A Note on Efficiency Gains from Auxiliary Samples*

Saraswata Chaudhuri†

Current version: July 13, 2013; First version: March 8, 2013.

Comments are welcome.

Abstract

We study the use of auxiliary samples for efficiency gains in the estimation of a finite-dimensional parameter value defined by a set of moment restrictions on a target population. The target population may be an artificial construct defined in terms of the observability of variables forming the moment vector. Multiple variables forming the moment vector may be missing jointly or individually from all but one of the primary and the auxiliary samples; hence there exist the so-called complete data cases. Efficiency gains are discussed in terms of the semiparametric efficiency bounds under the use of different auxiliary samples along with the primary sample as the observed data. The underlying population of the observed data may or may not contain the target population, and identification is obtained by a convenient version of the missing at random assumption.

JEL Classification: C13; C14; C31.

Keywords: Auxiliary Samples; Efficiency gain; Generalized method of moments; (Non-) Monotonically missing data; Semiparametric efficiency bound.

*We thank A. Prokhorov, SJ. Lee, P. Saha Chaudhuri, the seminar participants at U. Sydney (Business Analytics and Economics), U. New South Wales and U. Canterbury for helpful discussions.

†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

Consider a finite dimensional random variable $Z = (Z_1, \ldots, Z_R)$ partitioned in R blocks. Suppose that we have a collection of sample units where for each unit we only observe the random variables $(C, G_C(Z))$. $C \in \mathbb{C} := \{1, \ldots, R\}$ is the coarsening variable. $G_C(Z)$ is a transformation of Z. In most of this paper we will maintain $G_r(Z) := (Z_1, \ldots, Z_r)$ for $r = 1, \ldots, R$, that reflects a single hierarchy in the information content, i.e., monotone missingness (coarsening). A simple case of non-monotone missingness is also studied at the end of the paper.

The setup in this paper is based on Chen et al. (2008). 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 $C \perp Z$. $\mathcal{D}_\lambda := (C, Z : C \in \lambda)$ can be viewed as the target (sub-)population of $\mathcal{D} := (C, Z : C \in \mathbb{C})$. Some practical examples of this setup are discussed later in the introduction to give an idea of the scope of our paper.

There are two econometric issues of interest of which we will focus on the second one. The first issue is the point identification of β^0 based on the observed data. A common practice, that we also follow here, is identification through assumptions such as missing/coarsening at random introduced by and discussed in Little and Rubin (2002), Heitjan and Rubin (1991), Gill et al. (1997). The second issue, pioneered by Robins et al. (1994), is the optimal use of information contained in the observed data for efficient estimation of β^0 [also see Hahn (1998), Hirano et al. (2003), Chen et al. (2008), Graham (2011), etc. when $R = 2$].

We generalize the context of the second issue, i.e., efficient estimation of β^0, to the setup of (1). To facilitate the discussion, consider a partition of the observed data into primary and auxiliary samples for which we adopt the following notion that is similar to Chaudhuri and Min (2012) and differs slightly from Chen et al. (2008). When $\lambda \neq \mathbb{C}$, all sample units with $C \in \lambda$ are referred to as the primary sample and the rest constitute the auxiliary sample. On the other hand, when $\lambda = \mathbb{C}$, sample units with $G_C(Z) = Z$, i.e., the so-called complete data, are referred to as the primary sample (difference with Chen et al. (2008)) while the rest constitute the auxiliary samples.

The goal in this paper is to optimally use the observed data consisting of the primary and the auxiliary samples, and extract all the relevant information from the latter for the efficient estimation

of β^0. The notion of primary and auxiliary samples adopted here stresses on the efficiency gains coming from features of the joint distribution of Z that get revealed more precisely due to the availability of the auxiliary samples, and not on gains merely due to an increase in sample size.

We bypass the difficult (first) issue of point identification of β^0 by maintaining throughout the following convenient version of the missing/coarsening at random assumption: For $r = 1, \ldots, R$,

$$P(C = r|Z) \underbrace{=}_{\text{"at random"}} P(C = r|G_r(Z)) \underbrace{=}_{\text{convenient version}} P(C = r|G_1(Z)) \equiv P(C = r|Z_1). \qquad (2)$$

Since this is sufficient for our goal of describing the main features of optimally using the observed data for efficient estimation of β^0, we do not consider sophisticated variations of (2) that may be more realistic in many cases [see Robins and Rotnitzky (1995), Tsiatis (2006) and footnote 4.]

