A Note on Efficiency Gains from Multiple Incomplete Subsamples[*†]

Saraswata Chaudhuri[‡]

Current version: Jan 20, 2014; First version: March 8, 2013.

Comments are welcome.

Abstract

We demonstrate efficiency gains in estimation by optimally using multiple subsamples all but one of which are incomplete following a monotone pattern. The finite dimensional parameter of interest is defined by moment restrictions on a target population which is some arbitrary union of the possibly different subpopulations for the multiple subsamples. A form of the missing at random (MAR) assumption is made for identification. MAR also makes the information contained in each incomplete subsample usable and thus contributes to efficiency gains. We show that the characteristics and possibility of such efficiency gains can be very different from those in the two subsamples contexts that have been studied extensively in the literature. Implication of these results on possible sampling strategies is briefly noted. We also show that a set of unconditional and conditional moment restrictions exhausts all the relevant information in the subsamples and can be easily used in a Frisch-Waugh-Lovell type sequential way, by virtue of monotonicity, for efficient estimation of the parameter of interest.

JEL Classification: C13; C14; C31.

Keywords: Efficiency gain; Generalized method of moments; Incomplete subsamples; Monotonically missing data; Semiparametric efficiency bound; Sequential Projections.

[*] An older and longer version of the paper, containing additional results, was circulated as "A Note on Efficiency Gains from Auxiliary Samples" and is available from the author's web page.

[†] We thank A. Prokhorov, C. Muris, D. Guilkey, E. Renault, F. Lange, J. Hill, P. Saha Chaudhuri, SJ. Lee, V. Zinde-Walsh, the seminar participants at U. Sydney, U. New South Wales, U. Canterbury, West Virginia University, McGill, Concordia and the Midwest Econometrics Group meetings (2013) for helpful comments.

[‡] Department of Economics, CB 3305, University of North Carolina, Chapel Hill, NC 27519. Telephone: 919-966-3962. Fax: 919-966-4986. Email: saraswata_chaudhuri@unc.edu.

1 Introduction

We consider efficient estimation of parameters by using multiple subsamples, all but one of which are incomplete following a monotone pattern under a form of missingness at random (MAR). The finite dimensional parameter of interest is defined by moment restrictions on a target population which is some arbitrary union of the possibly different subpopulations for the multiple subsamples.

By "incomplete" we mean that the true values of some relevant random variables are missing for all units in the concerned subsample. Incompleteness that satisfies MAR can be due to sampling design such as selective followups in multi-phase surveys, non-response in surveys, attrition and/or refreshment samples in panel surveys, by definition of counterfactuals, measurement error, etc.

The MAR assumption and the existence of one complete subsample (and an additional overlap condition to be defined duely) allow us to focus on efficiency considerations following the seminal work of Robins et al. (1994), and sidestep the difficult and important issue of point identification of the parameter of interest that has also received attention in the related literature on data combination; see Ichimura and Martinez-Sanchis (2005), Ridder and Moffitt (2007) and the references therein.[1]

MAR is used if it can be argued that conditioning on certain observed variables makes the event of incompleteness independent of the true values of the variables that are unobserved in the concerned incomplete subsample. Whether or not such conditioning variables exist and hence MAR holds is a difficult question; see Little and Rubin (2002), Heitjan and Rubin (1991), Gill et al. (1997), Tsiatis (2006). Nevertheless, MAR has been successfully employed for point identification of parameters in contexts such as: (i) program evaluation: see Imbens (2004), Heckman and Vytlacil (2007) and the references therein; (ii) attrition in panel data: see Fitzgerald et al. (1998); (iii) non-classical measurement error: see Chen et al. (2005). Chen et al. (2008) generalize the framework of many of the above papers to moment conditions models and consider efficient estimation under MAR.

Our demonstration of efficiency gains by using multiple subsamples is based on the framework of Chen et al. (2008) with a technical extension that accommodates for a monotone pattern of multi-level incompleteness in the subsamples. Let us first formally describe the extended framework.

Consider a finite dimensional random variable $Z = (Z_1, \ldots, Z_R)$ partitioned in R blocks. Suppose we observe the random variables $(C, G_C(Z))$ for a group of units. $C \in \mathbb{C} := \{1, \ldots, R\}$ is the coarsening variable. $G_C(Z)$ is the coarsening transformation of Z defined as $G_r(Z) := (Z_1, \ldots, Z_r)$ for $r = 1, \ldots, R$, and reflects a single hierarchy in the information content, i.e., monotone coarsening or missingness. In this paper, a collection of units with the same level of coarsening is called a *subsample*.

[1] In a related work outside of the references, Nevo (2003) uses auxiliary data to correct for sample selection.

The r-th subsample is the collection of units for whom only $G_r(Z)$ is observed. $I(C = r)$ is the indicator of the r-th subsample. The R-th subsample is complete while the rest are incomplete satisfying the monotone pattern of multi-level incompleteness that resulted from monotone missingness.[2]

Now consider a function $m(Z;\beta) : \text{Support}(Z) \times \mathcal{B} \mapsto \mathbb{R}^{d_m}$, $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ where $d_\beta \leq d_m$. The parameter value of interest $\beta^0 \in \text{interior}(\mathcal{B})$ is defined as follows. Consider any given element $\lambda \in \Lambda$ where $\Lambda := \text{Power-Set}(\mathbb{C}) \setminus \{\text{empty set}\}$, and let

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

β^0 is defined as a function of λ and may not be same across $\lambda \in \Lambda$ unless C and Z are independent. λ represents the *target population* for the study. Specific examples are given in Section 3.

Point identification of β^0 follows from: (i) a suitable overlap assumption: $P(C = R|G_R(Z)) > 0$ almost surely in $G_R(Z)$, and (ii) a convenient version of the MAR assumption: For $r = 1, \ldots, R$,

$$P(C = r|Z) \underbrace{=}_{\text{standard MAR}}^{\text{[A]}} P(C = r|G_r(Z)) \underbrace{=}_{\text{convenient version}}^{\text{[B]}} P(C = r|G_1(Z)) \equiv P(C = r|Z_1). \tag{2}$$

This is because these two assumptions (i) and (ii) imply:

$$E\left[\frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \frac{I(C = R)}{P(C = R|G_1(Z))} m(Z;\beta)\right] = E[m(Z;\beta)|C \in \lambda].$$

The convenient version in (2)[B] is unrealistic in certain cases [see Appendix B]. However, we maintain it for a simple demonstration of how the additional information content of each subsample contributes to the possible efficiency gain due to the subsample's use in estimation of β^0 *for any target* $\lambda \in \Lambda$.

Technically, our setup is a straightforward extension of the vast literature on missing data, in particular under monotone missingness [see Robins and Rotnitzky (1995)], to cases where the parameter of interest is possibly defined only in terms of subpopulations. However, viewed from the perspective of the use of auxiliary samples for efficiency gains, our setup extends Chen et al. (2008)'s or Chen et al. (2005)'s two-level ($R = 2$) model in a simple way that provides additional insights useful for practical purposes. This will be highlighted in the context of collection and use of possibly incomplete subsamples, similar to multi-phase surveys, for efficiency gains. Our general results, however, remain applicable more broadly to other contexts of the monotone missing (at random) data framework.

Let us preview an implication of the extension beyond $R = 2$. Take for example $R = 3$ and $\lambda = 3$ in

[2]Consideration of efficiency under more general incompleteness resulting from non-monotone missingness is usually difficult [see Tsiatis (2006)]. Chaudhuri and Guilkey (2013) consider some specific examples of non-monotone missingness.

(1). (Section 3 presents scenarios under which such estimation is of interest.) The complete subsample ($C = 3$) can identify β^0 by virtue of MAR. The incomplete subsamples ($C = 1$) and ($C = 2$) can possibly help in efficiency gains. However, we show that neither ($C = 1$) nor ($C = 2$), when used with ($C = 3$), gives the efficiency gain. Interestingly, on the other hand, we also show that efficiency gain is possible when all three subsamples are used. This insight is available only by going beyond $R = 2$.

The rest of the paper and a summary of its contributions are as follows:

(I) In Section 2 we present the efficiency bounds (varying with λ) for estimation of β^0 defined in (1) and the MAR assumption in (2) for a general (finite) R. We also present bounds under the respective assumptions that $P(C = r|G_r(Z))$ for $r = 1, \ldots, R$ is: (a) completely known, and (b) known up to some unknown finite dimensional parameter. Unsurprisingly, similar to the results in Hahn (1998) and Chen et al. (2008) for $R = 2$, these additional assumptions lead to smaller (in matrix sense) bounds also for a general R when $\lambda \neq \mathbb{C}$. Asymptotic properties of the concerned estimators for β^0 are standard and, for brevity, we only provide suitable references without going into the details.

(II) In Section 3 we consider the simple case of $R = 3$ to demonstrate the possible efficiency gains from optimally using all the subsamples (as opposed to leaving out some similar to the example above) for estimation of β^0 under all choices of λ in (1). Efficiency comparisons are based on the bounds obtained in Section 2. The stress is on possible efficiency gains from features of the joint (conditional) distribution of Z that get revealed more precisely due to the availability of the subsamples and through the use of MAR, and not merely due to an increase in sample size. We also provide examples, with emphasis on the case of possible sampling strategies, where these results could be applicable.

(III) The exposition in Section 3 does not assume knowledge of $P(C = r|G_r(Z))$ to maintain broader appeal. However, the stress on possible sampling strategies makes it imperative that special attention be given to the case where $P(C = r|G_r(Z))$ is known. Under this assumption and a general R, we demonstrate in Section 4 how the additional information contained in each incomplete subsample contributes to efficiency gains by virtue of MAR. In particular, similar to Graham (2011) who considers the special case of $R = 2$ and $\lambda = \mathbb{C}$, we show that a set of unconditional and conditional moment restrictions exhausts all the information contained in (1) and (2). Specification of these conditional moment restrictions is new. Monotonicity in the conditioning sets of the latter moment restrictions is then used to invoke a Frisch-Waugh-Lovell type sequential application of the results of Brown and Newey (1998) to combine the unconditional and conditional moment restrictions. The result remains valid for the most common scenario in empirical research: $P(C = r|G_r(Z))$ is unknown and $\lambda = \mathbb{C}$.

(IV) The result of Section 4 also has a theoretical implication. While many papers discuss that

when $R = 2$ and $\lambda = \mathbb{C}$, plugging in an estimated $P(C = R|G_1(Z))$ instead of its true value in the inverse probability weighted estimator of β^0 gives smaller asymptotic variance [see Graham (2011) and the references therein], a similar discussion is apparently absent in the context where $\lambda \neq \mathbb{C}$ [except in Chen et al. (2008) with $R = 2$]. But $\lambda \neq \mathbb{C}$ is perhaps more interesting because the efficiency bounds are different depending on the knowledge of $P(C = r|G_1(Z))$ [see Section 2]. We show how our result provides a justification for related phenomena under the general framework considered in this paper.

Section 5 concludes. Proofs of all technical results are collected in Appendix A.

2 Efficiency bounds under monotone missing/coarsened data

In a R-level monotone missing data model we observe $(C, G_C(Z))$ where $C \in \mathbb{C} := \{1, \ldots, R\}$ and $G_C(Z)$ is such that $G_r(Z) := (Z_1, \ldots, Z_r)$ for $r = 1, \ldots, R$. We work with any given $\lambda \in \Lambda$ where $\Lambda :=$ Power-Set$(\mathbb{C}) \setminus \{$empty set$\}$. We maintain throughout the following assumption.

Assumption A

(A1) The observed data $\{C_i, G_{C_i}(Z_i)\}_{i=1}^N$ are i.i.d. copies of $(C, G_C(Z))$.

(A2) $P(C = r|G_1(Z)) > 0$ for $r = 1, \ldots, R-1$ and $P(C = R|G_1(Z)) > \kappa > 0$ almost surely in $G_1(Z)$.

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

See Devereux and Tripathi (2009) (pages 19-20) for discussion of (A1). $P(C = R|G_1(Z)) > \kappa > 0$ in (A2) is a strict version of the aforementioned overlap assumption [see Khan and Tamer (2010) and Chaudhuri and Hill (2013)]. The restrictions $P(C = r|G_1(Z)) > 0$ for $r = 1, \ldots, R-1$ only help to avoid more involved proofs peripheral to the main message. However $P(C = r) > 0$ for $r = 1, \ldots, R$ is intrinsic to the R-level missing data model. (A3) with $M_\lambda(\beta) := \frac{\partial}{\partial \beta'} E\left[m(Z;\beta)\middle| C \in \lambda\right]$ also works.

Now define the following quantities to be used to express the efficient influence functions whose variances will give the efficiency bounds in the propositions stated below:

$$\begin{aligned}
\varphi_{(1)}(C, G_C(Z); \beta) &:= E[m(Z;\beta)|G_1(Z)], \\
\varphi_{(r)}(C, G_C(Z); \beta) &:= \frac{I(C \geq r)}{P(C \geq r|G_1(Z))} \left(E[m(Z;\beta)|G_r(Z)] - E[m(Z;\beta)|G_{r-1}(Z)]\right),
\end{aligned}$$

for $r = 2, \ldots, R$. Unless confusing, we will drop the argument β from quantities evaluated at $\beta = \beta^0$.

