NONPARAMETRIC SCANNING FOR NONRANDOM MISSING DATA WITH CONTINUOUS INSTRUMENTAL VARIABLES

A PREPRINT

Arkaprabha Ganguli
Department of Statistics and Probability
Michigan State University
East Lansing, Michigan 48824, U.S.A.
gangulia@msu.edu

David Todem *
Department of Epidemiology and Biostatistics
Michigan State University
East Lansing, Michigan 48824, U.S.A.
todem@msu.edu

November 05, 2021

ABSTRACT

This article introduces a new instrumental variable approach for estimating unknown population parameters with data having nonrandom missing values. With coarse and discrete instruments, Shao and Wang (2016) proposed a semiparametric method that uses the added information to identify the tilting parameter from the missing data propensity model. A naive application of this idea to continuous instruments through arbitrary discretizations is apt to be inefficient, and maybe questionable in some settings. We propose a nonparametric method not requiring arbitrary discretizations but involves scanning over continuous dichotomizations of the instrument; and combining scan statistics to estimate the unknown parameters via weighted integration. We establish the asymptotic normality of the proposed integrated estimator and that of the underlying scan processes uniformly across the instrument sample space. Simulation studies and the analysis of a real data set demonstrate the gains of the methodology over procedures that rely either on arbitrary discretizations or moments of the instrument.

Keywords Continuous instrument variable · Dichotomizations · Empirical processes · Exponential tilting · Identifiability · Kernel regression.

1 Introduction

Analysis of incomplete data poses a recurring challenge in statistics. It is well-established that inferences may be misleading if the working model assumptions are not consistent with the underlying missing data mechanism (Molenberghs et al., 1997). Existing strategies oftentimes used to account for some degree of selection bias due to nonresponse, typically rely on the assumption that missingness can effectively be rendered independent of the outcome, upon conditioning on a sufficiently rich set of observed variables (Robins et al., 1995; Tsiatis & Davidian, 2007; Rubin, 1987; Little & Rubin, 2002; and Kim & Shao, 2021). This assumption is most often made in practice, and formally entails a missing at random (MAR) data mechanism. Without imposing additional conditions, such a mechanism is strictly untestable in view of observed data. Consequently, it is often prudent to conservatively assume that systematic differences between respondents and nonrespondents persist even after covariate adjustment, consistent with nonrandom nonresponse mechanism (Rotnitzky et al., 2001; Linero, 2017; and Yang et al., 2019).

Identifiability issues commonly arise with nonrandom missing data, where the parameters of the model for the missingness may not be jointly identifiable with those from the measurement model using only the observed data (Scharfstein et al., 1999; Todem et al., 2010; Zhao and Shao, 2015). Analytic strategies addressing these concerns include methods that rely on identification on parametric restrictions which may be subject to misspecification (Wu & Carroll, 1988; Molenberghs et al., 1997; Troxel et al., 1998 and Wang et al., 2014).

*Corresponding author : David Todem
A more recent method that has been shown to be essential for the model identification is the use of an instrumental variable (IV). At least two versions of this approach have appeared in the literature. While the first version treats the IV as being independent of the missing data indicator given the response (Shao & Wang, 2016), the second version rather treats the IV as being independent of the outcome given the missing data indicator (Tchetgen Tchetgen and Wirth, 2017). Although these formulations may have different substantive interpretations, they are fundamentally equivalent from a distributional viewpoint. In fact, one can argue that they are closely connected to the selection and pattern mixture model taxonomy. For this reason, we will adopt, in the spirit of Shao & Wang (2016), the first IV representation which allows a direct characterization of parameters indexing the law of data given covariates.

To fix ideas, let \( y \) denote a univariate outcome subject to nonresponse, and \( \delta \) the associated response status indicator with \( \delta = 1 \) indicating an observed value. Let \( z \) be a nonresponse IV, and \( u \) an auxiliary variable, both assumed to be continuous and fully observed. As such, \( z \) is assumed to be associated with \( \delta \) only through its dependence on \( y \), such that \( z \) and \( \delta \) are conditionally independent given \( y \). Consequently, \( z \) must predict a person’s propensity to have an observed outcome through its direct influence on the outcome itself; that is \( \text{pr}(\delta = 1 \mid y, u, z) = \text{pr}(\delta = 1 \mid y, u) \).

Letting \( x = (u, z) \) and \( \pi(y, x) = \text{pr}(\delta = 1 \mid y, x) \) and following the formulation of Kim & Yu (2011), we consider an exponential tilting model of the form,

\[
\pi(y, x) = \left[ 1 + \exp\{g(u) + \gamma y\} \right]^{-1},
\]

where \( \gamma \) is the tilting parameter measuring the extent of deviation of \( \delta \) from MAR, \( g \) is an unknown and completely unspecified function of \( u \). Unlike other approaches that rely on either fixing \( \gamma \) to some known value or estimating \( \gamma \) using external data (e.g., Kim and Yu, 2011; Zhao et al, 2013 and Tang et al, 2014), added information from IVs can be used to formulate additional estimating equations for \( \gamma \) while profiling \( g \). With discrete IVs, Shao & Wang (2016) showed that the profiled estimating equations can be used to consistently estimate \( \gamma \). A straightforward application of this idea to continuous IVs via arbitrary discretizations is apt to be suboptimal. Because of its dependency on thresholds, this approach may fail to provide a unique answer to the substantive question in some applications. It may further be subject to the theoretical complications arising with the data-dependent threshold placement. A comprehensive discussion on the danger of discretizing continuous variables can be found elsewhere (e.g., Altman, 1998; Royston et al., 2004 and Peng & Fine, 2008).

In this paper, we propose a novel and flexible approach that does not rely on arbitrary discretizations but involves scanning through continuous thresholding of the instrument and combining scan statistics to estimate the unknown parameter through weighted integration. The key to effective inference under continuous thresholding lies in its ability to improve efficiency, especially when there are local dependencies between \( z \) and \( \delta \) involving a small portion of the sample. An interesting feature of the scanning approach is that the efficient gain extends naturally to settings involving multiple instruments, albeit a heavy computational cost. Numerical simulations and an analysis of data generated from an income panel study demonstrate the gains of the methodology over procedures that rely either on arbitrary discretizations or moments of the instrument. Technical details are given in the Supplementary Material.

## 2 The scanning method

Suppose one is interested in estimating \( \theta \), a \( q \)-dimensional parameter vector from the law \( f(y \mid x) \) that uniquely solves \( E\{\eta(y, x, \theta) \mid x\} = 0 \), for a generic data function \( \eta(\cdot) \). To this end, assume that an independent and identically distributed sample \( \{y_i, x_i, \delta_i\}_{i=1}^n \) of \( \{y, x, \delta\} \) is readily available. As alluded to in the introduction, we exploit the information contained in instrument \( z \) to identify and estimate the tilting parameter \( \gamma \) from the propensity model in (1).

Let \( \Xi \subseteq \mathbb{R} \) be the support of \( z \), and \( \Xi^o \) its interior. A naive strategy is to discretize \( z \) using say \( p \geq 1 \) finite arbitrary thresholds \( \xi_1 < \xi_2 < \cdots < \xi_p \) of \( \Xi^o \) to construct \( p + 1 \) instrumental estimating functions,

\[
M_{\xi_j}(y, u, \delta; g, \gamma) = I\{\xi_{j-1} < z \leq \xi_j\} \left\{ \frac{\delta}{\pi(y, u; g, \gamma)} - 1 \right\}, \quad j = 1, \cdots, p + 1.
\]

Here we set \( \xi_0 = -\infty \) and \( \xi_{p+1} = +\infty \), and \( I\{A\} \) is the indicator function for event \( A \). Letting \( M_{\xi} \) be a column vector made of functions \( \{M_{\xi_j}\}_{j=1}^{p+1} \) obtained through stratification, we define the instrumental estimating equation,

\[
E\{M_{\xi}(y, u, \delta; g, \gamma) \mid u\} = 0.
\]

