Efficient estimation in sub and full populations with monotonically missing at random data*

Saraswata Chaudhuri†

Updated version: August 30, 2017

Abstract

We consider estimation of a parameter defined by moment restrictions on a target population characterized by the missingness pattern of monotonically missing at random data. Attrition or dropout in a, say, R-period study/survey typically generates such data. In this case, a generic target is the underlying population of the sample units dropping out in contiguous periods $a, \ldots, b$ where $a \leq b \in \{1, \ldots, R\}$. The semiparametric efficiency bound and the efficient influence function are obtained for the parameter of interest that nests the well-known special cases with targets $(a = 1, b = R)$ or $(R = 2, a = b = 1)$. The consideration of a generic target results in new insights on the usability and contribution of certain sample units toward efficient estimation.

Keywords: Attrition; Dropouts; Multi-phase sampling; Pattern-mixtures; Semiparametric efficiency; Generalized method of moments.

*I am very grateful to Daniel Farewell and Erica Moodie for helpful comments and discussions.

†Department of Economics, McGill University; and Cireq, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

1 Introduction

Estimation based on monotonically missing data has received special attention in the missing data literature; see, e.g., the textbooks by Little and Rubin (2002), Tsiatis (2006), etc. Since the pioneering work of Robins et al. (1994, 1995), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Holcroft et al. (1997), etc., efficient estimation in such cases has generally been considered under the missing at random, i.e., a selection on observables, assumption. The focus of efficient estimation has conventionally been on parameters defined by the joint distribution of the underlying variables in the full population, i.e., the population from which the full set of sample units were randomly drawn.

However, selection on observables leads to systematic differences in the sample units differing in terms of the variables observed in them. Consider, for example, attrition or dropout in a R-period study/survey where the decision to leave at the end of any period depends on the variables observed until the end of that period. Consider the two groups of sample units who leave at the end of the a-th and b-th periods ($1 \leq a < b \leq R$) respectively. The variables that should have been observed in periods $a+1, \ldots, b$ are missing for the former group.[1] Then the underlying sub-populations, call them a and b, for these two groups are generally different. (These are sub-populations of what we call the full population above.) It is interesting, at least for descriptive purposes [see Little (1993)] or sometimes even more substantively [see Diggle et al. (2007)], to study such sub-populations.

In this paper, we consider efficient estimation of parameters defined by the joint distribution of the variables in unions of contiguous sub-populations, namely, $[a, \ldots, b]$ where $a \leq b \in \{1, \ldots, R\}$. This includes the full population, i.e., when $a = 1, b = R$, and the individual sub-populations, i.e., when $a = b$, as special cases. We obtain the efficient influence functions and the efficiency bounds for all such cases under a unified framework, and closely analyze them. This is the contribution of our paper.

Our result provides new insights on the usability of the sample units toward efficient estimation. Roughly speaking, in the context of the attrition example above, we find that if interest lies in the sub-populations for the units who leave in periods $[a, \ldots, b]$, then the units who left before period a are not usable. We show that this happens because we do not: (i) allow any dimension-reduction in our selection on observables assumption, or (ii) assume that the conditional probabilities of missingness are known. This is the cost of working under a framework that is more realistic for analyzing cases of attrition. By contrast, Chaudhuri (2017) finds that all sample units are usable for estimation in all sub-populations if the conditional probabilities for missingness satisfy either an extreme dimension-reduction assumption pinned to the baseline period 1 (except in $a = 2, b = R$), or are assumed known.

[1] As in Diggle et al. (2007), we do not differentiate between attrition (subsequent outcomes unobserved) and dropouts (subsequent outcomes, but not the intended ones, observed), and treat both as attrition [c.f. Heckman et al. (1998)].

Of course, both dimension-reduction and known conditional probabilities are immaterial in the conventional case where the full population is of interest, i.e., when $a = 1, b = R$. However, they turn out to be of major importance when interest lies in the sub-populations, as is the case in our paper.

An efficient and doubly-robust estimator is readily obtained by using an estimating function based on the efficient influence function. It directly follows from Holcroft et al. (1997), Tsiatis (2006), etc. that this estimator has standard asymptotic properties resembling that of the typical doubly-robust estimators. If the unknown nuisance parameters in the estimating function are estimated nonparametrically, then the standard asymptotic properties of this estimator follow from, e.g., Chen et al. (2003). These results are well-known and hence we simply refer to them without further details.

We frequently refer to attrition/dropout since it is a common cause of monotonically missing data and has been studied extensively under the selection on observables assumption [see, e.g., Fitzgerald et al. (1996), Nicoletti (2006), etc. even in economics]. Our result, however, applies more broadly.

For example, consider the pattern mixture models of Glynn et al. (1986), Little (1993, 1994), etc. Thanks to Theorem 1 of Molenberghs et al. (1998), the selection on observables, i.e., the missing at random assumption maintained in our paper for the monotonically missing data is equivalent to the available case missing value assumption maintained for pattern mixture models. However, under either similar or different selection mechanisms, estimation in pattern mixture models are typically not carried out by the doubly-robust augmented inverse probability weighting estimators that directly follow from our result [see, e.g., Chapter 16, Molenberghs and Kenward (2007); Chapter 15, Little and Rubin (2002)]. Hence, our result on semiparametric efficiency may prove useful to this literature.

Our paper proceeds as follows. Section 2 establishes and discusses the main result. Section 3 is a Monte Carlo experiment studying the efficient estimator via simulations. Section 4 concludes. Appendix A contains expository endnotes. Appendix B contains the proof of our main result.

2 Efficiency bound and efficient estimation

2.1 Framework:

Let $Z := (Z_1', \ldots, Z_R')'$ where Z_r is a $d_r \times 1$ random vector and $\sum_{r=1}^{R} d_r$ is finite. Following Tsiatis (2006), let C be a random variable with support $\mathcal{C} := \{1, \ldots, R\}$ and $T_C(Z)$ a transformation defined as $T_r(Z) := (Z_1', \ldots, Z_r')'$ with dimension $(\sum_{s=1}^{r} d_s) \times 1$ for $r = 1, \ldots, R$. For example, in the context of attrition/dropout in a R-period study/survey, Z_r are the variables specific to period r, while $T_r(Z)$ are all the variables observed for a unit that leaves at the end of period r, i.e., with $C = r$, for $r = 1, \ldots, R$. Formally, let $O := (C, T_C'(Z))'$ denote what is observed for a unit in the sample.

We maintain a general selection on observables, i.e., a missing at random (MAR), assumption:

$$P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r|T_r(Z)) \text{ for } r = 1, \ldots, R. \tag{1}$$

This is the MAR assumption [see e.g. Robins and Rotnitzky (1995), Tsiatis (2006)] in the sense of Rubin (1976). It is important to recognize that (1) implies that for any $r = 2, \ldots, R$:

$$P(C \geq r|Z) = 1 - \sum_{j=1}^{r-1} P(C = j|T_j(Z)) = 1 - \sum_{j=1}^{r-1} P(C = j|T_{r-1}(Z)) = P(C \geq r|T_{r-1}(Z)) \tag{2}$$

only depends on $T_{r-1}(Z)$. In the context of attrition, this means that the decision to not leave at the end of the $(r-1)$-th period is independent of the thus far unobserved variables $Z_r, \ldots, Z_R$, once conditioned on the variables that have been already observed, i.e., $T_{r-1}(Z)$.

