A Note on Efficiency Gains from Multiple Incomplete Subsamples*

Saraswata Chaudhuri†

Current version: June 20, 2016. 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 $\widehat{\beta}$) is proposed. This GMM estimation involves unknown nuisance parameters whose estimation turns out to have only mild effects on the asymptotic properties of $\widehat{\beta}$. In particular, $\widehat{\beta}$ remains consistent and asymptotically normal under standard conditions if the estimator of the nuisance parameters converges in probability to any member of a wide class of functions which is not necessarily the truth. Moreover, $\widehat{\beta}$ is efficient if this limit function, irrespective of the rate of convergence to it, is the true nuisance parameters. These offer flexibility in practice for parametric or nonparametric estimation of the nuisance parameters. Several issues useful for practical implementation are discussed. Simulation evidence of the efficiency gains from using all the sub-samples, and the finite-sample behavior of $\widehat{\beta}$ is provided.

JEL Classification: C13; C14; C31.

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

*We are grateful to 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 U. Sydney, U. New South Wales, U. Canterbury, West Virginia University, McGill (Econ and Biostat), Concordia, Queen's, U. Montreal and the Midwest Econometrics Group meetings (2013) for helpful discussion and comments on various versions of the paper.

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

1 Introduction

Empirical work is often adversely affected by data that are missing due to survey non-response or measured with error. While big budgets for surveys could ideally solve such problems, cost considerations are important. They are perhaps even more relevant now because recent empirical research in economics is often based on data collected from field or laboratory experiments conducted by researchers facing varying degrees of budget constraint.

There are several well known approaches of dealing with the problems related to missing data or data measured with error. See, e.g., Carroll et al. (1995) and Little and Rubin (2002) for textbook treatments of the topic.[1]

Our paper considers the approach of planned incompleteness in the data whereby imperfect/mismeasured and less expensive variables are obtained for all the units in the sample, whereas survey-resources are spent to obtain perfect data – nothing missing or mismeasured – for a part of the sample. All of it happens by sampling design. We refer to the collection of sample units with perfect data as the complete sub-sample and the rest as incomplete sub-samples. As will happen in multi-phase surveys, there can be multiple distinct incomplete sub-samples. We allow the underlying populations for the sub-samples (call them sub-populations) to be systematically different depending on the sampling design, and the parameters of interest for the empirical work to be defined in terms of unions of some or all such sub-populations. (The idea behind these sub-populations is the same as Rubin (1977)'s enhancement of the classical double-sampling through the introduction of mixtures; also called pattern-mixtures by Little (1993, 1994).) Identification of these parameters is not a problem under planned incompleteness. Instead, the issue is how to combine the complete and incomplete sub-samples optimally to obtain efficient and easy to compute estimates for the parameters of interest. This is the topic and focus of our paper.

While even the long and short form census generate what can be viewed as planned complete and incomplete sub-samples, of interest to us are surveys of (much) smaller scale and where the incomplete sub-samples contain substantial relevant information so that their inclusion in the analysis could result in meaningful efficiency gains. Consider three well known scenarios from the recent literature for an initial idea on the premise of our discussion.

First, in laboratory experiments to elicit risk aversion and study its dependence on the size of the stake, budget constraint may necessitate that the experiment be conducted with real low stake and hypothetical high stake on a large group of subjects from a representative population whereas with higher real stakes on smaller subsets of the original subjects. See Holt and Laury (2002) for a variation of this strategy where, instead of the smaller subset, an entirely new small group of subjects was chosen for the high stakes (50 and 90$\times$ low stake).

Second, 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. The discussion focuses on the tradeoff in the choice of N (number of subjects) and

[1] One approach is to go back to the field and recollect the data that were missing or measured with error. For example, since 1992 the Panel Study of Income Dynamics initiated attempts to re-contact former non-response sample cases. This alleviated the problem, although partially since, starting in 1993, a family was no longer followed once it was a non-response for two consecutive waves. A second approach accepts the problem with the data and focuses on the development and use of sophisticated statistical tools to estimate the relevant parameters of interest either by allowing for them to be partially identified or by enforcing point identification by maintaining assumptions such as selection on observables or unobservables (but with some suitable exclusion restriction) that are ultimately non-testable without further strong assumptions. Both approaches attempt to correct for the data-problem ex-post and their effectiveness varies from case to case. In this paper we consider a third approach that deals with this problem ex-ante.

T (number of measurements including baseline and followups) at a given cost. Alternatively, it is not unreasonable to keep both N and T large but measure the relevant variables only for a subset of subjects at each followup.

Third, 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 from the respondents 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, given the cost concerns it seems reasonable to obtain the good but expensive measures for only subsets of subjects, and the other measures for everyone.

Finally, consider a rather simple example demonstrating how, under cost considerations, it may be beneficial to follow the approach of planned incompleteness of data by which we obtain the sample $\mathcal{S}$ below.

Example 1: Let (Y, X) be scalar random 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 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 $Var(\widehat{\beta}^\dagger) = \Delta/n^\dagger$ where $\Delta = Var(Y - X)$. On the other hand, a special case of our results gives an estimator of β based on $\mathcal{S}$ as:[3]

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

Now, let the cost of observing Y for an unit be 1 and that for X be c where $c > 1$. Let the allowed total expected cost for the sample be c^*, and thus $n^\dagger = \lfloor c^*/(1+c) \rfloor$ and $n = \lfloor c^*/(1+pc) \rfloor$ for a given c, c^* and p 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 $q = \Delta/E[Var(X|Y)] - 1$. 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 $p = 1/\sqrt{cq}$. ■

Such explicit optimality arguments will not be possible under the general setup that we consider in this paper.[5] Nevertheless, the approach of planned incompleteness of data to reduce the burden on the respondent (and the

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

surveyor) and thereby decrease the chance of unplanned non-response or measurement error that substantially complicates the analysis, has long been recognized as a cost-effective strategy and has been employed in various fields of research where data collection by the researcher has traditionally been more prevalent than in economics. Consider the following examples. The two/many method/measurement design is used in the 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 (CPS)) in economics. In 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-form, mainly 2D and 3D forms (e.g. the Occupational Information Network survey), design discussed by Graham et al. (1996), Graham et al. (2006), etc. in psychology and behavioral research.

Our paper generalizes the idea behind planned incompleteness in several directions and proposes a simple to implement efficient estimation strategy for the concerned parameters of interest under a broad class of estimation framework. To provide an overview of the generalization, let us consider a running example for concreteness.

Running Example: Consider two variables y and X that are, respectively, cheap and expensive (e.g. consumption according to regularly monitored personal diaries) to observe. Consider two other variables X_c and X_e that respectively denote a cheap (c) mismeasured X (e.g. seven day recall) and a relatively more expensive (e) mismeasured X (e.g. household diaries). Both X_c and X_e are related to X. They can also depend on y; precisely, they need not be simply surrogates. Define $Z_{(1)} = (y, X_c')'$, $Z_{(2)} = X_e$ and $Z_{(3)} = X$. Consider a sample where we observe $Z_{(1)}$ for all the units, $Z_{(2)}$ for a subset and $Z_{(3)}$ for a further subset. Hence, there are three sub-samples. One is complete. The two incomplete sub-samples respectively contain information on $Z_{(1)}, Z_{(2)}$ and only $Z_{(1)}$. The information contained in a sample unit decreases in the order the sub-samples are listed above. ■

This monotone pattern in the information content and the existence of the complete sub-sample are by design, and allow flexibility in defining the parameters of interest while ensuring their easy and efficient estimation without imposing strong parametric assumptions. A monotone pattern will exist here by definition if, e.g. there is no $Z_{(2)}$. It can also arise naturally in rotating panels, e.g., the rotation scheme in the CPS ensures that in any single month, one-eighth of the housing units are interviewed for the first time, another eighth for the second time, and so on. PQD does not impose the monotone pattern, e.g. it can accommodate for the sub-sample where we also observe $(y', X_e')'$. While Chatterjee and Li (2010) propose efficient estimation with a similar simple PQD under additional assumptions, efficient estimation quickly becomes infeasible under the generalizations that we are interested in this paper. (See Chaudhuri and Guilkey (2016) for a similar result in a general context and under weaker assumptions.) SQD and multi-form design do not require the monotone pattern or the complete

sub-sample; however, nor do they handle anything but independent (of $Z_{(1)}, Z_{(2)}, Z_{(3)}$) allocation of units to the various sub-samples. Our paper will allow for the dependence for the sake of generality and, more importantly, because it allows to fully exploit the multi-phase sequential nature of the sampling that can be useful in practice.[6]

Allowing for such dependence, however, makes the exact mechanism of obtaining the sub-samples important and opens up the possibility that the sub-populations corresponding to the sub-samples are systematically different. In the context of the running example, the latter means that the joint distribution of y and X can be different in these sub-populations, and thus, features of each of these joint distributions can be of interest at least for the descriptive purpose of the empirical work [also see e.g. page 126 of Little (1993)]. Let us elaborate.

Case 1: Consider the standard multi-phase sampling. Start with n sample units from a population. In phase 1, collect $Z_{(1)}$ for all. In phase 2, select a subset of units and collect $Z_{(2)}$. Let the size of this subset be $n - n_1$. In phase 3, for a further subset of units collect $Z_{(3)}$. Let its size be n_3. Therefore we now have n_1 units with only $Z_{(1)}$ (sub-sample 1), $n_2 := n - n_1 - n_3$ units with $Z_{(1)}, Z_{(2)}$ (sub-sample 2), and n_3 units with all $Z_{(1)}, Z_{(2)}, Z_{(3)}$ (sub-sample 3). This is a case of progressive enrichment of the original sample units. The selection of units in phases 2 and 3 can be done: (i) arbitrarily (e.g. tossing a coin), or (ii) depending on the value of $Z_{(1)}$ for the unit, or (iii) phase 2 selection can depend on the value of $Z_{(1)}$ for the unit while phase 3 on the value of $Z_{(1)}, Z_{(2)}$ for the unit. The last one generates what is commonly known as the missing at random (MAR) data. The second one also generates MAR data but is more restrictive. This can be useful in practice when there is not much time lag between phases 2 and 3 and hence the sampling design for these phases are done at the end of phase 1, i.e, before collecting $Z_{(2)}$ at all. We refer to it as the convenient MAR (CMAR) sampling. The first one, in spirit, corresponds to SQD, multi-form, rotating panels, etc. and we refer to it as the independent (INDEP) sampling. (i) (trivially), (ii) and (iii) are all examples of selection on observables. The original population is the only reasonable target under (i). However, under (ii) and (iii) the sub-populations are also interesting; after all, these sub-populations are closely associated with the sampling design made by the researcher/surveyor.

Case 2: There is a sample of size n of which n_3 units contain $Z_{(1)}$, $Z_{(2)}$, $Z_{(3)}$ (sub-sample 3) and $n - n_3$ units contain only $Z_{(1)}$. Given extra budget, collect $Z_{(2)}$ for a subset of the latter $n-n_3$ units and thus enrich this subset (of size, say, n_2) that now forms sub-sample 2 while the rest (of size, say, $n_1 = n - n_3 - n_2$) forms sub-sample 1.

Case 3: There is a sample of size $n_1 + n_3$ of which n_3 units contain $Z_{(1)}$, $Z_{(2)}$, $Z_{(3)}$ (sub-sample 3) and n_1 units contain only $Z_{(1)}$ (sub-sample 1). Given extra budget, collect $Z_{(1)}$, $Z_{(2)}$ for $n_2 = n - n_1 - n_2$ additional units (sub-sample 2). Initial interest generally lies in the combined sub-populations for sub-samples 1 and 3.

There are various other possibilities besides these three cases. However, the actual cost of observing variables is

[6] Consider a simple example demonstrating the usefulness in terms of efficiency of dependent instead of independent missing data. **Example 2:** 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 that 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 $D = 1$. 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 of β based on sample units with $D_i = 1$ and its asymptotic variance are, respectively: $\widehat{\beta} = \sum_{i=1}^n D_i X_i \left(Y_i - \sum_j D_j Y_j / \sum_j D_j\right) \Big/ \sum_{i=1}^n D_i X_i \left(X_i - \sum_j D_j X_j / \sum_j D_j\right)$ and $\text{Avar} = \sigma^2 / E[DX]\,(1 - E[DX]/E[D]) = p\sigma^2 / [qp(1)(p - qp(1))]$. If $P(D = 1|Y, X) = P(D = 1) = p$, implying $p(1) = p(0) = p$, then $\text{Avar} = \sigma^2/pq(1-q)$. On the other hand, $p(1) = p/(2q)$ minimizes the general Avar and the minimized value is $\text{Avar} = 4\sigma^2/p$, 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 (say, treatment) in the population – the essential idea behind stratification – to minimize variance. ■

the key for any of these cases to make sense in practice. Hence the decision to pursue a particular sampling strategy instead of the others has to be situation-specific, which is not unreasonable for surveys generally associated with field or laboratory experiments designed to study specific research questions. It is also not unreasonable for some version of selection on observables – MAR, CMAR, INDEP – to hold in all these cases and thereby link the sub-samples and make it possible to combine the sub-samples for the purpose of efficient estimation.

Our paper generalizes this idea and considers an arbitrary but finite number of R sub-samples which, in turn, allow for the parameters of interest to be defined in terms of any of the 2^R-1 possible unions of possibly R different sub-populations. We consider efficient estimation by combining the information from all the sub-samples. The idea of data combination in related settings has also appeared frequently in the recent economics literature. See e.g. Ichimura and Martinez-Sanchis (2005), Ridder and Moffitt (2007), Devereux and Tripathi (2009), Tripathi (2009), Abrevaya and Donald (2011), Muris (2014), Graham et al. (2016), and the references therein.

The rest of the paper is organized as follows. Section 2 establishes the efficiency bound for the parameters of interest (call them β) and illustrates how the information content of each sub-sample gets combined optimally for efficient estimation. Section 3 considers efficient semiparametric GMM estimation (call the estimator $\widehat{\beta}$). The estimation framework has special features that make semiparametric or even parametric estimation much less burdensome than usual. In particular, it allows the estimator of the nuisance parameters (call it $\widehat{h}$) to converge in probability to functions other than the truth without affecting the consistency of $\widehat{\beta}$; while a convergence of $\widehat{h}$ to the truth, irrespective of the rate, ensures efficiency for $\widehat{\beta}$. Useful issues for practical implementation are discussed. A Monte Carlo experiment in Section 4 documents what happens in finite samples. Section 5 concludes. The main results are proved in Technical Appendix A. Auxiliary results with proofs are collected in Technical Appendix B.

2 Framework and Combination of Sub-samples

2.1 Framework

Let $Z = (Z'_{(1)}, \ldots, Z'_{(R)})'$ be a random vector with $d_r \times 1$ element $Z_{(r)}$ for $r = 1, \ldots, R$ where $\sum_{r=1}^{R} d_r$ is finite. To model 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 the dimension of $T_C(Z)$, 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, i.e., $\{i = 1 \ldots, n : C_i = r\}$ with 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 rest are incomplete with a monotone pattern $T_1(Z) \subset T_2(Z) \subset \ldots \subset T_{R-1}(Z) \subset T_R(Z) = Z$ in their information content.

