Supplementary Material for: "A kernel-type regression estimator for NMAR response variables with applications to classification"

Majid Mojirsheibani and Arin Khudaverdyan

Department of Mathematics, California State University Northridge, CA, 91330, USA

We need the following regularity assumptions:

Assumption (A). For all $\varphi \in \mathcal{F}$, the selection probability function $\pi_\varphi(\mathbf{x}, y)$ in (3) satisfies $\inf_{\mathbf{x},y} \pi_\varphi(\mathbf{x}, y) =: \pi_{\min} > 0$.

Assumption (B). The kernel $\mathcal{K}$ used in (11) and (12) is regular and satisfies $\int_{\mathbb{R}^d} \mathcal{K}(\mathbf{x})\, d\mathbf{x} = 1$ and $\int_{\mathbb{R}^d} |x_i| \mathcal{K}(\mathbf{x})\, d\mathbf{x} < \infty$ for each $x_i \in (x_1, \cdots, x_d) = \mathbf{x}$. Also, the smoothing parameter h satisfies $h \to 0$ and $mh^d \to \infty$, as n (and thus m) $\to \infty$.

Assumption (C). The density function $f(\mathbf{x})$ of $\mathbf{X}$ 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 (D). For all $\varphi \in \mathcal{F}$, $E[\Delta\, \varphi(Y) | \mathbf{X} = \mathbf{x}] \geq \varrho_0$ for μ–a.e. $\mathbf{x}$, for some constant $\varrho_0 > 0$.

Assumption (E). The partial derivatives $\frac{\partial}{\partial x_i} E[\Delta | \mathbf{X} = \mathbf{x}]$ and $\frac{\partial}{\partial x_i} E[\Delta\, \varphi(Y) | \mathbf{X} = \mathbf{x}]$ exist for each $x_i \in (x_1, \cdots, x_d) = \mathbf{x}$ and are bounded on the compact support of f.

Assumption (F). The class $\mathcal{F}$ is a totally bounded class of functions $\varphi : [-L, L] \to (0, B]$, for some $B < \infty$ and $L < \infty$.

Here assumption (A) is quite standard in missing data literature and has been used, for example, by Kim and Yu (2011) in semiparametric estimation of mean functionals and by Shao and Wang (2016) in inverse weighting estimation. Assumption (B) is not restrictive at all since the choice of the kernel is at our discretion; stronger versions of this assumption appear in Kim and Yu (2011). The first part of Assumption (C) is often imposed in the literature on nonparametric regression to avoid unstable estimates of $m(\mathbf{x})$ in the tails of the density, f; see, for example, Cheng and Chu (1996) and Mojirsheibani (2007). Assumption (D) is quite mild and is justified by the fact that $E[\Delta\, \varphi(Y) | \mathbf{X}] = E[\varphi(Y) E(\Delta | \mathbf{X}, Y) | \mathbf{X}] \geq \pi_{\min} E[\varphi(Y) | \mathbf{X}]$ together with the fact that $\varphi(y) > 0$ for all y; this assumption is new. Assumption (E) is technical and versions of it have already been used in the literature; see, for example, Cheng and Chu (1996) and Mojirsheibani (2007). Assumption (F) is often used in nonparametric estimation; see, for example, Györfi et al (2002; Ch. 9).

In what follows, equation numbers (1) through (25) refer to those in the main article. To proceed with the proofs, we start by stating a number of lemmas.

Lemma 1 *The regression function* $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$ *admits the following representation*

$$m(\mathbf{x}) \;=\; m(\mathbf{x};\varphi^*) \;:=\; \eta_1(\mathbf{x}) + \frac{\psi_1(\mathbf{x};\varphi^*)}{\psi_2(\mathbf{x};\varphi^*)} \cdot (1 - \eta_2(\mathbf{x}))\,. \tag{26}$$

where the functions ψ_k *and* η_k, $k = 1, 2$, *are given by (6) and* φ^* *is as in (4).*

PROOF OF LEMMA 1.

Let $\pi_{\varphi^*}(\mathbf{x}, y)$ be as in (4) and observe that

$$1 - \pi_{\varphi^*}(\mathbf{X}, Y) = \frac{\exp\{g(\mathbf{X})\}\,\varphi^*(Y)}{1 + \exp\{g(\mathbf{X})\}\,\varphi^*(Y)} = \exp\{g(\mathbf{X})\}\,\varphi^*(Y)\,\pi_{\varphi^*}(\mathbf{X}, Y). \tag{27}$$

