\{1 - (1 + r^2)^{-p/2}\}}{(2\sqrt{\pi})^p\, r^2(2 + r^2)}.$$

The result in Example 1 can be extended to non-Gaussian kernels as summarized in the following result.

**Proposition 2.** *Assume the conditions in Theorem 1 with $r \leq 1$. Assume further that the function in (2.10) has continuous second-order derivative $\rho''(s) < 0$ for $0 < s < 1$. Then* $\text{AMISE}(\widehat{\mu}_{CK}) < \text{AMISE}(\widehat{\mu}_K)$ *if and only if*

$$c < \tau \frac{-\int_0^1 (1-t)^2 \rho''(rt) dt}{2(2+r^2)}.$$

### 2.3 Bandwidth selection

(A4) in Theorem 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 can apply the following $k$-fold cross-validation (CV) as described in Györfi et al. (2002). Let $\mathcal{G}_1, ..., \mathcal{G}_k$ be a random partition of the internal dataset with approximately equal size $n/k$, and let $\widehat{\mu}_{CK}^{(-j)}(\boldsymbol{u})$ be the estimator in (2.7) with bandwidths $l$ and $b$ but without using data $\{(Y_i, \boldsymbol{U}_i), i \in \mathcal{G}_j\}$, $j = 1, ..., k$. Then, over a reasonable range, we select $(l, b)$ that minimizes

$$\text{CV}(l, b) = \sum_{j=1}^{k} \sum_{i \in \mathcal{G}_j} \{\widehat{\mu}_{CK}^{(-j)}(\boldsymbol{U}_i) - Y_i\}^2. \tag{2.12}$$

When $n$ is not very large, there may be not enough validation terms in (2.12) and we may apply the following repeated sub-sampling cross-validation (RSCV) as an alternative. We independently create $\mathcal{G}_1, ..., \mathcal{G}_B$, where each $\mathcal{G}_j$ is a subset of the internal dataset with size $n_0$ and $n - n_0$ is comparable with $n$. Then, we select $(l, b)$ that minimizes $\text{CV}(l, b)$ in (2.12) with $k$ replaced by $B$. Note that $B$ can be a large number, not like the restricted $k$ in the $k$-fold CV.

### 2.4 Confidence intervals

Confidence intervals for $\mu(\boldsymbol{u})$ at a fixed $\boldsymbol{u}$ based on kernel estimation has been studied extensively in the literature (Fan and Gijbels, 1996; Eubank, 1999; Wasserman, 2006). The main technical difficulty is, regardless whether external information is utilized or not, how to handle the bias in the kernel estimator of $\mu(\boldsymbol{u})$. Note that the asymptotic bias $B_K(\boldsymbol{u})$ for the standard kernel estimation and $B_{CK}(\boldsymbol{u})$ for the proposed constrained kernel estimation are not zero unless $c = 0$, and $c > 0$ leads to the best convergence rate for any kernel estimation.

If we can successfully estimate $B_K(\boldsymbol{u})$ or $B_{CK}(\boldsymbol{u})$, then confidence intervals based on kernel estimation with bias correction can be applied. However, bias estimation is difficult (Hall, 1992; Wasserman, 2006). We suggest the idea of under smoothing (Hall, 1992; Wasserman, 2006), i.e., we choose bandwidths smaller than those chosen by CV (Section 2.3) for confidence intervals. Specifically, if $b$ and $l$ are selected by CV for the CK method, then we calculate $\widehat{\mu}_{CK}(\boldsymbol{u})$ with under smoothing bandwidths $c_l l$ and $c_b b$ in the first and second stages, respectively, where $0 < c_l \leq 1$ and $0 < c_b \leq 1$ are under smoothing constants, and set set a confidence interval $[\widehat{\mu}_{CK}(\boldsymbol{u}) - z_\alpha \widehat{V}_{CK}^{1/2}(\boldsymbol{u}),\ \widehat{\mu}_{CK}(\boldsymbol{u}) + z_\alpha \widehat{V}_{CK}^{1/2}(\boldsymbol{u})]$ for $\mu(\boldsymbol{u})$, where $\widehat{V}_{CK}$ is the variance estimator according to (2.8),

$$\widehat{V}_{CK}(\boldsymbol{u}) = \frac{\widehat{\sigma}^2_{CK}(\boldsymbol{u})}{\widehat{f}_U(\boldsymbol{u})} \int \left\{ \int \kappa(\boldsymbol{v} - r\boldsymbol{w})\kappa(\boldsymbol{w}) d\boldsymbol{w} \right\}^2 d\boldsymbol{v},$$

$\widehat{f}_U$ is the kernel density estimator of $f_U$, and

$$\widehat{\sigma}^2_{CK}(\boldsymbol{u}) = \sum_{i=1}^n \{Y_i - \widehat{\mu}_{CK}(\boldsymbol{U}_i)\}^2 \kappa_{\widetilde{b}}(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^n \kappa_{\widetilde{b}}(\boldsymbol{u} - \boldsymbol{U}_i),$$

for some bandwidth $\widetilde{b}$. When $\sigma^2(\boldsymbol{u})$ does not depend on $\boldsymbol{u}$, a simplified estimator is

$$\widehat{\sigma}^2_{CK} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \widehat{\mu}_{CK}(\boldsymbol{U}_i)\}^2.$$

Similarly, if we apply the standard kernel without using external information, then the under smoothing bandwidth is $c_b b$ for $\widehat{\mu}_K$ and the confidence interval is obtained by replacing $\widehat{\mu}_{CK}(\boldsymbol{u})$ with $\widehat{\mu}_K(\boldsymbol{u})$ and $\widehat{V}_{CK}(\boldsymbol{u})$ with

$$\widehat{V}_K(\boldsymbol{u}) = \frac{\widehat{\sigma}^2_K(\boldsymbol{u})}{\widehat{f}_U(\boldsymbol{u})}\int\{\kappa(\boldsymbol{v})\}^2 d\boldsymbol{v}.$$

The performance of this confidence interval is examined by simulation in Section 3.2.

