On classification with nonignorable missing data[1]

Majid Mojirsheibani[2]
Department of Mathematics, California State University, Northridge, CA 91330

Abstract

We consider the problem of kernel classification with nonignorable missing data. Instead of imposing a fully parametric model for the selection probability, which can be quite sensitive to the violations of model assumptions, here we consider a semiparametric exponential tilting selection probability model in the spirit of Kim and Yu (2011). In addition to the existing parameter estimators, we also develop some new estimators of the unknown components of the model that are particularly suitable for classification problems. We also study various strong optimality properties of the proposed kernel-type classifiers.

Keywords Kernel, missing data, regression, classification, convergence.

1 Introduction

In recent years, the problem of statistical estimation, prediction, and inference with nonignorable missing data has received considerable attention. This is the situation where, unlike the missing at random assumption, the probability that a variable is missing depends on the variable itself as well. Recent key results on nonignorable missing response data include the landmark paper of Kim and Yu (2011) as well as those of Zhao and Shao (2015), Shao and Wang (2016), Morikawa, et al. (2017), Zhao et al. (2017), Uehara and Kim (2018), Morikawa and Kim (2018), Morikawa and Kano (2018), Fang et al. (2018), O'Brien et al. (2018), Maity et al. (2019), Sadinle and Reiter (2019), Zhao et al. (2019), Yuan et al. (2020), Chen et al. (2020), and Liu et al. (2021). As discussed in virtually all of the above cited papers, nonignorable missing data mechanisms pose major challenges in terms of the estimation of various unknown quantities in the model.

In the context of predictive models (as in regression and classification), where a response variable Y is to be predicted based on the covariates $\boldsymbol{X} \in \mathbb{R}^d$, Kim and Yu (2011) considered the following model where Y may be missing nonignorably according to

$$\pi(\boldsymbol{x}, y) := P\left\{\delta = 1 | \boldsymbol{X} = \boldsymbol{x}, Y = y\right\} = \Big[1 + \exp\{g(\boldsymbol{x})\} \cdot \exp\{\gamma y\}\Big]^{-1}, \tag{1}$$

where the indicator variable $\delta = 0$ if Y is missing (and δ =1, otherwise). Here g is a completely unspecified function and γ is an unknown parameter. Additionally, Kim and Yu (2011) developed a kernel estimator of the function $\exp\{g(\boldsymbol{x})\}$ for the case of a known γ. As for the parameter γ, these authors assume that it can be estimated from an independent external data; this approach has been further studied and refined by, for example, Zhao, et al. (2013) and Tang, et al. (2014), who still need γ to be estimated from some available external data. To circumvent the requirement of external data, Shao and Wang (2016) showed that if the function $g(\boldsymbol{x})$ in (1) depends only on

[1] MSC: 62G08

[2] Email: majid.mojirsheibani@csun.edu

This work was supported by the NSF under Grant DMS-1916161 of Majid Mojirsheibani

some subset of the covariates in $\boldsymbol{x}$, then all unknown components of the model will be identifiable and can be consistently estimated; their proposed method is based on estimating equations.

The focus of this article is on the problem of nonparametric classification in the presence of nonignorable missing data, where we develop new kernel-type classification rules that are asymptotically strongly optimal under fairly standard assumptions. Here, optimality means that the error of the proposed classification rule converges to that of the theoretically best (but unknown) classifier. This problem has only been tackled for the simpler case of Missing At Random (MAR) scenarios in the literature; see, for example, Reese and Mojirsheibani (2017) and Mojirsheibani and Reese (2017). In fact, to the best of our knowledge, our results in this paper are the first to tackle the problem of nonparametric classification with nonignorable missing data.

Our contributions in this paper are two-fold. First, we propose an initial kernel-based classifier that takes into account the nonignorable selection probability in (1); we also derive probabilistic upper bounds on the performance of this classifier. These bounds depend on the quantity $P(|\widehat{\gamma}-\gamma| > c)$, where c is a fixed constant and $\widehat{\gamma}$ is any estimator of γ. As a result, the optimality of this classifier will depend on the quality of $\widehat{\gamma}$. For example, the estimator of Kim and Yu (2011) works well but it requires external data. Similarly, the estimator of Shao and Wang (2016) requires the function g in (1) to depend only on some parts of $\boldsymbol{x}$.

The second part of our contributions involves a new estimator of the nonignorability component $\varphi(y) := \exp\{\gamma y\}$ in (1). The new estimation approach, which also works for more general functions $\varphi(y)$, does not require any external data (as in Kim and Yu (2011). This approach also evades the conditions imposed by Shao and Wang (2016) on the function g in (1). Furthermore, we show that the corresponding revised kernel classifier is strongly asymptotically optimal with this new estimator. This improvement owes to the fact that our new estimator, which is based on the approximation theory of totally bounded classes of functions, is selected to minimize a measure of the empirical error of the proposed kernel classifier (see (15)). Our key results along these lines include the exponential performance bound in Theorem 3 and its consequence in terms of the strong optimality of the proposed classifier.

In passing, we also note that although our results are stated for kernel classifiers, similar results can be obtained if we replace kernels with other popular methods such as nearest neighbors and cubic histogram classification rules. However, due to page limitations and for the sake of concreteness, we only present kernel rules here. This paper is organized as follows. Section 2 presents the main results. Theorem 3 provides asymptotic exponential performance bounds on the deviations of the misclassification error of the proposed kernel classification rule from that of the theoretically optimal (but unknown) classifier. Such bounds in conjunction with the Borel-Cantelli lemma immediately yield strong (i.e., almost-sure) optimality results for the proposed classifiers. All proofs are deferred to Section 4. Furthermore, several numerical examples are presented in Section 3; these numerical results confirm the good finite-sample performance of the proposed kernel classifier.

2 Main results

2.1 Some background

To state our main results, consider the following standard two-group classification problem. Let $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \{0,1\}$ be a random pair where the class variable Y has to be predicted based on the covariate vector $\boldsymbol{X}$. In classification, one seeks to find a function (a classifier) $\psi : \mathbb{R}^d \to \{0,1\}$ for which the misclassification error, defined by $L(\psi) := P\{\psi(\boldsymbol{X}) \neq Y\}$, is as small as possible. The best classifier, denoted by ψ_{B}, is the one that has the smallest misclassification error, i.e., $P\{\psi_{\mathrm{B}}(\boldsymbol{X}) \neq Y\} = \min_{\psi:\mathbb{R}^d\to\{0,1\}} L(\psi)$. If we let η be the class conditional probability for class 1, i.e.,

$$\eta(\boldsymbol{x}) = E\left[Y \middle| \boldsymbol{X} = \boldsymbol{x}\right] = P\left\{Y = 1 \middle| \boldsymbol{X} = \boldsymbol{x}\right\}, \tag{2}$$

then it is straightforward to see that the best classifier is (see, for example, Devroye et al. (1996, Sec. 2))

$$\psi_B(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \eta(\boldsymbol{x}) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Clearly, in practice, the regression function $\eta(\boldsymbol{x})$ is not available because the distribution of $(\boldsymbol{X}, Y)$ is virtually always unknown. One general approach in nonparametric setups is to replace the unknown regression function $\eta(\boldsymbol{x})$ in (3) by some estimate, say $\widetilde{\eta}(\boldsymbol{x})$, based on the iid data $\mathcal{D}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$, and use the plug-in version of (3) given by $\widetilde{\psi}_n(\boldsymbol{x}) = 1$ if $\widetilde{\eta}(\boldsymbol{x}) > \frac{1}{2}$ (otherwise, $\widetilde{\psi}_n(\boldsymbol{x}) = 0$).

Now suppose that some of the Y_i's may be missing nonignorably, i.e., the probability that Y_i is missing depends on Y_i (and possibly on $\boldsymbol{X}_i$). In order to take this fact into account when constructing sample versions of (3), first observe that for each fix γ in (1), the regression function $\eta(\boldsymbol{x})$ in (2) can also be written as (see Lemma 3 in Sec. 4):

$$\eta(\boldsymbol{x}) = E\left[\delta Y \middle| \boldsymbol{X} = \boldsymbol{x}\right] + \frac{E\left[\delta Y \exp\{\gamma Y\} \middle| \boldsymbol{X} = \boldsymbol{x}\right]}{E\left[\delta \exp\{\gamma Y\} \middle| \boldsymbol{X} = \boldsymbol{x}\right]} \cdot E[1 - \delta | \boldsymbol{X} = \boldsymbol{x}]. \tag{4}$$

Next, let $\widehat{\gamma}$ be any estimator of γ; this could be, for example, the estimator proposed by Shao and Wang (2016), or the estimators discussed by Kim and Yu (2011) based on external data. In general, here we do not require $\widehat{\gamma}$ to be independent of the data $\mathbb{D}_n$. Now, consider the kernel-type estimator of $\eta(\boldsymbol{x})$ given by

$$\begin{aligned} \widehat{\eta}(\boldsymbol{x}) \quad = \quad & \frac{\sum_{i=1}^n \delta_i Y_i K((\boldsymbol{x} - \boldsymbol{X}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{x} - \boldsymbol{X}_i)/h)} + \frac{\sum_{i=1}^n \delta_i Y_i \exp\{\widehat{\gamma}\, Y_i\} K((\boldsymbol{x} - \boldsymbol{X}_i)/h)}{\sum_{i=1}^n \delta_i \exp\{\widehat{\gamma}\, Y_i\} K((\boldsymbol{x} - \boldsymbol{X}_i)/h)} \\ & \times \frac{\sum_{i=1}^n (1 - \delta_i) K((\boldsymbol{x} - \boldsymbol{X}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{x} - \boldsymbol{X}_i)/h)}, \end{aligned} \tag{5}$$

where $K : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel used with bandwidth h. Therefore, we have the plug-in kernel type classification rule

$$\widehat{\psi}_n(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \widehat{\eta}(\boldsymbol{x}) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

How good is the classifier $\widehat{\psi}_n(\boldsymbol{x})$ in (6) as compared to the optimal classifier ψ_{B} in (3)? To answer this, we will assume that the kernel K is *regular* (Devroye and Krzyżak (1989)): *A nonnegative* *kernel K is said to be regular if there are constants $b > 0$ and $r > 0$ such that $K(\boldsymbol{u}) \geq b\, I\{\boldsymbol{u} \in S_{0,r}\}$ and $\int \sup_{\boldsymbol{y}\in\boldsymbol{u}+S_{0,r}} K(\boldsymbol{y})\, d\boldsymbol{u} < \infty$, where $S_{0,r}$ is the ball of radius r with center at the origin.*

We also need the following assumption regarding the missing probability mechanism $\pi(\boldsymbol{x}, y)$ in (1), which is also used by Kim and Yu (2011). It implies, in a sense, that Y can be observed with a nonzero probability for all values of $\boldsymbol{x}$ and y.

Assumption A1. $\inf_{\boldsymbol{x},y} \pi(\boldsymbol{x}, y) =: \pi_{\min} > 0$, for some arbitrarily small $\pi_{\min}$.

Our first result below provides upper bounds on the performance of the proposed classifier, under rather standard assumptions, which can then be used to study the strong/weak optimality properties of this classifier (see Remark 1).

Theorem 1 *Let $\widehat{\gamma}$ be any estimator of γ in (1) and let $\widehat{\eta}(\boldsymbol{x})$ be the kernel estimator appearing in (5), where the kernel K is regular, and suppose that assumption A1 holds. Let $\widehat{\psi}_n(\boldsymbol{x})$ be the classifier defined via (6) and (5), and suppose that $h \to 0$ and $nh^d \to \infty$, as $n \to \infty$. Then, for every $\varepsilon > 0$, any distribution of $(\boldsymbol{X}, Y)$, and n large enough, one has*

$$P\left\{L_n(\widehat{\psi}_n) - L(\psi_B) > \varepsilon\right\} \;\leq\; c_1\, e^{-c_2 n} + c_3\, e^{-nc_4\varepsilon^2} + c_5\, P\left\{\left|\widehat{\gamma} - \gamma\right| > C_0\right\}, \tag{7}$$

where $c_1, \ldots, c_5$, and C_0 are positive constants not depending on n, $L(\psi_B) = P\{\psi_B(\boldsymbol{X}) \neq Y\}$, and $L_n(\widehat{\psi}_n) = P\big\{\widehat{\psi}_n(\boldsymbol{X}) \neq Y \,\big|\, \mathbb{D}_n\big\}$ is the conditional error of the classifier $\widehat{\psi}_n$.

Remark 1 *Due to the presence of the term $P\{|\widehat{\gamma} - \gamma| > C_0\}$ in (7), the bound given in Theorem 1 falls short of the classical exponential bounds established by Devroye and Krzyżak (1989) for the usual kernel estimators with no missing data. This is the price to pay to allow $\widehat{\gamma}$ to be any arbitrary estimator of γ in Theorem 1. Therefore, (7) does not guarantee the almost sure convergence of the error $L_n(\widehat{\psi}_n) \rightarrow^{a.s.} L(\psi_B)$, unless $P\{|\widehat{\gamma} - \gamma| > C_0\}$ goes to zero fast enough (e.g. $P\{|\widehat{\gamma} - \gamma| > C_0\} \leq n^{-\alpha}$ for some $\alpha > 1$). Of course, if $\widehat{\gamma} \rightarrow^{p} \gamma$ (as in, for example, Kim and Yu (2011) or Shao and Wang (2016)), then weak consistency follows immediately, i.e., $L_n(\widehat{\psi}_n) \rightarrow^{p} L(\psi_B)$.*