Technically, our setup is a straightforward extension of the vast literature on missing data to cases where the parameter of interest is possibly defined only in terms of sub-populations. 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 (missing data) model in a simple way that provides additional important insights useful for practical purposes.[1] This is our main message.

Let us preview an implication of the extension beyond two-level missingness that we consider in our paper. Take for example a three-level monotone missing data model ($R = 3$) where $\lambda = \{3\}$. Now the primary sample ($C = 3$) alone can point identify β^0. Auxiliary samples, aided by (2), may only help efficiency gains relative to what one would otherwise obtain based on the primary sample alone. However, we show that neither ($C = 1$) nor ($C = 2$), when used separately as an auxiliary sample along with the primary sample, gives the efficiency gain. Interestingly, on the other hand, we also show that ($C = 1$) and ($C = 2$) provide such improvements when used jointly as auxiliary samples. From a practical perspective, this makes a case for collecting presumably easier to obtain (than Z_3 due to monotonicity) information on (Z_2 and Z_1) and (Z_1) separately for additional sample units chosen based on Z_1.[2] Other such examples are also presented in Section 3.

Some examples:

It is useful at this point to give some explicit practical examples of the setup considered here. The examples vary by context but share the same flavor of monotonically missing data. For example 3 we specify the estimand β because it requires an *explicit* consideration of the potential outcome

[1] $\lambda = \{1\}$ represents Chen et al. (2008)'s verify-out-of-sample case but with auxiliary samples represented by $C = 2, \ldots, C = R$ with increasing information content. $\lambda = \mathbb{C}$ represents their verify-in-sample case because $E[m(Z;\beta)|C \in \mathbb{C}] \equiv E[m(Z;\beta)]$. The connection with the vast literature on validation data is discussed in Chen et al. (2008).

[2] Nevo (2003) makes a similar point in the context of correction for sample selection, i.e., the first issue of interest — point identification of β^0 — stated above. Hellerstein and Imbens (1999) consider both two issues of interest.

framework. The estimand in the other examples can be more general.

Example 1: Suppose information on Z_1, Z_2 and Z_3 are progressively more costly to obtain. A modification of the variable probability stratified sampling [see Wooldridge (1999)] can be employed. Collect Z_1 for randomly chosen individuals. Stratify the sample by one or more variables in Z_1. For each individual in the sample, collect (Z_2 and Z_3) or (Z_2) or nothing with non-trivial probability depending on the stratum of the individual.

Example 2: Attrition in panel studies where individuals who leave the panel never return leads to monotonically missing data. For example, let X_t and W respectively represent the (non-trivially) time varying and time invariant characteristics of an individual in period t for, say, $t = 1, 2, 3$. $Z_1 = (X_1, W)$ is observed for all. $Z_2 = X_2$ and $Z_3 = X_3$ are not observed if the individual leaves the panel after period 1 while $Z_3 = X_3$ is not observed if the individual leaves after period 2. Fitzgerald et al. (1998) motivated the missing at random assumption in the context of attrition. Our simplified setup can be modified to accommodate for possible serial correlation in X_t. As noted before, the convenient version (second equality) in (2) that does not require conditioning on Z_2 for ignorability is possibly an unrealistic characterization of the missingness of Z_3.

Example 3: Consider a causal model with potential outcomes defined as: $Y(m, x)$, $M(x)$ for $x = 0, 1$ and $m = 0, 1$. Let $\beta^0 = E[Y(0, 0)]$ (intercept term). We observe $X, M = XM(1) + (1 - X)M(0)$ and $Y = MXY(1, 1) + M(1-X)Y(1, 0) + (1-M)XY(0, 1) + (1-M)(1-X)Y(0, 0)$. Then taking $Z_1 = X$, $Z_2 = M(0)$ and $Z_3 = Y(0, 0)$ gives monotonically missing data along the nodes of the causal path: $[X = 0] \longrightarrow [M(0) = 0] \longrightarrow [Y(0, 0)]$. The convenient version in (2) imposes strong restrictions on the mediator M in this context. (To fit this directly to our setup, take $Z_2 = (1-X)M \equiv (1-X)M(0)$ and $Z_3 = (1-X)(1-M)Y \equiv (1-X)(1-M)Y(0, 0)$, i.e, simply consider a scenario where the other potential outcomes are never observed.)

Example 4: Let the setup be similar to example 1. However, let Z_2 and Z_3 be correctly measured values of some elements of Z_1. This is related to the idea of refreshment samples, but the selective followup is within the original sample.

