Edward H. Kennedy[1]

# Efficient Nonparametric Causal Inference with Missing Exposure Information

[1] Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA 15213-3815, USA, E-mail: edward@stat.cmu.edu

**Abstract:**
Missing exposure information is a very common feature of many observational studies. Here we study identifiability and efficient estimation of causal effects on vector outcomes, in such cases where treatment is unconfounded but partially missing. We consider a missing at random setting where missingness in treatment can depend not only on complex covariates, but also on post-treatment outcomes. We give a new identifying expression for average treatment effects in this setting, along with the efficient influence function for this parameter in a nonparametric model, which yields a nonparametric efficiency bound. We use this latter result to construct nonparametric estimators that are less sensitive to the curse of dimensionality than usual, e. g. by having faster rates of convergence than the complex nuisance estimators they rely on. Further we show that these estimators can be root-n consistent and asymptotically normal under weak nonparametric conditions, even when constructed using flexible machine learning. Finally we apply these results to the problem of causal inference with a partially missing instrumental variable.

**Keywords:** causal inference, missing data, efficiency theory, instrumental variable

**DOI:** 10.1515/ijb-2019-0087

**Received:** August 5, 2019; **Revised:** January 28, 2020; **Accepted:** February 17, 2020

## 1 Introduction

It is very common for there to be missing data in observational studies where causal effects are of interest. In this paper we consider studies where there is substantial missingness in an exposure variable. This is a very common feature of observational studies, and examples abound in the literature. For example, Zhang et al. [1] described the Consortium on Safe Labor observational study, where the goal was to estimate effects of mothers' body mass index on infant birthweight. There, covariate and outcome information was essentially always available, but body mass index data was only available for about half of the mothers. Shortreed and Forbes [2] used the Framingham Heart Study data to assess effects of physical activity on cardiovascular disease and mortality, but up to 30 % of subjects were missing physical activity information. Similarly, Ahn et al. [3] used the Molecular Epidemiology of Colorectal Cancer study to estimate effects of physical activity on colorectal cancer stage, but 20 % of subjects were missing physical activity data. Shardell and Hicks [4] described an analysis of the Baltimore Hip Studies involving older adults with hip fractures, where the goal was to assess effects of perceived mobility recovery on survival. However, this self-reported mobility measure was unavailable for 27 % of subjects. Molinari [5] and Mebane Jr and Poast [6] give numerous other examples of studies with missing exposure information, particularly in survey settings, e. g. from the National Longitudinal Survey of Youth, and the Health and Retirement Study. This is certainly a prevalent problem.

In fact the problem is even more widespread, since in instrumental variable studies one can view the instrument as a type of exposure (e. g. for the purpose of estimating intention-to-treat-style effects, as well as other instrumental variable estimands that require estimating instrument effects on both treatment and outcome). And it is similarly common for instrument values to be missing. For example, in a Mendelian randomization context Burgess et al. [7] used genetic variants as instrumental variables to study the effect of C-reactive protein on fibrinogen and coronary heart disease. However, data on these variants was missing for up to 10 % of subjects, due to difficulty in interpreting output from genotyping platforms. Mogstad and Wiswall [8] and Chaudhuri and Guilkey [9] give further examples of missing instruments from economics.

Although missing exposures and instruments are a prevalent problem, the proposed methods for dealing with this issue have relied on potentially restrictive modeling assumptions, and have been somewhat ad hoc in not considering optimal efficiency. For example Williamson et al. [10] and Zhang et al. [1] propose interesting

**Edward H. Kennedy** is the corresponding author.

semiparametric estimators, but they rely on parametric models for nuisance functions, and do not consider the question of efficiency in either nonparametric or semiparametric models. Zhang et al. [1] also only considers binary outcomes. Chaudhuri and Guilkey [9] discusses (semiparametric) efficiency theory, but only for a finite-dimensional parameter in a population moment condition depending on a known function. This means their results apply to classical linear models, but not to the fully nonparametric setting pursued here. Kennedy and Small [11] consider nonparametric efficiency theory in missing instrumental variable problems, but only in simpler settings with one-sided noncompliance and no covariates.

Thus we fill these gaps by giving a new identifying expression for average treatment effects of multivalued discrete exposures in the presence of complex confounding and missing exposure values, deriving the efficient influence function and corresponding nonparametric efficiency bounds, and constructing nonparametric estimators that can be $\sqrt{n}$-consistent and asymptotically normal, even if nuisance functions are estimated at slower rates via nonparametric machine learning tools. Finally we apply these general results to also address the problem of causal inference with a partially missing instrumental variable. Throughout we make use of a missing at random assumption used by previous authors, allowing exposure missingness to depend on post-exposure outcome information.

## 2 Missing exposures

In this section we consider the general problem of identification and efficient estimation of average treatment effects, when exposure values are missing at random, allowing the missingness mechanism to depend on both covariates and post-treatment outcome information.

### 2.1 Setup

Suppose we observe an iid sample $(\mathbf{O}_1, \ldots, \mathbf{O}_n) \sim \mathbb{P}$ where

$$\mathbf{O} = (\mathbf{X}, R, RZ, \mathbf{Y})$$

for $\mathbf{X} \in \mathbb{R}^d$ denoting covariate information, $Z \in \{z_1, \ldots, z_k\}$ a discrete treatment or exposure, $R \in \{0, 1\}$ an indicator for whether $Z$ is observed or not, and $\mathbf{Y} = (Y_1, \ldots, Y_p)^{\mathsf{T}} \in \mathbb{R}^p$ a vector of $p$ outcomes of interest. In general we use script characters to denote the support of a random variable, e. g. $\mathbf{X} \in \boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^d$. For notational simplicity we further define the nuisance functions

$$\begin{aligned}
\mu(\mathbf{y} \mid \mathbf{x}) &= \mathbb{P}(\mathbf{Y} \leq \mathbf{y} \mid \mathbf{X} = \mathbf{x}) \\
\pi(\mathbf{x}, \mathbf{y}) &= \mathbb{P}(R = 1 \mid \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) \\
\lambda_z(\mathbf{x}, \mathbf{y}) &= \mathbb{P}(Z = z \mid \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, R = 1).
\end{aligned}$$

Note that $\mu$ is the cumulative distribution function of the outcome given covariates, $\pi$ can be viewed as the missingness propensity score, and $\lambda$ the regression on covariates and outcomes of treatment among those for whom it is measured. We further define

$$\begin{aligned}
\boldsymbol{\beta}_z(\mathbf{x}) &= \mathbb{E}\{\mathbf{Y}\lambda_z(\mathbf{X}, \mathbf{Y}) \mid \mathbf{X} = \mathbf{x}\} = \int_{\mathcal{Y}} \mathbf{y}\lambda_z(\mathbf{x}, \mathbf{y})\, d\mu(\mathbf{y} \mid \mathbf{x}) \\
\gamma_z(\mathbf{x}) &= \mathbb{E}\{\lambda_z(\mathbf{X}, \mathbf{Y}) \mid \mathbf{X} = \mathbf{x}\} = \int_{\mathcal{Y}} \lambda_z(\mathbf{x}, \mathbf{y})\ d\mu(\mathbf{y} \mid \mathbf{x}).
\end{aligned}$$

The quantity $\boldsymbol{\beta}_z(\mathbf{x}) = \{\beta_z^1(\mathbf{x}), \ldots, \beta_z^p(\mathbf{x})\}^{\mathsf{T}}$ is a vector of the same dimension as $\mathbf{Y}$. We will see shortly that, under missing at random assumptions, $\gamma_z$ equals the propensity score, while $\boldsymbol{\beta}_z$ equals the product of the propensity score and outcome regression.

