A Note on Efficiency Gains from Multiple Incomplete Sub-samples *

Saraswata Chaudhuri†

Current version: March 7, 2017. First version: March 8, 2013.

Abstract

Cost-effective survey methods such as multi(R)-phase sampling typically generate samples that are collections of monotonic sub-samples, i.e., the variables observed for the units in sub-sample r are also observed for the units in sub-sample $r+1$ for $r = 1, \ldots, R-1$. These sub-samples are representative of sub-populations that can be systematically different if the selection of an unit in each phase of sampling depends on the observed variables for that unit from past phases. Our paper is about optimally combining all the sub-samples for efficient estimation of a finite dimensional parameter defined by moment restrictions on a target population that is an arbitrary union of some or all of these sub-populations. Only the R-th sub-sample is assumed to contain all the variables that are arguments of the moment function. Semiparametric efficiency bounds for estimation are obtained under a unified framework allowing for full generality of the selection on observables in the sampling design. Contribution of each sub-sample toward the efficiency bounds is analyzed. An easy to compute efficient GMM estimator (call it $\widehat{\beta}$), that falls under the class of the MINPIN estimators, is proposed. By virtue of the features of our estimation framework, that in a more general context would lead to the so-called double-robustness, the first-order asymptotic properties of $\widehat{\beta}$ are only mildly affected by the estimation of the nuisance parameters. Simulation evidence of (i) the efficiency gains from using all the sub-samples and (ii) the finite-sample behavior of $\widehat{\beta}$ is provided.

JEL Classification: C13; C14; C31.

Keywords: Planned-incompleteness, Incomplete sub-samples; Multi-phase sampling; Semiparametric efficiency; Generalized method of moments.

*This is a substantially revised version of the paper that was circulated earlier. Previous versions of the paper, some of which are available on the author's webpage, benefitted from the helpful comments of A. Prokhorov, C. Muris, D. Guilkey, D. Frazier, E. Renault, F. Lange, J. Hill, J. Haushofer, J. MacKinnon, J. Wooldridge, M. Carrasco, M. Chemin, P. Saha Chaudhuri, SJ. Lee, V. Zinde-Walsh, the seminar participants at Brown, Concordia, McGill (Econ and Biostat), Queen's, U. Canterbury, U. Montreal, U. New South Wales, UNC Chapel Hill, U. Sydney, West Virginia University and the Midwest Econometrics Group meetings (2013).

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

1 Introduction

Planned incompleteness in the data can be useful when conducting surveys under budget constraints. It cuts the cost of surveys by generating samples in which only a subset of the units contains all the intended variables, while the rest contain different collections of only the less expensive variables. This happens by plan, i.e., sampling design, and eliminates or at least reduces unplanned non-response or mismeasurement that could have otherwise complicated subsequent analyses of the data.[1] In this paper we provide a unified treatment of efficient estimation based on such planned incomplete data, while taking a very general form of the sampling design as given.

The sample units with all the variables form the complete sub-sample. The rest form the incomplete sub-samples. As it usually happens in multi-phase surveys (on which we focus), the sub-samples are assumed monotonic, i.e., variables observed for the units in sub-sample r are also observed for the units in sub-sample $r+1$ for $r = 1, \ldots, R-1$. The R-th sub-sample is complete. The underlying populations for the sub-samples, call them sub-populations, can be systematically different in terms of the joint distribution of the concerned variables.

Interest generally lies in the full population, i.e., the union of all the sub-populations. However, as emphasized in Little (1993), features of the joint distributions in the sub-populations can also be interesting. Providing a unified framework, of interest in this paper are finite dimensional parameters defined by moment restrictions on these joint distributions in a generic target population that is an arbitrary union of the sub-populations.

For flexibility in the design of the multiple phases in the survey, we maintain a general selection on observables, i.e., the missing at random assumption. Identification of the parameter of interest then follows by a simple target-population-dependent probability weighting using the complete sub-sample [Horvitz and Thompson (1952)].

Under this setup, the goal of our paper is to also use all the incomplete sub-samples and combine them with the complete sub-sample to obtain an efficient and easy to compute estimator for these parameters.

Our paper is primarily based on Robins et al. (1994) and Chen et al. (2008). Our setup requires extending Chen et al. (2008), who consider the case of $R = 2$, to: (i) more than one level of planned monotonic incompleteness ($R > 2$), and (ii) with a parameter of interest defined in terms of a generic target population that is an arbitrary union of the sub-populations. We also show in a Supplemental Appendix that depending on the target population, the case of $R = 2$ has different qualitative implications on the usefulness of certain sub-samples toward efficiency under planned versus unplanned incompleteness. This difference can be reconciled only when we consider $R > 2$.

The paper proceeds as follows. Section 2 gives an idea on the premise of our paper. Section 3 introduces the theoretical framework following Tsiatis (2006) and delineates it from the closely related papers. Our framework naturally calls for GMM estimators, and Section 3 also discusses how the introduction of the various sub-samples into the analysis would contribute to reducing the variance of such estimators. Part of this discussion is based on the results from Brown and Newey (1998) and Graham (2011). To make the discussion in Section 3 precise, we set the benchmark for the variance at the efficiency bound, i.e., we focus on GMM estimators whose asymptotic variance corresponds to the bound. Section 3 also establishes the relevant bounds. It should be noted that while monotonicity naturally arises from multi-phase surveys, it also helps us to obtain closed form expressions for the

[1] See, e.g Carroll et al. (1995) and Little and Rubin (2002) for various ways of dealing with mismeasured or missing data.

efficiency bounds that we use as the benchmarks. An Appendix contains all the proofs. Certain results under non-monotonicity and for the full population are presented in a companion paper Chaudhuri and Guilkey (2016).

Section 4 considers one such GMM estimator that is feasible and falls under the class of MINPIN estimators. Its asymptotic properties follow directly from Chen et al. (2003) with the allowance for weaker conditions by virtue of our estimation framework that, in a more general context, would lead to the so-called doubly-robust models introduced and studied by Robins and co-authors [see e.g. Scharfstein et al. (1999)]. Our presentation highlights this. See Cattaneo (2010) and Rothe and Firpo (2016) for insightful discussions on double-robustness.

The GMM estimator is illustrated with $R = 3$ in a linear regression. A simpler to compute one-step updating estimator that is as efficient as this GMM estimator is presented with illustration in a linear quantile regression.

It may be tempting to not use certain incomplete sub-samples for the GMM estimation if their inclusion does not lead to meaningful gain in efficiency, since their omission (in the present context) does not affect consistency of the GMM estimator. A Monte Carlo experiment in Section 5 studies this issue numerically in a linear regression.

Analytical results for such efficiency gain/loss are provided in a Supplemental Appendix, which also contains: (A) descriptive endnotes on planned incompleteness; (B) results under unplanned incompleteness, a topic not considered in the main text; and (C) simulation evidence of good finite-sample properties of the GMM estimator.

Some of the results in (B), i.e., Supplemental Appendix B, are provided mainly to resemble Hahn (1998) and Chen et al. (2008), while others in (B) help to reveal that a fundamental difference between the implications of planned versus unplanned incompleteness on efficiency gains can only be reconciled when we extend Chen et al. (2008)'s framework to $R > 2$, i.e., to more than one level of incompleteness. In particular, these results show that unlike under planned incompleteness, that always leads to efficiency gains (except in cases similar to those in Section 5 of Wooldridge (2007)) for any R, there may not be any efficiency gain under unplanned incompleteness depending on our target population of interest, when $R = 2$. However, this difference vanishes when $R > 2$.

Lastly, we note that recently there have been many insightful contributions to data combination in economics. See e.g. Ichimura and Martinez-Sanchis (2005), Ridder and Moffitt (2007), Devereux and Tripathi (2009), Tripathi (2009), Abrevaya and Donald (2017), Muris (2016), Graham et al. (2016), and the numerous influential references therein. It is primarily our focus on planned incompleteness and the allowance for the parameter of interest to be defined in terms of arbitrary unions of sub-populations that distinguish our paper from the rest. We hope that the easy to compute estimator, and the analytical and simulation evidence of efficiency gains from the optimal use of the sub-samples encourage the practice of planned incompleteness in data collection under cost considerations.

2 Motivation and the scope for planned incompleteness in this paper

First, consider an example showing how planned incompleteness helps with efficiency under cost considerations.

Example: Let (Y, X) be scalar variables with finite means and variances. Let the parameter of interest be $\beta = E[Y - X]$. Consider two random samples $\mathcal{S}^\dagger = \{Y_j, X_j\}_{j=1}^{n^\dagger}$ and $\mathcal{S} = \{Y_i, D_i, D_i X_i\}_{i=1}^{n}$ where D is binary. We observe X in $\mathcal{S}$ only when $D = 1$. Assume that $P(D = 1|Y, X) = P(D = 1) = p$.[2] The standard and, in this

[2]While $n^\dagger$ and n are non-random quantities, we allow, here and throughout, D to be random. Hence $n_D := \sum_{i=1}^{n} D_i \sim Bin(n, p)$,

case, efficient estimator of β based on $\mathcal{S}^\dagger$ is: $\widehat{\beta}^\dagger = \sum_{j=1}^{n^\dagger}(Y_j - X_j)/n^\dagger$ with $Var(\widehat{\beta}^\dagger) = Var(Y-X)/n^\dagger$. On the other hand, a special case of our results (Theorem 2, Chen et al. (2008)) gives an estimator of β based on $\mathcal{S}$ as:[3]

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_i}{p}(Y_i - X_i) + \left(1 - \frac{D_i}{p}\right)(Y_i - E[X|Y_i])\right\} \text{ with } Var\left(\widehat{\beta}\right) = \frac{1}{n}\left[Var(Y-X) + \frac{1-p}{p}E[Var(X|Y)]\right].$$

Now, let the cost of observing Y for an unit be 1 and that for X be $c > 1$. Let the allowed expected total cost for the sample be c^*. Thus, $n^\dagger = \lfloor c^*/(1+c)\rfloor$ and $n = \lfloor c^*/(1+pc)\rfloor$ for a given c, c^* and p, and where $\lfloor a\rfloor$ denotes the largest integer $\leq a$. Consider the problem of choosing p such that $Var(\widehat{\beta}) < Var(\widehat{\beta}^\dagger)$.[4] By simple calculations: $Var(\widehat{\beta}) < Var(\widehat{\beta}^\dagger) \iff p > 1/(cq)$ given $cq > 1$ where $q = Var(Y-X)/E[Var(X|Y)] - 1$. No solution exists if $cq \leq 1$. However, if $cq > 1$ and $p > 1/(cq)$, then the sample $\mathcal{S}$ is strictly advantageous over the sample $\mathcal{S}^\dagger$ under the premise of the stated problem. (If Y and X are normally distributed with unit variance and correlation ρ then $q = (1-\rho)/(1+\rho)$.) If $cq > 1$ and $n = c^*/(1+pc)$, $Var(\widehat{\beta})$ is minimized when $p = 1/\sqrt{cq}$. ■

Unfortunately, such explicit optimality arguments will not be possible under our general setup.[5] Nevertheless, planned incompleteness in the data has long been recognized as a cost-effective strategy and frequently employed in various fields of research where data collection by the researcher has traditionally been more prevalent than in economics. See Supplemental Appendix A1 for a brief review of instances of such planned incompleteness.

In the present paper we impose a monotone pattern in the planned incompleteness to resemble the data structure from multi-phase surveys. To facilitate the utilization of the multi-phase sequential nature of the sampling, either for variance reduction or for other purposes, we let the sampling design at each phase depend on the data observed until then. (See Supplemental Appendix A2 for an example showing how dependent, as opposed to independent, sampling may lead to variance reduction.) Lack of such dependence are treated as special cases.

For an idea on the scope of our paper, consider three scenarios that could, by plan/sampling-design, lead to a complete sub-sample and multiple incomplete sub-samples. The scenarios are presented in chronological order and, to avoid repetition, issues of interest to our paper are discussed here in the context of the first scenario only.

Scenario 1: Consider laboratory experiments following Holt and Laury (2002) to elicit risk aversion and study its dependence on the size of the stake (incentives). For concreteness, define $Z_{(1)}$ as the demographic variables and the responses under the low stake, and $Z_{(2)}$ and $Z_{(3)}$ as the responses respectively under 50× and 90× the low stake. Budget constraint may necessitate that the experiment be conducted such that $Z_{(1)}$ is collected in the first phase for all the subjects, $Z_{(2)}$ in the second phase for a subset, and $Z_{(3)}$ in the third phase for a further subset. Thus, the subjects can be partitioned into three sub-samples: the first and the second ones are incomplete since

i.e., the size of the complete sub-sample is random. This is in spirit similar to the familiar relationship between multinomial sampling and standard stratified sampling. It provides the technical convenience to consider a variety of cases under a unified framework.

[3] $E[X|Y]$ is unknown in practice and needs to be replaced by an estimator, say $\widehat{E}[X|Y]$. An important and desirable feature of our results is: as long as $\widehat{E}[X|Y]$ is consistent for $E[X|Y]$ uniformly in Support(Y), plugging this estimator in the formula for $\widehat{\beta}$ only makes the result asymptotic, i.e., (i) what is referred to as $Var(\widehat{\beta})$ turns out to be the $(1/n)$ times the asymptotic variance of $\widehat{\beta}$, and (ii) $\widehat{\beta}$ is no longer unbiased (neither will be $\widehat{\beta}^\dagger$ in more general cases), but is asymptotically unbiased and normally distributed.

[4] While the idea of choosing p that makes n_D random may seem odd, thanks to arguments as in e.g. Theorems 4.1 and 4.3 of Devereux and Tripathi (2009), similar insights follow if one equivalently considers choosing n_D (and hence $n = \lfloor c^* - cn_D\rfloor$) to resemble the standard survey design problems of choosing sample size, and subsequently estimating p jointly with β.

[5] It is rarely possible in setups more complex than Reilly (1996)'s who considers a two-stage(phase) design (e.g. case-control) where one collects the binary dependent variable and some inexpensive (categorical) covariates in the first phase, and more expensive covariates for a subset of subjects in the second. See Cochran (1977) (Chapter 12) for the standard pros and cons of double-sampling. Song et al. (2009) note that outcome(choice)-based samplings, even with continuous outcomes, neatly fit into the missing data setup.

they only contain $Z_{(1)}$ and $Z_{(1)}, Z_{(2)}$ respectively, while the third one is complete since it contains $Z_{(1)}, Z_{(2)}, Z_{(3)}$.

The parameters of interest can be the risk attitudes under each of these three stakes or any combinations of them, and this will exhaust all possibilities if the successive subsets of units are chosen independently of $Z_{(1)}, Z_{(2)}, Z_{(3)}$. However, if the successive units are chosen based on their past responses, then the sub-populations corresponding to the sub-samples are systematically different, causing the aforementioned parameters of interest to vary with the sub-populations. Depending on how the dependence in the sampling was introduced — that can reflect objectives such as to simply increase the precision of the estimates, or to over-sample subjects whose past responses suggest a particular type of risk attitude, etc. — these sub-population-specific parameters of interest can be interesting broadly or at least for descriptive purposes. The questions of interest to us can then be framed in this context as: if e.g. the risk attitude when presented with the highest stake (i.e., $90\times$ the low stake) in the full population or a sub-population is the parameter of interest, then (i) does each available sub-sample (not necessarily from the target) contribute incremental information for the estimation of this parameter?; and if so, (ii) how can we optimally use all the available sub-samples to obtain an efficient estimator for this parameter?

Scenario 2: McKenzie (2012) draws on the clinical trial literature and provides an analysis on the benefit in precision gains from multiple followup measurements in field experiments over the standard practice of a single baseline and a single followup. His discussion focuses on the tradeoff in the choice of n (number of subjects) and T (number of measurements including baseline and followups) at a given cost. Alternatively, one could keep both n and T large but measure the relevant variables only for a subset of subjects at each followup.

Scenario 3: In their editorial introduction, McKenzie and Rosenzweig (2012) note how the different measurements of the same variables can dramatically alter the conclusion of analyses using survey data. However, the "good" measures can be substantially more expensive. For example, collecting consumption data through maintaining a personal diary can be 6 to 10 times more expensive than a 7-day recall [see the last 3 rows of column 6 in Table 10 of Beegle et al. (2012)]. On the other hand, the 7-day recall with a short list of aggregated consumption items can understate food consumption by 30% as compared to personal diaries [compare rows 4 and 8 of column 2 in Table 2 of Beegle et al. (2012)]. Hence, faced with a budget constraint, one could obtain the good but expensive measures for only subsets of subjects, and the other measures for larger subsets or everyone.

Lastly, we note that while it is natural to consider the prototypical multi-phase sampling (as in the description of Scenario 1) in all cases, other types of multi-phases (loosely defined) can also be accommodated by our results under our maintained assumption of selection on observables (stated formally in (3) below). For example, one may start with the complete sub-sample and then, given additional budget, collect subsets of variables for new groups of sample units to form a monotonic collection of complete and incomplete sub-samples. Unlike in Scenario 1, where the successive phases progressively enriched the information content of certain sample units by collecting new variables, this is a case that involves deletion of variables from the future phases, resulting in incomplete information being collected for new sample units. (The selection on observables assumptions will trivially hold in this case if the sampling design is independent of the concerned variables.) Under all such cases, of interest to us in this paper is: how to optimally combine the complete and incomplete sub-samples for efficient estimation.

3 Framework and the Combination of Sub-samples

3.1 Framework

Let $Z := (Z'_{(1)}, \ldots, Z'_{(R)})'$ where $Z_{(r)}$ is a $d_r \times 1$ random vector for $r = 1, \ldots, R$, and $\sum_{r=1}^{R} d_r$ is finite. To model the monotonic pattern in the observability of the elements of Z, following Tsiatis (2006), consider a scalar variable C with support $\mathcal{C} := \{1, \ldots, R\}$ and a transformation $T_C(Z)$ defined as $T_r(Z) := (Z'_{(1)}, \ldots, Z'_{(r)})'$ of dimension $(\sum_{s=1}^{r} d_s) \times 1$ for $r = 1, \ldots, R$. The value of C determines $T_C(Z)$, i.e., how much of Z is observed.

Let $O := (C, T'_C(Z))'$ denote what is observed for an unit. The observed sample is $\{O_i := (C'_i, T'_{C_i}(Z_i))'\}_{i=1}^{n}$. The r-th sub-sample is the collection of units for whom $T_r(Z)$ is observed; it is of size $n_r := \sum_{i=1}^{n} I(C_i = r)$ for $r = 1, \ldots, R$. The R-th sub-sample is complete, i.e., $T_R(Z) = Z$. The other sub-samples are incomplete.

Now 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 $\lambda \in \Lambda$ where $\Lambda := \text{Power-Set}(\mathcal{C})$ excluding the empty set, let the parameter value of interest β^0_λ be defined as:

$$E[m(Z; \beta)|C \in \lambda] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0_\lambda. \tag{1}$$

β^0_λ is defined as a function of λ and may not be same across target populations $\lambda \in \Lambda$ if C and Z are dependent.

For a given β, the function $m(Z; \beta)$ can be evaluated from the observed sample only for the n_R units in the complete sub-sample, i.e., $I(C = R)m(Z; \beta)$. However, point identification of β^0_λ is still possible by the Horvitz-Thompson re-weighting if $P(C = R|T_R(Z)) > 0$ almost surely in $T_R(Z)$, since for any given β:

$$E\left[\frac{P(C \in \lambda|T_R(Z))}{P(C \in \lambda)} \frac{I(C = R)}{P(C = R|T_R(Z))} m(Z; \beta)\right] = E[m(Z; \beta)|C \in \lambda]. \tag{2}$$

To relate our setup to the (by now) standard terminology introduced in Chen et al. (2008) where the "primary sample" refers to the one whose population is the target λ and the "auxiliary sample" is the one which brings identification, our sub-sample R is always the auxiliary sample. Thus, if any element of λ contains R then (1) can be seen as Chen et al. (2008)'s "verify-in-sample" case, and otherwise as their "verify-out-of-sample" case.

To be general in characterizing the multi-phase nature of the sampling design associated with such complete and incomplete sub-samples, we maintain a general selection on observables assumption that: for $r = 1, \ldots, R$

$$\text{MAR:} \qquad P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r|T_r(Z)). \tag{3}$$

This is the MAR assumption [see e.g. Robins and Rotnitzky (1995), Tsiatis (2006)] in the sense of Rubin (1976). (3) implies that $P(C \geq r|Z) = 1 - P(C \leq r-1|Z) \overset{\text{by (3)}}{=} 1 - \sum_{j=1}^{r-1} P(C = j|T_j(Z)) \overset{\text{by (3)}}{=} 1 - \sum_{j=1}^{r-1} P(C = j|T_{r-1}(Z)) = P(C \geq r|T_{r-1}(Z))$ only depends on $T_{r-1}(Z)$. Hence, taking $r = R = 2$ in (3), it does not contradict the standard representation of MAR, $P(C = 2|Z) = P(C = 2|Z_{(1)})$, in economics where the focus has traditionally been on $R = 2$ [see e.g. Chen et al. (2005), Chen et al. (2008), Graham (2011), Graham et al. (2012)].

Given (3), the discussion in this section will focus on exploring the information content of each sub-sample and how all such information could be combined for the purpose of efficient estimation of β^0_λ defined in (1).

Naturally, all our technical results derived under the general condition (3) also hold under special cases such as (4) and (5), that we call convenient MAR (CMAR) and independent (INDEP) sampling respectively:

$$\text{CMAR:} \qquad P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r|T_1(Z)), \tag{4}$$

$$\text{INDEP:} \qquad P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r). \tag{5}$$

CMAR happens if the sampling design for the later phases is based only on the observed variables from the first phase (pilot phase). CMAR and MAR are the same in the commonly studied case of $R = 2$. INDEP happens if the sampling design is independent of Z. $\lambda = \mathcal{C}$ is the only target of interest under INDEP since β^0_λ does not vary with λ [also see Remark 1 on page 812 of Chen et al. (2008)]; however, that is not the case under CMAR.

(3) also covers other scenarios of practical importance. For example, letting $Z_{(r)} = (Z'_{(r1)}, Z'_{(r2)})'$ where $Z_{(rj)}$ is $d_{rj} \times 1$ for $j = 1, 2$ and $r = 1, \ldots, R$ and taking $m(Z; \beta) = m(Z_{(11)}, Z_{(21)}, \ldots, Z_{(R1)}; \beta)$ allow for the presence of auxiliary variables $(Z'_{(12)}, Z'_{(22)}, \ldots, Z'_{(R2)})'$ that do not enter $m(Z; \beta)$ but affect the observability of the variables involved in it. Further modifying (3) by instead assuming that $P(C = r|Z) = P(C = r|Z_{(12)}, Z_{(22)}, \ldots, Z_{(R2)})$ serves a similar purpose. One could also simply let e.g. $m(Z; \beta) = m(Z_{(R)}; \beta)$, $m(Z_{(1)}, Z_{(R)}; \beta)$, etc. to allow for the auxiliary variables in the sampling design (we do this in the simulation study in Section 5).[6]

Our theoretical framework is closely related to several papers, and it is important to note where we actually differ from them. Consider the following not-too-old representative examples under the non-Bayesian paradigm. (1) Whittemore (1997) considers maximum likelihood and Horvitz-Thompson estimators with data obtained by multi-phase sampling (and seems to prefer the latter) where the target is the full population, i.e, $\lambda = \mathcal{C}$. (2) Robins and Rotnitzky (1995) and Holcroft et al. (1997) consider optimally using all the sub-samples under a framework similar to ours but with $\lambda = \mathcal{C}$. (3) Lee et al. (2012) consider efficient semiparametric likelihood-based estimation with $\lambda = \mathcal{C}$ in multi-phase case-control studies when $T_{R-1}(Z)$ has a finite number of support points. (4) While the multi-valued treatment framework with $\lambda = \mathcal{C}$ considered in Cattaneo (2010) is generally related, it also differs in an important way because we actually allow the entire random vector Z to be the argument for each element of the vectorial moment function $m(Z; \beta)$, and thus for each element there can be R levels of hierarchy in observability. This creates a major difference in terms of efficiency bounds, efficient influence functions, etc. [see Chaudhuri and Guilkey (2016)]. (5) Our framework is also related to special cases of the literature on dynamic treatment regimes [see Robins (2004) and the references therein] but the focus is different.[7] We do not explore it for brevity. (6) Finally, Chen et al. (2005) and Chen et al. (2008) consider frameworks where β^0_λ is defined in the same way as (1), i.e., it can characterize the sub-populations also, for $R = 2$ and $\lambda = \{1\}$ (sub-population)

[6] One important scenario that we do not cover is that of Wooldridge (2002, 2007) where, in terms of our notation: $R = 2$, $T_1(Z) = Z_{(1)}$ is *empty*, $T_2(Z) = (Z'_{(21)}, Z'_{(22)})' = Z_{(2)} = Z$, $m(Z; \beta) = m(Z_{(21)}; \beta)$ and $P(C = 2|Z) = P(C = 2|Z_{(22)})$. A leading example is the variable probability sampling that discards all (or no) information about units with probability depending on, say, their $Z_{(22)}$ [see Wooldridge (1999)]. The author uses the Horvitz-Thompson approach to correct for selection bias in estimation based on the complete sub-sample. The key point is: there is essentially only one sub-sample under this scenario, and it is complete. Given our focus on optimally combining *all the sub-samples* for efficient estimation, it is possible that we do not lose much by this omission.