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:

$$E[m(Z; \beta)|C \in \lambda] = 0 \text{ for } \beta \in \mathcal{B} \iff \beta = \beta^0_\lambda. \tag{2.1}$$

β_λ^0 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., $I(C = R)m(Z;\beta)$ is the feasible version of the moment vector without further assumptions. However, point identification of β_λ^0 is still technically possible by the Horvitz-Thompson re-weighting under a suitable overlap assumption, $P(C = R|T_R(Z)) > 0$ almost surely in $T_R(Z)$, since for any given β:

$$E\left[\frac{P(C \in \lambda|T_R(Z))}{P(C \in \lambda)}\frac{I(C = R)}{P(C = R|T_R(Z))}m(Z;\beta)\right] = E[m(Z;\beta)|C \in \lambda]. \tag{2.2}$$

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 $r = 1, \ldots, R$

$$\text{MAR:} \qquad P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r|T_r(Z)). \tag{2.3}$$

This is the MAR assumption [see e.g. Robins and Rotnitzky (1995), Tsiatis (2006)] in the sense of Rubin (1976). Naturally (2.3) implies $P(C \geq r|Z) = 1 - P(C \leq r-1|Z) = 1 - P(C \leq r-1|T_{r-1}(Z)) = P(C \geq r|T_{r-1}(Z))$ only depends on $T_{r-1}(Z)$, and hence taking $r = R$ and $R = 2$, it does not contradict the standard representation of MAR, $P(C = 2|Z) = P(C = 2|Z_{(1)})$, in the econometrics literature 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 (2.3), 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_λ^0 defined in (2.1).

Naturally, all our theoretical results derived under the general condition (2.3) also hold under (2.4) and (2.5):

$$\text{CMAR:} \qquad P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r|T_1(Z)), \tag{2.4}$$

$$\text{INDEP:} \qquad P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r). \tag{2.5}$$

(2.3) 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 (2.3) by instead assuming that $P(C = r|Z) = P(C = r|Z_{(12)}, Z_{(22)}, \ldots, Z_{(R2)})$ serves a similar purpose. In the same vein one could also simply let e.g., $m(Z;\beta) = m(Z_{(R)};\beta)$, $m(Z_{(1)}, Z_{(R)};\beta)$, etc., which we do in the simulation study under (2.3)-(2.5) in Section 4.[7]

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.

[7]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.

(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 framework similar to ours 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.[8] 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 (2.1), i.e., it can characterize the sub-populations also, for $R = 2$ and $\lambda = \{1\}$ (sub-population) and $\{1, 2\}$ (full population). By contrast, we allow for a general R and expand the scope to consider all possible $(2^R - 1)$ sub-populations under a unified framework for a comprehensive treatment of the topic.

To consider all possible sub-populations under a general R in a unified and user-friendly manner, we assume:

$$P(C = r|T_r(Z) = t_r(z)) \text{ is known for all } t_r(z) \in \text{Support}(T_r(Z)) \text{ and for all } r = 1, \ldots, R. \tag{2.6}$$

While this assumption is restrictive for more general purposes, it is not unreasonable for our paper since the focus is on efficient estimation based on complete and incomplete samples drawn under a planned/known sampling design, and since surveys typically report the weights for each unit to indicate its representativeness.

It should still be noted that (2.3) and (2.6) jointly imply that $P(C = r|T_R(Z)) \equiv P(C = r|Z)$ is known for all $r = 1, \ldots, R-1$. This is a very powerful condition that, as evident from (2.2), solves any identification problem even allowing for selection on unobservables. (Recall that we do not impose the selection on observables condition (2.3) for identification purposes but for fully exploiting the multi-phase nature of the sampling design.) However, given our premise of incomplete-by-design sub-samples where a motivation behind the design is to guard against unplanned and, possibly, non-ignorable non-responses, attrition, etc. and thus rule out selection on unobservables, we will not consider the latter [see e.g. Tang et al. (2004)] although it would have opened up new possibilities.

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 $\underline{p} \in (0, 1)$.

[8] 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) in the Introduction. Thus our results are somewhat applicable to this literature, a proper treatment of which is however beyond the scope and focus of our paper.

(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)].[9] **2.** $P(C = R|T_R(Z)) > \underline{p} > 0$ in (A2) is a strict version of the aforementioned overlap assumption [see Khan and Tamer (2010) and Chaudhuri and Hill (2016)]. 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), etc.].

2.2 Combining the sub-samples for efficient estimation

To state our key result in Proposition 2.1 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:

$$\varphi_{r,\lambda}(O;\beta) := E\left[\frac{P(C \in \lambda|T_R(Z))}{P(C \in \lambda)} m(T_R(Z);\beta) \middle| T_r(Z) \right] \text{ for } r = 1, \ldots, R, \tag{2.7}$$

$$\begin{aligned} \varphi_\lambda(O;\beta) := & \frac{I(C = R)}{P(C = R|T_R(Z))} \varphi_{R,\lambda}(O;\beta) \\ & + \sum_{r=1}^{R-1} \left[\frac{I(C \geq R - r)}{P(C \geq R - r|T_{R-r}(Z))} - \frac{I(C \geq R - r + 1)}{P(C \geq R - r + 1|T_{R-r+1}(Z))} \right] \varphi_{R-r,\lambda}(O;\beta). \end{aligned} \tag{2.8}$$

Proposition 2.1 *Let* (2.1), (2.3), (2.6) *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* (2.8) *and* β^0_λ *is defined in* (2.1). *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*

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

Remarks:

1. This proposition extends Theorem 2 of Chen et al. (2008) to the case of a general R and a general target population ($C \in \lambda$) that can be arbitrary union of possibly different sub-populations from which the sub-samples are drawn. The result for a general R when the target is $\lambda = \mathcal{C}$ is not at all new and has been known since Robins and Rotnitzky (1995); Rotnitzky and Robins (1995), Robins et al. (1995). The novelty of the proposition lies precisely in allowing for more generality in the choice of the target population.

To make a slightly unrelated observation, consider the case when (2.6) does not hold, i.e., when $P(C = r|T_r(Z))$ is unknown. If $\lambda = \mathcal{C}$ then the proposition still holds without any change, giving the same efficiency bound and the same efficient influence function. If $\lambda = \{1\}$, the proposition continues to hold by merely changing the R-th term of $\varphi_\lambda(O;\beta)$ to $\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m(T_R(Z);\beta)|T_1(Z)] - \frac{I(C\geq 2)}{P(C\geq 2|T_2(Z))} \varphi_{1,\lambda}(O;\beta)$. See Tsiatis (2006) for a textbook

[9] 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 Example 1 in the Introduction, 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)].

treatment of the first scenario. Proofs for both scenarios are available from the author.

2. $\varphi_\lambda(O;\beta)$ in (2.8) 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:

$$\begin{aligned}
Cov(\text{term}_1, \text{term}_r) &= -Var(\text{term}_r) \text{ for } r = 2, \ldots, R \\
&= E\left[\left(\frac{1}{P(C \geq R-r+1|T_{R-r+1}(Z))}\right.\right. \\
&\qquad \left.\left. - \frac{1}{P(C \geq R-r+2|T_{R-r+2}(Z))}\right)\varphi_{R-r+1,\lambda}(O;\beta^0_\lambda)\varphi'_{R-r+1,\lambda}(O;\beta^0_\lambda)\right], \quad (2.9) \\
Cov(\text{term}_s, \text{term}_r) &= 0 \text{ for } s \neq r \neq 1, \\
\text{and hence } V_\lambda &= Var\left(\sum_{r=1}^{R} \text{term}_r\right) = Var(\text{term}_1) - \sum_{r=2}^{R} Var(\text{term}_r).
\end{aligned}$$

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 $r > 1$.

3. Although $Var(\varphi_{R-r+1,\lambda}(O;\beta)) \geq Var(\varphi_{R-r,\lambda}(O;\beta))$ (in a matrix sense), the order is not always preserved between $Var(\text{term}_r)$ and $Var(\text{term}_{r+1})$ for $r > 1$ because the latter variances are also affected by certain conditional probabilities in a nontrivial way (see (2.9)). This makes it difficult in general (see footnote 5) to construct an optimal design for obtaining the sub-samples even under the case of CMAR in (2.4) and INDEP in (2.5).

We conclude this subsection by looking into combining the sub-samples from alternative viewpoints but to the same effect. To this end let us first revert the order of the terms on the RHS of (2.8) and rewrite $\varphi_\lambda(O;\beta)$ as

$$\varphi_\lambda(O;\beta) = \varphi_{1,\lambda}(O;\beta) + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r(Z))}\left[\varphi_{r,\lambda}(O;\beta) - \varphi_{r-1,\lambda}(O;\beta)\right] \quad (2.10)$$

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 $(i = 1, \ldots, n : C_i \geq r)$ as is signified by the multiplier $I(C \geq r)$. (For the sake of the argument, define $T_0(Z)$ as a constant, and thus $\varphi_{0,\lambda}(O;\beta) := 0$.) 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 $r = 1, \ldots, R$. The terms on the RHS are uncorrelated. Thus 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 the above proposition. Accordingly, consider estimation of β_λ^0 based on:

$$E\left[\phi_{R,\lambda}(O;\beta)\right] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta_\lambda^0, \tag{2.11}$$

$$E\left[\phi_{R-r}(O)|T_{R-r}(Z)\right] = 0 \text{ almost surely } T_{R-r}(Z) \text{ for } r = 1,\ldots,R-1. \tag{2.12}$$

The functions $\phi_{R,\lambda}(O;\beta)$ and $\phi_{R-r}(O)$ are defined as follows:

$$\begin{aligned}\phi_{R,\lambda}(O;\beta) &:= \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta) \text{ [the IPW term from (2.8)]},\\ \phi_{R-r}(O) &:= I(C \geq R-r)\left[I(C \geq R-r+1) - P(C \geq R-r+1|C \geq R-r, T_{R-r}(Z))\right]\end{aligned}$$

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.)

The moment restriction in (2.11) already identifies β_λ^0, and GMM estimation based on it is the GMM-version of the Horvitz-Thompson method of obtaining IPW estimators. The moment restrictions in (2.12) do not involve β but bring additional information that captures the information content of the assumption (2.3) under the monotonic structure of the observed data, i.e., $T_{r-1}(Z) \subset T_r(Z)$. In particular, as evident from the multiplier $I(C \geq r)$ for the r-th moment $\phi_r(O)$ 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 (2.11) and (2.12) 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, (2.11) 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, i.e., $T_{r-1}(Z) \subset T_r(Z)$, of the conditioning set in (2.12). Proposition 2.2 links this efficient estimating function with the result in Proposition 2.1.

Proposition 2.2 *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. Under* (2.3) *and assumptions (A1) and (A2), $\varphi_\lambda(O;\beta)$ defined in* (2.8) *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 gives a moment function such that GMM based on it using the optimal weighting matrix results in a semiparametrically efficient estimator for β_λ^0 under the moment restrictions (2.11)-(2.12). (Also see Theorem 1 in Chamberlain (1992).) The repeated projection operation in its definition is the repeated application of equation (15) in Brown and Newey (1998) to efficiently combine the information from the unconditional moment restriction in (2.11) with all the information from the conditional moment restrictions in (2.12).[10] (This

[10] While not directly relevant and hence not shown here, due to the monotonic structure of the conditioning sets in (2.12), projections taken in any order $\phi_{\pi_1},\ldots,\phi_{\pi_R}$, where $(\pi_1,\ldots,\pi_R)$ is a permutation of $(1,\ldots,R)$, will generate the same quantity.

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 far more complex and thus requires an adequately rich choice for the sequence of functions $(\phi_{R-r}(O))_{r=1}^{R-1}$ in (2.12) to establish the equivalence result that allows an alternative viewpoint to appreciate the contribution of the sub-samples toward efficient estimation.[11]

3 Estimation and Inference

3.1 Estimation framework: Overview

Estimation of β_λ^0 can be done as a standard exercise in GMM by treating $\varphi_\lambda(O;\beta)$ in (2.8) 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 $\{O_i = (C_i', T_{C_i}'(Z_i))'\}_{i=1}^n$.[12] The other terms involve unknown conditional expectations $(\varphi_{r,\lambda}(O;\beta))_{r=1}^{R-1}$ and must be estimated prior to (or along with) the estimation of β_λ^0. Whether we treat the unknowns in each of these term to be of finite or infinite dimension determines whether the estimator of β_λ^0 is parametric or semiparametric, and we consider both under, by now, well-understood high level conditions.

Our estimation framework is a special case of Chen et al. (2003), to whom we appeal for discussing the asymptotic properties of the GMM estimator. However, two key features of the framework help us to weaken their general conditions in practically important ways. This will be our focus of discussion, and we will abstract from technical issues that, while important, are already discussed in Chen et al. (2003) and the references therein.

To facilitate the connection with Chen et al. (2003) and discuss these key features, we define a $d_m \times 1$ function:

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

where $h(O;\beta) = (h_1'(O;\beta),\dots h_{R-1}'(O;\beta))'$ are nuisance parameters, and $h_r(O;\beta) : \text{Support}(O) \times \mathcal{B} \mapsto \mathbb{R}^{d_m}$ belongs to a class of functions, call it $\mathcal{H}_r(\beta)$, for $r = 1,\dots,R-1$. If $(h_r(O;\beta))_{r=1}^{R-1} = (\varphi_{r,\lambda}(O;\beta))_{r=1}^{R-1}$ then

[11] In spirit, the idea behind augmenting the moment restriction (2.11), that already identifies β_λ^0 and can be used to obtain a $\sqrt{n}$-consistent estimator [see e.g. Wooldridge (2007)], by those in (2.12) 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)(HI), Devereux and Tripathi (2009), Tripathi (2011), Graham et al. (2012), etc. or HI, Nevo (2003), etc. in a different context. To see the connection note that under our setup this means estimating β_λ^0 by solving for β from $\sum_{i=1}^n q_i\varphi_{R,\lambda}(O_i,\beta) = 0$ where $q_i = I(C_i = R)/P(C = R|T_R(Z_i)) = q_{IPW,i}$, say, (instead of $1/n$ to reflect the non-representativeness of the complete sub-sample) if only (2.11) is used, whereas if the calibration/augmenting/auxiliary restrictions in (2.12) are also utilized, then $q_i = q_{IPW,i} + \sum_{r=1}^{R-1} a_{r,i}$ where e.g. $a_{r,i}$'s are recursively obtained using (3.3); and for $R = 2$, $a_{1,i} = q_{IPW,i}\Upsilon_{K_1}'(T_1(Z_i))(\sum_{j=1}^n \Upsilon_{K_1}(T_1(Z_j))\Upsilon_{K_1}'(T_1(Z_j)))^{-1}\sum_{l=1}^n(1-q_{IPW,l})\Upsilon_{K_1}(T_1(Z_l))$ 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 $\bar{q}_i = q_i/\sum_j q_j$ as the weights so that they necessarily add up to one. However, there is no guarantee that $\bar{q}_i \in [0,1]$ for all i (indeed it can be outside the interval for all i), which is not a desirable characteristic for weights. To avoid getting into the operational details of correcting for this undesirable characteristic by methods such as shrinkage or empirical likelihood that are peripheral to the main message of our paper, we will, henceforth, not be explicit in our discussion about such re-weighing of observations based on the extended set of moment restrictions.

[12] Since all the R terms of $\varphi_\lambda(O;\beta)$ involves the scalar multiple $1/P(C \in \lambda)$ which is a finite constant by assumption (A2), it can be safely ignored in the rest of the paper since GMM estimation is invariant to moment vectors up to scalar multiplication.