## 2.5 Robustness against heterogeneity in populations and extensions

We consider the situation where the populations for internal and external data are different. Let $R$ be the indicator for internal and external data. Let $(Y_i, \boldsymbol{U}_i, R_i)$, $i = 1, ..., N$, be iid with total sample size $N$, where $(Y_i, \boldsymbol{U}_i)$ with $R_i = 1$ are the observed internal data and $(Y_i, \boldsymbol{X}_i)$ with $R_i = 0$ are the external data but only summary statistics based on external data are available. Our interest is still to estimate the regression function for internal data population, i.e.,

$$\mu_1(\boldsymbol{u}) = \mathrm{E}(Y \mid \boldsymbol{U} = \boldsymbol{u}, R = 1), \tag{2.13}$$

which reduces to $\mu(\boldsymbol{u})$ in (2.1) when internal and external populations are the same.

The results obtained so far hold when internal and external populations are homogeneous, i.e., $(Y, \boldsymbol{X}, \boldsymbol{Z}) \perp R$, where $A \perp B$ denotes that $A$ and $B$ are independent. To what extent are the results robust against the violation of $(Y, \boldsymbol{X}, \boldsymbol{Z}) \perp R$?

With $R = 1$ and $R = 0$ indicating the internal and external data, respectively, constraint (2.3) should be replaced by

$$\mathrm{E}[\{\boldsymbol{\beta}_g^\top \boldsymbol{g}(\boldsymbol{X}) - \mu_1(\boldsymbol{U})\}\boldsymbol{g}(\boldsymbol{X})^\top | R = 1] = 0, \tag{2.14}$$

where

$$\boldsymbol{\beta}_g = [\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top | R = 0\}]^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})Y | R = 0\}, \tag{2.15}$$

because constraint (2.14) is used in the estimation of $\mu_1(\boldsymbol{u})$ in (2.13) with internal data (conditioning on $R = 1$), whereas $\boldsymbol{\beta}_g$ in (2.15) is the limit of estimator $\widehat{\boldsymbol{\beta}}_g$ based on external data (conditioning on $R = 0$). That is, if (2.14) holds, then all derived results hold after we replace (2.3) by (2.14) and constraint (2.5) by

$$\sum_{i=1}^{N} R_i\{\widehat{\boldsymbol{\beta}}_g^\top \boldsymbol{g}(\boldsymbol{X}_i) - \mu_i\}\boldsymbol{g}(\boldsymbol{X}_i)^\top = 0.$$

We now show that (2.14) holds under the condition

$$\mathrm{E}(Y \mid \boldsymbol{X}, R = 1) = \mathrm{E}(Y \mid \boldsymbol{X}, R = 0) \qquad \text{and} \qquad \boldsymbol{X} \perp R. \tag{2.16}$$

Under (2.16), $\boldsymbol{\beta}_g$ in (2.15) equals $[\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}(\boldsymbol{X})^\top | R = 1\}]^{-1}\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X})Y | R = 1\}$ (see the Supplementary Material) and, consequently,

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

i.e., (2.14) holds.

Therefore, the derived results so far are robust as long as (2.16) holds. Note that (2.16) is still much weaker than $(Y, \boldsymbol{X}, \boldsymbol{Z}) \perp R$ since the first equality in (2.16) involves only moment instead of distribution and (2.16) is actually implied by $(Y, \boldsymbol{X}) \perp R$.

Without (2.16), constraint (2.14) may not be satisfied and thus the derived results may not hold. Extensions may be possible if we have individual-level external data. Suppose that the first equality in (2.16) holds, and estimates of $\widehat{h}(\boldsymbol{x})$ of $h(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x})$ is available as external information. Then we may extend our method by replacing constraint (2.5) by

$$\sum_{i=1}^{N} R_i\{\mu_i - \widehat{h}(\boldsymbol{X}_i)\}\boldsymbol{g}(\boldsymbol{X}_i)^\top = 0. \tag{2.17}$$

Note that $\widehat{h}$ can be obtained if we have individual-level external data.

Finally, we consider an extension from a different direction. In Section 2.1, we only consider the summary-level external information from a linear regression. We can generalize it to any generalized estimating equation (GEE), such as logistic regression for a discrete response $Y$. Assume that the summary-level statistic $\widehat{\boldsymbol{\beta}}$ is a solution of the following GEE based on external data,

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

where $\boldsymbol{H}$ is a known $k$-dimensional function. As an analogy of (2.5), the following constraint for GEE summary-level information can be used,

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

## 3. Simulation Results

In this section, we present simulation results to examine the performance of our proposed CK estimator (2.7) and compare it with the standard kernel estimator (2.2) without using external information.

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

(i) bounded covariates: $X = BW_1 + (1 - B)W_2$ and $Z = BW_1 + (1 - B)W_3$, where $W_1$, $W_2$, and $W_3$ are identically distributed as uniform on $[-1, 1]$, $B$ is uniform on $[0, 1]$, and $W_1$, $W_2$, $W_3$, and $B$ are independent;

(ii) normal covariates: $(X, Z)$ is bivariate normal with means 0, variances 1, and correlation 0.5.

Conditioned on $(X, Z)$, the response $Y$ is normal with mean $\mu(X, Z)$ and variance 1, where $\mu(X, Z)$ follows one of the following four models:

M1. $\mu(X, Z) = X/2 - Z^2/4$;

M2. $\mu(X, Z) = \cos(2X)/2 + \sin(Z)$;

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

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

Note that all four models are nonlinear in $(X, Z)$; 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; internal and external data are indicated by $R = 1$ and $R = 0$, respectively, and given $(Y, X, Z)$, $R$ is generated according to $\mathrm{P}(R = 1 \mid Y, X, Z) = 1/\exp(1+2|X|)$. Under this setting, the unconditional $\mathrm{P}(R = 1)$ is between 10% and 15%.