[7] For example, consider an R-period experiment where at each period (after the first) either a treatment is assigned or the subject is dropped from the experiment (i.e., not observed further) depending on the history of observables for the subject until that period. Let $Z_{(r)}$ be the observables (including the outcome) from the r-th period and $C = r$ the subjects who received treatment until period r. (Causal interpretations require care in determining the conditioning set based on the elements of $Z_{(r)}$.) This simplistic representation establishes a relation with cases such as our discussion of Holt and Laury (2002) under Scenario 1. Thus our results are somewhat applicable to this literature, a proper treatment of which is however much beyond the scope and focus of our paper.

and $\{1,2\}$ (full population). By contrast, we allow for a general R and expand the scope to all possible (2^R-1) sub-populations (including $\lambda=\mathcal{C}$) under a unified framework for a comprehensive treatment of the topic.

To reflect that our paper focuses on planned incompleteness due to multi-phase surveys, we maintain that:

$$P(C=r|T_r(Z)=T_r(z)) \text{ is known for all } T_r(z)\in \text{Support}(T_r(Z)) \text{ and for all } r=1,\ldots,R. \tag{6}$$

While this assumption is restrictive for more general purposes, it is not unreasonable under planned incompleteness. (6) does not imply that $P(C=r)$ is known, except trivially under INDEP in (5). Therefore, our discussion of efficiency is based on (6) only and does not consider a known $P(C=r)$ as additional information.

Remark: As an aside, we note here that Supplemental Appendix B contains certain results under (3) but without imposing (6). Under the special case of (4) it also demonstrates, following Hahn (1998) and Chen et al. (2008), the differential in the efficiency bounds for estimation β^0_λ for any λ and R depending on if $P(C=r|T_r(Z)=T_r(z))$ is known, or partially known up to finite dimensional unknown parameters, or completely unknown.

The general discussion of our framework concludes by listing an assumption that we also maintain throughout.

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

Remarks: 1. (A1) is a standard assumption [see Tsiatis (2006)].[8] **2.** $P(C=R|T_R(Z))>\underline{p}>0$ in (A2) is a strict version of the overlap assumption [see Khan and Tamer (2010)]. The restrictions $P(C=r|T_R(Z))>0$ for $r=1,\ldots,R-1$ are not strictly required but help to avoid more involved proofs peripheral to the main message. However $P(C=r)>0$ for $r=1,\ldots,R$ is intrinsic to the R-level missing data model. **3.** (A3) allows for moment vectors $m(Z;\beta)$ that are not differentiable in β. We do impose differentiability of $E[m(Z;\beta)|C\in\lambda]$, which is standard [see Chen et al. (2003), Chen et al. (2008), Cattaneo (2010), etc.].

3.2 Combining the sub-samples for efficient estimation

To state our key result in Proposition 1 that provides the foundation for the rest of the paper, let us first, for a given $\lambda\in\Lambda$, define the following $d_m\times 1$ functions of the observed data O and the $d_\beta\times 1$ parameter β as:

$$\varphi_{r,\lambda}(O;\beta) := E\left[\frac{P(C\in\lambda|T_R(Z))}{P(C\in\lambda)}m(T_R(Z);\beta)\middle|T_r(Z)\right] \text{ for } r=1,\ldots,R, \tag{7}$$

$$\begin{aligned}\varphi_\lambda(O;\beta) := {}& \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta)\\ &+\sum_{r=1}^{R-1}\left[\frac{I(C\geq R-r)}{P(C\geq R-r|T_{R-r}(Z))}-\frac{I(C\geq R-r+1)}{P(C\geq R-r+1|T_{R-r+1}(Z))}\right]\varphi_{R-r,\lambda}(O;\beta).\end{aligned} \tag{8}$$

Note that, throughout we will use the notation $T_r(Z):=(Z'_{(1)},\ldots,Z'_{(r)})'$ for $r=1,\ldots,R$ and, thus $T_R(Z):=Z$.

[8] Devereux and Tripathi (2009) (Section 3) formally discuss (A1) and emphasize the technical convenience that it provides in allowing to treat the sample as i.i.d. from an enlarged population at the minor cost of making the sub-sample sizes random. As noted in the footnotes 2 and 4 for Example from Section 2, this technical convenience is similar to the more well known case of what a multinomial sampling representation provides over the standard stratified sampling [see e.g. page 50 in Tripathi (2011)].

Proposition 1 *Let* (1), (3), (6) *and assumption A hold. Let the* $d_m \times d_m$ *matrix* $V_\lambda := Var(\varphi_\lambda(O;\beta^0_\lambda))$ *be finite and positive definite where* $\varphi_\lambda(O;\beta)$ *is defined in* (8) *and* β^0_λ *is defined in* (1). *Then for* β^0_λ, *the asymptotic variance lower bound for* $\sqrt{n}(\widehat{\beta}-\beta^0_\lambda)$ *of any regular estimator* $\widehat{\beta}$ *is given by* $\Omega_\lambda := (M'_\lambda V_\lambda^{-1} M_\lambda)^{-1}$. *An estimator whose asymptotic variance equals* Ω_λ *has the asymptotically linear representation*

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta^0_\lambda) = -\Omega_\lambda M'_\lambda V_\lambda^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_\lambda(O_i;\beta^0_\lambda) + o_p(1).$$

Remarks:

1. This proposition generalizes Theorem 2 of Chen et al. (2008) to the case of a general R and a general target λ. When $R > 2$, their selection on observables assumption (Assumption 2) can be generalized either as MAR in (3) or as CMAR in (4), the latter being a special case of the former. Proposition 1 is obtained under MAR.

2. The result for a general R under MAR and when the target is $\lambda = \mathcal{C}$ has been known since Robins and Rotnitzky (1995); Rotnitzky and Robins (1995), Robins et al. (1995), Holcroft et al. (1997). Proposition 1 allows for any target λ. The key to this is our treatment of the term $P(C \in \lambda | T_R(Z))$ in (7) (immaterial when $\lambda = \mathcal{C}$). It simplifies to Chen et al. (2008)'s treatment when considering their verify-out-of-sample case, i.e., when $R = 2$ and $\lambda = \{1\}$ [see their equation (30)], from which, on the other hand, an extension to the general case considered in this paper possibly does not seem obvious *ex ante*. We provide more intuition on this in Proposition 2 below.

(Supplemental Appendix B presents generalizations ignoring the known $P(C = r|T_r(Z))$ assumption in (6).)

3. $\varphi_\lambda(O;\beta)$ in (8) belongs to the class of AIPW (Augmented Inverse Probability Weighted) estimating functions of Robins et al. (1994). The first term $\varphi_{R,\lambda}(O;\beta)$ is the IPW term based on the complete sub-sample. The rest are the augmentations due to the incomplete sub-samples: the r-th term represents the contribution of the $(R-r+1)$-th sub-sample. Each of these R terms are themselves unbiased estimating function for β^0_λ but only the first one, i.e., the IPW, is known without further assumptions. The augmentation terms reduce the variability of the IPW estimating function and thereby deliver the efficient AIPW estimating function. More precisely:

$$\begin{aligned}
Cov(\text{term}_1, \text{term}_r) &= -Var(\text{term}_r) \text{ for } r = 2,\ldots,R \\
&= E\left[\left(\frac{1}{P(C \geq R-r+1 | T_{R-r+1}(Z))} \right.\right. \\
&\qquad \left.\left. - \frac{1}{P(C \geq R-r+2 | T_{R-r+2}(Z))}\right) \varphi_{R-r+1,\lambda}(O;\beta^0_\lambda)\varphi'_{R-r+1,\lambda}(O;\beta^0_\lambda)\right], \qquad (9) \\
Cov(\text{term}_s, \text{term}_r) &= 0 \text{ for } s \neq r \neq 1, \\
\text{and hence } V_\lambda &= Var\left(\sum_{r=1}^{R} \text{term}_r\right) = Var(\text{term}_1) - \sum_{r=2}^{R} Var(\text{term}_r).
\end{aligned}$$

The $(R-r+1)$-th sub-sample's contribution to the efficiency for estimation of β^0_λ rises with $Var(\text{term}_r)$ for $r > 1$. While $Var(\varphi_{R-r+1,\lambda}(O;\beta)) \geq Var(\varphi_{R-r,\lambda}(O;\beta))$ (in a matrix sense), the order is however not always preserved when comparing the relative contribution of $Var(\text{term}_r)$ and $Var(\text{term}_{r+1})$ for $r > 1$ (see (9)). This makes it difficult in general (see footnote 5) to optimally design the sub-samples under MAR in (3) or even CMAR in (4).

Accordingly, Proposition 1 provides guidance on how one could combine a given set of complete and incomplete sub-samples, that might or might not have been obtained optimally, for the purpose of efficient estimation.

We conclude this section by looking into combining the sub-samples from alternative viewpoints but to the same effect. To this end, let us rearrange the terms on the right hand side (RHS) of (8) and rewrite $\varphi_\lambda(O;\beta)$ as

$$\varphi_\lambda(O;\beta) = \varphi_{1,\lambda}(O;\beta) + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r(Z))} \left[\varphi_{r,\lambda}(O;\beta) - \varphi_{r-1,\lambda}(O;\beta)\right] \tag{10}$$

to slice the contribution of the sub-samples differently. Consider the r-th term on the RHS. $\varphi_{r,\lambda}(O;\beta)$ and $\varphi_{r-1,\lambda}(O;\beta)$ differ due to $Z_{(r)}$, which is observed for all the $(R-r+1)$ sub-samples (i.e., for all the units $i = 1,\ldots,n : C_i \geq r$) as is signified by the multiplier $I(C \geq r)$. Thus, the contribution of all the R sub-samples toward estimation is represented in this r-th term in an incremental fashion only according to their ability in delivering an observable $Z_{(r)}$. This holds for each $r = 1,\ldots,R$, i.e., including the first term on the RHS of (10). Note that the R terms on the RHS are uncorrelated. Therefore, V_λ is the sum of the variances of the R terms:

$$V_\lambda = Var(\varphi_{1,\lambda}(O;\beta^0_\lambda)) + \sum_{r=2}^{R} E\left[\frac{Var(\varphi_{r,\lambda}(O;\beta^0_\lambda)|T_{r-1}(Z))}{P(C \geq r|T_r(Z))}\right].$$

The variance inflating factor $1/P(C \geq r|T_r(Z))$ accounts for the observability of $Z_{(r)}$ by varying inversely with the conditional probability of observing $Z_{(r)}$. There is no inflation for the first term since $Z_{(1)}$ is always observed.

Yet another way of looking at it is to design a set of extended moment restrictions whose information content, when combined efficiently, equals that in Proposition 1. Accordingly, consider estimation of β^0_λ based on:

$$E\left[\phi_{R,\lambda}(O;\beta)\right] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0_\lambda, \tag{11}$$

$$E\left[\phi_{R-r}(O)|T_{R-r}(Z)\right] = 0 \text{ almost surely } T_{R-r}(Z) \text{ for } r = 1,\ldots,R-1; \tag{12}$$

where:

$$\begin{aligned}
\phi_{R,\lambda}(O;\beta) &:= \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta) \quad \text{[the IPW term from (8)]},\\
\phi_{R-r}(O) &:= I(C \geq R-r)\left[I(C \geq R-r+1) - P(C \geq R-r+1|C \geq R-r, T_{R-r}(Z))\right]
\end{aligned}$$

for $r = 1\ldots,R-1$. (We wrote $I(C=R)$ as $I(C \geq R)$ and 1 as $I(C \geq 1)$ in the last line for notational brevity.)

The moment restriction in (11) already identifies β^0_λ (see (2)), and GMM estimation based on it is the GMM-version of the Horvitz-Thompson method of obtaining IPW estimators. The moment restrictions in (12) do not involve β but bring additional information that captures the information content of the MAR assumption (3) under the monotonic structure of the observed data. In particular, as evident from the multiplier $I(C \geq r)$ for the r-th moment function $\phi_r(O)$ for $r = 1,\ldots,R-1$, the corresponding moment restriction reflects the additional information that becomes available due to the observability of $Z_{(r)}$ that is observed only when $C \geq r$.

Efficiency results under moment restrictions of the form (11) and (12) follow from Chamberlain (1992) [also see Ai and Chen (2012)]. (To match the sequential moment restrictions in these references, define $T_0(Z)$ as a constant and consider, equivalently, (11) as expectation conditional on $T_0(Z)$.) In particular, the efficient estimating function can be obtained by repeated application of equation (15) in Brown and Newey (1998) [see also Theorem 2.1 of Graham (2011)] facilitated by the monotonic structure ($T_r(Z)$ nests $T_{r-1}(Z)$) of the conditioning set in (12). Proposition 2 below links this efficient estimating function with the result in Proposition 1.

Proposition 2 *Let* (3) *and assumptions (A1) and (A2) hold. For any* $r = 1, \ldots, R-1$*, denoting* $\phi_r(O)$ *by* ϕ_r*, define* $\overline{Proj}_{T_r}(Y|\,\phi_r) := Y - Proj_{T_r}(Y|\,\phi_r)$ *where* $Proj_{T_r}(Y|\,\phi_r) := E\left[Y\phi_r|T_r(Z)\right]\left(E\left[\phi_r^2|T_r(Z)\right]\right)^{-1}\phi_r$ *for any random variable* Y *such that the conditional expectations exist. Then* $\varphi_\lambda(O;\beta)$ *defined in* (8) *satisfies:*

$$\varphi_\lambda(O;\beta) = \overline{Proj}_{T_1}\left(\overline{Proj}_{T_2}\left(\ldots\overline{Proj}_{T_{R-2}}\left(\overline{Proj}_{T_{R-1}}\left(\phi_{R,\lambda}(O;\beta)|\,\phi_{R-1}\right)\Big|\,\phi_{R-2}\right)\ldots\Big|\,\phi_2\right)\Big|\,\phi_1\right).$$

Remark: The RHS above offers an alternative way of looking at the efficient influence function for β_λ^0 in Proposition 1. (Also see Theorem 1 in Chamberlain (1992).) The repeated projection operation in the RHS is the repeated application of equation (15) in Brown and Newey (1998) to efficiently combine the information from the unconditional and conditional moment restrictions in (11) and (12) respectively such that GMM based on the RHS with the optimal weighting matrix would result in a semiparametrically efficient estimator for β_λ^0 under (11)-(12). (This will not hold in general under non-monotonic structure of the data without further assumptions; see footnote 5 in Chaudhuri and Guilkey (2016).) Therefore, the original problem of efficiently combining the sub-samples boils down to an equivalent problem of efficiently combining a set of carefully chosen moments restrictions, a problem/idea that is perhaps more common in economics. Graham (2011) was first to establish a similar result for the case where $R = 2$ and the target was $\lambda = \mathcal{C}$. Our setup is more involved and thus requires an adequately rich choice for the sequence of functions $(\phi_{R-r}(O))_{r=1}^{R-1}$ in (12) to establish the equivalence result that allows an alternative viewpoint to appreciate the contribution of the sub-samples toward efficient estimation.

(Supplemental Appendix A3 contains a brief discussion of the related literature on moment augmentation in econometrics and a mention of its connection with the calibration literature in survey sampling.)

4 GMM estimation of the parameter of interest β_λ^0 defined in (1)

4.1 Estimation framework: Definitions and the Key feature

Estimation of β_λ^0 can be done as a standard exercise in GMM by treating $\varphi_\lambda(O;\beta)$ in (8) as the moment vector. However, only the first term $I(C = R)\varphi_{R,\lambda}(O;\beta)/P(C = R|T_R(Z))$, i.e., the IPW term, of $\varphi_\lambda(O;\beta)$ is feasible based on the observed data $\{O_i = (C_i', T_{C_i}'(Z_i))'\}_{i=1}^n$.[9] The other terms involve the unknown conditional expectations $(\varphi_{r,\lambda}(O;\beta))_{r=1}^{R-1}$ and must be estimated prior to the estimation of β_λ^0; thus the MINPIN estimation. Whether we treat the unknowns in each term as finite or infinite dimensional determines whether the estimator of β_λ^0 is parametric or semiparametric. We consider both under, by now, well-understood high level conditions.

Our estimation framework is a special case of Chen et al. (2003), from which it is straightforward to establish the standard asymptotic properties of the GMM estimator for β_λ^0. The estimator is defined below in (14).

However, our framework has a key feature that in a more general context would lead to the doubly-robust property introduced by Robins and co-authors [see Scharfstein et al. (1999)]. [See Cattaneo (2010), Rothe and Firpo (2016) and the references therein.] This feature ensures that the expectation of the average moment vector for the GMM estimation does not vary *at all* with the nuisance parameters inside a large nuisance parameter

[9] Since all the terms involve the scalar multiple $1/P(C \in \lambda)$, a positive constant by (A2), we safely ignore it for GMM estimation.

space defined below in (17). Thus, it helps to weaken Chen et al. (2003)'s conditions in practically important ways while ensuring that the estimation of the nuisance parameter has only mild effects on the asymptotic properties of the GMM estimator for β_λ^0. Our discussion below in Section 4.1-4.2 focuses on making this statement precise.

To consolidate notation following Chen et al. (2003), and guided by (8), define a $d_m \times 1$ function:

$$g(O;\beta,h(\beta)) := \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta) + \sum_{r=1}^{R-1}\left[\frac{I(C\geq r)}{P(C\geq r|T_r(Z))} - \frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1}(Z))}\right]h_r(\beta) \quad (13)$$

where $h(\beta) = (h_1'(\beta),\dots h_{R-1}'(\beta))'$ are the unknown nuisance parameters, and $h_r(\beta) : (C,Z,\beta) \mapsto \mathbb{R}^{d_m}$ belongs to a class of functions, call it $\mathcal{H}_r(\beta)$, for $r = 1,\dots,R-1$. Note that, if $h_r(\beta) = \varphi_{r,\lambda}(O;\beta)$ for $r = 1,\dots,R-1$ then $g(O;\beta,h(\beta)) = \varphi_\lambda(O;\beta)$ defined in (8). Denote the true $h_r(\beta)$ as $h_r^0(\beta) := \varphi_{r,\lambda}(O;\beta)$ for $r = 1,\dots,R-1$.

The GMM estimator: Take $g(O;\beta,h(\beta))$ as the moment vector for the GMM estimation. A feasible version of it requires that for $r = 1,\dots R-1$, the unknown $h_r(\beta)$ for the n_r observations $\{i = 1,\dots,n : C_i = r\}$, call it $h_{r,i}(\beta)$, be replaced with an estimator $\widehat{h}_{r,i}(\beta)$. Now, define the average moment vector and its expectation as:

$$G_n(\beta,h(\beta)) := \frac{1}{n}\sum_{i=1}^n g(O_i;\beta,(h_{1,i}'(\beta),\dots,h_{R-1,i}'(\beta))') \text{ and } G(\beta,h(\beta)) := E\left[G_n(\beta,h(\beta))\right].$$

Then, for a $d_m \times d_m$ symmetric weighting matrix W_n, the GMM estimator $\widehat{\beta}_\lambda(W_n)$ of β_λ^0 is defined as:

$$\widehat{\beta}_\lambda(W_n) \approx \arg\min_{\beta\in\mathcal{B}} \|G_n(\beta,\widehat{h}(\beta))\|_{W_n} \quad (14)$$

where for two conformable matrices A and B, $\|A\|_B := \sqrt{\text{trace}(A'BA)}$. Write it as $\|A\|$ if B is an identity matrix.

First-step estimation of the nuisance parameters: As an example, consider a naive series estimator $\widehat{h}_{r,i}(\beta) \equiv \widehat{h}_r(\beta;T_r(Z_i))$ for $h_{r,i}(\beta)$ for $\{i = 1,\dots,n : C_i = r\}$. $\widehat{h}_r(\beta;T_r(Z))$ operates on $T_r(Z)$ and is defined as

$$\widehat{h}_r(\beta;T_r(Z)) := \left(\sum_{j=1}^n \omega_{r+1,j}\widehat{h}_{r+1,j}(\beta)\Upsilon_{K_r}'(T_r(Z_j))\right)\left(\sum_{l=1}^n \omega_{r,l}\Upsilon_{K_r}(T_r(Z_l))\Upsilon_{K_r}'(T_r(Z_l))\right)^- \Upsilon_{K_r}(T_r(Z)) \quad (15)$$

for $r = 1,\dots,R-1$. Inside the first parentheses on the RHS, we take $\widehat{h}_{R,i}(\beta)$ as $\varphi_{R,\lambda}(O_i;\beta)$ as a convention. $\omega_{r,j} = I(C_j \geq r)/P(C \geq r|T_r(Z_j))$ are the balancing weights for the j-th observation for $j = 1,\dots,n$ and $r = 1,\dots,R-1$. $\Upsilon_{K_r}(T_r(Z))$ is a $K_r \times 1$ series of functions of $T_r(Z)$ [see e.g. Newey (1997), Chen (2007)]. $\widehat{h}_r(\beta;T_r(Z))$ is parametric when K_r is fixed for $r = 1,\dots,R-1$, and nonparametric when $K_r \to \infty$ as $n \to \infty$. Let $\widehat{h}(\beta) := (\widehat{h}_1'(\beta;T_1(Z)),\dots,\widehat{h}_{R-1}'(\beta;T_{R-1}(Z)))'$. We use this estimator for the simulation study in Section 5.

Probability limit of $\widehat{h}(\beta)$ and the Nuisance parameter space: Let $\mathcal{Q}(\beta)$ denote the functions in $\mathbb{R}^{d_m}$ of C,Z,β, and let $\mathcal{Q} := \{\mathcal{Q}(\beta) : \beta \in \mathcal{B}\}$ be, as in Chen et al. (2003), a vector space endowed with a pseudo-metric $\|.\|_\mathcal{Q}$, which is the sup-norm metric with respect to the argument β and a pseudo-metric with respect to the other arguments. It follows under standard conditions that $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_\mathcal{Q} = o_p(1)$ where

$$h^\dagger(\beta) := (h_1^{\dagger'}(\beta),\dots,h_{R-1}^{\dagger'}(\beta))' \text{ and } h_r^\dagger(\beta) := \Pi(\Pi(\dots\Pi(\varphi_{R,\lambda}(O;\beta)|T_{R-1}(Z))\dots|T_{r+1}(Z))|T_r(Z)), \quad (16)$$

and where $\Pi(Y|X) := E[YX'](E[XX'])^-X$ is the population least squares projection of a random variable Y on another variable X. Additionally, conditions similar to Assumptions 1 and 2 in Hahn (1997) give $h^\dagger(\beta) = h^0(\beta)$.

