Online supplemental material for:

A Note on Efficiency Gains from Multiple Incomplete Sub-samples

Saraswata Chaudhuri[15]

This supplemental appendix contains three sections: Appendices A, B and C. Appendix A (A.1-A.9) contains clarifying or descriptive endnotes from Sections 1-3. Appendix B contains the detailed version of the proofs of the results in Sections 2 and 3 of our paper. Abridged versions of these proofs were presented in our paper. Appendix C (C.1-C.7) provides formal statements and their proofs for the asymptotic properties of the efficient estimator in Section 4. This presentation allows for overidentified models. Appendix C also reports simulation results describing the finite-sample properties of the efficient estimator in the context of the Monte Carlo experiment in Section 5.

Additionally, Appendix C describes a simple one-step updating of any $\sqrt{n}$-consistent estimator (e.g., IPW estimator) to obtain an estimator that is asymptotically equivalent to the efficient estimator. A sketch of the proof for this efficiency is provided under standard regularity conditions. This updating is computationally convenient and can be easily performed in standard statistical softwares. We provide two illustrations of the efficient estimator: (i) a linear regression as in Section 5 where a closed form efficient estimator is available (so, no updating is required), and (ii) a linear quantile regression where the updating is useful due to the unavailability of closed form expressions.

Index:

- Appendix A: Descriptive endnotes: pp. 2-13

  A.1 Planned incomplete design: examples from economics and other fields: pp. 2-5

  A.2 Planned incomplete design: examples of optimality of the design: pp. 5-7

  A.3 The equivalence relation in the MAR condition in (1): pp. 7-8

  A.4 The equivalence relation in the planned incompleteness condition in (2): pp. 8-9

  A.5 Intermediate steps in equation (4): p. 10

  A.6 Relation of the framework in Section 2 with closely related technical papers: pp. 10-11

  A.7 Intermediate steps for Remark 1 following Proposition 1: pp. 11-12

  A.8 Proposition 2's connection with the calibration and econometrics literature: pp. 12-13

[15] Department of Economics, McGill University, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

A.9 The importance of the planned incompleteness condition (2) in Proposition 2: p. 13

- Appendix B: Detailed version of the proofs of the results in Section 2 and 3: pp. 13-22
- Appendix C: GMM estimation of β^0_λ defined in (3): pp. 22-39

C.1 This GMM estimation is a special case of Ai and Chen (2012): pp. 22-23

C.2 Estimation framework and the key feature: pp. 24-25

C.3 Asymptotic properties of the GMM estimator in (42): pp. 25-27

C.4 One step from the IPW estimator gives efficiency: pp. 27-28

C.5 Illustration of the GMM estimator when $R = 3$: pp. 28-29

C.6 Simulation evidence from Section 5 of the finite-sample properties of $\widehat{\beta}_\lambda$: pp. 29-34

C.7 Proofs: pp. 35-39

- References: pp. 39-43

Appendix A: Descriptive endnotes

A.1 Planned incomplete design: examples from economics and other fields

Examples from other fields

The adoption of the planned incomplete survey design is common in other fields to the extent that there are even established terminologies to refer to the different types of planned incompleteness.

The two/many-measurement-design is used in psychology where it is common to encounter an expensive “gold standard” measure and other inexpensive but less accurate measures for behavioral traits [see, e.g., Graham et al. (2006)]. Then, the gold standard measure is typically employed only on a subset of the study subjects while the other measures are employed on all. In other contexts, planned missing waves for pre-selected sample units in a panel have been extensively used since MacArdle and Woodcock (1997) to cut the cost of estimation of key quantities in psychology.[16] In yet other contexts, the multiple matrix sampling of Shoemaker (1973), that requires most units to respond only to parts of the full survey questionnaire, was extended as the split-questionnaire design (SQD) by Raghunathan and Grizzle (1995) in statistics, as the partial questionnaire design (PQD) by Wacholder et al. (1994) in biostatistics and epidemiology, and as the multi-forms surveys discussed by Graham et al. (1996), Graham et al. (2006), and others in psychology and behavioral research.

Examples from economics

[16] While this example may appear less familiar than the other two types of examples, note that the structure of the sample due to missing waves is actually similar to that from rotating panels with a single rotation. Rotating panels such as the Current Population Survey are common in economics [see Nijman et al. (1991) for an influential study].

The common theme in all these references is the cost cutting of surveys, which also applies to the field of economics. This is even more relevant now as the use of primary data, often under tight budgets, gets more common among economists. However, in spite of the promising early work of DiNardo et al. (2006) who point to the benefits of planned incompleteness, systematic adoption of planned incompleteness seems nonexistent in economics. Ad hoc adoptions can be found in laboratory and field experiments, and we list below a small number of representative examples of both types.

(1) In a highly cited paper in experimental economics, Holt and Laury (2002) run a laboratory experiment to elicit risk aversion for studying its dependence on the size of the stake. The experiment involved planned incompleteness whereby the low-stake experiments were first run on all subjects (phase one) and then the high-stake experiments were run on subsets of these subjects.

(2) Field experiments also typically involve follow-up rounds. We provide three recent examples:

(2a) Thornton (2008) studies an experiment in rural Malawi where the subjects where tested for their HIV status and given incentives to learn the results from a nearby centre. After the respondents had a chance to learn about the result (some did not), a follow-up interview was conducted on 75% (so, 25% incompleteness by plan) of the original subjects to record their sexual behavior and their response to an offer to buy up to 5 packages of 3 condoms using the .30 USD that was paid to them.

(2b) Ashraf et al. (2010) run an experiment in Zambia to differentiate between the screening and sunk-cost effects measured by the usage of clorin (purchased from the experimenter) to purify drinking water. In the first phase (baseline), the experimenter measures, among other variables, the chemical concentration of clorin in the households' drinking water. In the second phase (marketing), the experimenter offers to sell a bottle of clorin to the concerned households at less than market price. In the third phase (follow-up), the experimenter again measures, among other things, the clorin concentration. The data are monotonic in terms of incompleteness — the third phase was conducted only on those households who could be reached in the second phase (planned incompleteness) and there was also high attrition, particularly, in the third phase (unplanned incompleteness).

(2c) Ashraf et al. (2014) run an experiment in Zambia to study household bargaining power in terms of eventual fertility and usage of contraceptives when women were given access to contraceptives in the presence and absence of their husbands. The first phase is a baseline survey on women that also provided them with information on contraception and prevention of STD, and distributed condoms. In the second phase (experiment) the respondents were reached either in the presence or absence of their husbands (reflecting two types of treatments) and vouchers for injectable contraceptives were provided. In the third phase (follow-up) information was collected on the women's use of contraceptives, sexual behavior, fertility, etc. Interestingly, beside a small number of rather balanced attrition (unplanned incompleteness), the monotonicity in this data resulted primarily from planned incompleteness because the second phase was conducted on a much smaller subset of the respondents from the first phase owing, in the authors' words, to "overwhelmingly...resource constraints on the part of the investigators and a strict timeline for completion of the study"/"Not enough budget".

Other types of planned incompleteness in economics

Another source of planned incompleteness (and eventual monotonicity) in Ashraf et al. (2014)'s data is the decision to collect new variables during the follow-up and an additional round but *only* in focus groups with subsets of participants. In other words, by design, both the original and the new variables are observed for a subset of units (those in focus groups) in the data, while only the original variables are observed for the remaining units in the data. Relatedly, there can be cases where such new variables might have less accurate counterparts in the original variables, making the former subset (in the last sentence) a validation sample. An example is Beaman et al. (2015) who use an input survey to obtain such data. An important consequence of this that we highlight in our paper is that the joint distribution of the more and less accurate variables that are jointly observed in the validation sample can often be useful for efficiency gains in subsequent estimation (although Beaman et al. (2015) did not need to exploit it). A similar example with more and less accurate measures of consumption, but unfortunately no joint observability (not needed for the stated purpose of their paper), is Beegle et al. (2012) [see our Section 5 for more on it]. This is also an example that does not involve a time dimension unlike the other references presented here.

Other types of cases where planned incompleteness could be useful include McKenzie (2012) and Allcott and Rogers (2014). Monotonicity is natural (at least, not unnatural) in both types of cases.

McKenzie (2012) draws on the clinical trial literature and provides an analysis of the benefit in precision gains from multiple follow-up measurements in field experiments over the standard practice of a single baseline and a single follow-up. His discussion focuses on the tradeoff in the choice of n (number of subjects) and T (number of measurements including baseline and follow-ups) at a given cost. Alternatively, one could keep both n and T large but measure the relevant variables only for a subset of subjects at each follow-up exactly like the prototypical multi-phase sampling.

Allcott and Rogers (2014) consider a treatment that was applied to subjects for varying duration. Specifically, the treatment was applied, i.e., a "home energy report" (containing personalized energy use, social comparisons, and energy conservation information) was sent to subjects, over a period of time but was discontinued (and not reinstated) for subsets of subjects during the tenure. The authors study the effect of this treatment on the energy consumption of the subjects. Note that, in such cases, the treatment administrator need not choose the subset of subjects "exogenously" but could conceivably incorporate the subjects' past responses to the treatment in the choice decision.

Relation with our framework

While the details of estimation vary, all the studies cited above involve estimating expectations and, sometimes, regression coefficients. For example, consider, without loss of generality, the instrumental variables (IV) regression in equation (2) (p. 1848) in Thornton (2008) (our Example (2a)) that was run on 75% of the full sample, namely, on the subjects from the districts of Rumphi and Balaka and not from Mchinji [see their Tables 6 and 7]. Assume in the spirit of Table 7 that the district-level heterogeneity is captured by the intercept, and extend this assumption to the full sample so that the regression continues to hold in the population of the full sample simply by adding a dummy D for Mchinji as a regressor. Denoting the instruments, endogenous regressors, exogenous regressors and dependent variable by W, X_1, X_2 and y respectively, define the (moment) function:

$$m(y, X_1, X_2, W; \beta_1, \beta_2) := (W', X_2')'(y - X_1\beta_1 - X_2\beta_2).$$

The planned incompleteness due to the selective follow-up here is a case of missing y. Now, while the coefficient of D (in X_2) is unidentified, the results in our paper imply that if interest lies in the population of all three districts then the optimal use of the full sample is possible using the modified moment vector: $\frac{(1-D)}{1-P(D=1)} m(y, X_1, X_2, W; \beta_1, \beta_2) + \left(1 - \frac{(1-D)}{1-P(D=1)}\right) E[m(y, X_1, X_2, W; \beta_1, \beta_2)|X_1, X_2, W]$ instead of $\frac{(1-D)}{1-P(D=1)} m(y, X_1, X_2, W; \beta_1, \beta_2)$ that is "close" to what was used in Table 7.[17,18] (Feasibility issues of the modified moment vector, which also arise in Example 1 below (Appendix A.2), are addressed in detail in the sequel and can be skipped for now in this introductory discussion.) Our paper explores such optimal uses of the sample for efficient estimation in more general contexts.

A.2 Planned incomplete design: examples of optimality of the design

Example 1: Minimizing variance of estimator subject to a given expected cost of survey

Let (Y, X) be scalar variables with finite means and variances. Let the parameter of interest be $\beta = E[Y - X]$. Consider two random samples $\mathcal{S}^\dagger = \{Y_j, X_j\}_{j=1}^{n^\dagger}$ and $\mathcal{S} = \{Y_i, D_i, D_i X_i\}_{i=1}^{n}$ where D is binary. We observe X in $\mathcal{S}$ only when $D = 1$. Assume that $P(D = 1|Y, X) = P(D = 1) = p$.[19]

[17] Standard IV conditions such as $E[m(y, X_1, X_2, W; \beta_1^0, \beta_2^0)|X_2, W] = 0$ or $E[m(y, X_1, X_2, W; \beta_1^0, \beta_2^0)] = 0$ do not imply that $E[m(y, X_1, X_2, W; \beta_1^0, \beta_2^0)|X_1, X_2, W] = 0$ where β_1^0 and β_2^0 are the true values of β_1 and β_2. Hence, the modification in the moment vector is not moot, and it reduces the variability of the estimating function for β_1 and β_2.

[18] We say "close" to mean asymptotically equivalent. Note that Tables 6 and 7 suggest that the first stage was run on the full sample since only y is missing, while the second stage was run on the sample where $D = 0$. While this gives more precise first stage estimates than what our latter representation above gives, under standard assumptions both approaches actually give asymptotically equivalent estimates of the parameters of interest β_1 and β_2 that, in turn, are less precise than what our former representation above with the modified moment vector does.

[19] 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 (the sub-sample containing all the variables required to estimate β) 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.