2.2 A more general approach

Remark 1 together with Theorem 1 show that the performance of the initial proposed classifier $\widehat{\psi}_n(\boldsymbol{x})$ in (6) depends on the quality of $\widehat{\gamma}$. Furthermore, the currently available estimators of γ require some stringent conditions that may not hold in practice. For example, the estimator proposed by Kim and Yu (2011), and further refined by others such as Zhao, et al. (2013) and Tang, et al. (2014), still requires the availability of the missing y values for a randomly selected validation subset of the nonrespondents through external data. Similarly, for the estimator of Shao and Wang (2016), one must find a part $\mathbf{z}$ of the vector $\mathbf{x} = (\mathbf{u}, \mathbf{z})$ that is not involved in the function g in (1), i.e., one must work with some $g(\mathbf{u})$ instead of $g(\mathbf{x})$ in the definition of $\pi(\boldsymbol{x}, y)$ in (1). These assumptions are needed in order to be able to estimate various parameters consistently. In this section, we propose alternative estimators that still yield strongly optimal kernel classifier without the types of assumptions imposed in the above cited references. Our estimators are based on the approximation theory of totally bounded classes of functions. More specifically, consider the more general nonignorable missing probability model

$$\pi(\boldsymbol{x}, y) \;\equiv\; \pi_{\varphi}(\boldsymbol{x}, y) \;:=\; \Big[1 + \exp\{g(\boldsymbol{x})\} \cdot \varphi(y)\Big]^{-1}, \tag{8}$$

where $\varphi(y) > 0$ is unknown. To develop a theoretical framework for our estimator, we consider the situation where φ belongs to a totally bounded class of functions $\mathcal{F}$ and y may be any bounded variable. We then apply our results to the particular case of interest where $\varphi(y) = e^{\gamma y}$ as in (1) but with $y \in \{0, 1\}$. More specifically, let $\mathcal{F}$ be a given class of functions $\varphi : [-L, L] \longrightarrow (0, B]$, for some $B < \infty$. Fix $\varepsilon > 0$ and suppose that the finite collection of functions $\mathcal{F}_\varepsilon = \{\varphi_1, \ldots, \varphi_{N(\varepsilon)}\}$, $\varphi_i : [-L, L] \to (0, B]$, is an ε-cover of $\mathcal{F}$, i.e., for each $\varphi \in \mathcal{F}$, there is a $\varphi^\dagger \in \mathcal{F}_\varepsilon$ such that $\|\varphi - \varphi^\dagger\|_\infty < \varepsilon$; here, $\|\|_\infty$ is the usual supnorm. We also note that $\mathcal{F} \subset \bigcup_{1 \le i \le N(\varepsilon)} B(\varphi_i, \varepsilon)$, where $B(\varphi_i, \varepsilon)$ is the ball of functions centered at φ_i, with $\|\|_\infty$-radius equal to ε. The cardinality of the smallest ε-cover of $\mathcal{F}$ is called the *covering number* of the family $\mathcal{F}$ and will be denoted by $\mathcal{N}(\varepsilon, \mathcal{F})$. If $\mathcal{N}(\varepsilon, \mathcal{F}) < \infty$ for every $\varepsilon > 0$, then the family $\mathcal{F}$ is said to be *totally bounded* (with respect to $\|\|_\infty$). For more on such concepts from the approximation theory, one may refer to, for example, van der Vaart and Wellner (1996; p. 83). The following simple example illustrates this approach.

Consider the class of functions $\mathcal{F}$ of the form

$$\varphi(y) = \begin{cases} e^{\gamma y} & \text{if } \; |y| \le L, \;\; |\gamma| \le M, \quad \text{(for some } L < \infty \text{ and } M < \infty) \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

which is in the spirit of the model proposed by Kim and Yu (2011); see the term $\exp\{\gamma y\}$ in (1). Now, let

$$\gamma_i = \frac{2\,i\varepsilon}{L\exp\{ML\}}, \quad \text{where} \quad -\left\lfloor \frac{ML\exp\{ML\}}{\varepsilon} \right\rfloor \le i \le \left\lfloor \frac{ML\exp\{ML\}}{\varepsilon} \right\rfloor$$

and define the set

$$\Theta_\varepsilon = \left\{ \gamma_i = \frac{2\,i\varepsilon}{L\exp\{ML\}} \;\middle|\; -\left\lfloor \frac{ML\exp\{ML\}}{\varepsilon} \right\rfloor \le i \le \left\lfloor \frac{ML\exp\{ML\}}{\varepsilon} \right\rfloor \right\} \cup \{-M\} \cup \{M\}.$$

Then $\mathcal{F}_\varepsilon = \left\{ e^{\gamma y} \;\middle|\; -L \le y \le L, \quad \gamma \in \Theta_\varepsilon \right\}$ is an ε-cover of the family $\mathcal{F}$ of functions of the form (9). To see this, observe that if $\gamma^* \in [-M, M]$ with the corresponding function $\varphi^*(y) = e^{\gamma^* y} \in \mathcal{F}$, and if $\widetilde{\gamma} \in \Theta_\varepsilon$ is the closest value to γ^*, then for every $\varepsilon > 0$

$$\begin{aligned} \sup_{-L \le y \le L} \left| e^{\gamma^* y} - e^{\widetilde{\gamma} y} \right| &= \sup_{-L \le y \le L} \left| y \, \exp\{\overline{\gamma} y\} \right| \cdot \left| \widetilde{\gamma} - \gamma^* \right|, \quad \text{where } \; \overline{\gamma} \in (\widetilde{\gamma} \wedge \gamma^*, \, \widetilde{\gamma} \vee \gamma^*) \\ &\le L\exp\{ML\} \cdot \left| \widetilde{\gamma} - \gamma^* \right| \\ &\le L\exp\{ML\} \cdot \frac{\varepsilon}{L\exp\{ML\}} \\ &= \varepsilon\,, \end{aligned}$$

where we have used the fact that the distance between γ^* and its nearest value in Θ_ε is bounded by $(L\exp\{ML\})^{-1}\varepsilon$. Therefore, the class $\mathcal{F}$ is totally bounded and its ϵ-covering number is bounded by the quantity $2\lfloor ML\exp\{ML\}\varepsilon^{-1} \rfloor + 3$. Of course, in the case of classification with $y \in \{0, 1\}$, the constant L becomes 1 everywhere.

To construct our estimators, for each $\varphi \in \mathcal{F}$ define the classifier

$$\psi_\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \; \mathcal{R}(\boldsymbol{x}; \varphi) > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $\mathcal{R}(\boldsymbol{x};\varphi) := E\big[Y\big|\boldsymbol{X}=\boldsymbol{x}\big]$. The function $\mathcal{R}(\boldsymbol{x};\varphi)$ can also be expressed as (see Lemma 3)

$$\mathcal{R}(\boldsymbol{x};\varphi) \;=\; E\left[\delta Y\big|\boldsymbol{X}=\boldsymbol{x}\right] + \frac{E\left[\delta\, Y\,\varphi(Y)\big|\boldsymbol{X}=\boldsymbol{x}\right]}{E\left[\delta\,\varphi(Y)\big|\boldsymbol{X}=\boldsymbol{x}\right]}\cdot E[1-\delta|\boldsymbol{X}=\boldsymbol{x}]. \tag{11}$$

Here, $\mathcal{F}$ is any class of functions of the form $\varphi:[0,1]\to(0,B]$ for some finite B. Also let φ^* be the true value of φ, i.e., we have

$$L^* \;:=\; P\big\{\psi_{\varphi^*}(\boldsymbol{X})\neq Y\big\} \;=\; \inf_{\psi:\,\mathbb{R}^d\to\{0,1\}} P\big\{\psi(\boldsymbol{X})\neq Y\big\}; \qquad \text{also put} \qquad \psi^*(\boldsymbol{x}) := \psi_{\varphi^*}(\boldsymbol{x}) \tag{12}$$

i.e., ψ_{φ^*}, which is obtained by substituting φ^* for φ in (10) and (11), is the best classifier. Of course, φ^* may or may not be in $\mathcal{F}$. Now, let $\mathbb{D}_n = \{(\boldsymbol{X}_1, Y_1, \delta_1), \dots, ((\boldsymbol{X}_n, Y_n, \delta_n)\}$ represent the data, where $\delta_i = 0$ if Y_i is missing (and $\delta_i = 1$ otherwise). To present our estimators and various classifiers, start by randomly splitting $\mathbb{D}_n$ into a training sample $\mathbb{D}_m$ of size m and a testing sequence $\mathbb{D}_\ell$ of size $\ell = n-m$. Here, $\mathbb{D}_m\cup\mathbb{D}_\ell = \mathbb{D}_n$ and $\mathbb{D}_m\cap\mathbb{D}_\ell = \varnothing$. The choices of m and ℓ will be discussed later. Now, for each $\varphi\in\mathcal{F}$, define the sample based classification rule constructed based on $\mathbb{D}_m$ by

$$\widehat{\psi}_{m,\varphi}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi) \;>\; \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for each } \varphi\in\mathcal{F}, \tag{13}$$

where $\widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi)$ is the kernel estimator of the regression function $\mathcal{R}(\boldsymbol{x};\varphi)$, based on $\mathbb{D}_m$, given by

$$\begin{aligned}\widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi) \;&=\; \frac{\sum_{i:(\boldsymbol{X}_i,Y_i,\delta_i)\in\mathbb{D}_m}\delta_i Y_i K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i:(\boldsymbol{X}_i,Y_i,\delta_i)\in\mathbb{D}_m} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} \\ &\quad + \left[\frac{\sum_{i:(\boldsymbol{X}_i,Y_i,\delta_i)\in\mathbb{D}_m}\delta_i Y_i\,\varphi(Y_i)K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i:(\boldsymbol{X}_i,Y_i,\delta_i)\in\mathbb{D}_m}\delta_i\,\varphi(Y_i)K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}\right. \\ &\qquad\qquad \left.\times\,\frac{\sum_{i:(\boldsymbol{X}_i,Y_i,\delta_i)\in\mathbb{D}_m}(1-\delta_i)K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i:(\boldsymbol{X}_i,Y_i,\delta_i)\in\mathbb{D}_m} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}\right].\end{aligned} \tag{14}$$

Next, for each $\varphi\in\mathcal{F}$, let $\widehat{L}_\ell(\widehat{\psi}_{m,\varphi})$ be the weighted empirical error of $\widehat{\psi}_{m,\varphi}$ committed on the testing sequence $\mathbb{D}_\ell$, i.e., for each $\varphi\in\mathcal{F}$,

$$\widehat{L}_\ell(\widehat{\psi}_{m,\varphi}) \;\;:=\;\; \ell^{-1}\sum_{i:\,(\boldsymbol{X}_i,Y_i,\delta_i)\,\in\,\mathbb{D}_\ell}\frac{\delta_i}{\widehat{\pi}_\varphi(\boldsymbol{X}_i,Y_i)}\,I\big\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i)\neq Y_i\big\}, \tag{15}$$

where, for each $\varphi\in\mathcal{F}$ the quantity $\widehat{\pi}_\varphi$ in the denominator of (15) is given by

$$\widehat{\pi}_\varphi(\boldsymbol{X}_i,Y_i) = \left[1 + \frac{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\,\in\mathbb{D}_m}(1-\delta_j)K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\,\in\mathbb{D}_m}\delta_j\,\varphi(Y_j)\,K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}\cdot\varphi(Y_i)\right]^{-1}. \tag{16}$$

We note that (16) is justified as a kernel estimator by the fact that the term $\exp\{g(\boldsymbol{x})\}$ in (8) can alternatively be written as

$$\exp\{g(\boldsymbol{x})\} = \frac{E\big[1-\delta\big|\boldsymbol{X}=\boldsymbol{x}\big]}{E\big[\delta\,\varphi(Y)\big|\boldsymbol{X}=\boldsymbol{x}\big]}. \tag{17}$$

Remark 2 below provides some explanation for the type of empirical error employed in (15). Next, fix $\varepsilon>0$ and let $\mathcal{F}_\varepsilon = \{\varphi_1,\dots,\varphi_{N(\varepsilon)}\}\subset\mathcal{F}$ be any ε-cover of $\mathcal{F}$. Then, our estimate of the unknown function φ is given by

$$\widehat{\varphi}\equiv\widehat{\varphi}_n = \underset{\varphi\in\mathcal{F}_\varepsilon}{\operatorname{argmin}}\;\widehat{L}_\ell(\widehat{\psi}_{m,\varphi}), \tag{18}$$

where we note that $\widehat{\varphi}$ depends on the entire data $\mathbb{D}_n$. Finally, our proposed classifier under the general setup of Section 2.2 is

$$\widehat{\psi}_{n,\widehat{\varphi}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \widehat{\mathcal{R}}_m(\boldsymbol{x};\widehat{\varphi}) \ > \ \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

where $\widehat{\mathcal{R}}_m(\boldsymbol{x};\widehat{\varphi})$ is obtained from $\widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi)$ upon replacing the function φ by $\widehat{\varphi}$ everywhere in (14). In our notation, the presence of n at $\widehat{\psi}_{n,\widehat{\varphi}}$ in (19) signifies the fact that it depends on the whole data. Of course, if the estimator $\widehat{\varphi}$ is replaced by a nonrandom function $\varphi \in \mathcal{F}$, then the notation $\widehat{\psi}_{n,\widehat{\varphi}}$ will immediately reduce to $\widehat{\psi}_{m,\varphi}$, which is given by (13).

Remark 2 *Our definition of the empirical error in (15) looks quite different from the more usual empirical error* $\overline{L}_\ell(\widehat{\psi}_{m,\varphi}) := \ell^{-1}\sum_{i:\,(\boldsymbol{X}_i,Y_i,\delta_i)\,\in\,\mathbb{D}_\ell} I\big\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\big\}$ *that counts the number of errors committed by the classifier* $\widehat{\psi}_{m,\varphi}$ *on the testing sequence* $\mathbb{D}_\ell$*. This is because* $\overline{L}_\ell(\widehat{\psi}_{m,\varphi})$ *is not necessarily computable (since some of the* Y_i*'s are missing). Furthermore, working with the alternative quantity,* $\overline{L}^*_\ell(\widehat{\psi}_{m,\varphi}) := \ell^{-1}\sum_{i:\,(\boldsymbol{X}_i,Y_i,\delta_i)\,\in\,\mathbb{D}_\ell} \delta_i\, I\big\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\big\}$*, is not going to resolve the issue as it is no longer an unbiased estimator of the error probability* $L_m(\widehat{\psi}_{m,\varphi}) := P\big\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}) \neq Y \big| \mathbb{D}_m\big\}$*, (not even asymptotically), which is due to the fact that the expected value of* $\overline{L}^*_\ell$*, conditional on* $\mathbb{D}_m$*, is not equal to* $L_m(\widehat{\psi}_{m,\varphi})$*. Of course, the more natural choice is to use* $\pi_\varphi(\boldsymbol{X}_i, Y_i)$ *instead of* $\widehat{\pi}_\varphi(\boldsymbol{X}_i, Y_i)$ *in the denominator of (15), but* π_φ *is unknown.*

To study the performance of the proposed classifier $\widehat{\psi}_{n,\widehat{\varphi}}(\boldsymbol{x})$ in (19) and the closeness and convergence of its error rate to that of the optimal classifier ψ^* given by (12), first let

$$L_n(\widehat{\psi}_{n,\widehat{\varphi}}) \ = \ P\big\{\widehat{\psi}_{n,\widehat{\varphi}}(\boldsymbol{X}) \neq Y \big| \mathbb{D}_n\big\} \quad \text{and} \quad L^* \ = \ L(\psi^*) \ = \ P\{\psi^*(\boldsymbol{X}) \neq Y\}, \tag{20}$$

and observe the fundamental decomposition

$$L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - L^* \ = \ \Big[L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - \inf_{\varphi\in\mathcal{F}_\varepsilon} L(\psi_\varphi)\Big] + \Big[\inf_{\varphi\in\mathcal{F}_\varepsilon} L(\psi_\varphi) - L^*\Big], \tag{21}$$

where $L(\psi_\varphi) \ = \ P\{\psi_\varphi(\boldsymbol{X}) \neq Y\}$ is the misclassification error of any classifier ψ_φ of the form (10). The first bracketed term in (21) is referred to as the estimation error, whereas the second term is the approximation error. To investigate their properties, we first state a number of assumptions which are quite standard in kernel regression estimation:

Assumption A2. The probability density function $f(\boldsymbol{x})$ of $\boldsymbol{X}$ is compactly supported and is bounded away from zero and infinity on its compact support. Additionally, the first-order partial derivatives of f exist and are bounded on the interior of its support.