Now, writing $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}] = E[Y\Delta|\mathbf{X} = \mathbf{x}] + \frac{E[Y(1-\Delta)|\mathbf{X}=\mathbf{x}]}{E[1-\Delta|\mathbf{X}=\mathbf{x}]} \cdot E[1 - \Delta|\mathbf{X} = \mathbf{x}]$, one finds

$$\begin{aligned}
\frac{E[Y(1-\Delta)|\mathbf{X}]}{E[1-\Delta|\mathbf{X}]} \quad &= \quad \frac{E\big[E\big\{Y(1-\Delta)|\mathbf{X}, Y\big\}\big|\mathbf{X}\big]}{E\big[E\big\{1-\Delta|\mathbf{X}, Y\big\}\big|\mathbf{X}\big]} = \frac{E\big[Y\big\{1-\pi_{\varphi^*}(\mathbf{X}, Y)\big\}\big|\mathbf{X}\big]}{E\big[1-\pi_{\varphi^*}(\mathbf{X}, Y)\big|\mathbf{X}\big]} \\
&\overset{\text{by (27)}}{=} \quad \frac{E\big[Y\exp\{g(\mathbf{X})\}\,\varphi^*(Y)\,\pi_{\varphi^*}(\mathbf{X}, Y)\big|\mathbf{X}\big]}{E\big[\exp\{g(\mathbf{X})\}\,\varphi^*(Y)\,\pi_{\varphi^*}(\mathbf{X}, Y)\big|\mathbf{X}\big]} = \frac{E\big[Y\varphi^*(Y)\,\Delta\big|\mathbf{X}\big]}{E\big[\varphi^*(Y)\,\Delta\big|\mathbf{X}\big]} = \frac{\psi_1(\mathbf{X};\varphi^*)}{\psi_2(\mathbf{X};\varphi^*)}\,.
\end{aligned}$$

The proof of the lemma now follows from this last result.

□

To state the next lemma, let $\mathcal{F}$ be a totally bounded class of functions $\varphi : [-L, L] \to (0, B]$, for some $B < \infty$. Also, for each $\varphi \in \mathcal{F}$, define

$$m(\mathbf{x};\varphi) \;=\; \eta_1(\mathbf{x}) + \frac{\psi_1(\mathbf{x};\varphi)}{\psi_2(\mathbf{x};\varphi)} \cdot (1 - \eta_2(\mathbf{x}))\,. \tag{28}$$

where ψ_k and η_k, $k = 1, 2$, are given by (6).

Lemma 2 *Let* $m(\mathbf{x};\varphi_j)$, $j = 1, 2$, *be as in (28). Then, under assumption (D), one has*

$$E\Big|m(\mathbf{X};\varphi_1) - m(\mathbf{X};\varphi_2)\Big| \;\le\; C \cdot \sup_{-L\le y\le L} \big|\varphi_1(y) - \varphi_2(y)\big|,$$

where the constant $C > 0$ *can be taken to be* $C = 2L/\varrho_0$, *with* ϱ_0 *as in assumption (D).*

PROOF OF LEMMA 2.

Let ψ_k, $k = 1, 2$, be as in (6) and observe that

$$\begin{aligned}
\Big|m(\mathbf{x};\varphi_1) - m(\mathbf{x};\varphi_2)\Big| \;&=\; \left|\frac{-\psi_1(\mathbf{x};\varphi_1)}{\psi_2(\mathbf{x};\varphi_1)} \cdot \frac{\psi_2(\mathbf{x};\varphi_1) - \psi_2(\mathbf{x};\varphi_2)}{\psi_2(\mathbf{x};\varphi_2)} + \frac{\psi_1(\mathbf{x};\varphi_1) - \psi_1(\mathbf{x};\varphi_2)}{\psi_2(\mathbf{x};\varphi_2)}\right| \\
&\quad \times E\big[1 - \Delta|\mathbf{X} = \mathbf{x}\big] \\
&\le\; \frac{1}{\psi_2(\mathbf{x};\varphi_2)}\big\{L\big|\psi_2(\mathbf{x};\varphi_1) - \psi_2(\mathbf{x};\varphi_2)\big| + \big|\psi_1(\mathbf{x};\varphi_1) - \psi_1(\mathbf{x};\varphi_2)\big|\big\},
\end{aligned}$$

where we used the fact $|\psi_1(\mathbf{x};\varphi_1)|/|\psi_2(\mathbf{x};\varphi_1)| \le L|\psi_2(\mathbf{x};\varphi_1)|/|\psi_2(\mathbf{x};\varphi_1)| = L$ (because $\varphi_k > 0$). But, since $|Y\Delta| \le L$, one finds

$$|\psi_1(\mathbf{x};\varphi_1) - \psi_1(\mathbf{x};\varphi_2)| \le E\left[|\Delta Y| \cdot \big|\varphi_1(Y) - \varphi_2(Y)\big| \,\big|\mathbf{X} = \mathbf{x}\right] \le L \sup_{-L\le y\le L} \big|\varphi_1(y) - \varphi_2(y)\big|.$$

Similarly, $|\psi_2(\mathbf{x};\varphi_1) - \psi_2(\mathbf{x};\varphi_2)| \le \sup_{-L\le y\le L} \big|\varphi_1(y) - \varphi_2(y)\big|$. On the other hand, by assumption (D) we have $\psi_2(\mathbf{x};\varphi_2) \ge \varrho_0 > 0$, for μ–a.e. $\mathbf{x}$. Therefore

$$\Big|m(\mathbf{x};\varphi_1) - m(\mathbf{x};\varphi_2)\Big| \le (2L/\varrho_0) \sup_{-L\le y\le L} \big|\varphi_1(y) - \varphi_2(y)\big|.$$

The lemma follows now by integrating both sides of this inequality with respect to $\mu(d\mathbf{x})$.