Note that S2 is for the scenario in Section 2.5.

### 3.1 Mean integrated square error

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

$$\mathrm{MISE} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{T}\sum_{t=1}^{T}\{\widehat{\mu}_1^{(s)}(\boldsymbol{U}_{s,t}) - \mu_1(\boldsymbol{U}_{s,t})\}^2, \tag{3.1}$$

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 (2.13), 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.

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

$$\text{IMP} = 1 - \frac{\min\{\text{MISE}(\widehat{\mu}_{CK}) \text{ over all CK methods}\}}{\text{MISE}(\widehat{\mu}_K)}. \tag{3.2}$$

In all cases, we use the Gaussian kernel as introduced in Example 1. The bandwidths $b$ and $l$ in (2.7) 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 (2.12), which produces a decent bandwidth that can be applied to real data.

In application, we cannot choose $\boldsymbol{g}$ in constraint (2.5) since it is given as part of the external information. In simulation, we can try different $\boldsymbol{g}$'s to see the effect on the CK method. Under setting S1, we consider four different choices of $\boldsymbol{g}$, i.e., $\boldsymbol{g}(X) = 1$, $(1, X)^\top$, $(1, \widehat{h}(X))^\top$, and $(1, X, \widehat{h}(X))^\top$, where $\widehat{h}$ is a kernel estimator of $h(x) = \mathrm{E}(Y|X = x)$.

The simulated MISE defined in (3.1) based on $S = 500$ 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. We also calculate the integrated bias by simulation, which is given by (3.1) with $\{\widehat{\mu}_1^{(s)}(\boldsymbol{U}_{s,t}) - \mu_1(\boldsymbol{U}_{s,t})\}^2$ replaced by $\widehat{\mu}_1^{(s)}(\boldsymbol{U}_{s,t}) - \mu_1(\boldsymbol{U}_{s,t})$. The results are shown in Table A1 of the Appendix.

From Table 1, we can see that the proposed CK estimator may be substantially better (in terms of MISE) than the standard kernel estimator without using external information. The improvement in efficiency IMP defined in (3.2) is often over 10% and can be as high as 72%. The bandwidths selected by CV work well although they may not achieve the best efficiency gain. The three choices of $\boldsymbol{g}$ functions in constraint (2.5), 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 any definite superiority of one over the other. Thus, $\boldsymbol{g}(X) = (1, X)^\top$ is recommended for its simplicity.

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 the scenario in which the internal and external populations are different as described in Section 2.5. We study two CK estimators: $\widehat{\mu}_{CK}$ with constraint (2.5), which is incorrect since (2.16) does not hold, and $\widehat{\mu}_{CK}$ with constraint (2.17), which is asymptotically valid (Section 2.5). The simulated MISE based on $S = 500$ replications is shown in Table 2.

From Table 2, the estimator using constraint (2.17) is correct and more efficient than estimators without using external information. The CK estimator using constraint (2.5) is biased as (2.16) 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 (2.17).

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

### 3.2 Confidence intervals at some covariate values

The second part of the simulation studies performance of approximate 95% confidence intervals described in Section 2.4 by applying the CK and standard kernel with under smoothing. We consider setting 1 with simulation size $S = 1,000$. Table 3 shows simulated coverage probability (CP) and length of confidence intervals and the bias of kernel estimators at some values of $\boldsymbol{u}$. Note that the length is proportional to the simulation average of estimation squared error and thus it indicates the efficiency of the kernel estimator as well as confidence interval. Values of under smoothing scales $c_b$ and $c_l$ (see Section 2.4) and the true $\mu(\boldsymbol{u})$ are also included in Table 3.

From Table 3, in the case where covariates are bounded, all confidence intervals perform well in terms of CP. The intervals based on the CK method has much shorter lengths than the intervals based on standard kernel without using external information. For normally distributed covariates, the intervals do not have very good CP in a few cases, indicating that the asymptotic theory has not kicked in, although the CK interval is generally shorter than the interval based on the standard kernel.

## 4. Application: An Example

We apply the proposed method to the University of Queensland Vital Signs Dataset (UQVSD) for intensive care patients (Liu et al., 2012), used as the internal dataset. The response $Y$ under consideration is the systolic blood pressure, a critical biomarker for health conditions. We are interested in how $Y$ is affected by the following two covariates collected via a sensor-

gas analysis, the inspired oxygen (inO2) and end-tidal oxygen (etO2) concentration. In addition, we consider three other covariates, heart rate, respiratory rate, and blood oxygen saturation. Because the sample size is only $n = 32$, it is important to find a help from external data.

We utilize the Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson et al., 2016) as an external dataset with a large sample size 54,060. This dataset is a freely available digital health record database with information of patients needing critical care. Since both data sets study intensive care units, they can be considered as samples from the same population or from similar populations. The external dataset MIMIC-III, however, does not have covariates inO2 and etO2, although both datasets share the same response $Y$ and covariates heart rate, respiratory rate, and blood oxygen saturation. Thus, inO2 and etO2 are considered as two components of $\boldsymbol{Z}$.

Since the sample size for internal dataset is only 32, we want to use kernel regression with a lower dimension and, thus, consider a linear combination of heart rate, respiratory rate, and blood oxygen saturation as a one-dimensional covariate $X$. The coefficients of this linear combination are from the first eigenvector of the well-known sufficient dimension reduction algorithm SAVE (Cook and Weisberg, 1991; Shao et al., 2007), from which the first eigenvector provides more than 94% variability. Therefore, the kernel regression uses a three-dimensional covariate $\boldsymbol{U}$.

