Annals of the Institute of Statistical Mathematics
https://doi.org/10.1007/s10463-023-00872-8

Statistical inference using regularized M-estimation in the reproducing kernel Hilbert space for handling missing data

Hengfang Wang[1] · **Jae Kwang Kim**[2]

Received: 11 November 2022 / Revised: 5 February 2023 / Accepted: 20 February 2023

Abstract
Imputation is a popular technique for handling missing data. We address a nonparametric imputation using the regularized M-estimation techniques in the reproducing kernel Hilbert space. Specifically, we first use kernel ridge regression to develop imputation for handling item nonresponse. Although this nonparametric approach is potentially promising for imputation, its statistical properties are not investigated in the literature. Under some conditions on the order of the tuning parameter, we first establish the root-n consistency of the kernel ridge regression imputation estimator and show that it achieves the lower bound of the semiparametric asymptotic variance. A nonparametric propensity score estimator using the reproducing kernel Hilbert space is also developed by the linear expression of the projection estimator. We show that the resulting propensity score estimator is asymptotically equivalent to the kernel ridge regression imputation estimator. Results from a limited simulation study are also presented to confirm our theory. The proposed method is applied to analyze air pollution data measured in Beijing, China.

Keywords Imputation · Kernel ridge regression · Missing at random · Propensity score

1 Introduction

Missing data is a universal problem in statistics. Ignoring cases with missing values can lead to misleading results (Kim and Shao 2021; Little and Rubin 2019). Two popular approaches for handling missing data are imputation and propensity

✉ Jae Kwang Kim
jkim@iastate.edu

[1] School of Mathematics and Statistics, Fujian Normal University, No. 8 Xuefu South Road, Shangjie, Fuzhou 350117, Fujian, China

[2] Department of Statistics, Iowa State University, 2438 Osborn Dr., Ames, IA 50011, USA

Published online: 27 April 2023

score weighting. Both approaches are based on some assumptions about the data structure and the response mechanism. To avoid potential biases due to model misspecification, instead of using strong parametric model assumptions, nonparametric approaches are preferred as they do not depend on explicit model assumptions.

In principle, any prediction technique can be used to impute missing values using the responding units as a training sample. However, statistical inference with an imputed estimator is not straightforward. Treating the imputed data as if observed and applying the standard estimation procedure may result in misleading inference, leading to an underestimation of the variance of the imputed point estimators. How to incorporate the uncertainty of the estimated parameters in the final inference is challenging, especially for nonparametric imputation because the model parameter is implicitly defined.

For nonparametric imputation, Cheng (1994) used kernel-based nonparametric regression for imputation and established the root-n consistency of the imputed estimator. Chen and Shao (2001) considered nearest neighbor imputation and discuss its variance estimation. Wang and Chen (2009) employed the kernel smoothing approach to make empirical likelihood inference with missing values. Yang and Kim (2020) considered predictive mean matching for imputation and established its asymptotic properties. Sang et al. (2022) proposed semiparametric fractional imputation using Gaussian mixtures.

For nonparametric propensity score estimation, Hainmueller (2012) proposed so-called the entropy balancing method to find the propensity score weights using the Kullback-Leibler information criterion with a finite dimensional basis function. Chen et al. (2013) established the root-n consistency of the kernel-based nonparametric propensity score estimator. Chan et al. (2016) generalized the entropy balancing method of Hainmueller (2012) further to develop a general calibration weighting method that satisfies the covariance balancing property with increasing dimensions of the control variables. They further showed the global efficiency of the proposed calibration weighting estimator. Zhao (2019) generalized the idea further and developed a unified approach of covariate balancing propensity score method. Tan (2020) developed regularized calibrated estimation of propensity scores with high-dimensional covariates. Although nonparametric kernel regression can be used to construct a nonparametric propensity score estimation, as in Chen et al. (2013), it is not clear how to generalize it to a wider function space to obtain a nonparametric propensity score estimation.

In this paper, we consider regularized M-estimation as a tool for nonparametric imputation and also for the nonparametric propensity score function. The kernel ridge regression Hastie et al. (2009); Shawe-Taylor and Cristianini (2004) is an example of regularized M-estimation for a modern regression technique. By using a regularized M-estimator in reproducing kernel Hilbert space (RKHS), kernel ridge regression can estimate the regression function with the complex reproducing Hilbert kernel space while a regularized term makes the original infinite-dimensional estimation problem viable (Wahba 1990). Due to its flexibility in the choice of kernel functions, kernel ridge regression is very popular in machine learning. van de Geer (2000); Mendelson (2002); Zhang (2005);

Koltchinskii (2006); Steinwart et al. (2009); Zhang and Simon (2023) studied the error bounds for the estimates of the kernel ridge regression method.

While the kernel ridge regression (KRR) is a promising tool for handling missing data, its statistical inference is not investigated in the literature. We aim to fill this important research gap in the missing data literature by establishing the statistical properties of the KRR imputation estimator. Specifically, we obtain root-n consistency of the KRR imputation estimator in some popular functional Hilbert spaces.

Because the KRR is a general tool for nonparametric regression with flexible assumptions, the proposed imputation method can be widely used to handle missing data without employing parametric model assumptions. Variance estimation after the KRR imputation is a challenging but important problem. To our knowledge, this is the first paper to consider the kernel ridge regression technique for imputation and to discuss its variance estimation rigorously.

The regularized M-estimation technique in RKHS is also used to obtain a nonparametric propensity score function to handle missing data. By utilizing the linear smoother form of the KRR imputation, we can easily find the propensity score weights for the responding units. The resulting propensity score estimator is equivalent to projection estimation using kernel ridge regression. The propensity score weights approximately satisfy the model calibration. The resulting estimator achieves optimality in the sense of Robins (1994) and the propensity weights are constructed from the same kernel functions for nonparametric imputation. The propensity weights can also be used to estimate the influence function for variance estimation.

The paper is organized as follows. In Sect. 2, the basic setup and the KRR method are introduced. In Sect. 3, the root-n consistency of the KRR imputation estimator is established. The propensity score estimation is discussed in Sect. 4. Results from a limited simulation study are presented in Sect. 5. An illustration of the proposed method to a real data example is presented in Sect. 6. Some concluding remarks are made in Sect. 7.

2 Basic setup

Consider the problem of estimating $\theta = E(Y)$ from an independent and identically distributed sample $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ of random vector $(\boldsymbol{X}, Y)$, where Y is a real-valued random variable and $\boldsymbol{X}$ is a d-dimensional random variable that serves as auxiliary information. Instead of always observing y_i, suppose that we observe y_i only if $\delta_i = 1$, where δ_i is the response indicator function of the unit i taking values in $\{0, 1\}$. The auxiliary variables $\boldsymbol{x}_i$ are always observed. We assume that the response mechanism is missing at random (MAR) in the sense of Rubin (1976). Specifically, given the auxiliary information $\boldsymbol{X}$, the response variable Y and the missing indicator variable δ are conditionally independent, that is, $Y \perp \delta \mid \boldsymbol{X}$.