Proposition 1 *Let assumption A and* (2) *hold. For the given $\lambda \in \Lambda$, define*

$$\varphi_\lambda(C, G_C(Z); \beta) := \frac{I(C \in \lambda)}{P(C \in \lambda)} \varphi_{(1)}(C, G_C(Z); \beta) + \frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \sum_{r=2}^{R} \varphi_{(r)}(C, G_C(Z); \beta).$$

Denoting $\varphi_\lambda(C, G_C(Z)) := \varphi_\lambda(C, G_C(Z); \beta^0)$*, assume that* $V_\lambda := Var(\varphi_\lambda(C, G_C(Z)))$ *is a* $d_m \times d_m$ *finite positive definite matrix where, in terms of the full data* (C, Z)*,*

$$V_\lambda = E\left[\frac{P(C \in \lambda|Z_1)}{P^2(C \in \lambda)} E[m(Z)|Z_1]E[m(Z)'|Z_1] + \frac{P^2(C \in \lambda|Z_1)}{P^2(C \in \lambda)} \sum_{r=2}^{R} \frac{Var\left(E[m(Z)|Z_1, \ldots, Z_r]|Z_1, \ldots, Z_{r-1}\right)}{P(C \geq r|Z_1)}\right].$$

Then for β^0 *defined by* (1) *for the given* λ*, the asymptotic variance lower bound for* $\sqrt{N}(\widehat{\beta} - \beta^0)$ *of any regular estimator* $\widehat{\beta}$ *is given by* $\Omega_\lambda := (M_\lambda' V_\lambda^{-1} M_\lambda)^{-1}$*. An estimator whose asymptotic variance equals* Ω_λ *has the asymptotically linear representation*

$$\begin{aligned} \sqrt{N}(\widehat{\beta} - \beta^0) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_\lambda(C_i, G_{C_i}(Z_i)) + o_p(1), \text{ where} \\ \psi_\lambda(C, G_C(Z)) &:= -\Omega_\lambda^{-1} M_\lambda' V_\lambda^{-1} \varphi_\lambda(C, G_C(Z); \beta^0). \end{aligned}$$

Remarks

(i) Proposition 1 extends Theorem 1 of Chen et al. (2008) to multi-level monotone missing/coarsened data and allows the parameter value of interest β^0 to be defined in terms of a variety of subpopulations. In Section 3 we show that this gives new insights on efficiency gains due to additional incomplete subsamples that are unavailable from a two-level missing data model.

(ii) The expression of the efficient influence function $\psi(C, Z)$ emphasizes the incremental contribution of each level of coarsened data, or in other words, each subsample. We revisit this in Section 4. On the other hand, the total contribution of each subsample, revisited in Section 3, can be emphasized by noting that the key term $\varphi_\lambda(C, G_C(Z))$ in $\psi(C, Z)$ also has the standard representation

$$\begin{aligned} \varphi_\lambda(C, G_C(Z); \beta) &= \frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \frac{I(C = R)}{P(C = R|G_1(Z))} m(Z; \beta) \\ &+ \frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \sum_{r=2}^{R-1} \left\{\frac{I(C \geq r)}{P(C \geq r|G_1(Z))} - \frac{I(C \geq r+1)}{P(C \geq r+1|G_1(Z))}\right\} E[m(Z; \beta)|G_r(Z)] \\ &+ \left\{\frac{I(C \in \lambda)}{P(C \in \lambda)} - \frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \frac{I(C \geq 2)}{P(C \geq 2|G_1(Z))}\right\} E[m(Z; \beta)|G_1(Z)] \end{aligned} \quad (3)$$

as a member of the augmented inverse probability weighted (AIPW) class introduced by Robins et al. (1994). The first line of (3) is the IPW part. The last two lines represent the augmentation. AIPW forms generally have beneficial consequences on the estimation of β^0 [see Tsiatis (2006)].

Estimation of β^0 can be done as a standard exercise in semiparametric GMM as follows. Plug in suitable nonparametric estimators $\widehat{P}(C \geq r|G_1(Z))$ and $\widehat{E}[m(Z; \beta)|G_r(Z)]$, for $r = 1, \ldots, R$ in places

of the respective unknown nuisance functions to obtain a feasible version of $\varphi_\lambda(C, G_C(Z); \beta)$ for each β. Denote it by $\widehat{\varphi}_\lambda(C, G_C(Z); \beta)$. A semiparametric GMM estimator of β^0 is then obtained as

$$\widehat{\beta}_\lambda^{GMM}(W) = \arg\min_{\beta \in \mathcal{B}} \widehat{\varphi}_{\lambda,N}(C, G_C(Z); \beta)' W_N \widehat{\varphi}_{\lambda,N}(C, G_C(Z); \beta),$$

where $\widehat{\varphi}_{\lambda,N}(C, G_C(Z); \beta) := \frac{1}{N} \sum_{i=1}^N \widehat{\varphi}_\lambda(C_i, G_{C_i}(Z_i); \beta)$. W_N is some positive semi-definite weighting matrix such that $W_N \xrightarrow{P} W$. Conditions required for the suitable rate of convergence of the first-step nonparametric estimators that would lead to consistency and asymptotic normality of $\widehat{\beta}_\lambda^{GMM}(W)$ and, importantly, not affect the asymptotic variance are clearly stated in Newey (1997), Chen et al. (2008), Cattaneo (2010), Rothe and Firpo (2012), etc. for various choices of nonparametric estimators. (We do not repeat them for brevity.) $\widehat{\beta}_\lambda^{GMM}(W)$ is semiparametrically efficient if $W = V_\lambda^{-1}$.

On the other hand, a parametric GMM estimator of β^0 is obtained if instead one plugs in first-step parametric estimators $\widetilde{P}(C \geq r|G_1(Z))$ and $\widetilde{E}[m(Z;\beta)|G_r(Z)]$, for $r = 1, \ldots, R$. The estimator of β^0 is consistent if either the parametric model for $P(C \geq r|G_1(Z))$ or for $E[m(Z;\beta)|G_r(Z)]$ is correct for $r = 1, \ldots, R$. This is the double robustness property introduced by Robins et al. (1994), Scharfstein et al. (1999), etc.[3] When parametric models for $P(C \geq r|G_1(Z))$ and $E[m(Z;\beta)|G_r(Z)]$ are both correct and $W = V_\lambda^{-1}$, the estimator has asymptotic variance equal to the efficiency bound Ω_λ.

The practice of using parametric models for $P(C \geq r|G_1(Z))$ requires qualification. Similar to Hahn (1998) and Chen et al. (2008), when $\lambda \neq \mathbb{C}$ the efficiency bound is actually less than Ω_λ (in a matrix sense) if $P(C \geq r|G_1(Z))$ is known up to some finite-dimensional parameter, as was required for the local efficiency of the parametric GMM estimator above. It is even lesser if $P(C \geq r|G_1(Z))$ is completely known. This may arise when missingness is by design, as is the case in Sections 3 (partly) and 4. We present the efficiency bounds under these two scenarios in Propositions 2 and 3 below.

Proposition 2 *Let assumption A and* (2) *hold. Assume* $P(C = r|G_1(Z))$ *is known for* $r = 1, \ldots, R$. *Using a subscript* $[k]$ *to represent that* $P(C = r|G_1(Z))$ *is known, and for the given* $\lambda \in \Lambda$, *define*

$$\varphi_{\lambda[k]}(C, G_C(Z); \beta) := \frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \sum_{r=1}^R \varphi_{(r)}(C, G_C(Z); \beta).$$

Denoting $\varphi_{\lambda[k]}(C, G_C(Z)) := \varphi_{\lambda[k]}(C, G_C(Z); \beta^0)$, *assume that* $V_{\lambda[k]} := Var(\varphi_{\lambda[k]}(C, G_C(Z)))$ *is a*

[3] Appendix B.6 of an older version of this paper (available from the author's web page) demonstrates double robustness. See Chaudhuri and Min (2012) for more on parametric GMM estimation where the parameter of interest is defined possibly in terms of subpopulations, but only in a two-level missing data model. Rothe and Firpo (2012) and Chaudhuri and Guilkey (2013) show that even semiparametric GMM estimators based on the doubly robust form of the influence function is asymptotically efficient under milder than standard requirements on the convergence of preliminary nonparametric estimators of the nuisance parameters. See Robins and Ritov (1997) for more general discussion.

$d_m \times d_m$ finite positive definite matrix where, in terms of the full data (C, Z),

$$V_{\lambda[k]} = V_\lambda - \frac{P(C \in \lambda|Z_1)(1 - P(C \in \lambda|Z_1))}{P^2(C \in \lambda)} E[m(Z)|Z_1]E[m(Z)|Z_1]'.$$

Then for β^0 defined by (1) for the given λ, the asymptotic variance lower bound for $\sqrt{N}(\widehat{\beta} - \beta^0)$ of any regular estimator $\widehat{\beta}$ is given by $\Omega_{\lambda[k]} := (M_\lambda' V_{\lambda[k]}^{-1} M_\lambda)^{-1}$. An estimator whose asymptotic variance equals $\Omega_{\lambda[k]}$ has the asymptotically linear representation

$$\begin{aligned} \sqrt{N}(\widehat{\beta} - \beta^0) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_{\lambda[k]}(C_i, G_{C_i}(Z_i)) + o_p(1), \text{ where} \\ \psi_{\lambda[k]}(C, G_C(Z)) &:= -\Omega_{\lambda[k]}^{-1} M_\lambda' V_{\lambda[k]}^{-1} \varphi_{\lambda[k]}(C, G_C(Z); \beta^0). \end{aligned}$$

Remarks

(i) $V_{\lambda[k]}$ shows the improvement over the case where $P(C = r|G_1(Z))$ is unknown.

(ii) There is no improvement when $\lambda = \mathbb{C}$. This is because the identities $I(C \in \mathbb{C}) = 1, P(C \in \mathbb{C}) = 1$ and $P(C \in \mathbb{C}|G_1(Z)) = 1$ imply that $\psi_\lambda(C, G_C(Z)) = \psi_{\lambda[k]}(C, G_C(Z))$.

Proposition 3 *Let assumption A and (2) hold. Assume $P(C = r|G_1(Z)) = P(C = r|G_1(Z); \gamma^0)$ for some $\gamma^0 \in \Gamma \subset \mathbb{R}^{d_\gamma}$ where $P(C = r|G_1(Z); \gamma)$ is known up to the finite-dimensional unknown γ for $r = 1, \ldots, R$. Let $S_\gamma(C|G_1(Z)) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|G_1(Z))} \frac{\partial P(C=r|G_1(Z);\gamma^0)}{\partial \gamma}$ denote the score function for γ evaluated at $\gamma = \gamma^0$, and assume that $E\left[S_\gamma(C|G_1(Z))S_\gamma(C|G_1(Z))'\right]$ is positive definite. Using a subscript $[pk]$ to represent that $P(C = r|G_1(Z))$ is partially known, and for the given $\lambda \in \Lambda$, define*

$$\varphi_{\lambda[pk]}(C, G_C(Z); \beta) := \varphi_{\lambda[k]}(C, G_C(Z); \beta) + Proj\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z; \beta)|G_1(Z)] \middle| S_\gamma(C|G_1(Z)) \right)$$

where, for two random variables Y and X, $Proj(Y|X)$ denotes the population least squares projection of Y on X. Denoting $\varphi_{\lambda[pk]}(C, G_C(Z)) := \varphi_{\lambda[pk]}(C, G_C(Z); \beta^0)$, assume that $V_{\lambda[pk]} := Var(\varphi_{\lambda[pk]}(C, G_C(Z)))$ is a $d_m \times d_m$ finite positive definite matrix where, in terms of the full data (C, Z),

$$\begin{aligned} V_{\lambda[pk]} &= V_{\lambda[k]} + B\left(E\left[S_\gamma(C|Z_1)S_\gamma(C|Z_1)'\right]\right)^{-1} B', \\ &= V_\lambda - Var\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|Z_1] - Proj\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|Z_1] \middle| S_\gamma(C, Z_1) \right) \right); \end{aligned}$$

and $B := E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z; \beta)|G_1(Z)] S_\gamma(C|G_1(Z))'\right] = E\left[\frac{E[m(Z)|Z_1]}{P(C \in \lambda)} \sum_{r \in \lambda} \frac{\partial P(C=r|Z_1;\gamma^0)}{\partial \gamma'}\right]$. Then for β^0 defined by (1) for the given λ, the asymptotic variance lower bound for $\sqrt{N}(\widehat{\beta} - \beta^0)$ of any regular

estimator $\widehat{\beta}$ *is given by* $\Omega_{\lambda[pk]} := (M_\lambda' V_{\lambda[pk]}^{-1} M_\lambda)^{-1}$. *An estimator whose asymptotic variance equals* $\Omega_{\lambda[pk]}$ *has the asymptotically linear representation*

$$\begin{aligned}\sqrt{N}(\widehat{\beta} - \beta^0) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_{\lambda[pk]}(C_i, G_{C_i}(Z_i)) + o_p(1), \text{ where} \\ \psi_{\lambda[pk]}(C, G_C(Z)) &:= -\Omega_{\lambda[pk]}^{-1} M_\lambda' V_{\lambda[pk]}^{-1} \varphi_{\lambda[pk]}(C, G_C(Z); \beta^0).\end{aligned}$$

Remarks

(i) The first line in the expression of $V_{\lambda[pk]}$ shows the loss in efficiency relative to the case where $P(C = r|G_1(Z))$ is completely known. On the other hand, the second line shows the gain in efficiency relative to the case where $P(C = r|G_1(Z))$ is unknown.

(ii) There is no improvement when $\lambda = \mathbb{C}$. In this case, $\psi_{\lambda[pk]}(C, G_C(Z)) = \psi_\lambda(C, G_C(Z))$ since $\text{Proj}\left(\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m(Z)|Z_1] \middle| S_\gamma(C, Z_1)\right) := B\left(E[S_\gamma(C, Z_1) S_\gamma(C, Z_1)']\right)^{-1} S_\gamma(C, Z_1) = 0$. To see this note that the identities $I(C \in \mathbb{C}) = 1$ and $P(C \in \mathbb{C}) = 1$ imply that when $\lambda = \mathbb{C}$,

$$B = E\left[\frac{I(C \in \mathbb{C})}{P(C \in \mathbb{C})} E[m(Z)|Z_1] S_\gamma(C, Z_1)'\right] = E\left[E[m(Z)|Z_1] E[S_\gamma(C, Z_1)|Z_1]'\right] = 0$$

because, by definition of conditional score, $E[S_\gamma(C, Z_1)|Z_1] = 0$. Hence, $V_\lambda = V_{\lambda[k]} = V_{\lambda[pk]}$.