Since we have all external individual-level data, we use them in two ways. The first way uses constraint (2.5) in which $\boldsymbol{g}^\top = (1, X)$ and $\widehat{\boldsymbol{\beta}}_g$ is the least squares estimator under a linear regression between $Y$ and covariate $X$. The second way is to consider constraint

(2.17) to allow the populations from two datasets to be different. For comparison, we also include the standard kernel estimator (2.2). All the bandwidths are selected via the RSCV with $B = 100$ and $n_0 = 3$ (Section 2.3).

Figures 1-2 show two plots of the fitted kernel regression of $Y$ to three covariates, $X$, inO2, and etO2, using three kernel methods previously described. Since we cannot produce a four-dimensional figure for $Y$ and three covariates, Figure 1 shows the relationship among $Y$, $X$, and etO2 when inO2 is fixed at three quartiles, 61.2, 67.7, and 77.9, while Figure 2 shows the relationship among $Y$, $X$, and inO2 when etO2 is fixed at three quartiles, 56.3, 63.7, and 72.0. Table 4 shows 95% confidence intervals for systolic blood pressure under some selected covariate values with under smoothing scale $c_b = 0.8$, $c_l = 1$ and simplified variance estimator $\widehat{\sigma}^2_{CK}$ in Section 2.4.

It can be seen that the CK provides a clean pattern for the relationship among $Y$ and covariates while the standard kernel regression without using external information provides vague and flat regressions. Furthermore, the CK provides shorter confidence intervals.

## 5. Discussion

Curse of dimensionality is a well-known problem for nonparametric methods. Thus, the proposed CK method in Section 2 is intended for low dimensional covariate $\boldsymbol{U}$, i.e., $p$ is small. If $p$ is not small, then we should reduce the dimension of $\boldsymbol{U}$ prior to applying the CK, or any kernel methods. For example, consider a single index model assumption (Li, 1991), i.e., $\mu(\boldsymbol{U})$ in (2.1) is assumed to be

$$\mu(\boldsymbol{U}) = \mu(\boldsymbol{\eta}^\top \boldsymbol{U}), \tag{5.1}$$

where $\boldsymbol{\eta}$ is an unknown $p$-dimensional vector. The well-known SIR technique (Li, 1991) can be applied to obtain a consistent and asymptotically normal estimator $\widehat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ in (5.1). 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 (5.1) (Cook and Weisberg, 1991; Li and Wang, 2007; Shao et al., 2007; Xia et al., 2002; Ma and Zhu, 2012). In fact, we reduce the dimension using the method in Cook and Weisberg (1991) and Shao et al. (2007) in the example (Section 4).

We turn to the dimension of $\boldsymbol{X}$ in the external dataset. In the situation where (2.16) holds, constraint (2.5) 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, then we may consider the following approach. Instead of using constraint (2.5), we use component-wise constraints

$$\sum_{i=1}^{n} \{\mu_i - \widehat{h}^{(k)}(X_i^{(k)})\} \boldsymbol{g}_k(X_i^{(k)})^\top = 0, \qquad k = 1, ..., q, \tag{5.2}$$

where $X_i^{(k)}$ is the $k$th component of $\boldsymbol{X}_i$, $\boldsymbol{g}_k(X^{(k)})$ is a function of $X^{(k)}$, and $\widehat{h}^{(k)}(X_i^{(k)})$ equals $\widehat{\boldsymbol{\beta}}_{g_k}^\top \boldsymbol{g}_k(X^{(k)})$ when (2.5) is used. More constraints are involved in (5.2), but estimation only involves one dimensional $X^{(k)}$, $k = 1, ..., q$.

The kernel $\kappa$ we adopted in (2.2), (2.4), and (2.7) 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 Bierens (1987) 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

Fan et al. (1997) 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 (5.2) 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 2.5.

## Supplementary Material

The supplementary material contains an appendix with all technical lemmas and proofs and some additional numerical results.

## Acknowledgements

The authors would like to thank two anonymous referees for helpful comments and suggestions. Jun Shao's research was partially supported by the National Natural Science Foundation of China (11831008) and the U.S. National Science Foundation (DMS-1914411).

## References

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

Breslow, N. E. and R. Holubkov (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*(2), 447–461.

Chatterjee, N., Y.-H. Chen, P. Maas, and R. J. Carroll (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association 111*(513), 107–117.

Chen, Y.-H. and H. Chen (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(3), 449–460.

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

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association 87*(418), 376–382.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing* (2nd ed.). CRC Press.

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

Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics 20*(4), 2008–2036.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*. Routledge.

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

Hall, P. (1992). Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density. *The Annals of Statistics 20*(2), 675 – 694.

Johnson, A. E. W., T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data 3*(1), 160035.

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

Lawless, J., J. Kalbfleisch, and C. Wild (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(2), 413–438.

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

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

Liu, D., M. Görges, and S. A. Jenkins (2012). University of queensland vital signs dataset:

Development of an accessible repository of anesthesia patient monitoring data for research. *Anesthesia & Analgesia 114*(3).

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

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

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

Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis 73*(2), 166–179.

Qin, J., H. Zhang, P. Li, D. Albanes, and K. Yu (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika 102*(1), 169–180.

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

Scott, A. J. and C. J. Wild (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika 84*(1), 57–71.

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

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

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

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*(453), 185–193.