Our goal is to estimate the mean $\boldsymbol{\psi}_z = \mathbb{E}(\mathbf{Y}^z) = \{\mathbb{E}(Y_1^z), \ldots, \mathbb{E}(Y_p^z)\}^{\mathsf{T}} \in \mathbb{R}^p$ of the outcomes that would have been observed under treatment level $z \in \mathcal{Z}$. It is well-known that this equals

$$\boldsymbol{\psi}_z = \int_{\mathcal{X}} \mathbb{E}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, Z = z)\, d\mathbb{P}(x) \tag{1}$$

under the following standard causal assumptions:

$$(\text{Consistency.})\ \mathbf{Y} = \mathbf{Y}^Z \tag{2}$$

$$(Z - \text{Positivity.})\ \mathbb{P}\{\epsilon < \mathbb{P}(Z = z \mid \mathbf{X}) < 1 - \epsilon\} = 1 \text{ for all } z \in \mathcal{Z} \tag{3}$$

$$(Z-\text{Exchangeability.})\ Z \perp\!\!\!\perp \mathbf{Y}^z \mid \mathbf{X} \ \text{ for all } \ z \in \mathscr{Z} \tag{4}$$

These assumptions have been discussed extensively in the literature [12, 13], so we refer the reader elsewhere for more details.

Crucially, when treatment $Z$ is missing for some subjects, expression (1) is still not identified even under (2)–(4), since $Z$ is not observed unless $R = 1$. We consider identification under missing at random conditions used for example by Chaudhuri and Guilkey [10], Williamson et al. [9] and Zhang et al. [1], which are:

$$(R-\text{Exchangeability.}) \quad R \perp\!\!\!\perp Z \mid \mathbf{X}, \mathbf{Y} \tag{5}$$

$$(R-\text{Positivity.})\ \mathbb{P}\{\epsilon < \mathbb{P}(R=1 \mid \mathbf{X}, \mathbf{Y}) < 1-\epsilon\} = 1 \tag{6}$$

Note that the missing at random condition (5) allows missingness in treatment $Z$ to depend on the post-treatment outcome $\mathbf{Y}$; this will be important if the outcome captures some information about the missingness mechanism beyond the covariates. An alternative missing-at-random assumption would be $R \perp\!\!\!\perp (Z, \mathbf{Y}) \mid \mathbf{X}$; note however that this implies our $R$-exchangeability assumption, as well as the further testable implication that $R \perp\!\!\!\perp Y \mid \mathbf{X}$. Therefore our assumption is strictly weaker.

Figure 1 uses directed acyclic graphs to illustrate two different data generating processes that satisfy exchangeability conditions (4) and (5). The first represents a process where missingness occurs prior to the outcome (e. g. subjects miss a visit when they would have contributed treatment information); the second represents a process where missingness occurs after the outcome (e. g. survey non-response or data corruption after measurement).

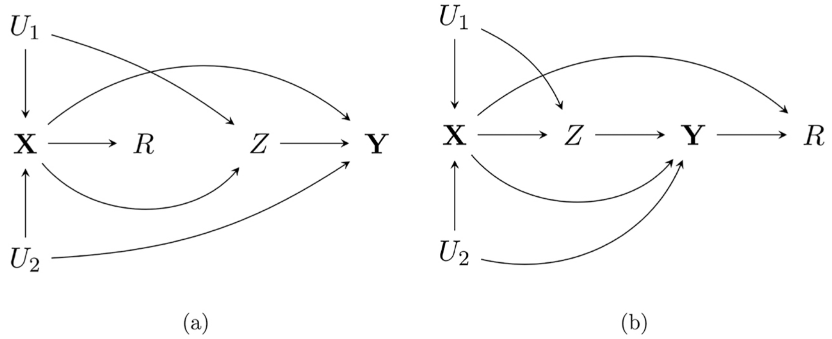

**Figure 1:** Two directed acyclic graphs for which the required exchangeability assumptions (4) and (5) hold, where $(U_1, U_2)$ are unmeasured and $Z$ is only observed when $R = 1$. In graph (a) missingness can be viewed as occurring prior to the outcome, while in (b) it can be viewed as occurring after. The variable $U_2$ can be represented as the potential outcome $\mathbf{Y}^0$.

## 2.2 Identification & efficiency theory

Our first result gives a new identifying expression for $\boldsymbol{\psi}_z$ under the causal and missing at random assumptions above. This essentially follows from the important facts that, under (5)–(6), the propensity score is given by

$$\gamma_z(\mathbf{x}) = \mathbb{P}(Z = z \mid \mathbf{X} = \mathbf{x})$$

(note this means $Z$-positivity (3) is equivalent to $\gamma_z$ being bounded away from zero and one), and that the outcome regression satisfies

$$\boldsymbol{\beta}_z(\mathbf{x}) = \gamma_z(\mathbf{x})\mathbb{E}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, Z = z).$$

**Proposition 1.**
*Under the causal assumptions (2)–(4) and the missing at random assumptions (5)–(6), it follows that*

$$\boldsymbol{\psi}_z = \mathbb{E}(\mathbf{Y}^z) = \mathbb{E}\left\{\frac{\boldsymbol{\beta}_z(\mathbf{X})}{\gamma_z(\mathbf{X})}\right\}.$$

**Proof.**

We have

$$\begin{aligned}\mathbb{E}(\mathbf{Y}^z) &= \int_{\mathcal{X}} \mathbb{E}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, Z = z)\, d\mathbb{P}(\mathbf{x}) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{\mathbf{y}\mathbb{P}(Z = z \mid \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})}{\mathbb{P}(Z = z \mid \mathbf{X} = \mathbf{x})}\, d\mu(\mathbf{y} \mid \mathbf{x})\, d\mathbb{P}(\mathbf{x}) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{\mathbf{y}\, \lambda_z(\mathbf{x}, \mathbf{y})}{\int_{\mathcal{Y}} \lambda_z(\mathbf{x}, \mathbf{y}) d\mu(\mathbf{y} \mid \mathbf{x})}\, d\mu(\mathbf{y} \mid \mathbf{x})\, d\mathbb{P}(\mathbf{x}) = \mathbb{E}\left\{\frac{\boldsymbol{\beta}_z(\mathbf{X})}{\gamma_z(\mathbf{X})}\right\}\end{aligned}$$

where the first equality follows by the causal assumptions (2)–(4), the second by Bayes' rule, the third by the missing at random assumptions (5)–(6) and iterated expectation, and the fourth by definition. □

Interestingly, although the complete-data functional (4) does not depend on the observational treatment process, its identified version under the missing at random assumptions does. Intuitively, this occurs because when $Z$ is missing, one cannot simply condition on $(\mathbf{X}, Z)$ anymore, and instead the distribution of $Z$ given $\boldsymbol{X}$ needs to be constructed by marginalizing over that of $Z$ given $(\mathbf{X}, \mathbf{Y})$ among those with $Z$ observed.