3 Optimally using multiple subsamples for efficiency gains

In this section we consider the simple case of $R = 3$ to demonstrate the possible efficiency gains from optimally using all the subsamples (as opposed to leaving out some) for estimation of β^0 under all choices of the target $\lambda \in \Lambda := \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \mathbb{C} = \{1, 2, 3\}\}$ in (1).

Let us first present two categories of toy examples where such considerations of efficiency gains may be of practical interest. These examples are simple extensions of those presented in Chen et al. (2004) who also provide the references for each example. We work with the second category for the sake of concreteness of demonstration in this section. The first category is revisited in Section 4.

Toy Example 1: Examples under this category have the common characteristic that the availability of all units or all relevant information on each unit for the study happens selectively in phases possibly under time and budget considerations. Possible incompleteness in subsamples is by design and hence under the surveyor's control. The rationale for incompleteness is appreciated, for example, under the premise that the cost of collecting Z_1 is less than Z_2, which is less than Z_3.

Toy Example 1A: Similar to Example 1 in Chen et al. (2004), suppose that we have a sample of

N units with only Z_1. A stratified sampling procedure can be used for multiphase collection of Z_2 and Z_3 as follows: Let $U_i \overset{\text{i.i.d.}}{\sim} U(0,1)$ be independent of Z_{1i}, Z_{2i}, Z_{3i} for all $i = 1, \ldots, N$. For the i-th unit collect Z_2 if $U_i \leq p_2(Z_{1i}) + p_3(Z_{1i})$, and Z_3 if $U_i \leq p_3(Z_{1i})$ where $p_2(Z_1), p_3(Z_1) \in (0,1)$ are measurable functions of Z_1, and $p_2(Z_1)+p_3(Z_1) < 1$. Taking $m(Z;\beta) = m(Z_2, Z_3; \beta)$ with target $\lambda = \mathbb{C}$ is a direct extension of Chen et al. (2004)'s Example 1. Alternatively, in multiphase experiments, Z_2 can be a treatment (randomized as above) and its outcome; and same for Z_3. However, in such cases, randomization for Z_3 generally depends on Z_2 in ways that make version [A] of (2) more appropriate than version (2)[B]. Both examples have a flavor of variable probability multiphase sampling and, in principle, have similar pros and cons as double sampling methods [see Cochran (1977), Chapter 12].

Toy Example 1B: Similar to Example 4 in Chen et al. (2004), consider a survey with three waves where Z_2 was revised (or included in the survey) starting from the second wave while Z_3 from the third wave. Z_1 is known from all waves. Example 5 of Chen et al. (2004) is also related if instead the first wave is some large scale survey such as census with information on Z_1 (for example, demographics) only, while the second and third waves are small scale household surveys with, respectively, additional information on Z_2 (for example, consumption), and on Z_2 and Z_3 (for example, income/wealth). Chen et al. (2004) discuss identification that follows from comparability of the surveys/waves such that assumptions similar to (2) hold. Our results then allow for any λ as target.

Toy Example 1C: Combination of the above two is possible. Let Z_1, Z_2 and Z_3 be observed at the beginning of time periods 1, 2 and 3 of a *unit's tenure in a study* that requires 3 periods to be done. Suppose the study starts with N_3 individuals randomized based on Z_1 by, for example, variable probability sampling [see Wooldridge (1999)]. Subsequent availability of funds allows to add, with randomization based on their Z_1, N_2 individuals at the beginning of period 2 and N_1 individuals at the beginning of period 3. If no units leave, at the beginning of period 3 from the start of the study the researcher observes only Z_1 for N_1 units who started in period 3, only (Z_1, Z_2) for N_2 individuals who started in period 2, and all (Z_1, Z_2, Z_3) who started in period 1. The former two can be thought of as refreshment samples, however, not in the sense of Hirano et al. (2001) who study panel attrition based on both observables (MAR) and unobservables. Similar to Example 3 in Chen et al. (2004) one could also consider attrition (with no return) in waves two and three in a panel. Although in this case (2)[B] is not appropriate since it does not recognize the dynamic nature of attrition. Instead, application of MAR typically requires version [A] of (2) or some explicit sequential version of MAR. See Appendix B or Robins et al. (1995), Robins and Rotnitzky (1995), Fitzgerald et al. (1998), etc.

Toy Example 2: In the examples under Toy example 1, incompleteness of the subsamples was by

the survey design. However, incompleteness often happens due to the unit's (agent's) choice. Consider one such scenario similar to Example 2 of Chen et al. (2004). Suppose we have a random sample of N units. Of them, N_1 do not have a college education ($C = 1$), N_2 went to a public college ($C = 2$), while $N_3 = N - N_1 - N_2$ went to a private college ($C = 3$). We observe demographic and other pre-college variables (denote them all by Z_1) and wages (Y) of all N units. We observe college grades (Z_2) for all $N - N_1$ units who went to college. Let the counterfactuals: $Y(C = 1)$, $Y(C = 2)$, and $Y(C = 3)$ representing the potential wages with no college, public college and private college education respectively, be well defined for all units. The observed wage is: $Y = \sum_{j=1}^{3} I(C = j)Y(C = j)$.

The parameter of interest can be $\beta^0 := E[Y(C = 3)|C \in \lambda]$ for any target λ. β^0 is defined by (1) when $m(Z; \beta) = Y(C = 3) - \beta$. Taking $Z_3 = Y(C = 3)$ (and ignoring $Y(C = 2)$ and $Y(C = 1)$) results in the monotone missing data. (2) identifies β^0 based on the observed data as:

$$\beta^0 = E\left[\frac{P(C \in \lambda|Z_1)}{P(C \in \lambda)} \frac{I(C = 3)}{P(C = 3|Z_1)} Y\right].$$

However, we argue that if MAR (2) is being used for identification of β^0, it is imperative that we also extract all the information that becomes available from the subsamples by virtue of MAR and use that for efficiency gains. One source of such information is the observed post-treatment variables (Z_2).

In the rest of this section we use Proposition 1 to demonstrate such efficiency gains, point out when they are possible and when not, under the premise of Toy Example 2.

For example, let $\lambda = \{1\}$. So the parameter of interest β^0 is the expected wage of those who did not go to college if they had gone to private college. Consider two estimators: (i) the efficient estimator $\widehat{\beta}^{\mathbb{C}}_{\lambda=\{1\}}$ based on the N observations from the entire observed data (i.e., all the subsamples), and (ii) the efficient estimator $\widehat{\beta}^{\{1,3\}}_{\lambda=\{1\}}$ based on $N_1 + N_3$ observations from the *suboptimal choice* of two subsamples ($C = 1$) and ($C = 3$), i.e., the subsample from the target and the complete subsample. The estimators are indexed by the target population and the population for the used subamples. Using Proposition 1 with $R = 3$ for $\widehat{\beta}^{\mathbb{C}}_{\lambda=\{1\}}$, and Theorem 1 of Chen et al. (2008) for $\widehat{\beta}^{\{1,3\}}_{\lambda=\{1\}}$ gives:

$$\widehat{\beta}^{\mathbb{C}}_{\lambda=\{1\}} \overset{\text{Asym.}}{\sim} \text{Normal}\left(\beta^0, \frac{\Omega^{\mathbb{C}}_{\lambda=\{1\}}}{N}\right) \text{ where } \Omega^{\mathbb{C}}_{\lambda=\{1\}} := \left(M'_{\lambda=\{1\}} \left(V^{\mathbb{C}}_{\lambda=\{1\}}\right)^{-1} M_{\lambda=\{1\}}\right)^{-1},$$

$$\widehat{\beta}^{\{1,3\}}_{\lambda=\{1\}} \overset{\text{Asym.}}{\sim} \text{Normal}\left(\beta^0, \frac{\Omega^{\{1,3\}}_{\lambda=\{1\}}}{N_1 + N_3}\right) \text{ where } \Omega^{\{1,3\}}_{\lambda=\{1\}} := \left(M'_{\lambda=\{1\}} \left(V^{\{1,3\}}_{\lambda=\{1\}}\right)^{-1} M_{\lambda=\{1\}}\right)^{-1}$$

where $V^{\mathbb{C}}_{\lambda=\{1\}} = V_{\lambda=\{1\}}$ defined in Proposition 1 ($R = 3$), while $V^{\{1,3\}}_{\lambda=\{1\}}$ is defined in Corollary 4.

Application of Chen et al. (2008) generally depends on the substitution pattern, in the present context, between no college education ($C = 1$) and private college ($C = 3$) when the option of public college ($C = 2$) is unavailable. However, our focus is on the less involved problem of using versus not using a particular subsample (say, ($C = 2$)) for estimation. This is emphasized in the formulae above by stating the size of the used sample for estimation. Hence, Theorem 1 of Chen et al. (2008) or, equivalently, our Proposition 1 with $R = 2$ and the original probability distribution, but conditional on $C \in \{1, 3\}$, give $\Omega^{\{1,3\}}_{\lambda=\{1\}}$. This is similar to proportional substitution in the unit's (agent's) choice.

Under the maintained assumptions, $\left(\Omega^{\{1,3\}}_{\lambda=\{1\}}/(N_1+N_3)\right)^{-1}\left(\Omega^{\{1,3\}}_{\lambda=\{1\}}/(N_1+N_3) - \Omega^{\mathbb{C}}_{\lambda=\{1\}}/N\right)$ is positive semidefinite if and only if $\left(V^{\{1,3\}}_{\lambda=\{1\}}/(N_1+N_3)\right)^{-1}\left(V^{\{1,3\}}_{\lambda=\{1\}}/(N_1+N_3) - V^{\mathbb{C}}_{\lambda=\{1\}}/N\right)$ is positive semidefinite. We work with the latter for simplicity.

Also for simplicity of exposition, let $d_m = d_\beta = 1$. Note that while $\widehat{\beta}^{\{1,3\}}_{\lambda=\{1\}}$ is based on fewer observations $(N_1 + N_3)$ instead of N; the assumptions of proportional substitution, (2) and A(2) do not give a proportional loss in efficiency, i.e., it should not be expected that:

$$\begin{aligned}
\lim_{N\to\infty} \Delta^{\{1,3\} \text{ v/s } \mathbb{C}}_{\lambda=\{1\}}(N) &= 1 - P(C \in \{1,3\}) = P(C = 2) \text{ [see, for example, Corollary 4(b)]} \\
\text{where } \Delta^{\{1,3\} \text{ v/s } \mathbb{C}}_{\lambda=\{1\}}(N) &:= \left(V^{\{1,3\}}_{\lambda=\{1\}}/(N_1+N_3)\right)^{-1}\left(V^{\{1,3\}}_{\lambda=\{1\}}/(N_1+N_3) - V^{\mathbb{C}}_{\lambda=\{1\}}/N\right).
\end{aligned}$$

In the context of the present example, $\lim_{N\to\infty} \Delta^{\{1,3\} \text{ v/s } \mathbb{C}}_{\lambda=\{1\}}(N)$ captures the (scaled) relevant information in the unused subsample ($C = 2$). (What we mean by relevant/key information will be clear from the expressions of efficiency loss given in the corollaries below.) This is how we define the loss in efficiency in the current context from using $\widehat{\beta}^{\{1,3\}}_{\lambda=\{1\}}$ instead of $\widehat{\beta}^{\mathbb{C}}_{\lambda=\{1\}}$.

More generally, while comparing other estimators for other choices of λ, a similar quantity is used to capture the relevant unused information and hence the efficiency loss. The candidates for comparisons and the choice of λ are respectively stated in superscripts and subscripts of the concerned expressions.

We demonstrate such efficiency losses (when they occur) for various choices of λ in the form of corollaries in Corollary 4 – 9. The proofs use Proposition 1, and are briefly stated in the Appendix. Details can be found in an older version of this paper available from the author's web page.

Corollary 4 *Let $\lambda = \{1\}$. The following hold under the assumptions of Proposition 1:*

(a) $V^{\{1,3\}}_{\lambda=\{1\}} = \frac{P(C\in\{1,3\})}{P(C=1)} E\left[E[m(Z)|Z_1]E[m(Z)'|Z_1] + \frac{P(C=1|Z_1)}{P(C=3|Z_1)} Var(m(Z)|Z_1)\middle| C = 1\right]$.

(b) $\lim_{N\to\infty} \Delta^{\{1,3\} \text{ v/s } \mathbb{C}}_{\lambda=\{1\}}(N) = \frac{P(C\in\{1,3\})}{V^{\{1,3\}}_{\lambda=\{1\}}} E\left[\frac{P^2(C=1|Z_1)}{P^2(C=1)} \frac{P(C=2|Z_1)}{P(C=3|Z_1)P(C\geq 2|Z_1)} Var(E[m(Z)|Z_1, Z_2]|Z_1)\right]$.

Corollary 5 *Let $\lambda = \{2\}$. The following hold under the assumptions of Proposition 1:*

(a) $V^{\{2,3\}}_{\lambda=\{2\}} = \frac{P(C\in\{2,3\})}{P(C=2)} E\left[E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2] + \frac{P(C=2|Z_1)}{P(C=3|Z_1)} Var(m(Z)|Z_1,Z_2)\Big| C=2\right]$.

(b) $\lim_{N\to\infty} \Delta^{\{2,3\}\ v/s\ \mathbb{C}}_{\lambda=\{2\}}(N) = \frac{P(C\in\{2,3\})}{V^{\{2,3\}}_{\lambda=\{2\}}} E\left[\frac{P^2(C=2|Z_1)}{P^2(C=2)} \frac{P(C=3|Z_1)}{P(C=2|Z_1)P(C\geq 2|Z_1)} Var(E[m(Z)|Z_1,Z_2]|Z_1)\right]$.