Under MAR, we can develop a nonparametric estimator $\widehat{m}(\boldsymbol{x})$ of $m(\boldsymbol{x}) = E(Y \mid \boldsymbol{x})$ and construct the following imputation estimator for θ:

$$\widehat{\theta}_I = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\widehat{m}(\boldsymbol{x}_i)\right\}. \tag{1}$$

If $\widehat{m}(\boldsymbol{x})$ is constructed using the kernel-based nonparametric regression method, we can express

$$\widehat{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n}\delta_i K_h(\boldsymbol{x}_i,\boldsymbol{x})y_i}{\sum_{i=1}^{n}\delta_i K_h(\boldsymbol{x}_i,\boldsymbol{x})} \tag{2}$$

where $K_h(\cdot,\cdot)$ is the kernel function with bandwidth h. Specifically, $K_h(\boldsymbol{x}_i,\boldsymbol{x}) = K(\boldsymbol{x}_i/h,\boldsymbol{x})/h$. Under some suitable choice of bandwidth h, Cheng (1994) first established the root-n consistency of the imputation estimator (1) with nonparametric function in (2). However, kernel-based regression imputation in (2) is applicable only when the dimension of $\boldsymbol{x}$ is small.

In this paper, we extend the work of Cheng (1994) by considering a more general type of nonparametric imputation, called kernel ridge regression imputation. The kernel ridge regression (KRR) can be understood using the reproducing kernel Hilbert space theory (Aronszajn 1950) and can be described as

$$\widehat{m} = \arg\min_{m\in\mathcal{H}}\left[\sum_{i=1}^{n}\delta_i\left\{y_i - m(\boldsymbol{x}_i)\right\}^2 + \lambda\|m\|_{\mathcal{H}}^2\right], \tag{3}$$

where $\|m\|_{\mathcal{H}}^2$ is the norm of m in the reproducing kernel Hilbert space $\mathcal{H}$ and $\lambda(>0)$ is a tuning parameter for regularization. Here, the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is induced by a kernel function, i.e.,

$$\langle f, K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}} = f(\boldsymbol{x}),$$

for any $\boldsymbol{x}\in\mathcal{X}\subset\mathbb{R}^d, f\in\mathcal{H}$, namely, the reproducing property of $\mathcal{H}$. Naturally, this reproducing property implies the $\mathcal{H}$ norm of f: $\|f\|_{\mathcal{H}} = \langle f,f\rangle_{\mathcal{H}}^{1/2}$. Schölkopf et al. (2002) provides a comprehensive overview of machine learning techniques using reproducing kernel functions.

One canonical example of such a functional Hilbert space is the Sobolev space. Specifically, assuming that the domain of such functional space is [0, 1], the Sobolev space of order ℓ can be denoted as

$$\mathcal{W}_2^{\ell} = \left\{f : [0,1]\to\mathbb{R} \mid f, f^{(1)},\ldots,f^{(\ell-1)}\in\mathbb{C}[0,1],\quad f^{(\ell)}\in L^2[0,1]\right\},$$

where $\mathbb{C}[0,1]$ denotes the absolutely continuous function on [0, 1]. One possible norm for this space can be

$$\|f\|_{\mathcal{W}_2^{\ell}}^2 = \sum_{q=0}^{\ell-1}\left\{\int_0^1 f^{(q)}(t)dt\right\}^2 + \int_0^1\left\{f^{(\ell)}(t)\right\}^2 dt.$$

In this section, we employ the Sobolev space of second order as the approximation function space. For a Sobolev space of order ℓ, we have the kernel function.

$$K(x,y) = \sum_{q=0}^{\ell-1} k_q(x)k_q(y) + k_\ell(x)k_\ell(y) + (-1)^\ell k_{2\ell}(|x-y|),$$

where $k_q(x) = (q!)^{-1}B_q(x)$ and $B_q(\cdot)$ is the Bernoulli polynomial of order q. The smoothing spline method is a special case of the kernel ridge regression method.

By the representer theorem for reproducing kernel Hilbert space (Wahba 1990), the estimate in (3) lies in the linear span of $\{K(\cdot,\boldsymbol{x}_i), i = 1,\ldots,n\}$. Specifically, we have

$$\widehat{m}(\cdot) = \sum_{i=1}^{r} \widehat{\alpha}_{i,\lambda} K(\cdot,\boldsymbol{x}_i), \tag{4}$$

where

$$\widehat{\boldsymbol{\alpha}}_\lambda = \left(\boldsymbol{\Delta}_n \mathbf{K} + \lambda \boldsymbol{I}_n\right)^{-1} \boldsymbol{y},$$

where $\boldsymbol{\Delta}_n = \text{ diag } (\delta_1,\ldots,\delta_n)$, $\mathbf{K} = (K(\boldsymbol{x}_i,\boldsymbol{x}_j))_{ij} \in \mathbb{R}^{n\times n}$, $\boldsymbol{y} = (y_1,\ldots,y_n)^{\mathrm{T}}$ and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

The tuning parameter λ is selected through generalized cross-validation in kernel ridge regression, where the criterion for λ is

$$\text{GCV } (\lambda) = \frac{n^{-1}\left\|\left\{\boldsymbol{I}_n - \boldsymbol{A}(\lambda)\right\}\boldsymbol{y}\right\|_2^2}{n^{-1}\text{tr}(\boldsymbol{I}_n - \boldsymbol{A}(\lambda))}, \tag{5}$$

and $\boldsymbol{A}(\lambda) = \boldsymbol{\Delta}_n\mathbf{K}(\boldsymbol{\Delta}_n\mathbf{K} + \lambda\boldsymbol{I}_n)^{-1}\boldsymbol{\Delta}_n$. The value of λ that minimizes the criterion (5) is used to select the tuning parameter.

Using the kernel ridge regression imputation in (3), we can obtain the imputed estimator in (1). Because $\widehat{m}(\boldsymbol{x})$ in (4) is a nonparametric regression estimator of $m(\boldsymbol{x}) = E(Y \mid \boldsymbol{x})$, we can expect that this imputation estimator in (1) is consistent for $\theta = E(Y)$ under missing at random, as long as $\widehat{m}(\boldsymbol{x})$ is a consistent estimator of $m(\boldsymbol{x})$. Surprisingly, it turns out that the consistency of $\widehat{\theta}_I$ to θ is of order $O_p(n^{-1/2})$, while the pointwise convergence rate for $\widehat{m}(\boldsymbol{x})$ to $m(\boldsymbol{x})$ is slower. This phenomenon is consistent with the result of Cheng (1994) for kernel-based nonparametric regression imputation.

We aim to establish two goals: (i) find sufficient conditions for the root-n consistency of the KRR imputation estimator and give a formal proof; (ii) find a linearization variance formula for the KRR imputation estimator. The first part is presented formally in Theorem 1 in Sect. 3. For the second part, we employ a simple algebra to obtain a consistent estimator of $\omega(\boldsymbol{x}) = \{\pi(\boldsymbol{x})\}^{-1}$ in the linearized version of $\widehat{\theta}_I$, where $\pi(\boldsymbol{x}) = E(\delta \mid \boldsymbol{x})$. The estimation of $\omega(\boldsymbol{x})$ will be presented in Sect. 4.