The next result gives a crucial von Mises-type expansion for the parameter from Proposition 1, which lays the foundation for the efficiency bound and estimation results to come. This result can be viewed as giving a distributional Taylor expansion for the functional $\boldsymbol{\psi}_z$.

**Lemma 1.**
*The functional $\boldsymbol{\psi}_z = \boldsymbol{\psi}_z(\mathbb{P})$ from Proposition 1 admits the expansion*

$$\boldsymbol{\psi}_z(\mathbb{P}) - \boldsymbol{\psi}_z(\mathbb{P}) = \int_O \{\phi_z(\mathbf{O}; \overline{\mathbb{P}}) - \boldsymbol{\psi}_z(\overline{\mathbb{P}})\}(d\overline{\mathbb{P}} - d\mathbb{P}) + R_2(\overline{\mathbb{P}}, \mathbb{P})$$

*where*

$$\phi_z(\mathbf{O}; \mathbb{P}) = \left\{\frac{\mathbf{Y} - \boldsymbol{\beta}_z(\mathbf{X})/\gamma_z(\mathbf{X})}{\gamma_z(\mathbf{X})}\right\}\left[\frac{R\{1(Z = z) - \lambda_z(\mathbf{X}, \mathbf{Y})\}}{\pi(\mathbf{X}, \mathbf{Y})} + \lambda_z(\mathbf{X}, \mathbf{Y})\right] + \frac{\boldsymbol{\beta}_z(\mathbf{X})}{\gamma_z(\mathbf{X})}$$

*and*

$$\begin{aligned}R_2(\overline{\mathbb{P}}, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}\Bigg( &\left\{\frac{\mathbf{Y} - \overline{\boldsymbol{\beta}}_z(\mathbf{X})/\overline{\gamma}_z(\mathbf{X})}{\overline{\gamma}_z(\mathbf{X})}\right\}\left\{\frac{\pi(\mathbf{X}, \mathbf{Y}) - \overline{\pi}(\mathbf{X}, \mathbf{Y})}{\overline{\pi}(\mathbf{X}, \mathbf{Y})}\right\}\left\{\lambda_z(\mathbf{X}, \mathbf{Y}) - \overline{\lambda}_z(\mathbf{X}, \mathbf{Y})\right\} \\ &+ \left[\left\{\frac{\boldsymbol{\beta}_z(\mathbf{X}) - \overline{\boldsymbol{\beta}}_z(\mathbf{X})}{\gamma_z(\mathbf{X})}\right\} + \frac{\overline{\boldsymbol{\beta}}_z(\mathbf{X})}{\overline{\gamma}_z(\mathbf{X})}\left\{\frac{\overline{\gamma}_z(\mathbf{X}) - \gamma_z(\mathbf{X})}{\gamma_z(\mathbf{X})}\right\}\right]\left\{\frac{\gamma_z(\mathbf{X}) - \overline{\gamma}_z(\mathbf{X})}{\overline{\gamma}_z(\mathbf{X})}\right\}\Bigg).\end{aligned}$$

**Proof.**
Here we drop the $z$ argument throughout to ease notation. Note we can write

$$\begin{aligned}\phi(\mathbf{O}; \mathbb{P}) &= \left[\frac{R\{1(Z = z) - \lambda(\mathbf{X}, \mathbf{Y})\}}{\pi(\mathbf{X}, \mathbf{Y})} + \lambda(\mathbf{X}, \mathbf{Y})\right]\left\{\frac{\mathbf{Y} - \boldsymbol{\beta}(\mathbf{X})/\gamma(\mathbf{X})}{\gamma(\mathbf{X})}\right\} + \frac{\boldsymbol{\beta}(\mathbf{X})}{\gamma(\mathbf{X})} \\ &= \frac{1}{\gamma(\mathbf{X})}\left[\frac{R\mathbf{Y}}{\pi(\mathbf{X}, \mathbf{Y})}\{1(Z = z) - \lambda(\mathbf{X}, \mathbf{Y})\} + \{\mathbf{Y}\lambda(\mathbf{X}, \mathbf{Y}) - \boldsymbol{\beta}(\mathbf{X})\}\right] \\ &\quad - \frac{\boldsymbol{\beta}(\mathbf{X})}{\gamma(\mathbf{X})^2}\left[\frac{R\{1(Z = z) - \lambda(\mathbf{X}, \mathbf{Y})\}}{\pi(\mathbf{X}, \mathbf{Y})} + \{\lambda(\mathbf{X}, \mathbf{Y}) - \gamma(\mathbf{X})\}\right] + \frac{\boldsymbol{\beta}(\mathbf{X})}{\gamma(\mathbf{X})}.\end{aligned}$$

Therefore, letting $\mathbb{E} = \mathbb{E}_{\mathbb{P}}$ and dropping $(\mathbf{X},\ \mathbf{Y})$ arguments from $(\pi, \lambda, \boldsymbol{\beta}, \gamma)$ to further ease notation, we have

$$\begin{aligned}&\psi(\bar{\mathbb{P}}) - \psi(\mathbb{P}) + \int_O \{\phi(\mathbf{O}; \bar{\mathbb{P}}) - \psi(\bar{\mathbb{P}})\} d\mathbb{P} \\ &= \mathbb{E}\left[\frac{1}{\bar{\gamma}}\left\{\frac{R\mathbf{Y}}{\bar{\pi}}(1_z - \bar{\lambda}) + (\mathbf{Y}\bar{\lambda} - \bar{\boldsymbol{\beta}})\right\} - \frac{\bar{\boldsymbol{\beta}}}{\bar{\gamma}^2}\left\{\frac{R(1_z - \bar{\lambda})}{\bar{\pi}} + (\bar{\lambda} - \bar{\gamma})\right\} + \left(\frac{\bar{\boldsymbol{\beta}}}{\bar{\gamma}} - \frac{\boldsymbol{\beta}}{\gamma}\right)\right] \\ &= \mathbb{E}\left[\left(\frac{\mathbf{Y} - \bar{\boldsymbol{\beta}}/\bar{\gamma}}{\bar{\gamma}}\right)\left(\frac{\pi - \bar{\pi}}{\bar{\pi}}\right)(\lambda - \bar{\lambda}) + \left(\frac{\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}}{\bar{\gamma}}\right) - \frac{\bar{\boldsymbol{\beta}}}{\bar{\gamma}}\left(\frac{\gamma - \bar{\gamma}}{\bar{\gamma}}\right) + \left(\frac{\bar{\boldsymbol{\beta}}}{\bar{\gamma}} - \frac{\boldsymbol{\beta}}{\gamma}\right)\right] \\ &= \mathbb{E}\left[\left(\frac{\mathbf{Y} - \bar{\boldsymbol{\beta}}/\bar{\gamma}}{\bar{\gamma}}\right)\left(\frac{\pi - \bar{\pi}}{\bar{\pi}}\right)(\lambda - \bar{\lambda}) + \left(\frac{\boldsymbol{\beta}}{\gamma} - \frac{\bar{\boldsymbol{\beta}}}{\bar{\gamma}}\right)\left(\frac{\gamma - \bar{\gamma}}{\bar{\gamma}}\right)\right] = R_2(\bar{\mathbb{P}}, \mathbb{P})\end{aligned}$$