The standard and, in this case, efficient estimator of β based on $\mathcal{S}^\dagger$ is:

$$\widehat{\beta}^\dagger = \sum_{j=1}^{n^\dagger} (Y_j - X_j)/n^\dagger \text{ with } Var(\widehat{\beta}^\dagger) = \Delta/n^\dagger$$

where $\Delta := Var(Y - X)$. On the other hand, the result in this paper gives an infeasible version of the efficient estimator of β based on $\mathcal{S}$ as:

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

$\widehat{\beta}$ is infeasible because $E[X|Y]$ is unknown in practice. A feasible version of $\widehat{\beta}$ plugs in an estimator $\widehat{E}[X|Y]$ for $E[X|Y]$ in the expression for $\widehat{\beta}$. An important and desirable feature of our results that is repeatedly emphasized in Appendix C is that as long as $\widehat{E}[X|Y]$ is consistent for $E[X|Y]$ uniformly in Support(Y), plugging $\widehat{E}[X|Y]$ in the expression for $\widehat{\beta}$ only makes the result asymptotic, i.e., (i) what is referred to as $Var(\widehat{\beta})$ turns out to be $(1/n)$ times the asymptotic variance of the feasible $\widehat{\beta}$, and (ii) the feasible $\widehat{\beta}$ is no longer unbiased but is asymptotically unbiased and normally distributed.

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

Example 2: Variance reduction through dependent as opposed to independent sampling

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

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

and

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

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

A.3 The equivalence relation in the MAR condition in (1)

Lemma 11 *Let $P(C = r|T_R(Z)) > 0$ for each $r = 1, \ldots, R$. Then, $P(C = r|C \geq r, T_R(Z)) = P(C = r|C \geq r, T_r(Z))$ for $r = 1, \ldots, R$ if and only if $P(C = r|T_R(Z)) = P(C = r|T_r(Z))$ for $r = 1, \ldots, R$.*

Proof: We assume only $P(C = r|T_R(Z)) > 0$ for each $r = 1, \ldots, R$ for simplicity to avoid cases with 0/0. The proof follows by induction. We first show the "if" part and then the "only if" part.

"if:" Let $P(C = r|T_R(Z)) = P(C = r|T_r(Z))$ for $r = 1, \ldots, R$. Therefore, $P(C = 1|C \geq 1, T_R(Z)) \equiv P(C = 1|T_R(Z)) = P(C = 1|T_1(Z)) \equiv P(C = 1|C \geq 1, T_1(Z))$. Now, suppose that $P(C = j|C \geq j, T_R(Z)) = P(C = j|C \geq j, T_j(Z))$ for $j = 1, \ldots, r$ for some $r = 1, \ldots, R-1$. This will imply that $P(C = r+1|C \geq r+1, T_R(Z)) = P(C = r+1|C \geq r+1, T_{r+1}(Z))$ because:

$$\begin{aligned}
P(C = r+1|C \geq r+1, T_R(Z)) &= \frac{P(C = r+1|T_R(Z))}{P(C \geq r+1|T_R(Z))} \\
&= \frac{P(C = r+1|T_R(Z))}{1 - \sum_{j=1}^{r} P(C = j|T_R(Z))} \\
&= \frac{P(C = r+1|T_{r+1}(Z))}{1 - \sum_{j=1}^{r} P(C = j|T_j(Z))} \\
&= \frac{P(C = r+1|T_{r+1}(Z))}{1 - \sum_{j=1}^{r} P(C = j|T_{r+1}(Z))} \\
&= \frac{P(C = r+1|T_{r+1}(Z))}{P(C \geq r+1|T_{r+1}(Z))} \\
&= P(C = r+1|C \geq r+1, T_{r+1}(Z))
\end{aligned}$$

where the first, second, fifth and sixth equalities follow by definition, and the third and fourth equalities follow from the assumed conditions once we note that $T_j(Z)$ is nested by $T_{j+1}(Z)$ for all $j = 1, \ldots, R-1$.

"only if:" Let $P(C = r|C \geq r, T_R(Z)) = P(C = r|C \geq r, T_r(Z))$ for $r = 1, \ldots, R$. Therefore, $P(C = 1|T_R(Z)) \equiv P(C = 1|C \geq 1, T_R(Z)) = P(C = 1|C \geq 1, T_1(Z)) \equiv P(C = 1|T_1(Z))$. Now, suppose that $P(C = j|T_R(Z)) = P(C = j|T_j(Z))$ for $j = 1, \ldots, r$ for some $r = 1, \ldots, R-1$. This will imply that $P(C = r+1|T_R(Z)) = P(C = r+1|T_{r+1}(Z))$ because:

$$\begin{aligned}
P(C = r+1|T_R(Z)) &= P(C = r+1, C \geq r+1|T_R(Z)) \\
&= P(C = r+1|C \geq r+1, T_R(Z))P(C \geq r+1|T_R(Z)) \\
&= P(C = r+1|C \geq r+1, T_R(Z))\left(1 - \sum_{j=1}^{r} P(C = j|T_R(Z))\right) \\
&= P(C = r+1|C \geq r+1, T_{r+1}(Z))\left(1 - \sum_{j=1}^{r} P(C = j|T_j(Z))\right) \\
&= P(C = r+1|C \geq r+1, T_{r+1}(Z))P(C \geq r+1|T_r(Z)) \\
&= P(C = r+1|T_{r+1}(Z))
\end{aligned}$$

where the first three equalities follow by definition, the fourth equality follows by the assumed conditions, and the last two equalities are simply the reverse steps of the first three equalities coupled with the fact that $T_j(Z)$ is nested by $T_{j+1}(Z)$ for all $j = 1, \ldots, R-1$. ■

A.4 The equivalence relation in the planned incompleteness condition in (2)

Lemma 12 *Let* (1) *hold and also* $P(C = r|T_R(Z)) > 0$ *for each* $r = 1, \ldots, R$. *Then,* $P(C = r|C \geq r, T_r(Z))$ *is known for* $r = 1, \ldots, R$ *if and only if* $P(C = r|T_r(Z))$ *is known for* $r = 1, \ldots, R$.

Proof: The proof follows by induction exactly like the proof of Lemma 11. For the "if" part, when showing that the result holds for $r+1$ assuming that it holds for $j = 1, \ldots, r$, we have:

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

as before due to (1). The RHS is known by the assumed conditions. Hence the LHS is known.

For the "only if" part, when showing that the result holds for $r+1$ assuming that it holds for $j = 1, \ldots, r$, we have:

$$P(C = r+1|T_{r+1}(Z)) = P(C = r+1|C \geq r+1, T_{r+1}(Z))\left(1 - \sum_{j=1}^{r} P(C = j|T_j(Z))\right)$$

as before due to (1). The RHS is known by the assumed conditions. Hence the LHS is known. ■

Remark: At this stage, it is important to list two useful relations that are both related to the steps in the proofs of Lemmas 11 and 12, and also used repeatedly in the proofs in Appendices A and B.

Relation 1: (1) implies that

$$P(C \geq r|T_R(Z)) = P(C \geq r|T_{r-1}(Z)). \tag{31}$$

This follows by noting that:

$$\begin{aligned}
P(C \geq r|T_R(Z)) &= 1 - \sum_{j=1}^{r-1} P(C = j|T_R(Z)) \\
&= 1 - \sum_{j=1}^{r-1} P(C = j|T_j(Z)) \\
&= 1 - \sum_{j=1}^{r-1} P(C = j|T_{r-1}(Z)) \\
&= 1 - P(C \leq r-1|T_{r-1}(Z)) = P(C \geq r|T_{r-1}(Z))
\end{aligned}$$

where the first equality follows by definition, the second by (1), the third by (1) and the nested structure of $T_j(Z)$'s, while the fourth and the fifth by definition.

Note that, taking $R = 2$ in (31) implies that $P(C = 2|T_2(Z)) = P(C = 2|T_1(Z))$, the conventional MAR assumption found in the econometrics literature that has traditionally focused on $R = 2$ [see, e.g., Chen et al. (2005), Chen et al. (2008), Graham (2011), Graham et al. (2012)]. Looking at the complement events in (31) equivalently gives (31) as $P(C \leq r-1|T_R(Z)) = P(C \leq r-1|T_{r-1}(Z))$, which perhaps better indicates the generality of the selection on variables condition in our paper that can accommodate for all sorts of dimension reductions including the extreme reduction CMAR in (10) and the no reduction in Barnwell and Chaudhuri (2018).

Relation 2: For any function $\nu(Z)$ such that $E|\nu(Z)| < \infty$, (1) implies that:

$$\begin{aligned}
E\left[\frac{I(C \geq r)}{P(C \geq r|T_r(Z))}\nu(Z)\right] &= E\left[\frac{P(C \geq r|Z)}{P(C \geq r|T_r(Z))}\nu(Z)\right] \\
&= E\left[\frac{P(C \geq r|T_r(Z))}{P(C \geq r|T_r(Z))}\nu(Z)\right] = E[\nu(Z)]
\end{aligned} \tag{32}$$

where the first equality follows by the law of iterated expectations and the second one by (1). As a consequence of (31), one can instead write (32) as:

$$\begin{aligned}
E\left[\frac{I(C \geq r)}{P(C \geq r|T_{r-1}(Z))}\nu(Z)\right] &= E\left[\frac{P(C \geq r|T_r(Z))}{P(C \geq r|T_{r-1}(Z))}\nu(Z)\right] \\
&= E\left[\frac{P(C \geq r|T_{r-1}(Z))}{P(C \geq r|T_{r-1}(Z))}\nu(Z)\right] = E[\nu(Z)].
\end{aligned}$$

A.5 Intermediate steps in equation (4)

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

The first and third equalities follow by the law of iterated expectations, and the rest by definition.

Importantly, note that, the MAR condition in (1) and the planned incompleteness condition in (2) are not required for this relation in (4) to hold. However, as noted in the discussion around equations (1) and (2) that led to (4), the MAR condition in (1), in particular, is required to implement this relation in practice for the estimation of β by the IPW or the efficient estimator.

A.6 Relation of the framework in Section 2 with closely related technical papers

We delineate the framework in Section 2 from the following not-too-old representative examples under the non-Bayesian paradigm. **(a)** 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, $\lambda = \mathcal{C}$. **(b)** Robins and Rotnitzky (1995) and Holcroft et al. (1997) consider optimally using all the sub-samples under a framework similar to ours but with $\lambda = \mathcal{C}$. **(c)** 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. **(d)** 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., and is discussed in details in Chaudhuri and Guilkey (2016) (p. 686). **(e)** Dardanoni et al. (2011) consider a multiple regression framework with regressors missing non-monotonically under an assumption that implies that the regression coefficients do not vary across the populations of the sub-samples. So, they focus on $\lambda = \mathcal{C}$ and, unlike in our paper and the references cited in (a)-(d) and (f) (below), use of their complete sub-sample without correction for selection does not cause any bias in estimation.[20] Similarly, if one extends Abrevaya and Donald (2017) to the case of multiple incomplete sub-samples, then each sub-population would still be representative of $\lambda = \mathcal{C}$. **(f)** Finally, Chen et al. (2005) and Chen et al. (2008) consider frameworks where β^0_λ is defined exactly as in (3) for $R = 2$ and $\lambda = \{1\}$ (sub-population) and $\{1, 2\}$ (full population).

By contrast in one way or the other to (a)-(f), our setup: (i) allows for a general R, (ii) expands the scope to all $(2^R - 1)$ sub-populations (including $\lambda = \mathcal{C}$), (iii) introduces a dynamically updated sampling design via MAR, and (iv) provides the new insights available only from letting $R > 2$.

In this regard, it is also important to recall that the references in (d)-(f) above or the well-known sampling designs like the SQD, PQD, etc. noted in Appendix A.1 either do not consider or do not have the scope to consider a key feature of our framework, namely, sampling designs that are dynamically updated using the newly available information from more than one phase.

A.7 Intermediate steps for Remark 1 following Proposition 1

When $R = 2$ and $\lambda = \{1, 2\}$, (5) and (6) give:

$$
\begin{aligned}
\varphi_{\{1,2\}}(O;\beta) &= \frac{I(C=2)}{P(C=2|T_2(Z))} m(T_2(Z);\beta) \\
&\qquad + \left(\frac{I(C\geq 1)}{P(C\geq 1|T_1(Z))} - \frac{I(C=2)}{P(C=2|T_2(Z))} \right) E[m(T_2(Z);\beta)|T_1(Z)] \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))} m(T_2(Z);\beta) + \left(1 - \frac{I(C=2)}{P(C=2|T_1(Z))} \right) E[m(T_2(Z);\beta)|T_1(Z)] \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))} \left(m(T_2(Z);\beta) - E[m(T_2(Z);\beta)|T_1(Z)] \right) + E[m(T_2(Z);\beta)|T_1(Z)]
\end{aligned}
$$

where the second equality follows from (31). The last line is the expression from Chen et al. (2008).

[20] Bias arises due to problems with the imputed values if the same estimation is done in the incomplete sub-samples by replacing the missing regressors with their imputed values. To improve the precision of the unbiased estimator based on the complete sub-sample, they recommend Bayesian model averaging using the unbiased and biased estimates. While this approach should be very useful in many cases, it is a difficult proposition to compare it with the results in our paper and the other references here that all solve a different optimization problem: minimize asymptotic variance for *asymptotically unbiased* estimators. We thank a referee for pointing out this useful reference that we had missed earlier.