□

Next, for any $\varepsilon > 0$, let $\mathcal{F}_\varepsilon$ be any ε-cover of $\mathcal{F}$ (see Section 2.1). Also, let $m(\mathbf{x};\varphi)$ and $\widehat{L}_{m,\ell}(\varphi)$ be as in (28) and (14), and define the quantities

$$\varphi_\varepsilon = \operatorname*{argmin}_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\varphi) - Y\big|^2 \quad \text{and} \quad \widehat{\varphi}_\varepsilon = \operatorname*{argmin}_{\varphi\in\mathcal{F}_\varepsilon} \widehat{L}_{m,\ell}(\varphi). \tag{29}$$

Lemma 3 *Let $\widehat{m}_m(\mathbf{x};\varphi)$, $\widehat{L}_{m,\ell}(\varphi)$, and $m(\mathbf{x};\varphi)$ be as in (10), (14), and (28), respectively. Also, let φ_ε and $\widehat{\varphi}_\varepsilon$ be as in (29). Then, under the conditions of Theorem 1, for every $\varepsilon > 0$ we have*

$$\begin{aligned} E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big|^2 \Big|\mathbb{D}_n\right] \;\le\; & \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|E\left[\Big|\widehat{m}_m(\mathbf{X};\varphi) - Y\Big|^2 \Big|\mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\varphi)\right| \\ & + \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_{m,\ell}(\varphi) - E\Big|m(\mathbf{X};\varphi) - Y\Big|^2\right| \\ & + \varepsilon\cdot C_1 \cdot \sqrt{E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big|^2 \Big|\mathbb{D}_n\right]} \end{aligned} \tag{30}$$

where C_1 is a positive constant not depending on ℓ, m, n, or ε.

PROOF OF LEMMA 3.

Observe that $E\big[|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - Y|^2|\mathbb{D}_n\big] = E\big[|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)|^2|\mathbb{D}_n\big] + E|m(\mathbf{X};\varphi_\varepsilon) - Y|^2 + 2E\big[\big(\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\big)\big(m(\mathbf{X};\varphi_\varepsilon) - Y\big)\big|\mathbb{D}_n\big]$. Also, let φ^* be as in (4) and note that

$$\begin{aligned} & E\left[\Big(\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big)\Big(m(\mathbf{X};\varphi_\varepsilon) - Y\Big)\Big|\mathbb{D}_n\right] \\ & = E\left[\Big(\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big)\Big(m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*) + m(\mathbf{X};\varphi^*) - Y\Big)\Big|\mathbb{D}_n\right] \\ & = E\left[\Big(\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big)\Big(m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\Big)\Big|\mathbb{D}_n\right], \end{aligned}$$

where we have used the fact that in view of (5), $E[Y|\mathbf{X} = \mathbf{x}] := m(\mathbf{x}) = m(\mathbf{x};\varphi^*)$. Therefore

$$E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big|^2 \Big|\mathbb{D}_n\right]$$

$$
\begin{aligned}
&= \Big\{ E\Big[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - Y\Big|^2 \Big| \mathbb{D}_n\Big] - E\Big|m(\mathbf{X};\varphi_\varepsilon) - Y\Big|^2 \Big\} \\
&\quad - 2E\Big[\Big(\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big)\Big(m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\Big)\Big|\mathbb{D}_n\Big] \\
&:= \mathbf{I}_n + \mathbf{II}_n. \qquad (31)
\end{aligned}
$$

Now, observe that

$$
\begin{aligned}
\mathbf{I}_n &= E\Big[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - Y\Big|^2 \Big| \mathbb{D}_n\Big] - \inf_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\varphi) - Y\big|^2 \\
&= \sup_{\varphi\in\mathcal{F}_\varepsilon} \Big\{ E\Big[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - Y\Big|^2 \Big| \mathbb{D}_n\Big] - \widehat{L}_{m,\ell}(\varphi) + \widehat{L}_{m,\ell}(\varphi) - \widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon) \\
&\qquad + \widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon) - E\big|m(\mathbf{X};\varphi) - Y\big|^2 \Big\}, \quad \text{(where } \widehat{L}_{m,\ell}(\varphi) \text{ is as in (14))} \\
&\le \Big(E\Big[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - Y\Big|^2 \Big| \mathbb{D}_n\Big] - \widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon)\Big) + \sup_{\varphi\in\mathcal{F}_\varepsilon} \Big|\widehat{L}_{m,\ell}(\varphi) - E\big|m(\mathbf{X};\varphi) - Y\big|^2\Big|,
\end{aligned}
$$

