Electronic Journal of Statistics
Vol. 3 (2009) 1021–1038
ISSN: 1935-7524
DOI: 10.1214/09-EJS420

Empirical measures for incomplete data with applications

Shojaeddin Chenouri[†]

Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada,
e-mail: schenouri@uwaterloo.ca

Majid Mojirsheibani[*‡]

School of Mathematics and Statistics, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada,
e-mail: majidm@math.carleton.ca

Zahra Montazeri[§]

Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, 451 Smyth Road, Ontario, K1H 8M5, Canada,
e-mail: zmontaze@uottawa.ca

Abstract: Methods are proposed to construct empirical measures when there are missing terms among the components of a random vector. Furthermore, Vapnik-Chevonenkis type exponential bounds are obtained on the uniform deviations of these estimators, from the true probabilities. These results can then be used to deal with classical problems such as statistical classification, via empirical risk minimization, when there are missing covariates among the data. Another application involves the uniform estimation of a distribution function.

AMS 2000 subject classifications: Primary 60G50, 62G15; secondary 62H30.
Keywords and phrases: Exponential bounds, Vapnik-Chervonenkis, distribution function, classification, consistency.

Received May 2009.

Contents

*Corresponding author.
†Research supported by a grant from the Natural Sciences and Engineering Research Council of Canada.
‡Research supported by a grant from the Natural Sciences and Engineering Research Council of Canada.
§Postdoctoral fellow.

1. Introduction

Consider the random vector $\mathbf{X}^T = (\mathbf{Y}^T, \mathbf{Z}^T) \in \mathbb{R}^{d+p}$, where $\mathbf{Y} \in \mathbb{R}^d$, $d \geq 1$, is always observable but $\mathbf{Z} \in \mathbb{R}^p$, $p \geq 1$, may be missing. Now let $t : \mathbb{R}^{d+p} \mapsto \mathbb{R}^s$, $s \geq 1$, be a given (known) function and, for any measurable set $A \subset \mathbb{R}^s$, consider the estimation of

$$\mu(A) = P\{t(\mathbf{X}) \in A\} . \tag{1}$$

When $t(\mathbf{X}) = \mathbf{X}$ and $A = (-\infty, x_1] \times \cdots \times (-\infty, x_{d+p}]$ then (1) is the usual empirical cumulative distribution function of the random vector $\mathbf{X}$. Alternatively, taking $t(\mathbf{X}) = \mathbf{Z}$ and $A = (-\infty, z_1] \times \cdots \times (-\infty, z_p]$, gives the empirical c.d.f. of the subvector $\mathbf{Z}$.

Let $\mathcal{D}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be a random sample, where $\mathbf{X}_i \overset{\text{i.i.d.}}{=} \mathbf{X}$, $i = 1, \ldots, n$. Clearly, when every $\mathbf{X}_i^T = (\mathbf{Y}_i^T, \mathbf{Z}_i^T)$ is fully observable, (i.e., there are no missing $\mathbf{Z}_i$'s), one can use the classical empirical version

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I\{t(\mathbf{X}_i) \in A\} . \tag{2}$$

Now the celebrated inequality of Vapnik and Chervonenkis (1971) can be used to obtain uniform (in A) performance bounds on the deviations of $\mu_n(A)$ from $\mu(A)$. More specifically, let $\mathcal{A}$ be a class of measurable sets $\{A \,|\, A \subset \mathbb{R}^s\}$ and define

$$S(\mathcal{A}, n) = \max_{t(\mathbf{x}_1), \ldots, t(\mathbf{x}_n) \in \mathbb{R}^s} \{\#\,\text{of different sets in}\, \{\{t(\mathbf{x}_1), \ldots, t(\mathbf{x}_n)\} \cap A \,|\, A \in \mathcal{A}\}\} .$$

Here, the combinatorial quantity $S(\mathcal{A}, n)$, called the n^{th} shatter coefficient of the class $\mathcal{A}$, measures the richness/massiveness of the class $\mathcal{A}$, and is always bounded by 2^n. The following result is well-known and goes back to Vapnik and Chervonenkis (1971).

Theorem 1. *Let μ and μ_n be as above. Then, for every $\epsilon > 0$ and every $n \geq 1$,*

$$P\left\{\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| > \epsilon\right\} \leq 8\, S(\mathcal{A}, n)\, e^{-n\,\epsilon^2/32} .$$

For a proof of the above result based on symmetrization arguments see, for example, Devroye et al. (1996). Using more complicated techniques, it is possible to improve the constants that appear in the exponent of Theorem 1; see Devroye (1982) for more on this. Other relevent results along these lines are those of Talagrand (1994), Dudley (1978), and Massart (1990).

Remark 1. *When the collection of sets $\mathcal{A}$ is uncountable, the measurability of the supremum in the above theorem can become an important issue. One way to deal with the measurability problem, in general, is to work with outer probability; see, for example, the monograph by van der Vaart and Wellner (1996). In the rest of this article we shall assume that the supremum functionals do satisfy measurability conditions.*

In passing we also note that if $n^{-1}\log(S(\mathcal{A}, n)) \to 0$, as $n \to \infty$, then by the Borel-Cantelli lemma $\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \xrightarrow{a.s.} 0$, i.e. the so-called uniform law of large numbers.

To deal with the general case where there are missing $\mathbf{Z}_i$'s (recall that $\mathbf{X}^T = (\mathbf{Y}^T, \mathbf{Z}^T)$) among the data, we start by defining the random variable

$$\delta = \begin{cases} 1 & \text{if } \mathbf{Z} \text{ is not missing} \\ 0 & \text{otherwise}\,. \end{cases} \tag{3}$$

Then the data $\mathcal{D}_n$ may be represented by

$$\mathcal{D}_n = \{(\mathbf{X}_1, \delta_1), \ldots, (\mathbf{X}_n, \delta_n)\} = \{(\mathbf{Y}_1, \mathbf{Z}_1, \delta_1), \ldots, (\mathbf{Y}_n, \mathbf{Z}_n, \delta_n)\}\ .$$

Clearly, the estimator μ_n in (2) is no longer computable because some of the $\mathbf{Z}_i$'s may be missing. In order to revise (2) appropriately we also need to take into account the *missing probability mechanism*, i.e., the quantity $P\{\delta = 1 \,|\, \mathbf{Y}, \mathbf{Z}\} = E(\delta \,|\, \mathbf{Y}, \mathbf{Z})$. If the missing probability mechanism satisfies

$$P\{\delta = 1 \,|\, \mathbf{Y}, \mathbf{Z}\} = P\{\delta = 1\}$$

then it is said to be *Missing Completely At Random*, (MCAR). Of course the MCAR assumption is rather unrealistic and restrictive. A more widely used assumption in the literature is the *Missingness At Random* assumption, MAR, where one has

$$P\{\delta = 1 \,|\, \mathbf{Y}, \mathbf{Z}\} = P\{\delta = 1 \,|\, \mathbf{Y}\}\ , \tag{4}$$

i.e., the probability that $\mathbf{Z}$ is missing does not depend on $\mathbf{Z}$ itself. For more on these and other missing patterns one may refer to Little and Rubin (2002), p. 12. When the missing probability mechanism satisfies the MCAR assumption (unrealistic), one may just use the complete cases to estimate $\mu(A)$; (a complete case is a case where $\delta_i = 1$). For example, if $p := P\{\delta = 1\} \neq 0$ then one may consider the simple estimator

$$\widetilde{\mu}_n(A) = \frac{1}{n\,p}\sum_{i=1}^{n} \delta_i\, I\{t(\mathbf{X}_i) \in A\}\ , \tag{5}$$

provided that p is known. Under the MCAR assumption, the estimator (5) is unbiased for $\mu(A)$. To appreciate this simply observe that $E(\widetilde{\mu}_n(A)) = (n\,p)^{-1}\sum_{i=1}^{n} E\left[E\left(\delta_i\, I\{t(\mathbf{X}_i) \in A\} \,|\, \mathbf{X}_i\right)\right] = (n\,p)^{-1}\sum_{i=1}^{n} P\{t(\mathbf{X}_i) \in A\}\, p = \mu(A)$, because under the MCAR assumption, $E(\delta \,|\, \mathbf{X}) = P(\delta = 1) = p$. In fact more is true:

Theorem 2. *Let $\widetilde{\mu}_n(A)$ be as in (5) then for every $\epsilon > 0$ and every $n \geq 1$*

$$P\left\{\sup_{A\in\mathcal{A}} |\widetilde{\mu}_n(A) - \mu(A)| > \epsilon\right\} \leq c_1\, S(\mathcal{A},\, n)\, e^{-c_2\, n\, \epsilon^2}\,,$$

where $c_1 = 8$ and $c_2 = 2^{-1}p^2$.