When $R = 2$ and $\lambda = \{1\}$, (5) and (6) give:

$$\begin{aligned}
\varphi_{\{1\}}(O;\beta) &= \frac{I(C=2)}{P(C=2|T_2(Z))}\frac{P(C=1|T_2(Z))}{P(C=1)}m(T_2(Z);\beta) \\
&\quad + \left(\frac{I(C\geq 1)}{P(C\geq 1|T_1(Z))} - \frac{I(C=2)}{P(C=2|T_2(Z))}\right) E\left[\frac{P(C=1|T_2(Z))}{P(C=1)}m(T_2(Z);\beta)\middle| T_1(Z)\right] \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}m(T_2(Z);\beta) \\
&\quad + \left(1 - \frac{I(C=2)}{P(C=2|T_1(Z))}\right) E\left[\frac{P(C=1|T_1(Z))}{P(C=1)}m(T_2(Z);\beta)\middle| T_1(Z)\right] \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}\left(m(T_2(Z);\beta) - E[m(T_2(Z);\beta)|T_1(Z)]\right) \\
&\quad + \frac{P(C=1|T_1(Z))}{P(C=1)}E[m(T_2(Z);\beta)|T_1(Z)]
\end{aligned}$$

where the second equality follows from (31) and (1). The RHS of the last equality is the expression from Chen et al. (2008).

A.8 Proposition 2's connection with the calibration and econometrics literature

The idea behind using the moment restrictions in (9) to augment the moment restriction (8), that already identifies β^0_λ and can be used to obtain a $\sqrt{n}$-consistent estimator [see, e.g., Wooldridge (2007)], and thus achieving efficiency gains is the same as the idea of calibration in the survey sampling literature [see, e.g., Deville and Sarndal (1992)]. The same idea, in more economics-centric ways, has appeared in the econometrics literature also: see Back and Brown (1993), Imbens and Lancaster (1994), Hellerstein and Imbens (1999), Devereux and Tripathi (2009), Tripathi (2011), Graham et al. (2012), etc. or Hellerstein and Imbens (1999), Nevo (2003), etc. in another context. To see the connection, first note that under our setup this means estimating β^0_λ by solving for β from $\sum_{i=1}^n \omega_i \varphi_{R,\lambda}(O_i,\beta) = 0$ where $\omega_i = I(C_i = R)/P(C = R|T_R(Z_i)) = \omega_{IPW,i}$, say, (instead of $1/n$ to reflect the non-representativeness of the complete sub-sample) if only (8) is used. On the other hand, if the calibration/augmenting/auxiliary restrictions in (9) are also utilized, then $\omega_i = \omega_{IPW,i} + \sum_{r=1}^{R-1} a_{r,i}$ for some appropriate (and complicated) set of random functions $a_{r,i}$'s. For example, if $R = 2$, then $a_{1,i} = \omega_{IPW,i}\Upsilon'_{K_1}(T_1(Z_i))(\sum_{j=1}^n \Upsilon_{K_1}(T_1(Z_j))\Upsilon'_{K_1}(T_1(Z_j)))^{-1}\sum_{l=1}^n(1 - \omega_{IPW,l})\Upsilon_{K_1}(T_1(Z_l))$ where $\Upsilon_{K_1}(T_1(Z))$ is a $K_1 \times 1$ vector of some possibly orthogonalized series of functions (e.g., power series, splines, etc.) of $T_1(Z)$ with possibly $K_1 \to \infty$ as $n \to \infty$ [see Graham et al. (2012)]. One could instead use $\bar{\omega}_i = \omega_i/\sum_j \omega_j$ as the weights so that they necessarily add up to one. However, there is no guarantee that $\bar{\omega}_i \in [0,1]$ for all i (indeed it can be outside $[0,1]$ for all i), which is not a desirable characteristic for weights. We do not pursue corrections for this undesirable characteristic of the weights since they are peripheral to the main message of our paper.

A.9 The importance of the planned incompleteness condition (2) in Proposition 2

This importance becomes evident when the target is *not* the full population. Consider $R = 2$ and $\lambda = \{1\}$, and note that Proposition 2 gives:

$$\begin{aligned}
\varphi_{\{1\}}(O;\beta) &= \overline{\text{Proj}}_{T_1}\left(\phi_{2,\lambda}(O;\beta)|\,\phi_1\right) = \phi_{2,\lambda}(O;\beta) - \text{Proj}_{T_1}\left(\phi_{2,\lambda}(O;\beta)|\,\phi_1\right) \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}m(Z;\beta) \\
&\quad - \left\{\frac{P(C=1|T_1(Z))}{P(C=1)P(C=2|T_1(Z))}E[m(Z;\beta)|T_1(Z)]\right\}(I(C=2)-P(C=2|T_1(Z))) \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}(m(Z;\beta)-E[m(Z;\beta)|T_1(Z)]) \\
&\quad + \frac{P(C=1|T_1(Z))}{P(C=1)}E[m(Z;\beta)|T_1(Z)].
\end{aligned}$$

On the other hand, it is known from Case 1 in Theorem 1 of Chen et al. (2008) (or plugging in $R = 2$ and $\lambda = \{1\}$ in our Proposition 5, or, equivalently, Barnwell and Chaudhuri (2018)'s Proposition 1) that the corresponding quantity without (2) would be:

$$\begin{aligned}
\varphi_{\{1\}[u]}(O;\beta) &= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}(m(Z;\beta)-E[m(Z;\beta)|T_1(Z)]) \\
&\quad + \frac{I(C=1)}{P(C=1)}E[m(Z;\beta)|T_1(Z)].
\end{aligned}$$

Of course, $\varphi_{\{1\}[u]}(O;\beta) \neq \varphi_{\{1\}}(O;\beta)$, i.e., Proposition 2 does not generally apply when targets are sub-populations unless the planned incompleteness condition in (2) holds.

Appendix B: Detailed version of the proofs of the results in Section 2 and 3

Appendix B provides the detailed version of the proofs of the results in Sections 2 and 3. Abridged version of the same proofs were presented in our paper.

The proofs of Propositions 1, 3, 4 and 5 involve obtaining the semiparametric efficiency bound and the efficient influence function, under different assumptions, following Chen et al. (2008). They follow in two steps. Step 1 characterizes the tangent set for all regular parametric sub-models satisfying the semiparametric assumptions on the observed data. Step 2 obtains the efficient influence function and, thereby, the asymptotic variance lower bound as the expectation of its outer product. f and F denote the density and distribution functions, with the concerned random variables specified inside parentheses. $L_0^2(F)$ denotes the space of mean-zero, square integrable functions with respect to F.

Proof of Proposition 1:

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

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

To reflect our condition (2), i.e., $P(C = r|Z_{(1)}, \ldots, Z_{(r)})$ is known for $r = 1, \ldots, R$ and hence need not be accounted for in what follows, we do not index them by θ. (These quantities do not play a role in the proof of the present proposition but does so in the proof of our Propositions 4 and 5.)

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 then 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 dimensional linear combinations of $S_\theta(O)$ for all such smooth parametric sub-models, and it takes the form:

$$\mathcal{T} := a_1(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) a_r(Z_{(1)}, \ldots, Z_{(r)}), \tag{33}$$

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

STEP 2: Differentiating the moment conditions in (3) 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 (3) and (1) that:

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

Therefore, assumption (A3) now gives:

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

Pathwise differentiability follows if we can find a $\psi(O) \in \mathcal{T}$ such that:

$$E[\psi(O)S(O)'] = \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'}. \tag{34}$$

Let us conjecture that $\psi(O) = -M_\lambda^{-1}\varphi_\lambda(O;\beta_\lambda^0)$, and then verify (34) 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 left hand side (LHS) and, in accordance with the partition of $\varphi_\lambda(O)$ (we work with the alternative specification in (7) for convenience), write it as $\sum_{q=1}^{R} B_q$ where, for $q = 2,\ldots,R$:

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

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

Now, note that:

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

Using MAR in (1) in the first equality of the last line below and the fact that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ for $r > 1$ in the last equality of the last line below, we obtain that:

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

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

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

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

Now, we consider B_q. (1) gives for $q = 2, \ldots, R$:

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

Since $E[\varphi_{q,\lambda}|T_{q-1}] = \varphi_{q-1,\lambda}$, it follows by conditioning on T_{q-1} and from (32) that the first term on the RHS is 0. On the other hand, (31) and the fact that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ implies that the second term is:

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

Therefore, $B_q = E[ms(Z_{(q)}|T_{q-1})'|C \in \lambda]$ for $q = 2, \ldots, R$, combining which with B_1 verifies (34).

That $\psi(O) \in \mathcal{T}$ follows from matching terms as follows. (i) $-M_\lambda^{-1}\varphi_{1,\lambda}$ is a function of only $T_1 := Z_{(1)}$, and $E[\varphi_{1,\lambda}] = 0$ and, hence, satisfies the properties of $a_1(Z_{(1)})$ in (33). (ii) The r-th term ($r = 2, \ldots, R$, without the multiplier $I(C \geq r)$) on the RHS of $\psi(O)$ can be written as: $-\frac{1}{P(C \geq r|T_r)}M_\lambda^{-1}[\varphi_{r,\lambda} - \varphi_{r-1,\lambda}] = -\frac{1}{1-P(C \leq r-1|T_{r-1})}M_\lambda^{-1}[\varphi_{r,\lambda} - \varphi_{r-1,\lambda}]$ by (1) [also see (31)]. 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 a function of only T_r and it 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 (33).

Therefore, we have now verified that the projection of the influence function $-M_\lambda^{-1}m(Z; \beta_\lambda^0)$ on to the tangent set $\mathcal{T}$ is $\psi(O) := -M_\lambda^{-1}\varphi_\lambda(O; \beta_\lambda^0)$. Hence, $\psi(O)$ is the efficient influence function and, therefore, the asymptotic variance lower bound is $E[\psi(O)\psi(O)'] = M_\lambda^{-1}V_\lambda M_\lambda^{-1'} =: \Omega_\lambda$. ■

Proof of Proposition 2:

(i) Let us start with $r = 1$, i.e., the residual from the projection, $\overline{\text{Proj}}_{T_{R-1}}(\phi_{R,\lambda}(\beta)|\,\phi_{R-1})$, 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 all. Then we apply induction arguments. For brevity, write $\varphi_{R,\lambda}(O; \beta)$ as $\varphi_{R,\lambda}$ and $T_r(Z)$ as T_r.

First, note that direct computation and (1) along with (31) give:

$$\text{Proj}_{T_{R-1}}(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}) = \left[\frac{I(C = R)}{P(C = R|T_R)} - \frac{I(C \geq R-1)}{P(C \geq R-1|T_{R-1})}\right] E[\varphi_{R,\lambda}|T_{r-1}],$$

which implies that:

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

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

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

Thus, similar computation as above (and the use of (31)) gives for $r=2$:

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

which implies that:

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

For our proof by induction, first assume that the following holds for a general $r\in\{2,\dots,R-2\}$:

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

Now, once again using (31), note that:

$$E[\phi_{R-r-1}^2|T_{R-r-1}] = \frac{P(C\geq R-r|T_{R-r})P(C=R-r-1|T_{R-r-1})}{P(C\geq R-r-1|T_{R-r-1})},$$

and

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

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

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

(ii) The proof follows in the same way as that of Theorem 1 in Chamberlain (1992) or, more generally, as that of Theorem 1 of Ai and Chen (2012). Appendix B.1 makes the connection with Ai and Chen (2012) explicit. ■

Proof of Proposition 3:

This proof follows in the same way as that of Proposition 1. The efficient influence function in this case turns out to be exactly the same as in Proposition 1 if CMAR is imposed on the latter. ■

We present the proofs of Propositions 4 and 5 in reverse order because the proof for the latter makes a reference to that for the former. Certain details of lesser importance are omitted below because they were already made explicit in the proof of Proposition 1.

Proof of Proposition 5:

STEP 1: Consider a regular parametric sub-model indexed by θ for the joint distribution of the observed data $O=(C,T_C'(Z))'$. Because of CMAR in (10), the log of the distribution can be expressed in terms of the full data $(C,Z')'$ as:

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

Let the true distribution be $f(O)=f_{\theta_0}(O)$ for some θ_0. Using the same notations as before, the score function with respect to θ can be written in terms of $(C,Z')'$ as:

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

where $\dot{P}_\theta(C=r|Z_{(1)}):=\frac{\partial}{\partial\theta}P_\theta(C=r|Z_{(1)})$. Thus, the tangent space is characterized by functions of the form:

$$\mathcal{T}:=a_1(Z_{(1)})+\sum_{r=2}^{R}I(C\geq r)a_r(Z_{(1)},\dots,Z_{(r)})+\sum_{r=1}^{R}I(C=r)\frac{b_r(Z_{(1)})}{bb_r(Z_{(1)})},\tag{35}$$

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

To avoid notational clutter, in the rest of the proof we write $m(Z;\beta_\lambda^0)$ as m; $T_r(Z)$ as T_r for

$r = 1, \dots, R$; and also write $s(Z_{(r)}|Z_{(1)}, \dots, Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r = 2, \dots, R$.