Example 5: Let the setup be similar to examples 1 or 4. However, let the additional information be collected on new sample units in the spirit of genuine refreshment samples. Unlike in the previous examples, here the sample size for each missing/coarsened data group is not necessarily random [compare the verify-in-sample and verify-out-of-sample cases in Chen et al. (2008)].

Example 6: In the spirit of efficiency gains using presumably incomplete auxiliary samples in Section 3, we can consider the opposite scenario of example 5. Suppose the original sample contains

Z_1, Z_2, Z_3. New sample units with only Z_1 or (Z_1 and Z_2) can be added to the original sample.

Example 7: Unlike in examples 1, 4, 5 and 6 where the survey design leads to missing Z_2 and Z_3, these variables can be genuinely missing for various reasons even in cross sectional data.

Organization of the paper

The rest of the paper is organized as follows. Theoretical results for the monotone missing data model are provided in Section 2 where we present the efficiency bound for each choice of $\lambda \in \Lambda$ under Proposition 1. An insight that, in our opinion, makes these efficiency bounds more intuitive from both theoretical and computation perspectives was provided by Brown and Newey (1998) in the context of combining conditional and unconditional moment restrictions. We stress on this insight in our paper. For completeness and given the stress on designing surveys in the preceding examples, we also present efficiency bounds under a straightforward (since the original work of Hahn (1998) and Chen et al. (2008)) extension in terms of complete or partial parametric knowledge of $P(C = r|Z_1)$.

As noted earlier, the interesting consequences of Proposition 1 are described in Section 3. For concreteness we focus on a three-level monotone missing data model ($R = 3$), and present its implications for various choices of $\lambda \in \Lambda$ in the form of corollaries.

In Section 4 we consider a simple non-monotone missing data model similar to that in Chaudhuri and Guilkey (2013) and show the efficiency gains due to optimally using the observed data. The presentation is brief and the focus is on demonstrating improvements over the results in Sections 2 and 3 due to the incremental information available. All the proofs are collected in the Appendix.

2 Efficiency bounds under monotone missing/coarsened data

The observed data in the R-level monotone missing data model is $(C, G_C(Z))$ where $C \in \mathbb{C} := \{1, \ldots, R\}$ and $G_C(Z)$ is such that $G_r(Z) := (Z_1, \ldots, Z_r)$ for $r = 1, \ldots, R$. In this section we work with any given $\lambda \in \Lambda$ where $\Lambda :=$ Power-Set$(\mathbb{C}) \setminus \{\text{empty set}\}$. We maintain throughout the following assumption or minor variation of it that is clear from the context.

Assumption A

(A1) The observed data $\{C_i, G_{C_i}(Z_i)\}_{i=1}^N$ is 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 $M_\lambda(\beta) := E\left[\frac{\partial m(Z;\beta)}{\partial \beta'}\middle| C \in \lambda\right]$.

The restrictions on $P(C = r|G_1(Z))$ for $r = 1, \ldots, R-1$ in (A2) are not required for the identification of β^0 and the results presented here. They only help to avoid more involved proofs that are peripheral

to the message of the paper. However $P(C = r) > 0$ is intrinsic to the R-level missing model.

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:[3]

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

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

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) The proof is given in Appendix A. The expression for V_λ is derived in Appendix B.1.

(ii) 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 sub-

[3]The proper representation of $m(Z;\beta)$ in terms of the observed data is $m(G_R(Z);\beta)$. However, we continue to use the former to avoid any confusion due to our nonstandard use of $(C = R, G_R(Z))$ instead of $(C = \infty, G_\infty(Z))$ to denote the complete data [see Tsiatis (2006)]. We reserve the notation $(C = \infty, G_\infty(Z))$ exclusively for the complete data case in the non-monotone missing data model in Section 4 where it is important to make a distinction between the R-partition of Z and the number of levels of missingness in the model that is represented by C's support points.

populations. In Section 3 we show that this extension gives important insights on efficiency gains due to auxiliary samples that are unavailable from a two-level missing data model.

(iii) The above expression of the efficient influence function $\psi(C, Z)$ emphasizes the additional contribution of each level of coarsened data in a certain way that is useful for Section 3. However, the key term $\varphi_\lambda(C, G_C(Z))$ in $\psi(C, Z)$ also has the standard representation

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

as a member of the augmented inverse probability weighted (AIPW) class introduced by Robins et al. (1994), the consequences of which are beneficial for the purpose of estimation of β^0. The first line of (3) is the IPW part while the last two lines represent the augmentation.[4] More intuition on the structure of the influence function is provided in Appendix B.5.