where the second equality follows from from rearranging and iterated expectation (together with $\mathbb{E}(\mathbf{Y}\lambda \mid \mathbf{X}) = \boldsymbol{\beta}$ and $\mathbb{E}(\lambda \mid \mathbf{X}) = \gamma$), and the third since

$$\frac{\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}}{\overline{\gamma}} - \frac{\overline{\boldsymbol{\beta}}}{\overline{\gamma}}\left(\frac{\gamma-\overline{\gamma}}{\overline{\gamma}}\right) + \frac{\overline{\boldsymbol{\beta}}}{\overline{\gamma}} - \frac{\boldsymbol{\beta}}{\gamma} = \boldsymbol{\beta}\left(\frac{1}{\overline{\gamma}} - \frac{1}{\gamma}\right) - \frac{\overline{\boldsymbol{\beta}}}{\overline{\gamma}}\left(\frac{\gamma-\overline{\gamma}}{\overline{\gamma}}\right) = \left(\frac{\boldsymbol{\beta}}{\gamma} - \frac{\overline{\boldsymbol{\beta}}}{\overline{\gamma}}\right)\left(\frac{\gamma-\overline{\gamma}}{\overline{\gamma}}\right).$$

Note the above implies $\mathbb{E}\{\phi(\mathbf{O};\mathbb{P}) - \boldsymbol{\psi}_z(\mathbb{P})\} = 0$, which is also straightforward to see using iterated expectation. □

Lemma 1 has several important consequences. First, it suggests how one could correct the first-order bias of a plug-in estimator $\boldsymbol{\psi}_z(\widehat{\mathbb{P}})$, by estimating the first term in the expansion and subtracting it off. This is one way to view what semiparametric estimators (particularly of the "one-step" variety) based on influence functions are doing, and in fact the estimator presented in the next subsection does precisely this. Second, it essentially immediately yields the efficient influence function for $\boldsymbol{\psi}_z$. The next theorem states this result; after it we describe why the efficient influence function is useful here.

**Theorem 1.**
*Under a nonparametric model satisfying positivity conditions (3) and (6), the efficient influence function for $\boldsymbol{\psi}_z$ is given by $\boldsymbol{\phi}_z(\mathbf{O};\mathbb{P}) - \boldsymbol{\psi}_z$ as defined in Lemma 1.*

**Proof.**
Recall from Bickel et al. [14] and van der Vaart [15] that the nonparametric efficiency bound for a functional $\psi$ is given by the supremum of Cramer-Rao lower bounds for that functional across smooth parametric submodels. The efficient influence function is the mean-zero function whose variance equals the efficiency bound, and is given by the unique $\varphi$ that is a valid submodel score (or limit of such scores) satisfying pathwise differentiability, i. e.

$$\frac{d}{d\epsilon}\psi(\mathbb{P}_\epsilon)|_{\epsilon=0} = \int_{\mathcal{O}} \varphi(\mathrm{O};\mathbb{P})\left(\frac{d}{d\epsilon}\log d\mathbb{P}_\epsilon\right)|_{\epsilon=0}d\mathbb{P} \tag{7}$$

for $\mathbb{P}_\epsilon$ any smooth parametric submodel.

To see that $\boldsymbol{\phi} - \boldsymbol{\psi}$ is the efficient influence function for $\boldsymbol{\psi}$, first note that the expansion in Lemma 1 implies

$$\boldsymbol{\psi}_z(\mathbb{P}_\epsilon) - \boldsymbol{\psi}_z(\mathbb{P}) = \int_{\mathcal{O}} \big\{\boldsymbol{\phi}_z(\mathbf{O};\mathbb{P}) - \boldsymbol{\psi}_z(\mathbb{P})\big\}d\mathbb{P}_\epsilon - \mathbf{R}_\mathbf{z}(\mathbb{P},\mathbb{P}_\epsilon)$$

so differentiating with respect to $\varepsilon$ yields

$$\begin{aligned}\frac{d}{d\epsilon}\boldsymbol{\psi}_z(\mathbb{P}_\epsilon) &= \int_{\mathcal{O}} \big\{\boldsymbol{\phi}_z(\mathbf{O};\mathbb{P}) - \boldsymbol{\psi}_z(\mathbb{P})\big\}\frac{d}{d\epsilon}d\mathbb{P}_\epsilon - \frac{d}{d\epsilon}\mathbf{R}_\mathbf{z}(\mathbb{P},\mathbb{P}_\epsilon)\\ &= \int_{\mathcal{O}} \big\{\boldsymbol{\phi}_z(\mathbf{O};\mathbb{P}) - \boldsymbol{\psi}_z(\mathbb{P})\big\}\left(\frac{d}{d\epsilon}\log d\mathbb{P}_\epsilon\right)d\mathbb{P}_\epsilon - \frac{d}{d\epsilon}\mathbf{R}_\mathbf{z}(\mathbb{P},\mathbb{P}_\epsilon).\end{aligned}$$

The property (7) follows after evaluating at $\varepsilon = 0$, since

$$\frac{d}{d\epsilon}\mathbf{R}_\mathbf{z}(\mathbb{P},\mathbb{P}_\epsilon)\Big|_{\epsilon=0} = 0$$

by virtue of the fact that $\mathbf{R}_\mathbf{z}(\mathbb{P},\mathbb{P}_\epsilon)$ consists of only second-order products of errors between $\mathbb{P}_\epsilon$ and $\mathbb{P}$. Thus applying the product rule yields a sum of two terms, each of which is a product of a derivative term (which may not be zero at $\varepsilon = 0$) and an error term involving differences of components of $\mathbb{P}_\epsilon$ and $\mathbb{P}$ (which will be zero at $\varepsilon = 0$). Since our model is nonparametric, the tangent space is the entire Hilbert space of mean-zero finite-variance functions; hence there is only one influence function satisfying (7) and it is the efficient one [14–16]. Therefore $\boldsymbol{\phi} - \boldsymbol{\psi}$ must be the efficient influence function.