where the last line follows since $\widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon) \le \widehat{L}_{m,\ell}(\varphi)$ holds for all $\varphi \in \mathcal{F}_\varepsilon$ (because of the definition of $\widehat{\varphi}_\varepsilon$ in (29)). Therefore,

$$
\big|\mathbf{I}_n\big| \le \sup_{\varphi\in\mathcal{F}_\varepsilon} \left| E\Big[\Big|\widehat{m}_m(\mathbf{X};\varphi) - Y\Big|^2 \Big| \mathbb{D}_m\Big] - \widehat{L}_{m,\ell}(\varphi)\right| + \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_{m,\ell}(\varphi) - E\big|m(\mathbf{X};\varphi) - Y\big|^2\right|, \qquad (32)
$$

where conditioning on $\mathbb{D}_m$ in the above expression reflects the fact that $\widehat{m}_m(\mathbf{X};\varphi)$ depends on $\mathbb{D}_m$ only (and not the entire data $\mathbb{D}_n$). Furthermore, by Cauchy-Schwarz inequality

$$
\big|\mathbf{II}_n\big| \le 2\sqrt{E\Big[\big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\big|^2 \Big| \mathbb{D}_n\Big]} \cdot \sqrt{E\big|m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\big|^2}. \qquad (33)
$$

Now, let $\varphi^\dagger \in \mathcal{F}_\varepsilon$ be such that $\varphi^* \in B(\varphi^\dagger, \varepsilon)$; such a $\varphi^\dagger \in \mathcal{F}_\varepsilon$ exists because $\varphi^* \in \mathcal{F}$ and $\mathcal{F}_\varepsilon$ is an ε-cover of $\mathcal{F}$. Then, using the fact that $E\big|m(\mathbf{X};\varphi_\varepsilon) - Y\big|^2 = E\big|m(\mathbf{X};\varphi^*) - Y\big|^2 + E\big|m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\big|^2$, one finds

$$
\begin{aligned}
E\big|m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\big|^2 &= \inf_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\varphi) - Y\big|^2 - E\big|m(\mathbf{X};\varphi^*) - Y\big|^2 \\
&= \inf_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\varphi) - m(\mathbf{X};\varphi^*)\big|^2 \le E\big|m(\mathbf{X};\varphi^\dagger) - m(\mathbf{X};\varphi^*)\big|^2 \\
&\le 2CL\Big[\sup_{-L\le y\le L} \big|\varphi^\dagger(y) - \varphi^*(y)\big|\Big]^2, \quad \text{(by Lemma 2)} \\
&\le 2CL\cdot\varepsilon^2, \qquad (34)
\end{aligned}
$$

because $\varphi^* \in B(\varphi^\dagger, \varepsilon)$. Therefore, in view of (33) and (34), one finds

$$
\big|\mathbf{II}_n\big| \le \varepsilon \cdot C_1 \sqrt{E\Big[\big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\big|^2 \Big| \mathbb{D}_n\Big]}, \qquad (35)
$$

for a constant $C_1 > 0$ not depending on n or ε. Now Lemma 3 follows from (31), (32), and (35).

□

PROOF OF THEOREM 1

First observe that by Lemma 3 one can write

$$\begin{aligned}
&\int \left|\widehat{m}_m(\mathbf{x};\widehat{\varphi}_\varepsilon) - m(\mathbf{x};\varphi_\varepsilon)\right|^2 \mu(d\mathbf{x}) \\
&\quad \le \sup_{\varphi\in\mathcal{F}_\varepsilon} \left| E\left[\left|\widehat{m}_m(\mathbf{X};\widehat{\varphi}) - Y\right|^2 \middle| \mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\varphi)\right| + \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_{m,\ell}(\varphi) - E\middle|m(\mathbf{X};\varphi) - Y\right|^2\right| \\
&\qquad + \ \varepsilon\cdot C_1 \sqrt{\int \left|\widehat{m}_m(\mathbf{x};\widehat{\varphi}_\varepsilon) - m(\mathbf{x};\varphi_\varepsilon)\right|^2 \mu(d\mathbf{x})}\,, \qquad (36)
\end{aligned}$$

where, φ_ε and $\widehat{\varphi}_\varepsilon$ are as in (29). Therefore, in view of (36), for every constant $t > 0$

$$\begin{aligned}
P\left\{\int \left|\widehat{m}_m(\mathbf{x};\widehat{\varphi}_\varepsilon) - m(\mathbf{x};\varphi_\varepsilon)\right|^2 \mu(d\mathbf{x}) > t\right\} - P\left\{\int \left|\widehat{m}_m(\mathbf{x};\widehat{\varphi}_\varepsilon) - m(\mathbf{x};\varphi_\varepsilon)\right|^2 \mu(d\mathbf{x}) > \frac{t^2}{\varepsilon^2 c_4}\right\} \\
\le\ P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon} \left| E\left[\left|\widehat{m}_m(\mathbf{X};\varphi) - Y\right|^2 \middle| \mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\varphi)\right| > \frac{t}{3}\right\} \\
+ P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_{m,\ell}(\varphi) - E\middle|m(\mathbf{X};\varphi) - Y\right|^2\right| > \frac{t}{3}\right\}, \qquad (37)
\end{aligned}$$