This estimating function is made possible because of conditional independence between \( z \) and \( \delta \) given \( y \). For a fixed \( \gamma \), let \( g_{\gamma} \) be the function defined such that,

\[
\exp\{g_{\gamma}(u)\} = \frac{E\{(1 - \delta) \mid u\}}{E\{\delta \exp(\gamma y) \mid u\}}.
\]
Let $\gamma_0$ be the true value of $\gamma$ that solves (3). From Equation $E\{1'M_\xi(y, u, \delta; g, \gamma) \mid u\} = 0$, where $1$ is a $(p + 1)$-column vector of $1$, it can be readily shown that,

$$
\exp\{g(u)\} = \frac{E\{(1 - \delta) \mid u\}}{E\{\delta \exp(\gamma_0 y) \mid u\}},
$$

such that, $g = g_{\gamma_0}$. Because of the conditional expectations in (4), a kernel method may be used to estimate $g(u) = g_{\gamma_0}(u)$ provided $\gamma_0$ the true $\gamma$ is known or estimated from historical data. Because prior information on $\gamma_0$ is oftentimes unavailable in real applications, one may rely on $z$ for its estimation. First, we profile $g_{\gamma}(u)$ for fixed $\gamma$ as,

$$
\exp\{g_{\gamma}(u)\} = \frac{\sum_{i=1}^{n}(1 - \delta_i)K_h(u - u_i)}{\sum_{i=1}^{n}\delta_i \exp(\gamma y_i)K_h(u - u_i)},
$$

where $K_h(\cdot) = K(\cdot)/h$, with $K(\cdot)$ being a symmetric kernel function and $h$ a bandwidth. Estimation of $\gamma$ then follows by solving the sample version of equation (4),

$$
\frac{1}{n} \sum_{i=1}^{n} M_\xi(y_i, u_i, \delta_i; \hat{g}_\gamma, \gamma) = 0.
$$

Upon estimating $\gamma$ by $\hat{\gamma}$ and $g_{\gamma_0}$ with $\hat{g}_\gamma$, one may apply the inverse propensity weighting approach to estimate $\theta$. Specifically, an estimator of $\theta$ may be obtained by solving the equation,

$$
\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \eta(y_i, x_i, \theta)}{\pi(y_i, u_i; \hat{g}_\gamma, \gamma)} = 0.
$$

As $n \to \infty$, it can be shown under basic regularity conditions that the resulting estimator of $\theta$ converges asymptotically at the $\sqrt{n}$ rate to a normal distribution, despite $g(u)$ being nonparametric. An even stronger asymptotic result can be established for $\hat{\gamma}$ (Shao & Wang, 2016).

Because this estimation approach relies on arbitrary choices of thresholds and partial information from the instrument $z$, $I\{\xi_{j-1} < z \leq \xi_j\}, j = 1, \cdots, p + 1$, the resulting estimators of $\gamma$ and $\theta$ are likely to be suboptimal. This deficiency may be alleviated with a proper choice of $\{\xi_j\}_{j=1}^p$, which is generally a daunting task in practice without knowing the true data generating process. A natural and conservative strategy avoiding arbitrary choices of thresholds, is to vary the thresholds continuously over the sample space and accumulate results. Our idea then proceeds through scanning via continuous dichotomizations, which is computationally very efficient while maintaining information contained in the instrument $z$. This desirable feature of continuous dichotomizations relies on the property that the Borel $\sigma$-algebra for the real line $\mathbb{R}$ is generating by half-infinite intervals and by bounded intervals; that is $\sigma\{(a, b) : a, b \in \mathbb{R}\} = \sigma\{(\infty, b) \cap \Xi : b \in \mathbb{R}\} \cap \sigma\{(a, \infty) \cap \Xi : a, b \in \mathbb{R}\}$.

Operationally, we set $p = 1$ and regard estimators of $\gamma$ and $\theta$ in (3) and (6) as processes in single threshold $\xi$, which we denote $\hat{\gamma}_\xi$ and $\hat{\theta}_\xi$ to highlight this dependence. From these processes in $\xi$, we estimate $\theta$ as,

$$
\hat{\theta} = \arg\min_{\theta \in \Theta} \int_{\Xi^*} (\hat{\theta}_\xi - \theta)^T W_\xi (\hat{\theta}_\xi - \theta) d\xi,
$$

where $\Xi^* \subset \Xi^0$ with general requirements specified in (3). $W_\xi$ is a positive definite matrix of weights converging to $w_\xi$ uniformly in $\xi$. An estimator of $\theta$ is,

$$
\hat{\theta} = \left\{ \int_{\Xi^*} W_\xi d\xi \right\}^{-1} \int_{\Xi^*} \hat{\theta}_\xi W_\xi d\xi.
$$

(7)

Weights $W_\xi$ are chosen to ensure the stability of $\hat{\theta}$, with special cases being $\hat{V}(\hat{\theta}_\xi)^{-1}$ the inverse of the estimated variance-covariance of $\hat{\theta}_\xi$, and $\{1/\hat{p}_\xi + 1/(1 - \hat{p}_\xi)\}^{-1}I_q = \hat{p}_\xi(1 - \hat{p}_\xi)I_q$; $\hat{p}_\xi$ is the empirical version of $Pr(z \leq \xi_0)$, and $I_q$ the identity matrix of order $q$. While $\hat{V}(\hat{\theta}_\xi)^{-1}$ puts more weights on estimates that are less variable, $p_c(1 - p_c)I_q$ weighs more scans toward the median of $z$.

It is worth noting that while the current presentation focuses primarily on continuous instruments, the proposed method is generally applicable to IVs with an intrinsic order. A good example of such instruments is the class of ordinal categorical variables where the underlying ranking is automatically incorporated into the estimation. For such instruments, one would argue that the proposed scanning approach will generally outperform the method proposed by Shao & Wang (2016) which neglects the natural order of the instrument.
A competitive approach to the scanning method with continuous instruments, is the moment-based procedure, consisting of estimating $\gamma$ via the equation,

$$E \left\{ z^\ell \left( \frac{\delta}{\pi(y, u; g, \gamma)} - 1 \right) \mid u \right\} = 0, \ell = 1, \cdots, L.$$ \hspace{1cm} (8)

As we shall see hereafter, this approach has limited applications when $z$ is sparse albeit being quantitative. Furthermore, it is unclear how to choose $L$ or apply this approach to ordinal categorical instruments for which the scanning approach remains applicable.

### 3 Theory

Although the idea of scanning and integrating may appear intuitive, formally establishing the large sample properties of the resulting estimator is not trivial owing to the dependency among scan statistics. Consequently, a careful study using empirical process theory may be invoked for asymptotics. To this end, we adopt the primitive requirements of Shao & Wang (2016) pertaining to $\hat{\gamma}_\xi$ and $\hat{\theta}_\xi$ for each fixed $\xi \in \Xi^*$, which we regard as pointwise conditions. These weaker requirements are then strengthened with higher level conditions to guarantee the uniform convergence of empirical processes in $\xi$, $\hat{\gamma}_\xi$ and $\hat{\theta}_\xi$.

The following theorems establish key asymptotic results: uniform consistency and asymptotic normality of $\hat{\gamma}_\xi$; uniform asymptotic distribution of the stochastic process $\hat{\theta}_\xi$ and finally the asymptotic normality of our proposed estimator $\hat{\theta}$.

It is worth noting that $\hat{\gamma}_\xi$ for each fixed $\xi$, falls into the realm of ‘Semiparametric-M’ estimators for which general pointwise asymptotic results have been established (e.g., Khan & Powell, 2001). Theorems 1 and 2 give a stronger uniform results across the threshold space.

**Theorem 1** Assume $Pr(z \leq \inf \Xi^*) > 0$, $Pr(z \geq \sup \Xi^*) > 0$ and Conditions 1-5 in Appendix hold. Let $\gamma_{0\xi}$ be the unique solution to $m_\xi(g_\gamma, \gamma) = 0$, with $m_\xi(g_\gamma, \gamma) = E\{M_\xi(y, u, \delta; g, \gamma) \mid u\}$. If $\limsup_{\gamma \in \Lambda, \xi \in \Xi^*} \parallel m_\xi(g_\gamma, \gamma) \parallel < \infty$, then

$$\sup_{\xi \in \Xi^*} |\hat{\gamma}_\xi - \gamma_{0\xi}| \rightarrow 0, \text{ as } n \rightarrow \infty.$$  

**Theorem 2** Assume $Pr(z \leq \inf \Xi^*) > 0$, $Pr(z \geq \sup \Xi^*) > 0$ and Conditions 1-7 in Appendix hold. Then as $n \rightarrow \infty$, the random process $n^{1/2}(\hat{\gamma}_\xi - \gamma_{0\xi})$ in $\xi$ converges in distribution to a zero mean Gaussian process with some covariance kernel $\sigma_0(\cdot, \cdot)$.