Assumption A3. The kernel K satisfies $\int_{\mathbb{R}^d} K(\boldsymbol{x})\,d\boldsymbol{x} = 1$ and $\int_{\mathbb{R}^d} |x_i| K(\boldsymbol{x})\,d\boldsymbol{x} < \infty$, for $x_i \in (x_1,\cdots,x_d)' = \boldsymbol{x}$. Also, the smoothing parameter h satisfies $h \to 0$ and $mh^d \to \infty$, as $n \to \infty$.

Assumption A4. The partial derivatives $\frac{\partial}{\partial x_i}E[\delta|\boldsymbol{X} = \boldsymbol{x}]$ and $\frac{\partial}{\partial x_i}E[\delta\varphi(Y)|\boldsymbol{X} = \boldsymbol{x}]$ exist for $i = 1,\ldots,d$, and are bounded on the compact support of f.

Assumption A5. $E[\delta\,\varphi(Y)|\boldsymbol{X} = \boldsymbol{x}] \geq \varphi_{00}$, for μ–a.e. $\boldsymbol{x}$ and each $\varphi \in \mathcal{F}$, for some finite $\varphi_{00} > 0$.

Assumption A2 is often imposed in nonparametric regression to avoid having unstable estimates in the tails of f. Assumption A3 is not much of a constraint because the choice of K is at our discretion. We note that if $m = [cn]$ for some $c \in (0, 1)$, then requiring $nh^d \to \infty$ is equivalent to $mh^d \to \infty$, however, we do not want to impose such restrictions on the choice of m. Assumption A4 is technical; in fact, the first part of Assumption A4 has already been used in the literature (Cheng and Chu (1996)). Assumption A5 is not as restrictive as it appears because $E[\delta\,\phi(Y)|\boldsymbol{X}] = E\big[E\{\delta\varphi(Y)\big|\boldsymbol{X}, Y\}\big|\boldsymbol{X}\big] = E\big[\varphi(Y)E(\delta|\boldsymbol{X}, Y)\big|\boldsymbol{X}\big] \geq \pi_{\min}E\big[\varphi(Y)\big|\boldsymbol{X}\big]$, a.s. (by Assumption A1), and the fact that $\varphi(y) > 0$ for all y. Therefore, Assumption A5 is weaker than requiring $E[\varphi(Y)|\boldsymbol{X}] \geq \alpha_0 > 0$, a.s., for some $\alpha_0 > 0$. The following result deals with the estimation error, i.e. the error represented by the first bracketed term on the right side of (21). More specifically, we have

Theorem 2 *Let $\mathcal{F}$ be a totally bounded class of functions $\varphi : [0, 1] \to (0, B]$, for some $B < \infty$. Let the selection probability π be as in (8) and suppose that assumptions A1 – A5 hold. Then for every $\beta > 0$, every $\varepsilon > 0$, and n large enough,*

$$\begin{aligned} P\left\{L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - \inf_{\varphi\in\mathcal{F}_\varepsilon} L(\psi_\varphi) > \beta\right\} \;&\leq\; 2\,|\mathcal{F}_\varepsilon|\,e^{-\ell\beta^2/32} + c_{22}\,e^{-a_0\,m} \\ &\quad + \ell\,\big|\mathcal{F}_\varepsilon\big|\cdot\Big(c_{20}\,e^{-c_{21}\,mh^d\beta^2} + c_{17}\,e^{-c_{18}\,mh^d}\Big) \end{aligned}$$

where a_0 and c_{21} are positive constants depending on β, but not on m or ℓ. Here, c_{17}, c_{18}, and c_{20} are also positive constants, not depending on m or ℓ or β.

Theorem 2 in conjunction with the Borel-Cantelli lemma immediately yields the strong convergence of the estimation error (to zero), i.e., $L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - \inf_{\varphi\in\mathcal{F}_\varepsilon} L(\psi_\varphi) \to^{a.s.} 0$, whenever $(m\,h^d)^{-1}\log\ell \to 0$, as $n \to \infty$. To deal with the approximation error in (21), suppose that $\varphi^* \in \mathcal{F}$, (where φ^* is as in (12)), and let $\widetilde{\varphi} \in \mathcal{F}_\varepsilon$ be such that $\varphi^* \in B(\widetilde{\varphi}, \varepsilon)$. Such a $\widetilde{\varphi}$ exists because $\mathcal{F}_\varepsilon$ is an ε-cover of $\mathcal{F}$ and $\varphi^* \in \mathcal{F}$. Now let $\mathcal{R}(\boldsymbol{x}; \varphi)$ be as in (11) and observe that in view of the results of Devroye et al. (1996, p. 93)

$$\begin{aligned} \inf_{\varphi\in\mathcal{F}_\varepsilon} L(\psi_\varphi) - L^* \;&\leq\; L(\psi_{\widetilde{\varphi}}) - L^* \\ &\leq\; 2\int\Big|\mathcal{R}(\boldsymbol{x}; \widetilde{\varphi}) - \mathcal{R}(\boldsymbol{x}; \varphi^*)\Big|\,\mu(d\boldsymbol{x}) \\ &\leq\; \kappa_0 \sup_{0\leq y\leq 1}\big|\widetilde{\varphi}(y) - \varphi^*(y)\big|, \qquad \text{by Lemma 4} \\ &\leq\; \kappa_0\,\varepsilon, \qquad\qquad \text{for some constant } \kappa_0 > 0, \end{aligned} \tag{22}$$

where the last line follows because $\varphi^* \in B(\widetilde{\varphi}, \varepsilon)$. Now, the bound in (22) shows that the second term on the right side of (21) can converge to zero if ε is replaced by a decreasing sequence $\varepsilon_m \downarrow 0$, as $m \to \infty$. At the same time, ε_m should not converge to zero too rapidly because $\mathcal{F}_{\varepsilon_m} \to \infty$, as $\varepsilon_m \downarrow 0$, and, as a result, the first term on the right side of (21) may not necessarily converge to zero anymore (in view of the bound in Theorem 2). To address these points more formally, we first state the following theorem regarding the error difference $L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - L^*$, (i.e., the left side of (21)), followed by a corollary that captures the key points regarding the asymptotic choices of m, ℓ, and ε_m that yield strong optimality results for the proposed classifier.

Theorem 3 *Suppose that the assumptions of Theorem 2 hold. Then for every $\varepsilon_m > 0$ satisfying $\varepsilon_m \downarrow 0$, as $m \to \infty$, every $\beta > 0$, and n large enough*

$$\begin{aligned} P\Big\{L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - L^* > \beta\Big\} \;\; \leq \;\; & 2\,|\mathcal{F}_{\varepsilon_m}|\,\exp\{-A_{\beta,\varepsilon_m}\ell\} + c_{25}\,\exp\big\{-B_{\beta,\varepsilon_m} m\big\} \\ & + \ell\,\big|\mathcal{F}_{\varepsilon}\big|\cdot\Big(c_{26}\,\exp\big\{-C_{\beta,\varepsilon_m} mh^d\big\} + c_{27}\,\exp\big\{-c_{28}\,mh^d\big\}\Big), \end{aligned}$$

whenever $\varphi^* \in \mathcal{F}$, where A_{β,ε_m}, B_{β,ε_m}, and C_{β,ε_m} are positive constants depending on β and ϵ_m through the positive quantity $t_m := (\beta - \kappa_0\,\varepsilon_m)$ only, but not depending on m or β directly, and κ_0 is as in (22). Furthermore, c_{25}, c_{26}, c_{27}, and c_{28} are positive constants not depending on m, ℓ, β, or ε_m.

Theorem 3, which provides exponential bounds on the performance of the error of the proposed classifier, can be viewed as a more general version of the classical result of Devroye and Krzyzak for kernel classifiers (see Theorem 10.1 of Devroye et al. (1996)). Additionally, Theorem 3 in conjunction with the Borel-Cantelli lemma yields the strong optimality of the proposed classifier. More specifically, we have the following corollary

Corollary 1 *Suppose that the conditions of Theorem 3 hold. If $\varepsilon_m \downarrow 0$, as $n \to \infty$, and*

$$\ell^{-1}\,\log|\mathcal{F}_{\varepsilon_m}| \to 0, \qquad (mh^d)^{-1}\,\log\ell \to 0, \qquad \textit{and} \qquad (mh^d)^{-1}\,\log|\mathcal{F}_{\varepsilon_m}| \to 0,$$

then $\widehat{\psi}_{n,\widehat{\varphi}}$ is strongly optimal, i.e.,

$$L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - L^* \longrightarrow^{a.s.} 0.$$

Remark 3 *The choice of the bandwidth h is always important in practice. It is well-understood that the optimal bandwidth that minimizes quantities such as the MISE or ISE is not necessarily optimal in kernel classification (in the sense of minimizing the classification error); see, for example, Devroye et al. (1996; Sec. 25.9). In fact, a counter-example is presented in Theorem 25.9 of the cited monograph, where it is shown that the optimal bandwidth based on the MISE can result in large misclassification errors. As argued in Chapter 25 of the cited monograph, the optimal bandwidth h_{opt} is the one that minimizes the error $L_n(\widehat{\psi}_{n,\widehat{\varphi}})$ in (20) which is unfortunately always unknown; see Devroye et al. (1996; Sec. 25.1). In a similar vein, Hall and Kang (2005) noted that for kernel-based classification with univariate distributions and just two classes, the optimal bandwidth can be different for each class and its asymptotic magnitude can vary from terms of order $O(n^{-1/5})$ to $O(n^{-1/9})$ depending on the conditions imposed on the relationship between higher order derivatives of the marginal densities. Furthermore, their results show that there are no closed form expression for any one of their bandwidths. These issues are further compounded by the fact that finding a data-dependent bandwidth $\hat{h}_{opt}$ which is in some sense close to h_{opt} does not necessarily imply the closeness of the corresponding misclassification errors. Since, in classification, consistency (i.e., the convergence of $L_n(\widehat{\psi}_{n,\widehat{\varphi}})$ to L^*) is often the minimum requirement for a classifier, $\hat{h}_{opt}$ must be chosen in such a way that the resulting classifier will be consistent; see Devroye et al. (1996; p. 424). To that end, several methods have been proposed in the literature for finding data-dependent bandwidths that yield the minimum requirements; these methods include (i) minimizing the Apparent Error Rate that chooses $h > 0$ to minimize the error committed by the classifier on the data itself, and (ii) the cross-validation method; see Devroye et al. (1996; Ch. 25) for detail.*

3 Numerical Examples

In this section we perform some numerical work to study the performance of the proposed classifier.

EXAMPLE 1.
In this example, we consider the prediction of the class variable, $Y = 1$ or 0, based on the vector of covariates $\boldsymbol{X} \in \mathbb{R}^d$, where $d = 50$ and $d = 100$. If $Y = 1$ (i.e., class 1) then

$$\boldsymbol{X} \sim N_d(\mathbf{0}\,,\ c\Sigma),\ c > 0, \quad \text{where } \Sigma = (\sigma_{ij})\,, \quad \text{with } \sigma_{ij} = 2^{-|i-j|},\ \ i, j = 1, \ldots, d, \tag{23}$$

where $c > 0$ will be specified later. When $Y = 0$ (class 0) then $\boldsymbol{X}$ is a d-dim standard Cauchy random vector with independent components, i.e., the elements of $\boldsymbol{X}$ are independent standard Cauchy random variables. The unconditional class probabilities are $P\{Y = 1\} = 0.5 = P\{Y = 0\}$. The fact that both distributions are centered at zero makes the problem of classification rather challenging here. Next, we consider several response models.

Response Model A *[Nonignorable].*

$$\pi(\boldsymbol{x}, y) = \left[1 + \exp\left\{\alpha_0 + \textstyle\sum_{i=1}^d \alpha_i x_i + \gamma y\right\}\right]^{-1}.$$

Here, we consider four models under A. The choices of the coefficients below produce approximately 50% missing data:

Model A1.
For $d = 50$: $(\gamma, \alpha_0) = (0.5, -0.35)$, $\alpha_i = -0.11$ for $1 \le i \le \frac{d}{2}$, and $\alpha_i = 0.07$ for $\frac{d}{2} < i \le d$.
For $d = 100$: $(\gamma, \alpha_0) = (0.5, -0.35)$, $\alpha_i = -0.04$ for $1 \le i \le \frac{d}{2}$, and $\alpha_i = 0.06$ for $\frac{d}{2} < i \le d$.

Model A2.
For $d = 50$: $(\gamma, \alpha_0) = (1.5, -0.96)$, $\alpha_i = -0.03$ for $1 \le i \le \frac{d}{2}$, and $\alpha_i = 0.06$ for $\frac{d}{2} < i \le d$.
For $d = 100$: $(\gamma, \alpha_0) = (1.5, -0.98)$, $\alpha_i = -0.05$ for $1 \le i \le \frac{d}{2}$, and $\alpha_i = 0.06$ for $\frac{d}{2} < i \le d$.

Model A3.
For $d = 50$: $(\gamma, \alpha_0) = (2.5, -1.6)$, $\alpha_i = -0.03$ for $1 \le i \le \frac{d}{2}$, and $\alpha_i = 0.05$ for $\frac{d}{2} < i \le d$.
For $d = 100$: $(\gamma, \alpha_0) = (2.5, -1.6)$, $\alpha_i = -0.07$ for $1 \le i \le \frac{d}{2}$, and $\alpha_i = 0.08$ for $\frac{d}{2} < i \le d$.

Model A4.
For $d = 50$: $(\gamma, \alpha_0) = (5, -3.3)$, $\alpha_i = -0.04$ for $1 \le i \le \frac{d}{2}$, and $\alpha_i = 0.04$ for $\frac{d}{2} < i \le d$.
For $d = 100$: $(\gamma, \alpha_0) = (5, -3.4)$, $\alpha_i = -0.08$ for $1 \le i \le \frac{d}{2}$, and $\alpha_i = 0.11$ for $\frac{d}{2} < i \le d$.