When $p = P(\delta = 1)$ is unknown, it may be replaced by $\overline{p} = n^{-1}\sum_{i=1}^{n}\delta_i$ in (5), and the bound in Theorem 2 continues to hold with different constants $c_1 > 0$ and $c_2 > 0$.

In the rest of this article we shall focus on the popular (and more realistic) assumption of MAR missing mechanism, given by (4). In this case the empirical version of $\mu(A)$, given by

$$\widetilde{\mu}_n(A) = \frac{1}{n'}\sum_{i=1}^{n}\delta_i\, I\left\{t(\mathbf{X}_i) \in A\right\}\,,$$

where n' may be taken to be $n\,p$ or $\sum_{i=1}^{n}\delta_i$ or n, is no longer appropriate. This is because the resulting set-indexed empirical process $\{\widetilde{\mu}_n(A) - \mu(A)\,|\,A \in \mathcal{A}\}$ is not centered - not even asymptotically.

In the next section we propose methods for estimating $\mu(A)$, uniformly (in A), and derive counterparts of Theorem 1 for our proposed estimators. As an immediate consequence of our results, one can establish various Glivenko-Cantelli type theorems for incomplete data under the MAR assumption.

2. Main results

In this section we consider procedures to correct the naive estimator

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i\, I\left\{t(\mathbf{X}_i) \in A\right\}\,,$$

where the correction is done by weighting the complete cases by the inverse of the missing data probabilities $p(\mathbf{Y}) := P\left\{\delta = 1\,|\,\mathbf{Y}\right\}$, or its estimates. We recall that under the MAR assumption $P\left\{\delta = 1\,|\,\mathbf{Y},\,\mathbf{Z}\right\} = P\{\delta = 1\,|\,\mathbf{Y}\} := p(\mathbf{Y})$. To motivate our approaches we first consider the simple (but unrealistic) case where the function $p(\mathbf{y})$ is completely known. Now consider the revised estimator

$$\overline{\mu}_n(A) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{p(\mathbf{Y}_i)}\, I\left\{t(\mathbf{X}_i) \in A\right\}\,. \tag{6}$$

How good is $\overline{\mu}_n(A)$ as an estimator of $\mu(A)$? Clearly under the MAR assumption $E\left(\overline{\mu}_n(A)\right) = \mu(A)$. More importantly, we have

Theorem 3. *Let $\overline{\mu}_n$ be as above and suppose that $p_{min} = \min_{\mathbf{y}} p(\mathbf{y}) > 0$. Then for every $\epsilon > 0$ and every $n \geq 1$,*

$$P\left\{\sup_{A\in\mathcal{A}} |\overline{\mu}_n(A) - \mu(A)| > \epsilon\right\} \leq 8\, S(\mathcal{A},\, n)\, e^{-c_3\, n\, \epsilon^2},$$

where $c_3 = 2^{-1}\, p_{\min}^2 > 0$.

Of course in practice the function $p(\mathbf{y}) = P\{\delta = 1 \,|\, \mathbf{Y} = \mathbf{y}\} = E\,(\delta \,|\, \mathbf{Y} = \mathbf{y})$ is almost always unknown and must first be estimated. Here, we consider two possible estimators of $p(\mathbf{y})$: the first one is a kernel regression function estimator, whereas the second approach is based on the least-squares method.

2.1. A kernel-based method

Our first estimator of $p(\mathbf{Y}_i) = E(\delta_i \,|\, \mathbf{Y}_i)$ in (6) is

$$\widehat{p}(\mathbf{Y}_i) = \frac{\sum\limits_{j=1,\neq i} \delta_j\, \mathcal{K}\left(\frac{\mathbf{Y}_j - \mathbf{Y}_i}{h_n}\right)}{\sum\limits_{j=1,\neq i} \mathcal{K}\left(\frac{\mathbf{Y}_j - \mathbf{Y}_i}{h_n}\right)}, \tag{7}$$

with the convention $0/0 = 0$, where the function $\mathcal{K} : \mathbb{R}^d \mapsto \mathbb{R}$ is the kernel with smoothing parameter $h_n\, (\to 0,\ \text{as } n \to \infty)$. We then estimate $\mu(A) = P\,\{t(\mathbf{X}) \in A\}$ by

$$\widehat{\mu}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\widehat{p}(\mathbf{Y}_i)} I\,\{t(\mathbf{X}_i) \in A\}\ . \tag{8}$$

To study $\widehat{\mu}_n$ we first state some conditions.

C 1. $p_{\min} = \min_{\mathbf{y}} p(\mathbf{y}) > 0$.

C 2. The random vector $\mathbf{Y}$ has a compactly supported probability density function $f(\mathbf{y})$ and is bounded away from zero on its compact support. Furthermore, both f and its first-order partial derivatives are uniformly bounded.

C 3. The kernel $\mathcal{K}$ satisfies $\int_{\mathbb{R}^d} \mathcal{K}(\mathbf{u})\, d\mathbf{u} = 1$ and $\int_{\mathbb{R}^d} |\, u_i \,|\, \mathcal{K}(\mathbf{u})\, d\mathbf{u} \leq \infty$, for $i = 1, \ldots, d$. Furthermore, $h_n \to 0$ and $n\, h_n^d \to \infty$, as $n \to \infty$.

C 4. The partial derivatives $\frac{\partial}{\partial\, y_i} p(\mathbf{y})$ exist for $i = 1, \ldots, d$, and are bounded uniformly in $\mathbf{y}$ on the support of f.

Condition **C1** essentially states that the probability that $\mathbf{Z}$ can be observed (i.e., δ=1) will be nonzero (for all $\mathbf{Y} = \mathbf{y}$). Condition **C2** is often imposed in nonparametric regression in order to avoid having unstable estimates (in the tails of the pdf f of $\mathbf{Y}$). Condition **C3** is not a restriction since the choice of the kernel $\mathcal{K}$ is at our discretion. In fact, $\mathcal{K}$ only needs to be a proper density with a finite first absolute moment. Condition **C4**, which has also been used by Cheng and Chu (1996), p. 65, is technical.

Theorem 4. *Let $\widehat{\mu}_n(A)$ be as in (8). Then, under conditions C1-C4, for every $\epsilon > 0$ there is an n_o such that for all $n > n_o$,*

$$P\left\{\sup_{A\in\mathcal{A}} |\widehat{\mu}_n(A) - \mu(A)| > \epsilon\right\} \leq 8\, S(\mathcal{A},\, n)\, e^{-c_4\, n\, \epsilon^2} + 4\, n\, e^{-c_5\, n\, h_n^d\, \epsilon^2} + 4\, n\, e^{-c_6\, n\, h_n^d},$$

where c_4, c_5, and c_6 are positive constants not depending on n or ϵ.

The constants c_4, c_5, and c_6 that appear in Theorem 4 depend on the function p (as well as many other terms) through $p_{\min} = \min_{\mathbf{y}} p(\mathbf{y})$. In fact, the proof of the theorem makes it clear (with some more efforts) that one can always take

$$\begin{aligned} c_4 &= p_{\min}^2/8 \\ c_5 &= \frac{f_{\min}^2 p_{\min}^2}{8 \parallel \mathcal{K} \parallel_\infty (256 \parallel f \parallel_\infty + f_{\min}/6)} \wedge \frac{f_{\min}^2 p_{\min}^4}{2048 \parallel \mathcal{K} \parallel_\infty \parallel f \parallel_\infty} \\ c_6 &= \frac{4 f_{\min}^2}{8 \parallel \mathcal{K} \parallel_\infty (256 \parallel f \parallel_\infty + f_{\min}/6)} \wedge \frac{4 f_{\min}^2 p_{\min}^2}{2048 \parallel \mathcal{K} \parallel_\infty \parallel f \parallel_\infty}. \end{aligned}$$

The estimator $\widehat{\mu}_n$ defined via (8) and (7) is quite easy to compute in practice. However, the bound in Theorem 4 is not as tight as the one in Theorem 3. This is because of the presence of the term $n\, h_n^d$ that appears in the exponent of the bound of Theorem 4. In a sense, this shows that the effective sample size for the results of Theorem 4 is $n\, h_n^d$ (and not n).