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

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

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

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

Table 1: Simulated MISE (3.1) and IMP (3.2) with $S = 500$ under setting S1

| | | | | | $\widehat{\mu}_{CK}$ (2.7) with constraint (2.5), $\boldsymbol{g} =$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Covariate | Model | Test data | $b, l$ | $\widehat{\mu}_K$ (2.2) | 1 | $(1, X)$ | $(1,\widehat{h})$ | $(1, X, \widehat{h})$ | IMP % |
| Bounded | M1 | Sample | Best | 0.021 | 0.018 | 0.006 | 0.007 | 0.009 | 72.27 |
| | | | CV | 0.030 | 0.026 | 0.014 | 0.015 | 0.018 | 51.41 |
| | | Grid | Best | 0.046 | 0.043 | 0.018 | 0.019 | 0.024 | 61.12 |
| | | | CV | 0.067 | 0.063 | 0.040 | 0.040 | 0.046 | 40.59 |
| | M2 | Sample | Best | 0.046 | 0.037 | 0.036 | 0.033 | 0.029 | 36.30 |
| | | | CV | 0.051 | 0.046 | 0.044 | 0.043 | 0.040 | 22.27 |
| | | Grid | Best | 0.122 | 0.099 | 0.097 | 0.094 | 0.081 | 33.67 |
| | | | CV | 0.134 | 0.123 | 0.122 | 0.125 | 0.110 | 18.16 |
| | M3 | Sample | Best | 0.042 | 0.033 | 0.030 | 0.030 | 0.030 | 29.69 |
| | | | CV | 0.046 | 0.041 | 0.039 | 0.039 | 0.039 | 15.95 |
| | | Grid | Best | 0.101 | 0.088 | 0.086 | 0.088 | 0.081 | 20.20 |
| | | | CV | 0.120 | 0.110 | 0.110 | 0.113 | 0.107 | 10.51 |
| | M4 | Sample | Best | 0.022 | 0.018 | 0.007 | 0.008 | 0.009 | 67.20 |
| | | | CV | 0.030 | 0.027 | 0.016 | 0.015 | 0.018 | 47.53 |
| | | Grid | Best | 0.049 | 0.046 | 0.022 | 0.022 | 0.027 | 54.87 |
| | | | CV | 0.073 | 0.068 | 0.045 | 0.044 | 0.050 | 39.36 |
| Normal | M1 | Sample | Best | 0.067 | 0.060 | 0.050 | 0.049 | 0.062 | 27.57 |
| | | | CV | 0.077 | 0.069 | 0.061 | 0.061 | 0.076 | 21.10 |
| | | Grid | Best | 0.034 | 0.028 | 0.019 | 0.017 | 0.019 | 49.38 |
| | | | CV | 0.035 | 0.031 | 0.025 | 0.023 | 0.026 | 35.66 |
| | M2 | Sample | Best | 0.080 | 0.079 | 0.078 | 0.074 | 0.072 | 10.08 |
| | | | CV | 0.087 | 0.088 | 0.086 | 0.086 | 0.084 | 3.96 |
| | | Grid | Best | 0.053 | 0.053 | 0.052 | 0.051 | 0.049 | 8.10 |
| | | | CV | 0.063 | 0.065 | 0.063 | 0.069 | 0.066 | -0.00 |
| | M3 | Sample | Best | 0.090 | 0.090 | 0.088 | 0.091 | 0.092 | 2.36 |
| | | | CV | 0.099 | 0.098 | 0.097 | 0.102 | 0.102 | 2.05 |
| | | Grid | Best | 0.053 | 0.051 | 0.050 | 0.053 | 0.051 | 6.33 |
| | | | CV | 0.061 | 0.061 | 0.060 | 0.066 | 0.063 | 2.73 |
| | M4 | Sample | Best | 0.072 | 0.068 | 0.058 | 0.056 | 0.063 | 22.64 |
| | | | CV | 0.077 | 0.072 | 0.065 | 0.065 | 0.074 | 15.92 |
| | | Grid | Best | 0.034 | 0.030 | 0.024 | 0.021 | 0.021 | 39.89 |
| | | | CV | 0.036 | 0.034 | 0.029 | 0.026 | 0.028 | 27.44 |

Red indicates the best one among all methods.

Table 2: Simulated MISE(3.1) and IMP (3.2) with $S = 500$ under setting S2

| | | | | | $\widehat{\mu}_{CK}$ (2.7) with constraint | | |
|---|---|---|---|---|---|---|---|
| Covariate | Model | Test data | $b, l$ | $\widehat{\mu}_K$ (2.2) | (2.5) | (2.17) | IMP % |
| Bounded | M1 | Sample | Best | 0.021 | 0.014 | 0.006 | 72.77 |
| | | | CV | 0.028 | 0.015 | 0.015 | 48.49 |
| | | Grid | Best | 0.047 | 0.028 | 0.018 | 61.67 |
| | | | CV | 0.062 | 0.040 | 0.039 | 36.67 |
| | M2 | Sample | Best | 0.046 | 0.041 | 0.035 | 24.33 |
| | | | CV | 0.053 | 0.044 | 0.044 | 17.16 |
| | | Grid | Best | 0.123 | 0.103 | 0.095 | 23.29 |
| | | | CV | 0.136 | 0.123 | 0.124 | 9.23 |
| | M3 | Sample | Best | 0.042 | 0.036 | 0.030 | 27.89 |
| | | | CV | 0.045 | 0.039 | 0.038 | 15.45 |
| | | Grid | Best | 0.099 | 0.091 | 0.085 | 14.38 |
| | | | CV | 0.120 | 0.111 | 0.112 | 7.06 |
| | M4 | Sample | Best | 0.022 | 0.015 | 0.007 | 67.85 |
| | | | CV | 0.030 | 0.015 | 0.015 | 50.65 |
| | | Grid | Best | 0.049 | 0.032 | 0.022 | 54.14 |
| | | | CV | 0.070 | 0.044 | 0.043 | 38.58 |
| Normal | M1 | Sample | Best | 0.069 | 0.057 | 0.050 | 27.07 |
| | | | CV | 0.075 | 0.060 | 0.059 | 21.81 |
| | | Grid | Best | 0.034 | 0.024 | 0.019 | 44.34 |
| | | | CV | 0.035 | 0.025 | 0.024 | 29.56 |
| | M2 | Sample | Best | 0.082 | 0.082 | 0.079 | 3.15 |
| | | | CV | 0.087 | 0.086 | 0.087 | 0.72 |
| | | Grid | Best | 0.056 | 0.057 | 0.053 | 5.73 |
| | | | CV | 0.062 | 0.062 | 0.063 | -0.97 |
| | M3 | Sample | Best | 0.092 | 0.092 | 0.089 | 3.26 |
| | | | CV | 0.101 | 0.10 | 0.100 | 1.31 |
| | | Grid | Best | 0.054 | 0.054 | 0.050 | 7.34 |
| | | | CV | 0.061 | 0.060 | 0.059 | 3.00 |
| | M4 | Sample | Best | 0.070 | 0.062 | 0.057 | 17.69 |
| | | | CV | 0.079 | 0.068 | 0.067 | 14.96 |
| | | Grid | Best | 0.033 | 0.027 | 0.024 | 27.32 |
| | | | CV | 0.035 | 0.029 | 0.029 | 17.58 |