The definition $h_r^0(\beta) := \varphi_{r,\lambda}(O;\beta)$ below (13), or the limit $h_r^\dagger(\beta)$ in (16) suggests that it is natural to restrict the $h_r(\beta)$'s in (13) such that $\mathcal{H}_r(\beta)$ is a collection of functions (in $\mathbb{R}^{d_m}$) of $T_r(Z)$ and β. Thus, $\mathcal{H}_r(\beta) \subseteq \mathcal{H}_{r+1}(\beta)$. However, to highlight the key feature of our framework we will ignore the nesting but take advantage of MAR in (3) and the linearity of $g(O;\beta,h(\beta))$ in $h(\beta)$ in (13), and alternatively define the nuisance parameter space as

$$\mathcal{H} := \{\mathcal{H}(\beta) = \mathcal{H}_1(\beta) \times \ldots \times \mathcal{H}_{R-1}(\beta) : \beta \in \mathcal{B}\} \tag{17}$$
$$\text{where } \mathcal{H}_r(\beta) := \left\{ h_r \in \mathcal{Q}(\beta) \left| E\left[\left(\frac{I(C \geq r)}{P(C \geq r|T_r(Z))} - \frac{I(C \geq r+1)}{P(C \geq r+1|T_{r+1}(Z))} \right) h_r \right] = 0 \right. \right\}.$$

for $r = 1, \ldots, R-1$. Functions of (Z,β) belong in $\mathcal{H}$.[10] Functions of (C,Z,β) need not belong in $\mathcal{H}$.[11] Thus, $\widehat{h}(\beta)$ in (15) generally does not belong in $\mathcal{H}$, but its probability limit $h^\dagger(\beta)$ in (16) (or $h^0(\beta)$) does. Other standard restrictions on $\mathcal{H}_r(\beta)$, that are not essential to describing the key feature of our estimation framework, are assumed later via high level conditions [see Chen et al. (2008), Cattaneo (2010), etc. for primitive conditions].

Key Feature: (2), (3), (13) and (17) give, for any $\beta \in \mathcal{B}$ and any $h(.) \in \mathcal{H}$ [where $h(.)$ need not be $h(\beta)$],

$$G(\beta, h(.)) = E[\varphi_{R,\lambda}(O;\beta)] = E[m(Z;\beta)|C \in \lambda]. \tag{18}$$

So $G(\beta,h(.))$ does not depend on $h(.) \in \mathcal{H}$. This is the key feature of our framework. Its implications are:

(F1) For any $h(.) \in \mathcal{H}$, we have $G(\beta_\lambda^0, h(.)) = 0$ by (1). Therefore, for any $\beta \in \mathcal{B}$ and any $h(.), \bar{h}(.) \in \mathcal{H}$, it follows that $G(\beta,h(.)) - G(\beta_\lambda^0, \bar{h}(.)) = 0 \iff E[m(Z;\beta)|C \in \lambda] - E[m(Z;\beta_\lambda^0)|C \in \lambda] = 0 \iff \beta = \beta_\lambda^0$.

(F2) The partial derivative of $G(\beta,h(\beta))$ with respect to β, denote it by $G_\beta(\beta,h(\beta))$, satisfies $G_\beta(\beta,h(\beta)) = M_\lambda(\beta) := \frac{\partial}{\partial \beta'} E[m(Z;\beta)|C \in \lambda]$, and it exists whenever $M_\lambda(\beta)$ exists. [See assumption (A3)].

(F3) $G(\beta,h(.)) - G(\beta,\bar{h}(.)) = 0$ for any $\beta \in \mathcal{B}$ and $h(.), \bar{h}(.) \in \mathcal{H}$. Thus, the pathwise derivative of $G(\beta,h(.))$ with respect to $h(.)$, denote it by $G_h(\beta,h(.))$, exists at all $h(.) \in \mathcal{H}$, in all directions $[\bar{h}(.) - h(.)]$ for $\{h(.) + \tau(\bar{h}(.) - h(.)) : \tau \in [0,1]\} \subset \mathcal{H}$, and satisfies $G_h(\beta,h(.))[\bar{h}(.) - h(.)] = 0$. [Also see Chen et al. (2003).]

(F1) helps to verify the well-separability (of the true β) assumption for consistent estimation of β_λ^0 by $\widehat{\beta}_\lambda(W_n)$. It is even stronger since it indicates that $\widehat{h}(\beta)$ need not converge in probability to the true $h^0(\beta)$ but can converge to any $h^\dagger(\beta) \in$ interior$(\mathcal{H})$ [see (16)] without affecting the consistency of $\widehat{\beta}_\lambda(W_n)$ for β_λ^0 [see Proposition 3]. **(F2)** simplifies the Jacobian formula (and its estimation) in the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$ since it implies that $G_\beta(\beta_\lambda^0, h(\beta_\lambda^0)) = M_\lambda$. Finally, while it was already clear from (F1) that the asymptotic orthogonality condition, Assumption N(c), in Andrews (1994) is satisfied following his equations (4.9)-(4.11) if $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{Q}} = o_p(1)$ and $h^\dagger(\beta) \in$ interior$(\mathcal{H})$; **(F3)** is still stated in a way that makes it more convenient for us to verify condition (4.1.4) in Theorem 4.1 of Chen (2007). (Proofs of the results stated below proceed by verifying the conditions in Chen et al. (2003) or Chen (2007).) Hence, the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$ is unaffected by the estimation of $h(\beta)$ even if $\widehat{h}(\beta)$ converges at a rate slower than $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{Q}} = o_p(n^{-1/4})$; for example, $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{Q}} = o_p(1)$ will

[10] To see this, note that using the relationship (a) $P(C \geq r|Z) = P(C \geq r|T_{r-1}(Z))$, stated immediately after introducing MAR in (3), it follows that (b) $P(C \geq r|T_r(Z)) = P(C \geq r|T_{r-1}(Z))$ (by the law of iterated expectations), and thus for $r = 1, \ldots, R$: $E[\, I(C \geq r)/\, P(C \geq r|T_r(Z))|\, Z] \overset{\text{by (a)}}{=} E[\, P(C \geq r|T_{r-1}(Z))/\, P(C \geq r|T_r(Z))] \overset{\text{by (b)}}{=} E[\, P(C \geq r|T_r(Z))/\, P(C \geq r|T_r(Z))] = 1$.

[11] Any function h_r of (C,Z) can be written as $h_r(C,Z) = \sum_{j=1}^R a_j(Z) I(C=j)$ where $a_j(Z) := h_r(C=j,Z)$ and, therefore by the same manipulation as in the last footnote, the key expectation in the definition of $\mathcal{H}_r(\beta)$ is $E[a_r(Z)]$, which is not necessarily zero.

suffice. See Remark 2(iii) in Chen et al. (2003) and Theorem 5 in Cattaneo (2010). The scenario is actually stronger here since we do not even require that $h^\dagger(\beta) = h^0(\beta)$, the truth [see Proposition 4]. Of course, semiparametric efficiency for $\widehat{\beta}_\lambda(W_n)$ requires that $h^\dagger(\beta_\lambda^0) = h^0(\beta_\lambda^0)$, but the rate of convergence of the consistent $\widehat{h}(\beta)$ is still of no consequence as far as the first-order asymptotic properties of GMM estimators are concerned [see Corollary 5]. Naturally, all these nice implications of (18) also provide flexibility in estimating the nuisance parameters – parametrically based on misspecified models or nonparametrically under less than satisfactory conditions.

4.2 Asymptotic properties of the GMM estimator

For notational simplicity we again follow Chen et al. (2003) and write $(\beta, h(\beta))$ as (β, h), unless confusing.

Proposition 3 *Let* (1), (3), *and assumptions (A1) and (A2) hold. Let* $\{W_n\}$ *be a* $d_m \times d_m$ *positive semidefinite matrix such that* $W_n = W + o_p(1)$ *where* W *is a constant positive definite matrix. Assume:*

(B1) $\|G_n(\widehat{\beta}_\lambda(W_n), \widehat{h})\|_{W_n} \leq \inf_{\beta\in\mathcal{B}} \|G_n(\beta, \widehat{h})\|_{W_n} + o_p(1)$ *where* $\mathcal{B}$ *is a compact subset of* $\mathbb{R}^{d_\beta}$*;*

(B2) $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{Q}} = o_p(1)$ *for some* $h^\dagger(\beta) \in$ *interior*$(\mathcal{H})$ *for all* β*, and* $h^\dagger(\beta)$ *not necessarily equal to* $h^0(\beta)$*;*

(B3) for all sequences of positive numbers $\{\delta_n\}$ *with* $\delta_n = o(1)$*,*

$$\sup_{\beta\in\mathcal{B}, \|h-h^\dagger(\beta)\|_{\mathcal{Q}}\leq\delta_n} \frac{\|G_n(\beta,h) - G(\beta,h)\|}{1 + \|G_n(\beta,h)\| + \|G(\beta,h)\|} = o_p(1).$$

Then $\widehat{\beta}_\lambda(W_n) - \beta_\lambda^0 = o_p(1)$*.*

Proposition 4 *Let* (1), (3), (17) *and assumptions A hold. Let* $\{W_n\}$ *be a* $d_m \times d_m$ *positive semidefinite matrix such that* $W_n = W + o_p(1)$ *where* W *is a constant positive definite matrix. Let* $\beta_\lambda^0 \in$ *interior*$(\mathcal{B})$ *and* $h^\dagger(\beta) \in$ *interior*$(\mathcal{H})$ *for all* β*, but* $h^\dagger(\beta)$ *not necessarily equal to* $h^0(\beta)$*. For a small* $\delta > 0$ *define the neighborhoods* $\mathcal{B}_\delta := \{\beta \in \mathcal{B} : \|\beta - \beta_\lambda^0\| \leq \delta\}$ *and* $\mathcal{H}_\delta := \{h \in \mathcal{H} : \|h - h^\dagger(\beta)\|_{\mathcal{Q}} \leq \delta\}$*. (Nothing changes if the sup-norm with respect to* β *in* $\|.\|_{\mathcal{Q}}$ *is alternatively defined to be taken locally over* $\beta \in \mathcal{B}_\delta$ *instead* $\beta \in \mathcal{B}$*; see Chen et al. (2003).) Let* $\widehat{\beta}_\lambda^0(W_n) - \beta_\lambda^0 = o_p(1)$ *and* $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{Q}} = o_p(1)$*. Assume:*

(C1) $\|G_n(\widehat{\beta}_\lambda(W_n), \widehat{h})\|_{W_n} \leq \inf_{\beta\in\mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\|_{W_n} + o_p(n^{-1/2})$*;*

(C2) $G_\beta(\beta, h^\dagger)$ *exists for* $\beta \in \mathcal{B}_\delta$ *and is continuous at* $\beta = \beta_\lambda^0$ *(*$G_\beta(\beta_\lambda^0, h^\dagger)$ *is full column rank by (A3) and (F2));*

(C3) for all sequences of positive numbers $\{\delta_n\}$ *with* $\delta_n = o(1)$*,*

$$\sup_{\beta\in\mathcal{B}_{\delta_n}, h\in\mathcal{H}_{\delta_n}} \frac{\|G_n(\beta,h) - G(\beta,h) - G_n(\beta_\lambda^0, h^\dagger)\|}{n^{-1/2} + \|G_n(\beta,h)\| + \|G(\beta,h)\|} = o_p(1);$$

(C4) $\sqrt{n}G_n(\beta_\lambda^0, h^\dagger) \xrightarrow{d} N(0, \Sigma)$ *where* $\Sigma := E\left[g(O; (\beta_\lambda^0, h^\dagger))g(O; (\beta_\lambda^0, h^\dagger))'\right]$ *is finite.*

Then, for $M_\lambda := M(\beta_\lambda^0)$ *defined in assumption (A3),* $R_\lambda := M_\lambda' W M_\lambda$ *and* $S_\lambda := M_\lambda' W \Sigma W M_\lambda$*,*

$$\sqrt{n}(\widehat{\beta}_\lambda(W_n) - \beta_\lambda^0) = -R_\lambda^{-1} M_\lambda' W \sqrt{n} G_n(\beta_\lambda^0, h^\dagger) + o_p(1) \xrightarrow{d} N\left(0, R_\lambda^{-1} S_\lambda R_\lambda^{-1}\right).$$

Remark: Propositions 3 and 4 respectively establish the consistency and asymptotic normality of the GMM estimator defined in (14). We focus on showing how the key feature (18) helps to satisfy some of the conditions

from Theorem 1 in Chen et al. (2003) and Theorem 4.1 in Chen (2007). We assume their other conditions. Through its condition (4.1.4), as opposed to (4.1.4)', Theorem 4.1 in Chen (2007) broadens the scope of Theorem 2 in Chen et al. (2003). This is useful to highlight that Propositions 3 and 4 (and the subsequent results) do not depend on the rate of convergence $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{Q}} = o_p(1)$. Cattaneo (2010) also only requires $\|\widehat{h}(\beta) - h^0(\beta)\|_{\mathcal{Q}} = o_p(1)$ in a related context, and works with less high level conditions. (By virtue of (11), (12) and Proposition 2, a similar treatment is possible here if needed.[12]) Importantly, we allow $h^\dagger(\beta) \neq h^0(\beta)$ to emphasize that consistency and asymptotic unbiasedness of $\widehat{\beta}_\lambda(W_n)$ are robust to the estimation of the nuisance parameters $h(\beta)$ parametrically under misspecification or nonparametrically under less than satisfactory conditions.[13] (See Rothe and Firpo (2016) and the references therein for insightful discussions in the context of the truly doubly-robust models.)

Thus, the theoretical results confirm the intuitions from our discussion of the implications of the key feature, with the final bit, i.e., on efficiency, to be confirmed by the following result on efficient GMM estimation.

Corollary 5 *Under the assumptions of Proposition 4:*

(1) if $W = \Sigma^{-1}$ *then*

$$\sqrt{n}(\widehat{\beta}_\lambda(W_n) - \beta_\lambda^0) = -\left(M_\lambda'\Sigma^{-1}M_\lambda\right)^{-1} M_\lambda'\Sigma^{-1}\sqrt{n}G_n(\beta_\lambda^0, h^\dagger) + o_p(1) \xrightarrow{d} N\left(0, \left(M_\lambda'\Sigma^{-1}M_\lambda\right)^{-1}\right);$$

(2) if, additionally, $h^\dagger(\beta_\lambda^0) = h^0(\beta_\lambda^0)$ *then* $\Sigma = V_\lambda$ *as defined in Proposition 1, and letting* $\widehat{\beta}_\lambda := \widehat{\beta}_\lambda(W_n)$*,*

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -\left(M_\lambda'V_\lambda^{-1}M_\lambda\right)^{-1} M_\lambda'V_\lambda^{-1}\sqrt{n}G_n(\beta_\lambda^0, h^0) + o_p(1) \xrightarrow{d} N\left(0, \Omega_\lambda = \left(M_\lambda'V_\lambda^{-1}M_\lambda\right)^{-1}\right),$$

i.e., by Proposition 1, the estimator $\widehat{\beta}_\lambda$ *becomes semiparametrically efficient.*

Estimation of asymptotic variance: Consistent estimation of M_λ is simplified due to (F2) because one could completely ignore the unknown nuisance parameters and obtain an estimator by taking analytical derivative (if it exists) or numerical derivative only for the first term of $G_n(\beta, h)$. Consistency of $\widehat{M}_\lambda(\beta)$ for $M_\lambda(\beta)$ with numerical derivatives follows by Theorem 7.4 in Newey and McFadden (1994). Also see Section 5.3 of Cattaneo (2010).[14]

Standard conditions e.g. $g(O_i; (\beta, h))$ is continuous with probability approaching one in a neighborhood $\mathcal{N}$ of $(\beta_\lambda^0, h^\dagger)$ and $E\left[\sup_{(\beta,h)\in\mathcal{N}} \|g(O_i; (\beta, h))\|^2\right] < \infty$ [see Lemma 4.3 in Newey and McFadden (1994)], ensure that for any $\beta = \beta_\lambda^0 + o_p(1)$ and $h(\beta)$ such that $\|h(\beta) - h^\dagger(\beta)\|_{\mathcal{Q}} = o_p(1)$ (suffices if the sup-norm in $\|.\|_{\mathcal{Q}}$ with respect to β is only local), the estimator $\widehat{V}_\lambda(\beta, h) := \frac{1}{n}\sum_{i=1}^n g(O_i; (\beta, h))g(O_i; (\beta, h))' = \Sigma + o_p(1)$. Thus, the estimator $\widehat{\Omega}_\lambda(\widehat{\beta}_\lambda, \widehat{h}) := \left(\widehat{M}_\lambda'(\widehat{\beta}_\lambda)\widehat{V}_\lambda^{-1}(\widehat{\beta}_\lambda, \widehat{h})\widehat{M}_\lambda(\widehat{\beta}_\lambda)\right)^{-1}$ is consistent for the asymptotic variance in Corollary 5(1). If $h^\dagger(\beta_\lambda^0) = h^0(\beta_\lambda^0)$ then $\Sigma = V_\lambda$, and now $\widehat{\Omega}_\lambda(\widehat{\beta}_\lambda, \widehat{h})$ will be consistent for the asymptotic variance Ω_λ in Corollary 5(2). Any consistent (for the appropriate limit) estimator $(\widetilde{\beta}, \widetilde{h})$ ensures consistency of all these quantities.

[12] Recall that the residuals from the successive projections in Proposition 2, where the $j(= 1, \dots R-1)$-th projection involves the moment condition (11) and the $r = 1, \dots j$ conditions from (12), can be seen as moment vectors in the spirit of that for the efficient influence function estimator in Cattaneo (2010) [see the proof of Proposition 2 for clarity on this statement]. Therefore, alternatively, one could impose the less high level assumptions following Theorems 3 and 5 of Cattaneo (2010), and use the induction argument as in the proof of Proposition 2 to arrive at the results in Propositions 3 and 4 in a rigorous manner as in Cattaneo (2010).

[13] While maintained mainly as a shortcut to the asymptotic orthogonality condition (aided by $\mathcal{H}$), the assumption $h^0(.) \in$ interior($\mathcal{H}$) is perhaps awkward since what is interior($\mathcal{H}$) is rather vague without more structure. Nevertheless, the nonstandard (since defined backwards) definition of the nuisance parameter space $\mathcal{H}$ in (17) is general enough to contain the probability limits of the commonly used parametric and nonparametric estimators $\widehat{h}(.)$, and helps to highlight the essential idea behind the robustness of $\widehat{\beta}_\lambda(W_n)$.

[14] Each of the other $R-1$ terms of $G_n(\beta, h)$ can also be used singly or jointly (even along with the first term above) in the same way to consistently estimate $M_\lambda(\beta)$. These terms involve nuisance parameters that are essentially (progressively more) smoothed version of the function to be differentiated, and hence the resulting estimator of $M_\lambda(\beta)$ may display better properties in finite samples.

4.3 One step from the IPW estimator gives efficiency

The presence of β in possibly highly nonlinear form in all the R additive terms of $G_n(\beta, \widehat{h}(\beta))$ should not ideally be a drawback for computational purpose. If the GMM estimator has closed form (Illustration 1 below) then this is not an issue. However, if there is no closed form expression (Illustration 2 below), one could start with an easy to compute $\sqrt{n}$-consistent estimator for β_λ^0 and then update it in one step to obtain an estimator with the same asymptotic distribution as the efficient estimator in Corollary 5. For example, an IPW estimator based on the complete sub-sample and with identity (or some simple) weighting matrix is relatively easy to compute:

$$\widetilde{\beta} := \arg\min_{\beta \in \mathcal{B}} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{I(C_i = R)}{P(C = R|T_R(Z_i))} \varphi_{R,\lambda}(O_i; \beta) \right\| \equiv \arg\min_{\beta \in \mathcal{B}} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{I(C_i = R)}{P(C = R|Z_i)} \frac{P(C \in \lambda|Z_i)}{\widehat{P}(C \in \lambda)} m(Z_i; \beta) \right\|. \quad (19)$$

It is consistent under the assumptions of Proposition 3 [see e.g. Wooldridge (2002)]. Built-in routines in standard statistical softwares can be directly used or slightly modified to obtain this estimator for a wide variety of the moment vector $m(Z; \beta)$ (e.g. Illustration 2). Now a one-step estimator of β_λ^0 can be obtained by updating $\widetilde{\beta}$ as:

$$\widehat{\beta}_{1\text{step}} = \widetilde{\beta} - \widehat{\Omega}_\lambda^{-1}(\widetilde{\beta}, \widehat{h}(\widetilde{\beta})) \widehat{M}_\lambda'(\widetilde{\beta}) \widehat{V}_\lambda^{-1}(\widetilde{\beta}, \widehat{h}(\widetilde{\beta})) G_n(\widetilde{\beta}, \widehat{h}(\widetilde{\beta})) \quad (20)$$

where $\widehat{h}$ is as in (15), and $\widehat{M}_\lambda(\widetilde{\beta})$, $\widehat{V}_\lambda(\widetilde{\beta}, \widehat{h}(\widetilde{\beta}))$ and $\widehat{\Omega}_\lambda(\widetilde{\beta}, \widehat{h}(\widetilde{\beta}))$, defined below Corollary 5, are consistent estimators for M_λ, V_λ and Ω_λ respectively under the conditions noted therein. Under the assumptions of Corollary 5(2) (with $W_n \xrightarrow{P} V_\lambda^{-1}$), it follows that this one-step estimator is efficient, i.e., $\sqrt{n}\left(\widehat{\beta}_{1\text{step}} - \widehat{\beta}_\lambda(W_n)\right) = o_p(1)$.

Illustration of the GMM estimator when $R = 3$

Let the moment vector consist of two variables y and X. Let X_c and X_e be mismeasured X, possibly dependent also on y. Let $Z_{(1)} = (y, X_c')'$, $Z_{(2)} = X_e$ and $Z_{(3)} = X$, a data structure that can be justified if y and X_c ("c" for cheap) are cheap to observe, X_e ("e" for expensive) is more expensive but still cheaper to observe than X (e.g. true consumption) itself [see e.g. Scenario 3 in Section 2]. Abstract from W_n by taking $d_m = d_\beta$. We consider two cases where the moment vector respectively corresponds to: (1) a linear regression giving a closed form expression for the estimator, and (2) a linear quantile regression where the estimator is computed in one step as in (20).

Illustration 1: Linear regression in the target population λ

Consider a moment vector of the form $m(Z; \beta) = X(y - X'\beta)$. For $i = 1, \ldots, n$, let $T_{ji} = T_j(Z_i)$ for $j = 1, 2, 3$, $a_{3i} = I(C = 3)/P(C = 3|T_{3i})$, $a_{2i} = I(C \geq 2)/P(C \geq 2|T_{2i}) - a_{3i}$, $a_{1i} = 1 - a_{2i}$, $q = P(C \in \lambda|T_3(Z))$ and $q_i = P(C \in \lambda|T_{3i})$. Simple computation gives a closed form expression for the estimator $\widehat{\beta}_\lambda$ in (14) as:

$$\widehat{\beta}_\lambda = \left(\sum_{i=1}^{n} \left\{ a_{3i} q_i X_i X_i' + a_{2i} \widehat{E}\left[qXX'|T_{2i}\right] + a_{1i} \widehat{E}\left[qXX'|T_{1i}\right] \right\} \right)^{-1} \sum_{i=1}^{n} \left\{ a_{3i} q_i X_i y_i + a_{2i} \widehat{E}\left[qXy|T_{2i}\right] + a_{1i} \widehat{E}\left[qXy|T_{1i}\right] \right\}$$

where $\widehat{E}$ denotes the estimated conditional expectation (see e.g. (15)). While one could factor out y_i from all three terms inside the last pair of braces, our experience is that estimating the conditional expectation e.g. $E\left[qXy|T_{2i}\right]$ directly instead of using the form $E\left[qX|T_{2i}\right] y_i$ leads to smaller variance (in small samples) of the estimator $\widehat{\beta}_\lambda$.