$g(O;\beta,h(O;\beta)) = \varphi_\lambda(O;\beta)$ defined in (2.8). We denote $\varphi_{r,\lambda}(O;\beta)$ by $h_r^0(O;\beta)$, i.e., the "truth" for $h_r(O;\beta)$, for $r = 1,\ldots,R-1$. The definition of $\varphi_{r,\lambda}(O;\beta)$ (which is an expectation conditional on $T_r(Z)$) suggests that it is natural to assume $h(O;\beta) \in \mathcal{H}(\beta)$ where $\mathcal{H}_1(\beta) \subset \mathcal{H}_2(\beta) \subset \ldots, \mathcal{H}_{R-1}(\beta) = \mathcal{H}(\beta)$. However, such nesting is not necessary for the theoretical results (it can be useful in practice). Instead, we proceed with a general definition that, crucially for our purpose of discussing the key features, incorporates the MAR assumption (2.3) in it:

$$\mathcal{H}(\beta) := \mathcal{H}_1(\beta) \times \ldots \times \mathcal{H}_{R-1}(\beta) \text{ where, for a given } \beta \in \mathcal{B}, \tag{3.1}$$
$$\mathcal{H}_r(\beta) := \left\{ h_r : \text{Support}(O) \times \mathcal{B} \mapsto \mathbb{R}^{d_m} \,\middle|\, E\left[\left(\frac{I(C \geq r)}{P(C \geq r|T_r(Z))} - \frac{I(C \geq r+1)}{P(C \geq r+1|T_{r+1}(Z))}\right) h_r\right] = 0 \right\}.$$

(3.1) allows the nuisance parameters (and, as we will see, their estimates along with the probability limits) to be quite general, e.g., any function of Z by virtue of (2.3). Naturally, (2.3) implies that $(h_r^0(O;\beta) \in \mathcal{H}_r(\beta))_{r=1}^{R-1}$.

Notation: We follow Chen et al. (2003)'s notation closely. Unless confusing, for all β we write $h(O;\beta)$ as $h(\beta)$ by omitting O, and $(\beta, h(\beta))$ as (β, h) by omitting β. For simplicity we let $\mathcal{H}_r(\beta) = \mathcal{H}_r(\bar{\beta})$ for all $\beta, \bar{\beta} \in \mathcal{B}$ and write it as $\mathcal{H}_r$ for all $r = 1,\ldots,R-1$, and write $\mathcal{H}$ accordingly, following the definition in (3.1). We maintain that $\mathcal{H}$ is a vector space endowed with a pseudo-metric $\|.\|_{\mathcal{H}}$, which is the sup-norm metric with respect to the argument β and a pseudo-metric with respect to other arguments. For two conformable matrices A and B of finite dimensions, we write the Euclidean norm as $\|A\|_B = \sqrt{\text{trace}(A'BA)}$ and write $\|A\|$ if B is an identity matrix. We write the average moment vector and its expectation as:

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

The GMM estimator: Let $\widehat{h}(\beta) \equiv \widehat{h}(O;\beta)$ be an estimator of $h(\beta) \equiv h(O;\beta)$ for a given β. 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:

$$\widehat{\beta}_\lambda(W_n) := \arg\min_{\beta\in\mathcal{B}} \|G_n(\beta,\widehat{h})\|_{W_n}. \tag{3.2}$$

Practical considerations often dictate that the relationship be approximate, i.e., $\widehat{\beta}_\lambda(W_n)$ be an approximate minimizer. As a standard practice, our results on the asymptotic properties of $\widehat{\beta}_\lambda(W_n)$ will accommodate for such approximations and precisely state the required strength of the approximation in terms of the sample size n.

First-step estimation of the nuisance parameters: Given our focus on the first-order asymptotic properties of $\widehat{\beta}_\lambda(W_n)$, the choice of the method for first-step estimation of $h_r(\beta)$ is not important [see e.g. Proposition 1 in Newey (1994)], and we abstract from it. However, it is worth noting here that conventional first-step estimators will naturally belong in $\mathcal{H}_r(\beta)$ with probability approaching one as sample size $n \to \infty$. For example, consider the following naive (since non-orthogonalized) series estimator of $h_r^0(O_i;\beta)$ for the i-th observation (where, as evident from the definition of $g(O;\beta,h(O;\beta))$, we only require this estimator) for $\{i = 1,\ldots,n : C_i \geq r\}$:

$$\widehat{h}_r(O_i;\beta) := \left(\frac{1}{n}\sum_{j=1}^{n} \omega_{r+1,j}\widehat{h}_{r+1}(O_j;\beta)\Upsilon'_{K_r}(T_r(Z_j))\right)\left(\frac{1}{n}\sum_{l=1}^{n}\omega_{r,l}\Upsilon_{K_r}(T_r(Z_l))\Upsilon'_{K_r}(T_r(Z_l))\right)^{-}\Upsilon_{K_r}(T_r(Z_i)) \tag{3.3}$$

for $r = 1, \dots R-1$ and with $\widehat{h}_R(O_i;\beta)$ defined as a convention simply as $\varphi_{R,\lambda}(O_i;\beta)$. (We use this estimator in the simulation study in Section 4.) $\omega_{r,j} = I(C_j \geq r)/P(C \geq r|T_r(Z_j))$ are the balancing weights for the j-th observation for $j = 1, \dots, n$ and $r = 1, \dots, R-1$ [see e.g. Hirano and Imbens (2001)]. Let A^- denote a symmetric generalized inverse of a matrix A. $\Upsilon_{K_r}(T_r(Z))$ is a $K_r \times 1$ series of functions of $T_r(Z)$ [see e.g. Newey (1997), Chen (2007)]. This estimator 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 $\Pi(Y|X) := E[YX'](E[XX'])^- X$ be the population least squares projection of a random variable Y on another variable X. Then it follows under standard regularity conditions that $\|\widehat{h}_r(\beta) - h_r^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ where

$$h_r^\dagger(\beta) := \Pi(\Pi(\dots\Pi(\varphi_{R,\lambda}(O;\beta)|T_{R-1}(Z))\dots|T_{R-r+1}(Z))|T_{R-r}(Z)). \tag{3.4}$$

(2.3) implies that $h_r^\dagger(\beta) \in \mathcal{H}$. Conditions similar to Assumptions 1 and 2 in Hahn (1997) ensure that $h_r^\dagger(\beta) = h_r^0(\beta)$, the truth. Crucially, if $h_r^\dagger(\beta) \in \text{interior}(\mathcal{H}_r)$ then $\widehat{h}_r(\beta) \in \mathcal{H}_r$ with probability approaching one.

3.2 Estimation framework: Key features and their implications

The definition of $\mathcal{H}$ in (3.1) gives $G(\beta, h(\widetilde{\beta})) = E[\varphi_{R,\lambda}(O;\beta)] = E[m(Z;\beta)|C \in \lambda]$ for any $\beta, \widetilde{\beta} \in \mathcal{B}$ and any $h(\widetilde{\beta}) \in \mathcal{H}$. The last equality follows by (2.2). Therefore, (2.1) now implies what we refer to as the first key feature of our estimation framework:

$$G(\beta_\lambda^0, h(\beta)) = 0 \text{ for any } \beta \in \mathcal{B} \text{ and } h(\beta) \in \mathcal{H}. \tag{3.5}$$

Hence, irrespective of the introduction of the nuisance parameters, the same β_λ^0 defined in (2.1) is identified by instead working with the moment vector $g(O;\beta,h(\beta))$, since for any $\beta, \widetilde{\beta}, \bar{\beta} \in \mathcal{B}$ and any $h(\widetilde{\beta}), \bar{h}(\bar{\beta}) \in \mathcal{H}$:

$$G(\beta, h(\widetilde{\beta})) - G(\beta_\lambda^0, \bar{h}(\bar{\beta})) = 0 \iff E[m(Z;\beta)|C \in \lambda] - E[m(Z;\beta_\lambda^0)|C \in \lambda] = 0 \iff \beta = \beta_\lambda^0.$$

This helps to verify the standard well separability (of the true β) assumption for consistent estimation of β_λ^0 by $\widehat{\beta}_\lambda(W_n)$. It is even stronger in the sense that it indicates that the nuisance parameter estimator $\widehat{h}(\beta)$ need not converge in probability to the true $h^0(\beta)$ but can converge to any $h^\dagger(\beta) \in \text{interior}(\mathcal{H})$ [see. e.g., (3.4)] without affecting the consistency of $\widehat{\beta}_\lambda(W_n)$ for β_λ^0. This provides flexibility in estimation of the nuisance parameters parametrically based on misspecified models, or nonparametrically under less than satisfactory conditions.

It also implies that the partial derivative of $G(\beta,h)$ with respect to β, denote it by $G_\beta(\beta,h)$, satisfies

$$G_\beta(\beta,h) = M_\lambda(\beta) := \frac{\partial}{\partial \beta'} E\left[m(Z;\beta)|C \in \lambda\right], \tag{3.6}$$

and it exists whenever $M_\lambda(\beta)$ exists. By virtue of assumption (A3), this simplifies the Jacobian formula (and its estimation) in the asymptotic variance of the GMM estimator $\widehat{\beta}_\lambda(W_n)$ since (3.6) implies $G_\beta(\beta_\lambda^0, h) = M_\lambda$.

Now we turn to the second key feature of our framework that concerns the variability of $G(\beta,h)$ with respect

to h. While this was already implicit in the discussion above, we will now make it explicit since it provides important intuitions about the influence of the estimation of $\widehat{h}$ on the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$.[13] This feature is: for any $\beta \in \mathcal{B}$, $\widetilde{\beta} \in \mathcal{B}$, $\bar{\beta} \in \mathcal{B}$, $h(\widetilde{\beta}) = (h_1'(\widetilde{\beta}), \ldots, h_{R-1}'(\widetilde{\beta}))' \in \mathcal{H}$ and $\bar{h}(\bar{\beta}) = (\bar{h}_1'(\bar{\beta}), \ldots, \bar{h}_{R-1}'(\bar{\beta}))' \in \mathcal{H}$:

$$G(\beta, h(\widetilde{\beta})) - G(\beta, \bar{h}(\bar{\beta})) = 0, \tag{3.7}$$

i.e., for all $\beta \in \mathcal{B}$, the function $G(\beta, h)$ is not only linear in h, but also it does not vary at all with $h \in \mathcal{H}$. Trivially (3.7) implies that for all $\beta \in \mathcal{B}$, 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

$$G_h(\beta, h)[\bar{h} - h] = 0. \tag{3.8}$$

(The notation used is similar to Chen et al. (2003).) Therefore, conditional on the event $\widehat{h}(\beta) \in \mathcal{H}$ which happens with probability approaching one if $\|\widehat{h} - h^\dagger\|_{\mathcal{H}} = o_p(1)$ and $h^\dagger(\beta) \in \text{interior}(\mathcal{H})$ for all β, the second feature implies that for all $\beta \in \mathcal{B}$:

$$\|G(\beta, \widehat{h}) - G(\beta, h^\dagger) - G_h(\beta, h)[\widehat{h} - h^\dagger]\| = 0.$$

Therefore, conditions (4.1.3) and (4.1.4) in Theorem 4.1 of Chen (2007) hold. Hence the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$ is unaffected by the estimation of the nuisance parameters even if $\widehat{h}$ converges at a rate slower than $\|\widehat{h} - h^\dagger\|_{\mathcal{H}} = o_p(n^{-1/4})$. As we will see in the next subsection, all we require from the convergence of $\widehat{h}$ for this purpose is $\|\widehat{h} - h^\dagger\|_{\mathcal{H}} = o_p(1)$ (where the sup-norm with respect to β in $\|.\|_{\mathcal{H}}$ is taken over $\{\beta \in \mathcal{B} : \|\beta - \beta_\lambda^0\| = o(1)\}$); i.e., the same rate that, when global in β, also ensures consistency of $\widehat{\beta}_\lambda(W_n)$. The scenario is stronger here than Theorem 4.1 of Chen (2007) since we do not even require that the probability limit $h^\dagger = h^0$, the truth. These nice implications of the second feature also provide flexibility in estimation of the nuisance parameters parametrically based on misspecified models, or nonparametrically under less than satisfactory conditions.

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 estimator $\widehat{h}$ is still of no consequence as far as the first-order asymptotic properties of GMM estimators are concerned. See Chaudhuri and Guilkey (2016) for a related but less striking result when (2.6) does not hold. Rothe and Firpo (2015) discuss the higher order properties in a closely related context where estimation of h has an effect but of smaller order than in the case of general semiparametric estimation in moment conditions models [see Section 3.4.2 of our paper for a different perspective]. Also see Robins and Ritov (1997) for a general discussion. Finally, we note that all these nice properties of our estimation framework are intrinsically related to the so-called "double-robustness" property of estimating functions discussed in Scharfstein et al. (1999), although the term double-robust is not exactly appropriate for our estimation framework.

[13] It was implicit in the sense that since (3.5) holds for all $h \in \mathcal{H}$, it naturally satisfies the (asymptotic) orthogonality condition, Assumption N(c), in Andrews (1994) [see equations (4.9) and (4.10) in Andrews (1994)]. Therefore, (i) estimation of the unknown nuisance parameters should not affect the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$, (ii) the rate at which $\widehat{h}(\beta)$ converges to its probability limit, e.g., $\|\widehat{h} - h^\dagger\|_{\mathcal{H}} = o_p(n^{-1/4})$), turns out to be redundant either for the purpose in (i) or for the verification of the standard stochastic equicontinuity assumption that is typically imposed for establishing asymptotic normality of $\widehat{\beta}_\lambda(W_n)$.

3.3 Asymptotic properties of the GMM estimator

Now we demonstrate precisely all the intuitions provided in the above discussion of the key features.

Proposition 3.1 *Let* (2.1), (2.3), (3.1) *and assumptions (A1) and (A2) hold. Let* $\{W_n\}$ *be* $d_m \times d_m$ *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) $\|\widehat{h} - h^\dagger\|_{\mathcal{H}} = o_p(1)$ *for some* $h^\dagger \in$ *interior*$(\mathcal{H})$ *not necessarily equal to* h^0 *(see e.g.* (3.4)*);*

(B3) for all $\delta > 0$ *there exists* $\epsilon(\delta) > 0$ *such that* $\inf_{\|\beta-\beta^0_\lambda\|>\delta} \|G(\beta, h^\dagger)\| \geq \epsilon(\delta) > 0$*;*

(B4) for all sequences of positive numbers $\{\delta_n\}$ *with* $\delta_n = o(1)$*,*

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

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

Proposition 3.2 *Let* (2.1), (2.3), (3.1) *and assumptions A hold. Let* $\{W_n\}$ *be* $d_m \times d_m$ *positive semidefinite matrix such that* $W_n = W + o_p(1)$ *where* W *is a constant positive definite matrix. Let* $\beta^0_\lambda \in$ *interior*$(\mathcal{B})$ *and* $h^\dagger(\beta) \in$ *interior*$(\mathcal{H})$ *where* $h^\dagger(\beta)$ *is not necessarily* $h^0(\beta)$ *(see e.g.* (3.4)*). For a small* $\delta > 0$ *define the neighborhoods* $\mathcal{B}_\delta := \{\beta \in \mathcal{B} : \|\beta - \beta^0_\lambda\| \leq \delta\}$ *and* $\mathcal{H}_\delta := \{h(\beta) \in \mathcal{H} : \|h(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} \leq \delta\}$*. Let* $\widehat{\beta}^0_\lambda(W_n) - \beta^0_\lambda = o_p(1)$ *and* $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ *where the sup-norm with respect to* β *in* $\|.\|_{\mathcal{H}}$ *is taken for* $\beta \in \mathcal{B}_\delta$*. 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* $\beta = \beta^0_\lambda$ *(*$G_\beta(\beta^0_\lambda, h^\dagger)$ *is full column rank by (A3) and* (3.6)*);*

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

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

(C4) for some finite matrix Σ*,* $\sqrt{n}G_n(\beta^0_\lambda, h^\dagger) \xrightarrow{d} N(0, \Sigma)$*.*

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

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

Remark: Propositions 3.1 and 3.2 respectively establish consistency and asymptotic normality of the GMM estimator defined in (3.2). The assumptions in Proposition 3.1 are essentially identical to those in Theorem 1 of Chen et al. (2003) (their uniform continuity assumption (1.3) is directly satisfied by our (3.6)) except that we allow the probability limit of $\widehat{h}$ to be $h^\dagger \neq h^0$, as discussed in the last subsection. This is a remarkable property of the GMM estimator under our framework. The assumptions in Proposition 3.2 are similar to those in Theorem 2 of Chen et al. (2003). The similarity is actually more with Theorem 4.1 in Chen (2007) that partly generalizes