Red indicates the best one among all methods.

For CK estimator under all constraints, $\boldsymbol{g}(X) = (1, X)$.

Table 3: Simulated coverage probability (CP), length of confidence internals, bias of kernel estimator at some values of $\boldsymbol{u}$ ($S = 1,000$ under setting S1), and values of $\mu(\boldsymbol{u})$ and under smoothing scales $c_b$ and $c_l$.

| Covariate | Model | | $\widehat{\mu}_K$ (2.2) | $\widehat{\mu}_{CK}$ (2.7) | $\widehat{\mu}_K$ (2.2) | $\widehat{\mu}_{CK}$ (2.7) | $\widehat{\mu}_K$ (2.2) | $\widehat{\mu}_{CK}$ (2.7) |
|---|---|---|---|---|---|---|---|---|
| Bounded | M1 | CP | 0.94 | 0.95 | 0.95 | 0.95 | 0.94 | 0.96 |
| | | length | 0.94 | 0.38 | 0.81 | 0.42 | 0.94 | 0.37 |
| | | bias | 0.02 | -0.01 | -0.01 | -0.02 | -0.02 | 0.00 |
| | | $c_b$ | 0.30 | 0.50 | 0.30 | 0.80 | 0.30 | 0.50 |
| | | $c_l$ | | 1.00 | | 0.30 | | 1.00 |
| | | $\mu(\boldsymbol{u})$ | $\mu(-0.5,-0.5) = -0.31$ | | $\mu(0,0) = 0$ | | $\mu(0.5,0.5) = 0.19$ | |
| | M2 | CP | 0.95 | 0.95 | 0.94 | 0.95 | 0.93 | 0.95 |
| | | length | 0.86 | 0.58 | 0.75 | 0.63 | 0.86 | 0.52 |
| | | bias | 0.03 | 0.04 | -0.03 | -0.04 | -0.00 | -0.04 |
| | | $c_b$ | 0.50 | 0.80 | 0.50 | 0.30 | 0.50 | 0.80 |
| | | $c_l$ | | 0.80 | | 0.80 | | 1.00 |
| | | $\mu(\boldsymbol{u})$ | $\mu(-0.5,-0.5) = -0.21$ | | $\mu(0,0) = 0.5$ | | $\mu(0.5,0.5) = 0.75$ | |
| | M3 | CP | 0.94 | 0.95 | 0.94 | 0.95 | 0.95 | 0.95 |
| | | length | 0.82 | 0.60 | 0.70 | 0.52 | 1.17 | 0.62 |
| | | bias | 0.02 | 0.03 | -0.01 | -0.01 | -0.00 | -0.02 |
| | | $c_b$ | 0.50 | 1.00 | 0.50 | 0.30 | 0.30 | 0.10 |
| | | $c_l$ | | 0.30 | | 1.00 | | 1.00 |
| | | $\mu(\boldsymbol{u})$ | $\mu(-0.5,-0.5) = -0.04$ | | $\mu(0,0) = 0.5$ | | $\mu(0.5,0.5) = 0.92$ | |
| | M4 | CP | 0.95 | 0.96 | 0.94 | 0.95 | 0.95 | 0.95 |
| | | length | 0.93 | 0.38 | 0.82 | 0.43 | 0.93 | 0.48 |
| | | bias | 0.01 | -0.01 | -0.01 | -0.02 | -0.03 | -0.04 |
| | | $c_b$ | 0.30 | 0.50 | 0.30 | 0.80 | 0.30 | 0.80 |
| | | $c_l$ | | 1.00 | | 0.30 | | 0.30 |
| | | $\mu(\boldsymbol{u})$ | $\mu(-0.5,-0.5) = -0.07$ | | $\mu(0,0) = 0.25$ | | $\mu(0.5,0.5) = 0.43$ | |

Table 3: continued.