2.2. The least-squares method

Our second method uses the least-squares estimator of $p(\mathbf{y})$ to construct an empirical version of $\mu(A)$. More specifically, suppose that the regression function $p(\mathbf{y}) = E(\delta \,|\, \mathbf{Y} = \mathbf{y})$ belongs to a class $\mathcal{P}$ of functions $p : \mathbb{R}^d \mapsto [p_{\min},\, 1]$, where, as before, $p_{\min} = \min_{\mathbf{y}} p(\mathbf{y})$. Also, let $\breve{p}_{_{LS}}(\mathbf{y})$ be the least-squares estimator of $p(\mathbf{y})$, i.e.,

$$\breve{p}_{_{LS}} = \operatorname*{argmin}_{p\in\mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} |\, \delta_i - p(\mathbf{Y}_i)\, |^2 .$$

Then we have the following counterpart of (8).

$$\breve{\mu}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\breve{p}_{_{LS}}(\mathbf{Y}_i)} I\left\{t(\mathbf{X}_i) \in A\right\}. \tag{9}$$

To assess the performance of $\breve{\mu}_n(A)$, we employ results from the empirical process theory: For fixed $\mathbf{y}_1, \ldots, \mathbf{y}_n$, let $\mathcal{N}_1\,(\epsilon,\, \mathcal{P},\, (\mathbf{y}_i)_{i=1}^n)$ be the ϵ-covering number of $\mathcal{P}$ with respect to the empirical measure of the points $\mathbf{y}_1, \ldots, \mathbf{y}_n$. That is $\mathcal{N}_1\,(\epsilon,\, \mathcal{P},\, (\mathbf{y}_i)_{i=1}^n)$ is the cardinality of the smallest subclass of functions $\mathcal{P}_\epsilon = \{p_1, \ldots, p_{_{N(\epsilon)}} \,|\, p_i : \mathbb{R}^d \mapsto [p_{\min},\, 1]\}$ with the property that for every $p \in \mathcal{P}$ there is a $p^* \in \mathcal{P}_\epsilon$ such that $n^{-1} \sum_{i=1}^{n} |\, p(\mathbf{y}_i) - p^*(\mathbf{y}_i)\, | < \epsilon$. (See, for example, van der Vaart and Wellner (1996) p. 83, or Pollard (1984), p. 25.) The following result gives bounds on the uniform deviations of $\breve{\mu}_n(A)$ from $\mu(A)$.

Theorem 5. *Let $\breve{\mu}_n$ be as in (9). Then, under condition C1, for every $\epsilon > 0$ and every $n \geq 1$,*

$$P\left\{\sup_{A\in\mathcal{A}} |\breve{\mu}_n(A) - \mu(A)| > \epsilon\right\} \leq 8\, S(\mathcal{A}, n)\, e^{-c_7\, n\, \epsilon^2} + 8\, E\left[\mathcal{N}_1\left(p_{\min}^2\, \epsilon/32, \mathcal{P}, (\mathbf{Y}_i)_{i=1}^n\right)\right] e^{-c_8\, n\, \epsilon^2} + 8\, E\left[\mathcal{N}_1\left(p_{\min}^4\, \epsilon^2/256, \mathcal{P}, (\mathbf{Y}_i)_{i=1}^n\right)\right] e^{-c_9\, n\, \epsilon^4},$$

where $c_7 = p_{\min}^2/8$, $c_8 = p_{\min}^4/2048$, and $c_9 = p_{\min}^8/((16^2)(128))$.

The two methods of estimation discussed in this section are of course very different and the performance of each one depends on the function p. More specifically, if one is certain that p belongs to a known class of functions $\mathcal{P}$, then the least-squares method would be preferable. In this case the performance bound of Theorem 5 is nonasymptotic in that it holds for every $n \geq 1$. Unfortunately, if $p \notin \mathcal{P}$ then the conclusion of Theorem 5 will be incorrect. In this case (and in the general case where one has no knowledge of the class of functions $\mathcal{P}$), one can use the kernel estimator instead. There are, however, some theoretical drawbacks here: Theorem 4 is only asymptotic (it holds for large n). Furthermore, the kernel estimator requires more regularity conditions for the function p, (as reflected by Theorem 4).

2.3. An application

The results developed in Section 2 can be used to estimate a distribution function in the presence of missing covariates. This was briefly explained in the introduction section. Here we consider an application of our results to the problem of statistical classification. More specifically, let $(\mathbf{X}, W)$ be an $\mathbb{R}^{d+p} \times \{0, 1\}$-valued random pair, where $\mathbf{X} = (\mathbf{Y}^T, \mathbf{Z}^T)^T$, with $\mathbf{Y} \in \mathbb{R}^d$, $d \geq 1$, and $\mathbf{Z} \in \mathbb{R}^p$. The problem of statistical classification involves the prediction of W based on the vector of covariates $\mathbf{X}$. Formally, one seeks to find a classifier (a function) $\Psi : \mathbb{R}^{d+p} \mapsto \{0, 1\}$ for which the probability of misclassification (incorrect prediction), i.e., $P\{\Psi(\mathbf{X}) \neq W\}$ is as small as possible. It is a simple exercise to verify that the best classifier, i.e., the classifier with lowest misclassification probability, is given by

$$\Psi_B(\mathbf{x}) = \begin{cases} 1 & \text{if } P(W = 1 \,|\, \mathbf{X} = \mathbf{x}) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Since Ψ_B is virtually always unknown, one uses the data to construct a classifier. Given a random sample $\mathcal{D}_n = \{(\mathbf{X}_1, W_1), \ldots, (\mathbf{X}_n, W_n)\}$, where each pair $(\mathbf{X}_i, W_i)$ is fully observable, one tries to construct a classifier $\widehat{\Psi}_n$ in such a way that its misclassification error $L_n(\widehat{\Psi}_n) = P\left\{\widehat{\Psi}_n(\mathbf{X}) \neq W \,|\mathcal{D}_n\right\}$ is in some sense as small as possible. Let $L(\Psi) = P\{\Psi(\mathbf{X}) \neq W\}$. If $L_n(\widehat{\Psi}_n) \xrightarrow{a.s.}$

$\inf_{\Psi:\, \mathbb{R}^{d+p} \mapsto \{0,1\}} L(\Psi)$, we say $\widehat{\Psi}_n$ is strongly consistent. If the consequence holds in probability, $\widehat{\Psi}_n$ is said to be weakly consistent. Given a class $\mathbf{\Psi}$ of candidate classifiers Ψ, the principal of empirical risk minimization (ERM) chooses the classifier (from $\mathbf{\Psi}$) that minimizes the empirical error

$$\frac{1}{n}\sum_{i=1}^{n} I\left\{\Psi(\mathbf{X}_i) \neq W_i\right\} . \tag{10}$$

Now, consider the case where $\mathbf{Z}_i$ may be missing in $\mathbf{X}_i = (\mathbf{Y}_i^T, \mathbf{Z}_i^T)^T$. Clearly, the data can be represented by

$$\mathcal{D}_n = \{(\mathbf{Y}_1, \mathbf{Z}_1, W_1, \delta_1), \ldots, (\mathbf{Y}_n, \mathbf{Z}_n, W_n, \delta_n)\} ,$$

where $\delta_i = 0$ if $\mathbf{Z}_i$ is missing, (otherwise $\delta_i = 1$). First note that (10) cannot be computed because not every $\mathbf{X}_i$ is fully observable. Furthermore, using the complete cases alone will not work because

$$\frac{1}{n}\sum_{i=1}^{n} \delta_i\, I\left\{\Psi(\mathbf{X}_i) \neq W_i\right\}$$

is the empirical version of $E\left(\delta\, I\left\{\Psi(\mathbf{X}) \neq W\right\}\right)$, which is not the same as $P\left\{\Psi(\mathbf{X}) \neq W\right\}$. Our propose ERM-type estimator of the best classifier is given by

$$\widetilde{\Psi}_n = \underset{\Psi\in\mathbf{\Psi}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\widetilde{p}(\mathbf{Y}_i)}\, I\left\{\Psi(\mathbf{X}_i) \neq W_i\right\} , \tag{11}$$

where $\widetilde{p}(\mathbf{y})$ is either $p(\mathbf{y})$, (if $p(\mathbf{y})$ is known), or $\widehat{p}(\mathbf{y})$ in (7) or $\breve{p}_{LS}(\mathbf{y})$ of Section 2.2. Let Ψ^* be the best classifier in $\mathbf{\Psi}$, i.e. Ψ^* satisfies

$$P\left\{\Psi^*(\mathbf{X}) \neq W\right\} = \inf_{\Psi\in\mathbf{\Psi}} P\left\{\Psi(\mathbf{X}) \neq W\right\} .$$

How good is $\widetilde{\Psi}_n$ as an estimator of Ψ^*? To answer this question, let $L_n(\widetilde{\Psi}_n)$ be the error of the classifier $\widetilde{\Psi}_n$, i.e., $L_n(\widetilde{\Psi}_n) = P\left\{\widetilde{\Psi}_n(\mathbf{X}) \neq W \,\middle|\, \mathcal{D}_n\right\}$. Also let $\mathcal{A}_{\mathbf{\Psi}}$ be the class of all sets A of the from

$$A = \left\{\left\{\mathbf{x} \,\middle|\, \Psi(\mathbf{x}) = 1\right\} \times \{0\}\right\} \bigcup \left\{\left\{\mathbf{x} \,\middle|\, \Psi(\mathbf{x}) = 0\right\} \times \{1\}\right\}, \quad \Psi \in \mathbf{\Psi} .$$

Then, we have the following result.