The next three models do not satisfy the response probability assumption (1); they are intentionally included to examine the robustness of the proposed classifiers against departures from model assumptions.

Response Model B *[Nonignorable with Interaction].*

$$\pi_\varphi(\boldsymbol{x}, y) = \left[1 + \exp\{\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_1 y + \gamma y\}\right]^{-1}.$$

In each of the following four models under B, the coefficients are chosen to produce approximately 50% missing data (for both $d = 50$ and 100)

Model B1. $(\gamma, \alpha_0) = (0.5, -0.3)$, $\alpha_1 = 0.4$, $\alpha_2 = -0.3$, $\alpha_3 = -0.5$, and $\alpha_4 = -0.2$.

Model B2. $(\gamma, \alpha_0) = (1.5, -0.8)$, $\alpha_1 = 0.4$, $\alpha_2 = 0.1$, $\alpha_3 = -0.7$, and $\alpha_4 = -0.2$.

Model B3. $(\gamma, \alpha_0) = (2.5, -1.1)$, $\alpha_1 = -0.7$, $\alpha_2 = -0.8$, $\alpha_3 = -1.9$, and $\alpha_4 = -0.2$.

Model B4. $(\gamma, \alpha_0) = (5, -1.65)$, $\alpha_1 = -0.6$, $\alpha_2 = -0.7$, $\alpha_3 = -1.3$, and $\alpha_4 = -0.2$.

Response Model C *[Nonignorable Probit Model].*

$$\pi_\varphi(\boldsymbol{x}, y) = P\big\{Z \ \leq \ \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \gamma y\big\},$$

where $Z \sim N(0,1)$. The coefficients below yield about 50% missing rate (for both $d = 50$ and 100).

Model C1. $(\gamma, \alpha_0) = (0.5, -0.5)$, $\alpha_1 = -0.7$, $\alpha_2 = 0.2$, $\alpha_3 = -0.2$.

Model C2. $(\gamma, \alpha_0) = (1.5, -1.2)$, $\alpha_1 = -0.9$, $\alpha_2 = 0.3$, $\alpha_3 = -0.2$.

Model C3. $(\gamma, \alpha_0) = (2.5, -1.7)$, $\alpha_1 = 0.7$, $\alpha_2 = 0.1$, $\alpha_3 = -0.5$.

Model C4. $(\gamma, \alpha_0) = (5, -2.0)$, $\alpha_1 = 0.9$, $\alpha_2 = -0.3$, $\alpha_3 = -1.2$.

Response Model D *[Nonignorable sinusoidal].*

$$\pi(\boldsymbol{x}, y) = \Big[1 + \exp\big\{\alpha_0 + \textstyle\sum_{i=1}^d \alpha_i x_i\big\} \cdot \varphi(y)\Big]^{-1}, \ \text{ where } \varphi(y) \ = \ \sin(\gamma\pi y) + 0.1\,, \quad \gamma \in [0,\, 1/2],$$

which is as in (8). It is also the same as (1) but with $\exp\{\gamma y\}$ replace by the above function $\varphi(y)$. Here, we consider two models under D, each of which produces approximately 50% missing data:

Model D1.
For $d = 50$: $(\gamma, \alpha_0) = (0.2, 1.05)$, $\alpha_i = -0.05$ for $1 \leq i \leq \frac{d}{2}$, and $\alpha_i = 0.06$ for $\frac{d}{2} < i \leq d$.
For $d = 100$: $(\gamma, \alpha_0) = (0.2, 1.1)$, $\alpha_i = -0.06$ for $1 \leq i \leq \frac{d}{2}$, and $\alpha_i = 0.07$ for $\frac{d}{2} < i \leq d$.

Model D2.
For $d = 50$: $(\gamma, \alpha_0) = (0.4, 0.75)$, $\alpha_i = -0.08$ for $1 \leq i \leq \frac{d}{2}$, and $\alpha_i = 0.09$ for $\frac{d}{2} < i \leq d$.
For $d = 100$: $(\gamma, \alpha_0) = (0.4, 0.74)$, $\alpha_i = -0.06$ for $1 \leq i \leq \frac{d}{2}$, and $\alpha_i = 0.04$ for $\frac{d}{2} < i \leq d$.

Four classifiers are considered here: the proposed kernel classifier in (19), where γ is estimated using the proposed estimator as well as the approach of Kim and Yu (2011). These are denoted by $\widehat{\psi}_{n,\widehat{\varphi}}$ and $\widehat{\psi}_{\text{KY}}$, respectively. We also consider the classifier of Mojirsheibani and Reese (2017), denoted by $\widehat{\psi}_{\text{MR}}$, and the *complete-case* kernel classifier, $\widehat{\psi}_{cc}$, that only uses the complete cases. To construct various classifiers, we consider two sample sizes: $n =$ 200 and 400. As for the external data needed to construct the estimator of Kim and Yu, we employed 20% of the simulated missing Y_i values. Of course, in practice, one does not have external values of real data sets. For the new estimator of γ (equivalently, $\varphi(y) = \exp\{\gamma y\}$) we used the data-splitting approach outlined in Sec. 2.2 with $m = 0.7n$ and $\ell = 0.3n = n - m$, where γ was selected to minimize (15) over a grid of equally-spaced values of γ in $[-M, M]$. Here, we took $M = 20$ but a smaller value such as $M = 5$ would have been sufficient.

As for the choice of the kernel, we used the standard Gaussian kernel, where the bandwidth was determined using the cross-validation approach to minimize the empirical error of the classifier; see Remark 3 for this choice. Next, the misclassification error of each classifier is estimated based on an additional sample of 2000 observations generated in the same way as the original data (with 1000 from each class) and used as our "test" sample. The entire above process was repeated 500

Table 1: Misclassification errors for Example 1 when the dimension of $\boldsymbol{x}$ is 50 and $c=8$ in (23).

Missing Response	n	Model	$\widehat{\psi}_{n,\widehat{\varphi}}$	$\widehat{\psi}_{\mathrm{KY}}$	$\widehat{\psi}_{\mathrm{MR}}$	$\widehat{\psi}_{cc}$
A	200	A1	0.1268 (0.0023)	0.1354 (0.0025)	0.3090 (0.0013)	0.1413 (0.0025)
		A2	0.1359 (0.0027)	0.1497 (0.0028)	0.3471 (0.0016)	0.1540 (0.0028)
		A3	0.1564 (0.0029)	0.1566 (0.0029)	0.3778 (0.0011)	0.1602 (0.0028)
		A4	0.2157 (0.0036)	0.2232 (0.0037)	0.4228 (0.0012)	0.2325 (0.0037)
	400	A1	0.1069 (0.0019)	0.1117 (0.0015)	0.3047 (0.0010)	0.1208 (0.0015)
		A2	0.1162 (0.0016)	0.1229 (0.0017)	0.3297 (0.0007)	0.1332 (0.0017)
		A3	0.1065 (0.0013)	0.1163 (0.0016)	0.3503 (0.0007)	0.1283 (0.0016)
		A4	0.1839 (0.0021)	0.1847 (0.0023)	0.4216 (0.0011)	0.1993 (0.0023)
B	200	B1	0.1278 (0.0023)	0.1365 (0.0024)	0.3088 (0.0014)	0.1431 (0.0025)
		B2	0.1363 [0.0029]	0.1499 (0.0026)	0.3501 [0.0015]	0.1590 (0.0029)
		B3	0.1622 (0.0031)	0.1602 (0.0031)	0.3782 (0.0012)	0.1704 (0.0030)
		B4	0.2172 (0.0038)	0.2247 (0.0039)	0.4231 (0.0013)	0.2387 (0.0038)
	400	B1	0.1071 (0.0018)	0.1118(0.0015)	0.3046 (0.0009)	0.1201 (0.0016)
		B2	0.1154 (0.0015)	0.1219 (0.0016)	0.3291 (0.0007)	0.1330 (0.0018)
		B3	0.1070 (0.0014)	0.1177 (0.0017)	0.3509 (0.0007)	0.1294 (0.0017)
		B4	0.1848 (0.0022)	0.1858 (0.0023)	0.4221 (0.0010)	0.2049 (0.0024)
C	200	C1	0.1294 (0.0018)	0.1378 (0.0017)	0.3093 [0.0009]	0.1469 (0.0016)
		C2	0.1379 (0.0017)	0.1531 (0.0015)	0.3519 (0.0007)	0.1664 (0.0012)
		C3	0.1834 (0.0014)	0.1907 (0.0015)	0.3592 (0.0006)	0.2115 (0.0016)
		C4	0.2281 (0.0016)	0.2308 (0.0017)	0.4238 (0.0007)	0.2499 (0.0018)
	400	C1	0.1252 (0.0016)	0.1309 (0.0014)	0.3066 (0.0009)	0.1402 (0.0015)
		C2	0.1168 (0.0014)	0.1232 (0.0013)	0.3302 (0.0006)	0.1348 (0.0011)
		C3	0.1012 (0.0010)	0.1128 (0.0011)	0.3243 (0.0005)	0.1426 (0.0013)
		C4	0.1935 (0.0011)	0.1942 (0.0012)	0.4119 (0.0006)	0.2092 (0.0011)
D	200	D1	0.1342 (0.0025)	0.1476 (0.0030)	0.3548 (0.0015)	0.1598 (0.0031)
		D2	0.1485 (0.0031)	0.1469 (0.0027)	0.3456 (0.0015)	0.1592 (0.0028)
	400	D1	0.1159 (0.0022)	0.1202 (0.0023)	0.3448 (0.0011)	0.1356 (0.0024)
		D2	0.1397 (0.0026)	0.1417 (0.0023)	0.3449 (0.0010)	0.1543 (0.0024)

times and the average errors, over the 500 Monte Carlo runs, were computed for each classifier, each model, and each sample size n.

The results are summarized in Table 1 for 50-dim covariates with $c=8$ in (23). This value of c inflates the normal variances, thus making it more challenging to discriminate between Normal and Cauchy populations. The first two classifiers are based on the proposed kernel method, however $\widehat{\psi}_{n,\widehat{\varphi}}$ uses the data-splitting estimator of γ outlined in Sec. 2.2, whereas $\widehat{\psi}_{\mathrm{KY}}$ is based on the estimator $\widehat{\gamma}$ of Kim and Yu (2011) which requires external data. As Table 1 shows, the error of $\widehat{\psi}_{n,\widehat{\varphi}}$ is slightly better than that of $\widehat{\psi}_{\mathrm{KY}}$; *this is despite the fact that $\widehat{\psi}_{KY}$ uses some external data as well (it uses the values of 20% of the missing Y_i's).* Clearly, external data is not available in practice when dealing with real data sets. As the table shows, both of these classifiers perform better than $\widehat{\psi}_{\mathrm{MR}}$ (which requires the MAR assumption) and the *complete-case* kernel classifier, $\widehat{\psi}_{cc}$. These conclusions hold for the four response models A, B, C, and D. Table 2 presents the same analysis for the case of 100-dim covariates. Here we took $c=15$ in (23) in order to make the task of classification much more challenging (the variance of the normals is 15 here). The conclusion based on this table is essentially the same as before: the proposed kernel classifier $\widehat{\psi}_{n,\widehat{\varphi}}$ works well

Table 2: Misclassification errors for Example 1 when the dimension of $\boldsymbol{x}$ is 100 and $c = 8$ in (23).

Missing Response	n	Model	$\widehat{\psi}_{n,\widehat{\varphi}}$	$\widehat{\psi}_{\mathrm{KY}}$	$\widehat{\psi}_{\mathrm{MR}}$	$\widehat{\psi}_{cc}$
A	200	A1	0.1172 (0.0037)	0.1320 (0.0036)	0.3169 (0.0023)	0.1412 (0.0035)
		A2	0.1077 (0.0034)	0.1211 (0.0041)	0.3138 (0.0022)	0.1319 (0.0042)
		A3	0.1442 (0.0051)	0.1466 (0.0046)	0.3359 (0.0025)	0.1587 (0.0047)
		A4	0.1865 (0.0038)	0.1874 (0.0036)	0.3551 (0.0015)	0.1988 (0.0038)
	400	A1	0.1038 (0.0024)	0.1121 (0.0022)	0.3180 (0.0015)	0.1192 (0.0022)
		A2	0.0960 (0.0023)	0.1211 (0.0028)	0.3369 (0.0017)	0.1230 (0.0029)
		A3	0.1193 (0.0040)	0.1279 (0.0037)	0.3253 (0.0016)	0.1314 (0.0038)
		A4	0.1685 (0.0033)	0.1777 (0.0032)	0.3599 (0.0014)	0.1851 (0.0034)
B	200	B1	0.1174 (0.0036)	0.1321 (0.0035)	0.3174 (0.0023)	0.1385 (0.0034)
		B2	0.1080 (0.0038)	0.1225 (0.0043)	0.3140 (0.0024)	0.1274 (0.0045)
		B3	0.1871 (0.0054)	0.1882 (0.0048)	0.3560 (0.0027)	0.1944 (0.0047)
		B4	0.1882 (0.0039)	0.1893 (0.0038)	0.3571 (0.0016)	0.1984 (0.0036)
	400	B1	0.1039 (0.0025)	0.1123 (0.0023)	0.3179 (0.0015)	0.1174 (0.0022)
		B2	0.0958 (0.0022)	0.1198 (0.0026)	0.3365 (0.0018)	0.1245 (0.0027)
		B3	0.1198 (0.0041)	0.1287 (0.0039)	0.3258 (0.0016)	0.1337 (0.0039)
		B4	0.1712 (0.0037)	0.1805 (0.0035)	0.3611 (0.0014)	0.1921 (0.0033)
C	200	C1	0.1251 (0.0033)	0.1415 (0.0032)	0.3198 (0.0023)	0.1483 (0.0033)
		C2	0.1123 (0.0024)	0.1266 (0.0022)	0.3149 (0.0019)	0.1362 (0.0021)
		C3	0.1914 (0.0028)	0.1918 (0.0023)	0.3562 (0.0010)	0.1952 (0.0023)
		C4	0.1927 (0.0023)	0.1930 (0.0021)	0.3568 (0.0008)	0.2012 (0.0022)
	400	C1	0.1124 (0.0024)	0.1305 (0.0023)	0.3182 (0.0014)	0.1351 (0.0024)
		C2	0.1087 (0.0017)	0.1196 (0.0016)	0.3125 (0.0012)	0.1267 (0.0016)
		C3	0.1828 (0.0018)	0.1830 (0.0017)	0.3517 (0.0007)	0.1912 (0.0017)
		C4	0.1845 (0.0016)	0.1839 (0.0016)	0.3524 (0.0006)	0.1987 (0.0015)
D	200	D1	0.1433 (0.0037)	0.1569 (0.0034)	0.3482 (0.0013)	0.1697 (0.0035)
		D2	0.1683 (0.0051)	0.1691 (0.0049)	0.3437 (0.0024)	0.1817 (0.0049)
	400	D1	0.1282 (0.0032)	0.1344 (0.0029)	0.3402 (0.0010)	0.1487 (0.0030)
		D2	0.1455 (0.0041)	0.1594 (0.0038)	0.3431 (0.0014)	0.1731 (0.0037)

compared to the existing methods and it does so without requiring additional external data.