There is another way to delineate the additional contribution of each level of coarsened data. The basic insight comes from Brown and Newey (1998) and Graham (2011), and has been used by Chaudhuri and Guilkey (2013). For simplicity, consider an example where $R = 3$ and $\lambda = \mathbb{C} := \{1, 2, 3\}$. An inverse probability weighted moment vector based on the complete data $(C = 3, G_3(Z))$ (our notion of primary sample when $\lambda = \mathbb{C}$) that is unbiased for $E[m(Z; \beta)]$ under (2) is

$$\phi(C = 3, G_3(Z); \beta) := \frac{I(C = 3)}{P(C = 3 | G_1(Z))} m(Z; \beta).$$

Now we show that the efficiency results in Proposition 1 can alternatively be achieved by using the additional levels of coarsened data — $G_2(Z)$ and $G_1(Z)$ — to augment the moment restrictions

[4] $\psi_\lambda(C, G_C(Z))$ needs modification under sophisticated versions of (2) [see, for example, Tsiatis (2006)]. While we do not consider it in the rest of the paper for the purpose of simplicity, let us point out the modification required under the general form of the missing/coarsening at random assumption, i.e., the first equality in (2), with an example. Let $R = 3$ and $\lambda = \mathbb{C} := \{1, 2, 3\}$. Noting that, by virtue of (2), the identity $1 = \sum_{r=1}^{3} P(C = r | Z)$ gives $1 = \sum_{r=1}^{3} P(C = r | G_r(Z))$, (3) needs to be modified as follows [see Appendix B.4 for details]:

$$\begin{aligned}\varphi_{\mathbb{C}}(C, G_C(Z); \beta) &= \frac{I(C = 3)}{P(C = 3 | G_3(Z))} m(Z; \beta) + \left\{ \frac{I(C = 2) + I(C = 3)}{P(C = 2 | G_2(Z)) + P(C = 3 | G_3(Z))} - \frac{I(C = 3)}{P(C = 3 | G_3(Z))} \right\} \\ &\quad \times E[m(Z; \beta) | G_2(Z)] + \left\{ 1 - \frac{I(C = 2) + I(C = 3)}{P(C = 2 | G_2(Z)) + P(C = 3 | G_3(Z))} \right\} E[m(Z; \beta) | G_1(Z)] \\ &= E[m(Z; \beta) | G_1(Z)] + \frac{I(C = 2) + I(C = 3)}{P(C = 2 | G_2(Z)) + P(C = 3 | G_3(Z))} \left(E[m(Z; \beta) | G_2(Z)] - E[m(Z; \beta) | G_1(Z)] \right) \\ &\quad + \frac{I(C = 3)}{P(C = 3 | G_3(Z))} \left(m(Z; \beta) - E[m(Z; \beta) | G_2(Z)] \right).\end{aligned}$$

$$E[\phi(C=3,G_3(Z);\beta)]=0 \text{ for } \beta\in\mathcal{B}\Longleftrightarrow\beta=\beta^0 \tag{4}$$

with the following auxiliary restrictions implied by (2):

$$\begin{aligned} E\left[\phi_2(C,G(Z))|C\geq 2,G_2(Z)\right] &= 0 \text{ almost surely } G_2(Z), &(5)\\ E\left[\phi_1(C,G(Z))|G_1(Z)\right] &= 0 \text{ almost surely } G_1(Z). &(6)\end{aligned}$$

$\phi_2(C,G(Z))$ and $\phi_1(C,G(Z))$ are defined as follows:

$$\begin{aligned} \phi_2(C,G(Z)) &:= I(C\geq 2)\left[I(C=3)-P(C=3|C\geq 2,G_1(Z))\right],\\ \phi_1(C,G(Z)) &:= \left[I(C=3)-P(C=3|G_1(Z)),I(C=2)-P(C=2|G_1(Z))\right]'. \end{aligned}$$

Note that (5) and (6) reflect the observability of $G_2(Z)$ and $G_1(Z)$ respectively, and thus make it readily possible to appreciate the respective contributions of $G_2(Z)$ and $G_1(Z)$ to achieving semiparametric efficiency. We demonstrate this using the insights of Brown and Newey (1998) and Graham (2011) in a sequential manner similar to the application of the Frisch-Waugh-Lovell theorem.