Chen et al. (2003) and also provides an alternative high level assumption to assumptions (2.3)(i) and (2.4) of the latter (assumptions (4.1.3) and (4.1.4) in Chen (2007)'s Theorem 4.1 are automatically satisfied due to our (3.7) and (3.8)). However, unlike them, we can again allow the probability limit of $\widehat{h}$ to be $h^\dagger \neq h^0$. We repeat that the asymptotic unbiasedness and normality of the GMM estimator are not affected by the rate of convergence of $\widehat{h}$ to its probability limit. Thus the theoretical results confirm the intuitions from the last subsection, with the final bit of intuition to be confirmed by the following result on efficient GMM estimation.[14]

Corollary 3.3 *Under the assumptions of Proposition 3.2:*

(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 2.1, and*

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

i.e., by Proposition 2.1, the estimator $\widehat{\beta}_\lambda(W_n)$ *becomes semiparametrically efficient.*

3.4 Issues related to practical implementation

3.4.1 Estimation of the asymptotic variance in Corollary (3.3)

Consistent estimation of M_λ is simplified due to (3.6) 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 $G_n(\beta, h)$. The numerical derivative at any β can be obtained as in, e.g., page 2190 of Newey and McFadden (1994) such that the j-th ($j = 1, \ldots, d_\beta$) column of $M_\lambda(\beta)$ is estimated as follows:

$$\frac{1}{2n\varepsilon_{nj}} \sum_{i=1}^{n} \frac{I(C_i = R)}{P(C = R|T_R(Z_i))} \frac{P(C \in \lambda|T_R(Z_i))}{\widehat{P}(C \in \lambda)} \left[m(Z_i; \beta + \varepsilon_{nj}e_j) - m(Z_i, \beta - \varepsilon_{nj}e_j)\right]$$

where e_j is a $d_\beta \times 1$ vector with one at the j-th row and zeros elsewhere, and $\varepsilon_{nj} > 0$. If $\varepsilon_{nj} \to 0$ and $\varepsilon_{nj}\sqrt{n} \to \infty$ for $j = 1, \ldots, d_\beta$, then consistency of $\widehat{M}_\lambda(\beta)$ for $M_\lambda(\beta)$ follows by Theorem 7.4 in Newey and McFadden (1994) who subsequently discuss the practical issues related to the estimation [also see Section 5.3 of Cattaneo (2010)].[15]

Standard conditions, that are high level in our context, e.g., $g(O_i, \beta, h)$ is continuous with probability approaching one in a neighborhood $\mathcal{N}$ of (β_λ^0, h^0) and $E\left[\sup_{(\beta,h)\in\mathcal{N}} \|g(O_i, \beta, h)\|^2\right] < \infty$ [see Lemma 4.3 in Newey

[14] The results were presented under high level conditions of which, the primitive conditions for stochastic equicontinuity conditions (B4) and, in particular, (C3) have received the most attention in the (moderately) recent literature [see Section 4 of Chen et al. (2003) and the references therein]. Cattaneo (2010) takes a more direct approach than us with his assumptions 4, 6, and 7 in a similar context and work with the original moment vector ($m(Z; \beta)$ for our case). [Also see Chen et al. (2008).] However, since we do not have anything new to say about these issues, we abstract from them and instead focus on the novel features of our framework.

[15] 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.

and McFadden (1994)], ensure that for any $\beta = \beta^0_\lambda + o_p(1)$ and h such that $\|h - h^0\|_{\mathcal{H}} = o_p(1)$, the estimator

$$\widehat{V}_\lambda(\beta, h) := \frac{1}{n}\sum_{i=1}^{n} g(O_i, \beta, h) g(O_i, \beta, h)' = V_\lambda + o_p(1).$$

Therefore, the estimator $\left(\widehat{M}'_\lambda(\widehat{\beta}_\lambda)\widehat{V}^{-1}_\lambda(\widehat{\beta}_\lambda, \widehat{h})\widehat{M}_\lambda(\widehat{\beta}_\lambda)\right)^{-1}$ is consistent for the asymptotic variance in Corollary 3.3(1) if $\|\widehat{h} - h^\dagger\|_{\bar{H}} = o_p(1)$ [see e.g. (3.3)], and for that in Corollary 3.3(2) if $\|\widehat{h} - h^0\|_{\bar{H}} = o_p(1)$ where $\|.\|_{\bar{H}}$ is same as $\|.\|_{\mathcal{H}}$ but only in appropriately local neighborhoods of β and $h^\dagger$ or h^0 (as applicable).

It should also be noted that given a preliminary consistent but inefficient estimator $\widetilde{\beta}$ and a consistent estimator $\widetilde{h}$, the efficient weighting matrix for the GMM estimator in (3.2) can be obtained as $W_n^{\text{eff}} = \widehat{V}^{-1}_\lambda(\widetilde{\beta}, \widetilde{h})$.

3.4.2 Bootstrap estimator of asymptotic variance

Bootstrapping as follows avoids the computation of the numerical derivatives for estimation of the asymptotic variance. Independently for each $b = 1, \ldots, B$, draw $\{O^*_{ib}\}_{i=1}^n$ with replacement from the observed sample $\{O_i\}_{i=1}^n$. For each $\beta \in \mathcal{B}$ obtain $(\widehat{h}^*_{rb}(\beta))_{r=1}^{R-1}$ [see (3.3)] with the bootstrap sample. Obtain the bootstrap GMM estimator:

$$\widehat{\beta}^*_{\lambda,b} = \arg\min_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} g(O^*_{ib}; \beta, \widehat{h}^*_b(\beta)) - G_n(\widehat{\beta}, \widehat{h}) \right\|_{W_n^{\text{eff}}} \text{ for } b = 1, \ldots, B$$

and define $\bar{\beta}^*_B = \frac{1}{B}\sum_{b=1}^B \widehat{\beta}^*_{\lambda,b}$. (Non-bootstrap version W_n^{eff} is used for ease of computation.) The bootstrap estimator for Ω_λ is $\widehat{\Omega}^*_{\lambda,B} = \frac{n}{B}\sum_{b=1}^B \left(\widehat{\beta}^*_{\lambda,b} - \bar{\beta}^*_B\right)\left(\widehat{\beta}^*_{\lambda,b} - \bar{\beta}^*_B\right)'$. As $B \to \infty$ and n is fixed, $\widehat{\Omega}^*_{\lambda,B}/n$ is consistent almost surely for the (generally infeasible) ideal bootstrap variance $\text{Var}^*(\widehat{\beta}^*_\lambda)$ [see page 27, Andrews and Buchinsky (2000)] where Var^* is the variance with respect to the distribution P^* of the ideal bootstrap sample conditional on the original observed sample $\{O_i\}_{i=1}^n$, and $\widehat{\beta}^*_\lambda$ is the ideal bootstrap estimator. Theorem B in Chen et al. (2003) provides sufficient conditions (that could be relaxed as before for our purpose) to establish the convergence in distribution of $\sqrt{n}(\widehat{\beta}^*_\lambda - \widehat{\beta}_\lambda(W_n^{\text{eff}}))$ to $N(0, \Omega)$ in P^*-probability. Therefore, $n \times \text{Var}^*(\widehat{\beta}^*_\lambda)$ converges in probability to Ω_λ if $\lim_{c\to\infty} \sup_n E^*\left[I\left(\left\|\sqrt{n}(\widehat{\beta}^*_\lambda - \widehat{\beta}_\lambda(W_n^{\text{eff}}))\right\| > c\right)\left\|\sqrt{n}(\widehat{\beta}^*_\lambda - \widehat{\beta}_\lambda(W_n^{\text{eff}}))\right\|^2\right] = 0$ [see e.g. Theorem 5.9(i), Gut (2012)]. The uniform integrability, as assumed here directly for the estimator, is however a very high level condition and in general can be difficult to establish (and, thus, useless/ignored) in practice.

3.4.3 Bias from estimation of many nuisance parameters

The profiled moment vector $G_n(\beta, h(\beta))$ may suggest that efficient estimation of β^0_λ via the introduction and subsequent estimation of the many nuisance parameters $(h_r(\beta))_{r=1}^{R-1}$, that depend on β, leads to asymptotic bias. However, this is not the case. To see it, first note that each nuisance parameter, along with its scalar multiple, is essentially the contribution of each incomplete sub-sample [also see Remark 2 following Proposition 2.1]. Second, recall from the discussion (and proof) of Proposition 2.2 that the incremental information contributed by each incomplete sub-sample could alternatively be captured by the incremental information contributed by each additional moment restriction from (2.12). One could now obtain a GMM estimator of β^0_λ based on the

moment restrictions in (2.11) and (2.12) by considering a (naive) moment vector

$$\psi(O;\beta) = \Big[\phi'_{R,\lambda}(O;\beta),\ \phi_{R-1}(O)\times\Upsilon'_{K_{R-1}}(T_{R-1}(Z)),\quad \ldots,\quad \phi_2(O)\times\Upsilon'_{K_2}(T_2(Z)),\ \phi_1(O)\times\Upsilon'_{K_1}(T_1(Z))\Big]'$$

where, for $r = 1,\ldots,R-1$, $\Upsilon_{K_r}(T_r(Z))$ is a $K_r \times 1$ series of functions of $T_r(Z)$ [also see the discussion below (3.3)]. Following Hahn (1997), it is possible to provide sufficient conditions including the rate at which $K_r \to \infty$ as $n \to \infty$ so that the efficient GMM estimator based on $\psi(O;\beta)$ is asymptotically equivalent to the GMM estimator in Corollary 3.3(2). The crucial observation is: although with $K_r \to \infty$, the total number of moment restrictions $d_m + \sum_{r=1}^{R-1} K_r$ in $E[\psi(O;\beta^0_\lambda)] = 0$ also goes to infinity, but only the first d_m of them involve β. Hence, assuming for simplicity that $m(Z;\beta)$ is twice continuously differentiable in β, $(\partial/\partial\beta')\psi(O;\beta)$ has only the first d_m rows non-zero for any given $\sum_{r=1}^{R-1} K_r$, and accordingly for the higher order derivatives. Therefore, in the asymptotic expansion of $\sum_{i=1}^n \psi(O_i;\beta)/\sqrt{n}$ around $\sum_{i=1}^n \psi(O_i;\beta^0_\lambda)/\sqrt{n}$, the successive terms, all of which involve derivatives with respect to β, will not contain any contribution from any of the (d_m+j)-th (for all $j \geq 1$) rows of the moment vector $\psi(O;\beta)$ [also compare with (3.6)]. Thus, the moment restrictions from (2.12) and hence the nuisance parameters $(h_r(\beta))_{r=1}^{R-1}$ do not contribute to the asymptotic bias of the GMM estimator.

3.4.4 One step from 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 (see example 1 below) then this is not an issue. However, if there is no closed form expression, one could start with an easy to compute $\sqrt{n}$-consistent estimator for β^0_λ and then update it in one step to obtain an estimator with the same asymptotic distribution as the efficient estimator in Corollary 3.3. Typically, the following Horvitz-Thompson (IPW) estimator based on the complete sub-sample and with identity (or some simple) weighting matrix is relatively easy to compute:

$$\widetilde{\beta} := \arg\min_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^n \frac{I(C_i=R)}{P(C=R|T_R(Z_i))}\varphi_{R,\lambda}(O_i;\beta)\right\| \equiv \arg\min_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^n \frac{I(C_i=R)}{P(C=R|Z_i)}\frac{P(C\in\lambda|Z_i)}{\widehat{P}(C\in\lambda)}m(Z_i;\beta)\right\|. \quad (3.9)$$

Ignoring $\widehat{P}(C\in\lambda)$ has no consequence. 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 $m(Z;\beta)$ (see example 2 below).

Then a one-step estimator of β^0_λ can be obtained by updating $\widetilde{\beta}$ as follows:

$$\widehat{\beta}_{1\text{step}} = \widetilde{\beta} - \widehat{\Omega}^{-1}_\lambda(\widetilde{\beta},\widehat{h}(\widetilde{\beta}))\widehat{M}'_\lambda(\widetilde{\beta})\widehat{V}^{-1}_\lambda(\widetilde{\beta},\widehat{h}(\widetilde{\beta}))G_n(\widetilde{\beta},\widehat{h}(\widetilde{\beta})) \quad (3.10)$$

where $\widehat{M}_\lambda(\widetilde{\beta})$, $\widehat{V}_\lambda(\widetilde{\beta},\widehat{h}(\widetilde{\beta}))$ and $\widehat{\Omega}_\lambda(\widetilde{\beta},\widehat{h}(\widetilde{\beta}))$ are as defined in Section 3.4.1, and $\widehat{h}$ as in (3.3). Under the assumptions of Corollary 3.3 and Section 3.4.1, and with W_n an estimator of the efficient weighting matrix in Corollary 3.3(2), it follows that the one-step estimator is efficient, i.e.,

$$\sqrt{n}\left(\widehat{\beta}_{1\text{step}} - \widehat{\beta}_\lambda(W_n)\right) = o_p(1).$$

3.4.5 Two Examples

We consider two examples of the moment vector $m(Z;\beta)$ that respectively correspond to a simple linear regression giving a closed form expression for the estimator, and a linear quantile regression where the estimator is computed in one step following the above method. To avoid complicating the expressions by matters of secondary importance, we take $d_m = d_\beta$ and, continuing with the running example from the Introduction, $R = 3$, $Z_{(1)} = (y, X_c')'$, $Z_{(2)} = X_e$ and $Z_{(3)} = X$ where y is the dependent variable, X is the regressor, while X_c and X_e are two mismeasured values of X that are respectively cheap and expensive to observe and can depend on y also.

Example 1: Simple linear regression

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 $q_i = P(C \in \lambda|T_{3i})$. Then for any weighting matrix W_n, the GMM estimator $\widehat{\beta}_\lambda(W_n)$ in (3.2) is the same since it is a just identified model, so denote it by $\widehat{\beta}_\lambda$, and simple computation gives its closed form expression:

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

where $\widehat{E}$ denotes the estimated conditional expectation (see e.g. (3.3)). 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 of the estimator $\widehat{\beta}_\lambda$. A similar phenomenon was observed by Wang and Chen (2009) (page 493) and Chaudhuri and Min (2012) (page 45).

Example 2: Simple linear quantile regression

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 defined here, each notation used below is the same as in Example 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. (3.3)). (The ignored common denominator $P(C \in \lambda)$ will be adjusted for in the final step.) Let $\widetilde{\beta}$ denote the inefficient but $\sqrt{n}$-consistent estimator of β_λ^0 obtained from (3.9) by using this particular choice of the moment vector $m(Z;\beta)$. It is simple to obtain $\widetilde{\beta}$ since all commonly used statistical softwares provide in-built routine for weighted quantile regression which automatically gives the estimator with $(a_{3i})_{i=1}^n$ as weights. Estimate M_λ where $M_\lambda(\beta) = -(\partial/\partial\beta')E\left[XI(y - X'\beta < 0)|C \in \lambda\right]$ as discussed in Section 3.4.1. or possibly by using a post estimation command of the same in-built routine.[16] Therefore, since $d_m = d_\beta$, by using (3.10) we obtain the one-step estimator as:

$$\widehat{\beta}_{1\text{step}} = \widetilde{\beta} - \widehat{M}_\lambda^{-1}(\widetilde{\beta}) G_n(\widetilde{\beta}, \widehat{h}(\widetilde{\beta}))/\widehat{P}(C \in \lambda).$$

[16] Under standard assumptions and when $\lambda = \mathcal{C}$, we obtain the familiar expression from linear quantile regressions that $M_\lambda = E[XX' f_e(0|X)]$ where $e = y - X'\beta_\lambda^0$ and $f_e(0|X)$ its density at 0 conditional on X. Instead, we work with the general case.