Unlike $\hat{\gamma}_\xi$ which has a converging point dependent on $\xi$ both in probability and distribution, we show that $\hat{\theta}_\xi$ only converges in distribution and the centering point does not depend on $\xi$. For this reason, one would expect the scanning process to have an impact on precision but not on bias, at least in large samples.

**Theorem 3** Assume $Pr(z \leq \inf \Xi^*) > 0$, $Pr(z \geq \sup \Xi^*) > 0$ and Conditions 1-7 in Appendix hold. Then as $n \rightarrow \infty$, the random process $n^{1/2}(\hat{\theta}_\xi - \theta)$ in $\xi$, converges in distribution to a zero mean tight Gaussian process $\mathcal{G}$ with some covariance kernel $\sigma_1(\cdot, \cdot)$.

To prove Theorem 3 pertaining to the convergence of $v_n(\xi) = n^{1/2}(\hat{\theta}_\xi - \theta)$, one needs to show (Dudley, 1978); i) the sequence $(v_n(\xi_1), v_n(\xi_2), \cdots, v_n(\xi_k))$ converges in distribution to a centered multivariate normal distribution for every finite set of points $\{\xi_j\}_{j=1}^k$ of $\Xi$; and ii) the asymptotic equicontinuity condition

$$\forall \epsilon > 0, \lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} P[\sup_{||\xi - \xi'|| < \delta} \parallel v_n(\xi) - v_n(\xi') \parallel > \epsilon] = 0.$$  

**Theorem 4** Under Conditions 1-8 specified in Appendix and finite variance assumption specified in the Supplementary Material, $n^{1/2}(\hat{\theta} - \theta)$ the scaled version of the integrated estimator $\hat{\theta}$ converges in distribution to a centered normal distribution $N(0, \Sigma)$, as $n \rightarrow \infty$.

The asymptotic variance-covariance $\Sigma$ has a very complicated form and may require resampling techniques for estimation. Basic regularity conditions, including those pertaining to the kernel estimation of the unspecified function $g$, are relegated to the Appendix. Proofs of all theorems except theorem 2 which follows directly from that of Shao & Wang (2016), are provided in the Supplementary Material.
4 Simulation Study

We compare the small sample performance of the scanning procedure with that of competing methods for the estimation of the overall mean $\mu = E[y]$. Specifically, in addition to the scanning-based estimator $\hat{\mu}$ of $\mu$, our investigation focuses on the following alternatives:

- $\hat{\mu}_{0.0.2.0.4}$, the estimator with a single fixed threshold at quantile $\xi_0$ such that $P(z \leq \xi_0) = \alpha$, $\alpha \in \{0.2, 0.4, 0.6\}$, and with a slight abuse of notation we simply write $\hat{\mu}_\alpha$,
- $\hat{\mu}_{0.0.2.0.4}$ and $\hat{\mu}_{0.4.0.6}$, the estimator with two fixed thresholds at quantiles $\{0.2, 0.4\}$ and $\{0.4, 0.6\}$ of $z$,
- $\bar{y} = \sum_{i=1}^{n} y_i/n$, the sample mean of $y$ when there are no missing data,
- $\bar{y}_{\text{obs}} = \sum_{i=1}^{n} \delta_i y_i / \sum_{i=1}^{n} \delta_i$, the sample mean of the observed $y_i$, complete case analysis,
- $\hat{\mu}_\gamma$ the estimator of Kim & Yu (2011), with known tilting parameter $\gamma$ taking values $\gamma_0$ the true parameter, $\gamma_0/n$ the wrong tilting parameter, or $0$ for the MAR assumption,
- $\hat{\mu}$ the moment-based estimator with $L = 1$.

The estimator $\hat{\mu}$ of $\mu$ is obtained by setting $\Xi^* = (\xi_0, 1, \xi_0, 0.9)$, and weighing the scan statistics $\hat{\mu}_\xi = \sum_{i=1}^{n} \delta_i y_i / \pi(y_i; \hat{\xi}_\xi, \hat{\gamma}_\xi)$ by the inverse of their estimated variance. That is,

$$\hat{\mu} = \left\{ \int_{\Xi^*} \hat{V}^{-1}(\hat{\mu}_\xi) \, d\xi \right\}^{-1} \int_{\Xi^*} \hat{\mu}_\xi \hat{V}^{-1}(\hat{\mu}_\xi) \, d\xi.$$

Estimators $\hat{\mu}_{0.2}, \hat{\mu}_{0.4}, \hat{\mu}_{0.6}, \hat{\mu}_{0.2.0.4}$ and $\hat{\mu}_{0.4.0.6}$ are based on the approach in Shao & Wang (2016) for categorical instruments with two and three categories. The unspecified function $g(\cdot)$ in the exponential tilting model is estimated using a nonparametric Gaussian Kernel $K(u) = \exp(-u^2/2)/(2\pi)^{1/2}$ (Gasser & Müller, 1979). The bandwidths are selected adaptively to the sample stratification for estimation involving both fixed and continuous discretizations. For example with single thresholds, the bandwidths are selected depending on $\xi$, i.e., $h_{n,\xi} = 1.5\sigma_{u,\xi}/n_{\xi}^{-1/3}$ where $\sigma_{u,\xi}$ is the standard deviation of the observed variable $u_i$ with $z_i \leq \xi$; and $n_{\xi} = \sum_{i=1}^{n} I(z_i \leq \xi)$ is the stratum sample size.

Because the proposed method is applicable to instruments with an intrinsic order, the investigation is conducted for continuous and ordinal categorical instruments. For each data setting, we set the sample size to 200 and 500, and use 1000 Monte Carlo replicates to estimate the method’s relative bias, mean square errors, standard errors and coverage probability of confidence intervals (generated using normal approximations with bootstrapped standard errors).

4.1 Using continuous instrumental variable

We consider the following hierarchical structure for the full complete data:

$$z \sim N(0, 1), \ u \ | \ \{z\} \sim N(z, 1), \ y \ | \ {u, z} \sim N(E[y|u, z], 1).$$

Here $E[y|u, z] = \lbrack \sin(u^3) + 10u^2 \rbrack I[z \leq 0.84] + \lbrack 40 + \sin(u^3) + 10u \rbrack I[z > 0.84]$, which by averaging across $\{u, z\}$ gives the unconditional mean $\mu = 24.4$. Upon generating $\{z, u, y\}, \ we \ generate \ the \ missingness \ indicator \ \delta \ from \ Bernoulli \ distribution \ with \ probability \ \pi$. Four MNAR models ($M_1 - M_4$), and one MAR model ($M_5$) are entertained, resulting in the overall missing data rate of about 30%:

- $M_1 : \pi = 1/(1 + \exp(\alpha + \beta u + \gamma y))$ with $(\alpha, \beta, \gamma) = (2.8, -0.3, -0.6)$;
- $M_2 : \pi = 1/(1 + \exp(\alpha + \beta \sin(u) + \gamma y))$ with $(\alpha, \beta, \gamma) = (2.75, -0.2, -0.6)$;
- $M_3 : \pi = 1/(1 + \exp(\alpha + \beta u^2 + \gamma y))$ with $(\alpha, \beta, \gamma) = (2.8, -0.3, -0.6)$;
- $M_4 : \pi = 1/(1 + \exp(\alpha + \beta \exp(u) + \gamma y))$ with $(\alpha, \beta, \gamma) = (3, -0.3, -0.6)$;
- $M_5 : \pi = 1/(1 + \exp(\alpha + \beta \exp(u)))$ with $(\alpha, \beta) = (-0.9, -0.3)$;

Results from Tables 1 and 2 show that $\bar{y}_{\text{obs}}$ is biased and does not have good coverage probability under all missing data models considered. No other estimation method, except the benchmark $\bar{y}$, dominates the others uniformly across all data configurations, both in terms of bias and efficiency. The estimator $\hat{\mu}_{\alpha}$, which also serves as a benchmark when the true $\gamma$ is used, also has superior performance, both under the four MNAR and the MAR models. As expected, the naive estimators based on fixed single thresholds appear to have varying performance in realistic samples, that stabilizes with further discretizations. When pitched against the naive estimators, the scanning approach behaves robustly across most scenarios. This is consistent with early theoretical results that the scanning method enjoys some form of optimality as it accumulates information across all cutpoints. It is worth noting that when the 60th percentile is used, $\hat{\mu}_{0.6}$ outperforms $\hat{\mu}$ under all MNAR models. Because finding such an optimal cutpoint is a daunting task in
applications, continuous discretizing appears to be the most conservative strategy. This is especially true if there are local dependencies between $z$ and $\delta$ involving a small portion of the sample. As expected, $\bar{\mu}$, performs poorly when a wrong titling parameter is assumed. For example, $\hat{z}$ with $z=\bar{\mu}$ outperforms $\bar{\mu}$ in settings involving an instrument with a two-component mixture distribution.