EXAMPLE 2. *[The German Credit Data].*
This real data set consisting of 1000 individuals, 700 of whom have been identified as having "good credit", i.e., class 1, and the remaining 300 have "bad credit", which is class 0. A total of 24 numerical covariates are associated with each person. A full description of this data set can be found in the UCI repository of machine learning data sets at `https://archive.ics.uci.edu/ml/index.php`. To perform the analysis, we randomly selected 300 of the 1000 observations to be set aside as a test sequence to estimate the misclassification error of each classifier.

To compare the effectiveness of various methods, we deliberately deleted some of the y values according to the four response models discussed in Example 1. More specifically, for $j = 1, \ldots, 24$, we took $(\gamma, \alpha_0, \alpha_j)$ to be $(0.5, 2.4, -0.023)$ for Model A1, $(1.5, 1.2, -0.020)$ for Model A2, $(2.5, 0.7, -0.022)$ for Model A3, and $(5, -1.54, -0.021)$ for Model A4. Similarly, we took $(\gamma, \alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$ to be $(0.5, -0.15, -0.019, -0.019, -0.019, 0.14)$ for Model B1, $(1.5, 0.13, -0.08, -0.08, -0.08, 0.33)$ for Model B2, $(2.5, 2.1, -0.22, -0.22, -0.22, 0.67)$ for Model B3, $(5, 1.7, -0.28,$

$-0.28, -0.28, 0.56)$ for Model B4. For the response models under C we took $(\gamma, \alpha_0, \alpha_1, \alpha_2, \alpha_3)$ to be $(0.5, 2.4, -0.11, -0.11, -0.11)$ for Model C1, $(1.5, 1.85, -0.12, -0.12,\ -0.12)$ for Model C2, $(2.5, 1.53, -0.14, -0.14, -0.14)$ for Model C3, and $(5, 1.9, -0.23, -0.23, -0.23)$ for Model C4. Finally, for $1 \leq j \leq 24$, we took $(\gamma, \alpha_0, \alpha_j)$ to be $(0.2, -2.6, 0.03)$ for Model D1 and $(0.4, -2.95, -0.03)$ for Model D2. In each case, these values produced approximately 50% missing rates. Next, each classifier was constructed based on the sample of size 700 (using the same approach as in Example 1) and tested on the remaining sequence of 300 data values. Repeating this process 500 times, the average errors (over 500) and their standard errors were computed; these are reported in Table 3. The empirical effectiveness of the proposed classifier follows from the numerical results under $\widehat{\psi}_{n,\widehat{\varphi}}$ and $\widehat{\psi}_{\mathrm{KY}}$ of Table 3. Furthermore, unlike $\widehat{\psi}_{\mathrm{KY}}$, the classifier $\widehat{\psi}_{n,\widehat{\varphi}}$ does not require any external data in terms of access to some of the missing y values.

Table 3: Misclassification errors for the *German Credit* data of Example 2.

Missing					
Response	Model	$\widehat{\psi}_{n,\widehat{\varphi}}$	$\widehat{\psi}_{\mathrm{KY}}$	$\widehat{\psi}_{\mathrm{MR}}$	$\widehat{\psi}_{cc}$
A	A1	0.2746 (0.0024)	0.2798 (0.0023)	0.3810 (0.0014)	0.3015 (0.0024)
	A2	0.2842 (0.0026)	0.2881 (0.0026)	0.3918 (0.0015)	0.3113 (0.0025)
	A3	0.2984 (0.0030)	0.2998 (0.0031)	0.4152 (0.0014)	0.3275 (0.0030)
	A4	0.3261 (0.0038)	0.3293 (0.0037)	0.4325 (0.0016)	0.3718 (0.0037)
B	B1	0.2980 (0.0025)	0.2990 (0.0024)	0.3823 (0.0016)	0.3124 (0.0025)
	B2	0.2988 (0.0028)	0.2997 (0.0028)	0.3983 (0.0017)	0.3195 (0.0027)
	B3	0.3521 (0.0030)	0.3574 (0.0031)	0.4195 (0.0015)	0.3755 (0.0031)
	B4	0.3752 (0.0031)	0.3811 (0.0032)	0.4410 (0.0015)	0.4072 (0.0032)
C	C1	0.3185 (0.0019)	0.3214 (0.0017)	0.3945 (0.0008)	0.3450 (0.0018)
	C2	0.3261 (0.0016)	0.3288 (0.0016)	0.3998 (0.0007)	0.3375 (0.0017)
	C3	0.3718 (0.0015)	0.3807 (0.0014)	0.4153 (0.0007)	0.3989 (0.0014)
	C4	0.3850 (0.0017)	0.3895 (0.0017)	0.4340 (0.0008)	0.4160 (0.0018)
D	D1	0.3449 (0.0027)	0.3580 (0.0026)	0.3978 (0.0015)	0.3746 (0.0027)
	D2	0.3618 (0.0035)	0.3712 (0.0029)	0.4109 (0.0016)	0.3955 (0.0031)

Remark 4 *In the case of high-dimensional $\boldsymbol{x}$, one may face the curse of dimensionality in estimating $\eta(\boldsymbol{x})$ that often occurs in multivariate kernel estimation. It is well-understood that the curse of dimensionality manifests itself in the error of estimation, as seen in the expected L^2 error of kernel estimators, which is of order $\mathcal{O}(n^{-2/(d+2)})$ where d is the dimension (Györfi et al. (2002; Theorem 5.2)). To overcome this issue, lately substantial efforts have been made to develop ways for evading the curse of dimensionality. One main approach is through the use of vine copulas as proposed by Nagler and Czado (2016), Kraus and Czado (2017), and Noh et al. (2013).*

4 Proofs

To prove our main results, we start by stating a number of lemmas.

Lemma 1 *Let μ be any probability measure on the Borel sets of $\mathbb{R}^d$. If K is a regular kernel then* *there is a positive constant $\rho(K)$, depending on the kernel K but not n, such that for every $h > 0$*

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d} \int \frac{K((\boldsymbol{x}-\boldsymbol{u})/h)}{E[K((\boldsymbol{x}-\boldsymbol{X})/h)]}\,\mu(d\boldsymbol{x}) \;\le\; \rho(K)\,.$$

PROOF OF LEMMA 1

The proof can be found in, for example, Devroye and Krzyżak (1989; Lemma 1).

□

Lemma 2 *Let $(\boldsymbol{X},V),(\boldsymbol{X}_1,V_1),\ldots,(\boldsymbol{X}_n,V_n)$ be iid $\mathbb{R}^d\times[-L,L]$-valued random vectors, $0<L<\infty$, and define $\breve{m}_n(\boldsymbol{x})=\sum_{i=1}^n V_i\,K\big((\boldsymbol{x}-\boldsymbol{X}_i)/h\big)/\big\{nE[K((\boldsymbol{x}-\boldsymbol{X})/h)]\big\}$, where K is a regular kernel. If $h\to 0$ and $nh^d\to\infty$, as $n\to\infty$, then for every $\varepsilon>0$ and large enough n,*

$$P\left\{\int \Big|\breve{m}_n(\boldsymbol{x})-E[V|\boldsymbol{X}=\boldsymbol{x}]\Big|\mu(d\boldsymbol{x})>\varepsilon\right\} \;\le\; \exp\left\{-n\varepsilon^2/\big(64L^2\rho^2(K)\big)\right\}$$

where μ is the probability measure of $\boldsymbol{X}$ and $\rho(K)$ is as in Lemma 1.

PROOF OF LEMMA 2

The proof of this lemma appears in Györfi et al. (2002; Lemma 23.9).

Lemma 3 *Consider the random pair $(\boldsymbol{X},Y)\in\mathbb{R}^d\times\mathbb{R}$ where Y could be nonignorably missing according to (1). Let φ be any map of the form $\varphi:\mathbb{R}\to(0,\infty)$. Then, when the expectations exist, we have*

$$E[Y|\boldsymbol{X}=\boldsymbol{x}] \;=\; E[\delta Y|\boldsymbol{X}=\boldsymbol{x}]+\frac{E[\delta Y\,\varphi(Y)|\boldsymbol{X}=\boldsymbol{x}]}{E[\delta\,\varphi(Y)|\boldsymbol{X}=\boldsymbol{x}]}\cdot E[1-\delta|\boldsymbol{X}=\boldsymbol{x}], \tag{24}$$

PROOF OF LEMMA 3

The proof is straightforward and will not be given. □

Lemma 4 *Let $\mathcal{R}(\boldsymbol{x};\varphi_1)$ and $\mathcal{R}(\boldsymbol{x};\varphi_2)$ be as in (11), where $\varphi_1,\varphi_2:\,[0,1]\to(0,B]$ for some positive number B. Then, under assumption A5, one has*

$$E\Big|\mathcal{R}(\boldsymbol{X};\varphi_1)-\mathcal{R}(\boldsymbol{X};\varphi_2)\Big| \;\le\; c_7\cdot\sup_{0\le y\le 1}\big|\varphi_1(y)-\varphi_2(y)\big|,$$

where c_7 is a positive constant.

PROOF OF LEMMA 4

Define the quantities $S_j(\boldsymbol{X})=E[\delta\,Y\varphi_j(Y)|\boldsymbol{X}]$ and $T_j(\boldsymbol{X})=E[\delta\,\varphi_j(Y)|\boldsymbol{X}]$, $j=1,2$. Then it is straightforward to see that

$$\begin{aligned}\Big|\mathcal{R}(\boldsymbol{X};\varphi_1)-\mathcal{R}(\boldsymbol{X};\varphi_2)\Big| &= E[1-\delta|\boldsymbol{X}]\cdot\left|\frac{S_1(\boldsymbol{X})}{T_1(\boldsymbol{X})}-\frac{S_2(\boldsymbol{X})}{T_2(\boldsymbol{X})}\right|\\ &= \left|\frac{-S_1(\boldsymbol{X})}{T_1(\boldsymbol{X})}\cdot\frac{T_1(\boldsymbol{X})-T_2(\boldsymbol{X})}{T_2(\boldsymbol{X})}+\frac{S_1(\boldsymbol{X})-S_2(\boldsymbol{X})}{T_2(\boldsymbol{X})}\right|\cdot E\big[1-\delta|\boldsymbol{X}\big]\\ &\le \frac{1}{T_2(\boldsymbol{X})}\left(\Big|T_1(\boldsymbol{X})-T_2(\boldsymbol{X})\Big|+\Big|S_1(\boldsymbol{X})-S_2(\boldsymbol{X})\Big|\right).\end{aligned} \tag{25}$$

However, $T_2(\boldsymbol{X}) \geq \varphi_{00} > 0$, almost surely, by assumption A5. Furthermore,

$$|S_1(\boldsymbol{X}) - S_2(\boldsymbol{X})| \;\leq\; E\left[|\delta\, Y| \cdot \big|\varphi_1(Y) - \varphi_2(Y)\big|\,\big|\boldsymbol{X}\right] \;\leq\; \sup_{0\leq y\leq 1} \big|\varphi_1(y) - \varphi_2(y)\big|$$

Similarly, $|T_1(\boldsymbol{X}) - T_2(\boldsymbol{X})| \leq \sup_{0\leq y\leq 1} \big|\varphi_1(y) - \varphi_2(y)\big|$. Lemma 4 now follows from these bounds together with (25).

□

Lemma 5 *Let $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ be iid $\mathbb{R}^d \times [-L, L]$-valued random vectors, where Y could be nonignorably missing according to (1). Define the indicator variable $\delta_i = 0$ if Y_i is missing (otherwise $\delta_i = 1$). Also, let $\widehat{\gamma}$ be any estimator of γ in (1) and put*

$$\widehat{m}(\boldsymbol{x}) \;=\; \frac{\sum_{i=1}^n \delta_i Y_i K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} + \frac{\sum_{i=1}^n \delta_i Y_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n \delta_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} \times \frac{\sum_{i=1}^n (1-\delta_i) K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h)},$$

where K is a regular kernel. Suppose that assumption A1 holds. If $h \to 0$ and $nh^d \to \infty$, as $n \to \infty$, then for every $\varepsilon > 0$, every $1 \leq p < \infty$, and any distribution of $(\boldsymbol{X}, Y)$ satisfying $|Y| \leq L < \infty$,

$$P\left\{ \int \Big|\widehat{m}(\boldsymbol{x}) - E[Y|\boldsymbol{X}=\boldsymbol{x}]\Big|^p \mu(d\boldsymbol{x}) > \varepsilon \right\} \;\leq\; 4\, e^{-c_8 n} + 2\, e^{-c_9\, n\varepsilon^2} + 4\, P\left\{\big|\widehat{\gamma} - \gamma\big| > c_{10}\right\} \tag{26}$$

for n large enough, where μ is the probability measure of $\boldsymbol{X}$, and c_8, c_9, and c_{10} are positive constants not depending on n.