Theorem 6. *Let $\widetilde{\Psi}_n$ be as above. Then for every $\epsilon > 0$ there is an $n_o > 0$ such that for every $n > n_o$*

$$P\left\{L_n(\widetilde{\Psi}_n) - \inf_{\Psi\in\mathbf{\Psi}} L(\Psi) > \epsilon\right\} \leq 8\, S(\mathcal{A}_{\mathbf{\Psi}}, n)\, e^{-c_{19}\, n\, \epsilon^2} + r_n(\epsilon) ,$$

where

(i) if $\widetilde{p}(\mathbf{Y}_i) = \widehat{p}(\mathbf{Y}_i)$*, then under the conditions of Theorem 4,*

$$r_n(\epsilon) = 4\, n\, e^{-c_{20}\, n\, h_n^d\, \epsilon^2} + 4\, n\, e^{-c_{21}\, n\, h_n^d}\,,$$

and

(ii) if $\widetilde{p}(\mathbf{Y}_i) = \breve{p}_{_{LS}}(\mathbf{Y}_i)$*, then under the conditions of Theorem 5*

$$r_n(\epsilon) = 8\, E\left[\mathcal{N}_1\left(\frac{p_{\min}^2\, \epsilon}{64},\, \mathcal{P},\, (\mathbf{Y}_i)_{i=1}^n\right)\right] e^{-c_{22}\, n\, \epsilon^2} + 8\, E\left[\mathcal{N}_1\left(\frac{p_{\min}^4\, \epsilon^2}{1024},\, \mathcal{P},\, (\mathbf{Y}_i)_{i=1}^n\right)\right] e^{-c_{23}\, n\, \epsilon^4}\,.$$

Here c_{20}*,* c_{21}*,* c_{22}*, and* c_{23} *are positive constants not depending on* n *or* ϵ*.*

Remark 2. *Theorem 6 is based on the assumption that the missing probability mechanism satisfies the strong MAR assumption that* $P\{\delta_i = 1 \,|\, \mathbf{X}_i,\, W_i\} = P\{\delta_i = 1 \,|\, \mathbf{Y}_i\}$ *,* $i = 1, \ldots, n$*, where as before,* $\mathbf{X}_i = (\mathbf{Y}_i^T,\, \mathbf{Z}_i^T)^T$*. It is possible to relax this assumption to* $P\{\delta_i = 1 \,|\, \mathbf{X}_i,\, W_i\} = P\{\delta_i = 1 \,|\, \mathbf{Y}_i,\, W_i\}$ *, i.e., the probability that* $\mathbf{Z}_i$ *is missing can depend on both* $\mathbf{Y}_i$ *and* W_i*. In this case, one can revise* $\widetilde{\Psi}_n$ *in (11) by*

$$\widetilde{\Psi}_n = \underset{\Psi \in \mathbf{\Psi}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\widetilde{p}(\mathbf{Y}_i,\, W_i)} I\{\Psi(\mathbf{X}_i) \neq W_i\}\,,$$

where $\widetilde{p}(\mathbf{Y}_i,\, W_i)$ *is the estimator of* $p(\mathbf{Y}_i,\, W_i) = P\{\delta_i = 1 \,|\, \mathbf{Y}_i,\, W_i\}$*. Taking, for example,*

$$\widetilde{p}(\mathbf{Y}_i,\, W_i) = \frac{\sum\limits_{j=1,\, \neq i}^n \delta_j\, I\{W_j = W_i\}\, \mathcal{K}\left(\frac{\mathbf{Y}_j - \mathbf{Y}_i}{h_n}\right)}{\sum\limits_{j=1,\, \neq i}^n I\{W_j = W_i\}\, \mathcal{K}\left(\frac{\mathbf{Y}_j - \mathbf{Y}_i}{h_n}\right)}\,,$$

one can show, with some more efforts, that the conclusion of part (i) of Theorem 6 continues to hold with different constants $c_{20} > 0$ *and* $c_{21} > 0$*. Similar results can also be established for part (ii) of Theorem 6.*

In passing we also note that the size of $S(\mathcal{A}_{\mathbf{\Psi}},\, n)$ depends on the underlying class $\mathbf{\Psi}$. When, for example, $\mathbf{\Psi}$ is the popular class of linear classifiers

$$\Psi(\mathbf{x}) = \begin{cases} 1 & \text{if } a_{_0} + a_{_1}\, x_{_1} + \cdots + a_{_{d+p}}\, x_{_{d+p}} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $a_{_0},\, a_{_1},\, \ldots,\, a_{_{d+p}} \in \mathbb{R}$, then $S(\mathcal{A}_{\mathbf{\Psi}},\, n) \leq n^{d+p+1}$, (see, for example, chapter 13 of Devroye et al. (1996).). In this case, if we choose $\widetilde{p}(\mathbf{y}) = \widehat{p}(\mathbf{y})$, where $\widehat{p}$ is as in (7), then $L_n(\widetilde{\Psi}_n) \xrightarrow{a.s.} \inf\limits_{\Psi \in \mathbf{\Psi}} L(\Psi)$, as $n \to \infty$, provided that $\frac{log(n)}{n\, h_n^d} \to 0$.

Remark 3. *Strictly speaking, the classifier* $\widetilde{\Psi}_n$ *is not suitable if the new observation* $\mathbf{X}$*, (based on which* W *has to be predicted), also has missing covariates. I.e., in addition to the data* $\mathcal{D}_n$*, the new observation* $\mathbf{X} = (\mathbf{Y}^T,\, \mathbf{Z}^T)^T$ *is also allowed to have a missing* $\mathbf{Z}$*.*

3. Proofs

Proof of Theorem 3

The following proof employes the symmetrization argument of Dudley (1978), p. 925, and Pollard (1984), sec. II.3. (Also see van der Vaart and Wellner (1996), sec. 2.3.)

Let $\mathcal{D}'_n = \{(\mathbf{Y}'_1, \mathbf{Z}'_1, \delta'_1), \ldots, (\mathbf{Y}'_n, \mathbf{Z}'_n, \delta'_n)\}$ be a hypothetical sample, independent of $\mathcal{D}_n$, where $(\mathbf{Y}'_i, \mathbf{Z}'_i, \delta'_i) \overset{\text{i.i.d.}}{=} (\mathbf{Y}_1, \mathbf{Z}_1, \delta_1)$, $i = 1, \ldots, n$. Also, define

$$\overline{\mu}'_n(A) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta'_i}{p(\mathbf{Y}'_i)} I\{t(\mathbf{X}'_i) \in A\} .$$

Now, fix the data $\mathcal{D}_n$ and note that if $\sup_{A \in \mathcal{A}} |\overline{\mu}_n(A) - \mu(A)| > \epsilon$ then there is at least one set $A_\epsilon \in \mathcal{A}$ which depends on $\mathcal{D}_n$ (but not $\mathcal{D}'_n$) such that $|\overline{\mu}_n(A_\epsilon) - \mu(A_\epsilon \,|\, \mathcal{D}_n)| > \epsilon$, where

$$\mu(A_\epsilon \,|\, \mathcal{D}_n) = E\left[I\{t(\mathbf{X}) \in A_\epsilon\} \,|\, \mathcal{D}_n\right] .$$