It is also interesting to note that $\hat{\mu}$ outperforms $\bar{\mu}$ under $M_5$, but the underlying reason is unclear. Further simulations are conducted to compare $\hat{\mu}$ and $\bar{\mu}$ in settings involving an instrument with a two-component mixture distribution (results relegated to Supplementary Material). Essentially under zero-inflated continuous data, the moment-based estimator performs poorly regardless of the degree of the moment $L$ being chosen. This simulation result provides an excellent example of the limitation of this method in settings involving sparse data on the instrument.

### 4.2 Using ordinal instrumental variable

We now assume that the instrument $z$ is ordinal with 3 ordered levels $c_0 < c_1 < c_2$ but the distances between these categories are not known. Let $z^*$ be the underlying variable that generates $z$ such that $z = c_k \leftrightarrow z^* = k$, $k \in \{0, 1, 2\}$, with $p(z = c_k) = p_k$. Assuming $z^* \sim \text{Binomial}(2, 0.4)$, we consider the hierarchical structure for the full complete data:

$$
\begin{align*}
z & \sim \{(c_k, p_k)\}_{k=0}^2, 
\{z = c_k\} & \sim N(k, 1), 
y & \{u, z = c_k\} \sim N(E\{y|u, z = c_k\}, 1),
\end{align*}
$$

where $E\{y|u, z\} = [1 + 0.5(u - 1)^2]I\{z = c_0\} + u^2I\{z = c_1\} + [2 + (u - 2)^2]I\{z = c_2\}$.

The unconditional mean of $y$ is $\mu = 4.13$. Three MNAR models $M_1^*, M_2^*, M_3^*$ and one MAR model $M_4^*$ were considered, keeping the unconditional rate of missingness at 30%:

- $M_1^*$: $\pi = 1/\{1 + \exp(\alpha + \beta u + \gamma y)\}$ with $(\alpha, \beta, \gamma) = (0.4, -0.3, -0.2)$;
- $M_2^*$: $\pi = 1/\{1 + \exp(\alpha + \beta \sin(u) + \gamma y)\}$ with $(\alpha, \beta, \gamma) = (0.2, -0.9, -0.2)$;
- $M_3^*$: $\pi = 1/\{1 + \exp(\alpha + \beta u^2 + \gamma y)\}$ with $(\alpha, \beta, \gamma) = (0.5, -0.2, -0.1)$;
- $M_4^*$: $\pi = 1/\{1 + \exp(\alpha + \beta u)\}$ with $(\alpha, \beta) = (-0.4, -0.2)$.

### Table 1: Relative biases (RBs), Mean square errors (MSEs) for $n=200$ and 500; all RBs are multiplied by 100

| Estimator | Quantity | $n$ = | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ |
|-----------|----------|------|-------|-------|-------|-------|-------|
| $\hat{\mu}$ | MSE | 200 | 8.0  | 8.3  | 8.2  | 8.2  | 8.4  |
|           |        | 500 | 8.0  | 8.3  | 8.2  | 8.2  | 8.4  |
| $\hat{\mu}_{0.2}$ | MSE | 200 | 10.3 | 10.3 | 10.5 | 10.3 | 10.3 |
|           |        | 500 | 9.3  | 19.7 | 19.8 | 18.8 | 19.2 |
| $\hat{\mu}_{0.4}$ | MSE | 200 | 4.6  | 2.6  | 4.4  | 4.3  | 5.5  |
|           |        | 500 | 3.8  | 9.8  | 4.1  | 9.7  | 12.8 |
| $\hat{\mu}_{0.6}$ | MSE | 200 | 2.5  | 1.6  | 2.7  | 2.6  | 8.1  |
|           |        | 500 | 2.8  | 7.2  | 7.2  | 7.0  | 14.6 |
| $\hat{\mu}_{0.2,0.4}$ | MSE | 200 | 8.1  | 5.5  | 8.2  | 7.7  | 3.7  |
|           |        | 500 | 6.2  | 13.7 | 14.2 | 13.4 | 9.6  |
| $\hat{\mu}_{0.4,0.6}$ | MSE | 200 | 4.8  | 3.7  | 4.6  | 4.5  | 4.2  |
|           |        | 500 | 3.9  | 9.0  | 3.9  | 8.7  | 10.1 |
| $\hat{\mu}$ | MSE | 200 | 4.2  | 3.7  | 3.9  | 4.3  | 3.7  |
|           |        | 500 | 3.7  | 8.3  | 3.5  | 3.3  | 7.6  |
| $\bar{y}$ | MSE | 200 | 0.4  | 0.3  | 0.5  | 0.5  | 0.2  |
|           |        | 500 | 2.6  | 3.2  | 2.6  | 2.4  | 2.4  |
| $\bar{y}_{obs}$ | MSE | 200 | 40.5 | 40.8 | 40.5 | 39.2 | 39.3 |
|           |        | 500 | 103.4| 108.3| 99.9 | 92.1 | 91.2 |
| $\bar{\mu}_{\gamma_0}$ | MSE | 200 | 6.4  | 2.5  | 6.5  | 6.5  | 6.1  |
|           |        | 500 | 2.3  | 6.1  | 2.4  | 2.5  | 9.0  |
| $\bar{\mu}_{\gamma_0}/n$ | MSE | 200 | 16.6 | 18.4 | 16.7 | 16.4 | 16.1 |
|           |        | 500 | 23.1 | 24.6 | 23.6 | 23.7 | 22.7 |
| $\bar{\mu}_0$ | MSE | 200 | 17.5 | 18.8 | 17.5 | 17.2 | 16.9 |
|           |        | 500 | 23.9 | 26.2 | 24.4 | 25.3 | 24.3 |
Here the finite sample performance of the following estimators of $\mu$ are compared: $\hat{\mu}$ our proposed estimator along with $\tilde{\mu}$ the estimator proposed in Shao & Wang (2016) with 3 strata; $\bar{y}$, $\bar{y}_{obs}$, $\tilde{\mu}_{0}$, $\tilde{\mu}_{0}/n$ and $\hat{\mu}_{0}$ as previously described. Because instrument $z$ is not quantitative, the moment-based estimator $\tilde{\mu}$ is not applied. Results in Table 3 show that $\hat{\mu}$, the scanning-based estimator, performs almost similarly with $\tilde{\mu}$ in terms of bias and MSE. As expected, $\hat{\mu}$ consistently produces higher coverage probabilities of $\mu$ for all propensity models relative to $\tilde{\mu}$ (see Table 4). The other estimators continue to behave as expected. For example, the complete case analysis continues to underperform in terms of bias and MSE, especially when the missing data model deviates substantially from missing completely at random data mechanism.

### 5 Income panel data

We reanalyze income data that were generated from the Korean Labor and Income Panel Study. Detailed account of this survey and previous related analyses focusing on missingness can be found elsewhere (e.g., Wang et al, 2014; and Shao & Wang, 2016). In this survey, 2506 regular wage earners were interviewed for their 2006 monthly income regarded as the response $y$, in addition to their prior year income, gender (male versus female), age (years) and education level (high school and below versus above high school). For our analysis, the 2005 monthly income is treated as the $u$ variable. There were about 35% missing data on the $y$ variable, but the other variables were fully observed. Our analysis focuses on the overall average income in 2006, accounting for potentially nonrandom missingness. Indeed, missingness from self-reported income is well known to be informative in that both lower and higher ends of the income spectrum oftentimes do not report their income (Schenker et al., 2006).