4 Simulation Study

In this section we numerically study the benefit, if any, of using all the sub-samples for efficient estimation of β_λ. Alongside, we also study the performance of the efficient GMM estimator in finite samples.

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:

$$\text{Loss}(\beta_{\lambda,j}; s, s') = \lim_{n\to\infty} \frac{\text{Avar}(\widehat{\beta}^s_{\lambda,j})/n_s - \text{Avar}(\widehat{\beta}^{s'}_{\lambda,j})/n_{s'}}{\text{Avar}(\widehat{\beta}^{s'}_{\lambda,j})/n_{s'}} \text{ where } \lambda, s, s' = \Lambda \text{ and } j = 1, \ldots, d_\beta. \tag{4.1}$$

We always include $\{R\}$ in s, s'. In general, s, s' also include λ unless $\lambda = \mathcal{C}$. n_s and $n_{s'}$ are the total size of the combined sub-samples in s and s' respectively. For $j = 1, \ldots, d_\beta$ and $l = s, s'$, $\widehat{\beta}^l_{\lambda,j}$ is the j-th element of $\widehat{\beta}^l_\lambda$, 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.[17] Thus, the estimators that are not based on 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 thus reflects the information brought in by the additional sub-samples that are included in s' but not in s.

Results 1 and 2 in Technical Appendix B provide simplified analytical expressions for the loss in (4.1) under INDEP and CMAR. This may serve roughly as a point of reference for the estimated loss reported in Section 4.2.

4.1 Simulation Design

The framework of the running example from the Introduction in the context of Example 1 in Section 3.4.5 is employed to study this loss and also the finite-sample behavior of the estimators via simulations. Accordingly, we take $Z_{(1)} = (y, X_c)'$, $Z_{(2)} = X_e$ and $Z_{(3)} = X$. We generate 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 X_c and X_e have the same mean 0 as X, X_e is a less variable measure of X and hence it is possibly reasonable to maintain that X_e is more expensive to observe than X_c but less expensive than X. (It should, however, be noted that X_e does not bring in much new relevant information after controlling for y and X_c.) We take $\alpha = \delta = 1$.

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

$$\begin{aligned}
P(C = 1|Z_i) &= F_{t_1}(\gamma_c(X_{ci} + y_i - 1)) \\
P(C = 2|Z_i) &= (1 - P(C = 1|Z_i))(1 - F_{t_1}(\gamma_e X_{ei} + \gamma_c(X_{ci} + y_i - 2))) \\
P(C = 3|Z_i) &= 1 - P(C = 1|Z_i) - P(C = 2|Z_i)
\end{aligned}$$

[17] 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 2.1 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.

where $F_{t_1}(a)$ is the cumulative distribution function of a t_1-distributed random variable evaluated at $a \in \mathbb{R}$. We take $\gamma_c = \gamma_e = .25$ for MAR, $\gamma_c = .25, \gamma_e = 0$ for CMAR (i.e., (2.4)), and $\gamma_c = \gamma_e = 0$ for INDEP (i.e., (2.5)).

The parameters of interest, i.e., the Intercept (β_1) and Slope (β_2) are defined by (2.1) with the moment vector:

$$m(Z;\beta) = [y - \beta_1 - \beta_2 X, \quad X(y - \beta_1 - \beta_2 X)]'.$$

Their true values 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) mentioned in the Introduction). All results reported below are based on 10,000 Monte Carlo trials.

Apart from the GMM estimators based on various sub-samples, we also consider the 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 (3.9). While it is also based only on the complete sub-sample, it does not treat this sub-sample as representative of the population of interest, and hence the probability weighting. All other estimators utilize the incomplete sub-samples and thus involve unknown nuisance parameters $h_r(.)$. The $h_r(.)$'s are estimated using the series estimator in (3.3), and 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).

4.2 Simulation Results

Tables 2 and 3 list the estimated loss (in percent) defined in (4.1) for various s with respect to $s' = \{1,2,3\}$ 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. For the second row, however, this is not true because the additional 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).

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 Result 2(b) in Technical Appendix B 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. Indeed this loss computed by using the Monte Carlo variance of the estimators instead of their asymptotic variances, is also very small (even smaller) in this particular case.[18]

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 (4.1); and hence this shows the obvious benefit of using all the sub-samples for estimation of the parameters of interest. Perhaps more interestingly this also makes a case for collecting incomplete data for more sample units if the budget constraint does not allow the complete data collection for all. It is however important to have some idea on what kind of incompleteness is less harmful. Results 1 and 2 and the discussion following them in Technical Appendix B provide intuitions on this issue.

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). As a consequence 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. We found evidence (not reported here) of the Std based on the Avar formula for these estimators being smaller than the Monte Carlo standard deviation in many cases when $n = 1000$. One could use bootstrap, but a definitive study of why bootstrap will be more effective than the asymptotic variance formula in this context is beyond the scope of the paper.

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.

[18] While we do not study it here, it should be noted that if we do not impose the planned-missingness assumption (2.6) then under CMAR one can analytically show that $\text{Loss}(\beta_{\{3\}}; s = \{3\}, s' = \{1, 3\}) = \text{Loss}(\beta_{\{3\}}; s = \{3\}, s' = \{2, 3\}) = \text{Loss}(\beta_{\{2,3\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) = 0$ even when $E[m|Z_1] \neq 0$. On the other hand, $\text{Loss}(\beta_{\{3\}}; s = \{3\}, s' = \{1, 2, 3\}) > 0$. The result is interesting because the loss only gets reflected when we take $R > 2$ moving beyond the commonly studied scenario of $R = 2$ in which case there is no loss when the nesting sub-sample (s') does not contain any unit that is strictly more informative than all the units in the nested sub-sample (s). All these results follow by a similar application of Proposition B.3 presented in Technical Appendix B.

5 Conclusion

We considered estimation of a finite dimensional parameter whose definition depends on the joint distribution of variables that are not observed for various units in the sample following a monotone pattern of non-observability. Such patterns often arise from multi-phase sampling, and indeed by sampling design, conducted as a cost effective method of surveying a large number of units. We referred to the collection of sample units with the same level of observability of variables as a sub-sample, and allowed an arbitrary finite number of sub-samples (equivalently, phases in the multi-phase survey), but required one sub-sample to be complete. The mechanism of non-observability of variables was assumed known, as is likely under planned incompleteness in surveys, but was allowed to take a completely general form of selection on observables, i.e., MAR, to fully exploit the multi-phase nature of sampling that could be useful in practice. As a consequence of this generality, the sub-samples may differ systematically and thus may not be representative of the same population, but rather of different sub-populations. The parameter of interest was generically defined by moment restrictions with respect to the joint distribution of the variables in an arbitrary union of these sub-populations. Thus the setup allowed for a wide variety of parameters of interest. The moment vector was allowed to be non differentiable in the parameter.

We obtained the semiparametric efficiency bound for all such parameters of interest under a unified framework and provided a thorough discussion of how the information contained in various sub-samples contribute toward this efficiency bound. An efficient GMM estimator was proposed. The estimation framework possesses several key features that allowed us to establish consistency, asymptotic normality and efficiency of the GMM estimator under weaker than standard conditions. These weak conditions have practical relevance and make the estimation of the nuisance parameters involved in the GMM estimation less burdensome than usual. In particular, it turned out that the convergence in probability of the estimator for the nuisance parameters to functions (satisfying certain weak restrictions) that are not necessarily the truth does not affect the consistency, asymptotic normality and asymptotic unbiasedness of the GMM estimator. This offers flexibility in parametric estimation of the nuisance parameters. If, on the other hand, the estimator of the nuisance parameters converges to the truth then, irrespective of the rate of this convergence, the GMM estimator is semiparametrically efficient. This offers flexibility in nonparametric estimation of the nuisance parameters. Several issues related to practical implementation of the GMM estimator were discussed. A simulation study was presented to numerically confirm that the asymptotic theoretical results of our paper remain appropriate in reasonably finite samples.

Several important issues were, however, not covered in the paper. In particular, the topic of optimal sampling design under the generality of our framework deserves close attention. Slightly unrelated to the main focus of the current paper but of independent interest is the topic of establishing the efficiency bound under the generality of our framework but without imposing (2.6). These are left for immediate future research.

In summary, the paper provided a detailed treatment of estimation based on planned incomplete data under a very general framework with special attention toward issues related to practical implementation. It is our hope that the framework and the estimator put forward in this paper will be useful in applied research.

References

Abrevaya, J. and Donald, S. G. (2011). A GMM approach for dealing with missing data on regressors and instruments. Mimeo.

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.

Andrews, D. W. K. and Buchinsky, M. (2000). A Three-Step Method For Choosing The Number Of Bootstrap Repetitions. *Econometrica*, 68: 23–51.

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.

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.

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.

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.

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.

Chaudhuri, S. and Hill, J. B. (2016). Heavy Tail Robust Estimation and Inference for Average Treatment Effect. Technical report, University of North Carolina, Chapel Hill.

Chaudhuri, S. and Min, H. (2012). Doubly-Robust Parametric Estimation in Moment Conditions Models with Missing Data. Mimeo.

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.

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. (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.

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.

Gut, A. (2012). *Probability: A Graduate Course*. Springer.

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

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.

Hirano, K. and Imbens, G. (2001). Estimation of Causal Effects using Propensity Score Weighting : An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, 2: 259–278.

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.

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

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.

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.

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.

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.

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

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

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. (2014). Efficient GMM Estimation with a General Missing Data Pattern. Technical report, Simon Frasier University.

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

Newey, W. (1994). The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62: 1349–1382.

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.

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.

Pakes, A. and Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, 57: 1027–1057.

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

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

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

Robins, 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. (2015). 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.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72: 538–543.

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.

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.

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.

Tang, G., Little, R. J. A., and Raghunathan, T. E. (2004). Analysis of Multivariate Monotone Missing Data. In Lin, D. Y. and Heagerty, P., editors, *Proceedings of the Second Seattle Symposium on Biostatistics*. New York: Springer.

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.

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.

Wang, D. and Chen, S. X. (2009). Empirical Likelihood for Estimating Equation with Missing Values. *Annals of Statistics*, 37: 490–517.

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.

Tables with Simulation Results from Section 4

Target Popln. λ	Used Sample s	INDEP Sampling					
		Intercept			Slope		
		$n = 1000$	$n = 2000$	$n = 5000$	$n = 1000$	$n = 2000$	$n = 5000$
$\{1,2,3\}$	$\{3\}$	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$(\beta_\lambda; S = s, S = \{1,2,3\})$ (in percent) defined in (4.1) for $j = 1$ (Intercept) and $j = 2$ (Slope). The results are based on the analytically estimated Avar averaged over 10000 Monte Carlo trials.

Target Popln. λ	Used Sample s	CMAR Sampling						MAR Sampling					
		Intercept			Slope			Intercept			Slope		
		n			n			n			n		
		1000	2000	5000	1000	2000	5000	1000	2000	5000	1000	2000	5000
$\{1,2,3\}$	$\{3\}$:IPW	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
$\{1\}$	$\{3\}$:IPW	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
$\{2\}$	$\{3\}$:IPW	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
$\{1,3\}$	$\{3\}$:IPW	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
$\{2,3\}$	$\{3\}$:IPW	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$(\beta_{\lambda,j}; s, s' = \{1,2,3\})$ (in percent) defined in (4.1) for $j = 1$ (Intercept) and $j = 2$ (Slope). The results are based on the analytically estimated Avar averaged over 10,000 Monte Carlo trials.

Used Sample	$n = 1000$				$n = 2000$				$n = 5000$			
	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.

A Technical Appendix A: Proofs of reported results

Proof of Proposition 2.1:

The proof exactly follows the three steps in Chen et al. (2008). The first step characterizes the tangent set for all regular parametric submodels satisfying the semiparametric assumptions on the observed data. The second step obtains the efficient influence function for a given rotation of $m(Z;\beta)$. The last step 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 finite (e.g. one; see Stein (1956), page 188) dimensional parameter θ for the joint distribution of the observed data $O = (C', T_C'(Z))'$. So the log of joint density of the observed data can be expressed in terms of the full data $(C, Z')'$ as

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

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 submodels and can be characterized by functions of the form

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

where (each element of)the functions $a_1(Z_{(1)}) \in L_0^2(F(Z_{(1)}))$ 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 (2.1) 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 (2.2). 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 (2.1) and (2.3) that

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

Taking a full row rank A along with assumption A3 gives

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

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

$$E[\psi(A, O) S(O)'] = \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'}. \tag{A.2}$$

Let us conjecture that $\psi(A, O) = -(AM_\lambda)^{-1} A \varphi_\lambda(O;\beta_\lambda^0)$, and then verify (B.2) 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 (2.10) for

convenience), write it as $\sum_{q=1}^{R} B_q$ where

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

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 $r=2,\ldots,R$; and $s(Z_{(1)})$, interchangeably, as $s(Z_{(1)}|T_0)$. We hope that this is not unnecessarily confusing. Now note from the definitions of B_1 and B_q that

$$\begin{aligned} B_1 &= \sum_{r=1}^{R} E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]I(C\geq r)s(Z_{(r)}|T_{r-1})'\right] \\ &= \sum_{r=1}^{R} E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right](1-I(C\leq r-1))s(Z_{(r)}|T_{r-1})'\right] \\ &= E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right](1-I(C\leq 0))s(Z_{(1)})'\right] \text{ (since for } r>1\ s(Z_{(r)}|T_{r-1})\in L_0^2(F(Z_{(r)}|T_{r-1}))) \\ &= E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]s(Z_{(1)})'\right] \text{ (now recalling } T_1:=Z_{(1)}) \\ &= E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}ms(Z_{(1)})'\right]=E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}ms(Z_{(1)})'\right]=E[ms(Z_{(1)})'|C\in\lambda]. \end{aligned}$$

On the other hand, and now we skip steps that are similar to the above, (2.3) gives for $q=2,\ldots,R$:

$$\begin{aligned} B_q &= \sum_{r=1}^{q-1} E\left[\frac{I(C\geq q)}{P(C\geq q|T_q)}(\varphi_{q,\lambda}-\varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right]+\sum_{r=q}^{R} E\left[\frac{I(C\geq r)}{P(C\geq q|T_q)}(\varphi_{q,\lambda}-\varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] \\ &= \sum_{r=1}^{q-1} E\left[\frac{1-I(C\leq q-1)}{1-P(C\leq q-1|T_{q-1})}(\varphi_{q,\lambda}-\varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] \text{ (will use } E[\varphi_{q,\lambda}|T_{q-1}]=\varphi_{q-1,\lambda}) \\ &\quad +\sum_{r=q}^{R} E\left[\frac{1-I(C\leq r-1)}{1-P(C\leq q-1|T_{q-1})}(\varphi_{q,\lambda}-\varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] \text{ (will use } s(Z_{(r)}|T_{r-1})\in L_0^2(F(Z_{(r)}|T_{r-1}))) \\ &= E[\varphi_{q,\lambda}s(Z_{(q)}|T_{q-1})']=E[ms(Z_{(q)}|T_{q-1})'|C\in\lambda] \end{aligned}$$

where the first equality in the last line follows by using (2.3) along with the comments inside parentheses on lines two and three from below; while the last one follows by mimicking steps (with T_1 replaced by T_q) from the last line in the derivation of B_1 above. Combining the above verifies (B.2).

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 (A.1). (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 (2.3). 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 (A.1).