In the first step, use (5) and hence $G_2(Z)$ for efficiency gain with respect to what corresponds to (4). For this purpose consider the moment restrictions

$$E\left[\widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)\right]=0 \text{ for } \beta\in\mathcal{B}\Longleftrightarrow\beta=\beta^0 \tag{7}$$

where $\widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)$ is the residual from a projection and is defined as follows

$$\begin{aligned} &\widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)\\ :=\;& \phi(.;\beta)-E\left[\phi(.;\beta)\phi_2(.)|C\geq 2,G_2(Z)\right]\left(E\left[\phi_2^2(.)|C\geq 2,G_2(Z)\right]\right)^{-1}\phi_2(.)\\ =\;& \frac{I(C=3)}{P(C=3|G_1(Z))}\left(m(Z;\beta)-E\left[m(Z;\beta)|G_2(Z)\right]\right)+\frac{I(C\geq 2)}{P(C\geq 2|G_1(Z))}E\left[m(Z;\beta)|G_2(Z)\right]. \end{aligned}$$

Efficiency gain follows because $E\left[\phi(.;\beta^0)\phi(.;\beta^0)'\right]-E\left[\widetilde{\phi}(.;\beta^0)\widetilde{\phi}(.;\beta^0)'\right]$ is positive semi-definite. (Such computations are used repeatedly in the sequel.) Conditioning on $C\geq 2$ to define the auxiliary moment restriction (5) and hence $\widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)$ is only to stress on the observability of Z_2. It is superfluous for our purpose and the presentation remains valid without it.

In the second step use (6) for efficiency gain with respect to what corresponds to (7), and hence (4). For this purpose consider the moment restrictions

$$E\left[\widetilde{\widetilde{\phi}}(C,G(Z);\beta)\right]=0 \text{ for } \beta\in\mathcal{B}\Longleftrightarrow\beta=\beta^0. \tag{8}$$

where $\widetilde{\widetilde{\phi}}(C, G(Z); \beta)$ is the residual from another projection and is defined as follows

$$\begin{aligned}\widetilde{\widetilde{\phi}}(C, G(Z); \beta) &:= \widetilde{\phi}(.; \beta) - E\left[\widetilde{\phi}(.; \beta)\phi_1(.)'|G_1(Z)\right]\left(E\left[\phi_1(.)\phi_1(.)'|G_1(Z)\right]\right)^{-1}\phi_1(.) \\ &= \frac{I(C=3)}{P(C=3|G_1(Z))}\left(m(Z;\beta) - E\left[m(Z;\beta)|G_2(Z)\right]\right) \\ &\quad + \frac{I(C\geq 2)}{P(C\geq 2|G_1(Z))}\left(E\left[m(Z;\beta)|G_2(Z)\right] - E\left[m(Z;\beta)|G_1(Z)\right]\right) + E\left[m(Z;\beta)|G_1(Z)\right].\end{aligned}$$

Efficiency gain follows because $E\left[\widetilde{\phi}(.;\beta^0)\widetilde{\phi}(.;\beta^0)'\right] - E\left[\widetilde{\widetilde{\phi}}(.;\beta^0)\widetilde{\widetilde{\phi}}(.;\beta^0)'\right]$ is positive semi-definite. In fact the result is stronger. Comparing with Proposition 1 for the scenario where $R = 3$ shows that $\widetilde{\widetilde{\phi}}(C, G(Z); \beta) = \varphi_{\lambda=\mathbb{C}}(C, G_C(Z); \beta)$ and hence the second projection leads to semiparametric efficiency. The result remains the same if the order of the two projections are interchanged. Various extensions and a formal demonstration of the result that the semiparametric efficiency bound under (8) is equal to that under (4), (5) and (6) are possible following the original work of Graham (2011).

In Section 3, we study the same topic from a related point of view. We will consider various choices of the target population λ and demonstrate efficiency gain from the availability and, subsequently, optimal use of the auxiliary samples. The above example, in particular, is revisited in Corollary 9.

In this short note we do not study the asymptotic properties of the estimators of β^0. However, it is probably worthwhile to note that estimation of β^0 can be undertaken as a standard exercise in semiparametric GMM as follows. Plug in suitably chosen 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

$$\widehat{\beta}_\lambda^{GMM}(W) = \arg\min_{\beta\in\mathcal{B}} \widehat{\varphi}_{\lambda,N}(C, G_C(Z); \beta)' W_N \widehat{\varphi}_{\lambda,N}(C, G_C(Z); \beta), \tag{9}$$

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 on the rate of convergence of the first-step nonparametric estimators required for consistency and (importantly) asymptotic normality of $\widehat{\beta}_\lambda^{GMM}(W)$ are described in Newey (1997), Chen et al. (2008), Cattaneo (2010), Rothe and Firpo (2012), etc. for various choices of nonparametric estimators. $\widehat{\beta}_\lambda^{GMM}(W)$ is semiparametrically efficient if $W = V_\lambda^{-1}$.