| Covariate | Model | | $\widehat{\mu}_K$ (2.2) | $\widehat{\mu}_{CK}$ (2.7) | $\widehat{\mu}_K$ (2.2) | $\widehat{\mu}_{CK}$ (2.7) | $\widehat{\mu}_K$ (2.2) | $\widehat{\mu}_{CK}$ (2.7) |
|---|---|---|---|---|---|---|---|---|
| Normal | M1 | CP | 0.91 | 0.91 | 0.90 | 0.92 | 0.91 | 0.93 |
| | | length | 1.59 | 1.06 | 0.66 | 0.52 | 0.62 | 0.50 |
| | | bias | 0.11 | 0.15 | -0.01 | -0.03 | -0.04 | -0.03 |
| | | $c_b$ | 0.50 | 0.30 | 0.50 | 0.80 | 0.80 | 0.80 |
| | | $c_l$ | | 1.00 | | 0.30 | | 1.00 |
| | | $\mu(\boldsymbol{u})$ | $\mu(-1,1) = -0.75$ | | $\mu(0,0) = 0$ | | $\mu(1,1) = 0.25$ | |
| | M2 | CP | 0.95 | 0.94 | 0.87 | 0.85 | 0.91 | 0.93 |
| | | length | 1.03 | 1.04 | 0.73 | 0.67 | 0.59 | 0.59 |
| | | bias | 0.01 | -0.01 | -0.05 | -0.06 | -0.00 | -0.02 |
| | | $c_b$ | 1.00 | 1.00 | 0.50 | 0.50 | 1.00 | 0.80 |
| | | $c_l$ | | 0.80 | | 0.50 | | 1.00 |
| | | $\mu(\boldsymbol{u})$ | $\mu(-1,1) = 0.63$ | | $\mu(0,0) = 0.5$ | | $\mu(1,1) = 0.63$ | |
| | M3 | CP | 0.91 | 0.91 | 0.89 | 0.91 | 0.89 | 0.89 |
| | | length | 1.03 | 0.96 | 0.72 | 0.58 | 0.98 | 0.68 |
| | | bias | 0.18 | 0.13 | -0.00 | -0.01 | 0.05 | 0.08 |
| | | $c_b$ | 1.00 | 1.00 | 1.00 | 0.80 | 0.50 | 0.30 |
| | | $c_l$ | | 0.50 | | 0.30 | | 1.00 |
| | | $\mu(\boldsymbol{u})$ | $\mu(-1,1) = 0.63$ | | $\mu(0,0) = 0.5$ | | $\mu(1,1) = 0.63$ | |
| | M4 | CP | 0.91 | 0.90 | 0.89 | 0.91 | 0.90 | 0.94 |
| | | length | 1.69 | 1.32 | 0.69 | 0.54 | 0.66 | 0.53 |
| | | bias | 0.15 | 0.20 | -0.02 | -0.03 | -0.02 | -0.02 |
| | | $c_b$ | 0.50 | 0.80 | 0.50 | 0.80 | 0.80 | 1.00 |
| | | $c_l$ | | 0.30 | | 0.30 | | 0.80 |
| | | $\mu(\boldsymbol{u})$ | $\mu(-1,1) = -0.62$ | | $\mu(0,0) = 0.25$ | | $\mu(1,1) = 0.39$ | |

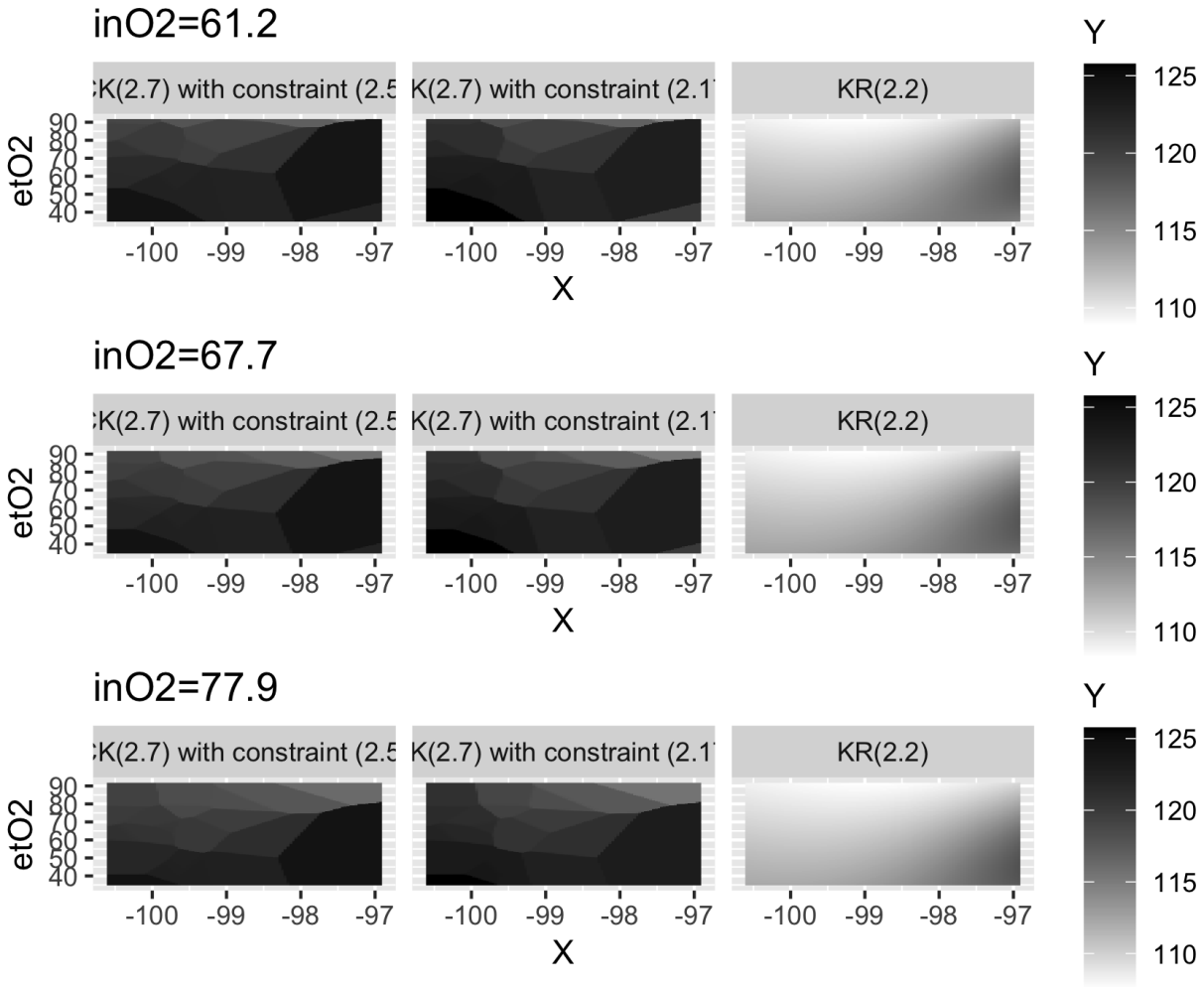


Figure 1: Plot of the fitted kernel regression of systolic blood pressure ($Y$) to etO2 and $X$, given inO2 equals to its first, second, and third quartiles.