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

$$(AM_\lambda)^{-1}A\ V_\lambda\ A'(AM_\lambda)^{-1\prime}$$

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 efficiency bound is $\Omega_\lambda:=(M_\lambda'V_\lambda^{-1}M_\lambda)^{-1}$ and the efficient influence function with variance equal to the efficiency bound is

$$\psi(A_*,O)=-\Omega_\lambda^{-1}M_\lambda'V_\lambda^{-1}\varphi_\lambda(O;\beta_\lambda^0). \ \blacksquare$$

Proof of Proposition 2.2:

Consider the first equality. 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 notational simplicity write $\varphi_{R,\lambda}(O;\beta)$ as $\varphi_{R,\lambda}$ and $T_r(Z)$ as T_r. We will use the following repeatedly:

$$P(C \geq R-r+1|T_{R-r+1}) = 1 - P(C \leq R-r|T_{R-r+1}) = 1 - P(C \leq R-r|T_{R-r}) = P(C \geq R-r+1|T_{R-r}) \quad \text{(A.3)}$$

for $r = 1\ldots, R$ where the second equality follows from (2.3).

Direct computation and (2.3) along with (A.3) give

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

Consider the under-braced part in the RHS of the expression for $\overline{\text{Proj}}_{T_{R-1}}(\phi_{R,\lambda}(\beta)|\,\phi_{R-1})$. Using $T_{R-1} \setminus T_{R-2} = Z_{R-1}$ and (2.3), 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 successive projections. (Terms with no contribution in successive projections are marked by under-braces in the sequel.) 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 (and use of (A.3)) gives for $r = 2$:

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

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

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

Now, once again using (A.3), note that

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

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

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

Proof of Proposition 3.1:

We closely follow the steps of the proof for Theorem 1 in Chen et al. (2003) with minor adjustments for the slightly weaker (and useful) conditions that are consequences of (3.7). The first adjustment is that we allow $\|\widehat{h} - h^\dagger\|_{\mathcal{H}} = o_p(1)$ where $h^\dagger \in \mathcal{H}$ need not be h^0. The second adjustment is that the uniform continuity assumption [(1.3) in Chen et al. (2003)] is automatically satisfied due to (3.7). An unrelated minor (since d_m is finite) adjustment is also required since we allow for $W_n \neq I_{d_m}$ or W. We write $\widehat{\beta}_\lambda(W_n)$ as $\widehat{\beta}$ and β^0_λ as β^0 for simplicity. Recall that (B3) implies that for all $\delta > 0$, $P(\|\widehat{\beta} - \beta^0\| > \delta) \leq P(\|G(\widehat{\beta}, h^\dagger)\| \geq \epsilon(\delta))$. Therefore, it is sufficient to show that $\|G(\widehat{\beta}, h^\dagger)\| = o_p(1)$. Assumption (B2) implies that $P(\widehat{h}(\beta) \in \mathcal{H}) \to 1$ as $n \to \infty$. The rest of the proof works conditional on this event, i.e., we use the fact that

$$\begin{aligned}
P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta)) &= P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta)|\widehat{h} \in \mathcal{H})P(\widehat{h} \in \mathcal{H}) + P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta)|\widehat{h} \notin \mathcal{H})P(\widehat{h} \notin \mathcal{H}) \\
&= P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta)|\widehat{h} \in \mathcal{H}) + o(1) \qquad \text{(A.4)}
\end{aligned}$$

as $n \to \infty$ and, instead, show that $\|G(\widehat{\beta}, h^\dagger)\| = o_p(1)$ conditional on $\widehat{h}(\beta) \in \mathcal{H}$. To this end, first note that

$$\|G(\widehat{\beta}, h^\dagger)\| \leq \|G(\widehat{\beta}, h^\dagger) - G(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h})\| + \|G_n(\widehat{\beta}, \widehat{h})\| = \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h})\| + \|G_n(\widehat{\beta}, \widehat{h})\| \quad \text{(A.5)}$$

where the inequality holds by the triangle inequality and the equality holds due to (3.7). Using (B4) and then (3.7):

$$\|G(\widehat{\beta},\widehat{h}) - G_n(\widehat{\beta},\widehat{h})\| \leq o_p(1)\{1 + \|G_n(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},\widehat{h})\|\} \leq o_p(1)\{1 + \|G_n(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},h^\dagger)\|\}.$$

Using this along with (A.5) gives

$$\begin{aligned}
\|G(\widehat{\beta},h^\dagger)\| \times (1 - o_p(1)) &\leq o_p(1) + \|G_n(\widehat{\beta},\widehat{h})\| \times (1 + o_p(1)) \\
&\leq o_p(1) + \|G_n(\widehat{\beta},\widehat{h})\|_{W_n} \times (1 + \|W_n^{-1} - W^{-1}\| + \|W^{-1} - I_{d_m}\|) \times (1 + o_p(1)) \\
&= o_p(1) + \|G_n(\widehat{\beta},\widehat{h})\|_{W_n} \times (c + o_p(1)) \leq o_p(1) + \inf_{\beta\in\mathcal{B}} \|G_n(\beta,\widehat{h})\|_{W_n} \times (c + o_p(1)) \quad \text{(A.6)}
\end{aligned}$$

where $c = 1 + \|W^{-1} - I_{d_m}\|$. The first equality in the last line follows since $W_n - W = o_p(1)$ and W is a constant (finite) positive definite matrix imply that W_n^{-1} exists with probability approaching one and $W_n^{-1} - W^{-1} = o_p(1)$ and hence $\|W_n^{-1} - W^{-1}\| = o_p(1)$ since d_m is finite, whereas finite and positive definite W and finite d_m imply that $c > 1$ is finite. The last inequality is due to (B1). Following similar steps again and letting $d = 1 + \|W - I_{d_m}\|$ (> 1 and finite), note that

$$\begin{aligned}
\|G_n(\beta,\widehat{h})\|_{W_n} &\leq \|G_n(\beta,\widehat{h})\| \times (d + o_p(1)) \\
&\leq \{\|G_n(\beta,\widehat{h}) - G(\beta,\widehat{h})\| + \|G(\beta,\widehat{h}) - G(\beta,h^\dagger)\| + \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\|\} \times (d + o_p(1)) \quad \text{(A.7)}
\end{aligned}$$

by using (3.5), i.e., $G(\beta^0,h) = 0$ for all $h \in \mathcal{H}$ (in the last term inside the braces). This is special feature of our setup; in Chen et al. (2003) this holds only at $h = h^0$. $\|G(\beta,\widehat{h}) - G(\beta,h^\dagger)\| = 0$ by (3.7). Using (B4) as before (in the first line)

$$\begin{aligned}
\|G_n(\beta,\widehat{h}) - G(\beta,\widehat{h})\| &\leq o_p(1)\{1 + \|G_n(\beta,\widehat{h})\| + \|G(\beta,h^\dagger)\| + o_p(1)\} = o_p(1) + \|G_n(\beta,\widehat{h})\| \times o_p(1) \\
&= o_p(1) + \|G_n(\beta,\widehat{h})\|_{W_n} \times (c + o_p(1)) \times o_p(1)
\end{aligned}$$

where the second line follows by and the same argument as in (A.6). Therefore, (A.7) gives

$$\begin{aligned}
\|G_n(\beta,\widehat{h})\|_{W_n} &\leq \{o_p(1) + \|G_n(\beta,\widehat{h})\|_{W_n} \times (c + o_p(1)) \times o_p(1) + \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\|\} \times (d + o_p(1)) \\
&= o_p(1) + \|G_n(\beta,\widehat{h})\|_{W_n} \times o_p(1) + \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\| \times (d + o_p(1))
\end{aligned}$$

and hence $\|G_n(\beta,\widehat{h})\|_{W_n} \times (1 - o_p(1)) \leq o_p(1) + \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\| \times (d + o_p(1))$ where all the $o_p(1)$ terms are uniform with respect to $\beta \in \mathcal{B}$. This implies

$$\inf_{\beta\in\mathcal{B}} \|G_n(\beta,\widehat{h})\|_{W_n} \leq \sup_{\beta\in\mathcal{B}} o_p(1) + \inf_{\beta\in\mathcal{B}} \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\|_W \times (d + \sup_{\beta\in\mathcal{B}} o_p(1)) = o_p(1)$$

since $\inf_{\beta\in\mathcal{B}} \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\|_W = 0$. Hence (A.6) and (A.4) give $\|G(\widehat{\beta},h^\dagger)\|_W = o_p(1)$. This completes the proof. ■

Proof of Proposition 3.2:

We closely follow the steps of the proof for Theorem 2 in Chen et al. (2003). Our conditions are, however, weaker since we use the special structure of our setup. In particular this is allowed by (3.6)-(3.8) and, along the course of the proof, we try to make their implications clear. We write $\widehat{\beta}_\lambda(W_n)$ as $\widehat{\beta}$ and β_λ^0 as β^0 for simplicity. The first step of the proof is to show the $\sqrt{n}$-consistency of $\widehat{\beta}$ and the final step is to show its asymptotic normality.

Since $\beta^0 \in$ interior($\mathcal{B}$), $h^\dagger(\beta) \in$ interior($\mathcal{H}$), $\widehat{\beta} - \beta = o_p(1)$ and $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$, we can choose a positive sequence $\delta_n = o_p(1)$ such that $P((\widehat{\beta},\widehat{h}) \in \mathcal{B}_{\delta_n} \times \mathcal{H}_{\delta_n}) \to 1$ as $n \to \infty$. For the δ in the statement of the proposition, $P(\mathcal{B}_{\delta_n} \times \mathcal{H}_{\delta_n} \subset \mathcal{B}_\delta \times \mathcal{H}_\delta) \to 1$ as $n \to \infty$. While to avoid repetition we do not make it explicit, it is important to keep in mind that as in the proof of Proposition 3.1, here also we work conditional on the event $\{(\widehat{\beta},\widehat{h}) \in \mathcal{B}_{\delta_n} \times \mathcal{H}_{\delta_n}\}$ which occurs with probability approaching one, i.e., we implicitly use arguments similar to (A.4) throughout the proof.

(C2) implies that there exists a constant $a > 0$ such that $P(a\|\widehat{\beta} - \beta^0\| \leq \|G(\widehat{\beta},h^\dagger)\|) \to 1$ as $n \to \infty$. Therefore, $\sqrt{n}$-consistency of $\widehat{\beta}$ follows if we can establish that $\|G(\widehat{\beta},h^\dagger)\| = O_p(n^{-1/2})$. To this end, note that

$$\begin{aligned}
\|G(\widehat{\beta},h^\dagger)\| &\leq \|G(\widehat{\beta},h^\dagger) - G(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},\widehat{h}) - G_n(\widehat{\beta},\widehat{h}) + G_n(\beta^0,h^\dagger)\| + \|G_n(\widehat{\beta},\widehat{h})\| + \|G_n(\beta^0,h^\dagger)\| \\
&= 0 + \|G(\widehat{\beta},\widehat{h}) - G_n(\widehat{\beta},\widehat{h}) + G_n(\beta^0,h^\dagger)\| + \|G_n(\widehat{\beta},\widehat{h})\| + O_p(n^{-1/2}) \quad \text{(A.8)}
\end{aligned}$$

where the first 0 follows from (3.7) and the last $O_p(n^{-1/2})$ from (C4). Now, by (C3) for the first inequality below,

$$\begin{aligned}
\|G(\widehat{\beta},\widehat{h}) - G_n(\widehat{\beta},\widehat{h}) + G_n(\beta^0,h^\dagger)\| &\leq o_p(1) \times \{n^{-1/2} + \|G_n(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},\widehat{h})\|\} \\
&\leq o_p(1) \times \{n^{-1/2} + \|G_n(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},\widehat{h}) - G(\widehat{\beta},h^\dagger)\| + \|G(\widehat{\beta},h^\dagger)\|\} \\
&= o_p(1) \times \{n^{-1/2} + \|G_n(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},h^\dagger)\|\}
\end{aligned}$$

where the last line follows by (3.7). Therefore, this along with (A.8) imply

$$\|G(\widehat{\beta},h^\dagger)\| \leq o_p(1) \times \{n^{-1/2} + \|G_n(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},h^\dagger)\|\} + \|G_n(\widehat{\beta},\widehat{h})\| + O_p(n^{-1/2})$$

which, further implies that (second inequality below follows using same arguments as in (A.6) with $c = 1 + \|W^{-1} - I_{d_m}\|$)

$$\begin{aligned} \|G(\widehat{\beta}, h^\dagger)\| \times (1 - o_p(1)) &\leq O_p(n^{-1/2}) + \|G_n(\widehat{\beta}, \widehat{h})\| \times (1 + o_p(1)) \\ &\leq O_p(n^{-1/2}) + \|G_n(\widehat{\beta}, \widehat{h})\|_{W_n} \times (c + o_p(1)) \\ &\leq O_p(n^{-1/2}) + \inf_{\beta \in \mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\|_{W_n} \times (c + o_p(1)) \end{aligned} \tag{A.9}$$

where the last line follows by (C1). Now, for $d = 1 + \|W - I_{d_m}\|$, recall from the first line of (A.7) that $\|G_n(\beta, \widehat{h})\|_{W_n} \leq \|G_n(\beta, \widehat{h})\| \times (d + o_p(1))$. On the other hand,

$$\begin{aligned} \|G_n(\beta, \widehat{h})\| &\leq \|G_n(\beta, \widehat{h}) - G(\beta, \widehat{h}) - G_n(\beta^0, h^\dagger)\| + \|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| + \|G(\beta, h^\dagger)\| + \|G_n(\beta^0, h^\dagger)\| \\ &\leq o_p(1) \times \{n^{-1/2} + \|G_n(\beta, \widehat{h})\| + \|G(\beta, \widehat{h})\|\} + 0 + \|G(\beta, h^\dagger)\| + O_p(n^{-1/2}) \end{aligned}$$

where the first term in the last line follows from (C3), the third term from (3.7) and the last one from (C4). Therefore,

$$\begin{aligned} \|G_n(\beta, \widehat{h})\| \times (1 - o_p(1)) &\leq \|G(\beta, \widehat{h})\| \times o_p(1) + \|G(\beta, h^\dagger)\| + O_p(n^{-1/2}) \\ &\leq \|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| \times o_p(1) + \|G(\beta, h^\dagger)\| \times (1 + o_p(1)) + O_p(n^{-1/2}) \\ &= \|G(\beta, h^\dagger)\| \times (1 + o_p(1)) + O_p(n^{-1/2}) \text{ [by (3.7)]} \\ &\leq \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\| \times (1 + o_p(1)) + \|G(\beta^0, h^\dagger)\| \times (1 + o_p(1)) + O_p(n^{-1/2}) \\ &= \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\| \times (1 + o_p(1)) + O_p(n^{-1/2}) \end{aligned}$$

since $G(\beta^0, h^\dagger) = 0$. Therefore, $\|G_n(\beta, \widehat{h})\|_{W_n} \leq \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\| \times (d + o_p(1)) + O_p(n^{-1/2})$ where all the o_p and O_p terms are uniform with respect to $\beta \in \mathcal{B}_\delta$. Hence, as in the proof of Proposition 3.1, noting that $\inf_{\beta \in \mathcal{B}} \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\| = 0$, it follows that $\inf_{\beta \in \mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\|_{W_n} = O_p(n^{-1/2})$ and, therefore, (A.9) gives $\|G(\widehat{\beta}, h^\dagger)\| = O_p(n^{-1/2})$ and subsequently $\widehat{\beta} - \beta^0 = O_p(n^{-1/2})$.