3 Main theory

Before we develop our main theory, we first introduce Mercer's Theorem.

Lemma 1 (Mercer's theorem) *Given a continuous, symmetric, positive definite kernel function* $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. *For* $\boldsymbol{x}, z \in \mathcal{X}$, *under some regularity conditions, Mercer's theorem characterizes K by the following expansion*

$$K(\boldsymbol{x}, z) = \sum_{j=1}^{\infty} \lambda_j \psi_j(\boldsymbol{x}) \psi_j(z),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ *is a non-negative sequence of eigenvalues,* $\{\psi_j\}_{j=1}^{\infty}$ *is an orthonormal basis (eigenfunctions) for* $L^2(\mathbb{P})$ *and* $\mathbb{P}$ *is the given distribution of* $\boldsymbol{X}$ *on* $\mathcal{X}$. *The eigenvalues and the eigenfunctions satisfy*

$$\lambda_j \psi_j(\boldsymbol{x}) = \int_{\mathcal{X}} K(\boldsymbol{x}, z) \psi_j(z) \mathbb{P}(dz), \quad \text{for } j = 1, 2, \ldots.$$

Furthermore, we make the following assumptions.

[A1] For some $k \geq 2$, there is a constant $\rho < \infty$ such that $E[\psi_j(\boldsymbol{X})^{2k}] \leq \rho^{2k}$ for all $j \in \mathbb{N}$, where $\{\psi_j\}_{j=1}^{\infty}$ are orthonormal basis by Karhunen-Loève expansion of Mercer's theorem.

[A2] The function $m \in \mathcal{H}$, and for $\boldsymbol{x} \in \mathcal{X}$, we have $E[\{Y - m(\boldsymbol{x})\}^2 \mid \boldsymbol{x}] \leq \sigma^2$, for some $\sigma^2 < \infty$.

[A3] The response mechanism is missing at random. Furthermore, the propensity score $\pi(\boldsymbol{x}) = P(\delta = 1 \mid \boldsymbol{x})$ is uniformly bounded away from zero. In particular, there exists a positive constant $c > 0$ such that $\pi(\boldsymbol{x}_i) \geq c$, for $i = 1, \ldots, n$.

The first assumption is a technical assumption that controls the tail behavior of $\{\psi_j\}_{j=1}^{\infty}$. Assumption 3 indicates that the noise has a bounded variance. Assumption 3 and Assumption 3 together aim to control the error bound of the kernel ridge regression estimate $\widehat{m}$. Furthermore, Assumption 3 means that the support for the respondents should be the same as the support of the original sample. Assumption 3 is a standard assumption for missing data analysis.

We further introduce the following lemma. Let $\boldsymbol{S}_{\lambda} = (\boldsymbol{I}_n + \lambda \mathbf{K}^{-1})^{-1}$ be the linear smoother for the KRR method. That is, $\widehat{m} = \boldsymbol{S}_{\lambda} \boldsymbol{\Delta}_n \boldsymbol{y}$ be the best predictor of $\boldsymbol{y}$ using the kernel ridge regression method, where $\boldsymbol{\Delta}_n = \text{diag}\,(\delta_1, \ldots, \delta_n)$. We now present the following lemma without proof, which is modified from Lemma 7 in Zhang et al. (2013).

Lemma 2 *Under* [A1]-[A2], *for a random vector* $\boldsymbol{z} = E(\boldsymbol{z}) + \sigma \boldsymbol{\varepsilon}$, *we have*

$$\boldsymbol{S}_{\lambda} \boldsymbol{z} = E(\boldsymbol{z} \mid \boldsymbol{x}) + \boldsymbol{a}_n,$$

where $\boldsymbol{a}_n = (a_1, \ldots, a_n)^{\mathrm{T}}$ *and*

$$a_i = \mathcal{O}_p\big(\lambda^{1/2} + \{\gamma(\lambda)\}^{1/2} n^{-1/2}\big), \tag{6}$$

for $i = 1, \ldots, n$, *as long as* $E(\|z_i\|_{\mathcal{H}})$ *and* σ^2 *is bounded from above, for* $i = 1, \ldots, n$, *where* $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ *are noise vector with mean zero and bounded variance and*

$$\gamma(\lambda) = \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda},$$

is the effective dimension and $\{\mu_j\}_{j=1}^{\infty}$ *are the eigenvalues of the kernel* K *used in* $\widehat{m}(\boldsymbol{x})$.

The first term in (6) denotes the order of the bias term, and the second term denotes the square root of the variance term. Specifically, we have the asymptotic mean square error for $\widehat{m}$,

$$\text{AMSE}\,(\widehat{m}) = O(1) \times \left\{\lambda \|m\|_{\mathcal{H}}^2 + n^{-1}\gamma(\lambda)\right\}. \tag{7}$$

For the ℓ-th order of Sobolev space, we have $\mu_j \leq Cj^{-2\ell}$ and

$$\gamma(\lambda) = \sum_{j=1}^{\infty}(1 + j^{2\ell}\lambda)^{-1} \leq O\left(\lambda^{-1/(2\ell)}\right). \tag{8}$$

Note that (7) is minimized when $\lambda \asymp \gamma(\lambda)/n$, which is equivalent to $\lambda \asymp n^{-2\ell/(2\ell+1)}$ under (8). The optimal rate $\lambda \asymp n^{-2\ell/(2\ell+1)}$ leads to

$$\text{AMSE}\,(\widehat{m}) = O(n^{-2\ell/(2\ell+1)}) \tag{9}$$

which is the optimal rate in Sobolev space, as discussed by Stone (1982).

To investigate the asymptotic properties of the kernel ridge regression imputation estimator, we express

$$\begin{aligned}\widehat{\theta}_I &= \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\widehat{m}(\boldsymbol{x}_i)\right\} \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{x}_i)}_{R_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{y_i - m(\boldsymbol{x}_i)\right\}}_{S_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\left\{\widehat{m}(\boldsymbol{x}_i) - m(\boldsymbol{x}_i)\right\}}_{T_n}.\end{aligned}$$

Therefore, as long as we can show

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{\frac{1}{\pi(\boldsymbol{x}_i)} - 1\right\}\left\{y_i - m(\boldsymbol{x}_i)\right\} + o_p(n^{-1/2}), \tag{10}$$

then we can establish the root-n consistency. The following theorem formally states the theoretical result. A proof of Theorem 1 is presented in the supplementary material.

Theorem 1 *Suppose Assumption* [A1]-[A3] *hold for a Sobolev kernel of order* ℓ, *as long as*

$$n\lambda \to 0, \quad n\lambda^{1/2\ell} \to \infty, \tag{11}$$

we have

$$n^{1/2}\left(\widehat{\theta}_I - \theta\right) \to N(0, \sigma^2),$$

where

$$\sigma^2 = \mathrm{Var}\{E(Y \mid \boldsymbol{x})\} + E\{\mathrm{Var}(Y \mid \boldsymbol{x})/\pi(\boldsymbol{x})\} = \mathrm{Var}(\eta)$$

with

$$\eta = m(\boldsymbol{x}) + \delta \frac{1}{\pi(\boldsymbol{x})}\{y - m(\boldsymbol{x})\}. \tag{12}$$

Remark 1 Note that the optimal rate $\lambda \asymp n^{-2\ell/(2\ell+1)}$ does not satisfy the first part of (11). To control the bias part, we need a smaller λ such as $\lambda = n^{-\kappa}$ with $\kappa > 1$. Similar conditions are used for bandwidth selection for nonparametric kernel regression with bandwidth h:

$$nh \to \infty \text{ and } n^{1/2}h^2 \to 0$$

for $\dim(\boldsymbol{x}) = 1$. See Wang and Chen (2009) for details.