Next, observe that

$$\begin{aligned}
P\left\{|\overline{\mu}'_n(A_\epsilon) - \mu(A_\epsilon \,|\, \mathcal{D}_n)| < \frac{\epsilon}{2} \,|\, \mathcal{D}_n\right\} &\geq 1 - \sup_{A \in \mathcal{A}} P\left\{|\overline{\mu}'_n(A) - \mu(A)| \geq \frac{\epsilon}{2}\right\} \\
&\geq 1 - \frac{4}{n\,\epsilon^2} \sup_{A \in \mathcal{A}} \mathrm{Var}\left(\frac{\delta_1}{p(\mathbf{Y}_1)} I\{t(\mathbf{X}_1) \in A\}\right) \\
&\quad \text{(via Chebysheff's inequality)} \\
&\geq 1 - \frac{4}{n\,p_{\min}^2\epsilon^2} \sup_{A \in \mathcal{A}} \left[E\left(\delta_1 I\{t(\mathbf{X}_1) \in A\}\right)\left(1 - E\left(\delta_1 I\{t(\mathbf{X}_1) \in A\}\right)\right)\right] \\
&\quad \text{(because } \delta\, I\{t(\mathbf{X}) \in A\} \text{ is a Bernoulli r.v.)} \\
&\geq 1 - \frac{4}{n\,p_{\min}^2\epsilon^2} \cdot \frac{1}{4} \geq \frac{1}{2}, \qquad \text{for } n\,\epsilon^2 \geq \frac{2}{p_{\min}^2} .
\end{aligned}$$

Therefore, for $n\,\epsilon^2 \geq 2/p_{\min}^2$

$$\begin{aligned}
\frac{1}{2} &\leq P\left\{|\overline{\mu}'_n(A_\epsilon) - \mu(A_\epsilon \,|\, \mathcal{D}_n)| < \frac{\epsilon}{2} \,\middle|\, \mathcal{D}_n\right\} \\
&\leq P\left\{-|\overline{\mu}'_n(A_\epsilon) - \overline{\mu}_n(A_\epsilon)| + \underbrace{|\overline{\mu}_n(A_\epsilon) - \mu(A_\epsilon \,|\, \mathcal{D}_n)|}_{>\epsilon} < \frac{\epsilon}{2} \,\middle|\, \mathcal{D}_n\right\} \\
&\leq P\left\{\sup_{A \in \mathcal{A}} |\overline{\mu}'_n(A) - \overline{\mu}_n(A)| > \frac{\epsilon}{2} \,\middle|\, \mathcal{D}_n\right\} . \qquad (12)
\end{aligned}$$

But the far left and far right sides of (12) do not depend on any particular A_ϵ and the chain of inequalities between them remain valid on the set

$\left\{ \sup_{A\in\mathcal{A}} \left| \overline{\mu}_n(A) - \mu(A) \right| > \epsilon \right\}$. Therefore, integrating the two sides with respect to the distribution of $\mathcal{D}_n$, over this set, one finds

$$P\left\{ \sup_{A\in\mathcal{A}} \left| \overline{\mu}_n(A) - \mu(A) \right| > \epsilon \right\} \leq 2\, P\left\{ \sup_{A\in\mathcal{A}} \left| \overline{\mu}'_n(A) - \overline{\mu}_n(A) \right| > \frac{\epsilon}{2} \right\}, \quad (13)$$

Next, let $R_1, \ldots, R_n$ be i.i.d. random variables, independent of $\mathcal{D}_n$ and $\mathcal{D}'_n$, where $P\{R_i = +1\} = P\{R_i = -1\} = 1/2$, and observe that

$$\begin{aligned}
&P\left\{ \sup_{A\in\mathcal{A}} \left| \overline{\mu}'_n(A) - \overline{\mu}_n(A) \right| > \frac{\epsilon}{2} \right\} \\
&= P\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \left[\frac{\delta'_i}{p(\mathbf{Y}'_i)} I\left\{ t(\mathbf{X}'_i) \in A \right\} - \frac{\delta_i}{p(\mathbf{Y}_i)} I\left\{ t(\mathbf{X}_i) \in A \right\} \right] \right| > \frac{\epsilon}{2} \right\} \\
&= P\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} R_i \left[\frac{\delta'_i}{p(\mathbf{Y}'_i)} I\left\{ t(\mathbf{X}'_i) \in A \right\} - \frac{\delta_i}{p(\mathbf{Y}_i)} I\left\{ t(\mathbf{X}_i) \in A \right\} \right] \right| > \frac{\epsilon}{2} \right\} \\
&\leq 2\, P\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \frac{R_i\, \delta_i}{p(\mathbf{Y}_i)} I\left\{ t(\mathbf{X}_i) \in A \right\} \right| > \frac{\epsilon}{4} \right\}.
\end{aligned}$$

This last expression together with (13) yield (upon conditioning on $\mathcal{D}_n$),

$$\begin{aligned}
&P\left\{ \sup_{A\in\mathcal{A}} \left| \overline{\mu}_n(A) - \mu(A) \right| > \epsilon \right\} \\
&\leq 4\, E\left[P\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \frac{R_i\, \delta_i}{p(\mathbf{Y}_i)} I\left\{ t(\mathbf{X}_i) \in A \right\} \right| > \frac{\epsilon}{4} \,\middle|\, \mathcal{D}_n \right\} \right]. \quad (14)
\end{aligned}$$

Now observe that for fixed $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the number of different vectors

$$\left(I\left\{ t(\mathbf{x}_1) \in A \right\}, \ldots, I\left\{ t(\mathbf{x}_n) \in A \right\} \right)$$

obtained, as A ranges over $\mathcal{A}$, is just the number of different sets in

$$\left\{ \{ t(\mathbf{x}_1), \ldots, t(\mathbf{x}_n) \} \cap A \,|\, A \in \mathcal{A} \right\},$$

and this number is bounded by $S(\mathcal{A}, n)$. Thus

$$\begin{aligned}
&P\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \frac{R_i\, \delta_i}{p(\mathbf{Y}_i)} I\left\{ t(\mathbf{X}_i) \in A \right\} \right| > \frac{\epsilon}{4} \,\middle|\, \mathcal{D}_n \right\} \\
&\leq S(\mathcal{A}, n) \cdot \sup_{A\in\mathcal{A}} P\left\{ \frac{1}{n} \left| \sum_{i=1}^{n} \frac{R_i\, \delta_i}{p(\mathbf{Y}_i)} I\left\{ t(\mathbf{X}_i) \in A \right\} \right| > \frac{\epsilon}{4} \,\middle|\, \mathcal{D}_n \right\}. \quad (15)
\end{aligned}$$

Since, conditional on $\mathcal{D}_n$, the term $\sum_{i=1}^{n} \frac{R_i\, \delta_i}{p(\mathbf{Y}_i)} I\left\{ t(\mathbf{X}_i) \in A \right\}$ is the sum of n independent zero-mean random variables, bounded by $-1/p_{\min}$ and $+1/p_{\min}$ (where

$p_{\min} = \min_{\mathbf{y}} p(\mathbf{y}) > 0$), one can use Hoeffding's inequality to conclude

$$P\left\{\frac{1}{n}\Big|\sum_{i=1}^{n}\frac{R_i\,\delta_i}{p(\mathbf{Y}_i)}\,I\left\{t(\mathbf{X}_i)\in A\right\}\Big| > \frac{\epsilon}{4}\,\Big|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\right\} \le 2\,e^{-(p_{\min}^2/2)\,n\,\epsilon^2}.$$

The above bound together with (14) and (15) give

$$P\left\{\sup_{A\in\mathcal{A}}\left|\overline{\mu}_n(A)-\mu(A)\right| > \epsilon\right\} \le 8\,S(\mathcal{A},\,n)\,e^{-(p_{\min}^2/2)\,n\,\epsilon^2}.$$

This proves the theorem for $n\,\epsilon^2 \ge 2/p_{\min}^2$. When $n\,\epsilon^2 < 2/p_{\min}^2$ the theorem is trivially true. □

To prove Theorem 4, we first state a lemma.

Lemma 1. *Put* $\widehat{\phi}(\mathbf{Y}_i) = (n-1)^{-1}\,h_n^{-d}\sum_{j=1,\neq i}^{n}\delta_j\,\mathcal{K}\left(\frac{\mathbf{Y}_j-\mathbf{Y}_i}{h_n}\right)$ *and let* $\phi(\mathbf{Y}) = f(\mathbf{Y})\,p(\mathbf{Y})$, *where* f *is the density of* $\mathbf{Y}$. *Then*

$$\left|E\left[\widehat{\phi}(\mathbf{Y})\,\middle|\,\mathbf{Y}\right]-\phi(\mathbf{Y})\right| \le c\,.\,h_n\,,$$

where c *is a positive constant not depending on* n.