Corollary 6 *Let $\lambda = \{3\}$. The following hold under the assumptions of Proposition 1:*

(a) $V^{\{3\}}_{\lambda=\{3\}} = E[m(Z)m(Z)'|C=3]$.

(b) $\lim_{N\to\infty} \Delta^{\{3\}\ v/s\ \{1,3\}}_{\lambda=\{3\}}(N) = \lim_{N\to\infty} \Delta^{\{3\}\ v/s\ \{2,3\}}_{\lambda=\{3\}}(N) = 0$.

(c) $\lim_{N\to\infty} \Delta^{\{3\}\ v/s\ \mathbb{C}}_{\lambda=\{3\}}(N) = \frac{P(C\in\{3\})}{V^{\{3\}}_{\lambda=\{3\}}} E\left[\frac{P(C=3|Z_1)}{P^2(C=3)} \frac{P(C=2|Z_1)}{P(C\geq 2|Z_1)} Var(E[m(Z)|Z_1,Z_2]|Z_1)\right]$.

Remarks:

(i) As can be seen from Corollaries 4(b), 5(b) and 6(c), the key information lost from not using all the subsamples for estimation is $Var(E[m(Z)|Z_1,Z_2]|Z_1)$. The resulting efficiency loss is a scaled expectation of this quantity, expressed in the corollaries as further scaled by the variance (note that $d_m = d_\beta = 1$) of the concerned estimator from the suboptimal choice of subsamples.

(ii) Corollary 6(b) shows that when $\lambda = \{3\}$, there is no efficiency loss from using only $(C=3)$, i.e., the complete subsample, instead of using jointly ($(C=1)$ and $(C=3)$) or ($(C=2)$ and $(C=3)$). The result corresponds to a two-level missing data model for which efficiency gains/losses have not been traditionally (and for the above reason, justifiably) considered when $\lambda = \{3\}$.[4] On the other hand, Corollary 6(c) shows that efficiency loss happens under more than two-level missingness if any subsample is unused. This insight becomes available only by going beyond two-level missingness.

Now let us consider cases where the target is heterogeneous in terms of observability of Z_1, Z_2 and Z_3.

Corollary 7 *Let $\lambda = \{1,3\}$. The following hold under the assumptions of Proposition 1:*

(a) $V^{\{1,3\}}_{\lambda=\{1,3\}} = E\left[E[m(Z)|Z_1]E[m(Z)'|Z_1] + Var(m(Z)|Z_1)\right| C\in\{1,3\}]$.

(b) $\lim_{N\to\infty} \Delta^{\{1,3\}\ v/s\ \mathbb{C}}_{\lambda=\{1,3\}}(N) = \frac{P(C\in\{1,3\})}{V^{\{1,3\}}_{\lambda=\{1,3\}}} E\left[\frac{P^2(C=\{1,3\}|Z_1)}{P^2(C=\{1,3\})} \frac{P(C=2|Z_1)}{P(C=3|Z_1)P(C\geq 2|Z_1)} Var(E[m(Z)|Z_1,Z_2]|Z_1)\right]$.

Corollary 8 *Let $\lambda = \{2,3\}$. The following hold under the assumptions of Proposition 1:*

(a) $V^{\{2,3\}}_{\lambda=\{2,3\}} = E\left[E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2] + Var(m(Z)|Z_1,Z_2)\right| C\in\{2,3\}]$.

(b) $\lim_{N\to\infty} \Delta^{\{2,3\}\ v/s\ \mathbb{C}}_{\lambda=\{2,3\}}(N) = 0$.

Corollary 9 *Let $\lambda = \mathbb{C} := \{1,2,3\}$. The following hold under the assumptions of Proposition 1:*

[4] For the cases where $\lambda = \{1\}$ or $\lambda = \{2\}$, efficiency loss from using only $(C=3)$ in two-level MAR missing data models is well documented in the literature through the AIPW form similar to (3) of the efficient influence functions. Hence in Corollaries 4 and 5 we do not consider $V^{\{3\}}_{\lambda=\{1\}}$ and $V^{\{3\}}_{\lambda=\{2\}}$ respectively and avoid redundancy in the demonstration.

(a) $V_{\lambda=\mathbb{C}}^{\{1,3\}} = P(C \in \{1,3\})E\left[\frac{Var(m(Z)|Z_1)}{P(C=3|Z_1)} + \frac{E[m(Z)|Z_1]E[m(Z)'|Z_1]}{P(C\in\{1,3\}|Z_1)}\right]$.

(b) $\lim_{N\to\infty} \Delta_{\lambda=\mathbb{C}}^{\{1,3\}\ v/s\ \mathbb{C}}(N) = \frac{P(C\in\{1,3\})}{V_{\lambda=\mathbb{C}}^{\{1,3\}}}(J_1 + J_2)$ *where*

$$J_1 := E\left[\frac{P(C=2|Z_1)Var(E[m(Z)|Z_1,Z_2]|Z_1)}{P(C=3|Z_1)P(C\geq 2|Z_1)}\right] \text{ and } J_2 := E\left[\frac{P(C=2|Z_1)E[m(Z)|Z_1]E[m(Z)'|Z_1]}{P(C\in\{1,3\}|Z_1)}\right].$$

(c) $V_{\lambda=\mathbb{C}}^{\{2,3\}} = P(C \in \{2,3\})E\left[\frac{Var(m(Z)|Z_1,Z_2)}{P(C=3|Z_1)} + \frac{E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2]}{P(C\in\{2,3\}|Z_1)}\right]$.

(d) $\lim_{N\to\infty} \Delta_{\lambda=\mathbb{C}}^{\{2,3\}\ v/s\ \mathbb{C}}(N) = \frac{P(C\in\{2,3\})}{V_{\lambda=\mathbb{C}}^{\{2,3\}}} E\left[\frac{P(C=1|Z_1)}{P(C\geq 2|Z_1)}E[m(Z)|Z_1]E[m(Z)'|Z_1]\right]$.

Remarks:

(i) Corollary 7 is similar to Corollary 4, the main difference in the expressions for the loss being the scale, i.e., $\frac{P^2(C=\{1,3\}|Z_1)}{P^2(C=\{1,3\})}$ as opposed to $\frac{P^2(C=\{1\}|Z_1)}{P^2(C=\{1\})}$, that is related to the concerned target λ.

(ii) Corollary 8 is similar to Corollary 6(a,b). There is no loss in efficiency from leaving out the subsample $(C = 1)$ instead of using the entire observed data.

(iii) Corollary 9 (a) and (c) are fundamentally different from the rest. The target population *strictly* contains the union of the subpopulations from which the subsamples used for estimation in these two cases are drawn. The efficiency loss from using the subsamples $(C = 1)$ and $(C = 3)$ instead of the entire observed data is presented in Corollary 9(b). It has two (additive) components: The component with J_1 is similar to the efficiency loss discussed so far (for example, in Corollaries 4(b) and 7(b)). However, the component with J_2 is new; it is the additional penalty from not using the subsample $(C = 2)$ that comes from a subpopulation of the target $\lambda = \mathbb{C}$. Similarly, the only efficiency loss in Corollary 9(d) is also due to a similar penalty from not using the subsample $(C = 1)$ drawn from a subpopulation of the target $\lambda = \mathbb{C}$ (compare with Corollary 8(b) where there is no loss). The insight of this additional penalty was unavailable from the other corollaries.

4 Each subsample through an additional moment restriction

In this section we study the contribution of each subsample toward efficiency gains in an alternative way. Using MAR in (2) we express the moment restrictions in (1) in terms of a function based on the complete subsample $(C = R)$. The incremental information in each incomplete subsample $(C = r) : r < R$, that is usable due to MAR, is then captured by an additional moment restriction.

The following functions $\phi_\lambda^{R-r}(C \geq R - r, G_{R-r}(Z); \beta)$ for $r = 0, 1 \ldots, R-1$ will be used for constructing the aforementioned set of moment restrictions. We maintain the conventions that $I(C \geq$

$R) \equiv I(C = R)$ and $I(C \geq 1) \equiv 1$. Define, for $r = 0$, the $d_m \times 1$ dimensional function:

$$\phi_\lambda^R(C \geq R, G_R(Z); \beta) \quad := \quad \frac{P(C \in \lambda | G_1(Z))}{P(C \in \lambda)} \frac{I(C = R)}{P(C = R | G_1(Z))} m(G_R(Z) \equiv Z; \beta).$$

For each $r = 1, \ldots, R-1$ define, $\phi_\lambda^{R-r}(C \geq R - r, G_{R-r}(Z); \beta) \equiv \phi^{R-r}(C \geq R - r, G_{R-r}(Z))$ where

$$\phi^{R-r}(C \geq R - r, G_{R-r}(Z)) := I(C \geq R - r) \left[I(C \geq R - r + 1) - P(C \geq R - r + 1 | C \geq R - r, G_{R-r}(Z)) \right]$$

does not depend on β or λ. Both β and λ are henceforth dropped from its arguments.

Now consider the following $d_m + (R-1)$ moment restrictions:

$$E\left[\phi_\lambda^R(C \geq R, G_R(Z); \beta)\right] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0, \tag{4}$$

$$E\left[\phi^{R-r}(C \geq R - r, G_{R-r}(Z)) | G_{R-r}(Z)\right] = 0 \text{ almost surely } G_{R-r}(Z) \text{ for } r = 1, \ldots, R-1. \tag{5}$$

Assumptions (2) and A(2) imply that (4) holds if and only if (1) holds. Call (4) the main moment restrictions. The $R-1$ conditional moment restrictions in (5) do not involve β or λ. Call them the auxiliary moment restrictions. (5) may contain additional relevant information not contained in (4).

Brown and Newey (1998) [equation 15] discuss a general method of using such additional information in the form of auxiliary restrictions for efficiency gains in the estimation of β^0. See Graham (2011) [Theorem 2.1] for an important application of this method when $R = 2$ and $\lambda = \mathbb{C}$.

Our setup is a bit more involved because the conditioning set $G_{R-r}(Z)$ is different in each restriction (5).[5] This is required to make precise the contribution of each incomplete subsample, i.e., each level of coarsened data to the estimation of β^0 through each auxiliary moment restriction. However, the *monotonicity* of the conditioning sets in (5), i.e., $G_{R-r}(Z) \setminus G_{R-r-1}(Z) = Z_{R-r}$ for $r = 1, \ldots, R-1$, facilitates the use of the additional information in a Frisch-Waugh-Lovell type sequential application of the method discussed in Brown and Newey (1998). This is shown in Proposition 10 below.

For any $r = 1, \ldots, R-1$, denoting $\phi^r(C \geq r, G_r(Z))$ by ϕ^r, define

$$\begin{aligned} \overline{\text{Proj}}_{G_r}\left(Y | \, \phi^r\right) &:= Y - \text{Proj}_{G_r}\left(Y | \, \phi^r\right), \text{ where} \\ \text{Proj}_{G_r}\left(Y | \, \phi^r\right) &:= E\left[Y \phi^r | G_r(Z)\right] \left(E\left[\phi^r \phi^r | G_r(Z)\right]\right)^{-1} \phi^r \end{aligned}$$

[5] The convenient version of MAR, i.e., (2)[B] implies that the auxiliary moment restrictions indeed hold for the same conditioning set $G_1(Z)$. However, it is important that the larger conditioning sets be maintained for each $r = 1, \ldots, R-1$, so that the demonstration below is not ad-hoc but adheres to an efficiency benchmark already established in Section 2.

for any random variable Y such that the conditional expectations exist.

Proposition 10 *Under assumptions* (2) *and A(2),* $\varphi_{\lambda[k]}(C, G_C(Z); \beta)$ *defined in the statement of Proposition 2 satisfies the following relationships:*

$$\begin{aligned}\varphi_{\lambda[k]}(C, G_C(Z); \beta) &= \overline{Proj}_{G_1}\left(\overline{Proj}_{G_2}\left(\ldots \overline{Proj}_{G_{R-2}}\left(\overline{Proj}_{G_{R-1}}\left(\phi^R_\lambda(\beta)\middle|\, \phi^{R-1}\right)\middle|\, \phi^{R-2}\right)\ldots\middle|\, \phi^2\right)\middle|\, \phi^1\right) \\ &= \overline{Proj}_{G_{R-1}}\left(\overline{Proj}_{G_{R-2}}\left(\ldots \overline{Proj}_{G_2}\left(\overline{Proj}_{G_1}\left(\phi^R_\lambda(\beta)\middle|\, \phi^1\right)\middle|\, \phi^2\right)\ldots\middle|\, \phi^{R-2}\right)\middle|\, \phi^{R-1}\right).\end{aligned}$$

Remark: Now we discuss two implications of Proposition 10 after recalling the following well known relationship from Propositions 1–3 for the most often studied scenario $\lambda = \mathbb{C}$:

$$\varphi_{\lambda=\mathbb{C},[k]}(C, G_C(Z); \beta) = \varphi_{\lambda=\mathbb{C}}(C, G_C(Z); \beta) = \varphi_{\lambda=\mathbb{C},[pk]}(C, G_C(Z); \beta). \tag{6}$$

Hence, although the focus of this section is on the case where $P(C = r|G_r(Z))$ is known, a broader appeal to cases with unknown or partially known $P(C = r|G_r(Z))$ is possible by the above relation.

The main implication of Proposition 10, in the context of our paper, is on the rationale behind possible sampling strategies. We discuss this now. Recall that, of the two categories of toy examples, the demonstration in Section 3 focused on the second one. In this section, the discussion focuses on the examples under Toy Example 1. All of these examples have the common flavor that the incompleteness of the subsamples, possibly for cost reduction, is by design. Hence, unless otherwise stated, we maintain that $P(C = r|G_r(Z))$, as constructed in Toy Example 1A, is known to the surveyor/researcher.