Illustration 2: Linear quantile regression in the target population λ

Consider a moment vector of the form $m(Z;\beta) = X\left(\tau - I(y - X'\beta < 0)\right)$ for some fixed $\tau \in (0,1)$. Unless redefined here, each notation used below is the same as that in Illustration 1. For any (β, h) define:

$$g(O_i;(\beta,h)) = a_{3i}q_i m(T_{3i};\beta) + [a_{2i} - a_{3i}]E[qm(T_3;\beta)|T_{2i}] + [1 - a_{2i}]E[qm(T_3;\beta)|T_{1i}],$$

and accordingly define $g(O_i;(\beta,\widehat{h}))$ and $G_n(\beta,\widehat{h})$ replacing the conditional expectations in $g(O_i;(\beta,h))$ by their estimators (see e.g. (15)). (The ignored common denominator $P(C \in \lambda)$ will be adjusted for in the final step; c.f. footnote 9.) Let $\widetilde{\beta}$ denote the inefficient but $\sqrt{n}$-consistent estimator of β_λ^0 obtained from (19) by using this particular choice of the moment vector $m(Z;\beta)$. It is simple to obtain $\widetilde{\beta}$ since commonly used statistical softwares provide built-in routine for weighted quantile regression which automatically gives the estimator with $(a_{3i}q_i/\sum_j a_{3j}q_j)_{i=1}^n$ as weights. Estimate M_λ where $M_\lambda(\beta) = -(\partial/\partial\beta')E\left[XI(y - X'\beta < 0)|C \in \lambda\right]$ possibly by using a post estimation command of the same built-in routine, or as discussed below Corollary 5. Therefore, since $d_m = d_\beta$, by using (20) we obtain the one-step estimator as: $\widehat{\beta}_{1\text{step}} = \widetilde{\beta} - \widehat{M}_\lambda^{-1}(\widetilde{\beta})G_n(\widetilde{\beta},\widehat{h}(\widetilde{\beta}))/\widehat{P}(C \in \lambda)$.

5 Simulation Study

Now we numerically study the benefit, if any, of using all the sub-samples for efficient estimation of β_λ. For a precise measure of benefit, define the efficiency loss associated with the j-th element from estimating β_λ based on a collection of sub-samples denoted by s instead of another collection of sub-samples denoted by s' as:

$$\text{Loss}(\beta_{\lambda,j};s,s') = \lim_{n\to\infty} \frac{\frac{1}{n_{\{s\}}}\text{Avar}(\widehat{\beta}_{\lambda,j}^{s}) - \frac{1}{n_{\{s'\}}}\text{Avar}(\widehat{\beta}_{\lambda,j}^{s'})}{\frac{1}{n_{\{s'\}}}\text{Avar}(\widehat{\beta}_{\lambda,j}^{s'})} \text{ where } \lambda, s, s' = \Lambda \text{ and } j = 1,\ldots,d_\beta. \tag{21}$$

$n_{\{s\}} := \sum_{r\in s} n_r = \sum_{r\in s}\sum_{i=1}^n I(C_i = r)$ and $n_{\{s'\}} := \sum_{r\in s'} n_r = \sum_{r\in s'}\sum_{i=1}^n I(C_i = r)$ are the size of the combined sub-samples in s and s' respectively. For $j = 1,\ldots,d_\beta$ and $l = s, s'$, $\widehat{\beta}_{\lambda,j}^{l}$ is the j-th element of $\widehat{\beta}_\lambda^l$, the efficient GMM estimator of β_λ based on the sub-samples in l. Avar is the asymptotic variance. These estimators and the Avar's are computed ignoring the existence of the other sub-samples.[15] Thus, the estimators not using all the sub-samples are not penalized for the sub-optimal use of (available) information since they are actually efficient if the sub-samples they use were the only available sub-samples. Letting s be included in s', the loss in (21) thus reflects the information brought in by the additional sub-samples that are included in s' but not in s.

Analytical expressions for this loss under INDEP in (5) and CMAR in (4) are intuitive, and are provided in Corollaries 10 and 11 in Supplemental Appendix B2 to serve as a rough reference to put our simulation results into perspective. Analogous expressions are also provided there in Corollary 12 without imposing the assumption of planned incompleteness in (6). As noted earlier while discussing the importance of considering $R > 2$, relaxing (6) ends up causing the results to be qualitatively different for certain target populations λ.

For identification (see (2)) we always include $\{R\}$ in s, s'. Unless $\lambda = \mathcal{C}$, we include λ in s, s' as a convention.

[15] For example, if $\lambda = \{1\}$ and $s = \{1, R\}$, then we replace $P(C \in \lambda|T_R(Z))$ and $P(C \in \lambda)$ in the result of Proposition 1 by $P(C \in \lambda|T_R(Z), C \in \{1,R\}) = P(C \in \lambda|T_R(Z))/P(C \in \{1,R\}|T_R(Z))$ and $P(C \in \lambda|C \in \{1,R\}) = P(C \in \lambda)/P(C \in \{1,R\})$ respectively, as if there exist only two sub-samples 1 and R. We employ the substitution pattern from multinomial/conditional logit.

5.1 Simulation Design

The framework of Illustration 1 (Section 4.3) is employed. Accordingly, we take $Z_{(1)} = (y, X_c)'$, $Z_{(2)} = X_e$ and $Z_{(3)} = X$. We consider an arbitrary data generating process and draw n i.i.d. copies of these variables as follows:

$$y_i = \alpha + \delta X_i + \epsilon_i, \quad X_{ci} = X_i + I(y_i > 0)\sqrt{2}\epsilon_{ci}, \quad X_{ei} = X_i + I(y_i > 0)\epsilon_{ei}$$

where $\epsilon_i, \epsilon_{ci}, \epsilon_{ei}, X_i$ are mutually independent and i.i.d. $N(0,1)$ for all $i = 1, \ldots, n$. While $E[X] = E[X_e] = E[X_c]$, X_e is a less variable measure of X. Therefore, to justify the data structure it is possibly reasonable to maintain that X_e is more expensive to observe than X_c but less expensive to observe than X. We take $\alpha = \delta = 1$.

We generate the variable $C_i \in \mathcal{C} := \{1, 2, 3\}$ as i.i.d. copies for $i = 1, \ldots, n$ such that:

$$P(C = 1|Z_i) = F_{t_1}(\gamma_c(X_{ci} + y_i - 1)), \quad P(C = 2|Z_i) = (1 - P(C = 1|Z_i))(1 - F_{t_1}(\gamma_e X_{ei} + \gamma_c(X_{ci} + y_i - 2))),$$

and $P(C = 3|Z_i) = 1 - P(C = 1|Z_i) - P(C = 2|Z_i)$ where $F_{t_1}(a)$ is the cumulative distribution function of a t_1-distributed random variable evaluated at $a \in \mathbb{R}$. We consider the selection mechanisms MAR in (3), CMAR in (4) and INDEP in (5) by taking $\gamma_c = \gamma_e = .25$ for MAR, $\gamma_c = .25, \gamma_e = 0$ for CMAR, and $\gamma_c = \gamma_e = 0$ for INDEP.

The parameters of interest, i.e., the Intercept (β_1) and Slope (β_2) are defined by (1) with the moment vector: $m(Z; \beta) = [y - \beta_1 - \beta_2 X, \quad X(y - \beta_1 - \beta_2 X)]'$ similar to that in Illustration 1 (Section 4.3). The true values of $(\beta_1, \beta_2)'$ under INDEP are $(1, 1)'$, i.e., $(\alpha, \delta)'$, always. The same holds under CMAR and MAR when $\lambda = \{1, 2, 3\}$. However, otherwise it is difficult to analytically obtain the true values, and given that the study of bias is not focus of this paper, in those cases we take the following, listed in Table 1, as the (roughly) true values.

Target	CMAR Sampling					MAR Sampling				
λ	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,3\}$	$\{2,3\}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,3\}$	$\{2,3\}$
Intercept	1.1375	0.7602	1.0087	1.1006	0.8652	1.1375	0.7624	0.9991	1.09856	0.8652
Slope	0.9630	0.9318	0.9562	0.9675	0.9685	0.9630	0.9239	0.9473	0.9628	0.9685

Table 1: Obtained as average over 10,000 Monte Carlo trials of ordinary least squares estimates of Intercept and Slope from the regression of y on X based on the appropriate sub-sample ($s = \lambda$) when the total sample size $n = 1$ million.

Now the sub-samples are made incomplete by deleting X_i if $C_i \neq 3$ and X_{ei} if $C_i = 1$ for $i = 1, \ldots, n$. Averaged over 10,000 Monte Carlo trials, $n_1/n \approx .5$ and $n_2/n \approx .31$ where $n_j = \sum_{i=1}^{n} I(C_i = j)$ for $j = 1, 2, 3$.

We consider $n = 1000, 2000, 5000$, which is not necessarily impractical (e.g. $n = 4000$ in Beegle et al. (2012)).

The nuisance parameters $h_r(.)$'s in the GMM estimators are estimated using the series estimator in (15). We always use cubic polynomials of all the elements of $(1, T_r(Z)')'$ for $\Upsilon_{K_r}(T_r(Z))$ for $r = 1, 2$ *irrespective* of n. Thus, one could alternatively consider the GMM estimators to be parametric in the sense of Ackerberg et al. (2012).

5.2 Simulation Results

Tables 2 and 3 list the estimated loss (in percent) defined in (21) for various s with respect to $s' = \{1, 2, 3\}$ (the one we recommend) under INDEP, and CMAR and MAR respectively. If all the sub-samples contained the same variables then these losses should more or less reflect the smaller than n size of the collection of sub-samples in s. For example, the first row of Table 2 would be $100 \times (1/n_3 - 1/n)/(1/n) \approx 426$, and similarly the second and third rows would be approximately 27 and 100 respectively. The actual loss will invariably be smaller in the

first and third rows because the units in the additional sub-samples in $s' = \{1,2,3\}$ that are not in $s = \{3\}$ and $s = \{2,3\}$, i.e., the sub-samples $\{1,2\}$ and $\{1\}$ respectively, are uniformly worse in terms of information content than those in s. This is however not true for the second row since the extra sub-sample in s' is $\{2\}$, and an unit in it is more informative than that in the sub-sample $\{1\}$ but less than that in the other sub-sample $\{3\}$ in s. Thus, it is not clear a priori in this case, i.e., $s = \{1,3\}$, if the actual loss will be smaller or larger. All these intuitions are clearly reflected in the tables, not only for INDEP (Table 2) but also for CMAR and MAR (Table 3).

Target Popln.	Used Sample	INDEP Sampling					
		Intercept			Slope		
λ	s	$n = 1000$	$n = 2000$	$n = 5000$	$n = 1000$	$n = 2000$	$n = 5000$
$\{1,2,3\}$	$\{3\}$	158	156	155	107	103	100
$\{1,2,3\}$	$\{1,3\}$	33	32	32	24	23	22
$\{1,2,3\}$	$\{2,3\}$	33	32	33	22	21	21

Table 2: Reported are estimated Loss$(\beta_{\lambda,j}; s, \{1,2,3\})$ (in percent) defined in (21) for $j = 1$ (Intercept) and $j = 2$ (Slope). The results are based on the analytically estimated Avar averaged over 10,000 Monte Carlo trials.

Target	Used	CMAR Sampling						MAR Sampling					
		Intercept			Slope			Intercept			Slope		
Popln.	Sample	n			n			n			n		
λ	s	1000	2000	5000	1000	2000	5000	1000	2000	5000	1000	2000	5000
$\{1,2,3\}$	$\{3\}$	159	158	157	128	125	122	169	167	164	140	134	130
$\{1,2,3\}$	$\{1,3\}$	39	39	38	42	39	38	42	42	42	47	43	41
$\{1,2,3\}$	$\{2,3\}$	44	42	42	45	45	45	45	43	42	49	47	47
$\{1\}$	$\{3\}$	129	127	124	109	107	106	137	133	129	110	109	106
$\{1\}$	$\{1,3\}$	27	26	25	18	17	17	31	30	30	20	19	18
$\{2\}$	$\{3\}$	174	173	169	131	134	133	175	172	168	152	149	146
$\{2\}$	$\{2,3\}$	25	24	23	4	5	5	23	22	20	4	4	3
$\{3\}$	$\{3\}$	156	155	154	107	105	103	153	153	152	101	99	97
$\{3\}$	$\{1,3\}$	37	36	37	32	30	28	36	36	35	27	24	23
$\{3\}$	$\{2,3\}$	41	40	40	33	32	32	41	40	40	37	34	33
$\{1,3\}$	$\{3\}$	138	136	133	111	108	106	142	139	137	110	108	104
$\{1,3\}$	$\{1,3\}$	30	29	29	22	20	19	32	32	31	22	21	19
$\{2,3\}$	$\{3\}$	176	176	174	133	132	131	185	185	182	152	148	145
$\{2,3\}$	$\{2,3\}$	35	35	35	16	16	16	37	36	35	17	16	16

Table 3: Reported are estimated Loss$(\beta_{\lambda,j}; s, \{1,2,3\})$ (in percent) defined in (21) for $j = 1$ (Intercept) and $j = 2$ (Slope). The results are based on the analytically estimated Avar averaged over 10,000 Monte Carlo trials.

There are cases like $\lambda = \{2\}$, $s = \{2,3\}$ under CMAR and MAR sampling where the loss for the Slope estimator is minimal and close to zero, and this is in spite of the fact that the estimator based on $s = \{2,3\}$ uses roughly half the number of observations used by the estimator based on $s' = \{1,2,3\}$. Note that, while Corollary 11(b) in Supplemental Appendix B2 implies a zero loss only if $E[m(Z;\beta^0_\lambda)|Z_1] = 0$, which is not true under our design, it does not rule out small losses either. The loss can, however, be quite substantial in many cases. The loss would be even larger if we had not modified the GMM estimation according to what we stated below (21); and hence this shows the obvious benefit of using all the sub-samples for the estimation of the parameter of interest.

Supplemental Appendix C reports simulation evidence of good finite-sample properties of the GMM estimator used here under all the cases considered and under the same simulation design. This lends credibility to the above numerical results on efficiency loss. Overall, the simulation study also makes a case for the collection of planned incomplete data for more sample units if budget constraints do not allow the complete data collection for all.

Appendix: Proofs of the results in Sections 3 and 4

For convenient reference, we list two relationships that are byproducts of MAR in (3) and are repeatedly used in the proofs. For any $r = 1, \ldots, R$ and function $\nu(T_r(Z))$ such that $E|\nu(T_r(Z))| < \infty$:

$$P(C \geq r|T_r(Z) = P(C \geq r|Z) = P(C \geq r|T_{r-1}(Z)) \tag{22}$$

$$E\left[\frac{I(C \geq r)}{P(C \geq r|T_r(Z))}\nu(T_r(Z))\right] = E[\nu(T_r(Z)]. \tag{23}$$

(Convention: $T_0(Z)$ is constant.) (22)-(23) were already explained in the main text, respectively, immediately after introducing MAR in (3) and in footnote 11 following the definition of the nuisance tangent space in (17).

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. f and F denote the density and distribution functions, and the concerned random variables are specified inside parentheses. $L_0^2(F)$ denotes the space of mean-zero, square integrable functions with respect to F.

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 can be expressed in terms of the full data $(C, Z')'$ as

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

θ_0 is 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

$$S_\theta(O) = s_\theta(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$$

where $s_\theta(Z_{(1)}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_{(1)})$ and $s_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$. We will omit the subscript θ from the quantities evaluated at $\theta = \theta_0$. 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

$$\mathcal{T} := a_1(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) a_r(Z_{(1)}, \ldots, Z_{(r)}), \tag{24}$$

where $a_1(Z_{(1)}) \in L_0^2(F(Z_{(1)}))$ and $a_r(Z_{(1)}, \ldots, Z_{(r)}) \in L_0^2(F(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}))$.

STEP - 2: The moment conditions in (1) for a given $\lambda \in \Lambda$ are equivalent to the requirement that for any $d_\beta \times d_m$ matrix A, the following just-identified system of moment conditions holds:

$$AE[m(Z;\beta_\lambda^0)|C \in \lambda] = AE\left[\frac{P(C \in \lambda|Z)}{P(C \in \lambda)} \frac{I(C = R)}{P(C = R|Z)} m(Z;\beta_\lambda^0)\right] = 0.$$

where the first equality follows from (2). Differentiating with respect to θ under the integral, and noting that $P(C \in \lambda|Z)$ (which is known) does not depend on θ but $P(C \in \lambda)$ (which is unknown) does, we obtain by using

(1) and (3) that

$$0 = AM_\lambda \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} + AE\left[m(Z;\beta_\lambda^0) \left\{ s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})' \right\} \middle| C \in \lambda \right].$$

Taking a full row rank A along with assumption (A3) gives

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} = -(AM_\lambda)^{-1} AE\left[m(Z;\beta_\lambda^0) \left\{ s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})' \right\} \middle| C \in \lambda \right].$$

Therefore, for the given A, any regular estimator for β_λ^0 will be asymptotically linear with influence function of the form $-(AM_\lambda)^{-1}Am(Z;\beta_\lambda^0)$. Now, for the given A, we can obtain the projection of this influence function on to the tangent set $\mathcal{T}$ in (24) if we can find a $\psi(A,O) \in \mathcal{T}$ such that

$$E[\psi(A,O)S(O)'] = \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'}. \tag{25}$$

Let us conjecture that $\psi(A,O) = -(AM_\lambda)^{-1}A\varphi_\lambda(O;\beta_\lambda^0)$, and then verify (25) by equivalently showing that

$$E[\varphi_\lambda(O;\beta_\lambda^0)S(O)'] = E\left[m(Z;\beta_\lambda^0) \left\{ s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})' \right\} \middle| C \in \lambda \right].$$

Consider the LHS and, in accordance with the partition of $\varphi_\lambda(O)$ (we work with the alternative specification in (10) for convenience), write it as $\sum_{q=1}^{R} B_q$ where

$$B_1 := E\left[\varphi_{1,\lambda}(O;\beta_\lambda^0)S(O)'\right], \text{ and } B_q := E\left[\frac{I(C \geq q)}{P(C \geq q|T_q(Z))} \left[\varphi_{q,\lambda}(O;\beta_\lambda^0) - \varphi_{q-1,\lambda}(O;\beta_\lambda^0)\right] S(O)' \right] \text{ for } q = 2,\ldots,R.$$

To avoid notational clutter, in the rest of STEP-2 we will write $m(Z;\beta_\lambda^0)$ as m; $T_q(Z)$ as T_q; $\varphi_{q,\lambda}(O;\beta_\lambda^0)$ as $\varphi_{q,\lambda}$ for $q = 1,\ldots,R$; and also write $s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r = 2,\ldots,R$. Now, note that

$$B_1 = E\left[E\left[\frac{P(C \in \lambda|T_R)}{P(C \in \lambda)} m \middle| T_1 \right] I(C \geq r) s(Z_{(1)})' \right] + \sum_{r=2}^{R} E\left[E\left[\frac{P(C \in \lambda|T_R)}{P(C \in \lambda)} m \middle| T_1 \right] I(C \geq r) s(Z_{(r)}|T_{r-1})' \right].$$

However, since for $r > 1$ we know that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$, it follows by MAR in (3) that

$$\begin{aligned} & \sum_{r=2}^{R} E\left[E\left[\frac{P(C \in \lambda|T_R)}{P(C \in \lambda)} m \middle| T_1 \right] I(C \geq r) s(Z_{(r)}|T_{r-1})' \right] \\ = & \sum_{r=2}^{R} E\left[E\left[\frac{P(C \in \lambda|T_R)}{P(C \in \lambda)} m \middle| T_1 \right] (1 - I(C \leq r-1)) s(Z_{(r)}|T_{r-1})' \right] = 0. \end{aligned}$$

This is the first observation. On the other hand, since $T_1 := Z_{(1)}$, we have the second observation that

$$E\left[E\left[\frac{P(C \in \lambda|T_R)}{P(C \in \lambda)} m \middle| T_1 \right] s(Z_{(1)})' \right] = E\left[\frac{P(C \in \lambda|T_R)}{P(C \in \lambda)} m s(Z_{(1)})' \right] = E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} m s(Z_{(1)})' \right].$$

Combining the two observations it follows that $B_1 = E[ms(Z_{(1)})'|C \in \lambda]$.

Now we consider B_q. (3) gives for $q = 2,\ldots,R$:

$$B_q = \sum_{r=1}^{q-1} E\left[\frac{I(C \geq q)}{P(C \geq q|T_q)} (\varphi_{q,\lambda} - \varphi_{q-1,\lambda}) s(Z_{(r)}|T_{r-1})' \right] + \sum_{r=q}^{R} E\left[\frac{I(C \geq r)}{P(C \geq q|T_q)} (\varphi_{q,\lambda} - \varphi_{q-1,\lambda}) s(Z_{(r)}|T_{r-1})' \right].$$

Since $E[\varphi_{q,\lambda}|T_{q-1}] = \varphi_{q-1,\lambda}$, it follows by conditioning on T_{q-1} and from (23) that the first term on the RHS is 0. On the other hand, (22) and the fact that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ imply that the second term

$$\sum_{r=q}^{R} E\left[\frac{1-I(C\leq r-1)}{1-P(C\leq q-1|T_{q-1})}(\varphi_{q,\lambda}-\varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] = E[\varphi_{q,\lambda}s(Z_{(q)}|T_{q-1})'] = E[ms(Z_{(q)}|T_{q-1})'|C\in\lambda].$$

Therefore, for $q = 2,\ldots,R$ we have $B_q = E[ms(Z_{(q)}|T_{q-1})'|C\in\lambda]$, combining which with B_1 verifies (25).

That $\psi(A,O)\in\mathcal{T}$ follows from matching terms as follows. (i) $-(AM_\lambda)^{-1}A\varphi_{1,\lambda}$ is only a function of $T_1 := Z_{(1)}$ and $E[\varphi_{1,\lambda}] = 0$ and, hence, satisfies the properties of $a_1(Z_{(1)})$ in (24). (ii) The r-th term ($r = 2,\ldots,R$, without the multiplier $I(C\geq r)$) on the RHS of $\psi(A,O)$ can be written as

$$-\frac{1}{P(C\geq r|T_r)}(AM_\lambda)^{-1}A\left[\varphi_{r,\lambda}-\varphi_{r-1,\lambda}\right] = -\frac{1}{1-P(C\leq r-1|T_{r-1})}(AM_\lambda)^{-1}A\left[\varphi_{r,\lambda}-\varphi_{r-1,\lambda}\right]$$

by (3) [also see (22)]. Hence, by definition of φ_r, taking expectation of the RHS of the above equation conditional on $T_{r-1} := (Z_{(1)},\ldots,Z_{(r-1)})'$ gives 0. Therefore, this term is only a function of T_r that is also in $L_0^2(F(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)}))$, and hence satisfies the properties of $a_r(Z_{(1)},\ldots,Z_{(r)})$ in (24).

STEP - 3: For a given A, we verified that the projection of the influence function $-(AM_\lambda)^{-1}Am(Z;\beta_\lambda^0)$ on to the tangent set $\mathcal{T}$ is $\psi(A,O) := -(AM_\lambda)^{-1}A\varphi_\lambda(O;\beta_\lambda^0)$. The asymptotic variance of $\psi(A,O)$ is $(AM_\lambda)^{-1}A\,V_\lambda\,A'(AM_\lambda)^{-1'}$ where $V_\lambda := Var\left(\varphi_\lambda(O;\beta_\lambda^0)\right) = E[\varphi_\lambda(O;\beta_\lambda^0)\varphi_\lambda(O;\beta_\lambda^0)']$. Therefore, the efficient influence function is obtained by minimizing the above variance with respect to A. Standard arguments give that the minimizer is $A_* = M_\lambda' V_\lambda^{-1}$. Hence, the variance lower bound is $\Omega_\lambda := (M_\lambda' V_\lambda^{-1} M_\lambda)^{-1}$ and the efficient influence function with variance equal to the variance lower bound is $\psi(A_*,O) = -\Omega_\lambda M_\lambda' V_\lambda^{-1}\varphi_\lambda(O;\beta_\lambda^0)$. ■

Proof of Proposition 2: Let us start with $r = 1$, i.e., the residual from the projection, $\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)$, inside the innermost parenthesis on the RHS. We will also consider $r = 2$ so that the pattern in the form of the residuals from the successive projections inside the first few innermost parentheses is clear to the reader. Then we apply induction arguments. For simplicity write $\varphi_{R,\lambda}(O;\beta)$ as $\varphi_{R,\lambda}$ and $T_r(Z)$ as T_r.