where $c_4 = (3C_1)^2$ with C_1 as in (36). But observe that for every constant $\beta > 0$

$$\begin{aligned}
&P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_{m,\ell}(\varphi) - E\left[\left|\widehat{m}_m(\mathbf{X};\varphi) - Y\right|^2 \middle| \mathbb{D}_m\right]\right| > \beta\right\} \qquad (38) \\
&= P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \Big(\Delta_i + \frac{(1-\Delta_i)\delta_i}{p_n}\Big)\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2 - E\left[\left|\widehat{m}_m(\mathbf{X};\varphi) - Y\right|^2 \middle| \mathbb{D}_m\right]\right| > \beta\right\}.
\end{aligned}$$

On the other hand, for every $i \in \boldsymbol{\mathcal{I}}_\ell$, we have

$$\begin{aligned}
&E\left[\Big(\Delta_i + \frac{(1-\Delta_i)\delta_i}{p_n}\Big)\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2 \middle| \mathbb{D}_m\right] \\
&= E\left[\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2 \left(E\left\{\Delta_i \middle| \mathbb{D}_m, \mathbf{X}_i, Y_i, \delta_i\right\} + \frac{\delta_i}{p_n} E\left\{1-\Delta_i \middle| \mathbb{D}_m, \mathbf{X}_i, Y_i, \delta_i\right\}\right) \middle| \mathbb{D}_m\right] \\
&= E\left[\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2 \pi_{\varphi^*}(\mathbf{X}_i, Y_i) \middle| \mathbb{D}_m\right] + \frac{E(\delta_i)}{p_\ell} E\left[\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2 \big(1 - \pi_{\varphi^*}(\mathbf{X}_i, Y_i)\big) \middle| \mathbb{D}_m\right] \\
&= E\left[\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2 \middle| \mathbb{D}_m\right],
\end{aligned}$$

because δ_i is independent of the data $\mathbb{D}_n$ (with $E(\delta_i) = p_n$), and the fact that Δ_i is independent of $\mathbb{D}_m$ for all $i \in \boldsymbol{\mathcal{I}}_\ell$. Moreover, in view of (10), one finds $|\widehat{m}_m(\mathbf{x};\varphi)| \le 2L$. Thus, conditional on $\mathbb{D}_m$, the terms $\big(\Delta_i + (1-\Delta_i)\delta_i/p_n\big)\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2$, $i \in \boldsymbol{\mathcal{I}}_\ell$, are independent nonnegative random variables bounded by $9L^2/p_n$. Therefore

$$(38) \ \le\ |\mathcal{F}_\varepsilon| \sup_{\varphi\in\mathcal{F}_\varepsilon} E\left[P\left\{\left|\widehat{L}_{m,\ell}(\varphi) - E\left[\left|\widehat{m}_m(\mathbf{X};\varphi) - Y\right|^2 \middle| \mathbb{D}_m\right]\right| > \beta \middle| \mathbb{D}_m\right\}\right]$$

$$\leq \quad 2\big|\mathcal{F}_\varepsilon\big| \exp\Big\{ -2\ell\,\beta^2\, p_n^2/(81L^4)\Big\}, \tag{39}$$

via Hoeffding's inequality. To complete the proof, we also need to deal with the last probability statement on the right side of (37). To that end, let $\widehat{L}_{m,\ell}(\varphi)$ be as in (14) and define the quantities

$$Q_{n,1}(\varphi) \;=\; \left| \widehat{L}_{m,\ell}(\varphi) - \frac{1}{\ell} \sum_{i\in \boldsymbol{\mathcal{I}}_\ell} \Big(\Delta_i + \frac{(1-\Delta_i)\delta_i}{p_n}\Big) \Big| m(\mathbf{X}_i;\varphi) - Y_i\Big|^2 \right| \tag{40}$$

$$Q_{n,2}(\varphi) \;=\; \left| \frac{1}{\ell} \sum_{i\in \boldsymbol{\mathcal{I}}_\ell} \Big(\Delta_i + \frac{(1-\Delta_i)\delta_i}{p_n}\Big) \Big| m(\mathbf{X}_i;\varphi) - Y_i\Big|^2 - E\big| m(\mathbf{X};\varphi) - Y\big|^2 \right| \tag{41}$$

and observe that for every $\beta > 0$,

$$\begin{aligned} &P\left\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} \left| \widehat{L}_{m,\ell}(\varphi) - E\Big| m(\mathbf{X};\varphi) - Y\Big|^2 \right| > \beta \right\} \\ &\qquad \leq\; P\Big\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} \big|Q_{n,1}(\varphi)\big| > \frac{\beta}{2}\Big\} + P\Big\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} \big|Q_{n,2}(\varphi)\big| > \frac{\beta}{2}\Big\} \quad =: \quad Q_{n,1} + Q_{n,2}\,. \end{aligned} \tag{42}$$