Remark 2 Theorem 1 is presented for a Sobolev kernel, and any kernel whose eigenvalues have the same tail behavior as Sobolev of order ℓ also has the result as Theorem 1. For sub-Gaussian kernel whose eigenvalues satisfy

$$\mu_j \le c_1 \exp(-c_2 j^2),$$

where c_1, c_2 are positive constants, we can establish similar results. To see this, note that

$$\begin{aligned}
\gamma(\lambda) &= \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} \\
&\le c_2^{-1/2}\{-\log(\lambda)\}^{1/2} + \frac{1}{\lambda}\int_{c_2^{-1/2}\{-\log(\lambda)\}^{1/2}} \exp(-c_2 z^2) dz \\
&\le c_2^{-1/2}\{-\log(\lambda)\}^{1/2} + O(1),
\end{aligned}$$

where the second term in the last equation can be obtained by the Gaussian tail bound inequality. Therefore, as long as $n\lambda \to 0$ and $n\{-\log(\lambda)\}^{-1/2} \to \infty$, we have $n^{-1}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{a} = o_p(n^{-1/2})$ and the root-n consistency can be established.

Remark 3 Using $\widehat{m}(\boldsymbol{x}_i)$, we can also construct the projection estimator $\widehat{\theta}_p = n^{-1}\sum_{i=1}^n \widehat{m}(\boldsymbol{x}_i)$. It can be shown that the projection estimator is asymptotically equivalent to the imputation estimator $\widehat{\theta}_I$ in (1) under the assumptions of Theorem 1.

Note that the asymptotic variance of the imputation estimator is equal to $n^{-1}\sigma^2$, which is the lower bound of the semiparametric asymptotic variance discussed in

Robins et al. (1994). Thus, the kernel ridge regression imputation is asymptotically optimal. The main term (12) in the linearization in Theorem 1 is called the influence function (Hampel 1974). The term influence function is motivated by the fact that to the first order $\eta_i = m(\boldsymbol{x}_i) + \delta_i\{\pi(\boldsymbol{x}_i)\}^{-1}\{y_i - m(\boldsymbol{x}_i)\}$ is the influence of a single observation on the estimator $\widehat{\theta}_I$.

The influence function in (12) can be used for variance estimation of the KRR imputation estimator $\widehat{\theta}_I$. The idea is to estimate the influence function $\eta_i = m(\boldsymbol{x}_i) + \delta_i\{\pi(\boldsymbol{x}_i)\}^{-1}\{y_i - m(\boldsymbol{x}_i)\}$ and apply the standard variance estimator using $\widehat{\eta}_i$. To estimate η_i, we need an estimator of $\{\pi(\boldsymbol{x})\}^{-1}$, or namely, $\omega(\boldsymbol{x}_i)$. Once $\widehat{\omega}(\boldsymbol{x})$ is constructed, we can use

$$\widehat{\mathrm{V}} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\widehat{\eta}_i - \bar{\eta}_n\right)^2$$

as a variance estimator of $\widehat{\theta}_I$ in (1), where

$$\widehat{\eta}_i = \widehat{m}(\boldsymbol{x}_i) + \delta_i\widehat{\omega}_i\{y_i - \widehat{m}(\boldsymbol{x}_i)\},$$

and $\bar{\eta}_n = n^{-1}\sum_{i=1}^{n}\widehat{\eta}_i$. How to estimate the propensity score function $\omega(\boldsymbol{x}) = \{\pi(\boldsymbol{x})\}^{-1}$ will be discussed in the next section.

4 Propensity score estimation

We now consider the estimation of the propensity weight function $\omega(\boldsymbol{x}) = \{\pi(\boldsymbol{x})\}^{-1}$ using kernel ridge regression.

To motivate the proposed propensity weight function, note that $\widehat{m}(x)$ is a linear function of y_i. Let $\boldsymbol{S}_\lambda = (\boldsymbol{I}_n + \lambda\mathbf{K}^{-1})^{-1} = (s_{ij}) \in \mathbb{R}^{n\times n}$. We can express

$$\widehat{m}(\boldsymbol{x}_i) = \sum_{j=1}^{n}\delta_j s_{ij} y_j.$$

Now, the result (10) can be expressed as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left(1-\delta_i\right)\left\{\sum_{j=1}^{n}\delta_j s_{ij} y_j - m(\boldsymbol{x}_i)\right\}\\ =&\frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{\omega(\boldsymbol{x}_i) - 1\right\}\left\{y_i - m(\boldsymbol{x}_i)\right\} + o_p(n^{-1/2}).\end{aligned} \tag{13}$$

We use (13) as a key condition to find $\widehat{\omega}(\boldsymbol{x})$. If $\widehat{\omega}(\boldsymbol{x})$ satisfies the model calibration approximately

$$n^{-1}\sum_{i=1}^{n}\delta_i\widehat{\omega}_i m(\boldsymbol{x}_i) = n^{-1}\sum_{i=1}^{n}m(\boldsymbol{x}_i) + o_p(n^{-1/2}), \tag{14}$$

then (13) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\left(1-\delta_i\right)\sum_{j=1}^{n}\delta_j s_{ij} y_j = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{\omega(\boldsymbol{x}_i)-1\right\}y_i + o_p(n^{-1/2}). \tag{15}$$

Using $n^{-1}\sum_{i=1}^{n}\delta_i\{y_i-\widehat{m}(\boldsymbol{x}_i)\} = o_p(n^{-1/2})$, we can express (15) as

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_j s_{ij} y_j = \frac{1}{n}\sum_{i=1}^{n}\delta_i\omega(\boldsymbol{x}_i)y_i + o_p(n^{-1/2}). \tag{16}$$

Thus, we can construct $\widehat{\omega}(\boldsymbol{x})$ from (16) by ignoring the $o_p(n^{-1/2})$ term. That is, we obtain

$$\widehat{\omega}(\boldsymbol{x}_j) = \sum_{i=1}^{n} s_{ij} \tag{17}$$

as the final propensity score weights where s_{ij} is the (i, j)-th element in the smoothing matrix $\boldsymbol{S}_\lambda = (\boldsymbol{I}_n + \lambda\mathbf{K}^{-1})^{-1}$ of the kernel ridge regression. Thus, the resulting propensity score function uses the same RKHS to construct the nonparametric imputation.