Under (1), we consider sub-populations $[a, \ldots, b]$, equivalently, $(a \leq C \leq b)$, for $a, b \in \{1, \ldots, R\}$. If $a = b = r$ then, in the context of attrition, this is the hypothetical sub-population from which the units who left at the end of period r can be viewed as being randomly drawn. If $a < b$, then this is the sub-population for the units who left in the periods $a, a+1, \ldots, b$. Convention: those who stay until the end, leave at the end of period R. Thus, if $a = 1$ and $b = R$, then this is the full population.

The underlying distributions of Z, denote them by $F_{Z|(a \leq C \leq b)}(z)$, in these sub-populations are typically different, and we define the parameter of interest in (3) as a finite dimensional feature of $F_{Z|(a \leq C \leq b)}(z)$ as follows. Consider a function $m(Z; \beta) : \text{Support}(Z) \times \mathcal{B} \mapsto \mathbb{R}^{d_m}$, $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ where $d_\beta \leq d_m$. For a given $a, b \in \{1, \ldots, R\}$ $(a \leq b)$, let the parameter value of interest β^0 be defined as:

$$E[m(Z; \beta)|a \leq C \leq b] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0. \tag{3}$$

Our goal is the efficient estimation of β^0. To define efficient we first obtain the efficiency bound for β^0, which is our primary contribution. For this, we additionally maintain the following assumptions.

Assumption A:

(A1) The observed sample units $\{O_i := (C_i, T'_{C_i}(Z_i))'\}_{i=1}^n$ are i.i.d. copies of $O := (C, T'_C(Z))'$.

(A2) $(P(C = r|T_R(Z)))_{r=1}^{R-1} > 0$ and $P(C = R|T_R(Z)) > \underline{p}$ almost surely in $T_R(Z)$ for a fixed $\underline{p} \in (0, 1)$.

(A3) $M := \frac{\partial}{\partial \beta'} E\left[m(Z; \beta^0)|a \leq C \leq b\right]$ is a $d_m \times d_\beta$ finite matrix of full column rank.

A1 rules out dependence and heterogeneity across sample units once viewed as random draws from O. A2 is necessary to ensure that the efficiency bound is finite [see Khan and Tamer (2010)]. A3 allows for $m(Z; \beta)$ to be non-differentiable in β, but does require that its expectation be differentiable.

2.2 Efficiency bound for estimation of β^0:

Proposition 1 is our main result. It provides the efficiency bound for β^0 along with the efficient influence function. The key quantity (and the novel contribution) for this purpose is defined as follows:

$$\begin{aligned}\varphi(O;\beta) \;:=\; & \sum_{r=b+1}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)} \left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) \\ & + \sum_{r=a+1}^{b} \frac{I(C\geq r)}{P(C\geq r|T_r)} \frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)} \left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) \\ & + \sum_{r=a}^{b} \frac{I(C=r)}{P(a\leq C\leq b)} E[m(T_R;\beta)|T_r]. \end{aligned} \tag{4}$$

In (4) and onward we use the following notation, unless confusing. $I(C\geq R)\equiv I(C=R)$, $T_r\equiv T_r(Z)$ (note that $T_R\equiv Z$), and $m(Z;\beta)\equiv E[m(T_R;\beta)|T_R]$. If $b=R$, then the indices (e.g., in summations) running from $b+1$ to R are void. If $a=b$ then the similar indices running from $a+1$ to b are void, while those running from a to b contain only one term and it corresponds to a (equivalently b).

Proposition 1 *Let* (1), (3) *and assumption A hold. Let the* $d_m\times d_m$ *matrix* $V:=Var(\varphi(O;\beta^0))$ *be finite and positive definite where* β^0 *and* $\varphi(O;\beta)$ *are as defined in* (3) *and* (4) *respectively. Then the asymptotic variance lower bound for* $\sqrt{n}(\widehat{\beta}-\beta^0)$ *of any regular estimator* $\widehat{\beta}$ *for* β^0 *is given by* $\Omega:=(M'V^{-1}M)^{-1}$. *An estimator* $\widehat{\beta}$ *whose asymptotic variance equals* Ω *has the asymptotically linear representation:*

$$\sqrt{n}(\widehat{\beta}-\beta^0)=-\Omega M'V^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(O_i;\beta^0)+o_p(1). \;\blacksquare$$

Remark 1: The following special cases of this general result are already well-known in the literature. (i) Taking $R=2$ and $a=b=1$ gives the result of Case (1) in Theorem 1 of Chen et al. (2008):

$$\varphi(O;\beta)=\frac{I(C=2)}{P(C=2|T_1)}\frac{P(C=1|T_1)}{P(C=1)}(m(T_2;\beta)-E[m(T_2;\beta)|T_1])+\frac{I(C=1)}{P(C=1)}E[m(T_2;\beta)|T_1].$$

(ii) $R=2$, $a=1$ and $b=2$ give Case (2) in Theorem 1 of Chen et al. (2008) [see Robins et al. (1994)]:

$$\varphi(O;\beta)=\frac{I(C=2)}{P(C=2|T_1)}(m(T_2;\beta)-E[m(T_2;\beta)|T_1])+E[m(T_2;\beta)|T_1].$$

(iii) On the other hand, taking a general R and setting $a=1, b=R$ give [see Appendix A.1]:

$$\varphi(O;\beta)=\sum_{r=1}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right)+E[m(T_R;\beta)|T_1],$$

the well-known result for the full population under monotone missing at random data as in Robins and Rotnitzky (1995); Rotnitzky and Robins (1995), Holcroft et al. (1997). Also see Tsiatis (2006).

Remark 2: If we focus on an individual sub-population, i.e., $a = b$, then for a general R, $\varphi(O;\beta)$ is:

$$\sum_{r=a+1}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r)} \frac{P(C = a|T_a)}{P(C = a)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) + \frac{I(C = a)}{P(C = a)} E[m(T_R;\beta)|T_a]$$

by Proposition 1. Thus, $\varphi(O;\beta) = I(C = R)m(T_R;\beta)/P(C = R)$ in the special case when $a = b = R$.

Remark 3: The terms involving $E[m(Z;\beta)|T_r]$ for $r < a$ do not appear in the expression for $\varphi(O;\beta)$ in (4). This is because, implicit in (1) and, hence, in Proposition 1 is a restriction on dimension-reduction: We do not allow $P(C = r|Z) = P(C = r|T_s)$ for any $s < r = 2, \ldots, R-1$, which would be additional information that is generally unrealistic in cases of attrition. Roughly speaking, we do not let the more recent past become irrelevant to the selection process conditional on the less recent past.