First, note that direct computation and (3) along with (22) give

$$\begin{aligned}
\text{Proj}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right) &= \left[\frac{I(C=R)}{P(C=R|T_R)} - \frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})}\right]E[\varphi_{R,\lambda}|T_{r-1}] \\
\Rightarrow \overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right) &= \frac{I(C=R)}{P(C=R|T_R)}\underbrace{(\varphi_{R,\lambda}-E[\varphi_{R,\lambda}|T_{R-1}])} + \frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})}E[\varphi_{R,\lambda}|T_{R-1}].
\end{aligned}$$

Consider the under-braced part in the RHS of the expression for $\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)$. Using $T_{R-1}\setminus T_{R-2} = Z_{R-1}$ and (3), note that $E\left[(\varphi_{R,\lambda}-E[\varphi_{R,,\lambda}|T_{R-1}])\,\phi_{R-2}|T_{R-2}\right]$ is a $d_m\times 2$ matrix of zeros, and hence has no contribution in the successive projections. (Terms with no contribution in the successive projections are marked by under-braces in this proof.) On the other hand,

$$E\left[\frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})}E[\varphi_{R,\lambda}|T_{R-1}]\phi_{R-2}\middle|T_{R-2}\right] = \frac{P(C=R-2|T_{R-2})}{P(C\geq R-2|T_{R-2})}E[\varphi_{R,\lambda}|T_{R-2}].$$

Thus, similar computation as above (and the use of (22)) gives for $r = 2$:

$$\begin{aligned}
& \text{Proj}_{T_{R-2}}\left(\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\Big|\,\phi_{R-2}\right) = \left[\frac{I(C \geq R-1)}{P(C \geq R-1|T_{R-1})} - \frac{I(C \geq R-2)}{P(C \geq R-2|T_{R-2})}\right] E[\varphi_{R,\lambda}|T_{R-2}] \\
\Rightarrow \quad & \overline{\text{Proj}}_{T_{R-2}}\left(\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\Big|\,\phi_{R-2}\right) \\
= \quad & \sum_{s=0}^{1} \frac{I(C \geq R-s)}{P(C \geq R-s|T_{R-s})} \underbrace{\left(E[\varphi_{R,\lambda}|T_{R-s}] - E[\varphi_{R,\lambda}|T_{R-s-1}]\right)} + \frac{I(C \geq R-2)}{P(C \geq R-2|T_{R-2})} E[\varphi_{R,\lambda}|T_{R-2}].
\end{aligned}$$

To prove the proposition by induction let us assume that the following holds for a general $r \in \{2, \ldots, R-2\}$:

$$\begin{aligned}
& \overline{\text{Proj}}_{T_{R-r}}\left(\ldots \overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\ldots\Big|\,\phi_{R-r}\right) \\
= \quad & \sum_{s=0}^{r-1} \frac{I(C \geq R-s)}{P(C \geq R-s|T_{R-s})} \underbrace{\left(E[\varphi_{R,\lambda}|T_{R-s}] - E[\varphi_{R,\lambda}|T_{R-s-1}]\right)} + \frac{I(C \geq R-r)}{P(C \geq R-r|T_{R-r})} E[\varphi_{R,\lambda}|T_{R-r}].
\end{aligned}$$

Now, once again using (22), note that

$$\begin{aligned}
& E[\phi_{R-r-1}^2|T_{R-r-1}] = \frac{P(C \geq R-r|T_{R-r})P(C = R-r-1|T_{R-r-1})}{P(C \geq R-r-1|T_{R-r-1})}, \text{ and} \\
& E[\overline{\text{Proj}}_{T_{R-r}}\left(\ldots \overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\ldots\Big|\,\phi_{R-r}\right)\phi_{R-r-1}|T_{R-r-1}] = \frac{P(C = R-r-1|T_{R-r-1})}{P(C \geq R-r-1|T_{R-r-1})}[\varphi_{R,\lambda}|T_{R-r-1}].
\end{aligned}$$

Hence the proof follows by indiction since the form is also valid for $r+1$, i.e.,

$$\begin{aligned}
& \overline{\text{Proj}}_{T_{R-r-1}}\left(\ldots \overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\ldots\Big|\,\phi_{R-r-1}\right) \\
= \quad & \sum_{s=0}^{r} \frac{I(C \geq R-s)}{P(C \geq R-s|T_{R-s})} \left(E[\varphi_{R,\lambda}|T_{R-s}] - E[\varphi_{R,\lambda}|T_{R-s-1}]\right) + \frac{I(C \geq R-r-1)}{P(C \geq R-r-1|T_{R-r-1})} E[\varphi_{R,\lambda}|T_{R-r-1}]. \ \blacksquare
\end{aligned}$$

Detailed proofs for Proposition 3 and 4 are presented in an older version Chaudhuri (2016).

Proof of Proposition 3: The proof follows by verifying the conditions of Theorem 1 in Chen et al. (2003). (B1) assumes their condition (1.1). Given (F1), our (B2) and (B3) assume their conditions (1.4) and (1.5) respectively. Their condition (1.2) holds by (1) and (18) [see (F1)]. Their condition (1.3) holds by (18) [see (F3)]. ■

Proof of Proposition 4: The proof follows by verifying the conditions of Theorem 4.1 in Chen (2007). (C1) assumes condition (4.1.1). Condition (4.1.2) (i) is satisfied by (C2) [see (A3) and (F2)], while (4.1.2)(ii) is satisfied by the same logic and by (A3), and because W is positive definite. (4.1.3) holds by (18) [see (F3)]. (4.1.5) and (4.1.6) are assumed in (C3) and (C4). Finally, to verify (4.1.4), i.e., $\sqrt{n}\|G(\beta, \widehat{h}) - G(\beta, h^\dagger) - G_h(\beta, h^\dagger)[\widehat{h}(\beta) - h^\dagger(\beta)]\| = o_p(1)$ for any $\beta \to \beta_\lambda^0$, we note that LHS $= \sqrt{n}\|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\|$ by (F3). Then we follow equations (4.9)-(4.11) in Andrews (1994) to note that the RHS in the last equation is by construction less than or equal to

$$\sup_{h \in \mathcal{H}} \sqrt{n}\|G(\beta, h) - G(\beta, h^\dagger)\| \times 1(\|\widehat{h}(b) - h^\dagger(b)\|_{\mathcal{Q}} \leq \epsilon) + \sqrt{n}\|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| \times 1(\|\widehat{h}(b) - h^\dagger(b)\|_{\mathcal{Q}} > \epsilon)$$

for some $\epsilon > 0$ since $\|\widehat{h}(b) - h^\dagger(b)\|_{\mathcal{Q}} = o_p(1)$ and $h^\dagger(.) \in \text{interior}(\mathcal{H})$. In the above expression, the first term is 0 by (18) and (F1), whereas the second term is $o_p(1)$. This verifies (4.1.4) and, therefore, completes the proof. ■

Proof of Corollary 5: (1) Standard, and hence omitted. (2) Follows since $g(O; \beta, h^0(\beta)) = \varphi_\lambda(O; \beta)$ in (8). ■

References

Abrevaya, J. and Donald, S. G. (2017). A GMM approach for dealing with missing data on regressors and instruments. Forthcoming in Review of Economics and Statistics.

Ackerberg, D., Chen, X., and Hahn, J. (2012). A Practical Asymptotic Variance Estimator For Two-Step Semiparametric Estimators. *The Review of Economics and Statistics*, 94: 481–498.

Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170: 442–457.

Andrews, D. W. K. (1994). Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. *Econometrica*, 62: 43–72.

Beegle, K., Weerdt, J. D., Friedman, J., and Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics*, pages 3 – 18.

Brown, B. and Newey, W. (1998). Efficient Semiparametric Estimation of Expectations. *Econometrica*, 66: 453–464.

Carroll, R., Ruppert, D., and Stefanski, L. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, New York.

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

Chamberlain, G. (1992). Comment: Sequential Moment Restrictions In Panel Data. *Journal of Business and Economic Statistics*, 10: 20–26.

Chaudhuri, S. (2016). 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. (2007). Large Sample Sieve Estimation Of Semi-Nonparametric Models. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume VIB, chapter 76, pages 5550–5632. Elsevier Science Publisher.

Chen, X., Hong, H., and Tamer, E. (2005). Measurement Error Models with Auxiliary Data. *Review of Economic Studies*, 72: 343–366.

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.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, 3 edition.

Devereux, P. J. and Tripathi, G. (2009). Optimally combining censored and uncensored datasets. *Journal of Econometrics*, 151: 17–32.

Graham, B. S. (2011). Efficiency Bounds for Missing Data Models with Semiparametric Restrictions. *Econometrica*, 79: 437 – 452.

Graham, B. S., Pinto, C., and Egel, D. (2012). Inverse Probability Tilting for Moment Condition Models with Missing Data. *Review of Economic Studies*, 79: 1053 – 1079.

Graham, B. S., Pinto, C. C. D. X., and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting. *Journal of Business and Economic Statistics*, 34: 288–301.

Hahn, J. (1997). Efficient estimation of panel data models with sequential moment restrictions. *Journal of Econometrics*, 79: 1–21.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66: 315–331.

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.

Holt, C. A. and Laury, S. K. (2002). Risk Aversion and Incentive Effects. *The American Economic Review*, 92: 1644–1655.

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

Ichimura, I. and Martinez-Sanchis, E. (2005). Identification and Estimation of GMM Models by Combining Two Data Sets. Working Paper.

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

Lee, A. J., Scott, A. J., and Wild, C. J. (2012). Efficient estimation in multi-phase case-control studies. *Biometrika*, 97: 361–374.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ.

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

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99: 210–221.

McKenzie, D. and Rosenzweig, M. (2012). Preface for symposium on measurement and survey design. *Journal of Development Economics*, 98: 1–2.

Muris, C. (2016). Efficient GMM Estimation with a General Missing Data Pattern. Technical report, Simon Frasier University.

Newey, W. (1997). Convergence rates and asymptotic normality of series estimators. *Journal of Econometrics*, 79: 147–168.

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.

Reilly, M. (1996). Optimal Sampling Strategies for Two-Stage Studies. *American Journal of Epidemiology*, 143: 92–100.

Ridder, G. and Moffitt, R. (2007). The Econometrics of Data Combination. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6B, chapter 75, pages 5470–5547. Elsevier Science Publisher.

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

Robins, J. M. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. In Lin, D. Y. and Heagerty, P., editors, *Proceedings of the Second Seattle Symposium on Biostatistics*. New York: Springer.

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.

Song, R., Zhou, H., and Kosorok, M. R. (2009). On semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, 96: 1–8.

Tripathi, G. (2009). Optimally combining censored and uncensored datasets. *Journal of Econometrics*, 151: 17–32.

Tripathi, G. (2011). Moment-based inference with stratified data. *Econometric Theory*, 27: 47–73.

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

Whittemore, A. S. (1997). Multistage Sampling Designs and Estimating Equations. *Journal of Royal Statistical Society, Series B*, 59: 589–602.

Wooldridge, J. (1999). Asymptotic Properties of Weighted M-estimators for Variable Probability Samples. *Econometrica*, 69: 1385–1406.

Wooldridge, J. (2002). Inverse Probability Weighted M-Estimation for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal*, 1: 117–139.

Wooldridge, J. (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics*, 141(2): 1281–1301.

Supplemental Appendix to "A Note on Efficiency Gains from Multiple Incomplete Sub-samples"

Supplemental Appendix A contains descriptive endnotes from Sections 2 and 3. Supplemental Appendix B relaxes the assumption of known $P(C = r|T_r(Z) = T_r(z))$ and presents certain auxiliary results that are either extensions of the results in Section 3, or have interesting implications in the context of our discussions from Sections 3 and 5. Indeed, it also shows that the latter interesting implications only become evident once we consider $R > 2$ (unlike in Chen et al. (2008)) and allow the target population to be as general as it is in our paper. Supplemental Appendix C contains evidence of good finite-sample behavior of the GMM estimator in the context of the simulation study from Section 5, and thus lends credibility to the discussion of efficiency gain/loss in Section 5.

Supplemental Appendix A: Descriptive endnotes

A1. A brief literature review and connection with our paper (reference: Section 2)

A2. A simple example showing the variance reduction through dependent sampling (reference: Section 2)

A3. Connection with the literature on moment augmentation and calibration (reference: Section 3)

Supplemental Appendix B: Auxiliary Results and Proofs

B1. Certain efficiency bounds under MAR in (3) but without imposing known $P(C = r|T_r(Z) = T_r(z))$ in (6)

B2. Results under CMAR in (4)

- **(a)** Efficiency bounds under CMAR when $P(C = r|T_r(Z) = T_r(z))$ is known, partially known up to finite dimensional parameters, and completely unknown: similar to e.g. Hahn (1998) and Chen et al. (2008)
- **(b)** Analytical expressions for efficiency gains due to the use of certain incomplete sub-samples when $R = 3$: known versus unknown $P(C = r|T_r(Z) = T_r(z))$ have very different implications on such gains

Supplemental Appendix C: Simulation evidence of finite-sample properties of $\widehat{\beta}_\lambda$ used in Section 5

Supplemental Appendix A: Descriptive endnotes

A1. A brief literature review and connection with our paper

The idea of planned incompleteness in the data is nonstandard in economics. However, commonly used survey techniques could possibly be modified easily to obtain the data structure considered in this paper. For example, if the information collected in the first phase of a variable probability sampling [see Wooldridge (1999)] is retained for all the units (and not only for those being followed in the next phase), then it leads to our data structure with $R = 2$. If this is continued for more than two phases, in which case not retaining the information of the discarded units in each phase would seem quite unreasonable, then it leads to our data structure with $R > 2$.

Let us note here a few common (but a very much incomplete list of) instances of planned incompleteness in the data from various fields of research. The two/many method/measurement design is used in psychology and behavioral research where it is common to encounter a "gold standard" (good) measure and other inexpensive

but less accurate measures for behavioral traits [see e.g. Graham et al. (2006)] just as multiple measures of e.g. consumption are common in economics [see e.g. Beegle et al. (2012)]. See Carroll and Wand (1991), Lee and Sepanski (1995), etc. for sophisticated use of validation data to deal with measurement error. In a different context, MacArdle and Woodcock (1997) demonstrate the usefulness of planned missing waves in panel in estimating key quantities of interest in the psychology literature. Nijman et al. (1991) demonstrate the same for the rotating panels (e.g. the Current Population Survey) in economics. In yet another context, the multiple matrix sampling of Shoemaker (1973) was extended as the split-questionnaire design (SQD) of Raghunathan and Grizzle (1995) in the statistics literature, the partial questionnaire design (PQD) of Wacholder et al. (1994) in epidemiology, and the multi-forms, mainly 2-D and 3-D forms, (e.g. the Occupational Information Network survey), discussed by Graham et al. (1996), Graham et al. (2006) in psychology and behavioral research.

Note that PQD can accommodate for non-monotonicity of up to 4 levels, and Chatterjee and Li (2010) propose efficient estimation in this setup under certain additional assumption. Chaudhuri and Guilkey (2016), unfortunately oblivious of the development in Chatterjee and Li (2010), but in a different (from PQD) context, provide a generalization by relaxing some of these assumptions. However, efficient estimation or even obtaining the expression for the efficiency bound (which we will use as the benchmark) quickly becomes intractable under the generalizations that are of interest to our present paper. SQD and multi-form design also do not require the monotone pattern or even the complete sub-sample. However, nor do they handle anything but independent (of the observed data) allocation of units to the various sub-samples.

A2. A simple example showing the variance reduction through dependent sampling

Consider estimating the parameter β from a regression model $Y = \alpha + \beta X + \epsilon$ where Y and X are scalar random variables. For simplicity, let $X \sim Bin(1, q)$ and let the model error $\epsilon \sim (0, \sigma^2)$ be independent of X. Let $\mathcal{S} = \{D_i, D_i Y_i, X_i\}_{i=1}^n$ where D is a binary variable such that we observe Y in $\mathcal{S}$ only when $D = 1$. (We switch the missing variable from X to Y in this example, unlike in most of our paper, so that we can consider a simple unweighted estimator without bothering about bias due to the possible non-representativeness of the units with $D_i = 1$ [see Wooldridge (2007)].) Let $p(j) = E[D|X = j]$ for $j = 0, 1$. Then $p := E[D] = qp(1) + (1 - q)p(0)$ and $E[DX] = qp(1)$. The ordinary least squares estimator $\widehat{\beta}$ of β, based on sample units with $D_i = 1$, and the asymptotic variance of $\widehat{\beta}$ are, respectively:

$$\widehat{\beta} = \sum_{i=1}^{n} D_i X_i \left(Y_i - \sum_j D_j Y_j / \sum_j D_j \right) \Bigg/ \sum_{i=1}^{n} D_i X_i \left(X_i - \sum_j D_j X_j / \sum_j D_j \right)$$

and

$$\text{Avar} = \sigma^2 / E[DX] \left(1 - E[DX]/E[D]\right) = p\sigma^2 / [qp(1)(p - qp(1))] \, .$$

If $P(D = 1|Y, X) = P(D = 1) = p$, implying that $p(1) = p(0) = p$, then Avar $= \sigma^2/pq(1-q)$. On the other hand, $p(1) = p/(2q)$ minimizes the general Avar and the minimized value is Avar $= 4\sigma^2/p$, which is strictly smaller than $\sigma^2/pq(1-q)$ unless $q = 1/2$. Hence, by virtue of making D dependent on X, optimally, one could correct for the non-50-50 assignment of X in the population – the essential idea behind stratification – to minimize variance.

A3. Connection with the literature on moment augmentation and calibration

The idea behind using the moment restrictions in (12) to augment the moment restriction (11), that already identifies β_λ^0 and can be used to obtain a $\sqrt{n}$-consistent estimator [see e.g. Wooldridge (2007)], and thus achieving efficiency gains is the same as the idea of calibration in the survey sampling literature [see e.g. Deville and Sarndal (1992)]. The same idea, in more economics-centric ways, has appeared in the econometrics literature also: see Back and Brown (1993), Imbens and Lancaster (1994), Hellerstein and Imbens (1999), Devereux and Tripathi (2009), Tripathi (2011), Graham et al. (2012), etc. or Hellerstein and Imbens (1999), Nevo (2003), etc. in another context. To see the connection, first note that under our setup this means estimating β_λ^0 by solving for β from $\sum_{i=1}^n \omega_i \varphi_{R,\lambda}(O_i,\beta) = 0$ where $\omega_i = I(C_i = R)/P(C = R|T_R(Z_i)) = \omega_{IPW,i}$, say, (instead of $1/n$ to reflect the non-representativeness of the complete sub-sample) if only (11) is used. On the other hand, if the calibration/augmenting/auxiliary restrictions in (12) are also utilized, then $\omega_i = \omega_{IPW,i} + \sum_{r=1}^{R-1} a_{r,i}$ where e.g. $a_{r,i}$'s are recursively obtained using (15). For example, if $R = 2$, then $a_{1,i} = \omega_{IPW,i}\Upsilon'_{K_1}(T_1(Z_i))(\sum_{j=1}^n \Upsilon_{K_1}(T_1(Z_j))\Upsilon'_{K_1}(T_1(Z_j)))^{-1}\sum_{l=1}^n(1-\omega_{IPW,l})\Upsilon_{K_1}(T_1(Z_l))$ where $\Upsilon_{K_1}(T_1(Z))$ is a $K_1 \times 1$ vector of some possibly orthogonalized series of functions (e.g. power series, splines, etc.) of $T_1(Z)$ with possibly $K_1 \to \infty$ as $n \to \infty$ [see Graham et al. (2012)]. One could instead use $\bar{\omega}_i = \omega_i / \sum_j \omega_j$ as the weights so that they necessarily add up to one. However, there is no guarantee that $\bar{\omega}_i \in [0,1]$ for all i (indeed it can be outside $[0,1]$ for all i), which is not a desirable characteristic for weights. We do not pursue corrections for this undesirable characteristic of the weights since they are peripheral to the main message of our paper.

Supplemental Appendix B: Auxiliary results and proofs

All the proofs are collected at the end of the Supplemental Appendix.

B1. Results under MAR in (3) without planned incompleteness assumption in (6)

We are unable to provide a concise but general result such as Proposition 1 under MAR in (3) when we completely relax the planned incompleteness assumption in (6). (We can provide such general results under CMAR in (4); we present them in Supplemental Appendix B2.) However, even under MAR, we can obtain concise results for certain choices of λ. In this subsection we provide the results for two such choices that are most appealing in practice: $\lambda = \mathcal{C}$ and $\lambda = \{1\}$. As noted in Remark 2 below Proposition 1, the result for $\lambda = \mathcal{C}$ is already known; indeed the result is identical to the one that presented in our Proposition 1 [also see Hahn (1998) and Chen et al. (2008) for the case when $R = 2$]. This is perhaps the most commonly studied case in this literature. On the other hand, the result for $\lambda = \{1\}$ appears to be new in this general context. It is also practically useful. For example, if we consider attrition in panel data then it is possible that the major attrition takes place at the end of the first period. Our result in this context thus establishes the efficiency benchmark for estimation of the features of the joint distribution in this sub-population ($\lambda = \{1\}$) who chose to leave the panel at the end of the first period.

For the choice of $\lambda = \mathcal{C}$ and $\lambda = \{1\}$, respectively, define in the spirit of the alternative expression in (10) as:

$$\bar{\varphi}_{\mathcal{C}}(O;\beta) := E[m(T_R(Z);\beta)|T_1(Z)] + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r(Z))} \left(E[m(T_R(Z);\beta)|T_r(Z)] - E[m(T_R(Z);\beta)|T_{r-1}(Z)]\right)$$

$$\bar{\varphi}_{\{1\}}(O;\beta) := \frac{I(C=1)}{P(C=1)} E[m(T_R(Z);\beta)|T_1(Z)] + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r(Z))} \left[\varphi_{r,\{1\}}(O;\beta) - \varphi_{r-1,\{1\}}(O;\beta)\right]$$

where $\varphi_{r,\lambda}(O;\beta)$ for a general λ, including $\lambda = \{1\}$, is defined in (7). To benefit from concise expressions (and, hopefully, clarity), we have used (i) $E[m(T_R);\beta)|T_R(Z)]$ and $m(T_R(Z);\beta)$, (ii) the event $\{C \geq R\}$ and $\{C = R\}$, and (iii) the event $\{C \in \{1\}\}$ and $\{C = 1\}$ interchangeably. We hope this is not unduly confusing to the reader.

Proposition 6 *Let* (1), (3), *and assumption A hold. For the respective choices* $\lambda = \mathcal{C}$ *and* $\lambda = \{1\}$, *let the* $d_m \times d_m$ *matrix* $\bar{V}_\lambda := Var(\bar{\varphi}_\lambda(O;\beta^0_\lambda))$ *be finite and positive definite where* β^0_λ *is defined in* (1). *Then for* β^0_λ, *the asymptotic variance lower bound for* $\sqrt{n}(\widehat{\beta} - \beta^0_\lambda)$ *of any regular estimator* $\widehat{\beta}$ *is given by* $\bar{\Omega}_\lambda := \left(M'_\lambda \bar{V}_\lambda^{-1} M_\lambda\right)^{-1}$. *An estimator whose asymptotic variance equals* $\bar{\Omega}_\lambda$ *has the asymptotically linear representation*

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

B2. Results under CMAR in (4): with and without planned incompleteness in (6)

Now we focus on the CMAR condition in (4), i.e., $P(C = r|Z) = P(C = r|T_1(Z))$ for all $r = 1, \ldots, R$, instead of the MAR condition in (3). Recall that CMAR is more restrictive than MAR but less restrictive than INDEP (5).