By construction, we have the following.

$$\sum_{i=1}^{n}\delta_i\widehat{\omega}(\boldsymbol{x}_i)y_i = \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i s_{ji} y_i = \sum_{j=1}^{n}\widehat{m}(\boldsymbol{x}_j). \tag{18}$$

Condition (18) is called the self-efficiency condition. Thus, by self-consistency, the PS estimator $\widehat{\theta}_{\mathrm{PS}} = n^{-1}\sum_{i=1}^{n}\delta_i\widehat{\omega}(\boldsymbol{x}_i)y_i$ is equivalent to the projection estimator and, by Remark 3, satisfies

$$\widehat{\theta}_{\mathrm{PS}} = \frac{1}{n}\sum_{i=1}^{n}\left[m(\boldsymbol{x}_i) + \frac{\delta_i}{\pi(\boldsymbol{x}_i)}\left\{y_i - m(\boldsymbol{x}_i)\right\}\right] + o_p(n^{-1/2})$$

under the assumptions in Theorem 1.

It remains to show that the propensity score function satisfies the approximate model calibration in (14). The model calibration was first discussed by Wu and Sitter (2001) when the mean function $m(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{x})$ is known. A sketched proof for (14) is presented in Appendix.

If an explicit form of $m(\boldsymbol{x})$ is known, then we can directly use the function in the calibration constraint, as in Wu and Sitter (2001). Our proposed estimator does not require the knowledge of the mean function. The only requirement is that the mean function lies in the reproducing kernel Hilbert space that the kernel function is generating. Also, the uniform function calibration considered in Wong and Chan (2018) achieves $O_p(n^{-1/2})$ which is higher than our convergence rate in (14). By imposing a function calibration that has nothing to do with Y, the uniform function calibration pays the price.

5 Simulation study

To compare with existing methods and to evaluate the finite-sample performance of the proposed imputation method and its variance estimator, we conducted a limited simulation study. In this simulation, we consider the continuous study variable with three different data-generating models. In the three models, we keep the response rate around 60% and $\mathrm{Var}(Y) \approx 10$. Also, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^{\mathrm{T}}$ are generated independently element-wise from the uniform distribution on the support (1, 3). In model A, we use a linear regression model $y_i = 3 + 2.5x_{i1} + 2.75x_{i2} + 2.5x_{i3} + 2.25x_{i4} + \sigma\epsilon_i$ to obtain y_i, where $\{\epsilon_i\}_{i=1}^n$ are generated from standard normal distribution and $\sigma = 3^{1/2}$. In model B, we use $y_i = 3 + (1/35)x_{i1}^2 x_{i2}^3 x_{i3} + 0.1x_{i4} + \sigma\epsilon_i$ to generate data with a nonlinear structure. The model C for generating the study variable is $y_i = 3 + (1/180)x_{i1}^2 x_{i2}^3 x_{i3} x_{i4}^2 + \sigma\epsilon_i$.

In addition to $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$, we consider two response mechanisms. The response indicator variable δ's for each mechanism are independently generated from different Bernoulli distributions. In the first response mechanism, the probability for the Bernoulli distribution is $\mathrm{logit}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + 2.5)$, where $\boldsymbol{\beta} = (-1.1, 0.5, -0.25, -0.1)^{\mathrm{T}}$ and $\mathrm{logit}\,(p) = \log\{p/(1-p)\}$. In the second response mechanism, the probability for the Bernoulli distribution is $\mathrm{logit}(-0.3 + 0.7x_1^2 - 0.5x_2 - 0.25x_3 - 0.25x_4)$. We consider two sample sizes $n = 500$ and $n = 1,000$.

The reproducing kernel Hilbert space we employed in the simulation study is the second-order Sobolev space. In particular, we used the tensor product RKHS to extend a one-dimensional Sobolev space to the multidimensional space. From each sample, we consider four imputation methods: kernel ridge regression (KRR), and the others are the kernel imputation method (Cheng 1994), B-spline, and linear regression. For the kernel imputation method (KI), we use the Gaussian kernel, and the bandwidth selection method is the expected Kullback–Leibler cross-validation (Hurvich et al. 1998). The kernel imputation method is performed with the aid of 'np' package in R Hayfield and Racine (2008). For the B-spline method, we employ the generalized additive model by R package 'mgcv' (Wood 2017). Specifically, we used a cubic spline with 15 knots for each coordinate with a restricted maximum likelihood estimation method. We used $B = 1,000$ Monte Carlo samples in the simulation study.

The simulation results of the four point estimators for the first response mechanism and for the second response mechanism are summarized in Figs. 1 and 2, respectively. The simulation results in Figs. 1 and 2 show that three methods show similar results under the linear model (model A) except for the KI method. The nonparametric kernel regression using Nadaraya-Watson method is biased when the dimension of X is large. Also, the multi-dimensional bandwidth selection may lead to unstable estimation. Meanwhile, kernel ridge regression imputation estimators show robust performance under nonlinear models (models B and C). In addition, KRR imputation estimations provide negligible biases in all scenarios.

In addition, we have computed the proposed variance estimators under kernel ridge regression imputation with the corresponding kernel. In Table 1, the

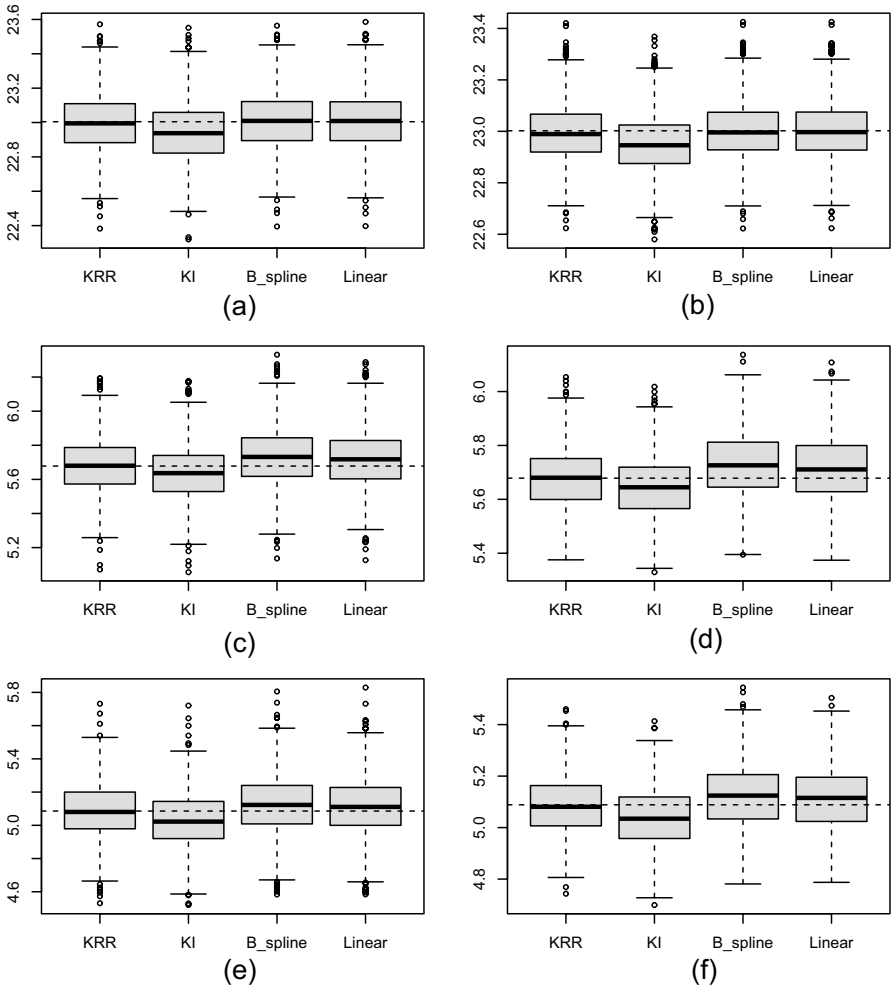

Fig. 1 Boxplots with four estimators for model A ((**a**) for $n = 500$ and (**b**) for $n = 1000$), model B ((**c**) for $n = 500$ and (**d**) for $n = 1000$) and model C ((**e**) for $n = 500$ and (**f**) for $n = 1000$) under first response mechanism with true values (dashes)