Several estimators of the overall 2006 monthly income using age as instrument were generated alongside the complete case analysis, the naive analysis relying on single thresholds. Additionally, sensitivity analysis varying the instruments was also conducted. Results are reported in Table 5. The naive estimates are generally less precise than the proposed...
Table 3: Relative biases (RB), Mean square errors (MSE) for n=200 and 500; all RBs are multiplied by 100

| Estimator | Quantity | $n =$ | $M_1^*$ | $M_2^*$ | $M_3^*$ | $M_4^*$ |
|-----------|----------|-------|--------|--------|--------|--------|
| $\hat{\mu}$ | RB       | 200   | 0.4    | 0.3    | -0.3   | -0.1   | 0.6    | 0.5    | -0.3   | -0.2 |
|           | MSE      | 200   | 0.2    | 0.1    | 0.2    | 0.1    | 0.2    | 0.1    | 0.2    | 0.1  |
| $\tilde{\mu}$ | RB      | 200   | 0.8    | 0.2    | -2.7   | -2.0   | 0.8    | 0.6    | -0.5   | -0.4 |
|           | MSE      | 200   | 0.2    | 0.1    | 0.2    | 0.1    | 0.2    | 0.1    | 0.2    | 0.1  |
| $\bar{y}$ | MSE      | 200   | 0.0    | -0.1   | -0.1   | 0.0    | -0.2   | -0.2   | -0.2   | -0.2 |
| $\bar{y}_{obs}$ | MSE   | 200   | 0.1    | 0.1    | 0.1    | 0.1    | <0.1   | 0.1    | <0.1   |

Table 4: Standard Error (SE), Coverage Probability (CP) for n=200 and 500; all CPs are multiplied by 100

| Estimator | Quantity | $n =$ | $M_1^*$ | $M_2^*$ | $M_3^*$ | $M_4^*$ |
|-----------|----------|-------|--------|--------|--------|--------|
| $\hat{\mu}$ | SE       | 200   | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2  |
|           | CP       | 200   | 95.0   | 94.3   | 94.4   | 95.2   | 94.5   | 93.1   | 93.5   | 94.5 |
| $\tilde{\mu}$ | SE     | 200   | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2  |
|           | CP       | 200   | 90.8   | 90.2   | 86.7   | 87.4   | 90.1   | 91.1   | 91.0   | 94.0 |
| $\bar{y}$ | SE       | 200   | 0.2    | 0.2    | 0.2    | 0.2    | 0.2    | 0.2    | 0.2    | 0.2  |
|           | CP       | 200   | 92.5   | 92.8   | 92.9   | 93.0   | 92.6   | 92.5   | 92.6   | 92.5 |
| $\bar{y}_{obs}$ | SE | 200   | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2  |
|           | CP       | 200   | 44.0   | 8.4    | 83.0   | 59.7   | 36.1   | 3.4    | 91.6   | 88.1 |
| $\tilde{\mu}_{\gamma_0}$ | SE  | 200   | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2  |
|           | CP       | 200   | 91.4   | 93.4   | 93.4   | 92.3   | 92.9   | 93.5   | 91.9   | 92.5 |
| $\tilde{\mu}_{\gamma_0}/n$ | SE | 200   | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2  |
|           | CP       | 200   | 90.7   | 81.5   | 87.4   | 68.8   | 93.2   | 92.7   | 91.9   | 92.5 |
| $\tilde{\mu}_0$ | SE  | 200   | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2    | 0.3    | 0.2  |
|           | CP       | 200   | 90.6   | 81.5   | 87.3   | 68.6   | 93.2   | 92.7   | 91.9   | 92.5 |

scanning-based estimates, demonstrating the need to accumulate information across the instrument. Surprisingly, the estimate under the complete case analysis has a reduced standard error albeit being substantially larger than the other estimates.

6 Discussion

The proposed scanning approach is broadly applicable to instruments with an intrinsic order, including quantitative and ordinal categorical variables. Because it naturally embeds the underlying order into the estimation process, it typically will outperform methods that ignore such information in efficiency. An interesting feature of the methodology is that it can be combined with multiple instruments not requiring scanning (e.g., categorical instruments). It can also be applied to multiple instruments jointly requiring scanning, but the computational cost for such an undertaking is often
Table 5: Estimates and standard errors of average monthly income in 2006, treating Age as a continuous instrumental variable, and additional sensitivity analysis results for various instruments

| Method                          | Additional Instruments | Estimates (SE) |
|---------------------------------|------------------------|----------------|
| Complete case analysis          | no instrument          | 205.7 (2.8)    |
|                                 |                        | 189.7 (2.3)    |
| Scanning across age             | gender                 | 185.5 (3.2)    |
|                                 | education              | 185.9 (3.7)    |
|                                 | gender, education      | 184.7 (0.6)    |
|                                 |                        | 190.1 (5.1)    |
| Single age threshold (37)       | gender                 | 184.9 (3.1)    |
|                                 | education              | 183.9 (3.6)    |
|                                 | gender, education      | 184.7 (4.0)    |
|                                 |                        | 193.3 (11.5)   |
| Multiple age thresholds (32,44) | gender                 | 186.2 (4.6)    |
|                                 | education              | 186.9 (5.6)    |
|                                 |                        | 190.3 (9.3)    |
| 1st moment of age               | gender                 | 183.7 (8.4)    |
|                                 | education              | 187.3 (4.9)    |
|                                 | gender, education      | 185.5 (4.3)    |

An alternative approach that shares this feature to some extent is the moment-based estimator, which is only applicable to quantitative instruments.

Computationally, an advantage of dichotomization is that it eliminates the need for a second step generalized moment estimator of \( \gamma \). Obviously for a more complicated propensity model such as \( \pi(y, u) = [1 + \exp\{g(u) + \gamma_1y + \gamma_2y^2\}]^{-1} \), the use of dichotomization may not be helpful with identification. In which case, it may be necessarily to entertain continuous thresholding with multiple cutpoints to ensure identifiability of \( \gamma_1 \) and \( \gamma_2 \).

Because \( z \) serves as a missing data instrument for \( y \), it must meet the basic requirement of being related to \( y \), and being excluded from the propensity model. Consequently, the law of \( f(y \mid x) \) must be related to \( z \). As such, it may be of interest to evaluate the effect of \( y \) on \( z \). We may adopt, as a byproduct of the scanning approach, the quantity

\[
r = \left\{ \int_{\Xi^*} p_\xi(1 - p_\xi) d\xi \right\}^{-1} \int_{\Xi^*} p_\xi(1 - p_\xi) \{ E\{y \mid z \leq \xi\} - E\{y \mid z > \xi\} \} d\xi,
\]

as a nonparametric measure of the effect of instrument \( z \) on \( y \). Obviously, if \( y \) and \( z \) are non associated \( r = 0 \). Here \( E\{y \mid z \in A\} \) is the mean of \( y \) restricted to the population stratified by \( A \) of \( z \). To estimate \( E\{y \mid z \in A\} \) for a stratum \( A \) of \( z \), one may use \( \{n_A\}^{-1} \sum_{i=1}^n f\{z_i \in A\} \delta y_i / \{\pi(y_i, u_i; \hat{g}_\gamma, \hat{\gamma}_\xi), \xi \in \Xi^* \} \), with \( n_A = \sum_{i=1}^n I\{z_i \in A\} \). Thus for \( \xi^* \in \Xi^* \), estimators of \( E\{y \mid z \leq \xi^* \} \) and \( E\{y \mid z > \xi^* \} \) then follow.

The proposed method relies essentially on a readily available instrument, which is often unrealistic in practice. Some technical non-design based guidelines for the choice of viable instruments can be found elsewhere (e.g., Shao & Wang, 2016; and Zhao et al., 2021). Additionally, the proposed method uses the Kernel-based estimation of the unknown function \( g(\cdot) \), which may prove computationally difficult in the case of high-dimensional auxiliary variables, \( u \). In such settings, any sufficient dimension reduction techniques may be entertained (e.g., Cook and Li, 2002; Wang et al., 2020; and Zhao et al., 2021).

Acknowledgements

The authors are grateful to Professor Jun Shao for sharing the income panel data. This research is supported by NSF grant DMS-1916339 and NIH/NIDCR grant R03DE027108.
Appendix

Key regularity conditions

Condition 1 The space $\Gamma$ of $\gamma$ and $\Theta$ that of $\theta$ are all compact.