(a) First, we obtain the efficiency bound for estimation of β^0_λ for any choice of λ as in Proposition 1, unlike which, however, the results are now applicable also to case of unplanned incompleteness.[16] In particular, the results are for a general λ under CMAR where $P(C = r|Z)$ is: (i) completely unknown (as in Proposition 6), (ii) completely known (a special case of Proposition 1), and (iii) known up to a finite dimensional parameter (an intermediate case between (i) and (ii)). Extending the original work of Hahn (1998) and Chen et al. (2008) to our general framework, we also clearly show that the bound is smallest under (ii), followed by (iii) and then (i).

(b) Next, we provide as corollaries the analytical expressions under CMAR and INDEP for the efficiency gains/loss defined in (21). Under the planned incompleteness in (6), these results are a rough reference to put the simulation results in Section 5 into perspective. Then, by relaxing (6) under CMAR, we show that the expressions thus obtained provide new intuitions on efficiency gains/loss that actually only becomes available if $R > 2$.

Define $\omega_r(T_1(Z)) := I(C \geq r)/P(C \geq r|T_1(Z))$ for all $r = 1, \ldots, R$. We use the convention $I(C \geq R) := I(C = R)$. Define $q_\lambda(T_1(Z)) := P(C \in \lambda|T_1(Z))/P(C \in \lambda)$ and $q_\lambda := I(C \in \lambda)/P(C \in \lambda)$. For any variables Y and X, let $\Pi(Y|X) := E[YX'](E[XX'])^{-1}X$ denote the population least squares projection when it exists.

Proposition 7 *Let* (1), (3), (6) *and assumption A hold. Define*

$$\varphi_\lambda(O;\beta) := q_\lambda(T_1(Z)) \left\{ E[m(T_R(Z);\beta)|T_1(Z)] + \sum_{r=2}^{R} \omega_r(T_1(Z)) \left(E[m(T_R(Z);\beta)|T_r(Z)] - E[m(T_R(Z);\beta)|T_{r-1}(Z)]\right) \right\}.$$

[16] While we noted earlier why CMAR can be useful under planned incompleteness, it turns out that CMAR has also been used under scenarios (not necessarily always involving surveys) where the incompleteness is unplanned; e.g. the second attrition analysis (page 145) in Fitzgerald et al. (1998), or under the name of sequential ignorability for identification in mediation analysis in Imai et al. (2010b), Imai et al. (2010a), Shpitser and Tchetgen (2012, 2014), etc.

Let $V_\lambda := Var(\varphi_\lambda)$ be a $d_m \times d_m$ finite positive definite matrix. Then for β_λ^0, the asymptotic variance lower bound for $\sqrt{n}(\widehat{\beta} - \beta_\lambda^0)$ of any regular estimator $\widehat{\beta}$ is given by $\Omega_\lambda := \left(M_\lambda' V_\lambda^{-1} M_\lambda\right)^{-1}$. An estimator whose asymptotic variance equals Ω_λ has the asymptotically linear representation

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -\Omega_\lambda M_\lambda' V_\lambda^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_\lambda(O_i; \beta_\lambda^0) + o_p(1).$$

Proposition 8 *Let* (1), (3) *and assumption A hold. Assume $P(C = r|T_1(Z)) = P(C = r|T_1(Z); \gamma^0)$ for some $\gamma^0 \in \Gamma \subset \mathbb{R}^{d_\gamma}$ where $P(C = r|T_1(Z); \gamma)$ is known up to the finite-dimensional unknown γ for $r = 1, \ldots, R$. Let $S_\gamma(C|T_1(Z)) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|T_1(Z))} \frac{\partial}{\partial \gamma} P(C = r|T_1(Z); \gamma^0)$ denote the score function for γ evaluated at $\gamma = \gamma^0$, and assume that $E\left[S_\gamma(C|T_1(Z)) S_\gamma(C|T_1(Z))'\right]$ is positive definite. Define*

$$\varphi_{\lambda[pu]}(C, T_C(Z); \beta) := \varphi_\lambda(C, T_C(Z); \beta) + \Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(T_R(Z); \beta)|T_1(Z)] \middle| S_\gamma(C|T_1(Z))\right)$$

where the subscript $[pu]$ denotes that $P(C = r|T_1(Z))$ is partially unknown, i.e., the finite dimensional parameter γ is unknown. Let $V_{\lambda[pu]} := Var(\varphi_{\lambda[pu]})$ be a $d_m \times d_m$ finite positive definite matrix. Then for β_λ^0, the asymptotic variance lower bound for $\sqrt{n}(\widehat{\beta} - \beta_\lambda^0)$ of any regular estimator $\widehat{\beta}$ is given by $\Omega_{\lambda[pu]} := \left(M_\lambda' V_{\lambda[pu]}^{-1} M_\lambda\right)^{-1}$. An estimator whose asymptotic variance equals $\Omega_{\lambda[pu]}$ has the asymptotically linear representation

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -\Omega_{\lambda[pu]} M_\lambda' V_{\lambda[pu]}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{\lambda[pu]}(O_i; \beta_\lambda^0) + o_p(1).$$

Proposition 9 *Let* (1), (3) *and assumption A hold. Define*

$$\varphi_{\lambda[u]}(O; \beta) := q_\lambda E[m(T_R(Z); \beta)|T_1(Z)] + q_\lambda(T_1(Z)) \sum_{r=2}^{R} \omega_r(T_1(Z)) \left(E[m(T_R(Z); \beta)|T_r(Z)] - E[m(T_R(Z); \beta)|T_{r-1}(Z)]\right)$$

where the subscript $[u]$ denotes that $P(C = r|T_1(Z))$ is unknown. Let $V_{\lambda[u]} := Var(\varphi_{\lambda[u]})$ be a $d_m \times d_m$ finite positive definite matrix. Then for β_λ^0, the asymptotic variance lower bound for $\sqrt{n}(\widehat{\beta} - \beta_\lambda^0)$ of any regular estimator $\widehat{\beta}$ is given by $\Omega_{\lambda[u]} := \left(M_\lambda' V_{\lambda[u]}^{-1} M_\lambda\right)^{-1}$. An estimator whose asymptotic variance equals $\Omega_{\lambda[u]}$ has the asymptotically linear representation

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -\Omega_{\lambda[u]} M_\lambda' V_{\lambda[u]}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{\lambda[u]}(O_i; \beta_\lambda^0) + o_p(1).$$

For concise notation, we henceforth write $m(Z; \beta_\lambda^0)$ as m and $T_r(Z)$ as T_r for $r = 1, \ldots, R$ unless confusing.

Remarks

1. Proposition 7 is a special case of Proposition 1. Proposition 8 is an intermediate result that partially relaxes the planned incompleteness assumption in (6). On the other hand, Proposition 9 fully relaxes (6) but, unlike Proposition 6, it can still provide the bounds for a general choice of the target (sub-)population $\lambda \in \Lambda$ because it works under CMAR in (4) instead of MAR.

2. It is straightforward to see from Propositions 7-9 that:

$$\begin{aligned}
V_{\lambda[u]} &= E\left[\frac{P(C\in\lambda|T_1)}{P^2(C\in\lambda)}E[m|T_1]E[m'|T_1]+\frac{P^2(C\in\lambda|T_1)}{P^2(C\in\lambda)}\sum_{r=2}^{R}\frac{Var\left(E[m|T_r]|T_{r-1}\right)}{P(C\geq r|T_1)}\right]\\
V_{\lambda} &= V_{\lambda[u]}-\frac{P(C\in\lambda|T_1)(1-P(C\in\lambda|T_1))}{P^2(C\in\lambda)}E[m|T_1]E[m|T_1]'\\
V_{\lambda[pu]} &= V_{\lambda}+B\left(E\left[S_\gamma(C|T_1)S_\gamma(C|T_1)'\right]\right)^{-1}B'\\
&= V_{\lambda[u]}-Var\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]-\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]\middle|S_\gamma(C,T_1)\right)\right)
\end{aligned}$$

where $B:=E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]S_\gamma(C|T_1)'\right]=E\left[\frac{E[m|T_1]}{P(C\in\lambda)}\sum_{r\in\lambda}\frac{\partial}{\partial\gamma'}P(C=r|T_1;\gamma^0)\right]$. Therefore, $V_\lambda=V_{\lambda[pu]}=V_{\lambda[u]}$ if $\lambda=\mathcal{C}$ and, otherwise, $V_\lambda\leq V_{\lambda[pu]}\leq V_{\lambda[u]}$ in the matrix sense. This ordering of the asymptotic variances establishes the result for a general R and a general choice of the target (sub-)population λ by generalizing the result with $R=2$ and $\lambda=\{1\}$, $\lambda=\{1,2\}$ from Chen et al. (2008) [also see Hahn (1998)].

As noted earlier, now let us consider the analytical expressions for the loss in (21). The expressions are complicated under MAR (3) and not very intuitive; however, they are very intuitive and self-explanatory under CMAR (4) and INDEP (5). So we list them for INDEP and CMAR as corollaries to Propositions 7 and 9.

For simplicity let $R=3$ and $d_\beta=d_m=1$. Let $V_\lambda^{s'}$ denote $Var(\varphi_\lambda(O;\beta_\lambda^0))$ when the latter is modified according to the discussion below (21). To avoid clutter, we write $m(Z;\beta_\lambda^0)$ as m, $T_r(Z)$ as T_r, $P(C=r)$ as p_r, $P(C=r|T_1)$ as $p_r(T_1)$, $P(C\in\{r,t\})$ as p_{rt} and $P(C\in\{r,t\}|T_1)$ as $p_{rt}(T_1)$ for $r,t=1,2,3$.

Corollary 10 *Let* (1), (5) *and assumption A hold. Under INDEP in* (5), β_λ^0 *is the same for all* $\lambda\in\Lambda$, *and* $p_r(T_1)=p_r$ *for all* $r=1,\ldots,R$, *and thus there is no distinction between planned versus unplanned incompleteness. Taking* $\lambda=\mathcal{C}:=\{1,2,3\}$, *and assuming that the concerned variances exist, the following hold as* $n\to\infty$*:*

(a) $Loss(\beta_\lambda;s=\{3\},s'=\{1,3\})\times V_\lambda^{\{1,3\}}=\frac{p_1}{p_3}E\left[E[m|T_1]^2\right]$.

(b) $Loss(\beta_\lambda;s=\{3\},s'=\{2,3\})\times V_\lambda^{\{2,3\}}=\frac{p_2}{p_3}E\left[E[m|T_2]^2\right]$.

(c) $Loss(\beta_\lambda;s=\{1,3\},s'=\{1,2,3\})\times V_\lambda^{\{1,2,3\}}=\frac{p_2}{p_{13}}E\left[E[m|T_1]^2\right]+\frac{p_2}{p_3p_{23}}E\left[Var(E[m|T_2]|T_1)\right]$.

(d) $Loss(\beta_\lambda;s=\{2,3\},s'=\{1,2,3\})\times V_\lambda^{\{1,2,3\}}=\frac{p_1}{p_{23}}E\left[E[m|T_1]^2\right]$.

Corollary 11 *Let* (1), (4), (6) *and assumption A hold. Under CMAR* ((4)) *we do not consider* $s=\{3\}$ *for brevity, unless* $\lambda=\{3\}$. *Assuming that the concerned variances exist, the following hold as* $n\to\infty$*:*

(a) $Loss(\beta_{\{1\}};s=\{1,3\},s'=\{1,2,3\})\times V_{\{1\}}^{\{1,2,3\}}=E\left[\frac{p_1(T_1)p_2(T_1)}{p_1}\left\{\frac{E[m|T_1]^2}{p_{13}(T_1)}+\frac{Var(E[m|T_2]|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle|C=1\right]$.

(b) $Loss(\beta_{\{2\}};s=\{2,3\},s'=\{1,2,3\})\times V_{\{2\}}^{\{1,2,3\}}=E\left[\frac{p_1(T_1)p_2(T_1)}{p_2p_{23}(T_1)}E[m|T_1]^2\middle|C=2\right]$.

(c1) $Loss(\beta_{\{3\}};s=\{3\},s'=\{1,3\})\times V_{\{3\}}^{\{1,3\}}=E\left[\frac{p_{13}}{p_3}\frac{p_1(T_1)}{p_{13}(T_1)}E[m|T_1]^2\middle|C=3\right]$.

(c2) $Loss(\beta_{\{3\}};s=\{3\},s'=\{2,3\})\times V_{\{3\}}^{\{2,3\}}=E\left[\frac{p_{23}}{p_3}\frac{p_2(T_1)}{p_{23}(T_1)}E[m|T_2]^2\middle|C=3\right]$.

(c3) $Loss(\beta_{\{3\}};s=\{1,3\},s'=\{1,2,3\})\times V_{\{3\}}^{\{1,2,3\}}=E\left[\frac{p_2(T_1)p_3(T_1)}{p_3}\left\{\frac{E[m|T_1]^2}{p_{13}(T_1)}+\frac{Var(E[m|T_2]|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle|C=3\right]$.

(c4) $Loss(\beta_{\{3\}};s=\{2,3\},s'=\{1,2,3\})\times V_{\{3\}}^{\{1,2,3\}}=E\left[\frac{p_1(T_1)p_3(T_1)}{p_3p_{23}(T_1)}E[m|T_1]^2\middle|C=3\right]$.

(d) $Loss(\beta_{\{1,3\}};s=\{1,3\},s'=\{1,2,3\})\times V_{\{1,3\}}^{\{1,2,3\}}=E\left[\frac{p_2(T_1)p_{13}(T_1)}{p_{13}}\left\{\frac{E[m|T_1]^2}{p_{13}(T_1)}+\frac{Var(E[m|T_2]|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle|C\in\{1,3\}\right]$.

(e) $Loss(\beta_{\{2,3\}}; s=\{2,3\}, s'=\{1,2,3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} E[m|T_1]^2 \middle| C \in \{2,3\}\right]$.

(f1) $Loss(\beta_{\{1,2,3\}}; s=\{1,3\}, s'=\{1,2,3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)}{p_{13}(T_1)} E[m|T_1]^2 + \frac{p_2(T_1)}{p_3(T_1)p_{23}(T_1)} Var(E[m|T_2]|Z_1)\right]$.

(f2) $Loss(\beta_{\{1,2,3\}}; s=\{2,3\}, s'=\{1,2,3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} E[m|T_1]^2\right]$.

Remarks: To complement the discussion on efficiency in Section 5 of Wooldridge (2007), let us note here that Corollaries 10 and 11 imply that there may not always be a strict loss in efficiency in the sense of (21) (and under the premise of our discussion below it) when one does not use all the sub-samples. For example, if $E[m|T_2] \equiv E[m|Z_{(1)}, Z_{(2)}] = 0$, then there is never any loss in all the above cases. Similarly, there is no loss in Corollary 10 (a), (d) and Corollary 11 (b), (c1), (c4), (e), (f2) if only $E[m|T_1] \equiv E[m|Z_{(1)}] = 0$. None of these is the case with the model in our simulation study.

As an aside, $s=\{3\}$ corresponds to both the IPW estimator in (19) and the so-called complete case estimator (which does not use weights to correct for selection) that are numerically equivalent if $\lambda=\{3\}$ or under INDEP. (Otherwise, the complete case estimator is generally not consistent for β_λ^0.) The IPW estimator is not considered in Corollary 11 (except if $\lambda=\{3\}$) because we had already pointed out following Proposition 1 the precise additional information not used by the IPW estimator. Let us now see what happens under unplanned incompleteness.

Corollary 12 *Let* (1), (4) *(but not* (6)*) and assumption A hold. Under CMAR (*(4)*) we do not consider* $s=\{3\}$ *for brevity, unless* $\lambda=\{3\}$*. Assuming that the concerned variances exist, the following hold as* $n \to \infty$*:*

(a) $Loss(\beta_{\{1\}}; s=\{1,3\}, s'=\{1,2,3\}) \times V_{\{1\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)p_2(T_1)}{p_1p_3(T_1)p_{23}(T_1)} Var(E[m|T_2)|T_1) \middle| C=1\right]$.

(b) $Loss(\beta_{\{2\}}; s=\{2,3\}, s'=\{1,2,3\}) \times V_{\{2\}}^{\{1,2,3\}} = E\left[\frac{p_3(T_1)}{p_2p_{23}(T_1)} Var(E[m|T_2]|T_1) \middle| C=2\right]$.

(c1) $Loss(\beta_{\{3\}}; s=\{3\}, s'=\{1,3\}) \times V_{\{3\}}^{\{1,3\}} = 0$.

(c2) $Loss(\beta_{\{3\}}; s=\{3\}, s'=\{2,3\}) \times V_{\{3\}}^{\{2,3\}} = 0$.

(c3) $Loss(\beta_{\{3\}}; s=\{3\}$ *or* $s=\{1,3\}$ *or* $s=\{2,3\}, s'=\{1,2,3\}) \times V_{\{3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)}{p_3p_{23}(T_1)} Var(E[m|T_2]|T_1) \middle| C=3\right]$.

(d) $Loss(\beta_{\{1,3\}}; s=\{1,3\}, s'=\{1,2,3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)p_{13}(T_1)}{p_{13}p_3(T_1)p_{23}(T_1)} Var(E[m|T_2]|T_1) \middle| C \in \{1,3\}\right]$.

(e) $Loss(\beta_{\{2,3\}}; s=\{2,3\}, s'=\{1,2,3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = 0$.

(f1) $Loss(\beta_{\{1,2,3\}}; s=\{1,3\}, s'=\{1,2,3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)}{p_{13}(T_1)} E[m|T_1]^2 + \frac{p_2(T_1)}{p_3(T_1)p_{23}(T_1)} Var(E[m|T_2]|Z_1)\right]$.

(f2) $Loss(\beta_{\{1,2,3\}}; s=\{2,3\}, s'=\{1,2,3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} E[m|T_1]^2\right]$.

Remarks: First, it is not surprising that for results (f1) and (f2), the Corollaries 11 and 12 are identical since, as evident from Propositions 7 and 9, there is no difference between planned versus unplanned incompleteness when $\lambda = \mathcal{C}$. Second, note that leaving out incomplete sub-samples from estimation now results in zero loss under even weaker conditions, i.e., $Var(E[m|T_2]|T_1) = 0$ as opposed to $E[m|T_2] = 0$ under Corollary 11. For example, if $E[m|T_2]$ is constant almost surely in T_2, then leaving out sub-samples does not results in any loss under Corollary 12(a)-(e). On the other hand, for the same to hold under Corollary 11 it would require a stronger condition that $E[m|T_2] = 0$. Finally, we note that this difference in the implication on loss under planned versus unplanned incompleteness manifests more prominently when the sample used for estimation has only one level of incompleteness. This is evident from comparing the results (c1), (c2) and (e) in Corollaries 11 and 12 respectively.

Supplemental Appendix C: Simulation evidence of finite-sample properties of $\widehat{\beta}_\lambda$

Apart from the GMM estimators based on various sub-samples, we also consider the complete case (CC) and IPW estimators. The CC estimator is the default for the statistical softwares. This is based only on the complete sub-sample by treating it as representative of the population of interest. The IPW estimator is defined in (19).

Let us now consider the finite-sample properties of the estimators. We report them in Table 4 under INDEP, Tables 5 for Intercept and 6 for Slope under CMAR, and Tables 7 for Intercept and 8 for Slope under MAR. In particular, we focus on the following quantities computed as average over the 10,000 Monte Carlo trials: Mbias (deviation from the true values), Abias (absolute deviation from the true values), Std (standard deviation obtained as $\sqrt{(\text{estimated Avar})/(\text{size of the used sample})}$) and Size (rejection of the true value by a 5% two-sided t-test).

The CC and IPW estimators are numerically equivalent if $\lambda = \{3\}$ or under INDEP. Otherwise, as expected, CC can be badly biased (Mbias) since it does not recognize the sample-selection. As a result, the estimated size with CC can be large, in particular it can be 1 or close to 1 for the Intercept term for various target λ's. The other estimators are consistent under our assumptions, and their small Mbias and decreasing (with n) Std support this.

The ordering of variability of the estimators, as measured by Abias and Std, are as expected: always the largest when the used sample is $\{3\}$, and the smallest when the used sample is $\{1, 2, 3\}$. Comparison between the two estimators based on the used samples $\{1, 3\}$ and $\{2, 3\}$ is possible under INDEP or when $\lambda = \{3\}$ or $\lambda = \{1, 2, 3\}$. However, thanks to the definition of X_c and X_e, there is essentially no difference in the performance between these two estimators. Overall, under our simulation design all the estimators display good properties in finite samples, and thus lend credibility to above discussion of the simulation results on the efficiency loss/gain.

References

Back, K. and Brown, D. (1993). Implied Probabilities in GMM estimators. *Econometrica*, 61: 971–976.

Beegle, K., Weerdt, J. D., Friedman, J., and Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics*, pages 3 – 18.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of Royal Statistical Society, Series B*, 53: 573–585.

Chatterjee, N. and Li, Y. (2010). Inference in Semiparametric Regression Models Under Partial Questionnaire Design and Nonmonotone Missing Data. *Journal of the American Statistical Association*, pages 787 – 797.

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.

Devereux, P. J. and Tripathi, G. (2009). Optimally combining censored and uncensored datasets. *Journal of Econometrics*, 151: 17–32.

Deville, J. C. and Sarndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87: 376–382.

Graham, B. S., Pinto, C., and Egel, D. (2012). Inverse Probability Tilting for Moment Condition Models with Missing Data. *Review of Economic Studies*, 79: 1053 – 1079.

Graham, J. W., Hofer, S. M., and MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31: 197–218.

Graham, J. W., Taylor, B. J., Olchowski, A. E., and Cumsille, P. E. (2006). Planned Missing Data Designs in Psychological Research. *Psychological Methods*, 11: 323–342.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66: 315–331.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing Moment Restriction from Auxiliary Data by Weighting. *The Review of Economics and Statistics*, 81: 1–14.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, 15: 309–334.

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

Imbens, G. W. and Lancaster, T. (1994). Combining Micro and Macro Data in Microeconometric Models. *Review of Economic Studies*, 61: 655–689.

Lee, L. and Sepanski, J. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of American Statistical Association*, 90: 130–140.

MacArdle, J. J. and Woodcock, R. W. (1997). Expanding testretest designs to include developmental time-lag components. *Psychological Methods*, 2: 403–435.

Nevo, A. (2003). Using Weights to Adjust for Sample Selection When Auxiliary Information is Available. *Journal of Business and Economic Statistics*, 21: 43–52.

Nijman, T., Verbeek, M., and van Soest, A. (1991). The efficiency of rotating-panel designs in an analysis-of-variance model. *Journal of Econometrics*, 49: 373–399.

Raghunathan, T. E. and Grizzle, J. E. (1995). A Split Questionnaire Survey Design. *Journal of the American Statistical Association*, pages 54 – 63.

Shoemaker, D. M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Cambridge, MA: Ballinger.

Shpitser, I. and Tchetgen, E. J. T. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Biometrika*, 40: 1816–1845.

Shpitser, I. and Tchetgen, E. J. T. (2014). Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101: 849–864.

Tripathi, G. (2011). Moment-based inference with stratified data. *Econometric Theory*, 27: 47–73.

Wacholder, S., Carroll, R. J., Pee, D., and Gail, M. H. (1994). The Partial Questionnaire Design For Case-Control Studies. *Statistics in Medicine*, 13: 623 – 634.