Unlike in Chen et al. (2008)'s proof we use the same factorization of the joint density of O for all λ. For a given $\lambda \in \Lambda$, the following relation obtained by two different factorization of the joint distribution of $(I(C \in \lambda), T_1(Z) \equiv Z_{(1)})$ helps us to switch between different factorizations:

$$\begin{aligned} & s(T_1) + I(C \in \lambda)\frac{\dot{P}(C \in \lambda|T_1)}{P(C \in \lambda|T_1)} + I(C \notin \lambda)\frac{\dot{P}(C \notin \lambda|T_1)}{P(C \notin \lambda|T_1)} \\ = \; & I(C \in \lambda)\left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda)\right] + I(C \notin \lambda)\left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(T_1|C \notin \lambda)\right]. \end{aligned} \tag{36}$$

STEP 2: Differentiating (3) with respect to θ under the integral:

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

Then, as in the proof of Proposition 1, here we will need to correspondingly verify that:

$$E[\varphi_{\lambda[u]}^{\text{CMAR}}(O; \beta_\lambda^0)S(O)'] = E\left[m\left\{s(T_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})'\right\}\middle| C \in \lambda\right]. \tag{37}$$

We do this term by term for $\varphi_{\lambda[u]}^{\text{CMAR}}(O; \beta_\lambda^0)$ and show equality of the terms on the LHS and RHS.

Consider the first term of $\varphi_{\lambda[u]}^{\text{CMAR}}(O; \beta_\lambda^0)$. Since $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ for $r = 2, \dots, R$ by definition, we can use (10) to take conditional expectations and then write:

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

where the second line follows by using (36) to replace $s(T_1)$. The last line follows since, by using (10), we obtain that:

$$\begin{aligned} E\left[I(C \in \lambda)\sum_{r=1}^{R} I(C = r)\frac{\dot{P}(C = r|T_1)}{P(C = r|T_1)}\middle| T_1\right] &= \sum_{r \in \lambda} P(C = r|T_1)\frac{\dot{P}(C = r|T_1)}{P(C = r|T_1)} \\ &= \sum_{r \in \lambda} \dot{P}(C = r|T_1) = \dot{P}(C \in \lambda|T_1). \end{aligned}$$

Hence, now by repeatedly using (10) (e.g., first term on RHS of second equality) we obtain that:

$$\begin{aligned}
E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]S(O)'\right] &= E\left[E[m|T_1]|C\in\lambda\right]\frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)} + E\left[E[m|T_1]s(T_1|C\in\lambda)'|C\in\lambda\right] \\
&\quad - E\left[E[m|T_1]\frac{\dot{P}(C\in\lambda|T_1)'}{P(C\in\lambda)}\right] + E\left[E[m|T_1]\frac{\dot{P}(C\in\lambda|T_1)'}{P(C\in\lambda)}\right] \\
&= E\left[m|C\in\lambda\right]\frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)} + E\left[E[m|T_1]s(T_1|C\in\lambda)'|C\in\lambda\right] + 0 \\
&= 0 + E[ms(T_1|C\in\lambda)'|C\in\lambda] + 0 \qquad (38)
\end{aligned}$$

where the first zero in last line follows from (3). The second term follows by using (10) and noting that $E\left[E[m|T_1]s(T_1|C\in\lambda)'|C\in\lambda\right] = E\left[E[ms(T_1|C\in\lambda)'|T_1, C\in\lambda]|C\in\lambda\right] = E\left[ms(T_1|C\in\lambda)'|C\in\lambda\right]$.

Now consider the r-th term of $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ for $r = 2,\ldots,R$. By taking expectation conditional on $T_{r-1} \equiv (Z_{(1)},\ldots,Z_{(r-1)})$, and using (10) we obtain that:

$$\begin{aligned}
& E\left[\frac{P(C\in\lambda|T_1)}{P(C\in\lambda)}(E[m|T_r] - E[m|T_{r-1}])S(O)'\right] \\
= \;& E\left[\frac{P(C\in\lambda|Z_1)}{P(C\in\lambda)}(E[m|T_r] - E[m|T_{r-1}])\sum_{s=r}^{R} s(Z_{(s)}|T_{s-1})\right] \\
= \;& E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_r]s(Z_{(r)}|T_{r-1})'\right] \\
= \;& E\left[ms(Z_{(r)}|T_{r-1})'|C\in\lambda\right] \qquad (39)
\end{aligned}$$

by using that $s(Z_{(s)}|T_{s-1}) \in L^2_0(F(Z_{(s)}|T_{s-1}))$ for $s = r,\ldots,R$ by definition, and by (10).

Therefore, (38) and (39) verify (37). That $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ belongs to $\mathcal{T}$ in (35) can be shown as follows. (i) Match the term $a(Z_{(1)},\ldots,Z_{(r)})$ in $\mathcal{T}$ with the r-th term of $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ for $r > 1$. (ii) Distribute the first term $s(Z_{(1)})$ in $\mathcal{T}$ according to the relation (36) and match the term $I(C \in \lambda)s(Z_{(1)}|C \in \lambda)$ with the first term of $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ while keeping in mind that, by definition, $s(Z_{(1)}|C \in \lambda) \in L^2_0(F(Z_{(1)}|C \in \lambda))$. It is straightforward to verify that all the corresponding conditional expectations, as required by the definition in (35) and also (36), are zeros. The remaining terms in $\mathcal{T}$ (including the one due to distributing the terms in (ii)) are represented in $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ by zeros. ■

Proof of Proposition 4:

The references in the steps of this proof are mainly to that of Proposition 3 (i.e., effectively to that of Proposition 1) and to that of Proposition 5. To avoid notational clutter, when convenient, we write $m(Z;\beta^0_\lambda)$ as m; $T_r(Z)$ as T_r for $r = 1,\ldots,R$; and also write $s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r = 2,\ldots,R$.

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

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

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

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

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

Recognizing that $P(C = r|T_1) = P(C = r|T_1;\gamma^0)$ is known up to the finite (d_γ) dimensional parameter γ, alters the relationship in (36) as follows:

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

As before, differentiating (3) (equivalently, (4)) under the integral with respect to θ, and using the above relationship gives:

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

Therefore, utilizing the expression of the efficient influence function in Proposition 3 and its relation to that in Proposition 4, the verification of pathwise differentiability reduces to verifying that:

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

Note that $E\left[S_\gamma(C|T_1) \left\{ s(T_1)' + \sum_{r=2}^{R} I(C \geq r) s(Z_{(r)}|T_{r-1})' \right\} \right] = 0$. This follows (term by term) by using $E[S_\gamma(C|T_1)|T_1] = 0$ for term one; and then, for the other terms $r = 2, \ldots, R$, by noting that (10) implies that $E\left[S_\gamma(C|T_1) I(C \geq r) s(Z_{(r)}|T_{r-1})'\right] = E\left[S_\gamma(C|T_1)(1 - I(C \leq r-1)) s(Z_{(r)}|T_{r-1})'\right] = E\left[S_\gamma(C|T_1)(1 - P(C \leq r-1|T_1)) s(Z_{(r)}|T_{r-1})'\right] = 0$ since $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1})$.

Therefore, using the expression for $S(O)$, it follows that in the above equation (that contains the equality relationship to be verified), the LHS simplifies as:

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

Proofs of Corollary 6, 7, 8: Straightforward but tedious manipulations of the results of Propositions 3 and 5 give Corollaries 7 and 8 respectively [see Chaudhuri (2014) for the proof of the latter]. Corollary 6 follows by imposing INDEP on the result of either Proposition 3 or Proposition 5. ■

Appendix C: Generalized method of moments (GMM) estimation of β^0_λ

Sections C.1, C.4 and part of C.5 in Appendix C collect materials related to efficient estimation that were not presented in our paper. On the other hand, the materials in Sections C.2 and C.3 were presented in our paper but again presented here to make Appendix C self-contained. The proofs of all the results that were presented in abridged form in our paper are presented here in detail.

C.1 This GMM estimation is a special case of Ai and Chen (2012)

Recall that Proposition 2 shows that under (1), (2) and assumption A, the efficient influence function and the efficiency bound for the estimation of β^0_λ based on (3) are identical to those based on the sequential moment restrictions (8)-(9). Hence, one could perform the efficient GMM estimation of β^0_λ simply as a special case of the optimally weighted orthogonalized sieve minimum distance (SMD) estimator that was proposed by Ai and Chen (2012) in a more general context.

To see the connection with Ai and Chen (2012) more clearly, note that our unconditional moment restriction in (8) corresponds to equation (1) in Ai and Chen (2012) with their conditioning variable $X^{(1)}$ taken as a constant. Now, the simplifications for our setup follow because, unlike Ai and Chen (2012), we do not have any unknown nuisance parameters (thanks to (2)) and because in our setup β_λ only enters the unconditional moment restrictions. That is, in our setup the moment restrictions in (9) turn out to be truly auxiliary whose sole purpose is to assist in obtaining efficiency gains.

This results in equation (10) of Ai and Chen (2012) (using their notation) to become:

$$\begin{aligned} \alpha_0 &:= \inf_{\alpha\in\Theta} E\left\{m_1(X^{(1)},\alpha)'\Sigma_{01}(X^{(1)})^{-1}m_1(X^{(1)},\alpha)\right\}, \qquad (40)\\ \text{where} \quad m_1(X^{(1)},\alpha) &:= E\left[\varepsilon_1(Z,\alpha)|X^{(1)}\right] = E\left[\varepsilon_1(Z,\alpha)\right],\\ \Sigma_{01}(X^{(1)}) &:= E\left[\varepsilon_1(Z,\alpha_0)\varepsilon_1(Z,\alpha_0)'|X^{(1)}\right] = E\left[\varepsilon_1(Z,\alpha_0)\varepsilon_1(Z,\alpha_0)'\right], \end{aligned}$$

i.e.,

$$\alpha_0 := \inf_{\alpha\in\Theta} E\left[\varepsilon_1(Z,\alpha)'\right]\left(E\left[\varepsilon_1(Z,\alpha_0)\varepsilon_1(Z,\alpha_0)'|X^{(1)}\right]\right)^{-1} E\left[\varepsilon_1(Z,\alpha)\right].$$

Now, note that $\varepsilon_1(Z,\alpha)$ is Ai and Chen (2012)'s sequentially orthogonalized moment vector, i.e.,

$$\varepsilon_1(Z,\alpha) := \rho_1(Z;\alpha) - \sum_{t=2}^{T}\Gamma_{1,t}(X^{(t)})\varepsilon_t(Z,\alpha)$$

where $\varepsilon_T(Z;\alpha) := \rho_T(Z,\alpha)$ and for $t = 2,\ldots,T-1$, $\varepsilon_t(Z,\alpha)$ are the orthogonalized residuals:

$$\begin{aligned} \varepsilon_t(Z,\alpha) &:= \rho_t(Z;\alpha) - \sum_{s=t+1}^{T}\Gamma_{t,s}(X^{(s)})\varepsilon_s(Z,\alpha),\\ \text{where} \quad \Gamma_{t,s}(X^{(s)}) &:= E\left[\rho_t(Z;\alpha_0)\varepsilon_s(Z;\alpha_0)'|X^{(s)}\right]\left(E\left[\varepsilon_s(Z;\alpha_0)\varepsilon_s(Z;\alpha_0)'|X^{(s)}\right]\right)^{-1}. \end{aligned}$$

Therefore, thanks to our Proposition 2, $\varepsilon_1(Z,\alpha)$ and $\Sigma_{01}(X^{(1)})$ in Ai and Chen (2012) are our $\varphi_\lambda(O;\beta)$ and $V_\lambda := Var(\varphi_\lambda(O;\beta^0))$ respectively. Accordingly, the optimally weighted orthogonalized SMD estimator in equation (11) of Ai and Chen (2012), that is based on the sample counterpart of (40), is identical to the GMM estimator that uses the average estimated $\varphi_\lambda(O;\beta)$ as the moment vector and an estimator of V_λ^{-1} as the weighting matrix. We say "estimated $\varphi_\lambda(O;\beta)$" because, as is clear from the definition of $\varepsilon_1(Z,\alpha)$ entering $m_1(X^{(1)},\alpha) := E\left[\varepsilon_1(Z,\alpha)\right]$, this contains unknown conditional expectations (covariance and variances) as nuisance parameters that need to be estimated and, thereby, profiled out from the criterion function of the estimation of the parameter of interest.

The purpose of Section C.2 and C.3 below is to point out with some details that under this special case of Ai and Chen (2012) that is our setup, a key feature of $\varphi_\lambda(O;\beta)$ provides practically useful flexibility in the parametric or nonparametric estimation of these nuisance parameters.

This key feature is of independent interest even without any consideration of efficiency. Hence, for completeness, we will work under the setup of an over-identified model where β is $d_\beta \times 1$, $m(Z;\beta)$ is $d_m \times 1$ and $d_m \geq d_\beta$. However, it is important to remember that the results on efficiency bounds that were presented in our paper imposed the restriction that $d_\beta = d_m = d$ [see footnote 5]. In light of this, our discussion when $d_m > d_\beta$ can be viewed simply as highlighting the specialities of GMM estimation based on a special moment vector like $\varphi_\lambda(O;\beta)$. To avoid introducing new notation when $d_m > d_\beta$, we continue with the notation from our paper. It is easy to see how the expressions thus obtained below will simplify when $d_m = d_\beta$ to give the exact expressions presented in our paper.