But $\big(\Delta_i + ((1-\Delta_i)\delta_i/p_n\big) \leq 1/p_n$ and $|\widehat{m}_m(\mathbf{x};\varphi)| \leq 2L$. Therefore, in view of the definition of $\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)$ in (14) and the fact that $|a^2 - b^2| \leq |a-b||a+b|$, one can write

$$\begin{aligned} Q_{n,1} \;&\leq\; P\Bigg\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} \frac{1}{\ell} \sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \Big[\Big(\Delta_i + \frac{(1-\Delta_i)\delta_i}{p_n}\Big) \\ &\qquad\qquad \times \Big|\widehat{m}_m(\mathbf{X}_i;\varphi) - m(\mathbf{X}_i;\varphi)\Big| \Big| \widehat{m}_m(\mathbf{X}_i;\varphi) + m(\mathbf{X}_i;\varphi) - 2Y_i\Big| \Big] > \frac{\beta}{2} \Bigg\} \\ &\leq\; P\left\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} \left| \frac{1}{\ell} \sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \left[\frac{5L}{p_n} \Big| \widehat{m}_m(\mathbf{X}_i;\varphi) - m(\mathbf{X}_i;\varphi)\Big| \right] \right| > \frac{\beta}{2} \right\} \\ &\leq\; \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}_\varepsilon} \sum_{i\in\boldsymbol{\mathcal{I}}_\ell} P\Big\{ \Big| \widehat{m}_m(\mathbf{X}_i;\varphi) - m(\mathbf{X}_i;\varphi)\Big| > \frac{\beta\, p_n}{10L}\Big\}. \end{aligned} \tag{43}$$

Using tedious but straightforward arguments, it can be shown (see, for example, Lemma 4.1 of Mojirsheibani (2007)) that under assumptions (B), (C), and (E), for each $i \in \boldsymbol{\mathcal{I}}_\ell$

$$P\Big\{ \Big| \widehat{m}_m(\mathbf{X}_i;\varphi) - m(\mathbf{X}_i;\varphi)\Big| > \frac{\beta\, p_n}{10L}\Big\} \;\leq\; C_{10}\, \exp\big\{ -C_{11} m h^d p_n^2\, \beta^2\big\},$$

for n (and thus m) large enough, where C_{10} and C_{11} are positive constants not depending on $\mathbf{X}_i$, n, β, or φ. Putting this together with (43), we arrive at

$$Q_{n,1} \;\leq\; C_{10}\big|\mathcal{F}_\varepsilon\big|\, \ell \cdot \exp\big\{ -C_{11} m h^d p_n^2\, \beta^2\big\}. \tag{44}$$

To deal with the term $Q_{n,2}$ in (42), we note that the terms $\big(\Delta_i + (1-\Delta_i)\delta_i/p_n\big) \cdot \big|m(\mathbf{X}_i;\pi_\varphi) - Y_i\big|^2$ are independent bounded random variables taking values in $\big[0\,,\, 9L^2/p_n\big]$. Therefore an application

of Hoeffding's inequality gives

$$\begin{aligned} Q_{n,2} &\le |\mathcal{F}_\varepsilon| \sup_{\varphi\in\mathcal{F}_\varepsilon} P\left\{ \left| \frac{1}{\ell} \sum_{i\in\mathcal{I}_\ell} \Big(\Delta_i + \frac{(1-\Delta_i)\delta_i}{p_n}\Big) \Big| m(\mathbf{X}_i;\pi_\varphi) - Y_i \Big|^2 - E\Big| m(\mathbf{X};\pi_\varphi) - Y \Big|^2 \right| > \frac{\beta}{2} \right\} \\ &\le 2\,|\mathcal{F}_\varepsilon| \exp\Big\{ -2\,\ell\, p_n^2\, \beta^2 / \big(324 L^4\big) \Big\}. \end{aligned} \tag{45}$$

Putting together (42), (44), and (45), for every $\beta > 0$ and n large enough, one has

$$P\left\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} \left| \widehat{L}_{m,\ell}(\varphi) - E\Big| m(\mathbf{X};\pi_\varphi) - Y \Big|^2 \right| > \beta \right\} \le C_{12}\,|\mathcal{F}_\varepsilon| \Big(\ell\, e^{-C_{11} m h^d\, p_n^2\, \beta^2} + e^{-C_{13}\ell\, p_n^2\, \beta^2} \Big), \tag{46}$$

where C_{11}, C_{12}, and C_{13} are positive constants not depending on n or β. Now, for any decreasing sequence $0 < \varepsilon_n \downarrow 0$, as $n \to \infty$, let φ_{ε_n} be as in (29) but with ε changed to ε_n. Also, let $m(\mathbf{x};\varphi)$ be as in (28). Then the C_p-inequality (with $p=2$) in conjunction with the arguments that led to (34) in the proof of Lemma 3 yield