Wooldridge, J. (1999). Asymptotic Properties of Weighted M-estimators for Variable Probability Samples. *Econometrica*, 69: 1385–1406.

Wooldridge, J. (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics*, 141(2): 1281–1301.

Tables for Supplemental Appendix C:

Used	$n = 1000$				$n = 2000$				$n = 5000$			
Sample	Mbias	Abias	Std	Size	Mbias	Abias	Std	Size	Mbias	Abias	Std	Size
$\{3\}$	0.0003	0.0589	0.0723	0.0526	0.0007	0.041	0.0512	0.0528	0.0002	0.0257	0.0324	0.049
$\{1,3\}$	-0.0007	0.0426	0.0518	0.06	0.0006	0.03	0.0368	0.0579	0.0003	0.0186	0.0233	0.0478
$\{2,3\}$	-0.0003	0.0433	0.0519	0.0652	0.0003	0.0303	0.0368	0.0583	0.0001	0.0189	0.0234	0.0537
$\{1,2,3\}$	-0.0005	0.0384	0.045	0.0704	0.0003	0.0265	0.032	0.0585	0.0002	0.0163	0.0203	0.0511
$\{3\}$	0.0009	0.0587	0.072	0.0572	-0.0006	0.0412	0.0512	0.0579	-0.0005	0.0258	0.0324	0.0507
$\{1,3\}$	0.0054	0.0481	0.0556	0.0741	0.0019	0.0329	0.0398	0.0611	0.0006	0.0205	0.0253	0.0539
$\{2,3\}$	0.0028	0.0481	0.0553	0.0789	0.0012	0.033	0.0395	0.0651	0	0.0205	0.0252	0.0559
$\{1,2,3\}$	0.0041	0.0454	0.05	0.0885	0.0017	0.0303	0.0359	0.0695	0.0004	0.0189	0.0229	0.06

Table 4: Bias (Mbias), absolute bias (Abias), standard deviation (Std) of the estimators and the size of a 5% nominal level two-sided t-test (Size) are reported based on the average over 10,000 Monte Carlo trials under INDEP sampling. Target population $\lambda = \{1,2,3\}$. Top panel is for the Intercept ($\beta_{\lambda,1}$) and bottom for the Slope ($\beta_{\lambda,2}$) parameters.

CMAR Sampling. Parameter of interest is the Intercept ($\beta_{\lambda,1}$)

Target Popln. (λ)	Used Sample (s)	$n = 1000$				$n = 2000$				$n = 5000$			
		Mbias	Abias	Std	Size	Mbias	Abias	Std	Size	Mbias	Abias	Std	Size
$\{1\}$	$\{3\}$:CC	-0.1277	0.1299	0.0711	0.433	-0.1281	0.1283	0.0503	0.7244	-0.1294	0.1294	0.0319	0.9808
$\{1\}$	$\{3\}$:IPW	-0.0004	0.0583	0.0733	0.054	-0.0002	0.0416	0.052	0.0534	-0.0008	0.0265	0.0329	0.051
$\{1\}$	$\{1,3\}$	-0.001	0.0453	0.0545	0.0649	-0.0004	0.0317	0.0388	0.0584	-0.0005	0.02	0.0246	0.0501
$\{1\}$	$\{1,2,3\}$	0.0006	0.0415	0.0484	0.0711	0.0005	0.0284	0.0345	0.0613	0	0.0177	0.022	0.0502
$\{2\}$	$\{3\}$:CC	0.2496	0.2496	0.0711	0.9376	0.2492	0.2492	0.0503	0.9981	0.2479	0.2479	0.0319	1
$\{2\}$	$\{3\}$:IPW	0.0029	0.0623	0.0782	0.0527	0.0021	0.0445	0.0554	0.0502	-0.0002	0.0282	0.0351	0.0516
$\{2\}$	$\{2,3\}$	0.003	0.0439	0.0528	0.0618	0.0013	0.0302	0.0373	0.0586	0	0.0189	0.0237	0.0486
$\{2\}$	$\{1,2,3\}$	-0.0008	0.0411	0.0472	0.0712	-0.0008	0.0277	0.0335	0.0645	-0.0006	0.0172	0.0214	0.0549
$\{3\}$	$\{3\}$:CC	0.0011	0.0565	0.0711	0.0543	0.0007	0.0402	0.0503	0.0527	-0.0006	0.0257	0.0319	0.0517
$\{3\}$	$\{3\}$:IPW	0.0011	0.0565	0.0711	0.0543	0.0007	0.0402	0.0503	0.0527	-0.0006	0.0257	0.0319	0.0517
$\{3\}$	$\{1,3\}$	-0.0006	0.043	0.052	0.0612	-0.0003	0.03	0.0368	0.0591	-0.0005	0.0189	0.0234	0.0531
$\{3\}$	$\{2,3\}$	0.0014	0.0441	0.0528	0.0588	0.0007	0.0303	0.0373	0.0594	-0.0002	0.0188	0.0237	0.0491
$\{3\}$	$\{1,2,3\}$	0.0006	0.0375	0.0444	0.0659	0.0002	0.0257	0.0315	0.0575	-0.0001	0.016	0.02	0.0505
$\{1,3\}$	$\{3\}$:CC	-0.0908	0.0976	0.0711	0.2456	-0.0912	0.0927	0.0503	0.4479	-0.0925	0.0925	0.0319	0.8227
$\{1,3\}$	$\{3\}$:IPW	0	0.0576	0.0725	0.0533	0	0.041	0.0513	0.053	-0.0008	0.0262	0.0325	0.0513
$\{1,3\}$	$\{1,3\}$	-0.0009	0.0444	0.0535	0.0643	-0.0004	0.031	0.038	0.058	-0.0005	0.0196	0.0242	0.0504
$\{1,3\}$	$\{1,2,3\}$	0.0004	0.0401	0.047	0.0681	0.0003	0.0274	0.0334	0.0589	-0.0001	0.0171	0.0213	0.0514
$\{2,3\}$	$\{3\}$:CC	0.1446	0.1458	0.0711	0.5269	0.1442	0.1443	0.0503	0.8195	0.1429	0.1429	0.0319	0.9943
$\{2,3\}$	$\{3\}$:IPW	0.0023	0.0588	0.0739	0.053	0.0015	0.0419	0.0523	0.0514	-0.0004	0.0266	0.0331	0.0511
$\{2,3\}$	$\{2,3\}$	0.002	0.0429	0.0517	0.06	0.0008	0.0296	0.0366	0.0568	-0.0002	0.0185	0.0232	0.0503
$\{2,3\}$	$\{1,2,3\}$	0.0003	0.0379	0.0445	0.0664	-0.0001	0.0257	0.0315	0.0596	-0.0004	0.016	0.02	0.0516
$\{1,2,3\}$	$\{3\}$:CC	0.0098	0.0569	0.0711	0.0551	0.0094	0.0408	0.0503	0.0565	0.0081	0.0265	0.0319	0.0582
$\{1,2,3\}$	$\{3\}$:IPW	0.001	0.0579	0.0728	0.0528	0.0007	0.0413	0.0516	0.0535	-0.0006	0.0263	0.0327	0.0505
$\{1,2,3\}$	$\{1,3\}$	0.0003	0.0442	0.0532	0.0629	0.0004	0.0309	0.0378	0.0589	-0.0003	0.0194	0.024	0.0513
$\{1,2,3\}$	$\{2,3\}$	-0.0001	0.0453	0.0543	0.0609	-0.0001	0.031	0.0383	0.0586	-0.0006	0.0193	0.0243	0.0497
$\{1,2,3\}$	$\{1,2,3\}$	0.0004	0.0385	0.0452	0.0662	0.0002	0.0263	0.0321	0.06	-0.0002	0.0163	0.0204	0.0508

Table 5: Reported are the bias (Mbias), absolute bias (Abias), standard deviation (Std) of the estimators and the size of a 5% nominal level two-sided t-test (Size) based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

CMAR Sampling. Parameter of interest is the Slope ($\beta_{\lambda,2}$)

Target Popln. (λ)	Used Sample (s)	$n = 1000$				$n = 2000$				$n = 5000$			
		Mbias	Abias	Std	Size	Mbias	Abias	Std	Size	Mbias	Abias	Std	Size
{1}	{3}:CC	-0.0064	0.0606	0.0741	0.0603	-0.0075	0.0424	0.0525	0.0535	-0.0069	0.0274	0.0333	0.0587
{1}	{3}:IPW	-0.0004	0.0656	0.0793	0.0644	-0.001	0.0459	0.0567	0.0523	-0.0001	0.0292	0.0362	0.0545
{1}	{1, 3}	0.0064	0.0523	0.0595	0.082	0.0027	0.0353	0.0426	0.0617	0.0011	0.0224	0.0272	0.0584
{1}	{1, 2, 3}	0.0043	0.0499	0.0548	0.0909	0.0014	0.0335	0.0394	0.0688	0.0005	0.0208	0.0252	0.0586
{2}	{3}:CC	0.0248	0.0636	0.0741	0.0715	0.0237	0.0465	0.0525	0.0717	0.0243	0.0335	0.0333	0.1144
{2}	{3}:IPW	-0.001	0.0748	0.0888	0.0701	-0.0015	0.0521	0.0641	0.063	-0.0007	0.0334	0.0412	0.0572
{2}	{2, 3}	0.0051	0.0539	0.0595	0.0867	0.0018	0.0359	0.043	0.0627	0.001	0.0228	0.0277	0.0598
{2}	{1, 2, 3}	0.0021	0.0543	0.0584	0.0985	0	0.0357	0.0419	0.0688	0.0002	0.0224	0.027	0.0612
{3}	{3}:CC	0.0004	0.0604	0.0741	0.0602	-0.0007	0.0421	0.0525	0.0513	-0.0001	0.0268	0.0333	0.0536
{3}	{3}:IPW	0.0004	0.0604	0.0741	0.0602	-0.0007	0.0421	0.0525	0.0513	-0.0001	0.0268	0.0333	0.0536
{3}	{1, 3}	0.0066	0.051	0.0592	0.0732	0.0027	0.0341	0.0418	0.0597	0.0011	0.0218	0.0265	0.0572
{3}	{2, 3}	0.0001	0.0524	0.0594	0.0799	-0.0011	0.0353	0.0422	0.061	-0.0005	0.022	0.0269	0.057
{3}	{1, 2, 3}	0.0081	0.0463	0.0515	0.0873	0.0035	0.0308	0.0367	0.065	0.0016	0.0193	0.0234	0.0614
{1, 3}	{3}:CC	-0.0109	0.061	0.0741	0.0618	-0.012	0.043	0.0525	0.0586	-0.0114	0.0284	0.0333	0.0667
{1, 3}	{3}:IPW	-0.0001	0.0634	0.0771	0.0617	-0.0009	0.0444	0.055	0.052	-0.0001	0.0282	0.035	0.0542
{1, 3}	{1, 3}	0.0065	0.0512	0.0587	0.0785	0.0028	0.0345	0.0418	0.0614	0.0011	0.022	0.0266	0.0592
{1, 3}	{1, 2, 3}	0.0054	0.0482	0.0531	0.0889	0.0021	0.0323	0.0381	0.0675	0.0009	0.0201	0.0244	0.0577
{2, 3}	{3}:CC	-0.0119	0.0611	0.0741	0.0623	-0.013	0.0432	0.0525	0.0602	-0.0124	0.0287	0.0333	0.068
{2, 3}	{3}:IPW	-0.0003	0.0661	0.0799	0.0622	-0.0012	0.0461	0.0571	0.0577	-0.0004	0.0295	0.0365	0.0534
{2, 3}	{2, 3}	0.0032	0.0504	0.0564	0.0843	0.0008	0.0337	0.0404	0.0598	0.0004	0.0213	0.0259	0.0611
{2, 3}	{1, 2, 3}	0.0049	0.0478	0.0524	0.0906	0.0017	0.0316	0.0375	0.0662	0.0009	0.0198	0.024	0.0595
{1, 2, 3}	{3}:CC	-0.0434	0.0701	0.0741	0.1	-0.0445	0.0559	0.0525	0.1416	-0.0439	0.0469	0.0333	0.2623
{1, 2, 3}	{3}:IPW	-0.0005	0.062	0.0755	0.0611	-0.0012	0.0432	0.0537	0.053	-0.0003	0.0276	0.0341	0.0544
{1, 2, 3}	{1, 3}	0.0023	0.0522	0.0596	0.0764	0.0001	0.0348	0.0422	0.0649	0	0.0222	0.0269	0.0566
{1, 2, 3}	{2, 3}	-0.0024	0.0539	0.0603	0.0861	-0.0027	0.0364	0.0431	0.067	-0.0013	0.0227	0.0276	0.0597
{1, 2, 3}	{1, 2, 3}	0.0048	0.0453	0.05	0.0901	0.0017	0.0302	0.0358	0.0656	0.0008	0.0189	0.0229	0.0596

Table 6: Reported are the bias (Mbias), absolute bias (Abias), standard deviation (Std) of the estimators and the size of a 5% nominal level two-sided t-test (Size) based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

MAR Sampling. Parameter of interest is the Intercept ($\beta_{\lambda,1}$)

Target Popln. (λ)	Used Sample (s)	$n = 1000$				$n = 2000$				$n = 5000$			
		Mbias	Abias	Std	Size	Mbias	Abias	Std	Size	Mbias	Abias	Std	Size
{1}	{3}:CC	-0.1375	0.1389	0.0718	0.4826	-0.1387	0.1388	0.0509	0.7725	-0.1384	0.1384	0.0322	0.9889
{1}	{3}:IPW	-0.0008	0.0596	0.0749	0.0524	-0.0011	0.0425	0.0531	0.0509	-0.0003	0.0271	0.0336	0.0522
{1}	{1, 3}	-0.0032	0.0468	0.0558	0.0685	-0.0018	0.0323	0.0397	0.0529	-0.0007	0.0206	0.0253	0.055
{1}	{1, 2, 3}	0.0004	0.0425	0.0487	0.0732	0	0.0287	0.0348	0.0583	0	0.0182	0.0222	0.0548
{2}	{3}:CC	0.2376	0.2377	0.0718	0.9105	0.2364	0.2364	0.0509	0.9964	0.2367	0.2367	0.0322	1
{2}	{3}:IPW	0.0028	0.065	0.0812	0.0527	0.0007	0.0464	0.0576	0.054	0.0005	0.0295	0.0365	0.0539
{2}	{2, 3}	0.0034	0.0451	0.0543	0.0628	0.001	0.0318	0.0385	0.0556	0.0008	0.0198	0.0244	0.0531
{2}	{1, 2, 3}	-0.0014	0.0431	0.049	0.0778	-0.0016	0.0295	0.0349	0.0663	-0.0004	0.0182	0.0223	0.0568
{3}	{3}:CC	0.0009	0.0572	0.0718	0.0488	-0.0003	0.041	0.0509	0.0519	0	0.0259	0.0322	0.0522
{3}	{3}:IPW	0.0009	0.0572	0.0718	0.0488	-0.0003	0.041	0.0509	0.0519	0	0.0259	0.0322	0.0522
{3}	{1, 3}	-0.0022	0.0437	0.0526	0.0606	-0.0015	0.0304	0.0373	0.0538	-0.0006	0.0192	0.0236	0.0559
{3}	{2, 3}	0.001	0.0451	0.0535	0.0618	0.0002	0.0312	0.0379	0.0547	0.0002	0.0194	0.024	0.0507
{3}	{1, 2, 3}	-0.0003	0.0382	0.0451	0.0647	-0.0005	0.0263	0.032	0.0581	-0.0001	0.0165	0.0203	0.0553
{1, 3}	{3}:CC	-0.0985	0.104	0.0718	0.2831	-0.0997	0.1007	0.0509	0.5027	-0.0994	0.0994	0.0322	0.8681
{1, 3}	{3}:IPW	-0.0003	0.0587	0.0738	0.051	-0.0009	0.0419	0.0523	0.0504	-0.0003	0.0266	0.0331	0.0516
{1, 3}	{1, 3}	-0.003	0.0456	0.0545	0.0656	-0.0018	0.0315	0.0388	0.0527	-0.0007	0.02	0.0246	0.055
{1, 3}	{1, 2, 3}	0.0001	0.041	0.0474	0.0713	-0.0002	0.0278	0.0338	0.0574	-0.0001	0.0176	0.0215	0.0557
{2, 3}	{3}:CC	0.1348	0.1365	0.0718	0.4644	0.1336	0.1338	0.0509	0.7443	0.1339	0.1339	0.0322	0.9857
{2, 3}	{3}:IPW	0.0022	0.0598	0.0748	0.0518	0.0004	0.0427	0.053	0.0513	0.0002	0.0271	0.0336	0.0525
{2, 3}	{2, 3}	0.0015	0.0424	0.0518	0.0587	0.0002	0.0299	0.0366	0.0533	0.0003	0.0187	0.0232	0.0539
{2, 3}	{1, 2, 3}	-0.0001	0.0373	0.0443	0.0647	-0.0006	0.0259	0.0314	0.0561	-0.0001	0.0161	0.02	0.0551
{1, 2, 3}	{3}:CC	0	0.0572	0.0718	0.0487	-0.0012	0.041	0.0509	0.0518	-0.0009	0.0259	0.0322	0.0526
{1, 2, 3}	{3}:IPW	0.0008	0.0593	0.0743	0.0515	-0.0003	0.0423	0.0526	0.0525	-0.0001	0.027	0.0333	0.0512
{1, 2, 3}	{1, 3}	-0.0018	0.0451	0.054	0.0629	-0.0011	0.0312	0.0384	0.0531	-0.0005	0.0199	0.0244	0.0557
{1, 2, 3}	{2, 3}	-0.0007	0.0456	0.0546	0.0603	-0.0009	0.0314	0.0385	0.0543	-0.0003	0.0196	0.0244	0.0498
{1, 2, 3}	{1, 2, 3}	0.0001	0.0387	0.0453	0.0653	-0.0003	0.0265	0.0322	0.0562	-0.0001	0.0167	0.0205	0.0551

Table 7: Reported are the bias (Mbias), absolute bias (Abias), standard deviation (Std) of the estimators and the size of a 5% nominal level two-sided t-test (Size) based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

MAR Sampling. Parameter of interest is the Slope ($\beta_{\lambda,2}$)

Target Popln. (λ)	Used Sample (s)	$n = 1000$				$n = 2000$				$n = 5000$			
		Mbias	Abias	Std	Size	Mbias	Abias	Std	Size	Mbias	Abias	Std	Size
$\{1\}$	$\{3\}$:CC	-0.0147	0.0625	0.0758	0.0619	-0.0153	0.0448	0.0537	0.0596	-0.0156	0.0301	0.0341	0.0747
$\{1\}$	$\{3\}$:IPW	0	0.0651	0.0796	0.0604	0.0001	0.046	0.0569	0.055	-0.0002	0.0293	0.0363	0.0511
$\{1\}$	$\{1,3\}$	0.0067	0.0526	0.0601	0.0806	0.003	0.0356	0.043	0.0604	0.0012	0.0222	0.0275	0.0534
$\{1\}$	$\{1,2,3\}$	0.0037	0.0505	0.0549	0.0921	0.002	0.033	0.0394	0.0647	0.0009	0.0208	0.0253	0.0557
$\{2\}$	$\{3\}$:CC	0.0244	0.0645	0.0758	0.0672	0.0238	0.0472	0.0537	0.0753	0.0235	0.0335	0.0341	0.1103
$\{2\}$	$\{3\}$:IPW	-0.0005	0.0855	0.0999	0.0797	-0.0009	0.0606	0.0727	0.0632	0.0001	0.0382	0.0472	0.0564
$\{2\}$	$\{2,3\}$	0.0063	0.0602	0.0641	0.1018	0.0029	0.0411	0.0469	0.0772	0.0018	0.0256	0.0306	0.0652
$\{2\}$	$\{1,2,3\}$	0.0018	0.0612	0.0629	0.1167	0.0001	0.041	0.0461	0.0846	0.0005	0.0254	0.0301	0.0672
$\{3\}$	$\{3\}$:CC	0.001	0.0613	0.0758	0.0544	0.0004	0.0431	0.0537	0.0506	0.0001	0.0274	0.0341	0.0468
$\{3\}$	$\{3\}$:IPW	0.001	0.0613	0.0758	0.0544	0.0004	0.0431	0.0537	0.0506	0.0001	0.0274	0.0341	0.0468
$\{3\}$	$\{1,3\}$	0.0084	0.0519	0.0601	0.0762	0.0037	0.0348	0.0425	0.0566	0.0016	0.0217	0.027	0.0527
$\{3\}$	$\{2,3\}$	-0.0019	0.0558	0.0624	0.0822	-0.0011	0.0365	0.0441	0.0615	-0.0004	0.0231	0.028	0.0564
$\{3\}$	$\{1,2,3\}$	0.0082	0.049	0.0534	0.0895	0.0045	0.0318	0.0381	0.065	0.0021	0.0199	0.0243	0.0556
$\{1,3\}$	$\{3\}$:CC	-0.0145	0.0625	0.0758	0.0617	-0.0151	0.0448	0.0537	0.0595	-0.0154	0.03	0.0341	0.0736
$\{1,3\}$	$\{3\}$:IPW	0.0003	0.0637	0.0782	0.0591	0.0001	0.0449	0.0557	0.0528	-0.0001	0.0286	0.0354	0.0497
$\{1,3\}$	$\{1,3\}$	0.0072	0.0519	0.0596	0.0779	0.0032	0.035	0.0424	0.0591	0.0013	0.0218	0.0271	0.054
$\{1,3\}$	$\{1,2,3\}$	0.0049	0.0495	0.0539	0.0916	0.0027	0.0323	0.0386	0.0642	0.0012	0.0203	0.0248	0.0563
$\{2,3\}$	$\{3\}$:CC	-0.0202	0.0635	0.0758	0.0658	-0.0208	0.0462	0.0537	0.0682	-0.0211	0.0323	0.0341	0.096
$\{2,3\}$	$\{3\}$:IPW	0.0003	0.0705	0.0847	0.068	-0.0003	0.0498	0.0609	0.0582	0.0002	0.0314	0.0391	0.0517
$\{2,3\}$	$\{2,3\}$	0.0031	0.0527	0.0577	0.0909	0.0015	0.0356	0.0417	0.0713	0.0011	0.0223	0.0269	0.0598
$\{2,3\}$	$\{1,2,3\}$	0.0049	0.05	0.0534	0.1013	0.0022	0.0334	0.0387	0.0751	0.0013	0.0207	0.025	0.0602
$\{1,2,3\}$	$\{3\}$:CC	-0.0517	0.0747	0.0758	0.1115	-0.0523	0.0618	0.0537	0.1724	-0.0526	0.0542	0.0341	0.3419
$\{1,2,3\}$	$\{3\}$:IPW	0	0.0638	0.0782	0.0611	-0.0002	0.045	0.0557	0.0531	0	0.0285	0.0355	0.0506
$\{1,2,3\}$	$\{1,3\}$	0.0025	0.0535	0.0612	0.0786	0.0002	0.0363	0.0435	0.0623	0.0001	0.0225	0.0278	0.0518
$\{1,2,3\}$	$\{2,3\}$	-0.004	0.0562	0.0617	0.0909	-0.0024	0.0374	0.0441	0.0688	-0.001	0.0236	0.0284	0.0593
$\{1,2,3\}$	$\{1,2,3\}$	0.0045	0.0466	0.0505	0.0946	0.0023	0.0308	0.0364	0.0678	0.0012	0.0192	0.0234	0.0552

Table 8: Reported are the bias (Mbias), absolute bias (Abias), standard deviation (Std) of the estimators and the size of a 5% nominal level two-sided t-test (Size) based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

Proofs of the results in Supplemental Appendix B

All the proofs are very similar to that of Proposition 1, which considered the case $d_m > d_\beta$ in detail. For brevity, we now take $d_m = d_\beta$ and, in the presentation below, we primarily focus on the verifications involved in Step 2.