Proof.

$$\begin{aligned}E\left[\widehat{\phi}(\mathbf{Y})\,\middle|\,\mathbf{Y}\right]-\phi(\mathbf{Y}) &= h_n^{-d}\,E\left[\delta_1\,\mathcal{K}\left(\frac{\mathbf{Y}_1-\mathbf{Y}}{h_n}\right)\,\middle|\,\mathbf{Y}\right]-f(\mathbf{Y})\,p(\mathbf{Y})\\ &= h_n^{-d}\,E\left[\mathcal{K}\left(\frac{\mathbf{Y}_1-\mathbf{Y}}{h_n}\right).E\left(\delta_1\,\middle|\,\mathbf{Y},\,\mathbf{Y}_1\right)\,\middle|\,\mathbf{Y}\right]-f(\mathbf{Y})\,p(\mathbf{Y})\end{aligned}$$

But, since $\mathbf{Y}$ and $\mathbf{Y}_1$ are independent, $E(\delta_1\,|\,\mathbf{Y},\,\mathbf{Y}_1) = E(\delta_1\,|\,\mathbf{Y}_1) = p(\mathbf{Y}_1)$. Thus

$$\begin{aligned}E\left(\widehat{\phi}(\mathbf{Y})\,\middle|\,\mathbf{Y}\right)-\phi(\mathbf{Y}) &= h_n^{-d}\,E\left[(p(\mathbf{Y}_1)-p(\mathbf{Y}))\,\mathcal{K}\left(\frac{\mathbf{Y}_1-\mathbf{Y}}{h_n}\right)\,\middle|\,\mathbf{Y}\right]\\ &\quad+E\left[p(\mathbf{Y})\left(h_n^{-d}\,\mathcal{K}\left(\frac{\mathbf{Y}_1-\mathbf{Y}}{h_n}\right)-f(\mathbf{Y})\right)\,\middle|\,\mathbf{Y}\right]\\ &:= G_1(\mathbf{Y})+G_2(\mathbf{Y}) \qquad (\text{say})\,.\end{aligned} \tag{16}$$

Now a one-term Taylor expansion gives

$$G_1(\mathbf{Y}) = h_n^{-d}E\left[\left(\sum_{i=1}^{d}\frac{\partial p(\mathbf{Y}^*)}{\partial Y_i}\,(Y_{1,\,i}-Y_i)\right)\,\mathcal{K}\left(\frac{\mathbf{Y}_1-\mathbf{Y}}{h_n}\right)\,\middle|\,\mathbf{Y}\right]$$

where $Y_{1,\,i}$ and Y_i are the i^{th} components of $\mathbf{Y}_1$ and $\mathbf{Y}$, and $\mathbf{Y}^*$ is a point on

the interior of the line segment joining $\mathbf{Y}$ and $\mathbf{Y}_1$. Therefore,

$$\begin{aligned}\left|G_1(\mathbf{Y})\right| &\leq c_{10} \sum_{i=1}^{d} E\left[|Y_{1,i} - Y_i| \,.\, h_n^{-d}\, \mathcal{K}\left(\frac{\mathbf{Y}_1 - \mathbf{Y}}{h_n}\right) \,\middle|\, \mathbf{Y}\right] \\ &\qquad \left(\text{where } c_{10} = \bigvee_{i=1}^{d} \sup_{\mathbf{y}} \left|\frac{\partial p(\mathbf{y})}{\partial y_i}\right|\right) \\ &= c_{10} \sum_{i=1}^{d} \int_{\mathbb{R}} |y_i - Y_i|\, h_n^{-d}\, \mathcal{K}\left(\frac{\mathbf{y} - \mathbf{Y}}{h_n}\right) f(\mathbf{y})\, d\mathbf{y} \\ &\leq c_{10}\, \|f\|_\infty \sum_{i=1}^{d} \int_{\mathbb{R}^d} h_n\, |\, y_i\, |\, \mathcal{K}(\mathbf{y})\, d\mathbf{y} = |\text{const}|\, h_n\end{aligned}$$

As for $G_2(\mathbf{Y})$, we have

$$\begin{aligned}G_2(\mathbf{Y}) &= p(\mathbf{Y}) \int_{\mathbb{R}^d} h_n^{-d}\, \mathcal{K}\left(\frac{\mathbf{y} - \mathbf{Y}}{h_n}\right) [f(\mathbf{y}) - f(\mathbf{Y})]\, d\mathbf{y} \\ &= p(\mathbf{Y}) \int_{\mathbb{R}^d} [f(\mathbf{Y} + h_n\, \mathbf{y}) - f(\mathbf{Y})]\, \mathcal{K}(\mathbf{y})\, d\mathbf{y}\,.\end{aligned}$$

Therefore, using a one-term Taylor expansion yields

$$\left|G_2(\mathbf{Y})\right| \leq \left(d\, \|f'\|_\infty \sum_{i=1}^{d} \int_{\mathbb{R}^d} |\, y_i\, |\, \mathcal{K}(\mathbf{y})\, d\mathbf{y}\right) h_n\,.$$

This completes the proof of the lemma. □

Proof of Theorem 4

First note that for every $A \in \mathcal{A}$,

$$\frac{\delta_i}{\widehat{p}(\mathbf{Y}_i)} I\{t(\mathbf{X}_i) \in A\} = \frac{\delta_i}{p(\mathbf{Y}_i)} I\{t(\mathbf{X}_i) \in A\} - \frac{\delta_i}{\widehat{p}(\mathbf{Y}_i)} I\{t(\mathbf{X}_i) \in A\} \left[\frac{\widehat{p}(\mathbf{Y}_i)}{p(\mathbf{Y}_i)} - 1\right].$$

Now let $\overline{\mu}_n(A)$ be as in (6) and observe that

$$\begin{aligned}P\left\{\sup_{A\in\mathcal{A}} |\widehat{\mu}_n(A) - \mu(A)| > \epsilon\right\} &\leq P\left\{\sup_{A\in\mathcal{A}} |\overline{\mu}_n(A) - \mu(A)| > \frac{\epsilon}{2}\right\} \\ &\quad + P\left\{\frac{1}{n}\sum_{i=1}^{n} \left|\frac{1}{\widehat{p}(\mathbf{Y}_i)}\right| \left|\frac{\widehat{p}(\mathbf{Y}_i)}{p(\mathbf{Y}_i)} - 1\right| > \frac{\epsilon}{2}\right\} \\ &:= I_n + II_n, \qquad \text{(say)}\,. \end{aligned} \tag{17}$$

But by Theorem 3,

$$I_n \leq 8\, S(\mathcal{A}, n)\, e^{-(p_{\min}^2/8)\, n\, \epsilon^2}\,. \tag{18}$$

To deal with the term II_n, first note that

$$\begin{aligned} II_n &\le P\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\Big|\frac{1}{\widehat{p}(\mathbf{Y}_i)}\Big|\frac{\widehat{p}(\mathbf{Y}_i)}{p(\mathbf{Y}_i)}-1\Big| > \frac{\epsilon}{2}\right]\bigcap\bigcap_{i=1}^{n}\left[\widehat{p}(\mathbf{Y}_i)\ge\frac{p_{\min}}{2}\right]\right\} \\ &\quad + P\left\{\bigcup_{i=1}^{n}\left[\widehat{p}(\mathbf{Y}_i)<\frac{p_{\min}}{2}\right]\right\} \\ &\le \sum_{i=1}^{n} P\left\{\frac{2}{p_{\min}^2}\Big|\widehat{p}(\mathbf{Y}_i)-p(\mathbf{Y}_i)\Big| > \frac{\epsilon}{2}\right\} + \sum_{i=1}^{n} P\left\{\widehat{p}(\mathbf{Y}_i)<\frac{p_{\min}}{2}\right\} \\ &:= II_n(1)+II_n(2), \quad \text{(say)}\,. \end{aligned} \tag{19}$$