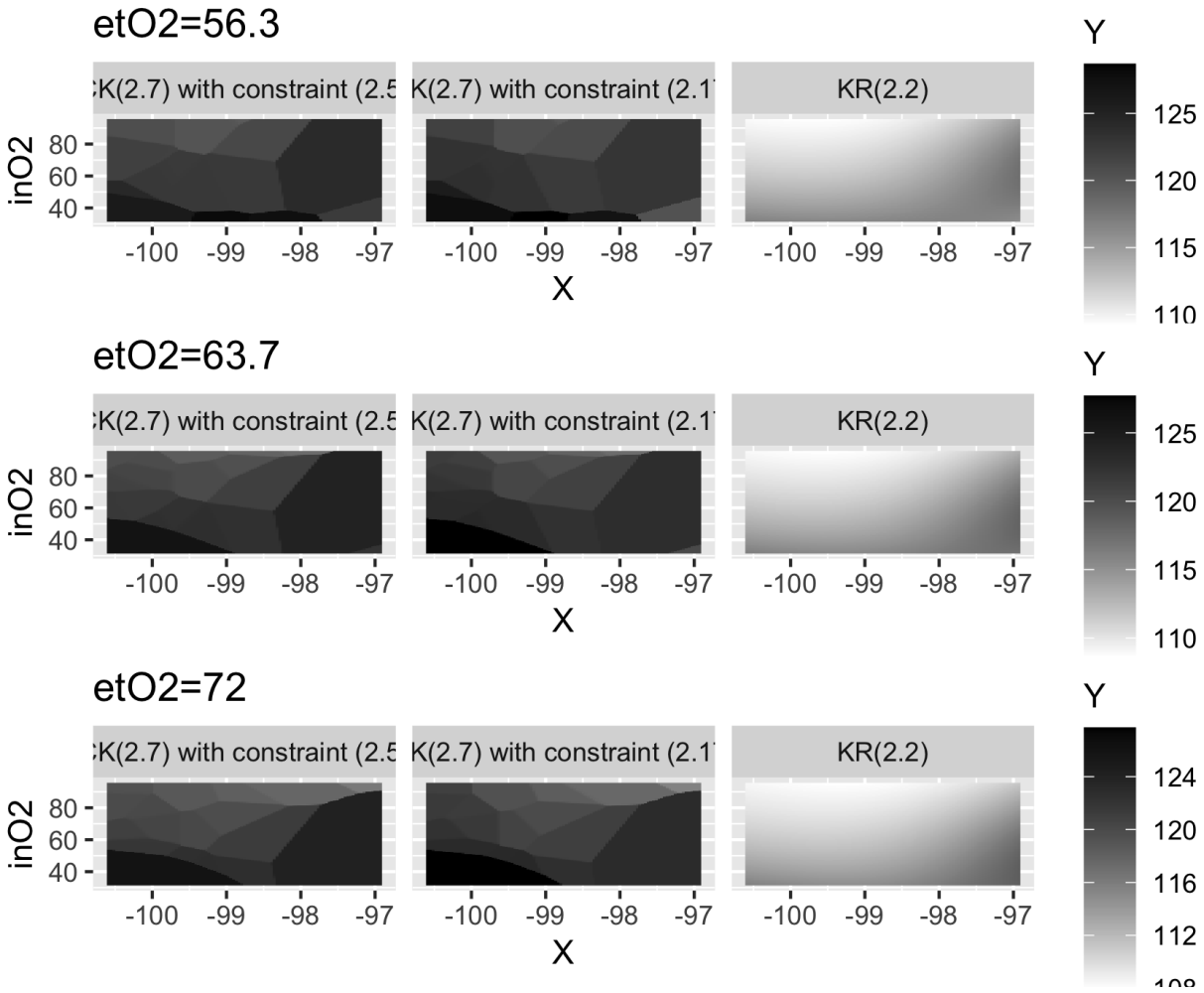


Figure 2: Plot of the fitted kernel regression of systolic blood pressure ($Y$) to inO2 and $X$, given etO2 equals to its first, second, and third quartiles.

Table 4: 95% confidence intervals of systolic blood pressure under selected covariate points with under smoothing scale $c_b = 0.8$, $c_l = 1$.

| Covariate value | | | | 95% confidence interval | | |
|---|---|---|---|---|---|---|
| $X$ | inO2 | etO2 | Method | lower | upper | length |
| -99.5 | 61.2 | 56.3 | $\widehat{\mu}_{CK}$ (2.7) | 121.12 | 124.40 | 3.28 |
| | | | $\widehat{\mu}_{CK}$ (2.17) | 121.68 | 125.00 | 3.32 |
| | | | $\widehat{\mu}_{K}$ (2.2) | 109.88 | 113.97 | 4.08 |
| -99.0 | 67.7 | 63.7 | $\widehat{\mu}_{CK}$ (2.7) | 116.67 | 123.68 | 7.01 |
| | | | $\widehat{\mu}_{CK}$ (2.17) | 117.05 | 124.13 | 7.08 |
| | | | $\widehat{\mu}_{K}$ (2.2) | 106.08 | 114.80 | 8.72 |
| -99.5 | 77.9 | 72.0 | $\widehat{\mu}_{CK}$ (2.7) | 118.09 | 122.26 | 4.17 |
| | | | $\widehat{\mu}_{CK}$ (2.17) | 118.48 | 122.70 | 4.22 |
| | | | $\widehat{\mu}_{K}$ (2.2) | 105.81 | 111.00 | 5.19 |