An equivalent way to derive this result, as suggested by an anonymous reviewer in a previous version of this manuscript, is to use results from Robins et al. [17]. Specifically, as in Theorem 7.2 of Tsiatis [16], one can take full-data efficient influence function for $\boldsymbol{\psi}_z$, inverse-probability-weight it for those with $R = 1$ and subtract off its projection onto the tangent space. This yields the same efficient influence function. □

The efficient influence function is important since its variance $\text{cov}\{\boldsymbol{\phi}_z(\mathbf{O};\mathbb{P}) - \boldsymbol{\psi}_z\}$ gives an efficiency bound for estimation of $\boldsymbol{\psi}_z$, providing a benchmark for efficient estimation. More precisely, following Bickel et al. [14], van der Vaart [15], and Tsiatis [16], this variance provides a local asymptotic minimax lower bound in the sense of Hajek and Le Cam, and tells us that the asymptotic variance of any regular asymptotically linear estimator can be no smaller (in that the difference in covariance matrices must be non-negative definite). Insofar as the bias-correction suggested earlier directly involves the efficient influence function, this object is also crucial for constructing estimators that have second-order bias and so can be $\sqrt{n}$-consistent and asymptotically normal even when the nuisance functions are estimated flexibly at slower rates of convergence. This feature will be detailed in the next subsection.

## 2.3 Estimation & inference

Here we present an estimator based on the functional expansion from Lemma 1, which is asymptotically efficient under weak nonparametric conditions.

To ease notation let $\boldsymbol{\phi}_z = \boldsymbol{\phi}_z(\mathbf{O}; \mathbb{P})$ and $\widehat{\boldsymbol{\phi}}_z = \boldsymbol{\phi}_z(\mathbf{O}; \widehat{\mathbb{P}})$ denote the true and estimated versions of the uncentered efficient influence function for $\boldsymbol{\psi}_z$. The estimator we study here is given by

$$\widehat{\boldsymbol{\psi}}_z = \mathbb{P}_n(\widehat{\boldsymbol{\phi}}_z)$$

where we use $\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{O}_i)$ to denote sample averages. Therefore the estimator $\widehat{\boldsymbol{\psi}}_z$ is simply the sample average of the estimated (uncentered) influence function values; equivalently we can write it as a bias-corrected version of the plug-in $\boldsymbol{\psi}_z(\widehat{\mathbb{P}})$, namely

$$\widehat{\boldsymbol{\psi}}_z = \boldsymbol{\psi}_z(\widehat{\mathbb{P}}) + \mathbb{P}_n(\widehat{\boldsymbol{\varphi}}_z)$$

where $\widehat{\boldsymbol{\varphi}}_z = \widehat{\boldsymbol{\phi}}_z - \boldsymbol{\psi}_z(\widehat{\mathbb{P}})$ is the estimated efficient influence function.

For simplicity, in the following results we assume the nuisance estimates $\widehat{\mathbb{P}}$ are constructed from a separate independent sample. In practice, one can split the sample, use part for fitting $\widehat{\mathbb{P}}$ and the other for constructing $\widehat{\boldsymbol{\phi}}_z$, and then swap so as to attain full efficiency based on the entire sample size $n$ rather than a fraction, e. g. $n/2$. This is the idea behind the sample-splitting methods used in other functional estimation problems [18–20]. Alternatively, if the same observations are used both for estimating $\widehat{\mathbb{P}}$ and constructing $\widehat{\boldsymbol{\phi}}_z$, one generally needs to rely on empirical process conditions to obtain the kinds of results we present here.

The next theorem gives the asymptotic properties of the estimator $\widehat{\boldsymbol{\psi}}_z$, and conditions under which it is $\sqrt{n}$-consistent and converging to a normal distribution with asymptotic variance equal to the nonparametric efficiency bound. In what follows, we let $\|f\|^2 = \mathbb{P}(f^2) = \int_O f(\mathbf{o})^2 \, d\mathbb{P}(\mathbf{o})$ denote the squared $L_2(\mathbb{P})$ norm.

**Theorem 2.**

*Assume*$\|\widehat{\boldsymbol{\phi}}_z - \boldsymbol{\phi}_z\| = o_{\mathbb{P}}(1)$*and*$\mathbb{P}(\epsilon < \widehat{\pi} < 1-\epsilon) = \mathbb{P}(\epsilon < \widehat{\gamma}_z < 1-\epsilon) = 1$. *Then*

$$\widehat{\boldsymbol{\psi}}_z - \boldsymbol{\psi}_z = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + \|\widehat{\pi} - \pi\| \|\hat{\lambda}_z - \lambda_z\| + \left(\|\hat{\boldsymbol{\beta}}_z - \boldsymbol{\beta}_z\| + \|\widehat{\gamma}_z - \gamma_z\|\right)\|\widehat{\gamma}_z - \gamma_z\|\right),$$

*and if* $\|\widehat{\pi} - \pi\| \|\hat{\lambda}_z - \lambda_z\| + \left(\|\hat{\boldsymbol{\beta}}_z - \boldsymbol{\beta}_z\| + \|\widehat{\gamma}_z - \gamma_z\|\right)\|\widehat{\gamma}_z - \gamma_z\| = o_{\mathbb{P}}(1/\sqrt{n})$, *we have*

$$\sqrt{n}(\widehat{\boldsymbol{\psi}}_z - \boldsymbol{\psi}_z) \rightsquigarrow N\Big(\mathbf{0}, cov(\boldsymbol{\phi}_z)\Big).$$

**Proof.**

Dropping $z$ subscripts to ease notation, we can write

$$\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi} = (\mathbb{P}_n - \mathbb{P})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) + (\mathbb{P}_n - \mathbb{P})\boldsymbol{\phi} + \mathbb{P}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}). \tag{8}$$

For the first term in (8) above, Lemma 2 in the Appendix (reproduced from Kennedy et al. [21]) implies that

$$(\mathbb{P}_n - \mathbb{P})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) = O_{\mathbb{P}}\left(\frac{\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|}{\sqrt{n}}\right) = o_{\mathbb{P}}(1/\sqrt{n})$$

where the last equality follows since $\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\| = o_{\mathbb{P}}(1)$ by assumption. The expansion from Lemma 1 now implies

$$\begin{aligned}\mathbb{P}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) &= \int_O \{\phi(\mathbf{O}; \widehat{\mathbb{P}}) - \boldsymbol{\psi}(\mathbb{P})\} d\mathbb{P} = R_2(\widehat{\mathbb{P}}, \mathbb{P}) \\ &\lesssim \|\hat{\pi} - \pi\| \left\|\hat{\lambda} - \lambda\right\| + \left(\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| + \|\hat{\gamma} - \gamma\|\right)\|\hat{\gamma} - \gamma\|\end{aligned}$$

where the last line uses Cauchy-Schwarz and the fact that $(\widehat{\gamma}, \widehat{\pi}, \gamma)$ are all bounded away from zero. This yields the result. □

Importantly, Theorem 2 shows that $\widehat{\boldsymbol{\psi}}_z$ attains faster rates than its nuisance estimators, and can be asymptotically efficient under weak nonparametric conditions. Specifically, as long as the influence function is consistently estimated in $L_2$ norm, the estimator $\widehat{\boldsymbol{\psi}}_z$ has a rate of convergence that is second-order in the nuisance estimation error, thus attaining faster rates than the nuisance estimators. Under standard $n^{-1/4}$-type rate conditions, the estimator is $\sqrt{n}$-consistent, asymptotically normal, and efficient. Importantly, these rates can plausibly be attained under nonparametric smoothness, sparsity, or other structural conditions (e. g. additive modeling or bounded variation assumptions, etc.). For example, if it is assumed that all $d$-dimensional nuisance functions lie in a Hölder class with smoothness index $s$ (i. e. partial derivatives up to order $s$ exist and are Lipschitz) then the assumption of Theorem 2 would be satisfied when $s > d/2$, i. e. the smoothness index is at least half the dimension. Alternatively, if the functions are $s$-sparse then one would need $s = o(\sqrt{n})$ up to log factors, as in Farrell [22]. Then asymptotically valid 95 % confidence intervals can be constructed via a simple Wald form, $\widehat{\boldsymbol{\psi}}_z \pm 1.96\sqrt{\text{diag}\{\text{cov}(\boldsymbol{\phi}_z)\}/n}$ The next result points out the double robustness of $\widehat{\boldsymbol{\psi}}_z$.