To contrast, consider an extreme dimension-reduction assumption: $P(C = r|Z) = P(C = r|T_1)$ that pins the selection to period 1, i.e., the baseline. (With $R = 3$, this reflects a common identification assumption in mediation analysis; see, e.g., Imai et al. (2010).) Under this, Proposition 9 in Chaudhuri (2017) establishes that the corresponding $\varphi(O;\beta)$ is (writing $E[m(Z;\beta)|T_r]$ as $m^\dagger(T_r;\beta)$ for brevity):

$$\varphi^\dagger_{[a,b]}(O;\beta) = \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r)} \frac{P(a \leq C \leq b|T_1)}{P(a \leq C \leq b)} (m^\dagger(T_r;\beta) - m^\dagger(T_{r-1};\beta)) + \frac{I(a \leq C \leq b)}{P(a \leq C \leq b)} m^\dagger(T_1;\beta). \quad (5)$$

Hence, under this dimension-reduction, all $E[m(Z;\beta)|T_r]$ $(r = 1, \ldots, R)$ are usable irrespective of a, b except if $a = 2, b = R$. It is also noteworthy that the terms $E[m(Z;\beta)|T_r]$ $(r = a, \ldots, b)$, i.e., those corresponding to the sub-populations of interest, contribute differently in this dimension-reduction case than in (the last two lines of) (4) unless $a = 1, b = R$, i.e., unless interest lies in the full population. Thus, while restrictive, Proposition 9 in Chaudhuri (2017) is not a special case of our Proposition 1.

Remark 4: Remark 3 is in direct contrast to the case of planned missingness in the data. To see the contrast, let MAR in (1) hold, i.e., $P(C = r|Z) = P(C = r|T_r)$ for $r = 1, \ldots, R$. In this context, by planned missingness we mean that $P(C = r|T_r)$ is known for $r = 1, \ldots, R$. Under this, Proposition 1 in Chaudhuri (2017) establishes that the corresponding $\varphi(O;\beta)$ is (writing $E\left[\frac{P(a \leq C \leq b|T_R)}{P(a \leq C \leq b)} m(T_R;\beta) \middle| T_r\right]$ as $m^*(T_r;\beta)$ for brevity):

$$\varphi^*_{[a,b]}(O;\beta) = \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r)} (m^*(T_r;\beta) - m^*(T_{r-1};\beta)) + m^*(T_1;\beta). \quad (6)$$

Thus, under planned missingness, nesting of the conditions $\{P(C = r|Z) = P(C = r|T_s) : s \leq r\}$ is automatic in the sense that all $E[m(Z;\beta)|T_r]$ $(r = 1, \ldots, R)$ are always usable irrespective of whether

dimension-reduction is allowed or not. As a result, while Chaudhuri (2017) also contrasts unplanned (with $P(C = r|Z) = P(C = r|T_1)$) and planned missingness, the contrast in our paper is more prominent because we maintain $P(C = r|Z) = P(C = r|T_r)$ without allowing for dimension-reduction.

Remark 5: Now consider a similarity with Chaudhuri (2017). It follows from (5) and (6) that:

$$\varphi^{\dagger}_{[a,b]}(O;\beta) = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} \varphi^{\dagger}_{[j,j]}(O;\beta) \quad \text{and} \quad \varphi^{*}_{[a,b]}(O;\beta) = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} \varphi^{*}_{[j,j]}(O;\beta),$$

which reflect the standard relationship among the moment restrictions in (3) that:

$$E[m(Z;\beta)|a \leq C \leq b] = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} E[m(Z;\beta)|C=j].$$

It turns out that the same also holds in our paper, i.e., with the $\varphi(O;\beta)$ defined in (4) [see Appendix A.2]. This is intuitive, and it presents a way of combining the efficient estimators for the individual sub-populations to obtain the efficient estimator for unions of such (contiguous) sub-populations.

Remark 6: $\varphi(O;\beta)$ in (4) is doubly-robust in the sense of Scharfstein et al. (1999) [see Appendix A.3]. This may lead to desirable properties for the estimator of β^0 based on the estimating function $\varphi(O;\beta)$ when the unknown nuisance functions, i.e., the conditional probabilities and expectations, are estimated parametrically or nonparametrically [see Robins et al. (1994), Robins and Ritov (1997), Bang and Robins (2005), Tan (2007), Cao et al. (2009), Rothe and Firpo (2016), etc.].

2.3 Efficient estimator of β, and estimation of its asymptotic variance:

Now we obtain the efficient generalized method of moments (GMM) estimator of β^0 and describe its asymptotic properties by appealing to existing results. Since these are not our contribution, we only provide an informal description by citing, for brevity, a rather incomplete set of relevant references.

Informed by Proposition 1, we treat $\varphi(O;\beta)$ as the moment vector. $\varphi(O;\beta)$ contains the unknown nuisance parameters — (i) the conditional probabilities $P(a \leq C \leq r|T_r) = \sum_{j=a}^{r} P(C=j|T_j)$ (by (1)) for $r = a, \ldots, b-1$, and $P(C \geq r|T_r)$ for $r = a+1, \ldots, R$, and (ii) the conditional expectations $E[m(T_R;\beta)|T_r]$ for $r = a, \ldots, R-1$ — and they need to be replaced by either parametric or nonparametric estimators for the GMM estimation of β^0. (The unknown scalar multiple $1/P(a \leq C \leq b)$ in $\varphi(O;\beta)$ can be replaced by an estimator $n/\sum_{i=1}^{n} I(a \leq C_i \leq b)$ without affecting anything else.)

For notational brevity, denote generically the parameters in (i) and (ii) by $p(.)$ and $q(.;\beta)$ respectively, and their estimators by $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ respectively. Note by inspecting (4) that estimation of the individual elements of $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are only required when their arguments, i.e., the conditioning

variables in the definition of the concerned individual elements of $p(.)$ and $q(.;\beta)$, are observed. Define

$$g(O_i, p(.), q(.;\beta);\beta) := \varphi(O_i;\beta), \quad \text{and} \quad \bar{g}_n(\beta, p(.), q(.;\beta)) = \frac{1}{n}\sum_{i=1}^{n} g(O_i, p(.), q(.;\beta);\beta).$$

Given $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$, and a $d_m \times d_m$ weighting matrix W_n, the GMM estimator $\widehat{\beta}_n(W_n)$ is defined as:

$$\widehat{\beta}_n(W_n) \approx \arg\min_{\beta\in\mathcal{B}} \ \bar{g}_n(\beta, \widehat{p}(.), \widehat{q}(.;\beta))' \ W_n \ \bar{g}_n(\beta, \widehat{p}(.), \widehat{q}(.;\beta)).$$

Straightforward extension of Holcroft et al. (1997), Tsiatis (2006), etc. gives sufficient conditions for consistency and asymptotic normality of $\widehat{\beta}_n(W_n)$ if $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are parametrically estimated. Remark 6 indicates that $\widehat{\beta}_n(W_n) \xrightarrow{P} \beta^0$ if the parametric model for either $p(.)$ or $q(.;\beta)$ is correctly specified. $\widehat{\beta}_n(W_n)$ is asymptotically efficient with asymptotic variance Ω, as in Proposition 1, if both parametric models are correctly specified and if the weighting matrix is efficient, i.e., if $W_n \xrightarrow{P} V^{-1}$.