To deal with $II_n(1)$, first note that

$$\Big|\widehat{p}(\mathbf{Y}_i)-p(\mathbf{Y}_i)\Big| = \Big|\frac{\widehat{\phi}(\mathbf{Y}_i)}{\widehat{f}(\mathbf{Y}_i)}-\frac{\phi(\mathbf{Y}_i)}{f(\mathbf{Y}_i)}\Big| \le \Big|\frac{\widehat{\phi}(\mathbf{Y}_i)-\phi(\mathbf{Y}_i)}{f(\mathbf{Y}_i)}\Big| + \Big|\frac{\widehat{f}(\mathbf{Y}_i)-f(\mathbf{Y}_i)}{f(\mathbf{Y}_i)}\Big|,$$

where $\widehat{\phi}(\mathbf{Y}_i)$ and $\phi(\mathbf{Y}_i)$ are as in Lemma 1, and $\widehat{f}(\mathbf{Y}_i)$ is the kernel density estimator of f at $\mathbf{Y}_i$, based on $\mathbf{Y}_1,\ldots,\mathbf{Y}_{i-1},\mathbf{Y}_{i+1},\ldots,\mathbf{Y}_n$. Thus

$$\begin{aligned} P\left\{\Big|\widehat{p}(\mathbf{Y}_i)-p(\mathbf{Y}_i)\Big| > \frac{p_{\min}^2\,\epsilon}{4}\right\} \le\; & P\left\{\Big|\frac{\widehat{\phi}(\mathbf{Y}_i)-\phi(\mathbf{Y}_i)}{f(\mathbf{Y}_i)}\Big| > \frac{p_{\min}^2\,\epsilon}{8}\right\} \\ & + P\left\{\Big|\frac{\widehat{f}(\mathbf{Y}_i)-f(\mathbf{Y}_i)}{f(\mathbf{Y}_i)}\Big| > \frac{p_{\min}^2\,\epsilon}{8}\right\} \\ & := \Pi_{n\,1}+\Pi_{n\,2}\,. \end{aligned} \tag{20}$$

But for n large enough, Lemma 1 implies that

$$\begin{aligned} \Pi_{n\,1} &\le P\left\{\Big|\widehat{\phi}(\mathbf{Y}_i)-E\left[\widehat{\phi}(\mathbf{Y}_i)\,\Big|\mathbf{Y}_i\right]+E\left[\widehat{\phi}(\mathbf{Y}_i)\,\Big|\mathbf{Y}_i\right]-\phi(\mathbf{Y}_i)\Big| > \frac{f_{\min}\,p_{\min}^2\,\epsilon}{8}\right\} \\ &\quad \text{(where } 0<f_{\min}:=\min_{\mathbf{y}} f(\mathbf{y}),\ \text{by condition C2)} \\ &\le P\left\{\Big|\widehat{\phi}(\mathbf{Y}_i)-E\left[\widehat{\phi}(\mathbf{Y}_i)\,\Big|\mathbf{Y}_i\right]\Big| + \frac{f_{\min}\,p_{\min}^2\,\epsilon}{16} > \frac{f_{\min}\,p_{\min}^2\,\epsilon}{8}\right\} \\ &\quad \text{(for } n \text{ large enough, by Lemma 1)} \\ &= E\left[P\left\{\frac{1}{n-1}\Big|\sum_{j=1,\neq i}^{n} T_j(\mathbf{Y}_i)\Big| > \frac{f_{\min}\,p_{\min}^2\,\epsilon}{16}\;\Big|\,\mathbf{Y}_i\right\}\right], \end{aligned}$$

where

$$T_j(\mathbf{Y}_i) = h_n^{-d}\left[\delta_j\,\mathcal{K}\left(\frac{\mathbf{Y}_j-\mathbf{Y}_i}{h_n}\right) - E\left(\delta_j\,\mathcal{K}\left(\frac{\mathbf{Y}_j-\mathbf{Y}_i}{h_n}\right)\,\Big|\,\mathbf{Y}_i\right)\right].$$

But, conditional on $\mathbf{Y}_i$, the terms T_j's are independent, zero-mean random variables, bounded by $-h_n^d\|\mathcal{K}\|_\infty$ and $+h_n^d\|\mathcal{K}\|_\infty$. Furthermore, Var $\left(T_j(\mathbf{Y}_i)\,\big|\mathbf{Y}_i\right) =$

$E\left(T_j^2(\mathbf{Y}_i)\,\middle|\mathbf{Y}_i\right) \le h_n^{-d}\,\|\mathcal{K}\|_\infty\,\|f\|_\infty$. Therefore, by Bennett's inequality (Bennett (1962)),

$$P\left\{\frac{1}{n-1}\Big|\sum_{j=1,\neq i}^{n} T_j(\mathbf{Y}_i)\Big| > \frac{f_{\min}\,p_{\min}^2\,\epsilon}{16}\;\Big|\,\mathbf{Y}_i\right\} \le 2\,\exp\left\{-\frac{(n-1)\,h_n^d\,f_{\min}^2\,p_{\min}^2\,\epsilon^2}{2\,\|\mathcal{K}\|_\infty\,(16^2\,\|f\|_\infty + f_{\min}\,p_{\min}^2\,\epsilon/48)}\right\},$$

which implies that, for n large enough, $\Pi_{n\,1} \le 2\exp\{-c(n-1)h_n^d\epsilon^2\}$, where $c > 0$ does not depend on n or ϵ. Here, we have also used the fact that in the far left side of (20) one only needs to consider $0 < \epsilon < 8/p_{\min}^2$ (because $|\widehat{p}(\mathbf{Y}_i) - p(\mathbf{Y}_i)| \le |\widehat{p}(\mathbf{Y}_i)| + |p(\mathbf{Y}_i)| \le 1 + 1 = 2$). Similarly, one can also show (in fact, with less effort) that for n large enough

$$\Pi_{n\,2} \le 2\,\exp\left\{\frac{-(n-1)h_n^d f_{\min}^2 p_{\min}^4 \epsilon^2}{1024\,\|\,\mathcal{K}\,\|_\infty \|\,f\,\|_\infty}\right\}.$$

Putting the above together, we have shown

$$II_n(1) \le 4\,n\,e^{-c_{11}\,n\,h_n^d\,\epsilon^2},$$

where $II_n(1)$ is as in (19), and $c_{11} > 0$ does not depend on n or ϵ. Finally the theorem follows by noticing that in $II_n(2)$,

$$P\left\{\widehat{p}(\mathbf{Y}_i) < p_{\min}/2\right\} \le P\left\{\left|\widehat{p}(\mathbf{Y}_i) - p(\mathbf{Y}_i)\right| > p_{\min}/2\right\}.$$

□

Proof of Theorem 5

Using the arguments employed in the proof of Theorem 4, one finds

$$P\left\{\sup_{A\in\mathcal{A}}\left|\breve{\mu}_n(A) - \mu(A)\right| > \epsilon\right\} \le I_n + II_n,$$

where I_n is as in (17) and

$$II_n = P\left\{\frac{1}{n}\sum_{i=1}^{n}\Big|\frac{1}{\breve{p}_{LS}(\mathbf{Y}_i)}\Big| \cdot \Big|\frac{\breve{p}_{LS}(\mathbf{Y}_i)}{p(\mathbf{Y}_i)} - 1\Big| > \frac{\epsilon}{2}\right\}.$$

By Theorem 3, we have $I_n \leq 8\, S(\mathcal{A},\, n)\, \exp\left\{-p_{\min}^2\, n\, \epsilon^2/8\right\}$. Also, observe that

$$\begin{aligned}
II_n &\leq P\left\{\frac{1}{n\, p_{\min}^2}\sum_{i=1}^{n}\left|\breve{p}_{_{LS}}(\mathbf{Y}_i) - p(\mathbf{Y}_i)\right| > \frac{\epsilon}{2}\right\} \\
&\leq P\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\breve{p}_{_{LS}}(\mathbf{Y}_i) - p(\mathbf{Y}_i)\right| - E\left[\left|\breve{p}_{_{LS}}(\mathbf{Y}) - p(\mathbf{Y})\right|\,\middle|\,\mathcal{D}_n\right]\right. \\
&\qquad \left. + E\left[\left|\breve{p}_{_{LS}}(\mathbf{Y}) - p(\mathbf{Y})\right|\,\middle|\,\mathcal{D}_n\right] > \frac{p_{\min}^2\,\epsilon}{2}\right\} \\
&\leq P\left\{\sup_{p'\in\mathcal{P}}\left|\frac{1}{n}\sum_{i=1}^{n}\left|p'(\mathbf{Y}_i) - p(\mathbf{Y}_i)\right| - E\left[\left|p'(\mathbf{Y}_i) - p(\mathbf{Y}_i)\right|\right]\right| > \frac{p_{\min}^2\,\epsilon}{4}\right\} \\
&\quad + P\left\{E\left[\left|\breve{p}_{_{LS}}(\mathbf{Y}) - p(\mathbf{Y})\right|\,\middle|\,\mathcal{D}_n\right] > \frac{p_{\min}^2\,\epsilon}{4}\right\} := II'_n + II''_n\,, \quad \text{(say)}\,. \qquad (21)
\end{aligned}$$