**Corollary 1.**
*Under the conditions of Theorem 2, the estimator $\widehat{\boldsymbol{\psi}}_z$ is consistent if either*

1. $\|\widehat{\gamma}_z - \gamma_z\| = o_{\mathbb{P}}(1) and \|\widehat{\pi} - \pi\| = o_{\mathbb{P}}(1)$, *or*
2. $\|\widehat{\gamma}_z - \gamma_z\| = o_{\mathbb{P}}(1) and \|\widehat{\lambda}_z - \lambda_z\| = o_{\mathbb{P}}(1)$.

Corollary 1 shows that $\widehat{\boldsymbol{\psi}}_z$ is doubly robust [23, 24], since it is consistent if either $\widehat{\pi}$ or $\widehat{\lambda}_z$ are (and $\widehat{\gamma}_z$ is). Note however that our formulation requires the propensity score $\gamma_z$ to be estimated consistently. This contrasts with the semiparametric approach of Zhang et al. [1], who construct an estimator that is consistent as long as two of three nuisance functions are estimated consistently. However, Zhang et al. [1] work under a different factorization of the likelihood, and impose parametric models on the partially observed propensity score and outcome regression functions. It is unclear whether our remainder can be written in a triply robust form, though we conjecture that results of Zhang et al. [1] would not hold in the fully nonparametric setting considered here. This and a more general study of triple robustness could be an important avenue for future work.

## 3 Application to missing instruments

Here we apply the theory from the previous section to identify and efficient estimate the local average treatment effect in instrumental variable studies with missing instrument values.

It is quite common for some instrument values to be missing in instrumental variable studies [8, 9, 11]. This setup fits in the proposed framework from the previous section as follows. We have $O = (\mathbf{X}, R, RZ, \mathbf{Y})$ where $Z \in \{0, 1\}$ is now an instrument, and $\mathbf{Y} = (A, Y)$ for $A \in \{0, 1\}$ a binary treatment and $Y \in \mathbb{R}$ an outcome of interest. Here the outcome $Y \in \mathbb{R}$ is a scalar, but a multivariate outcome presents no additional complications. Note also our slight abuse of notation in using bold $\mathbf{Y} = (A, Y)$ to denote a vector that contains the scalar outcome $Y$. Then we can write

$$\boldsymbol{\beta}_z(\mathrm{x}) = \{\beta_z^a(\mathbf{x}), \beta_z^y(\mathbf{x})\}^{\mathrm{T}}$$

where $\beta_z^t(\mathbf{x}) = \mathbb{E}\{T\lambda_z(\mathbf{X}, A, Y) \mid \mathbf{X} = \mathbf{x}\}$.

In addition to the causal assumptions (2)–(4) and missing at random assumptions (5)–(6) from before, we further make the instrumental variable assumptions:

$$\text{(Exclusion.)}\ \ Y^{za} = Y^a \tag{9}$$

$$\text{(Relevance.)}\ \ \mathbb{P}(A^{z=1} > A^{z=0}) \geq \epsilon > 0 \tag{10}$$

$$\text{(Monotonicity.)}\ \ \mathbb{P}(A^{z=1} \geq A^{z=0}) = 1 \tag{11}$$

Our first result identifies the local average treatment effect under the assumptions above.

**Proposition 2.**
*Under the causal assumptions (2)–(4), the missing at random assumptions (5)–(6), and the instrumental variable assumptions (9)–(11), it follows that*

$$\theta = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{z=1} > A^{z=0}) = \frac{\mathbb{E}\{\beta_1^y(\mathbf{X})/\gamma_1(\mathbf{X}) - \beta_0^y(\mathbf{X})/\gamma_0(\mathbf{X})\}}{\mathbb{E}\{\beta_1^a(\mathbf{X})/\gamma_1(\mathbf{X}) - \beta_0^a(\mathbf{X})/\gamma_0(\mathbf{X})\}}.$$

**Proof.**

It is well known [25, 26] that assumptions (9)–(11) imply

$$\theta = \frac{\mathbb{E}(Y^{z=1} - Y^{z=0})}{\mathbb{E}(A^{z=1} - A^{z=0})}$$

so the result follows from Proposition 1, after taking $\mathbf{Y} = (A, Y)$ and $\mathscr{Z} = \{0, 1\}$. □

Although we focus on the local average treatment effect, the same observed data functional can represent other treatment effects under varying assumptions (e. g. the effect on the would-be-treated under a no-effect-modification assumption as discussed for example by Hernán and Robins [27]). Thus our results equally apply to these other settings.

Now we go on to use the theory from the previous section to construct an efficient estimator of the local average treatment effect $\theta$. As with $\boldsymbol{\beta}_z(\mathbf{x})$, we can decompose the efficient influence function $\boldsymbol{\phi}_z$ from the previous section as

$$\boldsymbol{\phi}_z(\mathbf{O}) = \{\phi_z^a(\mathbf{O}), \phi_z^y(\mathbf{O})\}^{\mathrm{T}}$$

for the two outcomes $(A, Y) \in \mathbf{Y}$. As before we write $\boldsymbol{\phi}_z = \boldsymbol{\phi}_z(\mathbf{O}; \mathbb{P})$ and $\widehat{\boldsymbol{\phi}}_z = \boldsymbol{\phi}_z(\mathbf{O}; \widehat{\mathbb{P}})$ to ease notation, and suppose $\widehat{\mathbb{P}}$ is constructed from an independent sample. The proposed estimator is given by

$$\hat{\theta} = \frac{\mathbb{P}_n(\widehat{\phi}_1^y - \widehat{\phi}_0^y)}{\mathbb{P}_n(\widehat{\phi}_1^a - \widehat{\phi}_0^a)}.$$

This simply takes the ratio of the corresponding estimators for the effects of $Z$ on $A$ and $Y$, respectively.

The next result describes the asymptotic properties of the estimator $\hat{\theta}$, and gives conditions under which it is $\sqrt{n}$-consistent and asymptotically normal, akin to the earlier Theorem 2 for a general $\widehat{\boldsymbol{\psi}}_z$.