As before, let the cost of collecting $Z_1, Z_2, \ldots, Z_R$ increase progressively from Z_1 to Z_R. We will maintain that there are R populations that can possibly differ only on Z_1, but not on the distribution of $Z_R, \ldots, Z_2$ conditional on Z_1. This is stronger than version [B] of (2), but helps to discuss the second equality of Proposition 10. There are now two ways, represented by the two equalities of the proposition, to appreciate its implications on sampling strategies similar to those in Toy Example 1.

Consider the first equality. Suppose that observations on $Z_1, \ldots, Z_R$ are available for units from one population. Denote these units by $C = R$. Efficient estimation of β^0 is then based on the moment restrictions in (4). Suppose an additional set of units, for which $Z_1, \ldots, Z_{R-1}$ are observed, becomes available from another population. Denote them by $C = R - 1$. The residual from the innermost projection in the first equality, i.e., $\overline{\text{Proj}}_{G_{R-1}}\left(\phi^R_\lambda(\beta)\middle|\, \phi^{R-1}\right)$, provides the revision in the estimating equations and, hence, the resulting efficiency gain that happens due to the availability of the second subsample. Similarly, suppose a third set of units, for which $Z_1, \ldots, Z_{R-2}$ are observed,

becomes available from another population. Denote them by $C = R-2$. The residual from the second innermost projection in the first equality, i.e., $\overline{\text{Proj}}_{G_{R-2}}\left(\overline{\text{Proj}}_{G_{R-1}}\left(\phi_\lambda^R(\beta)\middle|\,\phi^{R-1}\right)\middle|\,\phi^{R-2}\right)$, provides the revision. Such revisions can be continued accordingly upon the successive availability of units by doing the successive projections. (See the proof of Proposition 10 for the explicit form of the revisions.)

Now consider the second equality. As before, suppose the units ($C = R$) are available. Efficient estimation of β^0 is then based on the moment restrictions in (4). Now suppose an additional set of units, for which only Z_1 is observed, is also available possibly from the other $R-1$ populations. The residual from the innermost projection in the second equality, i.e., $\overline{\text{Proj}}_{G_1}\left(\phi_\lambda^R(\beta)\middle|\,\phi^1\right)$, provides the revision. Now suppose the surveyor decides to enrich the units with $C = 2, \ldots, R-1$ by collecting Z_2. Then the residual from the second innermost projection in the second equality, i.e., $\overline{\text{Proj}}_{G_2}\left(\overline{\text{Proj}}_{G_1}\left(\phi_\lambda^R(\beta)\middle|\,\phi^1\right)\middle|\,\phi^2\right)$, provides the revision. If the units with $C = 3, \ldots, R-1$ are further enriched by collecting Z_3, then the residual from the third innermost projection in the second equality, i.e., $\overline{\text{Proj}}_{G_3}\left(\overline{\text{Proj}}_{G_2}\left(\overline{\text{Proj}}_{G_1}\left(\phi_\lambda^R(\beta)\middle|\,\phi^1\right)\middle|\,\phi^2\right)\middle|\,\phi^3\right)$, provides the revision. Such revisions can continue.

Importantly, in both equalities, the ordering of the projections does not matter due to monotonicity. Hence, for a given set of successively available units (equality 1) or enrichment of units (equality 2), the respective revisions are invariant to the ordering. This can be convenient for the surveyor in the consideration of the tradeoff between efficiency gains and the cost of survey while designing the survey.

The second implication of Proposition 10 is similar to the discussion in Graham (2011) who studies the case where $\lambda = \mathbb{C}$ and $R = 2$. Plugging in the known $P(C = R|G_1(Z))$ in (4) does not lead to efficient estimation for a general λ and R. The MAR assumption in (2) contains more information. Proposition 10 shows that this information can be accounted for by the auxiliary moment restrictions in (5) so that the efficiency bound in Proposition 2 is reached. Therefore, the moment restrictions in (4) and (5) exhaust all the information contained in the setup of Proposition 2. When $\lambda = \mathbb{C}$, the relationship in (6) broadens this implication to the cases where $P(C = r|G_1(Z))$ is completely unknown (Proposition 1) or is known up to a finite dimensional parameter (Proposition 3).

5 Conclusion

We discussed efficient estimation when the observed data is a collection of multiple subsamples all but one of which are incomplete following a monotone pattern. Such subsamples can be due to the agent's choice or by survey design. Our identifying assumption, MAR, is easier to justify for the latter.

The results on efficiency bounds allowed explicit demonstration of possible efficiency gains due

to the use of particular subsamples for estimation. Apart from its implications on possible sampling strategies that we discussed in this paper, these results could also be useful for empirical work. For example, since the incomplete subsamples are not required under MAR for consistent estimation, it can be tempting, and often justifiably so, to not use all/some of them because: (i) computational difficulty increases, and a stronger version of MAR is required with the inclusion of each incomplete subsample, (ii) subsamples with few observations may adversely affect the finite sample properties of the estimator. Our results are asymptotic and hence not useful for (ii). However, an informed judgement of whether to use a particular incomplete subsample for estimation in relation to (i) is aided by the demonstration of the actual information contained in each subsample that becomes usable by virtue of MAR.

References

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

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. (2013). GMM with Multiple Missing Variables. Technical report, University of North Carolina, Chapel Hill.

Chaudhuri, S. and Hill, J. B. (2013). Robust Estimation of Average Treatment Effect. Technical report, University of North Carolina, Chapel Hill.

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

Chen, X., 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. (2004). Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects. Mimeo.

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

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

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

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. In *Working Paper Series*. NBER.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997). Coarsening at Random: Characterizations, Conjectures and Counterexamples. In Lin, D. Y. and Fleming, T. R., editors, *Proceedings of The First Seattle Symposium in Biostatistics: Survival Analysis*, Lecture Notes in Statitsics, pages 255–294. New York: Springer-Verlag.

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

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

Heckman, J. and Vytlacil, E. (2007). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume VIB, chapter 71, pages 4875–5144. Elsevier Science Publisher.

Heitjan, D. F. and Rubin, D. B. (1991). Igonaribility and Coarse Data. *Annals of Statistics*, 19: 2244–2253.

Hirano, K., Imbens, G., Ridder, G., and Rubin, D. (2001). Attrition and Refreshment Samples. *Econometrica*, 69: 1645–1659.

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

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics*, 86: 4–29.

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

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

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

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

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

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

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

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

Rothe, C. and Firpo, S. (2012). Semiparametric Estimation and Inference Using doubly-Robust Moment Conditions. Mimeo.

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

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

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

Appendix A: Proof of the stated results

We use f and F to denote the density and distribution functions, and the concerned random variables are specified inside parentheses. We use $L_0^2(F)$ to denote the space of mean-zero, square integrable functions with respect to F. (C, Z) denotes the full data and $(C, G_C(Z))$ the observed data. In the sequel we always try to switch to notations in terms of the full data unless confusing.

Proof of Proposition 1:

We work with a generic $\lambda \in \Lambda$. We need to use the relationship (8) extensively. The proof consists of three steps that closely follow Chen et al. (2008) and hence the references therein. In the first step we characterize the tangent set for all regular parametric submodels satisfying the semiparametric assumptions on the observed data. In the second step we conjecture the form of the efficient influence function proving pathwise differentiability of β^0 and verifying that the efficient influence function lies in the tangent set. In the last step, we obtain the efficiency bound as the expectation of the outer product of the efficient influence function.

STEP - 1: Consider a regular parametric sub-model indexed by a finite-dimensional parameter θ for the joint distribution of the observed data $(C, G_C(Z))$. Recall that $C \in \mathbb{C} := (1, \ldots, R)$ and $G_r(Z) := (Z_1, \ldots, Z_r)$ for $r = 1, \ldots, R$ (meaning $G_r(Z) \setminus G_{r-1}(Z) = Z_r$). So the log of joint density of the observed data can be expressed in terms of the full data (C, Z) as

$$\begin{aligned} \log f_\theta(C, G_C(Z)) &= \log f_\theta(Z_1) + \sum_{r=1}^{R} I(C = r) \log P_\theta(C = r|Z_1) \\ &\quad + \sum_{r=2}^{R} I(C \geq r) \log f_\theta(Z_r|Z_1, \ldots, Z_{r-1}) \end{aligned}$$

where the first term in the last line uses assumption (2). θ_0 is the unique value of θ such that $f_{\theta_0}(C, G_C(Z))$ equals the true $f(C, G_C(Z))$, and accordingly for all the quantities. The score function with respect to θ can be written in terms of (C, Z) as

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

where $\dot{P}_\theta(C = r|Z_1) := \frac{\partial}{\partial\theta} P_\theta(C = r|Z_1)$, $s_\theta(Z_1) := \frac{\partial}{\partial\theta} \log f_\theta(Z_1)$ and $s_\theta(Z_r|Z_1, \ldots, Z_{r-1}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_r|Z_1, \ldots, Z_{r-1})$. Henceforth, we omit the subscript θ from the quantities evaluated at $\theta = \theta_0$.

Denoting all functions by $a(.)$ or $b(.)$ with arguments in parentheses to avoid introducing too many

notations, the tangent set for the model can be characterized by functions of the form:

$$\mathcal{T} := a(Z_1) + \sum_{r=1}^{R} I(C = r)\frac{b_r(Z_1)}{a_r(Z_1)} + \sum_{r=2}^{R} I(C \geq r)a(Z_1, \ldots, Z_r), \tag{7}$$

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

Unlike Chen et al. (2008) we use the same factorization of the joint density of $(C, G_C(Z))$ for all λ. For a given $\lambda \in \Lambda$, the following relation obtained by two different factorization of the joint distribution of $(I(C \in \lambda), G_1(Z)) \equiv (I(C \in \lambda), Z_1)$ helps us to switch between different factorizations:

$$\begin{aligned} & s(Z_1) + I(C \in \lambda)\frac{\dot{P}(C \in \lambda|Z_1)}{P(C \in \lambda|Z_1)} + I(C \notin \lambda)\frac{\dot{P}(C \notin \lambda|Z_1)}{P(C \notin \lambda|Z_1)} \\ = \; & I(C \in \lambda)\left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(Z_1|C \in \lambda)\right] + I(C \notin \lambda)\left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(Z_1|C \notin \lambda)\right]. \end{aligned} \tag{8}$$

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

$$AE[m(Z; \beta^0)|C \in \lambda] = 0.$$

Differentiating under the integral, and taking a full row rank A, we obtain by using (2) that

$$\begin{aligned} \frac{\partial \beta^0(\theta_0)}{\partial \theta'} &= -(AM_\lambda)^{-1}AE\left[m(Z; \beta^0)\,\frac{\partial \log f_{\theta_0}(Z|C \in \lambda)}{\partial \theta'}\middle| C \in \lambda\right] \\ &= -(AM_\lambda)^{-1}AE\left[m(Z; \beta^0)\left\{s(Z_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_r|Z_1, \ldots, Z_{r-1})'\right\}\middle| C \in \lambda\right]. \end{aligned}$$

For an arbitrary A, pathwise differentiability follows if we can find $\psi(A, C, G_C(Z)) \in \mathcal{T}$ such that

$$E[\psi(A, C, G_C(Z))S(C, G_C(Z))'] = \frac{\partial \beta^0(\theta_0)}{\partial \theta'}. \tag{9}$$

We do this by verifying (9) after conjecturing that $\psi(A, C, G_C(Z)) = -(AM_\lambda)^{-1}A\varphi_\lambda(C, G_C(Z))$. Verification of (9) is equivalent to showing that

$$E[\varphi_\lambda(C, G_C(Z))S(C, G_C(Z))'] = E\left[m(Z; \beta^0)\left\{s(Z_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_r|Z_1, \ldots, Z_{r-1})'\right\}\middle| C \in \lambda\right]. \tag{10}$$

We do this term by term for $\varphi_\lambda(C, G_C(Z))$ and show equality of the terms on the LHS and RHS.

Consider the first term of $\varphi_\lambda(C, G_C(Z))$. Since $s(Z_r|Z_1,\ldots,Z_{r-1}) \in L_0^2(F(Z_r|Z_1,\ldots,Z_{r-1}))$ for $r = 2,\ldots,R$ from (7), we can use (2) to take conditional expectations and then write

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

where the third line follows by using (8) to replace $s(Z_1)$. The last line follows by using (2) to see that $E\left[I(C\in\lambda)\sum_{r=1}^{R} I(C=r)\frac{\dot{P}(C=r|Z_1)}{P(C=r|Z_1)}\middle| Z_1\right] = \sum_{r\in\lambda} P(C=r|Z_1)\frac{\dot{P}(C=r|Z_1)}{P(C=r|Z_1)} = \sum_{r\in\lambda}\dot{P}(C=r|Z_1) = \dot{P}(C\in\lambda|Z_1)$. Repeated use of (2) gives

$$\begin{aligned}
& E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}\varphi_{(1)}(C,G_C(Z))S(C,G_C(Z))'\right] \\
= \;& E\left[E[m(Z;\beta^0)|Z_1]|C\in\lambda\right]\frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)} + E\left[E[m(Z;\beta^0)|Z_1]s(Z_1|C\in\lambda)'|C\in\lambda\right] \\
& - E\left[E[m(Z;\beta^0)|Z_1]\frac{\dot{P}(C\in\lambda|Z_1)'}{P(C\in\lambda)}\right] + E\left[E[m(Z;\beta^0)|Z_1]\frac{\dot{P}(C\in\lambda|Z_1)'}{P(C\in\lambda)}\right] \\
= \;& E\left[m(Z;\beta^0)|C\in\lambda\right]\frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)} + E\left[E[m(Z;\beta^0)|Z_1]s(Z_1|C\in\lambda)'|C\in\lambda\right] + 0 \\
= \;& 0 + E[m(Z;\beta^0)s(Z_1|C\in\lambda)'|C\in\lambda] + 0 \qquad (11)
\end{aligned}$$

where the first zero in last line follows from (1). The second term follows by using (2) and noting that $E\left[E[m(Z;\beta^0)|Z_1]s(Z_1|C\in\lambda)'|C\in\lambda\right] = E\left[E[m(Z;\beta^0)s(Z_1|C\in\lambda)'|Z_1, C\in\lambda]|C\in\lambda\right] = E\left[m(Z;\beta^0)s(Z_1|C\in\lambda)'|C\in\lambda\right]$.