Proof of Proposition 6: For Step 1, the difference is in the definition of the tangent space, and this is applicable to any choice of λ. For Step 2, however, we verify separately for $\lambda = \mathcal{C}$ and $\lambda = \{1\}$ that the (to be) proposed projection of the respective influence functions indeed belong in the tangent space.

STEP - 1: Conside a regular parametric sub-model indexed by θ for the joint distribution of the observed data $O = (C, T'_C(Z))'$. Its log density can be expressed in terms of the full data $(C, Z')'$ as

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

Let the true distribution be $f(O) = f_{\theta_0}(O)$ for some θ_0. Using the same notations as before, 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 $\dot{P}_\theta(C = r|.) := \frac{\partial}{\partial \theta} P_\theta(C = r|.)$. Thus the tangent space is characterized by functions of the form:

$$\mathcal{T} := a_1(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) a_r(Z_{(1)}, \ldots, Z_{(r)}) + \sum_{r=1}^{R} I(C = r) \frac{b_r(Z_{(1)}, \ldots, Z_{(r)})}{bb_r(Z_{(1)}, \ldots, Z_{(r)})},$$

where, respectively, $a_1(Z_{(1)}) \in L_0^2(F(Z_{(1)}))$, $a_r(Z_{(1)}, \ldots, Z_{(r)}) \in L_0^2(F(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}))$ for $r = 2, \ldots, R$, $\sum_{r=1}^{R} b_r(Z_{(1)}, \ldots, Z_{(r)})) = 0$, $\sum_{r=1}^{R} bb_r(Z_{(1)}, \ldots, Z_{(r)}) = 1$, and $\sum_{r=1}^{R} I(C = r) \frac{b_r(Z_1, \ldots, Z_r)}{bb_r(Z_1, \ldots, Z_r)} \in L_0^2(F(C|Z))$.

STEP - 2: $[d_m = d_\beta]$ Let us now consider the cases of $\lambda = \mathcal{C}$ and $\lambda = \{1\}$ separately.

Case $\lambda = \mathcal{C}$: Differentiating (1), in this case $E_\theta[m(Z; \beta_{\mathcal{C}}^0)] = 0$, with respect to θ at θ_0 we obtain that

$$\frac{\partial \beta_{\mathcal{C}}^0(\theta_0)}{\partial \theta'} = -M_{\mathcal{C}}^{-1} E\left[m(Z; \beta_{\mathcal{C}}^0) \frac{\partial \log f_{\theta_0}(Z)}{\partial \theta'}\right] = -M_{\mathcal{C}}^{-1} E\left[m(Z; \beta_{\mathcal{C}}^0) \left\{s(Z_{(1)}) + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})\right\}'\right],$$

where, as before, we omit the subscript θ_0 when a quantity (in the case the scores) is evaluated at $\theta = \theta_0$. We will also write $m(Z; \beta_{\mathcal{C}}^0) \equiv m(T_R(Z); \beta_{\mathcal{C}}^0)$ as m, and $T_r(Z)$ as T_r for $r = 1, \ldots, R$ for notational brevity.

To show pathwise differentiability as in the proof of Proposition 1, it suffices to verify that

$$E[\bar{\varphi}_{\mathcal{C}}(O) S(O)'] = E\left[m(Z) \left\{s(Z_{(1)}) + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})\right\}'\right]. \tag{26}$$

Note that, the LHS of (26) $= \sum_{q=1}^{R} B_q$ where

$$\begin{aligned} B_1 &:= E\left[E[m|T_1] S(O)'\right] \\ B_q &:= E\left[\frac{I(C \geq q)}{P(C \geq q|T_q)} \left(E[m|T_q] - E[m|T_{q-1}]\right) S(O)'\right] \text{ for } q = 2, \ldots, R. \end{aligned}$$

First, from the definition of $S(O)$ above and since $I(C \geq r) = 1 - I(C \leq r-1)$:

$$\begin{aligned} B_1 &= E[E[m|T_1]s(T_1)'] + \sum_{r=2}^{R} E[I(C \geq r)E[m|T_1]s(Z_{(r)}|T_{r-1})'] + \sum_{r=1}^{R} E\left[\frac{I(C=r)}{P(C=r|T_r)}E[m|T_1]\dot{P}(C=r|T_r)'\right] \\ &= E[ms(T_1)'] + \sum_{r=2}^{R} E[(1 - I(C \leq r-1))E[m|T_1]E[s(Z_{(r)}|T_{r-1})'|T_{r-1}]] + E[E[m|T_1]\sum_{r=1}^{R}\dot{P}(C=r|T_r)'] \\ &= E[ms(Z_{(1)})'] + 0 + 0 \end{aligned}$$

where the first term in the last line follows by the definition of T_1; the second term follows by using MAR in (3) and the definition of conditional score that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ while noting the definition of T_{r-1} for $r = 2, \ldots, R$ along with (22); and the third term follows by noting that $\sum_{r=1}^{R}\dot{P}(C=r|T_r) = 0$ since, by definition $1 = \sum_{r=1}^{R} P(C=r|T_R) = \sum_{r=1}^{R} P(C=r|T_r)$ where the last equality is due to MAR in (3).

Now we consider B_q and make two observations. First, since $E[E[m|T_q]|T_{q-1}] = E[m|T_{q-1}]$, we observe that:

$$E\left[\frac{I(C \geq q)}{P(C \geq q|T_q)}\left(E[m|T_q] - E[m|T_{q-1}]\right)s(T_1)'\right] = 0$$

by using (23). Furthermore, in the second line below, conditioning on T_R and using MAR in (3) give:

$$\begin{aligned} &\sum_{r=1}^{R} E\left[\frac{I(C \geq q)}{P(C \geq q|T_q)}\left(E[m|T_q] - E[m|T_{q-1}]\right)\frac{I(C=r)}{P(C=r|T_r)}E[m|T_1]\dot{P}(C=r|T_r)'\right] \\ = &\sum_{r=q}^{R} E\left[\frac{I(C=r)}{P(C \geq q|T_q)}\left(E[m|T_q] - E[m|T_{q-1}]\right)\frac{1}{P(C=r|T_r)}E[m|T_1]\dot{P}(C=r|T_r)'\right] \\ = &E\left[\left(E[m|T_q] - E[m|T_{q-1}]\right)\frac{\dot{P}(C \geq q|T_q)'}{P(C \geq q|T_q)}\right] = 0, \end{aligned}$$

which is the second observation. It follows by taking expectation conditional on T_{q-1} since, by (22), $P(C \geq q|T_q) = P(C \geq q|T_{q-1})$ while $\dot{P}(C \geq q|T_q) = -\dot{P}(C \leq q-1|T_{q-1})$ by noting that $\sum_{r=1}^{R}\dot{P}(C=r|T_r) = 0$. Based on the first and the second observations above, we have as we did in Step 2 of the proof of Proposition 1:

$$\begin{aligned} B_q &= \sum_{r=1}^{q-1} E\left[\frac{I(C \geq q)}{P(C \geq q|T_q)}\left(E[m|T_q] - E[m|T_{q-1}]\right)s(Z_{(r)}|T_{r-1})'\right] \\ &\quad + \sum_{r=q}^{R} E\left[\frac{I(C \geq r)}{P(C \geq q|T_q)}\left(E[m|T_q] - E[m|T_{q-1}]\right)s(Z_{(r)}|T_{r-1})'\right]. \end{aligned}$$

The first term on the RHS is 0 by using (23) since $E[E[m|T_q]|T_{q-1}] = E[m|T_{q-1}]$. On the other hand, (22) and the fact that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ imply that the second term on the RHS is

$$\sum_{r=q}^{R} E\left[\left(E[m|T_q] - E[m|T_{q-1}]\right)s(Z_{(r)}|T_{r-1})'\right] = E\left[\left(E[m|T_q] - E[m|T_{q-1}]\right)s(Z_{(q)}|T_{q-1})'\right] = E[ms(Z_{(q)}|T_{q-1})'].$$

Therefore, $q = 2, \ldots, R$ we have $B_q = E[ms(Z_{(q)}|T_{q-1})']$, combining which with B_1 verifies (26).

That $-M_{\mathcal{C}}^{-1}\bar{\varphi}_{\mathcal{C}}(O) \in \mathcal{T}$ follows by matching the first R terms of both (while noting that the properties for

the functions $a_1(T_1), \ldots, a_r(T_r)$ are satisfied by the corresponding terms in $-M_{\mathcal{C}}^{-1}\bar{\varphi}_{\mathcal{C}}(O)$), and finally taking the terms corresponding to the remaining ones in $\mathcal{T}$ as zeros in $-M_{\mathcal{C}}^{-1}\bar{\varphi}_{\mathcal{C}}(O)$.

Case $\lambda = \{1\}$: Differentiating (1), in this case $E_\theta[m(Z;\beta^0_{\{1\}})|C=1] = 0$, with respect to θ at θ_0 we obtain that

$$\begin{aligned}
\frac{\partial \beta^0_{\{1\}}(\theta_0)}{\partial \theta'} &= -M_{\{1\}}^{-1} E\left[m \frac{\partial \log f_{\theta_0}(Z)}{\partial \theta'} \middle| C=1 \right] \\
&= -M_{\{1\}}^{-1} E\left[m \left\{ s(T_1|C=1) + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1}, C=1) \right\}' \middle| C=1 \right] \\
&= -M_{\{1\}}^{-1} E\left[m \left\{ s(T_1|C=1) + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1}) \right\}' \middle| C=1 \right]
\end{aligned}$$

where $s(T_1|C=1)$ is the derivative of $\log f_\theta(T_1|C=1)$ with respect to θ at θ_0. Thus $E[s(T_1|C=1)|C=1] = 0$. Also, note that $C=1$ vanished in the last line from $s(Z_{(r)}|T_{r-1}, C=1)$ (the conditional scores, as defined before, but also conditional on $C=1$) because of MAR in (3) since the conditioning set already includes T_1. To see the connection between $s(T_1)$ that we have been using until now and $s(T_1|C=1)$ that was introduced above, note that the joint distribution of $(I(C=1), T_1')'$ gives the following two equivalent factorization of the score function:

$$\begin{aligned}
& s(T_1) + I(C=1)\frac{\dot{P}(C=1|T_1)}{P(C=1|T_1)} + I(C>1)\frac{\dot{P}(C>1|T_1)}{P(C>1|T_1)} \\
= \;& I(C=1)\left[\frac{\dot{P}(C=1)}{P(C=1)} + s(T_1|C=1) \right] + I(C>1)\left[\frac{\dot{P}(C>1)}{P(C>1)} + s(T_1|C>1) \right]. \qquad (27)
\end{aligned}$$

Also, as before, we always omit the subscript θ_0 when a quantity (in the case the scores) is evaluated at $\theta = \theta_0$; and write $m(Z;\beta^0_{\{1\}}) \equiv m(T_R(Z);\beta^0_{\{1\}})$ as m, and $T_r(Z)$ as T_r for $r = 1, \ldots, R$ for notational brevity.

To show pathwise differentiability as in the proof of Proposition 1, it suffices to verify that

$$E[\bar{\varphi}_{\{1\}}(O)S(O)'] = E\left[m \left\{ s(T_1|C=1) + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1}) \right\}' \middle| C=1 \right]. \qquad (28)$$

Noting that: (i) the only difference between the score function $S(O)$ in Proposition 1 and Proposition 6 is the additional third set of terms in the latter, and (ii) the similarity in $\varphi_{\{1\}}(O)$ and $\bar{\varphi}_{\{1\}}(O)$, (so we will omit the concerned steps from the previous proof to avoid repetition) the above verification will follow by showing that

$$E\left[\frac{I(C=1)}{P(C=1)} E[m|T_1] S(O)' \right] = E[m s(T_1|C=1)'|C=1], \qquad (29)$$

$$\sum_{q=2}^{R}\sum_{r=1}^{R} E\left[\frac{I(C \geq q)}{P(C \geq q|T_q)} \left[\varphi_{q,\{1\}}(O) - \varphi_{q-1,\{1\}}(O) \right] I(C=r) \frac{\dot{P}_\theta(C=r|T_r)}{P_\theta(C=r|T_r)} \right] = 0. \qquad (30)$$

(30) is obvious exactly following the steps that led to the second observation in the proof of the case for $\lambda = \mathcal{C}$ above with $\varphi_{q,\{1\}}(O)$ here playing the role of $E[m|T_q]$ there.

Now consider the LHS of (29) and first note that $\sum_{r=2}^{R} E[\frac{I(C=1)}{P(C=1)} E[m|T_1] I(C \geq r) s(Z_{(r)}|T_{r-1})'] = 0$ trivially,

whereas $E\left[\frac{I(C=1)}{P(C=1)}E[m|T_1]s(T_1)'\right] = E[ms(T_1)'|C=1]$. Therefore,

$$\begin{aligned}
E\left[\frac{I(C=1)}{P(C=1)}E[m|T_1]S(O)'\right] &= E\left[ms(T_1)'\,|\,C=1\right] + \sum_{r=1}^{R} E\left[\frac{I(C=1)}{P(C=1)}E[m|T_1]I(C=r)\frac{\dot{P}(C=r|T_r)'}{P(C=r|T_r)}\right] \\
&= E\left[\frac{I(C=1)}{P(C=1)}ms(T_1)' + \frac{I(C=1)}{P(C=1)}E[m|T_1]\frac{\dot{P}(C=1|T_1)'}{P(C=1|T_1)}\right] \\
&= E\left[\frac{I(C=1)}{P(C=1)}m\left\{s(T_1) + \frac{\dot{P}(C=1|T_1)}{P(C=1|T_1)}\right\}'\right] \text{ [by (3)]} \\
&= E\left[\frac{I(C=1)}{P(C=1)}m\left\{\frac{\dot{P}(C=1)}{P(C=1)} + s(T_1|C=1) - \frac{\dot{P}(C=1|T_1)}{P(C=1|T_1)} + \frac{\dot{P}(C=1|T_1)}{P(C=1|T_1)}\right\}'\right]
\end{aligned}$$

where the last line is obtained by substituting for $I(C=1)s(T_1)$ using (27). However, $\left[\frac{I(C=1)}{P(C=1)}m\frac{\dot{P}(C=1)'}{P(C=1)}\right] = 0$ by (1) when $\lambda = \{1\}$, whereas $E\left[\frac{I(C=1)}{P(C=1)}s(T_1|C=1)'\right] = E[ms(T_1|C=1)'|C=1]$, and thus (29) is now verified.

That $-M_{\{1\}}^{-1}\bar{\varphi}_{\{1\}}(O)$ belongs to $\mathcal{T}$ in (24) can be shown as follows. (i) Match the term $a_r(T_r)$ in $\mathcal{T}$ with the r-th term of $-M_{\{1\}}^{-1}\bar{\varphi}_{\{1\}}(O)$ for $r > 1$. (ii) Distribute the first term $s(Z_1)$ in $\mathcal{T}$ according to the relation (27) and match the term $I(C=1)s(T_1|C=1)$ with the first term of $-M_{\{1\}}^{-1}\bar{\varphi}_{\{1\}}(O)$ while keeping in mind that, by definition, $s(T_1|C=1) \in L_0^2(F(T_1|C=1))$. It is straightforward to verify that all the corresponding conditional expectations, as required by the definition of $\mathcal{T}$ and also (27), are zeros. Rest of the terms in $\mathcal{T}$ (including the one due to the distribution of terms in (ii)) are represented in $-M_{\{1\}}^{-1}\bar{\varphi}_{\{1\}}(O)$ by zeros. ■

Proposition 7 is a special case of Proposition 1, so the proof is omitted. We now prove Propositions 8 and 9 in reverse order to maintain the continuity of presentation from the above proof. The proofs of Corollaries 10-12 involve straightforward but tedious algebraic manipulations. We omit then for brevity but note that the proofs for the results in (f1) and (f2) require a minor change in the proofs of the original propositions that amounts to recognizing that for $j=1$ and $j=2$: $E[m] = E\left[\frac{I(C\in\{j,3\})}{P(C\in\{j,3\}|T_1)}m\right] = E\left[\frac{P(C\in\{j,3\})}{P(C\in\{j,3\}|T_1)}m\Big|\, C\in\{j,3\}\right]$.

Proof of Proposition 9: The proof here for a general $\lambda \in \Lambda$ is essentially the same as that of Proposition 6 for $\lambda = \{1\}$. CMAR in (4) adds a little more restriction on the tangent space, and allows what was done for $\lambda = \{1\}$ earlier to be done for all λ here. By utilizing CMAR, the score function and the tangent space are respectively:

$$\begin{aligned}
S_\theta(O) &= s_\theta(T_1) + \sum_{r=2}^{R} I(C\geq r)s_\theta(Z_{(r)}|T_{r-1}) + \sum_{r=1}^{R} I(C=r)\frac{\dot{P}_\theta(C=r|T_1)}{P_\theta(C=r|T_1)}, \\
\mathcal{T} &:= a_1(T_1) + \sum_{r=2}^{R} I(C\geq r)a_r(T_r) + \sum_{r=1}^{R} I(C=r)\frac{b_r(T_1)}{bb_r(T_1)},
\end{aligned}$$

where each term has the same property as before except that now additionally $\sum_{r=1}^{R} I(C=r)\frac{b_r(T_1)}{bb_r(T_1)} \in L_0^2(F(C|T_1))$.

For a given $\lambda \in \Lambda$, the following relation (similar to (27)) obtained by the two different factorization of the joint distribution of $(I(C\in\lambda), T_1)$ helps us to conveniently switch between the different factorizations (and quantities):

$$\begin{aligned}
& s(T_1) + I(C\in\lambda)\frac{\dot{P}(C\in\lambda|T_1)}{P(C\in\lambda|T_1)} + I(C\notin\lambda)\frac{\dot{P}(C\notin\lambda|T_1)}{P(C\notin\lambda|T_1)} \\
= \;& I(C\in\lambda)\left[\frac{\dot{P}(C\in\lambda)}{P(C\in\lambda)} + s(T_1|C\in\lambda)\right] + I(C\notin\lambda)\left[\frac{\dot{P}(C\notin\lambda)}{P(C\notin\lambda)} + s(T_1|C\notin\lambda)\right]. \qquad (31)
\end{aligned}$$

As before, but now using CMAR (unlike using the choice $\lambda = \{1\}$ as in Proposition 6), we obtain:

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} = -M_\lambda^{-1} E\left[m \left\{ s(T_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \middle| C \in \lambda \right]$$

and then verify that

$$E[\varphi_{\lambda[u]}(O) S(O)'] = E\left[m \left\{ s(T_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \middle| C \in \lambda \right].$$

Since $P(C = r|T_R) = P(C = r|T_1)$ by CMAR, simple inspection reveals that the only minor change in the proof now is in showing that $E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m|T_1] S(O)'\right] = E[m s(T_1|C \in \lambda)|C \in \lambda]\Big]$, and this will follow by (27) since

$$\begin{aligned}
E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m|T_1] S(O)'\right] &= E\left[m s(T_1)' \middle| C \in \lambda \right] + 0 + \sum_{r=1}^{R} E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m|T_1] I(C = r) \frac{\dot{P}(C = r|T_1)'}{P(C = r|T_1)}\right] \\
&= E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} m \left\{ s(T_1) + \frac{\dot{P}(C \in \lambda|T_1)}{P(C \in \lambda|T_1)} \right\}'\right] \text{ [by (3)]} \\
&= E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} m \left\{ \frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda) \right\}'\right] \text{ [by (31)]}.
\end{aligned}$$

Once again, since $\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} m \frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)}\right] = 0$ by (1) whereas $E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} s(T_1|C \in \lambda)'\right] = E[m s(T_1|C \in \lambda)'|C \in \lambda]$, the required verification is done. ■

Proof of Proposition 8: The references in the steps of this proof are to mainly to that of Proposition 7 (i.e., effectively to that of Proposition 1) and to that of Proposition 9.

As before, we obtain the score function for a parametric sub-model indexed by θ as

$$S_\theta(O) = s_\theta(T_1) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_{(r)}|T_{r-1}) + \sum_{r=1}^{R} \frac{I(C = r)}{P(C = r|T_1)} \left(\frac{\partial P(C = r|T_1; \gamma^0)}{\partial \gamma'} \frac{\partial \gamma^0}{\partial \theta'} \right)'.$$

Recall that $S_\gamma(C|T_1) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|T_1)} \frac{\partial}{\partial \gamma} P(C = r|T_1; \gamma^0)$. Let b denote a constant matrix with dimension same as that of $\frac{\partial \gamma^0}{\partial \theta'}$. Then the tangent set for the model is then characterized by the set of functions:

$$\mathcal{T} := a_1(T_1) + b' S_\gamma(C|T_1) + \sum_{r=2}^{R} I(C \geq r) a_r(T_r),$$

where $a_1(T_1) \in L_0^2(F(T_1))$, $S_\gamma(C|T_1) \in L_0^2(F(C|T_1))$ and $a_r(T_r) \in L_0^2(F(Z_{(r)}|T_{r-1}))$.

Recognizing that $P(C = r|T_1) = P(C = r|T_1; \gamma^0)$ is known up to the finite (d_γ) dimensional parameter γ, alters the relationship in (31) as follows

$$\begin{aligned}
& s(T_1) + \frac{\partial \gamma^{0\prime}}{\partial \theta} \left[I(C \in \lambda) \frac{\frac{\partial}{\partial \gamma} P(C \in \lambda|T_1; \gamma^0)}{P(C \in \lambda|T_1)} + I(C \notin \lambda) \frac{\frac{\partial}{\partial \gamma} P(C \notin \lambda|T_1; \gamma^0)}{P(C \notin \lambda|T_1)} \right] \\
= \; & I(C \in \lambda) \left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda) \right] + I(C \notin \lambda) \left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(T_1|C \notin \lambda) \right].
\end{aligned}$$

As before, differentiating (2) under the integral and using the above relationship give:

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} = -M_\lambda^{-1} E\left[\frac{P(C\in\lambda|T_1)}{P(C\in\lambda)} m\left\{s(T_1)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})'\right\}\right] - M_\lambda^{-1} E\left[E[m|T_1]\frac{\frac{\partial}{\partial\gamma'}P(C\in\lambda|T_1;\gamma^0)}{P(C\in\lambda)}\frac{\partial\gamma^0}{\partial\theta'}\right].$$

Therefore, utilizing the expression of the efficient influence function in Proposition 7 and its relation to that in Proposition 8, the verification of pathwise differentiability essentially boils to verifying that

$$E\left[\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1(Z)]\middle| S_\gamma(C|T_1(Z))\right) S(O)'\right] = E\left[E[m|T_1]\frac{\frac{\partial}{\partial\gamma'}P(C\in\lambda|T_1;\gamma^0)}{P(C\in\lambda)}\frac{\partial\gamma^0}{\partial\theta'}\right].$$

Note that $E\left[S_\gamma(C|T_1)\left\{s(T_1)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})'\right\}\right] = 0$ by using (term by term) that $E[S_\gamma(C|T_1)|T_1] = 0$ for term one; $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$, and then using (4) for the rest. Therefore, in the above equation (that contains the equality relationship to be verified), the LHS simplifies as

$$\begin{aligned}
LHS &= E\left[\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]\middle| S_\gamma(C|T_1)\right) S_\gamma(C|T_1)'\right]\frac{\partial\gamma^0}{\partial\theta'} \\
&= E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]S_\gamma(C|T_1)'\right]\frac{\partial\gamma^0}{\partial\theta'} \\
&= E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]\sum_{r=1}^{R}\frac{I(C=r)}{P(C=r|T_1)}\frac{\partial P(C=r|T_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'} \\
&= E\left[\frac{1}{P(C\in\lambda)}E[m|T_1]\sum_{r\in\lambda}\frac{P(C=r|T_1)}{P(C=r|T_1)}\frac{\partial P(C=r|T_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'} \\
&= E\left[\frac{1}{P(C\in\lambda)}E[m|T_1]\frac{\partial P(C\in\lambda|T_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'} \\
&= RHS. \ \blacksquare
\end{aligned}$$