Condition 2 The kernel $K(u)$ has bounded derivatives of order $d$ and satisfies $\int K(u)du = 1$, and has zero moments of orders up to $m-1$ and nonzero $m$-th order.

Condition 3 The true function of $g(u)$ is continuously differentiable and bounded on an open set containing the support of $u$.

Condition 4 The moment $E\{\exp(4\gamma u)\}$ is finite and the function $E\{\exp(4\gamma u)\}f(u)$ is bounded, where $f(u)$ is the marginal density of $u$.

Condition 5 The threshold dependent bandwidth $h_{\xi,n}$ is such that as $n \to \infty$, $\sup_{E \in \Xi^*} n h_{\xi,n} \to 0$, $\inf_{E \in \Xi^*} n h_{\xi,n} \to \infty$, $\inf_{E \in \Xi^*} n^{1/2} h_{\xi,n}^{p_u+2d}/\log n \to \infty$ and $\sup_{E \in \Xi^*} n h_{\xi,n}^{2m} \to 0$, where $p_u$ is the dimension of $u$.

Condition 6 For any threshold $\xi \in \Xi$, there is a vector of the functional $G^\xi(y,u,\delta,\omega)$ which is linear in $\omega = (\omega_1, \omega_2)^T$ and such that:

1. for small enough $||\omega - \omega_0||, ||\tilde{m}^\xi(y,u,\delta,\omega,\gamma) - m^\xi(y,u,\delta,\omega,\gamma_0) - G^\xi(y,u,\delta,\omega_0)|| \leq b^\xi(y,u,\delta)||\omega - \omega_0||^2$, where $\tilde{m}^\xi(y,u,\delta,\omega,\gamma) = I(z \leq \xi)\frac{1}{\delta} (1 + \exp(\gamma))\omega_1/\omega_2$, $\omega_0 = [E(1 - \delta | u) - E(\delta \exp(\gamma y)| u)|u]T$ and $\sup_{E \in \Xi} E(b^\xi(y,u,\delta)) < \infty$;

2. $||G^\xi(y,u,\delta,\omega)|| \leq c^\xi(y,u,\delta) ||w||$ and $\sup_{E \in \Xi} E(c^\xi(y,u,\delta)^2) \leq \infty$;

3. There exists an almost everywhere continuous function $\nu^\xi(u)$ with $\sup_{E \in \Xi} \int ||\nu^\xi(u)|| du \leq \infty$, $E(G^\xi(y,u,\delta,\omega)) = \int \nu^\xi(u)\omega(u)du$, for all $||\omega|| < \infty$ and $\sup_{E \in \Xi} E\{\sup_{||\xi|| \leq \epsilon} ||\nu^\xi(u + \xi)||^4\} < \infty$, for some $\epsilon > 0$.

Condition 7 For small enough $||\omega - \omega_0||$, $\tilde{m}^\xi(y,u,\delta,\omega,\gamma)$ is continuously differentiable in $\gamma$ for any threshold $\xi \in \Xi$, in a neighbourhood of $\gamma_0$, and there is $k^\xi(y,u,\delta)$ with $\sup_{E \in \Xi} E\{k^\xi(y,u,\delta)\} < \infty$ such that $||\nabla_\gamma \tilde{m}^\xi(y,u,\delta,\omega,\gamma) - \nabla_\gamma m^\xi(y,u,\delta,\omega,\gamma_0)|| \leq k^\xi(y,u,\delta)(||\gamma - \gamma_0||^r + ||\omega - \omega_0||^r)$, for any $r > 0$, $\Gamma^\xi = E\{\nabla_\gamma \tilde{m}^\xi(y,u,\delta,\omega,\gamma_0)\}$ exists and of full rank.

Condition 8 Assume that $V(\hat{\delta}_\xi)$ the variance of $\hat{\delta}_\xi$ is uniformly positive definite over $\xi \in \Xi^*$; that is $\inf_{E \in \Xi^*} \text{eigmin}V(\hat{\delta}_\xi) > 0$, where eigmin(.) denotes the smallest eigenvalue.

A sufficient condition for the previous assumption to hold is that the limiting point of $V(\hat{\delta}_\xi)$ is uniformly positive definite over $(\xi, \theta) \in \Xi^* \times \Theta$; that is $\inf_{E \in \Xi^* \times \Theta} \text{eigmin}(V(\hat{\delta}_\theta)) > 0$, where eigmin(.) denotes the smallest eigenvalue.

References

[1] G. Molenberghs, M. G. Kenward, and E. Lesaffre. The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84(1):33–44, 03 1997. ISSN 0006-3444. doi:10.1093/biomet/84.1.33 [https://doi.org/10.1093/biomet/84.1.33](https://doi.org/10.1093/biomet/84.1.33)

[2] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi:10.1080/01621459.1994.10476818 [https://doi.org/10.1080/01621459.1994.10476818](https://doi.org/10.1080/01621459.1994.10476818)

[3] Anastasios A. Tsiatis and Marie Davidian. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, 22(4):569–573, 11 2007.

[4] Donald B. Rubin. *Multiple Imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics. Wiley, New York, NY, 1987. ISBN 978-0-471-08705-2.

[5] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002. ISBN 0-471-18386-5.
[6] Jae Kwang Kim and Jun Shao. *Statistical methods for handling incomplete data*. CRC press, 2021.

[7] Antonio R Linero. Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness. *Biometrika*, 104(2):327–341, 2017.

[8] S Yang, L Wang, and P Ding. Causal inference with confounders missing not at random. *Biometrika*, 106(4):875–888, 09 2019. ISSN 0006-3444. doi:10.1093/biomet/asz048 URL https://doi.org/10.1093/biomet/asz048

[9] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models (C/R: P1121-1146). *Journal of the American Statistical Association*, 94: 1096–1120, 1999.

[10] Andrea Rotnitzky, Daniel Scharfstein, Ting-Li Su, and James Robins. Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, 57(1):103–113, 2001.

[11] Jiwei Zhao and Jun Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512):1577–1590, 2015.

[12] David Todem, Jason P. Fine, and Limin Peng. A global sensitivity test for evaluating statistical hypotheses with nonidentifiable models. *Biometrics*, 66:558–566, 2010.

[13] Margaret C. Wu and Raymond J. Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175–188, 1988. ISSN 0006341X, 15410420.

[14] Andrea B. Troxel, Stuart R. Lipsitz, and David P. Harrington. Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika*, 85:661–672, 1998b.

[15] Sheng Wang, Jun Shao, and Jae Kwang Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, pages 1097–1116, 2014.

[16] Jun Shao and Lei Wang. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187, 01 2016. ISSN 0006-3444. doi:10.1093/biomet/asv071 URL https://doi.org/10.1093/biomet/asv071

[17] Eric J. Tchetgen Tchetgen and Kathleen E. Wirth. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73(4):1123–1131, 2017.

[18] Jae Kwang Kim and Cindy Long Yu. A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 101(483):924–933, 2006.

[19] Douglas G. Altman. Categorizing continuous variables. In *Encyclopedia of Biostatistics*, P. Armitage and T. Colton (eds), pages 563–567. Chichester, U.K.: Wiley, 1998. ISBN 9780470011812.

[20] L Peng and JP Fine. Nonparametric tests for continuous covariate effects with multistate survival data. *Biometrics*, 64(4):1080–1089, 2008.

[21] Patrick Royston, Douglas G. Altman, and Willi Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, 2006. doi:https://doi.org/10.1002/sim.2331 URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2331

[22] Vern T Farewell, Brian D. M Tom, and Patrick Royston. The impact of dichotomization on the efficiency of testing for an interaction effect in exponential family models. *Journal of the American Statistical Association*, 99(467):822–831, 2004. doi:10.1198/016214504000001169 URL https://doi.org/10.1198/016214504000001169

[23] Shakeeb Khan and James L Powell. Two-step estimation of semiparametric censored regression models. *Journal of Econometrics*, 103(1):73–110, 2001. ISSN 0304-4076. doi:https://doi.org/10.1016/S0304-4076(01)00040-9 URL https://www.sciencedirect.com/science/article/pii/S0304407601000409

[24] R. M. Dudley. Central Limit Theorems for Empirical Measures. *The Annals of Probability*, 6(6):899 – 929, 1978. doi:10.1214/aop/1176995384 URL https://doi.org/10.1214/aop/1176995384

[25] Theo Gasser and Hans-Georg Müller. Kernel estimation of regression functions. In Th. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, pages 23–68, Berlin, Heidelberg, 1979. Springer Berlin Heidelberg. ISBN 978-3-540-38475-5.