To establish asymptotic normality, define the linearization $L_n(\beta) = G_n(\beta^0, h^\dagger) + M_\lambda(\beta - \beta^0)$. Note that the differences from the linearization in Chen et al. (2003) arise due to (3.6) and (3.8). This gives

$$\begin{aligned} \|G_n(\widehat{\beta}, \widehat{h}) - L_n(\widehat{\beta})\| &= \|G_n(\widehat{\beta}, \widehat{h}) - G_n(\beta^0, h^\dagger) - M_\lambda(\widehat{\beta} - \beta^0)\| \\ &= \|G_n(\widehat{\beta}, \widehat{h}) - G_n(\beta^0, h^\dagger) - G(\widehat{\beta}, \widehat{h}) + G(\widehat{\beta}, \widehat{h}) + G(\widehat{\beta}, h^\dagger) - G(\widehat{\beta}, h^\dagger) - M_\lambda(\widehat{\beta} - \beta^0)\| \\ &\leq \|G_n(\widehat{\beta}, \widehat{h}) - G_n(\beta^0, h^\dagger) - G(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h}) - G(\widehat{\beta}, h^\dagger)\| + \|G(\widehat{\beta}, h^\dagger) - M_\lambda(\widehat{\beta} - \beta^0)\| \\ &\leq \|G_n(\widehat{\beta}, \widehat{h}) - G_n(\beta^0, h^\dagger) - G(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, h^\dagger) - M_\lambda(\widehat{\beta} - \beta^0)\| \text{ [by (3.7)]} \\ &\leq o_p(1) \times \{1 + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h})\|\} + \|G(\widehat{\beta}, h^\dagger) - G(\beta^0, h^\dagger) - M_\lambda(\widehat{\beta} - \beta^0)\| \end{aligned}$$

where the term inside braces follow from (C3) and the inclusion of $G(\beta^0, h^\dagger)$ in the last term is innocuous since $G(\beta^0, h^\dagger) = 0$. Now, by the definition of M_λ, assumptions (C2), (A3) and (3.6), it follows that $\|G(\widehat{\beta}, h^\dagger) - G(\beta^0, h^\dagger) - M_\lambda(\widehat{\beta} - \beta^0)\| = o_p(\|\widehat{\beta} - \beta^0\|)$, which is $o_p(n^{-1/2})$ since $\widehat{\beta} - \beta^0 = O_p(n^{-1/2})$. On the other hand, the same steps from the top line of (A.9) until (almost) the end of the first part of the proof give $\|G_n(\widehat{\beta}, \widehat{h})\| \leq \inf_{\beta \in \mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\| + o_p(n^{-1/2}) = O_p(n^{-1/2})$. Finally, since $\|G(\widehat{\beta}, \widehat{h})\| \leq \|G(\widehat{\beta}, \widehat{h}) - G(\widehat{\beta}, h^\dagger)\| + \|G(\widehat{\beta}, h^\dagger)\| = O_p(n^{-1/2})$ because the first term is 0 by (3.7) and the second term is $O_p(n^{-1/2})$ from the first part of the proof, we obtain that $\|G_n(\widehat{\beta}, \widehat{h}) - L_n(\widehat{\beta})\| \leq o_p(n^{-1/2})$. Similarly, for $\bar{\beta} := \arg\min_\beta \|L_n(\beta)\|_W$, that, by construction, satisfies $\sqrt{n}(\bar{\beta} - \beta^0) = -(M_\lambda' W M_\lambda)^{-1} M_\lambda' W \sqrt{n} G_n(\beta^0, h^\dagger)$, we can show that $\|G_n(\bar{\beta}, \widehat{h}) - L_n(\bar{\beta})\| \leq o_p(n^{-1/2})$. Now that the proximity of $G_n(\beta, \widehat{h})$ and $L_n(\beta)$ has been established at $\widehat{\beta}$ and $\bar{\beta}$ respectively, the rest of the proof is to show that $\sqrt{n}(\bar{\beta} - \widehat{\beta}) = o_p(1)$. As was the case in Chen et al. (2003), this does not involve anything particularly related to the special feature of our setup (it only works with the linearization), and hence follows exactly in the same way as in the proof of Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989). ■

Proof of Corollary 3.3:

(1) This is standard and hence the proof is omitted.

(2) This follows by noting that $g(O; \beta, h^0(O; \beta)) = \varphi_\lambda(O; \beta)$ defined in (2.8). ■

B Technical Appendix B: Additional results

To support the discussion in Section 4, we provide the following results. First, under a simplified setup and assuming INDEP (2.5) and CMAR (2.4) respectively, we provide expressions for the loss defined in (4.1). Now recall that to further investigate the case of close to zero loss in our simulation results (when the moment restrictions we use do not support zero loss), we pointed out in footnote 18 a scenario under which the loss is actually zero in some cases without imposing any restriction on the moment vector or the $Z_{(r)}$'s. The second result provides a theoretical foundation for this footnote by establishing the efficiency bound under (2.4) for the case where $P(C = r|T_1(Z))$ is unknown. This assumption has also been used in the second attrition analysis (page 145) in Fitzgerald et al. (1998), or under the name of sequential ignorability for identification in mediation analysis in e.g. Imai et al. (2010b), Imai et al. (2010a), Shpitser and Tchetgen (2012, 2014). For the sake of completeness, we also establish the efficiency bounds under (2.4) for the cases where $P(C = r|T_1(Z))$ is

known (a special case of Proposition 2.1), and partially known up to finite dimensional parameters. In what follows we write Z_r instead of $Z_{(r)}$ for $r = 1, \ldots, R$ to simplify notation.

B.1 Results

B.1.1 The expression for loss in (4.1)

The expression for the loss in (4.1) can be complicated in general i.e. under MAR (2.3), but are simpler under CMAR (2.4) and INDEP (2.5), and we list them below. For simplicity let $R = 3$ and $d_\beta = d_m = 1$. The simplified expressions here do not act as the exact reference for the reported simulation results, but may be useful for an intuitive understanding of what constitutes the loss. These expressions are obtained from Proposition 2.1. Let $V_\lambda^{s'}$ denote the variance of $\varphi_\lambda(O; \beta_\lambda^0)$ when the latter is modified according to the discussion below (4.1). To avoid clutter, we write $m(Z; \beta_\lambda^0)$ as m, $P(C = r)$ as p_r, $P(C = r|Z_1)$ as $p_r(Z_1)$, $P(C \in \{r, t\})$ as p_{rt} and $P(C \in \{r, t\}|Z_1)$ as $p_{rt}(Z_1)$ for $r, t = 1, 2, 3$.

Result 1: Under INDEP β_λ^0 is the same for all $\lambda \in \Lambda$. Taking $\lambda = \mathcal{C} := \{1, 2, 3\}$, the following hold as $n \to \infty$:

(a) $\text{Loss}(\beta_\lambda; s = \{3\}, s' = \{1, 3\}) \times V_\lambda^{\{1,3\}} = \frac{p_1}{p_3} E\left[E[m|Z_1]^2\right]$.

(b) $\text{Loss}(\beta_\lambda; s = \{3\}, s' = \{2, 3\}) \times V_\lambda^{\{2,3\}} = \frac{p_2}{p_3} E\left[E[m|Z_1, Z_2]^2\right]$.

(c) $\text{Loss}(\beta_\lambda; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_\lambda^{\{1,2,3\}} = \frac{p_2}{p_{13}} E\left[E[m|Z_1]^2\right] + \frac{p_2}{p_3 p_{23}} E\left[Var(E[m|Z_1, Z_2]|Z_1)\right]$.

(d) $\text{Loss}(\beta_\lambda; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_\lambda^{\{1,2,3\}} = \frac{p_1}{p_{23}} E\left[E[m|Z_1]^2\right]$.

Result 2: Under CMAR we do not consider $s = \{3\}$ for brevity, unless $\lambda = \{3\}$. The following hold as $n \to \infty$:

(a) $\text{Loss}(\beta_{\{1\}}; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_{\{1\}}^{\{1,2,3\}} = E\left[\frac{p_1(Z_1)p_2(Z_1)}{p_1}\left\{\frac{E[m|Z_1]^2}{p_{13}(Z_1)} + \frac{Var(E[m|Z_1,Z_2)|Z_1)}{p_3(Z_1)p_{23}(Z_1)}\right\}\Big| C = 1\right]$.

(b) $\text{Loss}(\beta_{\{2\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_{\{2\}}^{\{1,2,3\}} = E\left[\frac{p_1(Z_1)p_2(Z_1)}{p_2 p_{23}(Z_1)} E[m|Z_1]^2 \Big| C = 2\right]$.

(c1) $\text{Loss}(\beta_{\{3\}}; s = \{3\}, s' = \{1, 3\}) \times V_{\{3\}}^{\{1,3\}} = \frac{p_{13}}{p_3} E\left[\frac{p_1(Z_1)}{p_{13}(Z_1)} E[m|Z_1]^2 \Big| C = 3\right]$.

(c2) $\text{Loss}(\beta_{\{3\}}; s = \{3\}, s' = \{2, 3\}) \times V_{\{3\}}^{\{2,3\}} = \frac{p_{23}}{p_3} E\left[\frac{p_2(Z_1)}{p_{23}(Z_1)} E[m|Z_1, Z_2]^2 \Big| C = 3\right]$.

(c3) $\text{Loss}(\beta_{\{3\}}; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_{\{3\}}^{\{1,2,3\}} = E\left[\frac{p_2(Z_1)p_3(Z_1)}{p_3}\left\{\frac{E[m|Z_1]^2}{p_{13}(Z_1)} + \frac{Var(E[m|Z_1,Z_2)|Z_1)}{p_3(Z_1)p_{23}(Z_1)}\right\}\Big| C = 3\right]$.

(c4) $\text{Loss}(\beta_{\{3\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_{\{3\}}^{\{1,2,3\}} = E\left[\frac{p_1(Z_1)p_3(Z_1)}{p_3 p_{23}(Z_1)} E[m|Z_1]^2 \Big| C = 3\right]$.

(d) $\text{Loss}(\beta_{\{1,3\}}; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(Z_1)p_{13}(Z_1)}{p_{13}}\left\{\frac{E[m|Z_1]^2}{p_{13}(Z_1)} + \frac{Var(E[m|Z_1,Z_2]|Z_1)}{p_3(Z_1)p_{23}(Z_1)}\right\}\Big| C \in \{1, 3\}\right]$.

(e) $\text{Loss}(\beta_{\{2,3\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(Z_1)}{p_{23}(Z_1)} E[m|Z_1]^2 \Big| C \in \{2, 3\}\right]$.

(f1) $\text{Loss}(\beta_{\{1,2,3\}}; s = \{1, 3\}, s' = \{1, 2, 3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(Z_1)}{p_{13}(Z_1)} E[m|Z_1]^2 + \frac{p_2(Z_1)}{p_3(Z_1)p_{23}(Z_1)} Var(E[m|Z_1, Z_2]|Z_1)\right]$.

(f2) $\text{Loss}(\beta_{\{1,2,3\}}; s = \{2, 3\}, s' = \{1, 2, 3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(Z_1)}{p_{23}(Z_1)} E[m|Z_1]^2\right]$.

Remarks:

1. It is therefore clear from Results 1 and 2 that there is not always a strict loss in efficiency in the sense of (4.1) and under the premise of our discussion below it, when one does not use all the sub-samples. For example, if $E[m|Z_1, Z_2] = 0$ then there is never any loss in all the above cases. This happens if, e.g., we consider a regression setup: $Z_3 = Z_1\beta + \epsilon$ with $E[\epsilon|Z_1] = 0$ and let Z_2 be any variable such that $E[\epsilon|Z_1, Z_2] = 0$ (e.g. a surrogate for Z_1). Similarly, there is no loss in Result 1 (a), (d) and Result 2 (b), (c1), (c4), (e), (f2) if only $E[m|Z_1] = 0$, which happens in the same regression setup but without any restriction on Z_2. These results complement the discussion on efficiency in Section 5 of Wooldridge (2007).

2. $s = \{3\}$ corresponds to both the IPW and CC estimators that are numerically equivalent if $\lambda = \{3\}$ or under INDEP. Otherwise, CC is generally not consistent. The IPW estimator is not considered in Result 2 (except if $\lambda = \{3\}$) because it was already discussed extensively following Proposition 2.1. The expressions under MAR are complicated and less intuitive since the conditioning set for the conditional probabilities are not always reduced to Z_1 as under CMAR (or a constant as under INDEP) and hence the probabilities factor out even less than they do under CMAR.

B.1.2 Theoretical foundation for footnote 18

To avoid notational clutter we sometimes omit O and β from the argument of all variables with the understanding that these variables are evaluated at $\beta = \beta_\lambda^0$. Our use of the notation $\{C \geq R\}$ and $E[m(Z; \beta)|T_R(Z)]$ for the sake of providing shorter expressions should be read as $\{C = R\}$ and $m(Z; \beta)$ respectively.

Proposition B.1 *Let* (2.1), (2.3), (2.6) *and assumption A hold. Define* $q_\lambda(T_1(Z)) := P(C \in \lambda|T_1(Z))/P(C \in \lambda)$ *and*

$$\varphi_{\lambda[k]}(O; \beta) := q_\lambda(T_1(Z)) \left\{ \sum_{r=1}^{R-1} \frac{I(C \geq R - r + 1)}{P(C \geq R - r + 1|T_1(Z))} \left(E[m(Z; \beta)|T_{R-r+1}(Z)] - E[m(Z; \beta)|T_{R-r}(Z)]\right) + E[m(Z; \beta)|T_1(Z)] \right\}$$

where the subscript [k] denotes that $P(C = r|T_1(Z))$ is known. Let $V_{\lambda[k]} := Var(\varphi_{\lambda[k]})$ be a $d_m \times d_m$ finite positive definite matrix. 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 $\Omega_{\lambda[k]} := (M'_\lambda V^{-1}_{\lambda[k]} M_\lambda)^{-1}$. An estimator whose asymptotic variance equals $\Omega_{\lambda[k]}$ has the asymptotically linear representation

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

Proposition B.2 *Let (2.1), (2.3) 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*

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

where least squares projection $\Pi(.|.)$ is defined above (3.4), the subscript [pk] denotes that $P(C = r|T_1(Z))$ is partially known, i.e., known up to a finite dimensional parameter γ. Let $V_{\lambda[pk]} := Var(\varphi_{\lambda[pk]})$ be a $d_m \times d_m$ finite positive definite matrix. 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 $\Omega_{\lambda[pk]} := (M'_\lambda V^{-1}_{\lambda[pk]} M_\lambda)^{-1}$. An estimator whose asymptotic variance equals $\Omega_{\lambda[pk]}$ has the asymptotically linear representation

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

Proposition B.3 *Let (2.1), (2.3) and assumption A hold. Define $q_\lambda(T_1(Z))$ as above, $q_\lambda := I(C \in \lambda)/P(C \in \lambda)$ and*

$$\varphi_{\lambda[u]}(O; \beta) := q_\lambda(T_1(Z)) \sum_{r=1}^{R-1} \frac{I(C \geq R - r + 1)}{P(C \geq R - r + 1|T_1(Z))} \left(E[m(Z; \beta)|T_{R-r+1}(Z)] - E[m(Z; \beta)|T_{R-r}(Z)]\right) + q_\lambda E[m(Z; \beta)|T_1(Z)]$$

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^0_λ, the asymptotic variance lower bound for $\sqrt{n}(\widehat{\beta} - \beta^0_\lambda)$ of any regular estimator $\widehat{\beta}$ is given by $\Omega_{\lambda[u]} := (M'_\lambda V^{-1}_{\lambda[u]} M_\lambda)^{-1}$. An estimator whose asymptotic variance equals $\Omega_{\lambda[u]}$ has the asymptotically linear representation

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

Remarks

(1) Proposition B.1 is a special case of Proposition 2.1. Proposition B.2 is an intermediate result that has not been used in our paper but is provided for the sake of completeness. It is unclear to us when the scenario of this proposition is appropriate in practice. Proposition B.3 is a new result that has been used in the discussion in Section 4. No general counterpart for Proposition B.3 is provided in our paper, although results from very limited choices of λ, e.g. $\lambda = \{1\}$ or $\lambda = \mathcal{C}$, under the generalization of (2.4) to (2.3) are available from the author. To our knowledge all these results are new in the sense that the literature has so far considered either $R = 2$ and $\lambda = \{1\}$, $\lambda = \mathcal{C}$ or a general R but $\lambda = \mathcal{C}$.