relative biases (in percentage) of the proposed variance estimator and the coverage rates of the proposed estimators under the nominal coverage rates 90% and 95% are presented. The relative biases of the variance estimator are relatively low for all scenarios, which confirms the validity of the proposed variance estimator. Furthermore, the interval estimators show good performances in terms of coverage rates.

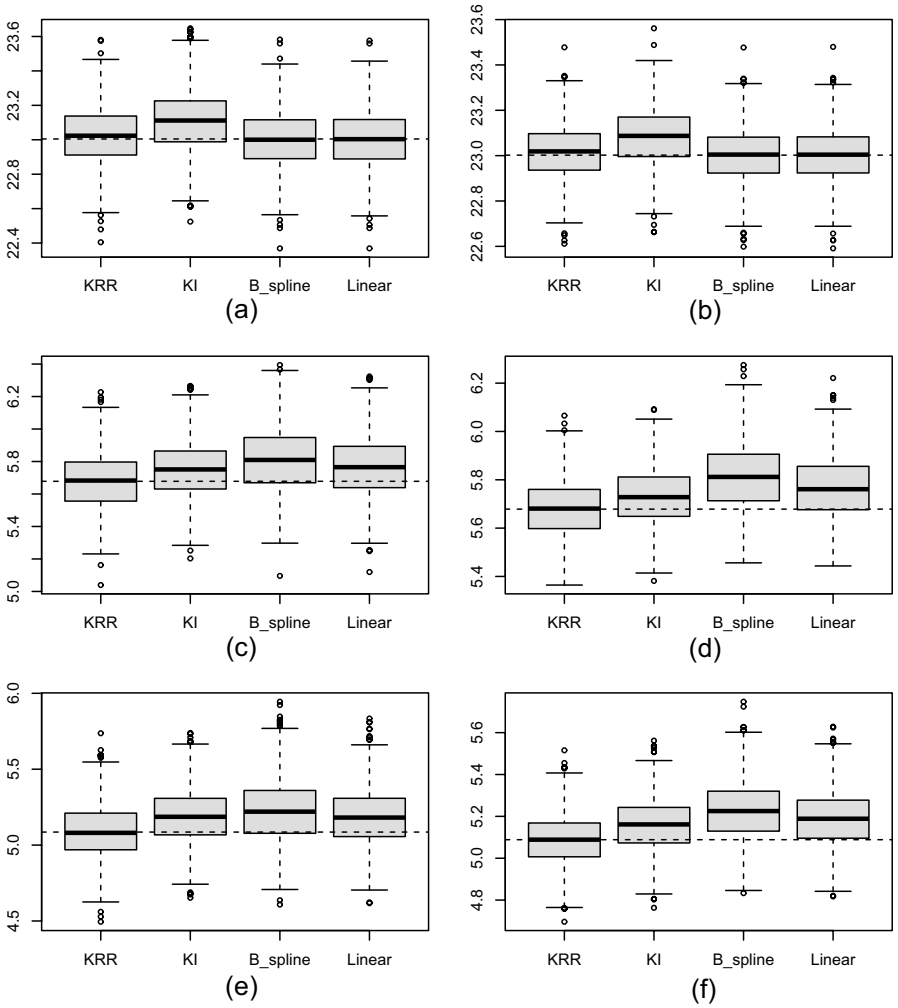

Fig. 2 Boxplots with four estimators for model A ((**a**) for $n = 500$ and (**b**) for $n = 1000$), model B ((**c**) for $n = 500$ and (**d**) for $n = 1000$) and model C ((**e**) for $n = 500$ and (**f**) for $n = 1000$) under second response mechanism with true values (dashes)

6 Application

We applied the kernel ridge regression with the kernel of second-order Sobolev space to study the $PM_{2.5}(\mu g/m^3)$ concentration measured in Beijing, China Liang et al. (2015). Hourly weather conditions: temperature, air pressure, cumulative wind speed, cumulative hours of snow and cumulative hours of rain are available from 2011 to 2015. Meanwhile, the averaged sensor response is subject to missingness. In December 2012, the missing rate of $PM_{2.5}$ is relatively high with missing rate 17.47%. We are interested in estimating the mean $PM_{2.5}$ in December with the

Table 1 Relative biases (R.B.) of the proposed variance estimator, coverage rates (C.R.) of the 90% and 95% confidence intervals for imputed estimators under kernel ridge regression with second-order Sobolev kernel for continuous responses

Model	Criteria	First missing mechanism		Second missing mechanism	
		n=500	n=1000	n=500	n=1000
A	R.B(%)	0.09	−2.80	3.40	2.74
	C.R.(90%)	90.30	89.95	90.25	90.60
	C.R.(95%)	95.50	94.95	95.20	95.45
B	R.B(%)	−2.77	−5.42	−6.07	−3.42
	C.R.(90%)	89.55	89.70	88.05	90.05
	C.R.(95%)	94.25	94.55	94.15	94.70
C	R.B(%)	−7.43	−3.97	−9.38	−2.29
	C.R.(90%)	87.95	88.70	88.80	89.50
	C.R.(95%)	93.35	94.20	93.95	95.15

imputed kernel ridge regression estimate. The point estimates and their 95% confidence intervals are presented in Table 2.

As a benchmark, the confidence interval computed from complete cases and confidence intervals for the imputed estimator under linear model (Kim and Rao 2009) are also presented there.

As we can see, the performances of kernel ridge regression imputation estimators are similar and created narrower 95% confidence intervals. Furthermore, the imputed $PM_{2.5}$ concentration during the missing period is relatively lower than the fully observed weather conditions on average. Therefore, if we only utilize the complete cases to estimate the mean of $PM_{2.5}$, the severeness of air pollution would be over-estimated.

7 Concluding remarks

We consider kernel ridge regression as a tool for nonparametric imputation and propensity score function estimation. The proposed kernel ridge regression imputation can be used as a general tool for nonparametric imputation. By choosing different kernel functions, different nonparametric imputation methods can be developed. Asymptotic properties of the propensity score estimator are also established. The unified theory developed in this paper enables us to make valid nonparametric statistical inferences about the population means under missing data.