One can also use first-step parametric estimators $\widetilde{P}(C \geq r|G_1(Z))$ and $\widetilde{E}[m(Z;\beta)|G_r(Z)]$, for $r = 1, \ldots, R$ to be plugged-in in $\varphi_\lambda(C, G_C(Z); \beta)$. The parametric GMM estimator of β^0 is obtained by replacing the plugged-in vector, denoted by $\tilde{\varphi}_\lambda(C, G_C(Z); \beta)$, in (9). The estimator 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 doubly robust property introduced by Robins et al. (1994), Scharfstein et al. (1999), etc.[5] 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.[6]

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 lesser than Ω_λ (in a matrix sense) if $P(C \geq r|G_1(Z))$ is known up to some finite-dimensional parameter, as is virtually required for the local efficiency of the parametric GMM estimator. It is even lesser if $P(C \geq r|G_1(Z))$ is completely known. Such cases may arise when the missingness/coarsening is by design. 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 $d_m \times d_m$ 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*

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

[5] See Appendix B.6 for the demonstration of the doubly robust property. See Chaudhuri and Min (2012) for more on parametric GMM estimation where the parameter of interest is defined possibly in terms of sub-populations, but only in a two-level missing data model.

[6] Another GMM estimator of β^0, known as the IPW-GMM estimator, can be obtained using the IPW part $\frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \frac{I(C=R)}{P(C=R|G_1(Z))} m(Z;\beta)$ of $\varphi_\lambda(C, G_C(Z); \beta)$ in (3) as the moment vector. When the unknown $P(C = R|G_1(Z))$ is replaced by a parametric estimator, the IPW-GMM estimator is consistent if the parametric model for $P(C = r|G_1(Z))$ is correct, but is still not efficient in general [see Wooldridge (2007)]. Replacing $P(C = R|G_1(Z))$ by a nonparametric estimator may lead to efficiency, but the convergence rate of the nonparametric estimator has to be controlled [see Hirano et al. (2003), Chen et al. (2008) and Cattaneo (2010)]. Robins and Ritov (1997), Rothe and Firpo (2012) and Chaudhuri and Guilkey (2013) discuss such requirement and its connection with the non-doubly robust form of the influence function.

Remarks

(i) The proof is given in Appendix A. The expression for $V_{\lambda[k]}$ is derived in Appendix B.2.

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

(iii) 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 $\psi_\lambda(C, G_C(Z)) = \psi_{\lambda[k]}(C, G_C(Z))$. Chen et al. (2008) discuss the anciliarity of $P(C = r|G_1(Z))$ in this context.

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*

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

where, for any two random variables Y *and* X, $\Pi(Y|X)$ *is used to denote the population least squares projection of* Y *on to the linear space spanned by* 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),

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

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 proof is given in Appendix A. The expression for $V_{\lambda[pk]}$ is derived in Appendix B.3.

(ii) 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.

(iii) 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 $\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m(Z)|Z_1]\middle| S_\gamma(C, Z_1)\right) := B\left(E[S_\gamma(C, Z_1)S_\gamma(C, Z_1)']\right)^{-1} S_\gamma(C, Z_1) = 0$. To see this, note that the identities $I(C \in \mathbb{C}) = 1$ and $P(C \in \mathbb{C}) = 1$ imply that

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

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 Auxiliary samples and efficiency gains

In this section we study the efficiency gains from the use of auxiliary samples by applying Proposition 1 with $R = 3$ for various choices of the target population $\lambda \in \Lambda := \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \mathbb{C} = \{1, 2, 3\}\}$ of the moment restrictions in (1). Denote this target population by $\mathcal{D}_\lambda := (C, Z : C \in \lambda)$.

The observed sample $\mathcal{S}^{\mathcal{O}} := \{C_i, G_{C_i}(Z_i)\}_{i=1}^N$ is a collection of N_1, N_2 and $N_3 = N - N_1 - N_2$ units with $(C = 1, G_1(Z) = Z_1)$, $(C = 2, G_2(Z) = (Z_1, Z_2))$ and $(C = 3, G_3(Z) = Z \equiv (Z_1, Z_2, Z_3))$ respectively from the full population $\mathcal{D} := (C, Z : C \in \mathbb{C})$ such that the proportion of units with $(C = r)$ satisfies $\text{plim}_{N\to\infty} N_r/N = P(C = r)$. The same holds for proportions conditional on Z_1.