When both these parametric models are correctly specified, the asymptotic variance of $\widehat{\beta}_n(W_n)$ can be estimated in the standard way (adjusting for whether W_n is efficient) by estimating: (i) M using possibly numerical derivatives (e.g., perturbation methods); and (ii) V by averaging the outer product of $g(O_i, \widehat{p}(.), \widehat{q}(.;\widehat{\beta}_n(W_n)); \widehat{\beta}_n(W_n))$ while ignoring that $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are actually estimated.

Otherwise, however, Ω is not the correct asymptotic variance. In particular, for (ii), one should take the estimation of $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ into account (as in Theorem 6.1 of Newey and McFadden (1994)), and instead use the routine modification described in Section 6.3 of Newey and McFadden (1994). Section B of the supplementary material for Cao et al. (2009) provides a comprehensive discussion of robust-sandwich-variance estimation meant specifically for similar scenarios [also see Tsiatis (2006)].

On the other hand, if nonparametric estimators $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are used then consistency of $\widehat{\beta}_n(W_n)$ (for β^0) and its asymptotic normality follow under the conditions of Theorems 1 and 2 of Chen et al. (2003). Chen et al. (2003) also provide the sufficient conditions under which the estimation of $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ has no effect on the asymptotic variance of $\widehat{\beta}_n(W_n)$. Rothe and Firpo (2016) established these results for the case of $a = 1, b = R = 2$ under weaker conditions by exploiting the double-robustness noted in Remark 6 [also see Proposition 2.3 in Chaudhuri and Guilkey (2016)]. $W_n \xrightarrow{P} V^{-1}$ gives efficiency of $\widehat{\beta}_n(W_n)$. Conditions similar to Theorems 6 and 7 of Cattaneo (2010) give consistency of the asymptotic variance estimator obtained based on averaging the outer product of $g(O_i, \widehat{p}(.), \widehat{q}(.;\widehat{\beta}_n(W_n)); \widehat{\beta}_n(W_n))$ and using, if needed, a numerical derivative-based estimate of M.

Formal statements and proofs for the above asymptotic properties are available from the author, but are not provided here for brevity since these are for long well-understood results in the literature.

3 Monte Carlo Experiment

We consider a setup reflecting the decision to stay or leave dynamically over periods in the context of programs (smoking cessation, weight loss), school, employment, experiments, surveys, market (if the unit is a firm), contracts, etc. (The selection mechanism below in (8) would reverse in some cases.)

The design and the data generating process (DGP) considered are as follows. Let, for $t = 1, \ldots, T$:

$$Y_t = \frac{1}{2}Y_{t-1} + \frac{1}{4}Y_{t-2} + \frac{1}{4}X_t + e_t, \quad \text{where} \quad X_t = X_{t-1} + v_t. \tag{7}$$

e_t and v_t are the model errors. Take X_0, Y_{-1}, Y_0 independently $N(1,1)$ as the initial state.

Define $R := T$. For $r = 1, \ldots, R$, let Y_r be the outcome from staying until the end of the r-th period in the program, and X_r the other variables for the r-th period.

Let the individual's expectation for the outcome in the r-th period be Y_r^*. Suppose that the individual decides to leave the program at the end of the r-th period, conditional on staying until then, provided that the actual outcome exceeds the expectation, i.e., $Y_r^* < Y_r$. More precisely, let

$$I(C = r) = I(Y_r^* < Y_r) \prod_{j=1}^{r-1} I(Y_j^* \geq Y_j) \text{ for } r = 1, \ldots, R-1, \text{ while } I(C = R) = 1 - \sum_{r=1}^{R-1} I(C = r). \tag{8}$$

We take $R = 3$. As usual, the analyst observes C but not Y_r^*. In terms of our notation for observability, this means: $Z_1 = (Y_{-1}, Y_0, Y_1, X_{-1}, X_0, X_1)'$, $Z_2 = (Y_2, X_2)'$ and $Z_3 = (Y_3, X_3)'$.

We assume that e_t and v_t are i.i.d. $N(0,1)$ for all t, while $u_r := Y_r^* - Y_r$ is i.i.d. $N(0, (2.5)^2)$ for all r.[2] (1) is imposed by maintaining that $e_t, v_t, u_r, X_0, Y_{-1}, Y_0$ are mutually independent for all t, r. This results in roughly 50% of the individuals with $C = 1$, 26% with $C = 2$, and 24% with $C = 3$.

To define β^0, we take the moment function in (3) as $m(Z; \beta) = Y_3 - \beta$, and consider the six different target sub-populations characterized by $(1 \leq C \leq 3)$, $(C = 1)$, $(C = 2)$, $(C = 3)$, $(1 \leq C \leq 2)$ and $(2 \leq C \leq 3)$ respectively, giving six different parameters of interest. The first parameter is the average period 3 outcome for all the units in the study; the second one is the average period 3 outcome for those who left after the first period had they continued until the end of period 3; and so on for the other parameters. The last one is the average period 3 outcome for those who left at the end of period 2 or 3 had those who left after period 2 continued until the end of period 3. We compute the "true value" of these parameters numerically by generating data from the above DGP with sample size 10 million, estimating the mean of Y_3 for each sub-population, and then averaging each mean over 10,000

[2]The larger variance for u_r relative to e_t (or v_t) is maintained to avoid the adverse consequences of the limited overlap problem under which the efficiency bound is infinite [recall assumption A2]. This effort was to an extent successful because we encountered estimated $P(C = 3|T_3)$ that is very close to zero only when $n = 100$. It happened in 12 out of 10,000 Monte Carlo trials. We ignore these trials in the sequel instead of taking routes as in, e.g., Cao et al. (2009).

Monte Carlo trials. Accordingly, the six different "true values", i.e., β^0's are: 1, 1.1709, .9617, .6858, 1.0994 and .8291 respectively. As evident from Table 1, the error in the above approximation is of a rather small order to seriously affect our subsequent analysis that is conducted with far smaller size.

The study of the efficient estimation under this setup is now conducted for sample sizes $n = 100, 200$ and 500 (also $n = 1000$ in Table 2). All results are reported below based on 10,000 Monte Carlo trials.

Target Population for β	Descriptive Statistics					
	Mean	Std $= 10^{-3}\times$	Median	IQR	Min	Max
$1 \leq C \leq 3$	1	.4860	1	.0007	.9982	1.0017
$C = 1$	1.1709	.6841	1.1709	.0009	1.1682	1.1735
$C = 2$	.9617	.9430	.9617	.0013	.9581	.9648
$C = 3$	.6858	.9769	.6858	.0013	.6817	.6895
$1 \leq C \leq 2$	1.0994	.5536	1.0994	.0007	1.0975	1.1012
$2 \leq C \leq 3$	.8291	.6800	.8291	.0009	.8265	.8316

Table 1: The "true" parameter value β^0 is approximated (column 2) for different target populations (column 1) based on averaging over 10,000 Monte Carlo trials the target-sample means obtained by using the same DGP and with sample size 10 million. Columns 3-7 list the standard deviation (Std), interquartile range (IQR), minimum (Min) and maximum (Max) of the estimator.

GMM estimation is conducted following Section 2.2 but ignoring the weighting matrix W_n since $d_m = d_\beta = 1$. The nuisance parameters, i.e., the conditional expectations (of Y_3) and the conditional probabilities are estimated by least squares and probit regressions respectively. For both, we specify the regression function as linear in the conditioning variables and do not include interactions.