$$\begin{aligned} \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x}) \Big|^2 \mu(d\mathbf{x}) &= \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x};\varphi_{\varepsilon_n}) + m(\mathbf{x};\varphi_{\varepsilon_n}) - m(\mathbf{x}) \Big|^2 \mu(d\mathbf{x}) \\ &\le 2 \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x};\varphi_{\varepsilon_n}) \Big|^2 \mu(d\mathbf{x}) + 2 \int \Big| m(\mathbf{x};\pi_{\varepsilon_n}) - m(\mathbf{x};\varphi^*) \Big|^2 \mu(d\mathbf{x}) \\ &\le 2 \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x};\varphi_{\varepsilon_n}) \Big|^2 \mu(d\mathbf{x}) + C_{14} \cdot \varepsilon_n^2\,, \end{aligned} \tag{47}$$

because $m(\mathbf{x}) = m(\mathbf{x};\varphi^*)$ (see Lemma 1), where C_{14} is a positive constant not depending on n. Therefore, in view of (47) and (37), for every constant $t > 0$, and with $c_{15} = C_{14}/2$, we have

$$\begin{aligned} &\frac{1}{2}\, P\left\{ \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x}) \Big|^2 \mu(d\mathbf{x}) > t \right\} \\ &\quad\le \frac{1}{2}\, P\left\{ \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x};\varphi_{\varepsilon_n}) \Big|^2 \mu(d\mathbf{x}) > t/2 - c_{15}\,\varepsilon_n^2 \right\} \\ &\quad\le P\left\{ \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x};\varphi_{\varepsilon_n}) \Big|^2 \mu(d\mathbf{x}) > t/2 - c_{15}\varepsilon_n^2 \right\} \\ &\qquad - P\left\{ \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x};\varphi_{\varepsilon_n}) \Big|^2 \mu(d\mathbf{x}) > (t/2 - c_{15}\,\varepsilon_n^2)^2/(c_4\,\varepsilon_n^2) \right\} \\ &\qquad\qquad \text{(for } n \text{ large enough, where } c_4 > 0 \text{ is as in the first line of the l.h.s. of (37))} \\ &\quad\le P\left\{ \sup_{\varphi\in\mathcal{F}_{\varepsilon_n}} \left| E\left[\Big| \widehat{m}_m(\mathbf{X};\varphi) - Y \Big|^2 \Big| \mathbb{D}_m \right] - \widehat{L}_{m,\ell}(\varphi) \right| > \frac{t/2 - c_{15}\,\varepsilon_n^2}{3} \right\} \\ &\qquad + P\left\{ \sup_{\varphi\in\mathcal{F}_{\varepsilon_n}} \left| \widehat{L}_{m,\ell}(\varphi) - E\Big| m(\mathbf{X};\varphi) - Y \Big|^2 \right| > \frac{t/2 - c_{15}\,\varepsilon_n^2}{3} \right\}, \end{aligned}$$

where the last inequality above follows from the bound in (37). Finally, choosing n large enough so that $(t/2 - c_{15}\,\varepsilon_n^2)/3 > t/12$ and using (38) and the bounds in (39) and (46), we find

$$P\left\{ \int \Big| \widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x}) \Big|^2 \mu(d\mathbf{x}) > t \right\} \le C_{15} \cdot |\mathcal{F}_{\varepsilon_n}| \Big(e^{-C_{16}\,\ell\, p_n^2\, t^2} + \ell\, e^{-C_{17}\, m h^d\, p_n^2\, t^2} \Big),$$

for n large enough where C_{15}, C_{16}, and C_{17} are positive constants not depending on m, ℓ, or t. This completes the proof of the theorem.

□

PROOF OF COROLLARY 1

The corollary follows from the Borel-Cantelli lemma in conjunction with (18), the bound in Theorem 1, and Remark 1.

□

PROOF OF THEOREM 2

We first note that, by Remark 1, it is sufficient to prove the theorem for the case of $p = 2$. The proof is along standard arguments and goes as follows. Observe that

$$\begin{aligned} E\big|\widehat{m}(\mathbf{X};\widehat{\varphi}_n) - m(\mathbf{X})\big|^2 &= E\left[\int_{\mathbb{R}^d}\Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\Big|^2\,\mu(d\mathbf{x}\right] \\ &= \int_0^\infty P\left\{\int_{\mathbb{R}^d}\Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\Big|^2\mu(d\mathbf{x}) > t\right\}dt \\ &= \int_0^{9L^2} P\left\{\int_{\mathbb{R}^d}\Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\Big|^2\mu(d\mathbf{x}) > t\right\}dt\,, \qquad (48) \end{aligned}$$