PROOF OF LEMMA 5.
Since, for every $p \geq 1$, $\big|\widehat{m}(\boldsymbol{x}) - E[Y|\boldsymbol{X}{=}\boldsymbol{x}]\big|^p \leq \big(\big|\widehat{m}(\boldsymbol{x})\big| + \big|E[Y|\boldsymbol{X} = \boldsymbol{x}]\big|\big)^{p-1} \big|\widehat{m}(\boldsymbol{x}) - E[Y|\boldsymbol{X} = \boldsymbol{x}]\big| \leq (3L)^{p-1} \cdot \big|\widehat{m}(\boldsymbol{x}) - E[Y|\boldsymbol{X} = \boldsymbol{x}]\big|$, it is sufficient to prove the lemma for the case of $p = 1$. The proof is along standard arguments and goes as follows. First observe that in view of Lemma 3, with $\varphi(y) = \exp\{\gamma y\}$, we have

$$\begin{aligned}
&\big|\widehat{m}(\boldsymbol{x}) - E[Y|\boldsymbol{X} = \boldsymbol{x}]\big| \\
&\leq \left| \frac{\sum_{i=1}^n \delta_i Y_i K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} - E[\delta Y|\boldsymbol{X} = \boldsymbol{x}] \right| \\
&\quad + \left| \frac{\sum_{i=1}^n \delta_i Y_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n \delta_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} \cdot \frac{\sum_{i=1}^n (1-\delta_i) K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} \right. \\
&\qquad\qquad \left. - \frac{E[\delta Y \exp\{\gamma Y\}|\boldsymbol{X} = \boldsymbol{x}]}{E[\delta \exp\{\gamma Y\}|\boldsymbol{X} = \boldsymbol{x}]} \cdot E[1-\delta|\boldsymbol{X} = \boldsymbol{x}] \right| \\
&=: |\mathbf{I}_n(\boldsymbol{x})| + |\mathbf{II}_n(\boldsymbol{x})|
\end{aligned} \tag{27}$$

By the results of Devroye and Krzyżak (1989), for every distribution of $(\boldsymbol{X}, \delta Y)$ with $|\delta Y| \leq L < \infty$ and every $\varepsilon > 0$, there is a positive constant b_1 depending on ε (but not n) such that for large enough n,

$$P\left\{ \int |\mathbf{I}_n(\boldsymbol{x})|\, \mu(d\boldsymbol{x}) > \frac{\varepsilon}{2} \right\} \leq e^{-b_1 n}. \tag{28}$$

Next, to deal with the term $|\mathbf{II}_n(\boldsymbol{x})|$ in (27), we note that since $\left|\frac{\sum_{i=1}^n \delta_i Y_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n \delta_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}\right| \leq \vee_{i=1}^n |Y_i| \leq L$, one finds

$$\begin{aligned}
|\mathbf{II}_n(\boldsymbol{x})| &= \left| \frac{\sum_{i=1}^n \delta_i Y_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n \delta_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} \cdot \left\{ \frac{\sum_{i=1}^n (1-\delta_i) K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} \right.\right. \\
&\qquad \left.\left. \pm E[1-\delta|\boldsymbol{X}=\boldsymbol{x}] \right\} - \frac{E[\delta Y \exp\{\gamma Y\}|\boldsymbol{X}=\boldsymbol{x}]}{E[\delta \exp\{\gamma Y\}|\boldsymbol{X}=\boldsymbol{x}]} \cdot E[1-\delta|\boldsymbol{X}=\boldsymbol{x}] \right| \\
&\leq \left| \frac{\sum_{i=1}^n \delta_i Y_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n \delta_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} - \frac{E[\delta Y \exp\{\gamma Y\}|\boldsymbol{X}=\boldsymbol{x}]}{E[\delta \exp\{\gamma Y\}|\boldsymbol{X}=\boldsymbol{x}]} \right| \\
&\qquad + L \cdot \left| \frac{\sum_{i=1}^n (1-\delta_i) K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h)} - E[1-\delta|\boldsymbol{X}=\boldsymbol{x}] \right| \\
&:= |\mathbf{II}_{n,1}(\boldsymbol{x})| + |\mathbf{II}_{n,2}(\boldsymbol{x})|. \qquad (29)
\end{aligned}$$

But, again by the results of Devroye and Krzyżak (1989), for every distribution of $(\boldsymbol{X}, \delta)$ and every $\varepsilon > 0$, there is a positive constant b_2 depending on ε such that for n large enough,

$$P\Big\{ \int |\mathbf{II}_{n,2}(\boldsymbol{x})|\, \mu(d\boldsymbol{x}) > \frac{\varepsilon}{4} \Big\} \leq e^{-b_2 n}. \qquad (30)$$

As for the term $|\mathbf{II}_{n,1}(\boldsymbol{x})|$ in (29), start by defining the quantities

$$\begin{aligned}
\widehat{\phi}_1(\boldsymbol{x}) &= \sum_{i=1}^n \delta_i Y_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h) \div \sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h) \qquad (31) \\
\widehat{\phi}_2(\boldsymbol{x}) &= \sum_{i=1}^n \delta_i \exp\{\widehat{\gamma} Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h) \div \sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h) \qquad (32) \\
\widetilde{\phi}_1(\boldsymbol{x}) &= \sum_{i=1}^n \delta_i Y_i \exp\{\gamma Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h) \div \sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h) \qquad (33) \\
\widetilde{\phi}_2(\boldsymbol{x}) &= \sum_{i=1}^n \delta_i \exp\{\gamma Y_i\} K((\boldsymbol{x}-\boldsymbol{X}_i)/h) \div \sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h) \qquad (34) \\
\phi_1(\boldsymbol{x}) &= E\big[\delta Y \exp\{\gamma Y\} \big| \boldsymbol{X}=\boldsymbol{x}\big] \qquad (35) \\
\phi_2(\boldsymbol{x}) &= E\big[\delta \exp\{\gamma Y\} \big| \boldsymbol{X}=\boldsymbol{x}\big]. \qquad (36)
\end{aligned}$$

Then it is straightforward to see that

$$\begin{aligned}
|\mathbf{II}_{n,1}(\boldsymbol{x})| &= \left| -\frac{\widehat{\phi}_1(\boldsymbol{x})}{\widehat{\phi}_2(\boldsymbol{x})} \cdot \frac{\widehat{\phi}_2(\boldsymbol{x}) - \phi_2(\boldsymbol{x})}{\phi_2(\boldsymbol{x})} + \frac{\widehat{\phi}_1(\boldsymbol{x}) - \phi_1(\boldsymbol{x})}{\phi_2(\boldsymbol{x})} \right| \\
&\leq (|\phi_2(\boldsymbol{x})|)^{-1} \left(L \left| \widehat{\phi}_2(\boldsymbol{x}) - \phi_2(\boldsymbol{x}) \right| + \left| \widehat{\phi}_1(\boldsymbol{x}) - \phi_1(\boldsymbol{x}) \right| \right) \\
&\qquad \text{(where we have used the fact that } \widehat{\phi}_1(\boldsymbol{x})/\widehat{\phi}_2(\boldsymbol{x}) \leq L) \\
&\leq \pi_{\min}^{-1} \exp\{|\gamma| L\} \cdot \left(L \left| \widehat{\phi}_2(\boldsymbol{x}) - \phi_2(\boldsymbol{x}) \right| + \left| \widehat{\phi}_1(\boldsymbol{x}) - \phi_1(\boldsymbol{x}) \right| \right),
\end{aligned}$$

where the last line follows since, by assumption A1, $\phi_2(\boldsymbol{X}) = E\big[E\big(\delta \exp\{\gamma Y\} \big| \boldsymbol{X}, Y\big) \big| \boldsymbol{X}\big] = E\big[\exp\{\gamma Y\} \cdot \pi(\boldsymbol{X}, Y) \big| \boldsymbol{X}\big] \geq \pi_{\min} \cdot \exp\{-|\gamma| L\}$. Therefore, for every $\varepsilon > 0$, one has

$$P\Big\{ \int |\mathbf{II}_{n,1}(\boldsymbol{x})|\, \mu(d\boldsymbol{x}) > \frac{\varepsilon}{4} \Big\} \quad \leq \quad P\Big\{ \int \left| \widehat{\phi}_1(\boldsymbol{x}) - \widetilde{\phi}_1(\boldsymbol{x}) \right| \mu(d\boldsymbol{x}) > \frac{\pi_{\min}\, \varepsilon}{16}\, e^{-|\gamma| L} \Big\}$$

$$
\begin{aligned}
&+P\Big\{\int \Big|\widetilde{\phi}_1(\boldsymbol{x})-\phi_1(\boldsymbol{x})\Big|\,\mu(d\boldsymbol{x}) > \frac{\pi_{\min}\,\varepsilon}{16}\,e^{-|\gamma|L}\Big\}\\
&+P\Big\{\int \Big|\widehat{\phi}_2(\boldsymbol{x})-\widetilde{\phi}_2(\boldsymbol{x})\Big|\,\mu(d\boldsymbol{x}) > \frac{\pi_{\min}\,\varepsilon}{16L}\,e^{-|\gamma|L}\Big\}\\
&+P\Big\{\int \Big|\widetilde{\phi}_2(\boldsymbol{x})-\phi_2(\boldsymbol{x})\Big|\,\mu(d\boldsymbol{x}) > \frac{\pi_{\min}\,\varepsilon}{16L}\,e^{-|\gamma|L}\Big\}\\
:=\;\; & \Delta_n(1)+\Delta_n(2)+\Delta_n(3)+\Delta_n(4). \qquad (37)
\end{aligned}
$$

Now, once again, by the results of Devroye and Krzyżak (1989), for n large enough,

$$
\Delta_n(2)\le e^{-b_3 n} \quad \text{and} \quad \Delta_n(4)\le e^{-b_4 n} \qquad (38)
$$

where b_3 and b_4 are positive constants depending on ε but not n. Next, to deal with the term $\Delta_n(1)$ in (37), observe that

$$
\begin{aligned}
\Big|\widehat{\phi}_1(\boldsymbol{x})-\widetilde{\phi}_1(\boldsymbol{x})\Big| \;\le\;& \left|\frac{\sum_{i=1}^n \delta_i Y_i\left(e^{\widehat{\gamma}Y_i}-e^{\gamma Y_i}\right)K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}{nE\big[K((\boldsymbol{x}-\boldsymbol{X})/h)\big]}\right|\\
&+\Bigg|\left[\sum_{i=1}^n \delta_i Y_i\left(e^{\widehat{\gamma}Y_i}-e^{\gamma Y_i}\right)K((\boldsymbol{x}-\boldsymbol{X}_i)/h)\right]\\
&\qquad\times\Big(\frac{1}{nE\big[K((\boldsymbol{x}-\boldsymbol{X})/h)\big]}-\frac{1}{\sum_{i=1}^n K((\boldsymbol{x}-\boldsymbol{X}_i)/h)}\Big)\Bigg|\\
=:\;\; & \big|U_{n,1}(\boldsymbol{x})\big|+\big|U_{n,2}(\boldsymbol{x})\big|. \qquad (39)
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\int \big|U_{n,1}(\boldsymbol{x})\big|\,\mu(d\boldsymbol{x}) \;\le\;& n^{-1}\sum_{i=1}^n \left|\delta_i Y_i\left(e^{\widehat{\gamma}Y_i}-e^{\gamma Y_i}\right)\right|\cdot\sup_{\boldsymbol{u}}\int\frac{K((\boldsymbol{x}-\boldsymbol{u})/h)}{nE\big[K((\boldsymbol{x}-\boldsymbol{X})/h)\big]}\,\mu(d\boldsymbol{x})\\
\le\;& n^{-1}L\,\rho(K)\sum_{i=1}^n\left|\delta_i\left(e^{\widehat{\gamma}Y_i}-e^{\gamma Y_i}\right)\right|, \quad \text{by Lemma 1.} \qquad (40)
\end{aligned}
$$

However, a one-term Taylor expansion gives $\big|e^{\widehat{\gamma}Y_i}-e^{\gamma Y_i}\big| = \big|Y_i\,\exp\{\widetilde{\gamma}Y_i\}\cdot(\widehat{\gamma}-\gamma)\big|\le L\,\exp\big\{|\widetilde{\gamma}-\gamma|L+\gamma Y_i\big\}\cdot\big|\widehat{\gamma}-\gamma\big|$, where $\widetilde{\gamma}$ is a point on the interior of the line segment joining $\widehat{\gamma}$ and γ. Therefore, for any constants $\varepsilon>0$ and $C>0$

$$
\begin{aligned}
&P\left\{\int \big|U_{n,1}(\boldsymbol{x})\big|\,\mu(d\boldsymbol{x}) > \frac{\pi_{\min}\,\varepsilon}{32\,e^{|\gamma|L}}\right\}\\
&\le\; P\left\{\left[\frac{L}{n}\,|\widehat{\gamma}-\gamma|\,\exp\big\{|\widetilde{\gamma}-\gamma|L\big\}\sum_{i=1}^n\delta_i e^{\gamma Y_i} \;>\; \frac{\pi_{\min}\,\varepsilon}{32L\,e^{|\gamma|L}\rho(K)}\right]\cap\Big[|\widehat{\gamma}-\gamma|\le C\Big]\right\}\\
&\qquad+\;P\Big\{|\widehat{\gamma}-\gamma|>C\Big\}\\
&\le\; nP\left\{\delta_1 e^{\gamma Y_1} > \frac{\pi_{\min}\,\varepsilon}{32CL^2e^{(C+|\gamma|)L}\rho(K)}\right\}+P\Big\{|\widehat{\gamma}-\gamma|>C\Big\}, \qquad (41)
\end{aligned}
$$

where, in arriving at (41), we have used the fact that $|\widetilde{\gamma}-\gamma| < |\widehat{\gamma}-\gamma|\le C$ holds on the set $\big\{|\widehat{\gamma}-\gamma|\le C\big\}$. Now, since $\delta_1\,e^{\gamma Y_1}\le e^{|\gamma|L}$, the first term on the right side of (41) becomes zero upon choosing $C>0$ to satisfy $(32CL^2e^{CL}\rho(K))^{-1}e^{-|\gamma|L}\,\pi_{\min}\,\varepsilon\ge e^{|\gamma|L}$. This choice of C yields

$$
P\left\{\int \big|U_{n,1}(\boldsymbol{x})\big|\,\mu(d\boldsymbol{x}) > \frac{\pi_{\min}\,\varepsilon}{32\,e^{|\gamma|L}}\right\} \;\le\; P\Big\{|\widehat{\gamma}-\gamma|>C\Big\}. \qquad (42)
$$

Next, we can handle the term $\left|U_{n,2}(\boldsymbol{x})\right|$ in (39) as follows. First note that

$$\left|U_{n,2}(\boldsymbol{x})\right| \leq L \cdot \max_{1\leq i\leq n}\left(\delta_i \left|e^{\widehat{\gamma} Y_i} - e^{\gamma Y_i}\right|\right) \cdot \left|\frac{\sum_{j=1}^{n} K((\boldsymbol{x}-\boldsymbol{X}_j)/h)}{nE\big[K((\boldsymbol{x}-\boldsymbol{X})/h)\big]} - 1\right|.$$

Therefore, using the arguments that led to (41), one finds for every $\varepsilon > 0$

$$\begin{aligned}
&P\left\{\int \left|U_{n,2}(\boldsymbol{x})\right| \mu(d\boldsymbol{x}) > \frac{\pi_{\min}\,\varepsilon}{32\, e^{|\gamma|L}}\right\}\\
&\leq\ P\left\{CLe^{CL} \max_{1\leq i\leq n}\left(\delta_i e^{\gamma Y_i}\right) \cdot \int \left|\frac{\sum_{j=1}^{n} K((\boldsymbol{x}-\boldsymbol{X}_j)/h)}{nE\big[K((\boldsymbol{x}-\boldsymbol{X})/h)\big]} - 1\right| \mu(d\boldsymbol{x})\ >\ \frac{\pi_{\min}\,\varepsilon}{32L\, e^{|\gamma|L}}\right\}\\
&\qquad +\ P\Big\{|\widehat{\gamma}-\gamma| > C\Big\}, \qquad \text{where } C \text{ is as in (42)}\\
&\leq\ \exp\left\{\frac{-n\pi_{\min}^{2}\,\varepsilon^{2}}{(32)^2(64)C^2L^4\rho(K)\cdot \exp\big\{2L\big(C+2|\gamma|\big)\big\}}\right\} + P\Big\{|\widehat{\gamma}-\gamma| > C\Big\}, \qquad (43)
\end{aligned}$$

for n large enough, by Lemma 2; it is the special case of Lemma 2 where $V_i \overset{a.s.}{=} 1$ for all $i = 1, \ldots, n$, and $E[V|\boldsymbol{X}=\boldsymbol{x}] = 1$. Putting together (43), (42), and (39), one arrives at

$$\Delta_n(1)\ \leq\ \exp\left\{\frac{-n\pi_{\min}^{2}\varepsilon^{2}}{(32)^2(64)C^2L^4\rho(K)\cdot \exp\big\{2L\big(C+2|\gamma|\big)\big\}}\right\} + 2P\Big\{|\widehat{\gamma}-\gamma| > C\Big\}, \qquad (44)$$

where C is as in (42). Similarly, it is straightforward to show that the term $\Delta_n(3)$ in (37) can also be bounded by (44). Now, Lemma 5 follows from this together with (44), (28), (29), (30), (37), and (38), where the constants c_8 and c_9 in Lemma 5 can be taken to be $c_8 = \min\{b_1, b_2, b_3, b_4\}$ and $c_9 = \big[(32)^2(64)C^2L^4\rho(K)\cdot\exp\big\{2L\big(C+2|\gamma|\big)\big\}\big]^{-1}\pi_{\min}^{2}$.