The parameter β for the sub-population $C = 3$ is not interesting for our purpose since the complete-case GMM estimator $\sum_{i=1}^n I(C_i = 3)Y_{3i}/\sum_{j=1}^n I(C_j = 3)$ already estimates it consistently and, by virtue of Proposition 1 (see Remark 2). Table 2 briefly summarizes the performance of the complete-case GMM estimator, and this is poor and misleading for all the targets except $C = 3$.

Target Population for β	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
	Std = .3140		Std = .2207		Std = .1386		Std = .0994	
	Bias	Size	Bias	Size	Bias	Size	Bias	Size
$1 \leq C \leq 3$	-.3148	16.8	-.3151	29.8	-.3155	62.5	-.3148	88.7
$C = 1$	-.4857	33.5	-.4860	59.1	-.4864	93.8	-.4857	99.8
$C = 2$	-.2765	14.1	-.2768	24.1	-.2772	51.9	-.2765	79.7
$C = 3$	-.0006	5.3	-.0009	5.2	-.0013	5.2	-.0006	4.8
$1 \leq C \leq 2$	-.4142	25.7	-.4145	46.8	-.4149	84.9	-.4142	98.7
$2 \leq C \leq 3$	-.1439	7.7	-.1442	9.9	-.1446	18.1	-.1439	30.8

Table 2: All results for the complete-case estimator are reported based on 10,000 Monte Carlo trials. Bias stands for the mean bias. The only representative population for the complete-case estimator is $C = 3$. Std stands for the Monte Carlo standard deviation, and since the estimator is identical for all sub-populations (and hence the bias for the other targets), there is only a single Std reported for each sample size n. Size stands for the empirical size of the asymptotic 5% two-sided t-test.

For the other five parameters, the results for the efficient GMM estimator from (2.3) is presented in Table 3. For each β and for each of the 10,000 Monte Carlo trials we obtain the estimate, subtract it from the corresponding "true value", and report the mean, median, minimum and maximum of this difference (as Bias, MedBias, MinBias, MaxBias respectively) over the 10,000 trials. The mean of the absolute value of this difference is reported as ABias, the Monte Carlo standard deviation as Std, the mean of the standard deviation based on the asymptotic variance formula as AStd, and the empirical size of the t-test based on Std and the asymptotic variance formula as Size and ASize respectively. The efficient GMM estimator performs quite well in all aspects even for relatively smaller sample sizes.

Target for β	$n = 100$								
	Bias	MedBias	MinBias	MaxBias	ABias	Std	AStd	Size	ASize
$1 \leq C \leq 3$	-.0037	-.0102	-1.4862	2.5896	.2544	.3254	.2512	5.3	13.8
$C = 1$	-.0041	-.0149	-2.6509	3.3278	.3177	.4092	.3336	5.1	11.8
$C = 2$	-.0068	-.0132	-2.4474	2.2725	.3484	.4501	.3986	5.3	8.2
$1 \leq C \leq 2$	-.0056	-.0147	-1.9473	2.8004	.2896	.3727	.2893	5.1	13.4
$2 \leq C \leq 3$	-.0034	-.0050	-1.6282	1.4262	.2492	.3187	.2803	5.1	8.5
	$n = 200$								
	Bias	MedBias	MinBias	MaxBias	ABias	Std	AStd	Size	ASize
$1 \leq C \leq 3$	-.0005	-.0042	-.9553	1.0449	.1650	.2078	.1884	5.1	8.1
$C = 1$	-.0009	-.0077	-1.7499	1.6389	.2015	.2551	.2436	5.0	6.4
$C = 2$	.0006	-.0021	-1.7145	3.1583	.2211	.2832	.2816	4.9	4.9
$1 \leq C \leq 2$	-.0009	-.0065	-1.2391	1.3064	.1836	.2321	.2139	4.9	7.5
$2 \leq C \leq 3$	.0001	-.0013	-.9836	1.7208	.1659	.2101	.2045	5.0	5.7
	$n = 500$								
	Bias	MedBias	MinBias	MaxBias	ABias	Std	AStd	Size	ASize
$1 \leq C \leq 3$	-.0020	-.0039	-.6830	.4939	.1009	.1269	.1208	5.0	6.7
$C = 1$	-.0024	-.0045	-.5654	.6081	.1225	.1540	.1546	5.0	5.3
$C = 2$	-.0015	-.0026	-1.6263	.7290	.1313	.1668	.1753	5.1	4.1
$1 \leq C \leq 2$	-.0023	-.0040	-.8461	.5669	.1103	.1392	.1358	5.2	6.0
$2 \leq C \leq 3$	-.0014	-.0016	-.9141	.4623	.1024	.1288	.1299	4.8	4.8

Table 3: All results for the efficient GMM estimator are reported based on 10,000 Monte Carlo trials.

To put the performance of the efficient GMM estimator into context, although it might not be fair, we report in Table 4 the same (except AStd and Asize) for the Horvitz and Thompson (1952)-type inverse probability weighting (IPW) estimator:

$$\sum_{i=1}^{n} \frac{I(C_i = 3)}{\widehat{P}(C = 3|T_{3i})} \frac{\widehat{P}(a \leq C_i \leq b|T_{bi})}{\sum_{j=1}^{n} I(a \leq C_j \leq b)} Y_{3i}.$$

Our specifications for the conditional probabilities nests the truth. Hence, this estimator is also consistent and asymptotically unbiased by virtue of (1). This is reflected by the Bias and the shrinking variability of the estimator as evident from MedBias, ABians and Std. Its efficiency loss with respect to the efficient GMM estimator is evident by comparing the corresponding Std, and this is expected

from Proposition 1. The loss can vary from 8.5% to 50% (with the median and average case both being 27%) across the numbers reported under Std in Tables 3 and 4. Indeed, a comparison of the other dispersion-related characteristics clearly demonstrates that the efficient GMM estimator is much less dispersed than the IPW estimator in all aspects and for all sample sizes considered in our experiment.

Target for β	$n = 100$						
	Bias	MedBias	MinBias	MaxBias	ABias	Std	Size
$1 \leq C \leq 3$	-.0449	-.0823	-1.5733	7.9008	.3080	.4278	3.7
$C = 1$	-.0593	-.1386	-2.2333	16.3040	.4159	.6143	3.2
$C = 2$	-.0579	-.1168	-4.5318	6.3902	.3987	.5611	4.1
$1 \leq C \leq 2$	-.0589	-.1184	-2.0413	10.6195	.3653	.5187	3.5
$2 \leq C \leq 3$	-.0305	-.0469	-1.9653	3.8259	.2743	.3678	4.2
	$n = 200$						
	Bias	MedBias	MinBias	MaxBias	ABias	Std	Size
$1 \leq C \leq 3$	-.0143	-.0338	-1.2238	3.4379	.2006	.2701	4.0
$C = 1$	-.0209	-.0606	-1.1614	4.6535	.2697	.3698	4.1
$C = 2$	-.0141	-.0421	-2.0402	10.4536	.2611	.3839	3.4
$1 \leq C \leq 2$	-.0189	-.0505	-1.4550	4.4394	.2345	.3223	3.8
$2 \leq C \leq 3$	-.0075	-.0174	-1.2727	6.0802	.1840	.2527	3.6
	$n = 500$						
	Bias	MedBias	MinBias	MaxBias	ABias	Std	Size
$1 \leq C \leq 3$	-.0044	-.0088	-.5559	1.0035	.1192	.1527	4.6
$C = 1$	-.0060	-.0213	-.6293	1.3960	.1610	.2084	4.5
$C = 2$	-.0041	-.0180	-.9816	2.4238	.1542	.2016	4.7
$1 \leq C \leq 2$	-.0056	-.0152	-.6295	1.2874	.1377	.1779	4.7
$2 \leq C \leq 3$	-.0027	-.0061	-.5498	1.3350	.1116	.1417	4.8

Table 4: All results for the IPW estimator are reported based on 10,000 Monte Carlo trials.