The primary sample $\mathcal{S}_{\lambda_p} \subset \mathcal{S}^{\mathcal{O}}$ is the collection of $N_{\lambda_p} := \sum_{r\in\lambda_p} N_r$ units for which $C \in \lambda_p$. The subscripts p and a are used to represent *primary* and *auxiliary*. $\mathcal{D}_{\lambda_p} := (C, Z : C \in \lambda_p)$ is the underlying sub-population for $\mathcal{S}_{\lambda_p}$. The notion of primary sample maintained in this paper dictates that $\lambda_p = \lambda$ when $\lambda \neq \mathbb{C}$, whereas $\lambda_p = \{3\}$ when $\lambda = \mathbb{C}$. We repeat that this is similar to Chaudhuri and Min (2012) but differs slightly from Chen et al. (2008).

An auxiliary sample is generally required for identification of β^0. However, when $\{3\} \in \lambda_p$, the units in $\mathcal{S}_{\lambda_p}$ can unbiasedly estimate $E[m(Z; \beta)|C \in \lambda]$ for any $\lambda \in \Lambda$ as

$$E\left[m(Z; \beta)|C \in \lambda\right] = E\left[\frac{P(C \in \lambda|Z_1)}{P(C \in \lambda)}\frac{I(C = 3)}{P(C = 3|Z_1)}m(Z; \beta)\right] \tag{10}$$

by virtue of (2) and assumption (A2), and given the relevant probabilities.

Our focus is on efficiency gains from the use of auxiliary samples; in particular, due to the Z-related information contained in them. Hence we abstract from the issue of identification by ignoring

the fact that $P(C \in \lambda|Z_1)$, $P(C = 3|Z_1)$ and $P(C \in \lambda)$ on the righthand side of (10) are unknown and has to be estimated, and by always maintaining that the auxiliary sample $\mathcal{S}_{\lambda_a}$ under consideration is a collection of N_{λ_a} units from $\mathcal{D}_{\lambda_a} := (C, Z : C \in \lambda_a)$ for some $\lambda_a \in \Lambda$ such that $\{3\} \in \lambda_p \cup \lambda_a$. Also, $\lambda_p \cap \lambda_a = \{\phi\}$ (empty set) under our notion of primary and auxiliary samples.

All the remaining units of the observed sample $\mathcal{S}^{\mathcal{O}}$, i.e., $\mathcal{S}^{\mathcal{O}} \setminus \mathcal{S}_{\lambda_p}$, can be used as auxiliary samples, in which case $N_{\lambda_a} = N - N_{\lambda_p}$. This is how we defined auxiliary samples in the introduction. However, computational convenience may encourage the use of only a subset of $S^{\mathcal{O}} \setminus \mathcal{S}_{\lambda_p}$ as the auxiliary sample. In the rest of this section we demonstrate the potential loss in efficiency when the used sample $\mathcal{S}^{\mathcal{U}} = (\mathcal{S}_{\lambda_p}, \mathcal{S}_{\lambda_a})$ is a subset of the observed sample $\mathcal{S}^{\mathcal{O}}$, and point out the key elements driving the efficiency loss. We hope this enables an informed judgement on when to choose computational convenience and the consequences of such choices.

To fix ideas consider an example where $\lambda = \lambda_p = \{1\}$ [Corollary 4]. Two choices of auxiliary samples $\mathcal{S}_{\lambda_a}$ correspond to $\lambda_a = \{3\}$ and $\lambda_a = \{2, 3\}$. Since an estimator of β^0 from the former is based on $N_1 + N_3$ sample units whereas that from the latter on a superset of $N = N_1 + N_2 + N_3$ sample units, it is trivial that the latter gives more efficiency.[7] Our interest, however, is in characterizing the efficiency gain by isolating the key information content from the richer auxiliary sample that drives the gain; and not merely in inferring gains from using more observations [Corollaries 4 – 9].

For this purpose, we compare the efficiency bounds under various choices of λ (equivalently λ_p) and λ_a in Corollaries 4 – 9. To account for the different sample sizes involved [see remark (i) of Corollary 4], the bounds are obtained by taking $\mathcal{D}_{\lambda_p} \cup \mathcal{D}_{\lambda_a} = (C, Z : C \in \lambda_p \cup \lambda_a)$ as the population for the used sample $\mathcal{S}^{\mathcal{U}} = (\mathcal{S}_{\lambda_p}, \mathcal{S}_{\lambda_a})$ or, equivalently, $(C, G_C(Z) : C \in \lambda_p \cup \lambda_a)$ as the observed data in Proposition 1.[8] For $\lambda \in \Lambda$, the efficiency bounds in all cases have the common form

$$\Omega_{\lambda}^{\lambda_p, \lambda_a} := \left(M_{\lambda}' \left(V_{\lambda}^{\lambda_p, \lambda_a} \right)^{-1} M_{\lambda} \right)^{-1}.$$

Therefore, we can focus on $V_{\lambda}^{\lambda_p, \lambda_a}$ for the discussion of efficiency gains from the auxiliary samples.