Now, using standard results from the empirical process theory (see, for example, Pollard (1984), p. 31), one finds

$$II'_n \leq 8\, E\left[\mathcal{N}_1\left(\frac{p_{\min}^2\,\epsilon}{32},\, \mathcal{P},\, (\mathbf{Y}_i)_{i=1}^n\right)\right] e^{-(p_{\min}^4/2048)n\epsilon^2}\,,$$

where $\mathcal{N}_1\left(p_{\min}^2\,\epsilon/32,\, \mathcal{P},\, (\mathbf{Y}_i)_{i=1}^n\right)$ is the ϵ-covering number of $\mathcal{P}$ with respect to the empirical measure. To deal with the term II''_n, put

$$\mathcal{S}_n(p) = \frac{1}{n}\sum_{i=1}^{n}(p(\mathbf{Y}_i) - \delta_i)^2\,,$$

and observe that by Cauchy-Schwartz inequality

$$\begin{aligned}
II''_n &\leq P\left\{E\left[\left|\breve{p}_{_{LS}}(\mathbf{Y}) - p(\mathbf{Y})\right|^2\,\middle|\,\mathcal{D}_n\right] > \frac{p_{\min}^4\,\epsilon^2}{16}\right\} \\
&= P\left\{E\left[\left|\breve{p}_{_{LS}}(\mathbf{Y}) - \delta\right|^2\,\middle|\,\mathcal{D}_n\right] - E\left[\left|p(\mathbf{Y}) - \delta\right|^2\right] > \frac{p_{\min}^4\,\epsilon^2}{16}\right\} \\
&\leq P\left\{2\sup_{p'\in\mathcal{P}}\left|\mathcal{S}_n(p') - E\left[\left|p'(\mathbf{Y}) - \delta\right|^2\right]\right| > \frac{p_{\min}^4\,\epsilon^2}{16}\right\}\,,
\end{aligned}$$

where the second line above follows from the fact that

$$E\left[\left|\breve{p}_{_{LS}}(\mathbf{Y}) - \delta\right|^2\,\middle|\,\mathcal{D}_n\right] = E\left[\left|\breve{p}_{_{LS}}(\mathbf{Y}) - p(\mathbf{Y})\right|^2\,\middle|\,\mathcal{D}_n\right] + E\left[\left|p(\mathbf{Y}) - \delta\right|^2\right]\,,$$

and the last line above follows from

$$
\begin{aligned}
&E\left[\left|\breve{p}_{LS}(\mathbf{Y})-\delta\right|^2\,\middle|\,\mathcal{D}_n\right]-E\left[\left|p(\mathbf{Y})-\delta\right|^2\right]\\
&=E\left[\left|\breve{p}_{LS}(\mathbf{Y})-\delta\right|^2\,\middle|\,\mathcal{D}_n\right]-\inf_{p'\in\mathcal{P}}E\left[\left|p'(\mathbf{Y})-\delta\right|^2\right],\quad\text{where}\;\;p(\mathbf{Y})=E(\delta\,|\,\mathbf{Y})\\
&=\sup_{p'\in\mathcal{P}}\Big\{E\left[\left|\breve{p}_{LS}(\mathbf{Y})-\delta\right|^2\,\middle|\,\mathcal{D}_n\right]-\mathcal{S}_n(\breve{p}_{LS})+\mathcal{S}(\breve{p}_{LS})\\
&\qquad\qquad-\mathcal{S}_n(p')+\mathcal{S}_n(p')-E\left[\left|p'(\mathbf{Y})-\delta\right|^2\right]\Big\}\\
&\le 2\sup_{p'\in\mathcal{P}}\left|\mathcal{S}_n(p')-E\left[\left|p'(\mathbf{Y})-\delta\right|^2\right]\right|,
\end{aligned}
$$

because $\mathcal{S}_n(\breve{p}_{LS})-\mathcal{S}_n(p')\le 0$ by the definition of $\breve{p}_{LS}$. Therefore, from the empirical process theory,

$$
II_n''\le 8\,E\left[\mathcal{N}_1\left(\frac{p_{\min}^4\,\epsilon^2}{256},\,\mathcal{P},\,(\mathbf{Y}_i)_{i=1}^n\right)\right]\,e^{-c_{17}\,n\,\epsilon^4},
$$

where $c_{17}=p_{\min}^8/((16^2)(128))$. □

Proof of Theorem 6

Let $\widehat{L}_n(\Psi)=\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{\delta_i}{\widetilde{p}(\mathbf{Y}_i)}\,I\{\Psi(\mathbf{X}_i)\neq W_i\}$ and observe that

$$
\begin{aligned}
L_n(\widetilde{\Psi}_n)-\inf_{\Psi\in\mathbf{\Psi}}L(\Psi)&=\left[L_n(\widetilde{\Psi}_n)-\widehat{L}_n(\widetilde{\Psi}_n)\right]+\left[\widehat{L}_n(\widetilde{\Psi}_n)-L(\Psi^*)\right],\\
&\qquad(\text{where }\Psi^*=\operatorname*{argmin}_{\Psi\in\mathbf{\Psi}}P\{\Psi(\mathbf{X})\neq W\})\\
&\le 2\sup_{\Psi\in\mathbf{\Psi}}\left|L(\Psi)-\widehat{L}_n(\Psi)\right|,
\end{aligned}
\tag{22}
$$

where the last line follows from the fact that $L_n(\Psi)=L(\Psi)$, for any nonrandom classifier Ψ, and the observation that $\widehat{L}_n(\widetilde{\Psi}_n)\le\widehat{L}(\Psi)$, for every $\Psi\in\mathbf{\Psi}$, (by the definition of $\widetilde{\Psi}_n$ in (11). Therefore, if we let $\mathcal{A}_{\mathbf{\Psi}}$ be the class of all sets of the form

$$
A=\big\{\{\mathbf{x}\,|\,\Psi(\mathbf{x})=1\}\times\{0\}\big\}\bigcup\big\{\{\mathbf{x}\,|\,\Psi(\mathbf{x})=0\}\times\{1\}\big\},\quad\Psi\in\mathbf{\Psi},
$$

then one finds

$$
\begin{aligned}
P\left\{L_n(\widetilde{\Psi}_n) - \inf_{\Psi\in\mathbf{\Psi}} L(\Psi) > \epsilon\right\} &\leq P\left\{\sup_{\Psi\in\mathbf{\Psi}} \left|\widehat{L}_n(\Psi) - L(\Psi)\right| > \frac{\epsilon}{2}\right\} \\
&= P\left\{\sup_{A\in\mathcal{A}_{\mathbf{\Psi}}} \left|\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\widetilde{p}(\mathbf{Y}_i)} I\left\{(\mathbf{X}_i, W_i)\in A\right\} - P\left\{(\mathbf{X}, W)\in A\right\}\right| > \frac{\epsilon}{2}\right\} \\
&\qquad \text{(Since } I\left\{(\mathbf{X}_i, W_i)\in A\right\} = I\left\{\Psi(\mathbf{X}_i)\neq W_i\right\}\text{)} \\
&\leq 8\, S(\mathcal{A}_{\mathbf{\Psi}}, n)\, e^{-c_{19}\, n\, \epsilon^2} + r_n(\epsilon),
\end{aligned}
$$

for n large enough, where $r_n(\epsilon)$ is as in the statement of the theorem (by Theorems 4 and 5).

References

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist Assoc.*, 57:33–45.

Cheng, P. E. and Chu, C. K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statist. Sinica*, 6:63–78. MR1379049

Devroye, L. (1982). Bounds on the uniform deviation of empirical meaures. *Journal of Multivariate Analysis*, 12:72–79. MR0650929

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York. MR1383093

Dudley, R. (1978). Central limit theorems for empirical measures. *Ann. Probab.*, 6:899–929. MR0512411

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. Wiley, New York. MR1925014

Massart, P. (1990). The tight constant in the Devoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18:1269–1283. MR1062069

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York. MR0762984

Talagrand, M. (1994). Sharper bounds for gaussian and empirical processes. *Ann. Probab.*, 22:28–76. MR1258865

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes, with Applications to Statistics*. Springer-Verlag, New York. MR1385671

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280.