Now consider the r-th term of $\varphi_\lambda(C, G_C(Z))$ for $r = 2,\ldots,R$. By taking expectation conditional on $G_{r-1}(Z) \equiv (Z_1,\ldots,Z_{r-1})$, and using (2) we obtain

$$\begin{aligned}
& E\left[\frac{P(C\in\lambda|G_1(Z))}{P(C\in\lambda)}\varphi_{(r)}(C,G_C(Z))S(C,G_C(Z))'\right] \\
= \;& E\left[\frac{P(C\in\lambda|Z_1)}{P(C\in\lambda)}\left(E[m(Z;\beta^0)|Z_1,\ldots,Z_r] - E[m(Z;\beta^0)|Z_1,\ldots,Z_{r-1}]\right)\sum_{s=r}^{R} s(Z_s|Z_1,\ldots,Z_{s-1})\right] \\
= \;& E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m(Z;\beta^0)|Z_1,\ldots,Z_r]s(Z_r|Z_1,\ldots,Z_{r-1})'\right] \\
= \;& E\left[m(Z;\beta^0)s(Z_r|Z_1,\ldots,Z_{r-1})'|C\in\lambda\right] \qquad (12)
\end{aligned}$$

by using $s(Z_s|Z_1,\ldots,Z_{s-1}) \in L_0^2(F(Z_s|Z_1,\ldots,Z_{s-1}))$ for $s = r,\ldots,R$ from (7), along with (2).

Therefore, (11) and (12) verify (10), and hence (9). That $\varphi_\lambda(C, G_C(Z))$ belongs to $\mathcal{T}$ in (7) can be shown as follows. (i) Match the term $a(Z_1,\ldots,Z_r)$ in $\mathcal{T}$ with the r-th term of $\varphi_\lambda(C, G_C(Z))$ for $r > 1$. (ii) Distribute the first term $s(Z_1)$ in $\mathcal{T}$ according to the relation (8) and match the term $I(C \in \lambda)s(Z_1|C \in \lambda)$ with the first term of $\varphi_\lambda(C, G_C(Z))$ while keeping in mind that, by definition, $s(Z_1|C \in \lambda) \in L_0^2(F(Z_1|C \in \lambda))$. It is straightforward to verify that all the corresponding conditional expectations, as required by the definition in (7) and also (8), are zeros. Rest of the terms in $\mathcal{T}$ (including the one due to the distribution of terms in (ii)) are represented in $\varphi_\lambda(C, G_C(Z))$ by zeros.

STEP - 3: So we have verified that any regular estimator for β^0 will be asymptotically linear with influence function of the form $-(AM_\lambda)^{-1}Am(Z;\beta^0)$. For a given A, the projection of the above influence function on to the tangent set $\mathcal{T}$ is $\psi(A, C, G_C(Z))$ which, therefore, is the efficient influence function given the A. The asymptotic variance of $\psi(A, C, G_C(Z))$ is

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

where $V_\lambda := Var\left(\varphi_\lambda(C, G_C(Z))\right) = E[\varphi_\lambda(C, G_C(Z))\varphi_\lambda(C, G_C(Z))']$. Therefore, the efficient influence function is obtained by minimizing the above variance with respect to A. Standard arguments give that the minimizer is $A_* = M_\lambda' V_\lambda^{-1}$. Hence the efficiency bound is $\Omega_\lambda := (M_\lambda' V_\lambda^{-1} M_\lambda)^{-1}$ and the efficient influence function with variance equal to the efficiency bound is

$$\psi_\lambda(C, G_C(Z)) := \psi(A_*, C, G_C(Z)) = -\Omega_\lambda^{-1} M_\lambda' V_\lambda^{-1} \varphi_\lambda(C, G_C(Z)).\ \blacksquare$$

Remark: We have already considered the over-identified case in detail. This only involves an optimal rotation. Hence, for brevity, we will not consider it anymore in the sequel and work under the just-identified setup $d_m = d_\beta$. As a consequence, STEP - 2 of the subsequent proofs will only involve verifying (10) for the appropriate $\varphi_{\{.\}}(C, G_C(Z))$ function (with appropriate subscript in $\{.\}$) defined in the statement of the concerned propositions. The modification for STEP - 3 required to fit the statement of the propositions is obvious and hence omitted.

Proof of Proposition 2: $[P(C = r|G_1(Z))$ is completely known for $r = 1,\ldots,R]$

STEP - 1: Working under the same factorization of the joint density of the observed data

$(C, G_C(Z))$ as in the proof of Proposition 1, we obtain the score function with respect to θ is

$$S_\theta(C, G_C(Z)) \quad = \quad s_\theta(Z_1) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_r | Z_1, \ldots, Z_{r-1})$$

since $P(C = r|Z_1)$ is completely known. Therefore, the tangent set for the model is characterized by the set of functions of the form:

$$\mathcal{T} := a(Z_1) + \sum_{r=2}^{R} I(C \geq r) a(Z_1, \ldots, Z_r), \tag{13}$$

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

STEP - 2: $(d_m = d_\beta)$ As before, differentiating (1) under the integral, (2) gives

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

Recognizing that $P(C = r|Z_1)$ is completely known alters the relationship in (8) as follows

$$s(Z_1) = I(C \in \lambda) \left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(Z_1 | C \in \lambda) \right] + I(C \notin \lambda) \left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(Z_1 | C \notin \lambda) \right].$$

This gives, by using (2) and noting that $E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} m(Z; \beta^0) \right] \frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} = E\left[m(Z; \beta^0) | C \in \lambda \right] \frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} = 0$ by (1), that

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

As in (10), pathwise differentiability follows by verifying that $\varphi_{\lambda[k]}(C, G_C(Z)) \in \mathcal{T}$ in (13) satisfies

$$E[\varphi_{\lambda[k]}(C, G_C(Z)) S(C, G_C(Z))'] = E \left[\frac{P(C \in \lambda | Z_1)}{P(C \in \lambda)} m(Z; \beta^0) \left\{ s(Z_1)' + \sum_{r=2}^{R} s(Z_r | Z_1, \ldots, Z_{r-1})' \right\} \right].$$

The verification for any given $\lambda \in \Lambda$ is exactly the same as that in the proof of Proposition 1 for the particular case $\lambda = \mathbb{C}$ (sample space of C) or as is better known: the verify-in-sample case of Chen et al. (2008). (Recall that in the context of the latter, conditioning on $C \in \mathbb{C}$ is superfluous and $I(C \in \mathbb{C}) \equiv 1$, $P(C \in \mathbb{C}) \equiv 1$, and $P(C \in \mathbb{C}|Z_1) \equiv 1$ for all Z_1.) This is obvious by comparing $\varphi_{\lambda[k]}(C, G_C(Z))$ and $\varphi_{\mathbb{C}}(C, G_C(Z))$ on one hand, and the expressions for $\frac{\partial \beta^0(\theta_0)}{\partial \theta'}$ for both on the other.

STEP - 3: This is obvious and hence omitted. ■

Proof of Proposition 3: $[P(C=r|G_1(Z)) = P(C=r|G_1(Z);\gamma^0)\ \forall\ r.\ \gamma^0 \in \Gamma \subset \mathbb{R}^{d_\gamma}$ unknown.]

STEP - 1: The same factorization of the joint density of the observed data $(C, G_C(Z))$ as in the proof of Proposition 1 gives the score function with respect to θ as

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

Recall that $S_\gamma(C|G_1(Z)) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|Z_1)} \frac{\partial P(C=r|Z_1;\gamma^0)}{\partial \gamma}$. Let B denote the constant matrix $\frac{\partial \gamma^0}{\partial \theta'}$. Then the tangent set for the model is characterized by the set of functions:

$$\mathcal{T} := a(Z_1) + B' S_\gamma(C|Z_1) + \sum_{r=2}^{R} I(C \geq r) a(Z_1, \ldots, G_r), \tag{14}$$

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

STEP - 2: $(d_m = d_\beta)$ As before, differentiating (1) under the integral, (2) gives

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

Recognizing that $P(C=r|Z_1) = P(C=r|Z_1;\gamma^0)$ is known up to the finite (d_γ) dimensional parameter γ, alters the relationship in (8) as follows

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

Now exactly following the corresponding steps in the proof of Proposition 2 we obtain that

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

where the first line follows exactly in the same way as in the corresponding step of the proof of Proposition 2. The second line follows from the modification of (8) above. The last line uses (2).

As in (10), pathwise differentiability follows by verifying that $\varphi_{\lambda[pk]}(C, G_C(Z)) \in \mathcal{T}$ in (14) satisfies

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

Comparing with the proof of Proposition 2, this boils down to verifying that

$$\begin{aligned} & E\left[\text{Proj}\left(\frac{I(C \in \lambda)}{P(C \in \lambda)}E[m(Z;\beta^0)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right) S(C, G_C(Z))'\right] \\ = \; & E\left[E[m(Z;\beta^0)|Z_1]\frac{\frac{\partial}{\partial\gamma'}P(C \in \lambda|Z_1;\gamma^0)}{P(C \in \lambda)}\frac{\partial\gamma^0}{\partial\theta'}\right]. \end{aligned} \tag{15}$$

Consider the LHS of (15). Note that $E\left[S_\gamma(C|Z_1)\left\{s(Z_1)' + \sum_{r=2}^{R} s(Z_r|Z_1, \ldots, Z_{r-1})'\right\}\right] = 0$ by using (term by term) that $E[S_\gamma(C|Z_1)|Z_1] = 0$ [for term 1]; $s(Z_r|Z_1, \ldots, Z_{r-1}) \in L_0^2(F(Z_r|Z_1, \ldots, Z_{r-1})$ and (2) [for the rest]. Therefore, the LHS of (15) becomes

$$\begin{aligned} & E\left[\text{Proj}\left(\frac{I(C \in \lambda)}{P(C \in \lambda)}E[m(Z;\beta^0)|Z_1]\middle| S_\gamma(C|Z_1)\right) S_\gamma(C|Z_1)'\right]\frac{\partial\gamma^0}{\partial\theta'} \\ = \; & E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)}E[m(Z;\beta^0)|Z_1]S_\gamma(C|Z_1)'\right]\left(E\left[S_\gamma(.)S_\gamma(.)'\right]\right)^{-1}E\left[S_\gamma(.)S_\gamma(.)'\right]\frac{\partial\gamma^0}{\partial\theta'} \\ = \; & E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)}E[m(Z;\beta^0)|Z_1]S_\gamma(C|Z_1)'\right]\frac{\partial\gamma^0}{\partial\theta'} \\ = \; & E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)}E[m(Z;\beta^0)|Z_1]\sum_{r=1}^{R}\frac{I(C=r)}{P(C=r|Z_1)}\frac{\partial P(C=r|Z_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'} \\ = \; & E\left[\frac{1}{P(C \in \lambda)}E[m(Z;\beta^0)|Z_1]\sum_{r\in\lambda}\frac{I(C=r)}{P(C=r|Z_1)}\frac{\partial P(C=r|Z_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'} \\ = \; & E\left[\frac{1}{P(C \in \lambda)}E[m(Z;\beta^0)|Z_1]\sum_{r\in\lambda}\frac{P(C=r|Z_1)}{P(C=r|Z_1)}\frac{\partial P(C=r|Z_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'} \\ = \; & E\left[\frac{1}{P(C \in \lambda)}E[m(Z;\beta^0)|Z_1]\frac{\partial P(C \in \lambda|Z_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'} \end{aligned}$$

because $\sum_{r\in\lambda}\frac{\partial P(C=r|Z_1;\gamma^0)}{\partial\gamma'} = \frac{\partial P(C\in\lambda|Z_1;\gamma^0)}{\partial\gamma'}$. Therefore, we have verified that the LHS of (15) is equal to the RHS. This completes the proof.

STEP - 3: This is obvious and hence omitted. ■

The proof of the corollaries uses Proposition 1 with $R = 2$ and appropriate conditioning on C for the case(s) involving the suboptimal choice of subsamples in each corollary, whereas Proposition 1 is applied with $R = 3$ for the optimal choice. Note that, by virtue of MAR in (2), any conditioning on C does not affect the joint distribution of (Z_2, Z_3) conditional on Z_1. Notations of Proposition 1 are suitably adapted to those of the corollaries such that differences among the scenarios are clear. We use m to denote $m(Z;\beta^0)$. See an older version available from the author's web page for details.