**Theorem 3.**

*Assume*$\|\widehat{\phi}_z^t - \phi_z^t\| = o_{\mathbb{P}}(1)$*for* $z \in \{0, 1\}$ *and*$t \in \{a, y\}$, *and*$\mathbb{P}(\epsilon < \widehat{\pi} < 1-\epsilon) = \mathbb{P}(\epsilon < \widehat{\gamma}_z < 1-\epsilon) = \mathbb{P}\{\mathbb{P}_n(\widehat{\phi}_1^a - \widehat{\phi}_0^a) > \epsilon\} = 1$. *Define*

$$S_{n,z} = \|\widehat{\pi} - \pi\|\|\hat{\lambda}_z - \lambda_z\| + \left(\max_{t \in \{a,y\}} \|\hat{\boldsymbol{\beta}}_z^t - \boldsymbol{\beta}_z^t\| + \|\widehat{\gamma}_z - \gamma_z\|\right)\|\widehat{\gamma}_z - \gamma_z\|.$$

*Then*

$$\hat{\theta} - \theta = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + S_{n,0} + S_{n,1}\right),$$

*and if* $S_{n,0} + S_{n,1} = o_{\mathbb{P}}(1/\sqrt{n})$, *we have*

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow N\left(0, var\left\{\frac{(\phi_1^y - \phi_0^y) - \theta(\phi_1^a - \phi_0^a)}{\mathbb{P}(\phi_1^a - \phi_0^a)}\right\}\right).$$

**Proof.**

Note that we have

$$\begin{aligned}
\hat{\theta} - \theta &= \frac{\mathbb{P}_n(\hat{\phi}_1^y - \hat{\phi}_0^y)}{\mathbb{P}_n(\hat{\phi}_1^a - \hat{\phi}_0^a)} - \frac{\mathbb{P}(\phi_1^y - \phi_0^y)}{\mathbb{P}(\phi_1^a - \phi_0^a)} \\
&= \frac{1}{\mathbb{P}_n(\hat{\phi}_1^a - \hat{\phi}_0^a)}\left[\{\mathbb{P}_n(\hat{\phi}_1^y - \hat{\phi}_0^y) - \mathbb{P}(\phi_1^y - \phi_0^y)\} - \theta\{\mathbb{P}_n(\hat{\phi}_1^a - \hat{\phi}_0^a) - \mathbb{P}(\phi_1^a - \phi_0^a)\}\right] \\
&= \mathbb{P}_n\left\{\frac{(\phi_1^y - \phi_0^y) - \theta(\phi_1^a - \phi_0^a)}{\mathbb{P}(\phi_1^a - \phi_0^a)}\right\} + o_{\mathbb{P}}(1/\sqrt{n}) \\
&\quad + \; O_{\mathbb{P}}\left(\|\hat{\pi} - \pi\| \max_z \left\|\hat{\lambda}_z - \lambda_z\right\| + \max_{z,t}\left\{\left(\left\|\hat{\boldsymbol{\beta}}_z^t - \boldsymbol{\beta}_z^t\right\| + \|\hat{\gamma}_z - \gamma_z\|\right)\|\hat{\gamma}_z - \gamma_z\|\right\}\right)
\end{aligned}$$

where the third line follows since

$$\mathbb{P}_n(\widehat{\phi}^t) - \mathbb{P}(\phi^t) = (\mathbb{P}_n - \mathbb{P})(\widehat{\phi}^t - \phi^t) + (\mathbb{P}_n - \mathbb{P})\phi^t + \mathbb{P}(\widehat{\phi}^t - \phi^t)$$

with the first term $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2 and the third remainder term from Theorem 2, and since $\mathbb{P}_n(\widehat{\phi}_1^a - \widehat{\phi}_0^a)$ is bounded away from zero with

$$\begin{aligned}
\mathbb{P}_n(\widehat{\phi}_1^a - \widehat{\phi}_0^a) - \mathbb{P}(\phi_1^a - \phi_0^a) &= (\mathbb{P}_n - \mathbb{P})(\widehat{\phi}_1^a - \widehat{\phi}_0^a) + \mathbb{P}\{(\widehat{\phi}_1^a - \widehat{\phi}_0^a) - (\phi_1^a - \phi_0^a)\} \\
&= O_{\mathbb{P}}(1/\sqrt{n}) + \max_z \|\widehat{\phi}_z^a - \phi_z^a\| = o_{\mathbb{P}}(1)
\end{aligned}$$

where the second equality follows from Lemma 2 and the central limit theorem. □

As before, the estimator $\hat{\theta}$ has a fast convergence rate that is second-order involving products of nuisance errors, so that under for example $n^{-1/4}$-type rates the estimator will be $\sqrt{n}$-consistent, asymptotically normal, and efficient. It is also doubly robust, as pointed out in the next corollary.

**Corollary 2.**

*Under the conditions of Theorem 3, the estimator $\hat{\theta}$ is consistent if either*

1. $\|\widehat{\gamma}_z - \gamma_z\| = o_{\mathbb{P}}(1)$*forz* $\in \{0, 1\}$ *and*$\|\widehat{\pi} - \pi\| = o_{\mathbb{P}}(1)$, *or*
2. $\|\widehat{\gamma}_z - \gamma_z\| = o_{\mathbb{P}}(1)$*and*$\|\widehat{\lambda}_z - \lambda_z\| = o_{\mathbb{P}}(1)$, *forz* $\in \{0, 1\}$.

To summarize, the above results extend the work of Chaudhuri and Guilkey [8], Mogstad and Wiswall [9], and Kennedy and Small [11], by providing an efficient nonparametric estimator of the instrumental variable estimand when some instrument values are missing, allowing adjustment for complex confounding via flexible data-adaptive estimators of the nuisance functions.

## 4 Discussion

In this paper we filled a gap in the literature by considering nonparametric identification, efficiency theory, and estimation of average treatment effects in the presence of complex confounding and missing exposure values, where the exposure missingness can depend not only on the covariates but also the outcome information. We derived the efficient influence function for the average treatment effect and corresponding nonparametric efficiency bounds, and constructed nonparametric estimators can attain these efficiency bounds under weak rate conditions on the nuisance estimators. This allows one to incorporate modern flexible regression and machine learning tools. We also apply our general results to the problem of causal inference with a partially missing instrumental variable, yielding a new estimator and efficiency bound in this problem as well.

There are several important avenues for future work. First, it will be useful to study finite-sample properties of the estimators proposed here, in comparison to the more parametric estimators proposed in earlier work. Relatedly, it would be useful to construct an efficient plug-in estimator using targeted maximum likelihood [28, 29], which would respect bounds on the parameter space, e. g. when $Y$ is bounded. Second, we restricted study to possibly multi-valued but discrete point treatments; it would be of interest to extend to treatments that are continuous [30, 31] or time-varying [32, 33]. This would also be useful for continuous instrumental variable problems [34] with instrument missingness. Further, identification, efficiency theory, and estimation are all more complicated in settings where there is simultaneous missingness in covariates, treatment, and outcome [35]; however, this also occurs often in practice and deserves deeper investigation. Lastly, we assumed exchangeability in the sense of the missing indicator $R$ being conditionally independent of the underlying exposure $Z$ given both covariates $\mathbf{X}$ and outcome $\mathbf{Y}$; it would be of interest to consider the case where we only assume $R \perp\!\!\!\perp Z \mid \mathbf{X}$. However, there average treatment effects are no longer point identified, and so one would need to consider bounds and/or sensitivity analysis.