4 Conclusion

The simulation results provided an overall encouraging picture for the performance of the efficient GMM estimator. This estimator falls under the class of the well-known doubly-robust augmented inverse probability weighting estimators for which the asymptotic theory is already well-documented. In other words, this estimator does not pose any new technical challenge to the practitioner.

On the other hand, our main result, i.e., Proposition 1, showed that this estimator is actually semiparametrically efficient for a wide variety of target populations characterized by the missingness of monotonically missing at random data. The discussion in Section 2 also documented that these targets nest those for which similar estimators have long been proposed and also employed in practice. Therefore, when the conventional likelihood-based methods are unsuitable or demand strong assumptions, we hope that our result and the above demonstration would encourage this alternative method of efficient GMM estimation for any target population that falls under the premise of our paper.

References

Bang, H. and Robins, J. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61: 962–972.

Cao, W., Tsiatis, A., and Davidian, M. (2009). Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. *Biometrika*, 96: 723–734.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155: 138–154.

Chaudhuri, S. (2017). A Note on Efficiency Gains from Multiple Incomplete Subsamples. Mimeo.

Chaudhuri, S. and Guilkey, D. K. (2016). GMM with Multiple Missing Variables. *Journal of Applied Econometrics*, 31: 678–706.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *Annals of Statistics*, 36: 808–843.

Chen, X., Linton, O., and van Keilegom, I. (2003). Estimation of Semiparametric Models when the Criteria Function is not Smooth. *Econometrica*, 71: 1591–1608.

Diggle, P., Farewell, D., and Henderson, R. (2007). Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *JRSS, Series C*, 56: 499–550.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1996). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. In *Working Paper Series*. NBER.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). *Selecion modeling versus mixture modeling with nonignorable nonresponses*, pages 115–142. Springer-Verlag, NY.

Heckman, J., Smith, J., and Taber, C. (1998). Accounting for Dropouts in Evaluations of Social Programs. *Review of Economics and Statistics*, LXXX: 1–14.

Holcroft, C., Rotnitzky, A., and Robins, J. M. (1997). Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference*, 65: 349–374.

Horvitz, D. and Thompson, D. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of American Statistical Association*, 47: 663–685.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25: 51–71.

Khan, S. and Tamer, E. (2010). Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica*, 78: 2021–2042.

Little, R, J. A. and Rubin, D. D. (2002). *Statistical Analysis with Missing Data*. Wiley - Interscience.

Little, R. J. A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association*, 88: 125–134.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81: 471–483.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. John Wiley and Sons.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, 52: 153–161.

Newey, W. K. and McFadden, D. L. (1994). Large Sample Estimation and Hypothesis Testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2212–2245. Elsevier Science Publisher.

Nicoletti, C. (2006). Nonresponse in dynamic panel data models. *Journal of Econometrics*, 132: 461–489.

Robins, J. and Ritov, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theroy for Semi-Parametric Models. *Statistics in Medicine*, 16: 285–319.

Robins, J. and Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of American Statistical Association*, 90: 122–129.

Robins, M., Rotnitzky, A., and Zhao, L. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of American Statistical Association*, 427: 846–866.

Robins, M., Rotnitzky, A., and Zhao, L. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of American Statistical Association*, 429: 106–121.

Rothe, C. and Firpo, S. (2016). Semiparametric Estimation and Inference Using doubly-Robust Moment Conditions. Mimeo.

Rotnitzky, A. and Robins, J. (1995). Semiparametric Regression Estimation in the Presence of Dependent Censoring. *Biometrika*, 82: 805–820.

Rubin, D. (1976). Inference and Missing Data. *Biometrika*, 63: 581–592.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94: 1096–1146.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22: 560–568.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Appendix A: Expository endnotes for Section 2

A.1 The expression of $\varphi(O;\beta)$ in (4) when $a=1, b=R$ [see Remark 1]:

$$
\begin{aligned}
\varphi(O;\beta) &= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} P(C\leq r-1|T_{r-1})\left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad + \sum_{r=1}^{R} I(C=r)E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} P(C\geq r|T_{r-1})\left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad + \sum_{r=1}^{R} I(C=r)E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \sum_{r=2}^{R} I(C\geq r)\left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) \text{ [by using (2)]} \\
&\quad + \sum_{r=1}^{R} I(C=r)E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \left\{\sum_{r=2}^{R} I(C=r)E[m(T_R;\beta)|T_r] - I(C\geq 2)E[m(T_R;\beta)|T_1]\right\} + \sum_{r=1}^{R} I(C=r)E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r]-E[m(T_R;\beta)|T_{r-1}]\right) + E[m(T_R;\beta)|T_1]. \ \blacksquare
\end{aligned}
$$

A.2 The weighted average of sub-population $\varphi(O;\beta)$'s as discussed in Remark 5:

Write $E[m(T_R;\beta)|T_r]$ as q_r for all r for brevity. Denote the expression for $\varphi(O;\beta)$ when $a=b=j$ (see Remark 2) by $\varphi_{[j,j]}(O;\beta)$. Then the weighted average with weights $\omega_j := P(C=j)/P(a\leq C\leq b)$ is:

$$\begin{aligned}\sum_{j=a}^{b}\omega_j\varphi_{[j,j]}(O;\beta) &= \sum_{j=a}^{b}\left\{\sum_{r=j+1}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}(q_r-q_{r-1})+\frac{I(C=j)}{P(a\leq C\leq b)}q_j\right\}\\ &= \sum_{r=b+1}^{R}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}(q_r-q_{r-1})\sum_{r=a}^{b}\frac{I(C=r)}{P(a\leq C\leq b)}q_r\\ &\quad+\sum_{j=a}^{b}\sum_{r=j+1}^{b}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}(q_r-q_{r-1})\end{aligned}$$

where the first term follows by (1). Matching the first two terms with the terms on lines one and three of (4), the demonstration will be complete if the third term is equal to the term on the second line of (4). This follows by interchanging the order of the summation (which is allowed here) and noting that:

$$\begin{aligned}\sum_{j=a}^{b}\sum_{r=j+1}^{b}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}(q_r-q_{r-1}) &= \sum_{r=a+1}^{b}\sum_{j=a}^{r-1}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}(q_r-q_{r-1})\\ &= \sum_{r=a+1}^{b}\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}(q_r-q_{r-1}),\end{aligned}$$

where the last line follows by (1). The final expression is equal to the term on the second line of (4). ■