Table 2 Point estimates (P.E.), standard error (S.E.) and 95% confidence intervals (C.I.) for imputed mean $PM_{2.5}$ in December, 2012 under kernel ridge regression

Estimator	P.E	S.E	95% C.I
Complete	109.20	3.91	(101.53, 116.87)
Linear	99.61	3.68	(92.39, 106.83)
KRR	101.92	3.50	(95.06, 108.79)

There are several possible extensions of the research. First, the theory can be extended to other nonparametric imputation methods, such as smoothing splines (Claeskens et al. 2009), thin plate spline (Wahba 1990), Gaussian process regression (Rasmussen and Williams 2006), or deep kernel learning (Bohn et al. 2019). The theoretical results in this paper can be used as building-blocks for establishing the statistical properties of these sophisticated nonparametric imputation methods. Second, instead of using ridge-type penalty term, one can also consider other penalty functions such as the smoothly clipped absolute deviation penalty (Fan and Li 2001) or adaptive lasso (Zou 2006). Such penalty functions can be potentially useful for handling high dimensional covariate problems. Also, the proposed method can be used for causal inference, including estimation of average treatment effect from observational studies (Morgan and Winship, 2014; Yang and Ding, 2020). Developing tools for causal inference using the kernel ridge regression-based propensity score method will be an important extension of this research.

Appendix

This Appendix contains the technical proof for Theorem 1 and a sketched proof for (14).

A Proof for Theorem 1

To prove our main theorem, we write

$$\begin{aligned}\widehat{\theta}_I &= \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\widehat{m}(\boldsymbol{x}_i)\right\} \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{x}_i)}_{R_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{y_i - m(\boldsymbol{x}_i)\right\}}_{S_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\left\{\widehat{m}(\boldsymbol{x}_i) - m(\boldsymbol{x}_i)\right\}}_{T_n}.\end{aligned}$$

Therefore, as long as we show

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{\frac{1}{\pi(\boldsymbol{x}_i)} - 1\right\}\left\{y_i - m(\boldsymbol{x}_i)\right\} + o_p(n^{-1/2}),$$

then the main theorem automatically holds.

To show (10), note that

$$\begin{aligned}\widehat{\boldsymbol{m}} &= \mathbf{K}\left(\boldsymbol{\Delta}_n\mathbf{K} + \lambda\boldsymbol{I}_n\right)^{-1}\boldsymbol{\Delta}_n\boldsymbol{y} \\ &= \mathbf{K}\left\{\left(\boldsymbol{\Delta}_n + \lambda\mathbf{K}^{-1}\right)\mathbf{K}\right\}^{-1}\boldsymbol{\Delta}_n\boldsymbol{y} \\ &= \left(\boldsymbol{\Delta}_n + \lambda\mathbf{K}^{-1}\right)^{-1}\boldsymbol{\Delta}_n\boldsymbol{y},\end{aligned}$$

where $\widehat{\boldsymbol{m}} = (\widehat{m}(\boldsymbol{x}_1), \ldots, \widehat{m}(\boldsymbol{x}_n))^{\mathrm{T}}$. Let $\boldsymbol{S}_\lambda = (\boldsymbol{I}_n + \lambda \mathbf{K}^{-1})^{-1}$, we have

$$\widehat{\boldsymbol{m}} = \left(\boldsymbol{\Delta}_n + \lambda \mathbf{K}^{-1}\right)^{-1} \boldsymbol{\Delta}_n \boldsymbol{y} = \boldsymbol{C}_n^{-1} \boldsymbol{d}_n,$$

where

$$\begin{aligned} \boldsymbol{C}_n &= \boldsymbol{S}_\lambda \left(\boldsymbol{\Delta}_n + \lambda \mathbf{K}^{-1}\right), \\ \boldsymbol{d}_n &= \boldsymbol{S}_\lambda \boldsymbol{\Delta}_n \boldsymbol{y}. \end{aligned}$$

By Lemma 2, we obtain

$$\begin{aligned} \boldsymbol{C}_n &= E(\boldsymbol{\Delta}_n \mid \boldsymbol{x}) + \boldsymbol{a}_n + \lambda \boldsymbol{S}_\lambda \mathbf{K}^{-1} \\ &= E(\boldsymbol{\Delta}_n \mid \boldsymbol{x}) + \boldsymbol{a}_n + \boldsymbol{S}_\lambda \{(\boldsymbol{I}_n + \lambda \mathbf{K}^{-1}) - \boldsymbol{I}_n\} \\ &= \boldsymbol{\Pi} + O_p(\boldsymbol{a}_n), \end{aligned} \tag{19}$$

where $\boldsymbol{\Pi} = \operatorname{diag}(\pi(\boldsymbol{x}_1), \ldots, \pi(\boldsymbol{x}_n))$ and $\gamma(\lambda)$ is the effective dimension of kernel K. Similarly, we have

$$\begin{aligned} \boldsymbol{d}_n &= E(\boldsymbol{\Delta}_n \boldsymbol{y} \mid \boldsymbol{x}) + O_p(\boldsymbol{a}_n) \\ &= \boldsymbol{\Pi} \boldsymbol{m} + O_p(\boldsymbol{a}_n). \end{aligned}$$

Now, writing

$$\widehat{\boldsymbol{m}} = \boldsymbol{m} + \boldsymbol{C}_n^{-1}(\boldsymbol{d}_n - \boldsymbol{C}_n \boldsymbol{m})$$

and using (19) and

$$\boldsymbol{d}_n - \boldsymbol{C}_n \boldsymbol{m} = O_p(\boldsymbol{a}) = o_p(\mathbf{1}_n),$$

by Taylor expansion, we have

$$\begin{aligned} \widehat{\boldsymbol{m}} &= \boldsymbol{m} + \{\boldsymbol{\Pi} + O_p(\boldsymbol{a}_n)\}^{-1}(\boldsymbol{d}_n - \boldsymbol{C}_n \boldsymbol{m}) \\ &= \boldsymbol{m} + \boldsymbol{\Pi}^{-1}\left(\boldsymbol{d}_n - \boldsymbol{C}_n \boldsymbol{m}\right) + o_p\left(\boldsymbol{a}_n\right) \\ &= \boldsymbol{m} + \boldsymbol{\Pi}^{-1}\left\{\boldsymbol{S}_\lambda \boldsymbol{\Delta}_n \boldsymbol{y} - \boldsymbol{S}_\lambda \left(\boldsymbol{\Delta}_n + \lambda \mathbf{K}^{-1}\right)\boldsymbol{m}\right\} + o_p\left(\boldsymbol{a}_n\right) \\ &= \boldsymbol{m} + \boldsymbol{\Pi}^{-1} \boldsymbol{S}_\lambda \boldsymbol{\Delta}_n (\boldsymbol{y} - \boldsymbol{m}) + O_p(\boldsymbol{a}_n), \end{aligned}$$

where the last equality holds because

$$\begin{aligned} \boldsymbol{S}_\lambda \lambda \mathbf{K}^{-1} \boldsymbol{m} &= \boldsymbol{S}_\lambda \left\{\left(\boldsymbol{I}_n + \lambda \mathbf{K}^{-1}\right) - \boldsymbol{I}_n\right\} \boldsymbol{m} \\ &= \boldsymbol{m} - \boldsymbol{S}_\lambda \boldsymbol{m} = O_p(\boldsymbol{a}_n). \end{aligned}$$