Proof of Corollary 4:

(a) Using Proposition 1 with $R = 2$, $\lambda = \{1\}$ and conditioning on $C \in \{1,3\}$ gives:

$$\begin{aligned}\varphi^{\{1,3\}}_{\lambda=\{1\}}(C, G_C(Z)) &= I(C \in \{1,3\}) \left[\frac{I(C=1)}{P(C=1|C \in \{1,3\})} E[m|Z_1] + \frac{P(C=1|C \in \{1,3\}, Z_1)}{P(C=1|C \in \{1,3\})} \right. \\ &\quad \left. \times \frac{I(C=3)}{P(C=3|C \in \{1,3\}, Z_1)} (m - E[m|Z_1]) \right].\end{aligned}$$

$I(C \in \{1,3\})$ is used *only to remind* the subpopulation. The expression for $V^{\{1,3\}}_{\lambda=\{1\}} = E\left[\varphi^{\{1,3\}}_{\lambda=\{1\}}(C, G_C(Z)) \right.$ $\left. \varphi^{\{1,3\}}_{\lambda=\{1\}}(C, G_C(Z))' | C \in \{1,3\}\right]$ follows from (details omitted from other corollaries for brevity):

$$\begin{aligned}
&= E\left[\varphi^{\{1,3\}}_{\lambda=\{1\}}(C, G_C(Z)) \varphi^{\{1,3\}}_{\lambda=\{1\}}(C, G_C(Z))' | C \in \{1,3\}\right] \\
&= E\left[\frac{I(C=1)E[m|Z_1]E[m|Z_1]'}{P^2(C=1|C \in \{1,3\})} + \frac{P^2(C=1|C \in \{1,3\}, Z_1)}{P^2(C=1|C \in \{1,3\})} \frac{I(C=3)Var\,(m|Z_1)}{P^2(C=3|C \in \{1,3\}, Z_1)} \middle| C \in \{1,3\} \right] \\
&= \frac{1}{P(C \in \{1,3\})} E\left[\frac{I(C=1)E[m|Z_1]E[m|Z_1]'}{P^2(C=1|C \in \{1,3\})} + \frac{P^2(C=1|C \in \{1,3\}, Z_1)}{P^2(C=1|C \in \{1,3\})} \frac{I(C=3)Var\,(m|Z_1)}{P^2(C=3|C \in \{1,3\}, Z_1)} \right] \\
&= \frac{1}{P(C=1)} E\left[\frac{I(C=1)E[m|Z_1]E[m|Z_1]'}{P(C=1|C \in \{1,3\})} + \frac{P^2(C=1|Z_1)}{P^2(C=3|Z_1)} \frac{I(C=3)Var\,(m|Z_1)}{P(C=1|C \in \{1,3\})} \right] \\
&= \frac{1}{P(C=1)} E\left[\frac{I(C=1)E[m|Z_1]E[m|Z_1]'}{P(C=1|C \in \{1,3\})} + \frac{P^2(C=1|Z_1)}{P(C=3|Z_1)} \frac{Var\,(m|Z_1)}{P(C=1|C \in \{1,3\})} \right] \\
&= \frac{1}{P(C=1)} E\left[\frac{I(C=1)E[m|Z_1]E[m|Z_1]'}{P(C=1|C \in \{1,3\})} + \frac{P(C=1|Z_1)}{P(C=3|Z_1)} \frac{I(C=1)Var\,(m|Z_1)}{P(C=1|C \in \{1,3\})} \right] \\
&= \frac{1}{P(C=1|C \in \{1,3\})} E\left[E[m|Z_1]E[m|Z_1]' + \frac{P(C=1|Z_1)}{P(C=3|Z_1)} Var\,(m|Z_1)\ |C=1 \right].
\end{aligned}$$

(b) Simple manipulation of terms from Proposition 1 with $R = 3$, $\lambda = \{1\}$ gives the result. ■

Proof of Corollary 5:

(a) Using Proposition 1 with $R = 2$, $\lambda = \{2\}$ and conditioning on $C \in \{2,3\}$ gives:

$$\begin{aligned}\varphi^{\{2,3\}}_{\lambda=\{2\}}(C, G_C(Z)) &= I(C \in \{2,3\}) \left[\frac{I(C=2)}{P(C=2|C \in \{2,3\})} E[m|Z_1, Z_2] + \frac{P(C=2|C \in \{2,3\}, Z_1)}{P(C=2|C \in \{2,3\})} \right. \\ &\quad \left. \times \frac{I(C=3)}{P(C=3|C \in \{2,3\}, Z_1)} (m - E[m|Z_1, Z_2]) \right].\end{aligned}$$

The expression for $V^{\{2,3\}}_{\lambda=\{2\}}$ follows from $V^{\{2,3\}}_{\lambda=\{2\}} = E\left[\varphi^{\{2,3\}}_{\lambda=\{2\}}(C, G_C(Z))\varphi^{\{2,3\}}_{\lambda=\{2\}}(C, G_C(Z))'|C \in \{2,3\}\right]$.

(b) Simple manipulation of terms from Proposition 1 with $R = 3$, $\lambda = \{2\}$ gives the result. ■

Proof of Corollary 6:

(a) Using Proposition 1 with $R = 1$, $\lambda = \{3\}$ and conditioning on $C \in \{3\}$ gives:

$$\varphi^{\{3\}}_{\lambda=\{3\}}(C, G_C(Z)) \;=\; I(C = 3)m.$$

The expression for $V^{\{3\}}_{\lambda=\{3\}}$ follows from $V^{\{3\}}_{\lambda=\{3\}} = E\left[\varphi^{\{3\}}_{\lambda=\{3\}}(C, G_C(Z))\varphi^{\{3\}}_{\lambda=\{3\}}(C, G_C(Z))'|C = 3\right]$.

(b) Using Proposition 1 with $R = 2$, $\lambda = \{3\}$ and conditioning on $C \in \{1,3\}$ gives:

$$\begin{aligned}
\varphi^{\{1,3\}}_{\lambda=\{3\}}(C, G_C(Z)) &= I(C \in \{1,3\})\left[\frac{I(C=3)}{P(C=3|C\in\{1,3\})}E[m|Z_1] + \frac{P(C=3|C\in\{1,3\},Z_1)}{P(C=3|C\in\{1,3\})}\right. \\
&\qquad \left. \times \frac{I(C=3)}{P(C=3|C\in\{1,3\},Z_1)}\left(m - E[m|Z_1]\right)\right] \\
&= I(C \in \{1,3\})\left[\frac{I(C=3)}{P(C=3|C\in\{1,3\})}E[m|Z_1] + \frac{I(C=3)\left(m - E[m|Z_1]\right)}{P(C=3|C\in\{1,3\})}\right] \\
&= \frac{I(C=3)}{P(C=3|C\in\{1,3\})}m.
\end{aligned}$$

This gives $V^{\{1,3\}}_{\lambda=\{3\}} = E\left[\varphi^{\{1,3\}}_{\lambda=\{3\}}(C, G_C(Z))\varphi^{\{1,3\}}_{\lambda=\{3\}}(C, G_C(Z))'|C \in \{1,3\}\right] = E[mm'|C = 3]/P(C = 3|C \in \{1,3\})$, and hence $\lim_{N\to\infty} \Delta^{\{3\}\text{ v/s }\{1,3\}}_{\lambda=\{3\}}(N) = 0$.

Similarly, using Proposition 1 with $R = 2$, $\lambda = \{3\}$ and conditioning on $C \in \{2,3\}$ gives:

$$\begin{aligned}
\varphi^{\{2,3\}}_{\lambda=\{3\}}(C, G_C(Z)) &= I(C \in \{2,3\})\left[\frac{I(C=3)}{P(C=3|C\in\{2,3\})}E[m|Z_1,Z_2] + \frac{P(C=3|C\in\{2,3\},Z_1)}{P(C=3|C\in\{2,3\})}\right. \\
&\qquad \left. \times \frac{I(C=3)}{P(C=3|C\in\{2,3\},Z_1)}\left(m - E[m|Z_1,Z_2]\right)\right] \\
&= \frac{I(C=3)}{P(C=3|C\in\{2,3\})}m.
\end{aligned}$$

This gives $V^{\{2,3\}}_{\lambda=\{3\}} = E\left[\varphi^{\{2,3\}}_{\lambda=\{3\}}(C, G_C(Z))\varphi^{\{2,3\}}_{\lambda=\{3\}}(C, G_C(Z))'|C \in \{2,3\}\right] = E[mm'|C = 3]/P(C = 3|C \in \{2,3\})$, and hence $\lim_{N\to\infty} \Delta^{\{3\}\text{ v/s }\{2,3\}}_{\lambda=\{3\}}(N) = 0$.

(c) Simple manipulation of terms from Proposition 1 with $R = 3$, $\lambda = \{3\}$ gives the result. ■

Proof of Corollary 7:

(a) Using Proposition 1 with $R = 2$, $\lambda = \{1,3\}$ and conditioning on $C \in \{1,3\}$ gives:

$$\varphi^{\{1,3\}}_{\lambda=\{1,3\}}(C, G_C(Z)) \;=\; I(C \in \{1,3\})\left[E[m|Z_1] + \frac{I(C=3)}{P(C=3|C\in\{1,3\},Z_1)}\left(m - E[m|Z_1]\right)\right].$$

The expression for $V^{\{1,3\}}_{\lambda=\{1,3\}}$ follows from $V^{\{1,3\}}_{\lambda=\{1,3\}} = E\left[\varphi^{\{1,3\}}_{\lambda=\{1,3\}}(C, G_C(Z))\varphi^{\{1,3\}}_{\lambda=\{1,3\}}(C, G_C(Z))'|C \in \{1,3\}\right]$.

(c) Simple manipulation of terms from Proposition 1 with $R = 3$, $\lambda = \{1,3\}$ gives the result. ■

Proof of Corollary 8:

(a) Using Proposition 1 with $R = 2$, $\lambda = \{2,3\}$ and conditioning on $C \in \{2,3\}$ gives:

$$\varphi^{\{2,3\}}_{\lambda=\{2,3\}}(C, G_C(Z)) \quad = \quad I(C \in \{2,3\})\left[E[m|Z_1, Z_2] + \frac{I(C=3)}{P(C=3|C \in \{2,3\}, Z_1)}\left(m - E[m|Z_1, Z_2]\right)\right].$$

The expression for $V^{\{2,3\}}_{\lambda=\{2,3\}}$ follows from $V^{\{2,3\}}_{\lambda=\{2,3\}} = E\left[\varphi^{\{2,3\}}_{\lambda=\{2,3\}}(C, G_C(Z))\varphi^{\{2,3\}}_{\lambda=\{2,3\}}(C, G_C(Z))'|C \in \{2,3\}\right]$.

(c) Simple manipulation of terms from Proposition 1 with $R = 3$, $\lambda = \{2,3\}$ gives the result. ■

Proof of Corollary 9:

(a) Here $\lambda = \mathbb{C}$ and the used sample is $(C = 1)$ and $(C = 3)$. Note that:

$$E[m(Z;\beta)] = E\left[\frac{P(C \in \{1,3\})}{P(C \in \{1,3\}|Z_1)} m(Z;\beta)\middle| C \in \{1,3\}\right]. \tag{16}$$

Since this proof is slightly different from the rest, let us briefly point out the connection with the proof of Proposition 1. Note that the same arguments lead to the score function (using the same notations):

$$\begin{aligned} S_\theta(C, G_C(Z)|C \in \{1,3\}) \quad = \quad & I(C \in \{1,3\}) s_\theta(Z_1|C \in \{1,3\}) + \sum_{r=1,3} I(C=r)\frac{\dot{P}_\theta(C=r|C \in \{1,3\}, Z_1)}{P_\theta(C=r|C \in \{1,3\}, Z_1)} \\ & + I(C=3) s_\theta(Z_2, Z_3|Z_1). \end{aligned}$$

Therefore, the tangent set for the model can be characterized by functions of the form:

$$\mathcal{T} := I(C \in \{1,3\}) a(Z_1|C \in \{1,3\}) + \sum_{r=1,3} I(C=r)\frac{b_r(C \in \{1,3\}, Z_1)}{a_r(C \in \{1,3\}, Z_1)} + I(C=3)a(Z), \tag{17}$$

where $a(Z_1|C \in \{1,3\}) \in L^2_0(F(Z_1|C \in \{1,3\}))$; $\sum_{r=1,3}(a_r(C \in \{1,3\}, Z_1), b_r(C \in \{1,3\}, Z_1)) = (1,0)$ for all Z_1 and $C \in \{1,3\}$, and $\sum_{r=1,3} I(C=r)\frac{b_r(C \in \{1,3\}, Z_1)}{a_r(C \in \{1,3\}, Z_1)} \in L^2_0(F(C|C \in \{1,3\}, Z_1))$; and $a(Z) \in L^2_0(F(Z_2, Z_3|Z_1))$.

Differentiating under the integral the identity $E[m] = 0$ with respect to θ (at $\theta = \theta^0$), it only remains to be shown similar to (10) in the proof of Proposition 1 that:

$$E[ms(Z)'] \quad = \quad E\left[\varphi^{\{1,3\}}_{\lambda=\mathbb{C}}(C, G_C(Z)) S(C, G_C(Z))'|C \in \{1,3\}\right], \tag{18}$$

$$\text{where } \varphi^{\{1,3\}}_{\lambda=\mathbb{C}}(C, G_C(Z)) \quad = \quad \frac{P(C \in \{1,3\})}{P(C \in \{1,3\}|Z_1)}\left[I(C \in \{1,3\})E[m|Z_1] + \frac{I(C=3)\left(m - E[m|Z_1]\right)}{P(C=3|C \in \{1,3\}, Z_1)}\right].$$

(18) follows directly using the same techniques (for example, (2) and the properties stated below (17)) as before, once we note that the differentiation of $E[m] = 0$ along with (16) and (2) imply that:

$$E[ms(Z)'] = E\left[\frac{P(C \in \{1,3\})}{P(C \in \{1,3\}|Z_1)} m \left\{ \underbrace{s(Z_2, Z_3|Z_1) + s(Z_1|C \in \{1,3\})}_{=s(Z|C \in \{1,3\})} \right\}' \middle| C \in \{1,3\} \right].$$

(b) Simple manipulation of terms from Proposition 1 with $R = 3$, $\lambda = \mathbb{C}$ gives the result.

(c) Follows in a similar way as part (a).

(d) Simple manipulation of terms from Proposition 1 with $R = 3$, $\lambda = \mathbb{C}$ gives the result. ■

Proof of Proposition 10:

Consider the first equality. Let us start with $r = 1$, i.e., the residual from the projection $\overline{\text{Proj}}_{G_{R-1}}\left(\phi_\lambda^R(\beta) \middle| \phi^{R-1}\right)$ inside the innermost parenthesis on the RHS. We will also consider $r = 2$, so that the pattern in the form of the residuals from the successive projections inside the first few innermost parentheses is clear to the reader. Then we apply induction arguments. For notational simplicity define $B := \frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)}$ and write $m(Z;\beta)$ as m and $G_s(Z)$ as G_s for all s.