A.3 Double-robustness of the expression of $\varphi(O;\beta)$ in (4) [see Remark 6]:

To see this, first replace the unknown $P(a\leq C\leq r|T_r)$ for $r=a+1,\ldots,b$ and $1/P(C\geq r|T_r)$ for $r=a+1,\ldots,R$ in (4) by any integrable functions of T_r, and then note that $E[\varphi(O;\beta^0)]=0$ by (3) (applied to the last line of (4)). On the other hand, rearranging the terms in (4) gives $\varphi(O;\beta)$ alternatively as:

$$\begin{aligned}&\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left[\frac{I(C\geq r)}{P(C\geq r|T_r)}m(T_R;\beta)+\sum_{r=b+1}^{R-1}\left(\frac{I(C\geq r)}{P(C\geq r|T_r)}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1})}\right)E[m(T_R;\beta)|T_r]\right]\\ &+\sum_{r=a}^{b}\left\{\left(\frac{I(C\geq r)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1})}\frac{P(a\leq C\leq r|T_r)}{P(a\leq C\leq b)}\right)\right.\\ &\qquad\left.+\frac{I(C=r)}{P(a\leq C\leq b)}\right\}E[m(T_R;\beta)|T_r]\end{aligned}$$

where $\{a\leq C\leq a-1\}$ is a null event. Now replacing the unknown $E[m(T_R;\beta)|T_r]$ by any $d_m\times 1$ integrable functions of T_r for $r=1,\ldots,R-1$, it follows by (1) and (3) that $E[\varphi(O;\beta^0)]=0$. ■

Appendix B: Proof of the main result

Notation: f and F denote the density and distribution functions, and the concerned random variables are specified inside parentheses. Their conditional counterparts are denoted accordingly using the obvious notation. $L_0^2(F)$ denotes the space of mean-zero, square integrable functions with respect to F.

Proof of Proposition 1: We follow the three steps in Chen et al. (2008). Step 1 characterizes the tangent set for all regular parametric sub-models satisfying the semiparametric assumptions on the observed data. Step 2 obtains the efficient influence function for a given rotation of $m(Z;\beta)$. Step 3 obtains the optimal rotation and, thereby, gives the efficiency bound as the expectation of the outer product of the efficient influence function.

STEP - 1: Consider a regular parametric sub-model indexed by a parameter θ for the distribution of the observed data $O = (C', T_C'(Z))'$. The log of the distribution, in terms of $(C, Z')'$, is:

$$\log f_\theta(O) = \log f_\theta(Z_1) + \sum_{r=2}^{R} I(C \geq r) \log f_\theta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \log P_\theta(C = r|Z_1, \ldots, Z_r).$$

Let θ_0 be the unique value of θ such that $f_{\theta_0}(O)$ equals the true $f(O)$, and accordingly for all the quantities. The score function with respect to θ can be written in terms of $(C, Z')'$ as:

$$S_\theta(O) = s_\theta(Z_1) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \frac{\dot{P}_\theta(C = r|Z_1, \ldots, Z_r)}{P_\theta(C = r|Z_1, \ldots, Z_r)}$$

where $s_\theta(Z_1) := \frac{\partial}{\partial\theta} \log f_\theta(Z_1)$, $s_\theta(Z_r|Z_1, \ldots, Z_{r-1}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_r|Z_1, \ldots, Z_{r-1})$ for $r = 2, \ldots, R$, and $\dot{P}_\theta(C = r|Z_1, \ldots, Z_r) := \frac{\partial}{\partial\theta} P_\theta(C = r|Z_1, \ldots, Z_r)$ for $r = 1, \ldots, R$. For Step-2, it is useful to note from (1) and (2) that for any $r = 2, \ldots, R$:

$$\dot{P}_\theta(C \geq r|Z) = -\dot{P}_\theta(C \leq r - 1|Z_1, \ldots, Z_{r-1}) = \dot{P}_\theta(C \geq r|Z_1, \ldots, Z_{r-1}). \tag{9}$$

The tangent set is the mean square closure of all d_β dimensional linear combinations of $S_\theta(O)$ for all such smooth parametric sub-models, and it takes the form:

$$\mathcal{T} := \nu_1(Z_1) + \sum_{r=2}^{R} I(C \geq r) \nu_r(Z_1, \ldots, Z_r) + \sum_{r=1}^{R} I(C = r) \omega_r(Z_1, \ldots, Z_r), \tag{10}$$

where $\nu_1(Z_1) \in L_0^2(F(Z_1))$ and $\nu_r(Z_1, \ldots, Z_r) \in L_0^2(F(Z_r|Z_1, \ldots, Z_{r-1}))$ for $r = 2, \ldots, R$, and $\omega_r(Z_1, \ldots, Z_r)$ is any square integrable function of $Z_1, \ldots, Z_r$ for $r = 1, \ldots, R$.

STEP - 2: For brevity we write $m(Z;\beta^0)$ as m, and drop the subscript θ from all quantities evaluated at θ^0. The moment conditions in (3) for a given a, b are equivalent to the requirement that for any $d_\beta \times d_m$ matrix A, the following just-identified system of moment conditions holds:

$$AE[m|a \leq C \leq b] = AE\left[\frac{P(a \leq C \leq b|Z)}{P(a \leq C \leq b)}\frac{I(C=R)}{P(C=R|Z)}m\right] = 0.$$

where the first equality follows from (1). Differentiating with respect to θ under the integral, we obtain

$$0 = AM\frac{\partial\beta^0(\theta_0)}{\partial\theta'} + AE\left[m\left\{s(Z)' + \frac{\dot{P}(a \leq C \leq b|Z)'}{P(a \leq C \leq b|Z)} - \frac{\dot{P}(a \leq C \leq b)'}{P(a \leq C \leq b)}\right\}\middle| a \leq C \leq b\right]$$

where $s(Z) := s(Z_1 + \sum_{r=2}^{R} s(Z_r|Z_1,\ldots,Z_{r-1})$ and $\dot{P}(a \leq C \leq b) := \frac{\partial}{\partial\theta}P_{\theta^0}(a \leq C \leq b)$. Taking a full row rank A along with (1), (3) and assumption (A3) gives:

$$\frac{\partial\beta^0(\theta_0)}{\partial\theta'} = -(AM)^{-1}A\left\{E\left[ms(Z)'|a \leq C \leq b\right] + \sum_{r=a}^{b} E\left[m\frac{\dot{P}(C=r|Z_1,\ldots,Z_r)'}{P(a \leq C \leq b)}\right]\right\}.$$

Now we establish that for the given A, $-(AM)^{-1}A\varphi(O;\beta^0)$ is the efficient influence function by showing that $E[-(AM)^{-1}A\varphi(O;\beta^0)S(O)'] = \frac{\partial\beta^0(\theta_0)}{\partial\theta'}$ and that $(AM)^{-1}A\varphi(O;\beta^0) \in \mathcal{T}$ defined in (10).