[26] Nathaniel Schenker, Trivellore E Raghunathan, Pei-Lu Chiu, Diane M Makuc, Guangyu Zhang, and Alan J Cohen. Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association*, 101(475):924–933, 2006.
[27] Puying Zhao, Lei Wang, and Jun Shao. Sufficient dimension reduction and instrument search for data with nonignorable nonresponse. *Bernoulli*, 27(2):930–945, 2021.

[28] R.Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455 – 474, 2002. doi:10.1214/aos/1021379861 URL https://doi.org/10.1214/aos/1021379861
Supplemental Material: Nonparametric scanning for nonrandom missing data with continuous instrumental variables

Proofs Of The Technical Results

Proof of Theorem 1
By Theorem 1 of Shao and Wang (2016), we have the pointwise consistency of \( \hat{\gamma}_\xi \) to \( \gamma_0\xi \). So we only need to show the uniform convergence of the process \( \hat{\gamma}_\xi \) in \( \xi \). For that we first recall the following definition and relevant theorem of stochastic equicontinuity from Newey & McFadden (1994). A sequence of random function \( Q_n(\theta) \) is stochastic equicontinuous if

\[
\forall \varepsilon > 0, \lim_{\delta \to 0} \lim_{n \to \infty} P \left[ \sup_{|\theta - \theta'| < \delta} |Q_n(\theta) - Q_n(\theta')| > \varepsilon \right] = 0
\]

And then by Lemma 2.8, Newey & McFadden (1994), for compact \( \Theta \) and continuous \( Q_0(\theta) \), \( \sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| \overset{P}{\to} 0 \) if and only if \( Q_n(\theta) \overset{P}{\to} Q_0(\theta) \forall \theta \in \Theta \) and \( Q_n(\theta) \) is stochastic equicontinuous.

Next, we consider \( \hat{Q}_n(\xi) = |\hat{\gamma}_\xi - \gamma_0\xi| \). Here the threshold variable \( \xi \in \Xi \) and to prove the uniform convergence we only consider a dense subset and the rational part \( \Xi^R \) of this compact set \( \Xi \). And we define, \( \Xi^R = \{ \xi, \xi' \in \Xi^R \text{ such that } |\xi - \xi'| < \delta \} \). Now,

\[
\lim_{\delta \to 0} \lim_{n \to \infty} P \left[ \sup_{|\xi - \xi'| < \delta} |\hat{\gamma}_\xi - \gamma_0\xi| - |\hat{\gamma}_{\xi'} - \gamma_0\xi'| > \varepsilon \right]
\]

\[
\leq \lim_{\delta \to 0} \lim_{n \to \infty} P \left[ \sup_{(\xi, \xi') \in \Xi^R} (|\hat{\gamma}_\xi - \gamma_0\xi| + |\hat{\gamma}_{\xi'} - \gamma_0\xi'|) > \varepsilon \right]
\]

\[
\leq \lim_{\delta \to 0} \lim_{n \to \infty} \sum_{\xi, \xi' \in \Xi^R} P \left[ |\hat{\gamma}_\xi - \gamma_0\xi| + |\hat{\gamma}_{\xi'} - \gamma_0\xi'| > \varepsilon \right]
\]

Here by applying dominated convergence theorem, we can take the \( \lim_{\delta \to 0} \) inside the sum and then the pointwise convergence of \( \hat{\gamma}_\xi \) to \( \gamma_0\xi \) gives the required stochastic continuity of \( \hat{Q}_n(\xi) \).

Therefore, we establish the uniform consistency of \( \hat{Q}_n(\xi) \), i.e. \( \sup_{\xi \in \Xi} |\hat{\gamma}_\xi - \gamma_0\xi| \overset{P}{\to} 0 \).

Proof of Theorem 3
Here, for the simplicity of the proof we consider \( \theta = \mu = E(Y) \). Following Dudley (1978), we will prove this in two steps: i) we will show that any finite dimensional distribution of the stochastic process \( n^{1/2}(\hat{\mu}_\xi - \mu) \) converges to a multivariate normal distribution; and ii) we will show its stochastic equicontinuity to prove the tightness of the asymptotic Gaussian process. Now, we prove the first part using the Cramér–Wold theorem, i.e. for any finite \( d \in \{1, 2, \ldots\}, t_1 X_{\xi_1} + t_2 X_{\xi_2} + \cdots + t_d X_{\xi_d} \) is convergent in distribution to \( t_1 X_{\xi_1} + t_2 X_{\xi_2} + \cdots + t_d X_{\xi_d} \), where \( X_{\xi_b} = n^{1/2}(\hat{\mu}_{\xi_b} - \mu) \) and \( (X_{\xi_1}, X_{\xi_2}, \ldots, X_{\xi_d}) \) follow some multivariate Gaussian distribution.

We consider the following expansion as in Shao and Wang (2016) of identically and independently distributed random variables with zero mean and finite variance.

For \( t_1, t_2, \ldots, t_d \in \mathbb{R}^d \),

\[
t_1 n^{1/2}(\hat{\mu}_{\xi_1} - \mu) + t_2 n^{1/2}(\hat{\mu}_{\xi_2} - \mu) + \cdots + t_d n^{1/2}(\hat{\mu}_{\xi_d} - \mu)
\]

\[
= n^{1/2} \sum_{i=1}^d \sum_{j=1}^d t_j \left( \frac{\delta(y_i - u_j)}{\pi(y_i, u_j; \gamma_0\xi)} + (v_i - \mu) + \zeta_i \right) + o_p(1) \overset{CLT}{\to} \sum_{j=1}^d t_j R_{\xi_j}
\]

where \( v_i = E(y_i / u_j) \) and \( R_{\xi_j} \sim N \left( 0, var \left\{ \frac{\delta(y-v)}{\pi(y, u_j; \gamma_0\xi)} + (v - \mu) + \zeta_i \right\} \right) \) for \( j \in \{1, 2, \ldots, d\} \).

Hence by Cramér–Wold theorem, any finite dimensional distribution of \( n^{1/2}(\hat{\mu}_\xi - \mu) \) is converging to a multivariate normal distribution.

Now, to establish the stochastic equicontinuity, we need to show
Then by applying CLT T we get the asymptotic normality of our proposed estimator.

Let denote, \( T_n(\delta) = \sup_{|\xi - \xi'| < \delta} |n^{1/2}(\hat{\mu}_\xi - \mu) - n^{1/2}(\hat{\mu}_{\xi'} - \mu)|. \)

If we can show that the limits in the stochastic equicontinuity conditions are interchangeable, then for any fixed \( n \) and for all \( \delta < \delta_0 \), where \( \delta_0 = \min\{t - t', t, t' \in \{z_{i(-1)}, z_{i(i)}\}, i = 2, 3, \ldots, n\} \), \( \hat{\mu}_\xi = \hat{\mu}_{\xi'} \) with probability 1 where \( |\xi - \xi'| < \delta \) and hence \( T_n(\delta) = 0 \). Hence the stochastic equicontinuity holds and the limiting Gaussian process is tight.

So we only need to show that the two limits are indeed interchangeable. By the Moore-Osgood Theorem on exchanging limits, the sufficient condition for that is: a) \( \lim_{\delta \to 0} T_n(\delta) = 0 \) uniformly on \( \delta \in \Delta, \Delta \subset R \) and b) \( \lim_{\delta \to 0} T_n(\delta) \) converges pointwise on \( n \in \mathbb{N} \). While (b) holds by the above mentioned argument with the specific choice of \( \delta_0 \) conditioned on the instrument \( z \); (a) can be proved by applying the continuous mapping theorem with the previously mentioned convergence of any finite dimensional distribution of \( n^{1/2}(\hat{\mu}_\xi - \mu) \) to multivariate Gaussian distribution.