C.2 Estimation framework and the key feature

To consolidate notation following Chen et al. (2003), and guided by (6), define a $d_m \times 1$ function:

$$g(O;\beta,h(\beta)) := \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta)+\sum_{r=1}^{R-1}\left[\frac{I(C\geq r)}{P(C\geq r|T_r(Z))}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1}(Z))}\right]h_r(\beta) \tag{41}$$

where $h(\beta) = (h_1'(\beta), \ldots h_{R-1}'(\beta))'$ are the unknown nuisance parameters, and each $h_r(\beta)$ belongs to a class of functions $(Z,\beta) \mapsto \mathbb{R}^{d_m}$, call it $\mathcal{H}_r(\beta)$, for $r = 1,\ldots,R-1$. Let $\mathcal{H} := \{\mathcal{H}_1(\beta)\times\ldots\times\mathcal{H}_{R-1}(\beta) : \beta \in \mathcal{B}\}$ be 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 the other arguments.

$g(O;\beta,h(\beta)) = \varphi_\lambda(O;\beta)$ defined in (6) if $h_r(\beta) = \varphi_{r,\lambda}(O;\beta)$ for $r = 1,\ldots,R-1$. Denote the true $h_r(\beta)$ as $h_r^0(\beta) := \varphi_{r,\lambda}(O;\beta)$ for $r = 1,\ldots,R-1$. While this suggests restricting $h_r(\beta)$ as $(T_r(Z),\beta) \mapsto \mathbb{R}^{d_m}$ for $r = 1,\ldots,R-1$, it turns out that letting $h_r(\beta)$ instead be a function of Z and β does not affect either consistency or asymptotic normality of the GMM estimator defined below.

In light of this discussion, now define the GMM average moment vector and its expectation as:

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

Then, given any standard parametric or nonparametric estimator $\widehat{h}(\beta)$ for $h(\beta)$ and any $d_m \times d_m$ symmetric weighting matrix W_n (possibly efficient), the GMM estimator $\widehat{\beta}_\lambda(W_n)$ of β_λ^0 is defined as:

$$\widehat{\beta}_\lambda(W_n) \approx \arg\min_{\beta\in\mathcal{B}} G_n(\beta,\widehat{h}(\beta))'W_n G_n(\beta,\widehat{h}(\beta)). \tag{42}$$

The key feature of our setup is the identity that for any $\beta \in \mathcal{B}$ and any $h(.) \in \mathcal{H}$ (that need not be $h(\beta)$):

$$G(\beta,h(.)) = E[\varphi_{R,\lambda}(O;\beta)] = E[m(Z;\beta)|C \in \lambda] \tag{43}$$

by (4), (1) and (41). That is, $G(\beta,h(.))$ does not depend on $h(.) \in \mathcal{H}$. Its main implications are:

(F1) $G(\beta_\lambda^0,h(.)) = 0$ for any $h(.) \in \mathcal{H}$ by also using (3). Also, for any $\beta \in \mathcal{B}$ and any $h(.),\bar{h}(.) \in \mathcal{H}$:

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

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

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

(F1) helps to verify the well-separability (of the true β) assumption for consistent estimation of β^0_λ by $\widehat{\beta}_\lambda(W_n)$. It is even stronger since it indicates that $\widehat{h}(\beta)$ need not converge in probability to the true $h^0(\beta)$ but can converge to any $h^\dagger(\beta) \in \text{interior}(\mathcal{H})$ without affecting the consistency of $\widehat{\beta}_\lambda(W_n)$ for β^0_λ [see Proposition 13]. (F2) simplifies the Jacobian formula (and its estimation) in the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$ since it implies that $G_\beta(\beta^0_\lambda, h(\beta^0_\lambda)) = M_\lambda$. Finally, while it was already clear from (F1) that the asymptotic orthogonality condition, Assumption N(c), in Andrews (1994) is satisfied following his equations (4.9)-(4.11) if $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ for any $h^\dagger(\beta) \in \text{interior}(\mathcal{H})$; (F3) is still stated in a way that makes it more convenient for us to verify condition (4.1.4) in Theorem 4.1 of Chen (2007). (Proofs of the results stated below proceed by verifying the conditions in Chen et al. (2003) or Chen (2007).) Hence, the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$ is unaffected by the estimation of $h(\beta)$ even if $\widehat{h}(\beta)$ converges at a rate slower than $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(n^{-1/4})$; for example, $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ will suffice. See Remark 2(iii) in Chen et al. (2003) and Theorem 5 in Cattaneo (2010). The scenario is actually stronger here since we do not even require that $h^\dagger(\beta) = h^0(\beta)$, the truth [see Proposition 14]. Of course, semiparametric efficiency for $\widehat{\beta}_\lambda(W_n)$ requires that $h^\dagger(\beta^0_\lambda) = h^0(\beta^0_\lambda)$, but the rate of convergence of the consistent $\widehat{h}(\beta)$ is still of no consequence as far as the first-order asymptotic properties of GMM estimators are concerned [see Corollary 15]. Naturally, all these nice implications of (43) also provide flexibility in estimating the nuisance parameters – (i) parametrically based on misspecified models, e.g., giving linear projections rather than conditional expectations or (ii) nonparametrically under less than satisfactory conditions that might prevent a faster than $n^{1/4}$-rate convergence of the estimator.

C.3 Asymptotic properties of the GMM estimator in (42)

For simplicity we follow Chen et al. (2003) and write $(\beta, h(\beta))$ as (β, h) unless confusing. Also, define $\|A\|_B := \sqrt{\text{trace}(A'BA)}$ for conformable matrices A and B. Write $\|A\| \equiv \|A\|_B$ if B is identity.

Proposition 13 *Let* (1), (3), *and assumptions (A1) and (A2) hold. Let* $\{W_n\}$ *be a* $d_m \times d_m$ *positive semidefinite matrix such that* $W_n = W + o_p(1)$ *where* W *is a constant positive definite matrix.*

Assume:

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

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

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

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

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

Proposition 14 *Let* (1), (3) *and assumptions A hold. Let* $\{W_n\}$ *be a* $d_m\times d_m$ *positive semidefinite matrix such that* $W_n = W + o_p(1)$ *where* W *is a constant positive definite matrix. Let* $\beta^0_\lambda\in$ *interior*$(\mathcal{B})$ *and* $h^\dagger(\beta)\in$ *interior*$(\mathcal{H})$ *for all* β*, but* $h^\dagger(\beta)$ *not necessarily equal to* $h^0(\beta)$*. For a small* $\delta>0$ *define the neighborhoods* $\mathcal{B}_\delta := \{\beta\in\mathcal{B}:\|\beta-\beta^0_\lambda\|\leq\delta\}$ *and* $\mathcal{H}_\delta := \{h\in\mathcal{H}:\|h-h^\dagger(\beta)\|_{\mathcal{H}}\leq\delta\}$*. (Nothing changes if the sup-norm with respect to* β *in* $\|.\|_{\mathcal{H}}$ *is alternatively defined to be taken locally over* $\beta\in\mathcal{B}_\delta$ *instead* $\beta\in\mathcal{B}$*; see Chen et al. (2003).) Let* $\widehat{\beta}_\lambda(W_n)-\beta^0_\lambda = o_p(1)$ *and* $\|\widehat{h}(\beta)-h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$*. Assume:*

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

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

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

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

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

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 13 and 14 respectively establish the consistency and asymptotic normality of the GMM estimator defined in (42). We focus on showing how the key feature (43) helps to satisfy some of the conditions from Theorem 1 in Chen et al. (2003) and Theorem 4.1 in Chen (2007). We assume their other conditions. Through its condition (4.1.4), as opposed to (4.1.4)', Theorem 4.1 in Chen (2007) broadens the scope of Theorem 2 in Chen et al. (2003). This is useful to highlight that Propositions 13 and 14 (and the subsequent results) do not depend on the rate of convergence $\|\widehat{h}(\beta) - h^{\dagger}(\beta)\|_{\mathcal{H}} = o_p(1)$. Importantly, we allow $h^{\dagger}(\beta) \neq h^0(\beta)$ to emphasize that consistency and asymptotic unbiasedness of $\widehat{\beta}_{\lambda}(W_n)$ are robust to the estimation of the nuisance parameters $h(\beta)$ parametrically under misspecification or nonparametrically under less than satisfactory conditions.

Thus, the theoretical results confirm the intuitions from our discussion of the implications of the key feature, except for the final bit, i.e., on efficiency, that is to be confirmed by the following result.

Corollary 15 *Under the assumptions of Proposition 14:*

(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 in Proposition 1, and letting* $\widehat{\beta}_{\lambda} := \widehat{\beta}_{\lambda}(W_n)$,

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

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

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

Standard conditions, e.g., $g(O_i; (\beta, h))$ is continuous with probability approaching one in a neighborhood $\mathcal{N}$ of $(\beta_{\lambda}^0, h^{\dagger})$ and $E\left[\sup_{(\beta,h)\in\mathcal{N}} \|g(O_i; (\beta, h))\|^2\right] < \infty$ [see Lemma 4.3 in Newey and McFadden (1994)], ensure that for any $\beta = \beta_{\lambda}^0 + o_p(1)$ and $h(\beta)$ such that $\|h(\beta) - h^{\dagger}(\beta)\|_{\mathcal{H}} = o_p(1)$ (suffices if the sup-norm in $\|.\|_{\mathcal{H}}$ with respect to β is only local), the estimator $\widehat{V}_{\lambda}(\beta, h) := \frac{1}{n}\sum_{i=1}^{n} g(O_i; (\beta, h))g(O_i; (\beta, h))' = \Sigma + o_p(1)$. Thus, the estimator $\widehat{\Omega}_{\lambda}(\widehat{\beta}_{\lambda}, \widehat{h}) := \left(\widehat{M_{\lambda}'}(\widehat{\beta}_{\lambda})\widehat{V}_{\lambda}^{-1}(\widehat{\beta}_{\lambda}, \widehat{h})\widehat{M_{\lambda}}(\widehat{\beta}_{\lambda})\right)^{-1}$ is consistent for the asymptotic variance in Corollary 15(1). If $h^{\dagger}(\beta_{\lambda}^0) = h^0(\beta_{\lambda}^0)$ then $\Sigma = V_{\lambda}$, and now $\widehat{\Omega}_{\lambda}(\widehat{\beta}_{\lambda}, \widehat{h})$ will be consistent for the asymptotic variance Ω_{λ} in Corollary 15(2). Any consistent (for the appropriate limit) estimator $(\widetilde{\beta}, \widetilde{h})$ ensures consistency of all these quantities.

C.4 One step from the IPW estimator gives asymptotic equivalence with $\widehat{\beta}_{\lambda}$

The presence of β in possibly highly nonlinear form in all the R additive terms of the average moment vector $G_n(\beta, \widehat{h}(\beta))$ should not ideally be a drawback for computational purpose. If the GMM estimator has a closed form (e.g., Illustration 1 below) then this is not an issue. However, if there is no closed form expression (e.g., Illustration 2 below), one could start with an easy to compute $\sqrt{n}$-consistent estimator for β_λ^0 and then update it in one step to obtain an estimator with the same asymptotic distribution as the estimator $\widehat{\beta}_\lambda$ in Corollary 15(2). For example, an IPW estimator based on the complete sub-sample $\{i = 1, \ldots, n : C_i = R\}$ and with the identity (or some simple) weighting matrix is relatively easy to compute:

$$\begin{aligned}\widetilde{\beta}_\lambda &:= \arg\min_{\beta \in \mathcal{B}} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{I(C_i = R)}{P(C = R | T_R(Z_i))} \varphi_{R,\lambda}(O_i; \beta) \right\| \\ &\equiv \arg\min_{\beta \in \mathcal{B}} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{I(C_i = R)}{P(C = R | Z_i)} \frac{P(C \in \lambda | Z_i)}{\widehat{P}(C \in \lambda)} m(Z_i; \beta) \right\|. \qquad (44)\end{aligned}$$

It is consistent under the assumptions of Proposition 13 [see, e.g., Wooldridge (2002)]. Built-in routines in standard statistical softwares can be directly used or slightly modified to obtain this estimator for a wide variety of the moment vector $m(Z; \beta)$ (e.g., Illustration 2 below). Now a one-step estimator of β_λ^0 can be obtained by updating $\widetilde{\beta}_\lambda$ as:

$$\widehat{\beta}_{\text{1step}} = \widetilde{\beta}_\lambda - \widehat{\Omega}_\lambda^{-1}(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda)) \widehat{M}_\lambda'(\widetilde{\beta}_\lambda) \widehat{V}_\lambda^{-1}(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda)) G_n(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda)) \qquad (45)$$

where $\widehat{h}(\widetilde{\beta}_\lambda)$ is a consistent estimator of $h^0(\beta_\lambda^0)$, and $\widehat{M}_\lambda(\widetilde{\beta}_\lambda)$, $\widehat{V}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda))$ and $\widehat{\Omega}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda))$, defined below Corollary 15, are consistent estimators for M_λ, V_λ and Ω_λ respectively under the conditions noted therein.[21]

Proposition 16 *Let all the conditions of Corollary 15(2) hold for $\widehat{\beta}_\lambda$, i.e., for the GMM estimator with the efficient weighting matrix. Additionally, let there be a first step estimator $\widetilde{\beta}_\lambda$ satisfying: $\sqrt{n}(\widetilde{\beta}_\lambda - \beta_\lambda^0) = O_p(1)$, $\widehat{M}_\lambda(\widetilde{\beta}_\lambda) = M_\lambda + o_p(1)$, $\widehat{V}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda)) = V_\lambda + o_p(1)$ and $\widehat{\Omega}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda)) = \Omega_\lambda + o_p(1)$. For simplicity, assume a slightly stronger version of the stochastic equicontinuity condition (C3) [see Proposition 14] as: $\sup_{\beta \in \mathcal{B}_{\delta_n}, h \in \mathcal{H}_{\delta_n}} \sqrt{n} \| G_n(\beta, h) - G(\beta, h) - G_n(\beta_\lambda^0, h^0) \| = o_p(1)$. Then, $\widehat{\beta}_{1step}$ defined in (45) satisfies: $\sqrt{n}\left(\widehat{\beta}_{1step} - \widehat{\beta}_\lambda\right) = o_p(1)$.*

C.5 Illustration of the GMM estimator when $R = 3$

To focus on the main components, we abstract from the weighting matrix W_n by taking $d_m = d_\beta$. We consider two cases where the moment vector respectively corresponds to: (1) a linear regression giving a closed form expression for the efficient estimator, and (2) a linear quantile regression where

[21] While the one-step estimator in (45) could be easily modified to allow for the possibility that $\widehat{h}(\widetilde{\beta}_\lambda)$ converges in probability to $h^\dagger(\beta_\lambda^0)$ instead of the truth $h^0(\beta_\lambda^0)$, we do not consider this here since, as evident from Corollary 15, semiparametric efficiency is usually not achieved unless $h^\dagger(\beta_\lambda^0) = h^0(\beta_\lambda^0)$.

the efficient estimator is computed in one step as in (45). As for a concrete scenario with $R = 3$, it may be useful to keep in mind the setup of our Monte Carlo experiment in Section 5.