(2) It is straightforward to see from Propositions B.1-B.3 that, in terms of the infeasible data $(C, Z')'$:

$$\begin{aligned}
V_{\lambda[u]} &= E\left[\frac{P(C \in \lambda|Z_1)}{P^2(C \in \lambda)} E[m(Z; \beta^0_\lambda)|Z_1] E[m(Z; \beta^0_\lambda)'|Z_1] + \frac{P^2(C \in \lambda|Z_1)}{P^2(C \in \lambda)} \sum_{r=2}^{R} \frac{Var\left(E[m(Z; \beta^0_\lambda)|Z_1, \ldots, Z_r]|Z_1, \ldots, Z_{r-1}\right)}{P(C \geq r|Z_1)}\right] \\
V_{\lambda[k]} &= V_{\lambda[u]} - \frac{P(C \in \lambda|Z_1)(1 - P(C \in \lambda|Z_1))}{P^2(C \in \lambda)} E[m(Z; \beta^0_\lambda)|Z_1] E[m(Z; \beta^0_\lambda)|Z_1]' \\
V_{\lambda[pk]} &= V_{\lambda[k]} + B\left(E\left[S_\gamma(C|Z_1)S_\gamma(C|Z_1)'\right]\right)^{-1} B' \\
&= V_{\lambda[u]} - Var\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z; \beta^0_\lambda)|Z_1] - \Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z; \beta^0_\lambda)|Z_1] \middle| S_\gamma(C, Z_1)\right)\right)
\end{aligned}$$

where $B := E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z; \beta^0_\lambda)|T_1(Z)] S_\gamma(C|T_1(Z))'\right] = E\left[\frac{E[m(Z; \beta^0_\lambda)|Z_1]}{P(C \in \lambda)} \sum_{r \in \lambda} \frac{\partial}{\partial \gamma'} P(C = r|Z_1; \gamma^0)\right]$. Therefore, $V_{\lambda[k]} = V_{\lambda[pk]} = V_{\lambda[u]}$ if $\lambda = \mathcal{C}$ and, otherwise, $V_{\lambda[k]} \leq V_{\lambda[pk]} \leq V_{\lambda[u]}$ in the matrix sense. This ordering of the asymptotic variances establishes the result for a general R and a general choice of target sub-population λ by generalizing the result with $R = 2$ and $\lambda = \{1\}$, $\lambda = \{1, 2\}$ from Chen et al. (2008).

B.2 Proofs

Proofs for Results 1 and 2: The proofs follow are simple applications of a slightly modified (following the discussion below (4.1)) version of Proposition 2.1. We do not provide them for brevity but they are available from the author. We only note here that the proof of Result 2 (f1) and (f2) requires minor change in the proof of Proposition 2.1 that amounts to recognizing that for $j = 2, 3$: $E[m(Z; \beta)] = E\left[\frac{I(C \in \{j,3\})}{P(C \in \{j,3\}|Z_1)} m(Z; \beta)\right] = E\left[\frac{P(C \in \{j,3\})}{P(C \in \{j,3\}|Z_1)} m(Z; \beta) \middle| C \in \{j, 3\}\right]$. ■

The proof of Proposition B.1 is a special case of that for Proposition 2.1 with $P(C = r|T_r(Z)) = P(C = 1|T_1(Z))$ for all $r = 1, \ldots, R$, and hence is omitted. We now prove Propositions B.2 and B.3 in reverse order since the latter proof requires more details, and more importantly since Proposition B.3 is directly related to footnote 18. We omit some details (e.g. we take $d_m = d_\beta$) that were laid out explicitly and elaborately in the proof of Proposition 2.1.

Proof of Proposition B.3: $[d_m = d_\beta]$

STEP - 1: For a parametric path θ of the joint distribution of O, the log of the density in terms of $(C, Z')'$ is

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

where $f_{\theta_0}(O)$ is the true joint density $f(O)$. The score function with respect to θ is

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

where $\dot{P}_\theta(C = r|Z_1) := \frac{\partial}{\partial\theta} P_\theta(C = r|Z_1)$, $s_\theta(Z_1) := \frac{\partial}{\partial\theta} \log f_\theta(Z_1)$ and $s_\theta(Z_r|Z_1, \ldots, Z_{r-1}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_r|Z_1, \ldots, Z_{r-1})$. Henceforth, we omit the subscript θ from the quantities evaluated at $\theta = \theta_0$. The tangent set for the model can then be characterized by functions of the form:

$$\mathcal{T} := a(Z_1) + \sum_{r=1}^{R} I(C = r) \frac{b_r(Z_1)}{a_r(Z_1)} + \sum_{r=2}^{R} I(C \geq r) a(Z_1, \ldots, Z_r),$$

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

For a given $\lambda \in \Lambda$, the following relation obtained by two different factorization of the joint distribution of $(I(C \in \lambda), Z_1)$ helps us to switch between different factorizations:

$$\begin{aligned} & s(Z_1) + I(C \in \lambda) \frac{\dot{P}(C \in \lambda|Z_1)}{P(C \in \lambda|Z_1)} + I(C \notin \lambda) \frac{\dot{P}(C \notin \lambda|Z_1)}{P(C \notin \lambda|Z_1)} \\ = \quad & I(C \in \lambda) \left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(Z_1|C \in \lambda) \right] + I(C \notin \lambda) \left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(Z_1|C \notin \lambda) \right]. \end{aligned} \tag{B.1}$$

STEP - 2: Differentiating $E[m(Z; \beta^0)|C \in \lambda] = 0$ under the integral we obtain by using (2.4) that

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} = -M_\lambda^{-1} E \left[m(Z) \left\{ s(Z_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_r|Z_1, \ldots, Z_{r-1})' \right\} \middle| C \in \lambda \right].$$

As in the proof of Proposition 2.1 we now verify that

$$E[\varphi_{\lambda[u]}(O) S(O)'] = E \left[m(Z) \left\{ s(Z_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_r|Z_1, \ldots, Z_{r-1})' \right\} \middle| C \in \lambda \right].$$

For the r-th term of $\varphi_{\lambda[u]}(O)$ where $r = 1, \ldots, R-1$: (2.4) and $E[E[m(Z)|T_{R-r+1}(Z)]|T_{R-r}(Z)] = E[m(Z)|T_{R-r}(Z)]$ give

$$\begin{aligned} & E \left[q_\lambda(T_1(Z)) \frac{I(C \geq R - r + 1)}{P(C \geq R - r + 1|T_1(Z))} \left(E[m(Z)|T_{R-r+1}(Z)] - E[m(Z)|T_{R-r}(Z)] \right) S(O)' \right] \\ = & \sum_{l=R-r+1}^{R} E \left[q_\lambda(T_1(Z)) \frac{I(C \geq R - r + 1)}{P(C \geq R - r + 1|T_1(Z))} \left(E[m(Z)|T_{R-r+1}(Z)] - E[m(Z)|T_{R-r}(Z)] \right) s(Z_{(l)}|T_{l-1}(Z))' \right] \\ = & E \left[q_\lambda(T_1(Z)) \frac{I(C \geq R - r + 1)}{P(C \geq R - r + 1|T_1(Z))} E[m(Z)|T_{R-r+1}(Z)] s(Z_{(R-r+1)}|T_{R-r}(Z))' \right] = E \left[m(Z) s(Z_{(R-r+1)}|T_{R-r}(Z))' | C \in \lambda \right] \end{aligned}$$

where the second last equality follows by $s(Z_{(R-r+l)}|T_{R-r}(Z)) \in L_0^2(F(Z_{(R-r+1)}|T_{R-r}(Z)))$, and the last one by (2.4).

Now consider the R-th term of $\varphi_{\lambda[u]}(O)$. Since $s(Z_r|Z_1, \ldots, Z_{r-1}) \in L_0^2(F(Z_r|T_{r-1}(Z)))$ for $r = 2, \ldots, R$:

$$\begin{aligned} & E \left[q_\lambda E[m(Z)|T_1(Z)] S(O)' \right] \\ = \quad & E \left[q_\lambda E[m(Z)|Z_1] \left\{ s(Z_1)' + \sum_{r=1}^{R} I(C = r) \frac{\dot{P}(C = r|Z_1)'}{P(C = r|Z_1)} \right\} \right] \\ = \quad & E \left[q_\lambda E[m(Z)|Z_1] \left\{ \frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(Z_1|C \in \lambda) - \frac{\dot{P}(C \in \lambda|Z_1)}{P(C \in \lambda|Z_1)} \right\}' \right] + E \left[E[m(Z; \beta^0)|Z_1] \frac{\dot{P}(C \in \lambda|Z_1)'}{P(C \in \lambda)} \right] \end{aligned}$$

where the first term in the last line follows by the use of (B.1) to replace $s(Z_1)$, whereas the last term follows by using (2.4) to

see that $E\left[I(C \in \lambda) \sum_{r=1}^{R} I(C=r) \frac{\dot{P}(C=r|Z_1)}{P(C=r|Z_1)} \Big| Z_1 \right] = \sum_{r \in \lambda} P(C=r|Z_1) \frac{\dot{P}(C=r|Z_1)}{P(C=r|Z_1)} = \sum_{r \in \lambda} \dot{P}(C=r|Z_1) = \dot{P}(C \in \lambda|Z_1)$. Repeated use of (2.4) simplifies the above equation as[19]

$$E\left[E[m(Z)|Z_1]|C \in \lambda\right] \frac{\dot{P}(C \in \lambda)'}{P(C \in \lambda)} + E\left[E[m(Z)|Z_1] s(Z_1|C \in \lambda)'|C \in \lambda\right] = 0 + E[m(Z) s(Z_1|C \in \lambda)'|C \in \lambda].$$

The zero in the RHS follows from (2.1). The second term follows by (2.4) and noting that $E\left[E[m(Z)|Z_1] s(Z_1|C \in \lambda)'|C \in \lambda\right] = E\left[E[m(Z) s(Z_1|C \in \lambda)'|Z_1, C \in \lambda]|C \in \lambda\right] = E\left[m(Z) s(Z_1|C \in \lambda)'|C \in \lambda\right]$. This completes the verification.

The rest of the proof is similar to the proof of Proposition 2.1. ■

Proof of Proposition B.2: [$d_m = d_\beta$. $P(C=r|T_1(Z)) = P(C=r|T_1(Z);\gamma^0)\ \forall\ r$. $\gamma^0 \in \Gamma \subset \mathbb{R}^{d_\gamma}$ unknown.]

STEP - 1: The same factorization of the joint density of O gives the score function with respect to θ as

$$S_\theta(O) \quad = \quad s_\theta(Z_1) + \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|Z_1)} \left(\frac{\partial P(C=r|Z_1;\gamma^0)}{\partial \gamma'} \frac{\partial \gamma^0}{\partial \theta'} \right)' + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_r|Z_1, \ldots, Z_{r-1}).$$

Recall that $S_\gamma(C|T_1(Z)) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|Z_1)} \frac{\partial}{\partial \gamma} P(C=r|Z_1;\gamma^0)$. Let b denote a $d_\gamma \times d_\theta$ constant matrix (note $\frac{\partial \gamma^0}{\partial \theta'}$). Then the tangent set for the model is characterized by the set of functions:

$$\mathcal{T} := a(Z_1) + b' S_\gamma(C|Z_1) + \sum_{r=2}^{R} I(C \geq r) a(Z_1, \ldots, T_r),$$

where $a(Z_1) \in L_0^2(F(Z_1))$, $S_\gamma(C|Z_1) \in L_0^2(F(C|Z_1))$ and $a(Z_1, \ldots, Z_r) \in L_0^2(F(Z_r|Z_1, \ldots, Z_{r-1}))$.

Recognizing that $P(C=r|Z_1) = P(C=r|Z_1;\gamma^0)$ is known up to the finite (d_γ) dimensional parameter γ, alters the relationship in (B.1) as follows

$$\begin{aligned} & s(Z_1) + \frac{\partial \gamma^{0\prime}}{\partial \theta} \left[I(C \in \lambda) \frac{\frac{\partial}{\partial \gamma} P(C \in \lambda|Z_1;\gamma^0)}{P(C \in \lambda|Z_1)} + I(C \notin \lambda) \frac{\frac{\partial}{\partial \gamma} P(C \notin \lambda|Z_1;\gamma^0)}{P(C \notin \lambda|Z_1)} \right] \\ = \quad & I(C \in \lambda) \left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(Z_1|C \in \lambda) \right] + I(C \notin \lambda) \left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(Z_1|C \notin \lambda) \right]. \end{aligned}$$

STEP - 2: Differentiating (2.2) under the integral, following the corresponding steps in the proof of Proposition B.3 and using the above relationship give:

$$\begin{aligned} \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} \quad = \quad & -M_\lambda^{-1} E\left[\frac{P(C \in \lambda|Z_1)}{P(C \in \lambda)} m(Z) \left\{ s(Z_1)' + \sum_{r=2}^{R} s(Z_r|Z_1, \ldots, Z_{r-1})' \right\} \right] \\ & -M_\lambda^{-1} E\left[E[m(Z)|Z_1] \frac{\frac{\partial}{\partial \gamma'} P(C \in \lambda|Z_1;\gamma^0)}{P(C \in \lambda)} \frac{\partial \gamma^0}{\partial \theta'} \right]. \end{aligned}$$

Therefore, utilizing the expression of the efficient influence function in Proposition B.1 and its relation to that in Proposition B.2, the verification of pathwise differentiability essentially boils to verifying that

$$E\left[\Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|T_1(Z)] \Big| S_\gamma(C|T_1(Z)) \right) S(O)' \right] = E\left[E[m(Z)|Z_1] \frac{\frac{\partial}{\partial \gamma'} P(C \in \lambda|Z_1;\gamma^0)}{P(C \in \lambda)} \frac{\partial \gamma^0}{\partial \theta'} \right].$$

Note that $E\left[S_\gamma(C|Z_1) \left\{ s(Z_1)' + \sum_{r=2}^{R} s(Z_r|Z_1, \ldots, Z_{r-1})' \right\} \right] = 0$ by using (term by term) that $E[S_\gamma(C|Z_1)|Z_1] = 0$ for term 1; $s(Z_r|Z_1, \ldots, Z_{r-1}) \in L_0^2(F(Z_r|Z_1, \ldots, Z_{r-1})$ and (2.4) for the rest. Hence the LHS of the above simplifies to $E\left[\Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|Z_1] \Big| S_\gamma(C|Z_1) \right) S_\gamma(C|Z_1)' \right] \frac{\partial \gamma^0}{\partial \theta'} = E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|Z_1] S_\gamma(C|Z_1)' \right] \frac{\partial \gamma^0}{\partial \theta'}$ giving

$$\begin{aligned} LHS \quad = \quad & E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|Z_1] \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|Z_1)} \frac{\partial P(C=r|Z_1;\gamma^0)}{\partial \gamma'} \right] \frac{\partial \gamma^0}{\partial \theta'} \\ = \quad & E\left[\frac{1}{P(C \in \lambda)} E[m(Z)|Z_1] \sum_{r \in \lambda} \frac{P(C=r|Z_1)}{P(C=r|Z_1)} \frac{\partial P(C=r|Z_1;\gamma^0)}{\partial \gamma'} \right] \frac{\partial \gamma^0}{\partial \theta'} \\ = \quad & E\left[\frac{1}{P(C \in \lambda)} E[m(Z)|Z_1] \frac{\partial P(C \in \lambda|Z_1;\gamma^0)}{\partial \gamma'} \right] \frac{\partial \gamma^0}{\partial \theta'} = RHS \ \blacksquare \end{aligned}$$

[19] It is precisely this step that creates a problem with establishing a similar general (for all λ) result under the less restrictive (2.3).