**Proof of Theorem**

Under the regularity condition where bootstraping gives a valid consistent variance estimator, we consider \( W(\xi) = \frac{1}{\mathbb{V}(\hat{\mu}_\xi)} \xrightarrow{a.s.} \frac{1}{\mathbb{V}(\mu)} = \rho(\xi) \) for any fixed \( \xi \). Now,

\[
 n^{1/2}(\hat{\mu} - \mu) = n^{1/2} \frac{\mathbb{E}[\hat{\mu}(\xi)w(\xi)d\xi]}{\mathbb{E}[w(\xi)d\xi]} = n^{1/2} \sum_{i=1}^{\infty} \frac{\mathbb{E}\left[ \frac{\delta_i(y_i - \mu_i)}{\sigma_0^2} + (v_i - \mu) + \xi_i \right] w(\xi)d\xi}{\mathbb{E}[w(\xi)d\xi]}
\]

Here we assume that \( E\left[ \mathbb{E}\left[ \frac{\delta_i(y_i - \mu_i)}{\sigma_0^2} + (v_i - \mu) + \xi_i \right] w(\xi)d\xi \right]^2 < \infty \)

Then by applying CLT we get the asymptotic normality of our proposed estimator.

**Additional Simulations - Using moments of the instrumental variable in estimating equation**

For continuous instruments, it was suggested in Shao and Wang (2016) to use moment-based estimating equations as an alternative approach to discretizing the instrument; for example replacing the indicator function in the estimating equations by any \( k \)th moment of \( z \) and using up to \( L \) estimating equations, \( \ell = 1, 2, \ldots, L \). The exact specification of this approach is given by Equation (8) in the main paper. A real difficulty in practice is to determine the optimal number of moments for the estimating equation without knowing the true data generating mechanism. This is critical as the performance of the moment-based approach may depend on the number of moments and the underlying structure of the data. In general, setting \( L \) equal to the dimension of the tilting parameter \( \gamma \) works reasonably well in practice. But for some other missing data settings, this method may not work well regardless of the moments of the instrument being used. Using a simulation study, we show that for the zero inflated instrumental variable, the moment-based approach does not work well in contrast to our proposed scanning method.

Here we consider the following incomplete data structure:

**Full data generating mechanism**

\( z = \Delta W_1 + (1 - \Delta)W_2, \) where \( \Delta \sim \text{Bernoulli}(0.7), W_1 = \delta_{(0)}, W_2 \sim N(0, 0.1) \)

\( u|\{z\} \sim N(z - 5, 1) \) and \( y|\{z, u\} \sim N(u^2 + 0.1z, 1) \)

Averaging with an MC approximation gives \( \mu_{\text{MC}} = 26 \) for \( E(y) \) the unconditional mean of \( y \).

**Missing data generating mechanism** We have considered the following propensity models, keeping the unconditional rate of missingness at 30%:

1. \( M_1^*: \pi_i = 1/[1 + \exp(\alpha + \beta u_i + \gamma y_i)] \) with \( \alpha, \beta, \gamma = (9.8, -0.5, -0.6); \)
2. \( M_2^*: \pi_i = 1/[1 + \exp(\alpha + \beta \exp(u_i) + \gamma y_i)] \) with \( \alpha, \beta, \gamma = (11.8, -0.5, -0.6); \)

In Tables 6 and 7 along with those of \( \mu \) our proposed estimator, \( \hat{y}_n, \hat{y}_{\text{obs}}, \hat{\mu}_{\text{MC}}, \hat{\mu}_{\text{MC}/n} \), and \( \hat{\mu}_0 \) as described in the main paper, the finite sample performance of the following three estimators of \( \mu \) are compared:

1. \( \hat{\mu}_1 \), the estimator using the 1st moment of \( z \) in the estimating equation, \( L = 1; \)
2. \( \hat{\mu}_{1,2} \), the estimator using the 1st two moments of \( z \) in the estimating equation, \( L = 2; \)
3. \( \hat{\mu}_{1,2,3} \), the estimator using the 1st three moments of \( z \) in the estimating equation, \( L = 3. \)
Table 6: Relative biases (RB), Mean square errors (MSE) for n=200 and 500; all RBs are multiplied by 100

| Estimator | Quantity | $n = 200$ | $n = 500$ | $n = 200$ | $n = 500$ |
|-----------|----------|-----------|-----------|-----------|-----------|
| $\hat{\mu}$ | RB       | -2.9      | -2.5      | -3.3      | -3.0      |
|           | MSE      | 2.2       | 1.3       | 2.4       | 1.5       |
| $\hat{\mu}_1$ | RB       | -11.5     | -10.7     | -11.3     | -11.3     |
|           | MSE      | 12.5      | 10.6      | 12.2      | 11.0      |
| $\hat{\mu}_{1,2}$ | RB       | -8.9      | -6.7      | 8.8       | -6.9      |
|           | MSE      | 10.1      | 6.6       | 9.8       | 6.8       |
| $\hat{\mu}_{1,2,3}$ | RB       | -9.2      | -6.8      | -9.5      | -7.2      |
|           | MSE      | 10.7      | 6.9       | 10.9      | 7.0       |
| $\tilde{\mu}_1$ | RB       | -0.1      | 0.0       | 0.0       | 0.0       |
|           | MSE      | 0.8       | 0.3       | 0.8       | 0.3       |
| $\tilde{y}$ | RB       | 16.9      | 17.0      | 16.9      | 16.9      |
|           | MSE      | 20.3      | 19.9      | 20.2      | 19.8      |
| $\hat{\mu}_{\gamma_0}$ | RB       | -4.2      | -3.4      | -3.9      | -3.3      |
|           | MSE      | 2.9       | 1.7       | 2.7       | 1.7       |
| $\hat{\mu}_{\gamma_0}/n$ | RB       | -6.1      | -4.8      | -6.2      | -5.2      |
|           | MSE      | 4.2       | 2.5       | 4.2       | 2.7       |
| $\bar{y}$ | RB       | -6.1      | -4.8      | -6.2      | -5.2      |
|           | MSE      | 4.2       | 2.5       | 4.3       | 2.7       |

Results show that the moment-based approach clearly fails here, contrarily to the scanning approach which works comparatively better for both propensity models.

Table 7: Standard Error (SE), Coverage Probability (CP) for n=200 and 500; all CPs are multiplied by 100

| Estimator | Quantity | $n = 200$ | $n = 500$ | $n = 200$ | $n = 500$ |
|-----------|----------|-----------|-----------|-----------|-----------|
| $\hat{\mu}$ | SE       | 0.7       | 0.4       | 0.7       | 0.5       |
|           | CP       | 64.0      | 52.3      | 63.4      | 45.1      |
| $\hat{\mu}_1$ | SE       | 1.4       | 1.3       | 1.3       | 1.3       |
|           | CP       | 44.9      | 43.3      | 37.6      | 33.1      |
| $\hat{\mu}_{1,2}$ | SE       | 1.4       | 1.3       | 1.3       | 1.2       |
|           | CP       | 57.5      | 72.8      | 51.4      | 64.6      |
| $\hat{\mu}_{1,2,3}$ | SE       | 1.3       | 1.2       | 1.3       | 1.2       |
|           | CP       | 49.6      | 66.7      | 41.6      | 61.1      |
| $\tilde{y}$ | SE       | 0.5       | 0.3       | 0.5       | 0.3       |
|           | CP       | 85.1      | 84.3      | 84.3      | 84.8      |
| $\tilde{y}_{obs}$ | SE       | 0.7       | 0.5       | 0.7       | 0.4       |
|           | CP       | 0.9       | 0.0       | 0.0       | 0.0       |
| $\bar{y}_{\gamma_0}$ | SE       | 0.7       | 0.5       | 0.7       | 0.5       |
|           | CP       | 52.3      | 37.6      | 54.9      | 41.8      |
| $\hat{\mu}_{\gamma_0}/n$ | SE       | 0.6       | 0.4       | 0.7       | 0.4       |
|           | CP       | 33.4      | 19.6      | 31.2      | 16.6      |
| $\bar{\mu}_0$ | SE       | 0.6       | 0.4       | 0.6       | 0.4       |
|           | CP       | 32.9      | 19.6      | 30.5      | 16.6      |
References For The Supplementary Material

[1] Shao, J., Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. Biometrika, 103(1), 175-187.

[2] Newey, W. K., McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of econometrics, 4, 2111-2245.

[3] Dudley, R. M. (1978). Central limit theorems for empirical measures. The Annals of Probability, 899-929.