Illustration 1: Linear regression in the target population λ

Consider a moment vector of the form $m(Z;\beta) = X(y - X'\beta)$. For $i = 1, \ldots, n$, let $T_{ji} := T_j(Z_i)$ for $j = 1, 2, 3$, $a_{3i} := I(C_i = 3)/P(C = 3|T_{3i})$, $a_{2i} := I(C_i \geq 2)/P(C \geq 2|T_{2i}) - a_{3i}$, $a_{1i} := 1 - a_{2i} - a_{3i}$, $q := P(C \in \lambda|T_3(Z))$ and $q_i := P(C \in \lambda|T_{3i})$. Simple computations give a closed form expression for the estimator $\widehat{\beta}_\lambda$ in (42) as:

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

where $\widehat{E}$ denotes the estimated conditional expectation. While one could factor out y_i from all three terms inside the last pair of braces under the setup of Section 5 (where y is always observed), our experience is that estimating the conditional expectations, 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 $\widehat{\beta}_\lambda$ in small samples.

Illustration 2: Linear quantile regression in the target population λ

Consider a moment vector of the form $m(Z;\beta) = X\left(\tau - I(y - X'\beta < 0)\right)$ for some fixed $\tau \in (0, 1)$. (The notation $a_{3i}, a_{2i}, a_{1i}, q_i$ and q remain the same as in Illustration 1.) For any (β, h) define:

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

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. (The ignored common denominator $P(C \in \lambda)$ will be adjusted for in the final step.) Let $\widetilde{\beta}_\lambda$ denote the inefficient but $\sqrt{n}$-consistent estimator of β^0_λ obtained from (44) by using this particular choice of the moment vector $m(Z;\beta)$. It is simple to obtain $\widetilde{\beta}_\lambda$ since commonly used statistical softwares provide built-in routine for weighted quantile regression which automatically gives the estimator with $(a_{3i}q_i/\sum_j a_{3j}q_j)_{i=1}^n$ as weights. Estimate M_λ where $M_\lambda(\beta) = -(\partial/\partial\beta')E\left[XI(y - X'\beta < 0)|C \in \lambda\right]$ using $\widetilde{\beta}_\lambda$ [see below Corollary 15]. Therefore, since $d_m = d_\beta$, by using (45) we obtain the one-step estimator as: $\widehat{\beta}_{\text{1step}} = \widetilde{\beta}_\lambda - \widehat{M}_\lambda^{-1}(\widetilde{\beta}_\lambda) G_n(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda))/\widehat{P}(C \in \lambda)$.

C.6 Simulation evidence from Section 5 of the finite-sample properties of $\widehat{\beta}_\lambda$

Besides the efficient estimators based on various sub-samples, we also consider the complete case (CC) and IPW [see (44)] estimators. The CC estimator is the default in the statistical softwares and is based only on the complete sub-sample ignoring its likely unrepresentative of the target population.

We consider certain finite-sample properties of all these estimators and 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. We focus on the following quantities computed as averages 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 IQR (interquartile range). Mean squared error is not reported but follows directly as Mbias^2+ Std^2.

The CC and IPW estimators are numerically equivalent if $\lambda = \{3\}$ or under INDEP. Otherwise, as expected, CC can be badly biased (Mbias) since it does not recognize the sample-selection.

The other estimators are consistent under our assumptions, and their small Mbias and decreasing (with n) Std support this. The ordering of the variability of the estimators, as measured by Abias, Std and IQR, 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 under CMAR and MAR if $\lambda = \{3\}$ or $\lambda = \{1, 2, 3\}$. In these cases, it seems that in spite of the poorer quality of information in the units of $\{1, 3\}$, its larger sample size makes it more desirable than $\{2, 3\}$. (Under our premise, $\{1, 3\}$ could still be less expensive than $\{2, 3\}$ to observe.)

Overall, under our simulation design all the estimators display good properties in finite samples, and thus lend credibility to the encouraging simulation results on the efficiency loss in Section 5.

Used	$n = 600$				$n = 1200$				$n = 1800$			
Sample	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{3\}$	-.0002	.0748	.0933	.1250	.0011	.0529	.0661	.0895	-.0003	.0436	.0540	.0739
$\{1, 3\}$	.0005	.0560	.0667	.0947	.0007	.0388	.0473	.0666	.0002	.0313	.0388	.0530
$\{2, 3\}$	-.0001	.0584	.0673	.0986	.0008	.0392	.0475	.0661	.0003	.0317	.0388	.0534
$\{1, 2, 3\}$	.0003	.0523	.0584	.0878	.0006	.0346	.0411	.0589	.0003	.0278	.0337	.0475
$\{3\}$	.0004	.0773	.0927	.1296	.0001	.0527	.0659	.0885	.0002	.0434	.0539	.0737
$\{1, 3\}$	.0090	.0641	.0714	.1069	.0038	.0425	.0510	.0715	.0028	.0345	.0418	.0579
$\{2, 3\}$	.0062	.0667	.0720	.1106	.0019	.0432	.0507	.0739	.0013	.0347	.0415	.0586
$\{1, 2, 3\}$	.0082	.0631	.0649	.1044	.0030	.0403	.0458	.0686	.0021	.0320	.0377	.0545

Table 4: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators under INDEP sampling are reported based on the average over 10,000 Monte Carlo trials. Target population $\lambda = \{1, 2, 3\}$. Top panel: Intercept parameter $\beta_{\lambda,1}$. Bottom panel: Slope parameter $\beta_{\lambda,2}$.

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

Target	Used	$n = 600$				$n = 1200$				$n = 1800$			
Popln. (λ)	Sample (s)	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{1\}$	$\{3\}$: CC	-.1283	.1353	.0917	.1274	-.1283	.1297	.0649	.0870	-.1288	.1291	.0530	.0713
$\{1\}$	$\{3\}$: IPW	-.0018	.0767	.0946	.1305	-.0005	.0542	.0670	.0907	-.0008	.0437	.0547	.0737
$\{1\}$	$\{1,3\}$	-.0027	.0617	.0701	.1025	-.0013	.0413	.0498	.0690	-.0008	.0335	.0408	.0561
$\{1\}$	$\{1,2,3\}$	.0001	.0579	.0629	.0964	.0001	.0378	.0443	.0642	-.0001	.0301	.0363	.0510
$\{2\}$	$\{3\}$: CC	.2490	.2491	.0917	.1274	.2490	.2490	.0649	.0870	.2485	.2485	.0530	.0713
$\{2\}$	$\{3\}$: IPW	.0040	.0814	.1006	.1393	.0023	.0577	.0714	.0977	.0012	.0469	.0583	.0788
$\{2\}$	$\{2,3\}$	.0036	.0596	.0685	.0998	.0017	.0398	.0481	.0673	.0004	.0322	.0394	.0542
$\{2\}$	$\{1,2,3\}$	-.0014	.0565	.0615	.0965	-.0014	.0366	.0431	.0616	-.0015	.0298	.0353	.0510
$\{3\}$	$\{3\}$: CC	.0005	.0742	.0917	.1274	.0005	.0523	.0649	.0870	.0000	.0423	.0530	.0713
$\{3\}$	$\{3\}$: IPW	.0005	.0742	.0402	.1274	.0005	.0523	.0285	.0870	.0000	.0423	.0233	.0713
$\{3\}$	$\{1,3\}$	-.0019	.0581	.0673	.0976	-.0011	.0389	.0475	.0650	-.0007	.0318	.0388	.0533
$\{3\}$	$\{2,3\}$	.0008	.0601	.0690	.1004	.0005	.0401	.0482	.0678	-.0003	.0320	.0394	.0539
$\{3\}$	$\{1,2,3\}$	.0001	.0518	.0579	.0863	-.0002	.0339	.0406	.0577	-.0004	.0274	.0332	.0463
$\{1,3\}$	$\{3\}$: CC	-.0914	.1074	.0917	.1274	-.0914	.0965	.0649	.0870	-.0919	.0938	.0530	.0713
$\{1,3\}$	$\{3\}$: IPW	-.0012	.0757	.0935	.1297	-.0002	.0535	.0662	.0893	-.0006	.0431	.0541	.0726
$\{1,3\}$	$\{1,3\}$	-.0025	.0603	.0690	.0999	-.0013	.0404	.0489	.0676	-.0008	.0328	.0401	.0554
$\{1,3\}$	$\{1,2,3\}$	-.0001	.0558	.0611	.0930	-.0001	.0364	.0430	.0618	-.0003	.0291	.0352	.0493
$\{2,3\}$	$\{3\}$: CC	.1440	.1483	.0917	.1274	.1440	.1447	.0649	.0870	.1435	.1436	.0530	.0713
$\{2,3\}$	$\{3\}$: IPW	.0026	.0770	.0952	.1317	.0016	.0544	.0675	.0919	.0008	.0441	.0551	.0746
$\{2,3\}$	$\{2,3\}$	.0018	.0583	.0673	.0978	.0009	.0389	.0472	.0659	-.0002	.0313	.0386	.0527
$\{2,3\}$	$\{1,2,3\}$	.0000	.0519	.0579	.0867	-.0004	.0339	.0406	.0579	-.0007	.0276	.0332	.0466
$\{1,2,3\}$	$\{3\}$: CC	.0092	.0744	.0917	.1274	.0092	.0528	.0649	.0870	.0087	.0428	.0530	.0713
$\{1,2,3\}$	$\{3\}$: IPW	.0004	.0761	.0939	.1308	.0007	.0536	.0666	.0894	.0001	.0434	.0544	.0733
$\{1,2,3\}$	$\{1,3\}$	-.0004	.0598	.0687	.0996	-.0002	.0400	.0486	.0670	-.0001	.0327	.0398	.0551
$\{1,2,3\}$	$\{2,3\}$	-.0013	.0619	.0711	.1033	-.0006	.0412	.0496	.0695	-.0013	.0328	.0404	.0549
$\{1,2,3\}$	$\{1,2,3\}$	.0000	.0531	.0587	.0884	-.0001	.0347	.0413	.0589	-.0004	.0280	.0338	.0472