where the last line follows because, by the definition of the estimator $\widehat{m}(\mathbf{x};\widehat{\varphi}_n)$ in (16), one has $\big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\big|^2 \le \big(\big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n)\big| + \big|m(\mathbf{x})\big|\big)^2 \le (2L+L)^2$. Therefore, by Theorem 1, for n large enough we have

$$\begin{aligned} (48) \le{}& \int_0^u dt \,+\, c_1\big|\mathcal{F}_{\varepsilon_n}\big|\cdot\left[\int_u^{9L^2} e^{-c_2\ell\, p_n^2\, t^2}\,dt \,+\, \ell\int_u^{9L^2} e^{-c_3\, mh^d\, p_n^2\, t^2}\,dt\right] \\ &\qquad \text{(where } c_1, c_2, \text{ and } c_3 \text{ are as in (1))} \\ \le{}& u + 2c_1\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\int_u^{9L^2} e^{-c_4(\ell\wedge mh^d)\,p_n^2 t^2}dt\,, \quad \text{where } c_4 = c_2\wedge c_3 \\ \le{}& u + \frac{2\,c_1\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell}{\sqrt{c_4(\ell\wedge mh^d)p_n^2}}\cdot\int_{u\sqrt{c_4(\ell\wedge mh^d)p_n^2}}^\infty e^{-v^2/2}\,dv \\ &\qquad \text{(which follows from the change of variable } v = t\,\sqrt{c_4(\ell\wedge mh^d)p_n^2}\,) \\ \le{}& u + \frac{2\,c_1\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell}{\sqrt{c_4(\ell\wedge mh^d)p_n^2}}\cdot\frac{\exp\big\{-c_4(\ell\wedge mh^d)p_n^2\,u^2/2\big\}}{u\,\sqrt{c_4(\ell\wedge mh^d)p_n^2}}, \qquad (49) \end{aligned}$$

where the last line follows from the upper bound of the Mill's ratio; see, for example, Mitrinovic (1970; p. 177). Next, let $c = 2\,c_1\big|\mathcal{F}_{\varepsilon_n}\big|\ell$ and $N = c_4(\ell\wedge mh^d)p_n^2/4$, and note that

$$(\text{right side of (49)}) \;=\; u + \frac{c}{4Nu}\,e^{-2Nu^2}. \qquad (50)$$

But the expression in (50) is approximately minimized by taking $u = \sqrt{\log(c)/(2N)}$, and the

corresponding minimum of (50) becomes

$$\sqrt{\frac{\log(c)}{2N}} + \sqrt{\frac{1}{8N\log(c)}} \;=\; \sqrt{\frac{c_5 + \log\ell + \log|\mathcal{F}_{\varepsilon_n}|}{c_6\,(\ell \wedge mh^d)\,p_n^2}} + \sqrt{\frac{1}{c_7\,(\ell \wedge mh^d)\,p_n^2\big[c_5 + \log\ell + \log|\mathcal{F}_{\varepsilon_n}|\big]}},$$

where c_5, c_6, and c_7 are positive constants not depending on m, ℓ, and n.

□

PROOF OF THEOREM 3

Part (i). By (23), we have

$$P\left\{\widehat{g}_n(\mathbf{X}) \neq Y \middle| \mathbb{D}_n\right\} - P\{g_{\mathrm{B}}(\mathbf{X}) \neq Y\} \;\leq\; 2E\left[\left|\widehat{m}(\mathbf{X};\widehat{\varphi}_n) - m(\mathbf{X})\right| \middle| \mathbb{D}_n\right] \tag{51}$$

Now part (i) follows from (51), Corollary 1 with $p=2$, and the Cauchy-Schwarz inequality.

Part (ii). Taking the expectation of both sides of (51), the result follows from Corollary 2 with $p=2$, together with the Cauchy-Schwarz inequality.

Part (iii). By a result of Audibert and Tsybakov (2007; Lemma 5.2), under the margin assumption (G), tedious algebra yields

$$P\left\{\widehat{g}_n(\mathbf{X}) \neq Y\right\} - P\{g_{\mathrm{B}}(\mathbf{X}) \neq Y\} \;\leq\; \left(E\left|\widehat{m}(\mathbf{X};\widehat{\varphi}_n) - m(\mathbf{X})\right|^2\right)^{\frac{1+\alpha}{2+\alpha}}, \tag{52}$$

where α is as in (25). The result now follows from Corollary 2 with $p=2$.

□

Acknowledgements

This work was supported by the National Science Foundation grant DMS-2310504 of Majid Mojirsheibani.

References

Audibert, J. Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers under the margin condition.*Ann. Statist.* **35** 608–633.

Cheng, P.E. and Chu, C.K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statistica Sinica,* **6** 63–78.

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

Mitrinovic, D. S. Analytic Inequalities. New York. Springer-Verlag, 1970.

Mojirsheibani, M. (2007). Nonparametric curve estimation with missing covariates: A general empirical process approach. *Journal of Statistical Planning and Inference*. **137** 2733–2758