□

Lemma 6 *Let $\mathcal{R}(\boldsymbol{x};\varphi)$ be as in (11) for a known function $\varphi : [0,1] \to (0, B]$ for some $B < \infty$. Also, let $\widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi)$ be the kernel estimator defined by (14), where the kernel K in (14) is regular. Suppose that assumption A1 holds. If $h \to 0$ and $mh^d \to \infty$, as $n \to \infty$ (and thus $m \to \infty$), then for every $\epsilon > 0$, every $1 \leq p < \infty$, and n large enough*

$$P\left\{\int \left|\widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi) - \mathcal{R}(\boldsymbol{x};\varphi)\right|^p \mu(\boldsymbol{x})\ >\ \epsilon\right\}\ \leq\ 4\, e^{-a_0\, m},$$

where a_0 is a positive constant depending on ϵ but not m or ℓ.

PROOF OF LEMMA 6

The proof of this lemma is similar to (and, in fact, much easier than) that of Lemma 5 and will not be given; it is easier because the function φ appearing in $\widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi)$ is fixed instead of being an estimator.

PROOF OF THEOREM 1

Let $\eta(\boldsymbol{x})$ and $\widehat{\eta}(\boldsymbol{x})$ be as in (4) and (5), respectively. Then (see Devroye et al. (1996, Corollary 6.1))

$$L_n(\widehat{\psi}_n) - L(\psi_{\mathrm{B}})\ \leq\ 2\int \left|\widehat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x})\right| \mu(d\boldsymbol{x}).$$

Thus, for every $\varepsilon > 0$,

$$P\left\{L_n(\widehat{\psi}_n) - L(\psi_{\mathrm{B}}) > \varepsilon\right\} \;\le\; P\left\{\int \left|\widehat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x})\right| \mu(d\boldsymbol{x}) > \frac{\varepsilon}{2}\right\}.$$

Theorem 1 now follows from Lemma 5 with $p = 1$.

□

PROOF OF THEOREM 2

Let $L_n(\widehat{\psi}_{n,\widehat{\varphi}})$ and L^* be as given in (20). Furthermore, for any $\varphi \in \mathcal{F}$ define

$$L_m(\widehat{\psi}_{m,\varphi}) \;=\; P\left\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}) \neq Y \middle| \mathbb{D}_m\right\} \quad \text{and} \quad L(\psi_\varphi) \;=\; P\{\psi_\varphi(\boldsymbol{X}) \neq Y\}, \tag{45}$$

where ψ_φ is as in (10). Now let $\widehat{L}_\ell(\widehat{\psi}_{n,\widehat{\varphi}})$ be the weighted empirical error of the proposed classifier $\widehat{\psi}_{n,\widehat{\varphi}}$ (see (15)) and observe that

$$\begin{aligned}
L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - \inf_{\varphi\in\mathcal{F}_\varepsilon} L(\psi_\varphi) \;&=\; \Big[L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - \widehat{L}_\ell(\widehat{\psi}_{n,\widehat{\varphi}})\Big] + \Big[\widehat{L}_\ell(\widehat{\psi}_{n,\widehat{\varphi}}) - \inf_{\varphi\in\mathcal{F}_\varepsilon} L_m(\widehat{\psi}_{m,\varphi})\Big]\\
&\qquad + \Big[\inf_{\varphi\in\mathcal{F}_\varepsilon} L_m(\widehat{\psi}_{m,\varphi}) - \inf_{\varphi\in\mathcal{F}_\varepsilon} L(\psi_\varphi)\Big]\\
&=: \; R(1) + R(2) + R(3).
\end{aligned} \tag{46}$$

But

$$R(1) \;\le\; \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|L_m(\widehat{\psi}_{m,\varphi}) - \widehat{L}_\ell(\widehat{\psi}_{m,\varphi})\right| \tag{47}$$

where we have taken into account the fact that upon replacing $\widehat{\varphi}$ (which depends on both $\mathbb{D}_m$ and $\mathbb{D}_\ell$ as shown in (18)) by φ, the error term $L_n(\widehat{\psi}_{n,\widehat{\varphi}})$ reduces to $L_m(\widehat{\psi}_{m,\varphi})$. Next, let $\breve{\varphi} = \operatorname{argmin}_{\varphi\in\mathcal{F}_\varepsilon} L_m(\widehat{\psi}_{m,\varphi})$; here, $\breve{\varphi}$ depends on $\mathbb{D}_m$ because $\widehat{\psi}_{m,\varphi}$ does. Then

$$\begin{aligned}
R(2) \;&=\; \widehat{L}_\ell(\widehat{\psi}_{n,\widehat{\varphi}}) - L_m(\widehat{\psi}_{m,\breve{\varphi}})\\
&\le\; \widehat{L}_\ell(\widehat{\psi}_{m,\breve{\varphi}}) - L_m(\widehat{\psi}_{m,\breve{\varphi}})\\
&\qquad \text{(since by (18), } \widehat{L}_\ell(\widehat{\psi}_{n,\widehat{\varphi}}) \le \widehat{L}_\ell(\widehat{\psi}_{m,\breve{\varphi}}),\ \forall\breve{\varphi} \in \mathcal{F}_\varepsilon\,;\ \text{furthermore } \widehat{\psi}_{n,\breve{\varphi}} = \widehat{\psi}_{m,\breve{\varphi}})\\
&\le\; \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_\ell(\widehat{\psi}_{m,\varphi}) - L_m(\widehat{\psi}_{m,\varphi})\right|.
\end{aligned} \tag{48}$$

Therefore, in view of (47) and (48),

$$\left|R(1) + R(2)\right| \;\le\; 2 \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_\ell(\widehat{\psi}_{m,\varphi}) - L_m(\widehat{\psi}_{m,\varphi})\right|. \tag{49}$$

As for the term $R(3)$ in (46), we note that with $\mathcal{R}(\boldsymbol{x};\varphi)$ and $\widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi)$ as in (11) and (14), respectively, one has

$$\begin{aligned}
R(3) \;&=\; \inf_{\varphi\in\mathcal{F}_\varepsilon} L_m(\widehat{\psi}_{m,\varphi}) - L(\psi_{\varphi'}), \quad \text{where } \varphi' = \operatorname*{argmin}_{\varphi\in\mathcal{F}_\varepsilon} L(\psi_\varphi)\\
&\le\; L_m(\widehat{\psi}_{m,\varphi'}) - L(\psi_{\varphi'})\\
&\le\; 2\int \left|\widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi') - \mathcal{R}(\boldsymbol{x};\varphi')\right| \mu(d\boldsymbol{x}),
\end{aligned} \tag{50}$$

where (50) follows from the results in Devroye et al. (1996, Corollary 6.1). Here, μ is the probability measure of $\boldsymbol{X}$. Therefore, by (46), (49), and (50), for every $\beta > 0$,

$$\begin{aligned} P\left\{ L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - \inf_{\varphi \in \mathcal{F}_\varepsilon} L(\psi_\varphi) > \beta \right\} \;&\leq\; P\left\{ 2 \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \widehat{L}_\ell(\widehat{\psi}_{m,\varphi}) - L_m(\widehat{\psi}_{m,\varphi}) \right| \;>\; \frac{\beta}{2} \right\} \\ &\quad + P\left\{ \int \left| \widehat{\mathcal{R}}_m(\boldsymbol{x};\varphi') - \mathcal{R}(\boldsymbol{x};\varphi') \right| \mu(\boldsymbol{x}) \;>\; \frac{\beta}{4} \right\} \\ &:= \; S_{n,1} + S_{n,2}. \end{aligned} \tag{51}$$

But, taking $p = 1$ in Lemma 6, one has, for n large enough

$$S_{n,2} \;\leq\; 4\, e^{-a\, m}, \tag{52}$$

where a is a positive constant depending on β but not on m or ℓ. To deal with the term $S_{n,1}$, first observe that with π_φ and $\widehat{\pi}_\varphi$ given by (8) and (16), the fact that

$$\frac{\delta_i\, I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\}}{\widehat{\pi}(\boldsymbol{X}_i, Y_i)} = \frac{\delta_i I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\}}{\pi(\boldsymbol{X}_i, Y_i)} \;-\; \delta_i I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\} \left[\frac{1}{\pi_\varphi(\boldsymbol{X}_i, Y_i)} - \frac{1}{\widehat{\pi}_\varphi(\boldsymbol{X}_i, Y_i)} \right]$$

implies that

$$\begin{aligned} S_{n,1} \;\;\leq\;\; & P\left\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \ell^{-1} \sum_{i\colon (\boldsymbol{X}_i, Y_i, \delta_i) \in \mathbb{D}_\ell} \frac{\delta_i I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\}}{\pi_\varphi(\boldsymbol{X}_i, Y_i)} - P\left\{ \widehat{\psi}_{m,\varphi}(\boldsymbol{X}) \neq Y \middle| \mathcal{D}_m \right\} \right| \;>\; \frac{\beta}{8} \right\} \\ & +\; P\left\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \ell^{-1} \sum_{i\colon \boldsymbol{X}_i, Y_i, \delta_i \in \mathbb{D}_\ell} I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\} \left[\frac{1}{\pi_\varphi(\boldsymbol{X}_i, Y_i)} - \frac{1}{\widehat{\pi}_\varphi(\boldsymbol{X}_i, Y_i)} \right] \right| \;>\; \frac{\beta}{8} \right\} \\ :=\;\; & P_n(1) + P_n(2). \end{aligned} \tag{53}$$

But, for each $(\boldsymbol{X}_i, Y_i, \delta_i) \,\in\, \mathbb{D}_\ell$, one has

$$\begin{aligned} E\left[\frac{\delta_i I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\}}{\pi_\varphi(\boldsymbol{X}_i, Y_i)} \middle| \mathbb{D}_m \right] \;&=\; E\left[E\left(\frac{\delta_i I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\}}{\pi_\varphi(\boldsymbol{X}_i, Y_i)} \middle| \mathbb{D}_m, \boldsymbol{X}_i, Y_i \right) \middle| \mathbb{D}_m \right] \\ &=\; E\left[\frac{I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\}}{\pi_\varphi(\boldsymbol{X}_i, Y_i)} E\big(\delta_i \big| \mathbb{D}_m, \boldsymbol{X}_i, Y_i\big) \middle| \mathbb{D}_m \right] \\ &=\; P\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}) \neq Y | \mathbb{D}_m\} \;=:\; L_m(\widehat{\psi}_{m,\varphi}) \end{aligned}$$

where the last line follows from the definition of $\pi_\varphi(\boldsymbol{X}_i, Y_i)$ and the independence of $\mathcal{D}_m$ and $\delta_i \in \mathcal{D}_\ell$. Therefore, if we define the quantity

$$\widetilde{L}_\ell(\widehat{\psi}_{m,\varphi}) \;\;:=\;\; \ell^{-1} \sum_{i\colon (\boldsymbol{X}_i, Y_i, \delta_i) \in \mathbb{D}_\ell} \frac{\delta_i}{\pi_\varphi(\boldsymbol{X}_i, Y_i)} I\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i) \neq Y_i\}, \qquad \varphi \in \mathcal{F}_\varepsilon\,, \tag{54}$$

where $\pi_\varphi(\boldsymbol{x}, y)$ is as in (8), then the term $P_n(1)$ in (53) can be written as

$$P_n(1) \;\;=\;\; E\left[P\left\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \widetilde{L}_\ell(\widehat{\psi}_{m,\varphi}) - L_m(\widehat{\psi}_{m,\varphi}) \right| \;>\; \frac{\beta}{8} \middle| \mathbb{D}_m \right\} \right], \;\; \text{(where } \widetilde{L}_\ell(\widehat{\psi}_{m,\varphi}) \text{ is as in (54))}$$

$$
\begin{aligned}
&\leq\; E\left[\big|\mathcal{F}_\varepsilon\big|\cdot\sup_{\varphi\in\mathcal{F}_\varepsilon}P\left\{\Big|\widetilde{L}_\ell(\widehat{\psi}_{m,\varphi})-L_m(\widehat{\psi}_{m,\varphi})\Big|\;>\;\frac{\beta}{8}\,\Big|\,\mathbb{D}_m\right\}\right]\\
&\leq\; 2\,\big|\mathcal{F}_\varepsilon\big|\,e^{-\ell\,\beta^2/32},\qquad \text{(by Hoeffding's (1963) inequality).}
\end{aligned}\tag{55}
$$