Table 5: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators of the Intercept parameter ($\beta_{\lambda,1}$) under CMAR sampling are reported 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	Used	$n = 600$				$n = 1200$				$n = 1800$			
Popln. (λ)	Sample (s)	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{1\}$	$\{3\}$: CC	-.0080	.0786	.0952	.1309	-.0050	.0546	.0677	.0924	-.0074	.0446	.0553	.0743
$\{1\}$	$\{3\}$: IPW	-.0027	.0846	.1008	.1422	.0014	.0592	.0727	.1005	-.0009	.0480	.0596	.0809
$\{1\}$	$\{1,3\}$	.0097	.0708	.0764	.1175	.0066	.0477	.0545	.0800	.0033	.0378	.0448	.0634
$\{1\}$	$\{1,2,3\}$	.0060	.0691	.0707	.1165	.0046	.0450	.0502	.0751	.0023	.0355	.0414	.0599
$\{2\}$	$\{3\}$: CC	.0232	.0801	.0952	.1309	.0262	.0582	.0677	.0924	.0238	.0480	.0553	.0743
$\{2\}$	$\{3\}$: IPW	-.0021	.0963	.1119	.1602	.0005	.0678	.0818	.1142	-.0018	.0546	.0675	.0911
$\{2\}$	$\{2,3\}$	.0068	.0726	.0760	.1218	.0045	.0480	.0545	.0803	.0028	.0386	.0451	.0647
$\{2\}$	$\{1,2,3\}$	.0031	.0748	.0755	.1260	.0018	.0486	.0533	.0821	.0007	.0385	.0440	.0648
$\{3\}$	$\{3\}$: CC	-.0012	.0782	.0952	.1309	.0018	.0544	.0677	.0924	-.0006	.0441	.0553	.0743
$\{3\}$	$\{3\}$: IPW	-.0012	.0782	.0417	.1309	.0018	.0544	.0297	.0924	-.0006	.0441	.0243	.0743
$\{3\}$	$\{1,3\}$	.0101	.0700	.0769	.1163	.0065	.0468	.0539	.0793	.0032	.0366	.0441	.0615
$\{3\}$	$\{2,3\}$	-.0009	.0729	.0777	.1224	.0010	.0467	.0543	.0786	-.0003	.0369	.0444	.0613
$\{3\}$	$\{1,2,3\}$	.0115	.0646	.0669	.1071	.0075	.0418	.0470	.0698	.0048	.0330	.0387	.0552
$\{1,3\}$	$\{3\}$: CC	-.0125	.0790	.0952	.1309	-.0095	.0550	.0677	.0924	-.0119	.0452	.0553	.0743
$\{1,3\}$	$\{3\}$: IPW	-.0023	.0820	.0984	.1381	.0016	.0572	.0706	.0973	-.0008	.0463	.0578	.0782
$\{1,3\}$	$\{1,3\}$	.0099	.0696	.0756	.1151	.0067	.0468	.0536	.0787	.0034	.0369	.0440	.0622
$\{1,3\}$	$\{1,2,3\}$	.0075	.0668	.0688	.1120	.0054	.0435	.0487	.0726	.0031	.0342	.0401	.0576
$\{2,3\}$	$\{3\}$: CC	-.0135	.0791	.0952	.1309	-.0105	.0551	.0677	.0924	-.0129	.0454	.0553	.0743
$\{2,3\}$	$\{3\}$: IPW	-.0017	.0854	.1016	.1426	.0011	.0596	.0733	.1010	-.0013	.0481	.0601	.0808
$\{2,3\}$	$\{2,3\}$	.0037	.0687	.0726	.1154	.0032	.0448	.0516	.0753	.0017	.0358	.0425	.0600
$\{2,3\}$	$\{1,2,3\}$	.0070	.0661	.0680	.1112	.0044	.0430	.0479	.0731	.0026	.0340	.0395	.0570
$\{1,2,3\}$	$\{3\}$: CC	-.0450	.0869	.0952	.1309	-.0420	.0643	.0677	.0924	-.0444	.0578	.0553	.0743
$\{1,2,3\}$	$\{3\}$: IPW	-.0025	.0801	.0967	.1335	.0012	.0558	.0692	.0947	-.0012	.0451	.0565	.0756
$\{1,2,3\}$	$\{1,3\}$	.0038	.0712	.0774	.1186	.0028	.0475	.0544	.0805	.0004	.0374	.0445	.0634
$\{1,2,3\}$	$\{2,3\}$	-.0037	.0745	.0786	.1240	-.0011	.0482	.0552	.0812	-.0021	.0382	.0453	.0641
$\{1,2,3\}$	$\{1,2,3\}$	.0066	.0628	.0648	.1048	.0047	.0409	.0458	.0691	.0027	.0322	.0377	.0548

Table 6: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators of the Slope parameter ($\beta_{\lambda,2}$) under CMAR sampling are reported 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	Used	$n = 600$				$n = 1200$				$n = 1800$			
Popln. (λ)	Sample (s)	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{1\}$	$\{3\}$: CC	-.1380	.1438	.0926	.1236	-.1394	.1402	.0656	.0867	-.1380	.1382	.0536	.0735
$\{1\}$	$\{3\}$: IPW	-.0026	.0771	.0963	.1283	-.0024	.0541	.0683	.0906	-.0005	.0454	.0559	.0769
$\{1\}$	$\{1,3\}$	-.0061	.0618	.0715	.1026	-.0030	.0426	.0510	.0720	-.0016	.0344	.0418	.0581
$\{1\}$	$\{1,2,3\}$	.0000	.0585	.0629	.0987	.0002	.0387	.0446	.0654	.0007	.0311	.0366	.0525
$\{2\}$	$\{3\}$: CC	.2371	.2376	.0926	.1236	.2357	.2357	.0656	.0867	.2371	.2371	.0536	.0735
$\{2\}$	$\{3\}$: IPW	.0038	.0849	.1043	.1432	.0010	.0587	.0742	.0981	.0011	.0490	.0607	.0828
$\{2\}$	$\{2,3\}$	.0060	.0617	.0709	.1033	.0025	.0412	.0496	.0698	.0016	.0332	.0405	.0557
$\{2\}$	$\{1,2,3\}$	-.0016	.0595	.0641	.0998	-.0012	.0391	.0448	.0653	-.0015	.0311	.0368	.0525
$\{3\}$	$\{3\}$: CC	.0004	.0742	.0926	.1236	-.0010	.0517	.0656	.0867	.0004	.0435	.0536	.0735
$\{3\}$	$\{3\}$: IPW	.0004	.0742	.0405	.1236	-.0010	.0517	.0287	.0867	.0004	.0435	.0235	.0735
$\{3\}$	$\{1,3\}$	-.0041	.0579	.0679	.0963	-.0021	.0397	.0480	.0667	-.0012	.0321	.0393	.0541
$\{3\}$	$\{2,3\}$	.0011	.0626	.0699	.1037	.0003	.0408	.0488	.0687	.0010	.0333	.0399	.0558
$\{3\}$	$\{1,2,3\}$	-.0003	.0529	.0588	.0892	-.0003	.0348	.0411	.0590	.0001	.0281	.0337	.0476
$\{1,3\}$	$\{3\}$: CC	-.0990	.1129	.0926	.1236	-.1004	.1038	.0656	.0867	-.0990	.1005	.0536	.0735
$\{1,3\}$	$\{3\}$: IPW	-.0017	.0760	.0949	.1271	-.0020	.0532	.0673	.0895	-.0003	.0447	.0550	.0755
$\{1,3\}$	$\{1,3\}$	-.0056	.0602	.0700	.1010	-.0028	.0415	.0498	.0701	-.0015	.0335	.0408	.0562
$\{1,3\}$	$\{1,2,3\}$	-.0002	.0564	.0613	.0956	.0000	.0373	.0433	.0632	.0005	.0300	.0355	.0507
$\{2,3\}$	$\{3\}$: CC	.1343	.1411	.0926	.1236	.1329	.1339	.0656	.0867	.1343	.1345	.0536	.0735
$\{2,3\}$	$\{3\}$: IPW	.0024	.0778	.0962	.1308	.0001	.0538	.0683	.0896	.0008	.0453	.0558	.0771
$\{2,3\}$	$\{2,3\}$	.0026	.0586	.0676	.0971	.0007	.0390	.0472	.0659	.0009	.0315	.0386	.0528
$\{2,3\}$	$\{1,2,3\}$	.0003	.0515	.0579	.0861	-.0001	.0342	.0405	.0579	-.0002	.0274	.0331	.0463
$\{1,2,3\}$	$\{3\}$: CC	-.0005	.0742	.0926	.1236	-.0019	.0517	.0656	.0867	-.0005	.0435	.0536	.0735
$\{1,2,3\}$	$\{3\}$: IPW	.0000	.0767	.0955	.1276	-.0012	.0534	.0678	.0904	.0002	.0450	.0554	.0763
$\{1,2,3\}$	$\{1,3\}$	-.0036	.0596	.0695	.0991	-.0018	.0411	.0493	.0695	-.0010	.0331	.0404	.0561
$\{1,2,3\}$	$\{2,3\}$	-.0012	.0632	.0716	.1044	-.0011	.0416	.0498	.0707	.0000	.0335	.0406	.0567
$\{1,2,3\}$	$\{1,2,3\}$	.0000	.0531	.0587	.0896	.0000	.0354	.0414	.0602	.0002	.0282	.0339	.0478

Table 7: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators of the Intercept parameter ($\beta_{\lambda,1}$) under MAR sampling are reported 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	Used	$n = 600$				$n = 1200$				$n = 1800$			
Popln. (λ)	Sample (s)	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{1\}$	$\{3\}$: CC	-.0164	.0802	.0974	.1328	-.0146	.0573	.0692	.0949	-.0149	.0469	.0566	.0769
$\{1\}$	$\{3\}$: IPW	-.0019	.0843	.1015	.1428	.0004	.0596	.0729	.0997	.0004	.0484	.0599	.0816
$\{1\}$	$\{1,3\}$	.0109	.0714	.0771	.1193	.0064	.0475	.0550	.0795	.0039	.0379	.0452	.0639
$\{1\}$	$\{1,2,3\}$	.0072	.0696	.0709	.1167	.0040	.0452	.0502	.0753	.0021	.0359	.0414	.0612
$\{2\}$	$\{3\}$: CC	.0227	.0811	.0974	.1328	.0245	.0596	.0692	.0949	.0242	.0495	.0566	.0769
$\{2\}$	$\{3\}$: IPW	-.0023	.1096	.1238	.1859	.0008	.0777	.0920	.1307	-.0001	.0630	.0765	.1055
$\{2\}$	$\{2,3\}$	.0108	.0824	.0810	.1358	.0059	.0547	.0590	.0915	.0026	.0436	.0493	.0737
$\{2\}$	$\{1,2,3\}$	.0042	.0846	.0801	.1404	.0021	.0546	.0579	.0923	-.0001	.0437	.0483	.0740
$\{3\}$	$\{3\}$: CC	-.0007	.0791	.0974	.1328	.0011	.0561	.0692	.0949	.0008	.0454	.0566	.0769
$\{3\}$	$\{3\}$: IPW	-.0007	.0791	.0426	.1328	.0011	.0561	.0303	.0949	.0008	.0454	.0248	.0769
$\{3\}$	$\{1,3\}$	.0134	.0705	.0779	.1163	.0076	.0468	.0548	.0781	.0047	.0371	.0448	.0620
$\{3\}$	$\{2,3\}$	-.0014	.0788	.0826	.1281	-.0014	.0500	.0569	.0836	-.0015	.0395	.0465	.0662
$\{3\}$	$\{1,2,3\}$	.0136	.0679	.0698	.1121	.0075	.0436	.0488	.0727	.0046	.0347	.0401	.0581
$\{1,3\}$	$\{3\}$: CC	-.0162	.0802	.0974	.1328	-.0144	.0573	.0692	.0949	-.0147	.0469	.0566	.0769
$\{1,3\}$	$\{3\}$: IPW	-.0017	.0824	.0999	.1389	.0006	.0582	.0715	.0978	.0004	.0472	.0587	.0796
$\{1,3\}$	$\{1,3\}$	.0115	.0705	.0767	.1174	.0067	.0469	.0545	.0781	.0041	.0373	.0447	.0619
$\{1,3\}$	$\{1,2,3\}$	.0088	.0683	.0700	.1136	.0049	.0442	.0493	.0735	.0028	.0351	.0406	.0592
$\{2,3\}$	$\{3\}$: CC	-.0219	.0811	.0974	.1328	-.0201	.0584	.0692	.0949	-.0204	.0483	.0566	.0769
$\{2,3\}$	$\{3\}$: IPW	-.0015	.0906	.1067	.1511	.0011	.0642	.0777	.1087	.0004	.0520	.0641	.0882
$\{2,3\}$	$\{2,3\}$	.0055	.0727	.0740	.1204	.0030	.0477	.0530	.0801	.0012	.0381	.0439	.0644
$\{2,3\}$	$\{1,2,3\}$	.0085	.0696	.0686	.1152	.0047	.0448	.0490	.0737	.0022	.0359	.0406	.0608
$\{1,2,3\}$	$\{3\}$: CC	-.0534	.0904	.0974	.1328	-.0516	.0704	.0692	.0949	-.0519	.0631	.0566	.0769
$\{1,2,3\}$	$\{3\}$: IPW	-.0021	.0824	.0997	.1391	.0006	.0582	.0716	.0987	.0003	.0472	.0587	.0793
$\{1,2,3\}$	$\{1,3\}$	.0047	.0730	.0791	.1222	.0025	.0482	.0559	.0802	.0012	.0383	.0459	.0649
$\{1,2,3\}$	$\{2,3\}$	-.0037	.0787	.0808	.1283	-.0028	.0505	.0564	.0841	-.0028	.0402	.0465	.0670
$\{1,2,3\}$	$\{1,2,3\}$	.0079	.0647	.0652	.1075	.0045	.0416	.0463	.0696	.0024	.0333	.0382	.0566

Table 8: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators of the Slope parameter ($\beta_{\lambda,2}$) under MAR sampling are reported based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

C.7 Proofs

For simplicity, we write β_λ as β. We follow the steps of the proof for Theorems 1 and 2 in Chen et al. (2003) with adjustments for the weaker conditions that are consequences of (43) [see (F1)-(F3)]. The main 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.