Therefore, we have

$$\begin{aligned} T_n &= n^{-1}\mathbf{1}_n^{\mathrm{T}}\big(\boldsymbol{I}_n - \boldsymbol{\Delta}_n\big)(\widehat{\boldsymbol{m}} - \boldsymbol{m}) \\ &= n^{-1}\mathbf{1}_n^{\mathrm{T}}\big(\boldsymbol{I}_n - \boldsymbol{\Delta}_n\big)\boldsymbol{\Pi}^{-1}\boldsymbol{S}_\lambda\boldsymbol{\Delta}_n(\boldsymbol{y} - \boldsymbol{m}) + O_p(n^{-1}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{a}_n) \\ &= n^{-1}\mathbf{1}_n^{\mathrm{T}}\big(\boldsymbol{I}_n - \boldsymbol{\Pi}\big)\boldsymbol{\Pi}^{-1}\boldsymbol{\Delta}_n(\boldsymbol{y} - \boldsymbol{m}) + O_p(n^{-1}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{a}_n) \\ &= n^{-1}\mathbf{1}_n^{\mathrm{T}}\big(\boldsymbol{\Pi}^{-1} - \boldsymbol{I}_n\big)\boldsymbol{\Delta}_n(\boldsymbol{y} - \boldsymbol{m}) + O_p(n^{-1}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{a}_n). \end{aligned}$$

For ℓ-th order of Sobolev space, we have

$$\begin{aligned} \gamma(\lambda) &= \sum_{j=1}^{\infty}\frac{1}{1+j^{2\ell}\lambda} \\ &\le \lambda^{-\frac{1}{2\ell}} + \sum_{\{j:j>\lambda^{-\frac{1}{2\ell}}\}}\frac{1}{1+j^{2\ell}\lambda} \\ &\le \lambda^{-\frac{1}{2\ell}} + \lambda^{-1}\int_{\lambda^{-\frac{1}{2\ell}}}^{\infty} z^{-2\ell}dz \\ &= \lambda^{-\frac{1}{2\ell}} + \frac{1}{2\ell-1}\lambda^{-\frac{1}{2\ell}} \\ &= O\Big(\lambda^{-\frac{1}{2\ell}}\Big). \end{aligned}$$

Additionally,

$$n^{-1}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{a}_n = O_p\big(\lambda^{1/2} + n^{-1}\{\gamma(\lambda)\}^{1/2}\big),$$

which implies that, as long as $n\lambda \to 0$ and $n\lambda^{1/2\ell} \to \infty$, holds, we have $n^{-1}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{a}_n = o_p(n^{-1/2})$ and (10) is established. □

B Justification of equation (14)

Note that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\delta_i\widehat{\omega}_i m(\boldsymbol{x}_i) =& n^{-1}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Delta}_n\boldsymbol{S}_\lambda\boldsymbol{m} = n^{-1}\boldsymbol{m}^{\mathrm{T}}\boldsymbol{S}_\lambda\boldsymbol{\Delta}_n\mathbf{1}_n \\ =& n^{-1}\boldsymbol{m}^{\mathrm{T}}(\mathbf{1}_n + \boldsymbol{a}_n) \\ =& \frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{x}_i) + o_p(n^{-1/2}), \end{aligned}$$

where the third equality holds by Lemma 2. □

Acknowledgments The authors thank the AE and two anonymous referees for very constructive comments. The research of the first author is supported by Fujian Provincial Department of Education (JAT210059). The research of the second author is partially supported by a grant from National Science Foundation (OAC-1931380) and a grant from the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68*(3), 337–404.

Bohn, B., Rieger, C., Griebel, M. (2019). A represented theorem for deep kernel learning. *Journal of Machine Learning Research, 20*, 1–32.

Chan, K. C. G., Yam, S. C. P., Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society, Series B, 78*(3), 673–700.

Chen, J., Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association, 96*(453), 260–269.

Chen, S. X., Qin, J., Tang, C. Y. (2013). Mann-whitney test with adjustments to pretreatment variables for missing values and observational study. *Journal of the Royal Statistical Society, Series B, 75*(1), 81–102.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association, 89*(425), 81–87.

Claeskens, G., Krivobokova, T., Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika, 96*(3), 529–544.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its Oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis, 20*, 25–46.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*(346), 383–393.

Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H. (2009). *The elements of statistical learning*: *Data mining, inference, and prediction* (Vol. 2). New York, NY, USA: Springer.

Hayfield, T., Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software, 27*(5), 1–32.

Hurvich, C. M., Simonoff, J. S., Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *60*(2), 271–293.

Kim, J. K., Rao, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika, 96*(4), 917–932.

Kim, J. K., Shao, J. (2021). Statistical methods for handling incomplete data, 2nd edition. New York: Chapman & Hall/CRC.

Koltchinskii, V., et al. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics, 34*(6), 2593–2656.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., Chen, S. X. (2015). Assessing Beijing's PM 2.5 pollution: Severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257.

Little, R. J. A., Rubin, D. B. (2019). Statistical analysis with missing data, 3rd edition. Hoboken, NJ: John Wiley & Sons.

Mendelson, S. (2002). Geometric parameters of kernel machines. In *International conference on computational learning theory*, pages 29–43. Springer.

Morgan, S. L., Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research*, 2nd edition. Cambridge: Cambridge University Press.

Rasmussen, C. E., Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and Methods, 23*(8), 2379–2412.

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*(427), 846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

Sang, H., Kim, J. K., Lee, D. (2022). Semiparametric fractional imputation using gaussian mixture models for handling multivariate missing data. *Journal of the American Statistical Association, 117*(538), 654–663.

Schölkopf, B., Smola, A. J., Bach, F. (2002). *Learning with Kernels*: *Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT press.

Shawe-Taylor, J., Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Steinwart, I., Hush, D. R., Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93.

Stone, C. (1982). Optimal global rates of converence for nonparametric regression. *The Annals of Statistics, 10*(4), 1040–1053.

Tan, Z. (2020). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika, 107*(1), 137–158.

van de Geer, S. A. (2000). *Empirical processes in M-estimation* (Vol. 6). Cambridge University Press.

Wahba, G. (1990). Spline models for observational data, volume 59. Siam.

Wang, D., Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics, 37*(1), 490–517.

Wong, R. K., Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika, 105*(1), 199–213.

Wood, S. (2017). *Generalized additive models: An introduction with R*, 2nd edition. New York: Chapman and Hall/CRC.

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

Yang, S., Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association, 115*(531), 1540–1554.

Yang, S., Kim, J. K. (2020). Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scandinavian Journal of Statistics, 47*, 839–861.

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation, 17*(9), 2077–2098.

Zhang, T., Simon, N. (2023). An online projection estimator for nonparametric regression in reproducing kernel hilbert spaces. *Statistica Sinica, 33*, 127–148.

Zhang, Y., Duchi, J., Wainwright, M. (2013). Divide and conquer kernel ridge regression. In *Conference on learning theory*, pages 592–617.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics, 47*(2), 965–993.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.