Next, to deal with the term $P_n(2)$ in (53), observe that

$$
\begin{aligned}
&P_n(2)\\
&\leq\; \big|\mathcal{F}_\varepsilon\big|\cdot\sup_{\varphi\in\mathcal{F}_\varepsilon}P\left\{\ell^{-1}\left|\sum_{i:\,(\boldsymbol{X}_i,Y_i,\delta_i)\,\in\,\mathbb{D}_\ell} I\big\{\widehat{\psi}_{m,\varphi}(\boldsymbol{X}_i)\neq Y_i\big\}\left[\frac{1}{\widehat{\pi}_\varphi(\boldsymbol{X}_i,Y_i)}-\frac{1}{\pi_\varphi(\boldsymbol{X}_i,Y_i)}\right]\right|>\frac{\beta}{8}\right\}\\
&\qquad\quad \big(\text{where } \widehat{\pi}_\varphi(\boldsymbol{X}_i,Y_i) \text{ and } \pi_\varphi(\boldsymbol{X}_i,Y_i) \text{ are as in (16) and (8)}\big)\\
&\leq\; \big|\mathcal{F}_\varepsilon\big|\cdot\sup_{\varphi\in\mathcal{F}_\varepsilon}\sum_{i:\,(\boldsymbol{X}_i,Y_i,\delta_i)\,\in\mathbb{D}_\ell}E\left[P\left\{\left|\frac{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\,\in\mathbb{D}_m}(1-\delta_j)K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\,\in\mathbb{D}_m}\delta_j\,\varphi(Y_j)\,K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}\right.\right.\right.\\
&\qquad\qquad\qquad\qquad\qquad\left.\left.\left.-\frac{E[1-\delta_i|\boldsymbol{X}_i]}{E[\delta_i\varphi(Y_i)|\boldsymbol{X}_i]}\right|\varphi(Y_i)>\frac{\beta}{8}\,\right|\boldsymbol{X}_i,Y_i\right\}\right],
\end{aligned}\tag{56}
$$

where we have replaced the term $\exp\{g(\boldsymbol{x})\}$ by (17) in the definition of $\pi_\varphi(\boldsymbol{x},y)$ in (8). Now, define the quantities

$$
\begin{aligned}
Q_1(\boldsymbol{X}_i) &:= E\big[1-\delta_i\big|\boldsymbol{X}_i\big] \qquad\text{and}\qquad Q_2(\boldsymbol{X}_i):=E\big[\delta_i\varphi(Y_i)\big|\boldsymbol{X}_i\big]\\
\widehat{Q}_1(\boldsymbol{X}_i) &= \frac{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\in\mathbb{D}_m}(1-\delta_j)K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\in\mathbb{D}_m}K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}\\
\widehat{Q}_2(\boldsymbol{X}_i) &= \frac{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\in\mathbb{D}_m}\delta_j\,\varphi(Y_j)K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\in\mathbb{D}_m}K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}
\end{aligned}
$$

and observe that since $\big|\frac{\widehat{Q}_1(\boldsymbol{X}_i)}{\widehat{Q}_2(\boldsymbol{X}_i)}-\frac{Q_1(\boldsymbol{X}_i)}{Q_2(\boldsymbol{X}_i)}\big|\leq\big|\frac{\widehat{Q}_1(\boldsymbol{X}_i)}{\widehat{Q}_2(\boldsymbol{X}_i)}\big|\cdot\big|\frac{\widehat{Q}_2(\boldsymbol{X}_i)-Q_2(\boldsymbol{X}_i)}{Q_2(\boldsymbol{X}_i)}\big|+\big|\frac{\widehat{Q}_1(\boldsymbol{X}_i)-Q_1(\boldsymbol{X}_i)}{Q_2(\boldsymbol{X}_i)}\big|$, one can bound the inner conditional probability in (56) as follows

$$
\begin{aligned}
&P\left\{\left|\frac{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\,\in\mathbb{D}_m}(1-\delta_j)K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}{\sum_{j:\,(\boldsymbol{X}_j,Y_j,\delta_j)\,\in\mathbb{D}_m}\delta_j\,\varphi(Y_j)\,K((\boldsymbol{X}_i-\boldsymbol{X}_j)/h)}-\frac{E[1-\delta_i|\boldsymbol{X}_i]}{E[\delta_i\varphi(Y_i)|\boldsymbol{X}_i]}\right|\varphi(Y_i)\;>\;\frac{\beta}{8}\,\right|\boldsymbol{X}_i,Y_i\right\}\\
&\leq\; P\left\{\left|\frac{\widehat{Q}_1(\boldsymbol{X}_i)}{\widehat{Q}_2(\boldsymbol{X}_i)}\right|\cdot\Big|\widehat{Q}_2(\boldsymbol{X}_i)-Q_2(\boldsymbol{X}_i)\Big|\;>\;\frac{\varphi_{00}\,\beta}{16B}\,\right|\boldsymbol{X}_i,Y_i\right\}\\
&\qquad+\;P\left\{\Big|\widehat{Q}_1(\boldsymbol{X}_i)-Q_1(\boldsymbol{X}_i)\Big|>\frac{\varphi_{00}\,\beta}{16B}\,\right|\boldsymbol{X}_i,Y_i\right\}\\
&:=\;\Delta_{ni}(1)+\Delta_{ni}(2)
\end{aligned}\tag{57}
$$

where φ_{00} is as in assumption A5. Using standard arguments, it can be shown that under assumptions A1 – A5, for n (and thus m) large enough, one has

$$
\Delta_{ni}(2)\;\leq\;C_{10}\,\exp\big\{-C_{11}\,mh^d\beta^2\big\},\tag{58}
$$

where C_{10} and C_{11} are positive constants not depending on m or ℓ. To deal with $\Delta_{ni}(1)$ in (57), first observe that

$$
\Delta_{ni}(1)\quad\leq\quad P\left\{\frac{2(1-\pi_{\min})}{\varphi_{00}/2}\cdot\Big|\widehat{Q}_2(\boldsymbol{X}_i)-Q_2(\boldsymbol{X}_i)\Big|\;>\;\frac{\varphi_{00}\,\beta}{16B}\,\right|\boldsymbol{X}_i,Y_i\right\}
$$

$$
\begin{aligned}
& + P\left\{\widehat{Q}_1(\boldsymbol{X}_i) > 2(1-\pi_{\min})\Big|\boldsymbol{X}_i, Y_i\right\} + P\left\{\widehat{Q}_2(\boldsymbol{X}_i) < \varphi_{00}/2\Big|\boldsymbol{X}_i, Y_i\right\} \\
=: \;& q_n(1) + q_n(2) + q_n(3) \qquad (59)
\end{aligned}
$$

Once again, as in (58), one obtains $q_n(1) = P\big\{\big|\widehat{Q}_2(\boldsymbol{X}_i) - Q_2(\boldsymbol{X}_i)\big| > \frac{B\,\varphi_{00}^2\,\beta}{64(1-\pi_{\min})}\Big|\boldsymbol{X}_i, Y_i\big\} \le C_{12}\exp\big\{-C_{13}\,mh^d\beta^2\big\}$, for n large enough, where C_{12} and C_{13} are positive constants not depending on m. Furthermore, $q_n(2) = P\big\{\widehat{Q}_1(\boldsymbol{X}_i) - Q_1(\boldsymbol{X}_i) > 2(1-\pi_{\min}) - Q_1(\boldsymbol{X}_i)\big|\boldsymbol{X}_i, Y_i\big\} \le P\big\{\big|\widehat{Q}_1(\boldsymbol{X}_i) - Q_1(\boldsymbol{X}_i)\big| > 1-\pi_{\min}\big|\boldsymbol{X}_i, Y_i\big\} \le C_{14}\exp\big\{-C_{15}\,mh^d\big\}$, for n large enough, where we used the fact that $Q_1(\boldsymbol{X}_i) \le 1-\pi_{\min}$. Similarly, one finds $q_n(3) \le C_{16}\exp\big\{-C_{17}\,mh^d\big\}$, for n large enough. Here C_{14} – C_{17} are positive constants not depending on n. Putting all the above together with (59), one arrives at

$$
\Delta_{ni}(1) \;\; \le \;\; C_{12}\exp\big\{-C_{13}\,mh^d\beta^2\big\} + C_{18}\exp\big\{-C_{19}\,mh^d\big\},
$$

for n large enough, where one can take $C_{18} = C_{14} + C_{16}$ and $C_{19} = C_{15} \wedge C_{17}$. This last bound in conjunction with (58), (57), (56), (55), and (53) implies that $S_{n,1}$ (in (51)) satisfies

$$
S_{n,1} \;\le\; 2\,|\mathcal{F}_\varepsilon|\,e^{-\ell\beta^2/32} + \ell\,|\mathcal{F}_\varepsilon|\left(c_{20}\,e^{-c_{21}\,mh^2\beta^2} + c_{17}\,e^{-c_{18}\,mh^2}\right) \qquad (60)
$$

where one can take $c_{20} = C_{10} + C_{12}$, $c_{21} = C_{11} \wedge C_{13}$, $c_{17} = C_{16}$, and $c_{18} = C_{17}$. Finally, the theorem follows from (60) and (52).

□

PROOF OF THEOREM 3

First note that for every $\beta > 0$, one can write

$$
\begin{aligned}
P\left\{L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - L^* > \beta\right\} &= P\left\{L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - \inf_{\varphi\in\mathcal{F}_{\varepsilon_m}} L(\psi_\varphi) > \beta - \big(\inf_{\varphi\in\mathcal{F}_{\varepsilon_m}} L(\psi_\varphi) - L^*\big)\right\} \\
&\le P\left\{L_n(\widehat{\psi}_{n,\widehat{\varphi}}) - \inf_{\varphi\in\mathcal{F}_{\varepsilon_m}} L(\psi_\varphi) > \beta - \kappa_0\,\varepsilon_m\right\},
\end{aligned}
$$

for some constant $\kappa_0 > 0$, where the last line follows by virtue of (22). Now, observe that we can choose m large enough so that $\beta - \kappa_0\,\varepsilon_m > 0$. Therefore, for n (and thus m) large enough, the result follows from Theorem 2.

□

PROOF OF COROLLARY 1

The proof follows from an application of the Borel-Cantelli lemma in conjunction with the bound in Theorem 3.

□

References

Chen, X., Diao, G., Qin, J. (2020) Pseudo likelihood-based estimation and testing of missingness mechanism function in nonignorable missing data problems. *Scand. J. Stat.* 47, 1377-1400.

Cheng, PE., Chu, CK. (1996) Kernel estimation of distribution functions and quantiles with missing data. Statistica Sinica 6:6378

Devroye, L., Györfi, L. (1985) Nonparametric density estimation: the L1 view. Wiley, New York

Devroye, L., Györfi, L., Lugosi, G. (1996) A probabilistic theory of pattern recognition. Springer-Verlag, New York

Devroye, L. Krzyżak, A. (1989) An equivalence theorem for L_1 convergence of kernel regression estimate. *J. Stat. Plan. Infer.*, 23, 71-82.

Fang, F., Zhao, J., Shao, J. (2018) Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. *Statist. Sinica*, 28, 1677-1701.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. A distribution-free theory of nonparametric regression. Springer-Verlag, New York, 2002.

Hall, P. Kang, K.H. (2005) Bandwidth choice for nonparametric classification. *Ann. Stat.*, 33, 284-306.

Kim, J.K., Yu, C.L. (2011) A semiparametric estimation of mean functionals with nonignorable missing data. *J. Am. Statist. Assoc.* 106, 157-65.

Kraus, D. and Czado, C. (2017). D-vine copula based quantile regression. Comp. Statist. Data Ana. 110, 1-18.

Liu, Z., Yau, C.-Y. (2021) Fitting time series models for longitudinal surveys with nonignorable missing data. *J. Statist. Plann. Inference* 214, 1-12.

Maity, A., Pradhan, V., Das, U. (2019) Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. *Amer. Statist.* 73, 340-349.

Mojirsheibani, M., Reese, T. (2017) Kernel regression estimation for incomplete data with applications. *Statist. Papers*, 58, 185-209.

Morikawa, K., Kano, Y. (2018) Identification problem of transition models for repeated measurement data with nonignorable missing values. *J. Multivariate Anal.* 165, 216-230.

Morikawa, K., Kim, J. K. Kano, Y. (2017) Semiparametric maximum likelihood estimation with data missing not at random. *Can. J. Statist.*, 45, 393-409.

Morikawa, K. and Kim, J. K. (2018) A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. *Stat. & Probab. Lett.*, 140, 1-6

Nagler, T. Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. J. Multivar. Anal. 151, 69-89.

Noh, H., Ghouch, A., Bouezmarni, T. (2013). Copula-based regression estimation and inference. J. Am. Statist. Assoc. 108, 676-688.

O'Brien, J., Gunawardena, H., Paulo, J., Chen, X., Ibrahim, J., Gygi, S., Qaqish, B. (2018) The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *Ann. Appl. Stat.* 12, 2075-2095.

Reese, T., Mojirsheibani, M. (2017) On the L_p norms of kernel regression estimators for incomplete data with applications to classification. *Statist. Methods & Appl.*, 26, 81-112.

Sadinle,., Reiter, J. (2019) Sequentially additive nonignorable missing data modelling using auxiliary marginal information. *Biometrika* 106, 889-911.

Shao, J., Wang, L. (2016) Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103, 175-187.

Tang, N., Zhao, P., Zhu, H. (2014) Empirical likelihood for estimating equations with nonignorably missing data. *Statist. Sinica*, 24, 723-47.

Uehara, M. and Kim, J.K. Semiparametric response model with nonignorable nonresponse. Preprint on arXiv:1810.12519 (2018). https://arxiv.org/abs/1810.12519v1

van der Vaart, A.W., Wellner, J.A., (1996) Weak Convergence and Empirical Processes with Applications to Statistics. Springer, New York.

Yuan, C., Hedeker, D., Mermelstein, R., Xie, H. (2020) A tractable method to account for high-dimensional nonignorable missing data in intensive longitudinal data. *Stat. Med.* 39, 2589-2605.

Zhao, J., Shao, J. (2015) Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *J. Am. Statist. Assoc* 110, 1577-1590.

Zhao, P., Tang, N., Qu, A. Jiang, D. (2017) Semiparametric estimating equations inference with nonignorable missing data. *Statist. Sinica* 27, 89-113.

Zhao, P., Wang, L., Shao, J. (2019) Empirical likelihood and Wilks phenomenon for data with nonignorable missing values. *Scand. J. Stat.* 46, 1003-1024.

Zhao, H., Zhao, P., Tang, N. (2013) Empirical likelihood inference for mean functionals with nonignorably missing response data. *Comp. Statist. Data Anal.* 66, 101-16.