Proof of Proposition 13: (F1) already implies the standard well-separability of β^0 by virtue of (3). Hence, for all $\delta > 0$ there exists $\epsilon(\delta) > 0$ such that $P(\|\widehat{\beta} - \beta^0\| > \delta) \leq P(\|G(\widehat{\beta}, h^\dagger)\| \geq \epsilon(\delta))$.

Therefore, to establish that $\widehat{\beta} \xrightarrow{P} \beta^0$, it is sufficient to show that $\|G(\widehat{\beta}, h^\dagger)\| = o_p(1)$. Assumption (B2) implies that $P(\widehat{h}(\beta) \in \mathcal{H}) \to 1$ uniformly in $\beta \in \mathcal{B}$ as $n \to \infty$. The rest of the proof works conditional on the sequence of events $\{\widehat{h}(\widehat{\beta}) \in \mathcal{H}\}$, i.e., we use the fact that:

$$\begin{aligned}
& P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta)) \\
= \; & P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta) | \widehat{h}(\widehat{\beta}) \in \mathcal{H}) P(\widehat{h}(\widehat{\beta}) \in \mathcal{H}) + P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta) | \widehat{h}(\widehat{\beta}) \notin \mathcal{H}) P(\widehat{h}(\widehat{\beta}) \notin \mathcal{H}) \\
= \; & P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta) | \widehat{h}(\widehat{\beta}) \in \mathcal{H}) + o(1) \qquad (46)
\end{aligned}$$

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

To this end, first note that:

$$\begin{aligned}
\|G(\widehat{\beta}, h^\dagger)\| & \leq \|G(\widehat{\beta}, h^\dagger) - G(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h})\| + \|G_n(\widehat{\beta}, \widehat{h})\| \\
& = \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h})\| + \|G_n(\widehat{\beta}, \widehat{h})\|. \qquad (47)
\end{aligned}$$

The inequality holds by the triangle inequality (kept implicit hereafter). The equality holds by (F3).

Using (B3) and then (F3), we obtain:

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

Using this along with (47) gives:

$$\begin{aligned}
& \|G(\widehat{\beta}, h^\dagger)\| \times (1 - o_p(1)) \\
\leq \; & o_p(1) + \|G_n(\widehat{\beta}, \widehat{h})\| \times (1 + o_p(1)) \\
\leq \; & o_p(1) + \|G_n(\widehat{\beta}, \widehat{h})\|_{W_n} \times (1 + \|W_n^{-1} - W^{-1}\| + \|W^{-1} - I_{d_m}\|) \times (1 + o_p(1)) \\
= \; & o_p(1) + \|G_n(\widehat{\beta}, \widehat{h})\|_{W_n} \times (c + o_p(1)) \\
\leq \; & o_p(1) + \inf_{\beta \in \mathcal{B}} \|G_n(\beta, \widehat{h})\|_{W_n} \times (c + o_p(1)) \qquad (48)
\end{aligned}$$

where $c = 1 + \|W^{-1} - I_{d_m}\|$. The equality in the above equations follows since (i) $W_n - W = o_p(1)$ for a constant positive definite matrix W implies that W_n^{-1} exists with probability approaching one and $W_n^{-1} - W^{-1} = o_p(1)$, and hence $\|W_n^{-1} - W^{-1}\| = o_p(1)$ as d_m is finite, (ii) a finite and positive definite W and a finite d_m imply that $c(> 1)$ is finite. The last inequality in (48) is due to (B1).

Following similar steps again and letting $d = 1 + \|W - I_{d_m}\|$ (> 1 and finite), note that:

$$\begin{aligned}
& \|G_n(\beta, \widehat{h})\|_{W_n} \\
\leq\ & \|G_n(\beta, \widehat{h})\| \times (d + o_p(1)) \\
\leq\ & \{\|G_n(\beta, \widehat{h}) - G(\beta, \widehat{h})\| + \|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| + \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\|\} \times (d + o_p(1)) \quad (49)
\end{aligned}$$

by using (43), i.e., $G(\beta^0, h) = 0$ for all $h \in \mathcal{H}$ (in the last term inside the braces). This is the special feature of our setup; whereas this holds only at $h = h^0$ in Chen et al. (2003). On the other hand, $\|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| = 0$ by (F3). Lastly, since $G(\beta^0, h^\dagger) = 0$ and $\|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| = 0$, we can use (B3) as before to obtain that:

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

where the second line follows by the same argument as in (48). Therefore, (49) 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 that:

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

since $\inf_{\beta \in \mathcal{B}} \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\|_W = 0$. So, by (46) and (48) it follows that $\|G(\widehat{\beta}, h^\dagger)\| = o_p(1)$. ■

Proof of Proposition 14: First, we show $\sqrt{n}$-consistency of $\widehat{\beta}$, and then 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 13, 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 (46) 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:

$$\begin{aligned}
\|G(\widehat{\beta}, h^\dagger)\| &\leq \|G(\widehat{\beta}, h^\dagger) - G(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h}) + G_n(\beta^0, h^\dagger)\| + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G_n(\beta^0, h^\dagger)\| \\
&= 0 + \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h}) + G_n(\beta^0, h^\dagger)\| + \|G_n(\widehat{\beta}, \widehat{h})\| + O_p(n^{-1/2}) \qquad (50)
\end{aligned}$$

where the first 0 follows from (F2) 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 (F3). Therefore, this along with (50) implies that:

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

which, further implies that (second inequality below follows using same arguments as in (48) with $c = 1 + \|W^{-1} - I_{d_m}\|$)

$$\begin{aligned}
\|G(\widehat{\beta}, h^\dagger)\| \times (1 - o_p(1)) &\leq O_p(n^{-1/2}) + \|G_n(\widehat{\beta}, \widehat{h})\| \times (1 + o_p(1)) \\
&\leq O_p(n^{-1/2}) + \|G_n(\widehat{\beta}, \widehat{h})\|_{W_n} \times (c + o_p(1)) \\
&\leq O_p(n^{-1/2}) + \inf_{\beta \in \mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\|_{W_n} \times (c + o_p(1)) \qquad (51)
\end{aligned}$$

where the last line follows by (C1). Now, for $d = 1 + \|W - I_{d_m}\|$, recall from the first line of (49) 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, i.e., the 0, from (F3), and the last one from (C4). Therefore,

$$\begin{aligned}
\|G_n(\beta,\widehat{h})\|\times(1-o_p(1)) &\leq \|G(\beta,\widehat{h})\|\times o_p(1)+\|G(\beta,h^\dagger)\|+O_p(n^{-1/2})\\
&\leq \|G(\beta,\widehat{h})-G(\beta,h^\dagger)\|\times o_p(1)+\|G(\beta,h^\dagger)\|\times(1+o_p(1))+O_p(n^{-1/2})\\
&= \|G(\beta,h^\dagger)\|\times(1+o_p(1))+O_p(n^{-1/2})\ \text{[by (F3)]}\\
&\leq \|G(\beta,h^\dagger)-G(\beta^0,h^\dagger)\|\times(1+o_p(1))+\|G(\beta^0,h^\dagger)\|\times(1+o_p(1))+O_p(n^{-1/2})\\
&= \|G(\beta,h^\dagger)-G(\beta^0,h^\dagger)\|\times(1+o_p(1))+O_p(n^{-1/2})
\end{aligned}$$

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 13, 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, (51) 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 (F2) and (F3). This gives:

$$\begin{aligned}
&\|G_n(\widehat{\beta},\widehat{h})-L_n(\widehat{\beta})\|\\
=\ &\|G_n(\widehat{\beta},\widehat{h})-G_n(\beta^0,h^\dagger)-M_\lambda(\widehat{\beta}-\beta^0)\|\\
=\ &\|G_n(\widehat{\beta},\widehat{h})-G_n(\beta^0,h^\dagger)-G(\widehat{\beta},\widehat{h})+G(\widehat{\beta},\widehat{h})+G(\widehat{\beta},h^\dagger)-G(\widehat{\beta},h^\dagger)-M_\lambda(\widehat{\beta}-\beta^0)\|\\
\leq\ &\|G_n(\widehat{\beta},\widehat{h})-G_n(\beta^0,h^\dagger)-G(\widehat{\beta},\widehat{h})\|+\|G(\widehat{\beta},\widehat{h})-G(\widehat{\beta},h^\dagger)\|+\|G(\widehat{\beta},h^\dagger)-M_\lambda(\widehat{\beta}-\beta^0)\|\\
\leq\ &\|G_n(\widehat{\beta},\widehat{h})-G_n(\beta^0,h^\dagger)-G(\widehat{\beta},\widehat{h})\|+\|G(\widehat{\beta},h^\dagger)-M_\lambda(\widehat{\beta}-\beta^0)\|\ \text{[by (F3)]}\\
\leq\ &o_p(1)\times\{1+\|G_n(\widehat{\beta},\widehat{h})\|+\|G(\widehat{\beta},\widehat{h})\|\}+\|G(\widehat{\beta},h^\dagger)-G(\beta^0,h^\dagger)-M_\lambda(\widehat{\beta}-\beta^0)\|
\end{aligned}$$

where the term inside braces follows 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 (F2), 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 (51) 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 (F3) 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 key 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 15:

(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 (6). ■

Proof of Proposition 16: Define $L_n(\beta) := G_n(\beta^0, h^0) + M_\lambda(\beta - \beta^0)$ and note that $\sqrt{n} L_n(\widetilde{\beta}) = O_p(1)$ by assumptions (A3), (C4) and since $\sqrt{n}(\widetilde{\beta} - \beta^0) = O_p(1)$. Therefore, using (F1), (F3) and also the stochastic equicontinuity condition from the statement of the proposition, we obtain that:

$$
\begin{aligned}
& \sqrt{n} \|G_n(\widetilde{\beta}, \widehat{h}) - L_n(\widetilde{\beta})\| \\
= \;& \sqrt{n} \|\{G_n(\widetilde{\beta}, \widehat{h}) - G(\widetilde{\beta}, \widehat{h}) - G_n(\beta^0, h^0)\} + \{G(\widetilde{\beta}, \widehat{h}) - G(\widetilde{\beta}, h^0(\beta^0))\} + \{G(\widetilde{\beta}, h^0(\beta^0) - M_\lambda(\widetilde{\beta} - \beta^0)\| \\
\leq \;& \sup_{\beta \in \mathcal{B}_{\delta_n}, h \in \mathcal{H}_{\delta_n}} \sqrt{n} \|G_n(\beta, h) - G(\beta, h) - G_n(\beta^0, h^0)\| + \sqrt{n} \|G(\widetilde{\beta}, \widehat{h}) - G(\widetilde{\beta}, h^0(\beta^0))\| \\
& \qquad + \|\sqrt{n} G(\beta^0, h^0) + (M_\lambda + o_p(1) - M_\lambda) \sqrt{n}(\widetilde{\beta} - \beta^0)\| \\
= \;& o_p(1) + 0 + (0 + o_p(1)) = o_p(1).
\end{aligned}
$$

Now, the proof completes since under the conditions of the proposition, the definition in (45) gives:

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\beta}_{\text{1step}} - \widetilde{\beta}\right) &= -\left(\Omega_\lambda^{-1} + o_p(1)\right)\left(M_\lambda' + o_p(1)\right)\left(V_\lambda^{-1} + o_p(1)\right)\left(\sqrt{n} L_n(\widetilde{\beta}) + o_p(1)\right) \\
&= -\Omega_\lambda^{-1} M_\lambda' V_\lambda^{-1}\left(\sqrt{n} G_n(\beta^0, h^0) + M_\lambda \sqrt{n}(\widetilde{\beta} - \beta^0)\right) + o_p(1) \\
&= \sqrt{n}\left(\widehat{\beta} - \beta^0\right) - \sqrt{n}\left(\widetilde{\beta} - \beta^0\right) + o_p(1) = \sqrt{n}\left(\widehat{\beta} - \widetilde{\beta}\right) + o_p(1). \blacksquare
\end{aligned}
$$

References

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

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.

Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104: 3003–3037.

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

Ashraf, N., Berry, J., and Shapiro, J. M. (2010). Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia. *American Economic Review*, 100: 2383–2413.

Ashraf, N., Field, E., and Lee, J. (2014). Household Bargaining and Excess Fertility: An Experimental Study in Zambia. *American Economic Review*, 104: 2210–2237.

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

Barnwell, J. L. and Chaudhuri, S. (2018). Efficient estimation in sub and full populations with monotonically missing at random data. Technical report, McGill University.

Beaman, L., Karlan, D., Thusbaert, B., and Udry, C. (2015). Self-Selection into Credit Markets: Evidence from Agriculture in Mali. Mimeo.

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*, 98: 3 – 18.

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

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

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

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

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

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

Dardanoni, V., Modica, S., and Peracchi, F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, 162: 362–368.

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.

DiNardo, J., McCrary, J., and Sanbonmatsu, L. (2006). Constructive Proposals for Dealing with Attrition: An Empirical Example. NBER and University of Michigan.

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, 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.

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.

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.

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

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

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.

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. 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.

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

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

Thornton, R. L. (2008). The Demand for, and Impact of, Learning HIV Status. *American Economic Review*, 98: 1829–1863.

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

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

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

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.