For this purpose, note by using (4) (and switching to the notation T_r for $(Z_1,\ldots,Z_r)$ when it helps brevity) that we can write $E[\varphi(O;\beta^0)S(O)'] = \sum_{i=1}^{3}\sum_{j=1}^{2} B_{ij}$ where:

$$\begin{aligned}
B_{11} &:= \sum_{r=b+1}^{R} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)D'\right],\\
B_{12} &:= \sum_{r=b+1}^{R} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
B_{21} &:= \sum_{r=a+1}^{b} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)D'\right],\\
B_{22} &:= \sum_{r=a+1}^{b} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
B_{31} &:= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a \leq C \leq b)}E[m|T_r]D'\right],\\
B_{32} &:= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a \leq C \leq b)}E[m|T_r]\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
D &:= s(Z_1) + \sum_{k=2}^{R} I(C \geq k)s(Z_k|T_{k-1}).
\end{aligned}$$

As noted above Proposition 1, we proceed with the understanding that if $b = R$ then $B_{11} = B_{12} = 0$,

and if $a = b$ then $B_{21} = B_{22} = 0$. Also, for notational brevity define T_0 as any constant, so that $s(Z_1) \equiv s(Z_1|T_0)$. First, note that:

$$\begin{aligned}
B_{11} &= \sum_{r=b+1}^{R}\sum_{k=1}^{r} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&\quad + \sum_{r=b+1}^{R}\sum_{k=r+1}^{R} E\left[\frac{I(C \geq k)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&= \sum_{r=b+1}^{R}\sum_{k=1}^{r} E\left[\frac{P(C \geq r|T_{r-1})}{P(C \geq r|T_{r-1})}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&\quad + \sum_{r=b+1}^{R}\sum_{k=r+1}^{R} E\left[\frac{P(C \geq k|T_{k-1})}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&= \sum_{r=b+1}^{R} E\left[\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}E[m|T_r]s(Z_r|T_{r-1})'\right] + 0 \\
&= E\left[\frac{I(a \leq C \leq b)}{P(a \leq C \leq b)}ms(Z_R, \ldots, Z_{b+1}|T_b)'\right] = E\left[ms(Z_R, \ldots, Z_{b+1}|T_b)'|a \leq C \leq b\right] \qquad (11)
\end{aligned}$$

where the third and fourth lines follow by (2), the fifth line follows by noting that for all $k = 1, \ldots, r-1$: $E\left[(E[m|T_r] - E[m|T_{r-1}])s(Z_k|T_{k-1})'\right] = E\left[E[(E[m|T_r] - E[m|T_{r-1}])s(Z_k|T_{k-1})'|T_{r-1}]\right] = 0$ while for $k \geq r+1$: $E\left[E[m|T_r]s(Z_k|T_{k-1})'\right] = E\left[E[m|T_r]E[s(Z_k|T_{k-1})'|T_{k-1}]\right] = 0$, and the sixth (last) line follows by (1) and the definition of score. Second, it now follows that:

$$B_{21} = \sum_{r=a+1}^{b} E\left[\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}E[m|T_{r-1}]s(Z_r|T_{r-1})'\right]$$

exactly following the steps that led to the fifth line in the expression for B_{11} in (11) above. Therefore,

$$\begin{aligned}
B_{21} &= \sum_{r=a+1}^{b}\sum_{k=a}^{r-1} E\left[\frac{P(C = k|T_k)}{P(a \leq C \leq b)}ms(Z_r|T_{r-1})'\right] \\
&= \sum_{r=a+1}^{b}\sum_{k=a}^{r-1} E\left[ms(Z_r|T_{r-1})'|C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \\
&= \sum_{k=a}^{b-1} E\left[m\sum_{r=k+1}^{b} s(Z_r|T_{r-1})'\middle| C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \\
&= \sum_{k=a}^{b-1} E\left[ms(Z_b, \ldots, Z_{k+1}|T_k)'|C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \qquad (12)
\end{aligned}$$

where the first line follows by (1), the second line follows by the same steps that gave the sixth line in (11), the third line follows by interchanging the order of summations (which is allowed here), and the fourth (last) line follows by the definition of score. Third, we consider B_{31} and note that using the

definition of score in the first line and the same argument as before in the second (last) line below:

$$\begin{aligned} B_{31} &= \sum_{r=a}^{b}\sum_{k=1}^{r} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(Z_k|T_{k-1})'\right] = \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(T_r)'\right] \\ &= \sum_{r=a}^{b} E\left[ms(T_r)'|C=r\right]\frac{P(C=r)}{P(a\leq C\leq b)}. \end{aligned} \tag{13}$$

Now we consider the terms B_{12}, B_{22} and B_{32} respectively. Accordingly, first note that:

$$\begin{aligned} B_{12} &= \sum_{r=b+1}^{R}\sum_{k=r}^{R} E\left[\frac{I(C=k)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right] \\ &= \sum_{r=b+1}^{R} E\left[\frac{1}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\sum_{k=r}^{R}\dot{P}(C=k|T_k)'\right] \\ &= \sum_{r=b+1}^{R} E\left[\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C\geq r|T_{r-1})'}{P(C\geq r|T_{r-1})}\right] \\ &= 0 \end{aligned} \tag{14}$$

where the second line follows by (1), the third follows line by (1), (2) and (9), and the fourth (last) line follows by taking expectation conditional on T_{r-1} for the r-th term inside the summation. Exactly following the same steps as in the above (recall the analogy with B_{11} and B_{12}) we obtain:

$$B_{22} = 0. \tag{15}$$

Lastly, as before, note that:

$$B_{32} = \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(C=r|T_r)}\frac{E[m|T_r]\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right] = E\left[m\sum_{r=a}^{b}\frac{\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right]. \tag{16}$$

Therefore, (11)-(16) imply that $E[-(AM)^{-1}A\varphi(O;\beta^0)S(O)'] = \frac{\partial\beta^0(\theta_0)}{\partial\theta'}$. Finally, by matching the first set of terms in $-(AM)^{-1}A\varphi(O;\beta^0)$ (i.e., those that correspond to line one in (4)) to the terms corresponding to $\nu_{b+1}(Z_1,\ldots,Z_{b+1}),\ldots,\nu_R(Z_1,\ldots,Z_R)$ in $\mathcal{T}$; the second set of terms (i.e., those that correspond to line two in (4)) to the terms corresponding to $\nu_a(Z_1,\ldots,Z_a),\ldots,\nu_b(Z_1,\ldots,Z_b)$ in $\mathcal{T}$; and the third set of terms (i.e., those that correspond to line three in (4)) to the terms corresponding to $\omega_a(Z_1,\ldots,Z_a),\ldots,\omega_b(Z_1,\ldots,Z_b)$ in $\mathcal{T}$; while matching zeros with the remaining terms in $\mathcal{T}$, it follows that $-(AM)^{-1}A\varphi(O;\beta^0)$ is the efficient influence function given A.

STEP - 3: Standard arguments give that $A_* := \arg\inf_A Var((AM)^{-1}A\varphi(O;\beta^0)) = M'V^{-1}$. Thus $(A_*M)^{-1}A_*\varphi(O;\beta^0)$ is the efficient influence function, and Ω is the efficiency bound. ■