## Acknowledgements

The author thanks Matteo Bonvini, Dylan Small, Mike Daniels, and Joe Hogan for helpful discussions, as well as an anonymous reviewer of a previous version of the manuscript.

## A Appendix

The following lemma from [21] is useful in proving Theorem 2.

**Lemma 2.**

*Let*$\hat{f}(\mathbf{o})$*be a function estimated from a sample*$\mathbf{O}^N = (\mathbf{O}_{n+1}, ..., \mathbf{O}_N)$, *and let*$\mathbb{P}_n$*denote the empirical measure over*$(\mathbf{O}_1, ..., \mathbf{O}_n)$, *which is independent of*$\mathbf{O}^N$. *Then*

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\hat{f} - f\|}{\sqrt{n}}\right).$$

**Proof.**

First note that, conditional on $\mathbf{O}^N$, the term in question has mean zero since

$$\mathbb{E}\Big\{\mathbb{P}_n(\hat{f}-f)\,\Big|\,\mathbf{O}^N\Big\} = \mathbb{E}(\hat{f}-f \mid \mathbf{O}^N) = \mathbb{P}(\hat{f}-f).$$

The conditional variance is

$$\mathrm{var}\Big\{(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)\,\Big|\,\mathbf{O}^N\Big\} = \mathrm{var}\Big\{\mathbb{P}_n(\hat{f}-f)\,\Big|\,\mathbf{O}^N\Big\} = \frac{1}{n}\mathrm{var}(\hat{f}-f \mid \mathbf{O}^N) \leq \|\hat{f}-f\|^2/n.$$

Therefore using Chebyshev's inequality we have

$$\mathbb{P}\left\{\frac{|(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)|}{\|\hat{f}-f\|/\sqrt{n}} \geq t\right\} = \mathbb{E}\left[\mathbb{P}\left\{\frac{|(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)|}{\|\hat{f}-f\|/\sqrt{n}} \geq t \,\Big|\, \mathbf{O}^N\right\}\right] \leq \frac{1}{t^2}.$$

Thus for any $\varepsilon > 0$ we can pick $t = 1/\sqrt{\epsilon}$ so that the probability above is no more than $\varepsilon$, which yields the result. □

## References

[1] Zhang Z, Liu W, Zhang B, Tang L, Zhang J. Causal inference with missing exposure information: methods and applications to an obstetric study. Stat Meth Med Res. 2016;25:2053–66.

[2] Shortreed SM, Forbes AB. Missing data in the exposure of interest and marginal structural models: a simulation study based on the framingham heart study. Stat Med. 2010;29:431–43.

[3] Ahn J, Mukherjee B, Gruber SB, Sinha S. Missing exposure data in stereotype regression model: application to matched case–control study with disease subclassification. Biometrics. 2011;67:546–58.

[4] Shardell M, Hicks GE. Statistical analysis with missing exposure data measured by proxy respondents: a misclassification problem within a missing-data problem. Stat Med. 2014;33:4437–452.

[5] Molinari F. Missing treatments. J Bus Econ Stat. 2010;28:82–95.

[6] Mebane Jr WR, Poast P. Causal inference without ignorability: identification with nonrandom assignment and missing treatment data. Political Anal. 2013;21:233–51.

[7] Burgess S, Seaman S, Lawlor DA, Casas JP, Thompson SG. Missing data methods in Mendelian randomization studies with multiple instruments. Am J Epidemiol. 2011;174:1069–76.

[8] Mogstad M, Wiswall M. Instrumental variables estimation with partially missing instruments. Econ Lett. 2012;114:186–9.

[9] Chaudhuri S, Guilkey DK. GMM with multiple missing variables. J Appl Econometrics. 2016;31:678–706.

[10] Williamson E, Forbes A, Wolfe R. Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. Stat Med. 2012;31:4382–400.

[11] Kennedy EH, Small DS. Paradoxes in instrumental variable studies with missing data and one-sided noncompliance. J French Stat Soc. 2017.

[12] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat. 2004;86:4–29.

[13] van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York: Springer, 2003.

[14] Bickel PJ, Klaassen CA, Ritov Y, Wellner JA. Efficient and adaptive estimation for semiparametric models. Baltimore: Johns Hopkins University Press, 1993.

[15] van der Vaart AW. Semiparametric statistics. In: Lectures on probability theory and statistics. Berlin Heidelberg: Springer, 2002:331–457.

[16] Tsiatis AA. Semiparametric theory and missing data. New York: Springer, 2006.

[17] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89:846–66.

[18] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double machine learning for treatment and causal parameters. arXiv preprint arXiv:1608.00060, 2016.

[19] Robins JM, Li L, Tchetgen Tchetgen EJ, van der Vaart AW. Higher order influence functions and minimax estimation of nonlinear functionals. Probability and Statistics: Essays in Honor of David A. Freedman, 2008:335–421.

[20] Zheng W, van der Laan MJ. Asymptotic theory for cross-validated targeted maximum likelihood estimation. UC Berkeley Division Biostat Working Paper Ser. 2010;273:1–58.

[21] Kennedy EH, Balakrishnan S, G'Sell M. Sharp instruments for classifying compliers and generalizing causal effects. The Ann Stat. 2019.

[22] Farrell MH. Robust inference on average treatment effects with possibly more covariates than observations. J Econometrics. 2015;189:1–23.

[23] J. M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. Proc Am Stat Assoc. 2000;1999:6–10.

[24] Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc. 1999;94:1096–120.

[25] Abadie A. Semiparametric instrumental variable estimation of treatment response models. J Econometrics. 2003;113:231–63.

[26] Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. Econometrica. 1994;62:467–75.

[27] Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology. 2006;17:360–72.

[28] van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. NYC: Springer, 2011.
[29] van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. UC Berkeley Division of Biostatistics Working Paper Series, 2006:212.
[30] Díaz I, van der Laan MJ. Population intervention causal effects based on stochastic interventions. Biometrics. 2012;68:541–9.
[31] Kennedy EH, Ma Z, McHugh MD, Small DS. Nonparametric methods for doubly robust estimation of continuous treatment effects. J R Stat Soc: Ser B. 2017;79:1229–45.
[32] Kennedy EH. Nonparametric causal effects based on incremental propensity score interventions. J Am Stat Assoc. 2019;114:645–56.
[33] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550–60.
[34] Kennedy EH, Lorch S, Small DS. Robust causal inference with continuous instruments using the local instrumental variable curve. J R Stat Soc: Ser B. 2019;81:121–43.
[35] Sun B, Tchetgen Tchetgen EJ. On inverse probability weighting for nonmonotone missing at random data. J Am Stat Assoc. 2018;113:369–79.