Direct computation and version (B) of (2) give:

$$\begin{aligned}
\text{Proj}_{G_{R-1}}\left(\phi_\lambda^R(\beta) \middle| \phi^{R-1}\right) &= B\left[\frac{I(C=R)}{P(C=R|G_1)} - \frac{I(C \geq R-1)}{P(C \geq R-1|G_1)}\right] E[m|G_{r-1}] \\
\Rightarrow \overline{\text{Proj}}_{G_{R-1}}\left(\phi_\lambda^R(\beta) \middle| \phi^{R-1}\right) &= B\left[\frac{I(C=R)}{P(C=R|G_1)} \underbrace{(m - E[m|G_{R-1}])} + \frac{I(C \geq R-1)}{P(C \geq R-1|G_1)} E[m|G_{R-1}]\right].
\end{aligned}$$

Consider the under-braced part in the RHS of the expression for $\overline{\text{Proj}}_{G_{R-1}}\left(\phi_\lambda^R(\beta) \middle| \phi^{R-1}\right)$. Using $G_{R-1} \setminus G_{R-2} = Z_{R-1}$ and (2), note that $E\left[(m - E[m|G_{R-1}])\,\phi^{R-2} | G_{R-2}\right]$ is a $d_m \times 2$ matrix of zeros, and hence has no contribution in successive projections. Similar computation gives for $r = 2$:

$$\begin{aligned}
&\overline{\text{Proj}}_{G_{R-2}}\left(\overline{\text{Proj}}_{G_{R-1}}\left(\phi_\lambda^R(\beta) \middle| \phi^{R-1}\right) \middle| \phi^{R-2}\right) \\
= \; & B\left[\sum_{s=0}^{r-1} \frac{I(C \geq R-s)}{P(C \geq R-s|G_1)} \underbrace{(E[m|G_{R-s}] - E[m|G_{R-s-1}])} + \frac{I(C \geq R-r)}{P(C \geq R-r|G_1)} E[m|G_{R-r}]\right]
\end{aligned}$$

where $E[m|G_R] \equiv m$ shortens the expression. Let the following hold for a general $r \in \{2, \ldots, R-2\}$:

$$\begin{aligned}
&\overline{\text{Proj}}_{G_{R-r}}\left(\ldots \overline{\text{Proj}}_{G_{R-1}}\left(\phi_\lambda^R(\beta) \middle| \phi^{R-1}\right) \ldots \middle| \phi^{R-r}\right) \\
= \; & B\left[\sum_{s=0}^{r-1} \frac{I(C \geq R-s)}{P(C \geq R-s|G_1)} \underbrace{(E[m|G_{R-s}] - E[m|G_{R-s-1}])} + \frac{I(C \geq R-r)}{P(C \geq R-r|G_1)} E[m|G_{R-r}]\right].
\end{aligned}$$

Now noting that

$$E[\phi^{R-r-1}\phi^{R-r-1}|G_{R-r-1}] = \frac{P(C \geq R-r|G_1)P(C = R-r-1|G_1)}{P(C \geq R-r-1|G_1)}, \text{ and}$$
$$E[\overline{\text{Proj}}_{G_{R-r}}\left(\ldots \overline{\text{Proj}}_{G_{R-1}}\left(\phi^R_\lambda(\beta)\middle|\, \phi^{R-1}\right)\ldots\middle|\, \phi^{R-r}\right)\phi^{R-r-1}|G_{R-r-1}] = \frac{P(C = R-r-1|G_1)}{P(C \geq R-r-1|G_1)}BE[m|G_{R-r-1}],$$

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

$$\begin{aligned} & \overline{\text{Proj}}_{G_{R-r-1}}\left(\ldots \overline{\text{Proj}}_{G_{R-1}}\left(\phi^R_\lambda(\beta)\middle|\, \phi^{R-1}\right)\ldots\middle|\, \phi^{R-r-1}\right) \\ = \; & B\left[\sum_{s=0}^{r}\frac{I(C \geq R-s)}{P(C \geq R-s|G_1)}\left(E[m|G_{R-s}] - E[m|G_{R-s-1}]\right) + \frac{I(C \geq R-r-1)}{P(C \geq R-r-1|G_1)}E[m|G_{R-r-1}]\right]. \end{aligned}$$

The second equality in the statement of the proposition follows due to monotonicity of the conditioning sets. It can also be verified directly like the first equality. ■

Appendix B: Version [A] of the MAR assumption in (2)

As noted following (2), the convenient version of MAR, i.e., (2)[B] may be unrealistic in certain cases. One such case is the most common scenario for monotone missingness: attrition with no return in panel studies. See the discussion toward the end of Toy Example 1C in Section 3. (2)[B] implies that attrition in, say, period 3 of a panel is independent of the period 3 and period 2 time varying characteristics of the units conditional on their period 1 characteristics. This is similar to the second attrition analysis (page 145) in Fitzgerald et al. (1998) who also noted the crude nature of this assumption.

A sequential approach with MAR that recognizes the dynamic nature of selection/attrition will typically modify this assumption such that attrition in period 3 is independent of the period 3 values of the time varying characteristics of the units conditional on their period 1 and period 2 characteristics.

Our focus in this paper was different. We set out to explore how incomplete subsamples can be combined optimally with a complete subsample for efficient estimation of some parameters of interest. Besides its role as the identifying assumption, version (2)[B] helped to demonstrate the benefits of such optimal combination for all choices of the target $\lambda \in \Lambda$. We noted that under certain scenarios possible sampling strategies satisfying (2)[B] can be designed for the purpose of cost effectiveness.

For completeness let us now report the counterpart of Proposition 1 under the more realistic version (2)[A] for a case common in empirical applications: $\lambda = \mathbb{C}$. The result is not new at least in the context of attrition. One can also accommodate for the other choices of the target λ similar to Proposition 1.

Proposition 11 *Let assumption A and (2)[A] hold. Denoting $m(Z;\beta)$ by $E[m(Z;\beta)|G_R(Z)]$, define*

$$\bar{\varphi}(C, G_C(Z); \beta) := E[m(Z;\beta)|G_1(Z)] + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|G_R(Z))} \left(E[m(Z;\beta)|G_r(Z)] - E[m(Z;\beta)|G_{r-1}(Z)]\right),$$

$$\text{and } \bar{V} := Var\left(\bar{\varphi}(.;\beta^0)\right) = \sum_{r=2}^{R} \frac{Var\left(E[m(Z)|G_r(Z)]|G_{r-1}(Z)\right)}{P(C \geq r|G_R(Z))} + E[m(Z)|G_1(Z)]E[m(Z)|G_1(Z)]'.$$

Let $\bar{V}$ be positive definite. Then for β^0 defined by (1), the asymptotic variance lower bound for $\sqrt{N}(\widehat{\beta}-\beta^0)$ of any regular estimator $\widehat{\beta}$ is given by $\Omega := (M'\bar{V}^{-1}M)^{-1}$. An estimator whose asymptotic variance equals Ω has the asymptotically linear representation

$$\begin{aligned} \sqrt{N}(\widehat{\beta} - \beta^0) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{\psi}(C_i, G_{C_i}(Z_i)) + o_p(1), \text{ where} \\ \bar{\psi}(C, G_C(Z)) &:= -\bar{\Omega}^{-1} M' \bar{V}^{-1} \bar{\varphi}_\lambda(C, G_C(Z); \beta^0). \end{aligned}$$

Proof of Proposition 11:

STEP - 1: In a regular parametric sub-model indexed by a finite-dimensional parameter θ for the joint distribution of $(C, G_C(Z))$, its log density can be expressed in terms of the full data (C, Z) as

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

under the conventions: $f_\theta(Z_1|Z_1, \ldots,) \equiv f_\theta(Z_1)$, $Z_j \equiv \varnothing$ for $j < 1$, and $I(C \geq 1) \equiv 1$. Using the same notations as before, the score function with respect to θ can be written in terms of (C, Z) as

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

As before, the tangent set for the model can be characterized by functions of the form:

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

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

STEP - 2: $[d_m = d_\beta]$ As before, differentiating (1) under the integral we obtain:

$$\frac{\partial \beta^0(\theta_0)}{\partial \theta'} = -M^{-1} E\left[m(Z;\beta^0) \frac{\partial \log f_{\theta_0}(Z)}{\partial \theta'}\right] = -AM^{-1} \sum_{r=1}^{R} E\left[m(Z;\beta^0) s(Z_r|Z_1, \ldots, Z_{r-1})'\right].$$

To show pathwise differentiability as in the other proofs, it suffices to verify that

$$E[\bar{\varphi}(C, G_C(Z))S(C, G_C(Z))'] = \sum_{r=1}^{R} E\left[m(Z;\beta^0)s(Z_r|Z_1,\ldots,Z_{r-1})'\right]. \tag{19}$$

Note that the LHS of (19) $= D + \sum_{q=2}^{R} B_q$ where

$$\begin{aligned} D &:= E\left[E[m(Z)|G_1(Z)]S(C, G_C(Z))'\right] \\ B_q &:= E\left[\frac{I(C \geq q)}{P(C \geq q|G_R(Z))}\left(E[m(Z)|G_q(Z)] - E[m(Z)|G_{q-1}(Z)]\right)S(C, G_C(Z))'\right]. \end{aligned}$$

To keep notations short let $m \equiv m(Z;\beta^0)$ and $G_r = G_r(Z)$ for all r. Now,

$$\begin{aligned} D &= E\left[E[m|G_1]\sum_{r=1}^{R}\dot{P}(C = r|G_r)\right] + \sum_{r=1}^{R} E\left[E[m|G_1]\left(1 - P(C \leq r-1|G_{r-1})\right)s(Z_r|Z_1,\ldots,Z_{r-1})'\right] \\ &= 0 + E[E[m|G_1]s(Z_1)'] = E[ms(Z_1)'] \end{aligned}$$

because $\sum_{r=1}^{R}\dot{P}(C = r|G_r) = 0$ and $s(Z_r|Z_1,\ldots,Z_{r-1}) \in L_0^2(F(Z_r|Z_1,\ldots,Z_{r-1}))$. In the last term in the RHS of the first line we switched from $I(C \geq r)$ to $1 - I(C \leq r-1)$ so the appropriate conditioning variables allow us to use the property $s(Z_r|Z_1,\ldots,Z_{r-1}) \in L_0^2(F(Z_r|Z_1,\ldots,Z_{r-1}))$. Similarly,

$$\begin{aligned} B_q &= \sum_{r=q}^{R} E\left[\frac{\dot{P}(C = j|G_j)}{P(C \geq q|G_R)}\left(E[m|G_q] - E[m|G_{q-1}]\right)\right] \\ &\quad + \sum_{r=1}^{R} E\left[\frac{I(C \geq \max(q,r))}{P(C \geq q|G_R)}\left(E[m|G_q] - E[m|G_{q-1}]\right)s(Z_r|Z_1,\ldots,Z_{r-1})'\right]. \end{aligned}$$

The RHS in the first line is $E\left[\frac{-\dot{P}(C \leq q-1|G_{q-1})}{1-P(C \leq q-1|G_{q-1})}\left(E[m|G_q] - E[m|G_{q-1}]\right)\right]$ by using $\dot{P}(C \geq q|G_R) = -\dot{P}(C \leq q-1|G_R) = -\dot{P}(C \leq q-1|G_{q-1})$. Taking expectation conditional on G_{q-1}, this term becomes zero. Hence,

$$\begin{aligned} B_q &= \sum_{r=1}^{q-1} E\left[\frac{I(C \geq q)}{P(C \geq q|G_R)}\left(E[m|G_q] - E[m|G_{q-1}]\right)s(Z_r|Z_1,\ldots,Z_{r-1})'\right] \\ &\quad + \sum_{r=q}^{R} E\left[\frac{I(C \geq r)}{P(C \geq q|G_R)}\left(E[m|G_q] - E[m|G_{q-1}]\right)s(Z_r|Z_1,\ldots,Z_{r-1})'\right] \\ &= \sum_{r=1}^{q-1} E\left[\frac{1 - I(C \leq q-1)}{1 - P(C \leq q-1|G_{q-1})}\left(E[m|G_q] - E[m|G_{q-1}]\right)s(Z_r|Z_1,\ldots,Z_{r-1})'\right] \\ &\quad + \sum_{r=q}^{R} E\left[\frac{I(C \geq r)}{P(C \geq q|G_R)}\left(E[m|G_q] - E[m|G_{q-1}]\right)s(Z_r|Z_1,\ldots,Z_{r-1})'\right]. \end{aligned}$$

The second line from below is $\sum_{r=1}^{q-1} E\left[\frac{1-P(C\leq q-1|G_{q-1})}{1-P(C\leq q-1|G_{q-1})}\left(E[m|G_q]-E[m|G_{q-1}]\right)s(Z_r|Z_1,\ldots,Z_{r-1})'\right] = 0$ by taking expectation conditional on G_{q-1}. Therefore,

$$\begin{aligned} B_q &= \sum_{r=q}^{R} E\left[\frac{1-P(C\leq r-1|G_{r-1})}{1-P(C\leq q-1|G_{q-1})}\left(E[m|G_q]-E[m|G_{q-1}]\right)s(Z_r|Z_1,\ldots,Z_{r-1})'\right] \\ &= E[E[m|G_q]s(Z_q|Z_1,\ldots,Z_{q-1})'] = E[ms(Z_q|Z_1,\ldots,Z_{q-1})'] \end{aligned}$$

by applying $s(Z_r|Z_1,\ldots,Z_{r-1}) \in L_0^2(F(Z_r|Z_1,\ldots,Z_{r-1}))$ for $r > q$ and $r \geq q$ to deal with $E[m|G_q]$ and $E[m|G_{q-1}]$ respectively. Verification follows since the LHS of (19) $= D + \sum_{q=2}^{R} B_q$.

STEP - 3: This is obvious and hence omitted. ■