[7] We work with the varying sample size framework to preserve the spirit that the cost of observing the elements of Z increases progressively from Z_1 to Z_2 to Z_3, a feature that we attempted to capture by monotone coarsening.

[8] Intuitively, this amounts to a second round of coarsening when $\mathcal{S}^{\mathcal{U}} \subset \mathcal{S}^{\mathcal{O}}$, if $\mathcal{S}^{\mathcal{O}}$ is assumed to be already available. For example, let $\lambda_p = \{3\}$ and $\lambda_a = \{1\}$. The second round of coarsening deletes the observed data with $C = 2$. This maintains the relative proportion (marginal and conditional on Z_1) of units with $C = 1$ and $C = 3$ similar to the well known IIA assumption in multinomial logit models, but changes the sample size involved in estimation. Our strategy to account for the issue of changed sample size is described in the text below. Another type of second round of coarsening can also be motivated from empirical practice. In the context of this example, the second type deletes Z_2 from the units with $C = 2$. This does not change the sample size involved but changes the relative proportion (marginal and conditional on Z_1) of units with $C = 1$ and $C = 3$. Here we continue with the first type of second round of coarsening because it allows to (i) consider a variety of cases under our simple setup without requiring that $\{1\} \in \lambda_p \cup \lambda_a$, and (ii) entertain the idea of collecting additional and presumably less costly auxiliary samples for efficiency gains.

Since $\lambda_p = \lambda$ when $\lambda \neq \mathbb{C}$, and $\lambda_p = \{3\}$ when $\lambda = \mathbb{C}$ let us, henceforth, denote $V_\lambda^{\lambda_p,\lambda_a}$ by $V_\lambda^{\lambda_a}$.

We point out the relevant feature of the joint distribution of Z that drives the efficiency gains due to the availability of the auxiliary samples. See Appendix A for the proofs of all the corollaries.

Corollary 4 *Let the target population in* (1) *be* $\lambda = \{1\}$*, and hence* $\lambda_p = \{1\}$*. Choices of* λ_a *considered are* $\lambda_a = \{3\}$ *and* $\lambda_a = \{2,3\}$*. Under assumption A and* (2)*,*

$$
\begin{aligned}
V_{\lambda=\{1\}}^{\lambda_a=\{3\}} &= \frac{E\left[E[m(Z)|Z_1]E[m(Z)'|Z_1] + \frac{P(C=1|Z_1)}{P(C=3|Z_1)}Var(m(Z)|Z_1)\middle| C=1\right]}{P(C=1|C\in\{1,3\})}, \\
V_{\lambda=\{1\}}^{\lambda_a=\{2,3\}} &= \frac{V_{\lambda=\{1\}}^{\lambda_a=\{3\}}}{P(C\in\{1,3\})} \\
&\quad - E\left[\frac{P^2(C=1|Z_1)}{P^2(C=1)}\left[\frac{1}{P(C=3|Z_1)} - \frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right].
\end{aligned}
$$

Remarks

(i) The first term in $V_{\lambda=\{1\}}^{\lambda_a=\{2,3\}}$ is a scaled version of $V_{\lambda=\{1\}}^{\lambda_a=\{3\}}$. Consider the sampling variance (taken as the asymptotic variance divided by the sample size involved) of the two semiparametrically efficient estimators based on $(\mathcal{S}_{\lambda_p=\{1\}}, \mathcal{S}_{\lambda_a=\{3\}})$ and $(\mathcal{S}_{\lambda_p=\{1\}}, \mathcal{S}_{\lambda_a=\{2,3\}})$ respectively. The scale $P(C\in\{1,3\})$ nullifies the advantage of the latter from using a larger sample size because

$$
\frac{N_{\lambda_p=\{1\}} + N_{\lambda_a=\{3\}}}{N_{\lambda_p=\{1\}} + N_{\lambda_a=\{2,3\}}} = \frac{N_{\lambda_p=\{1\}} + N_{\lambda_a=\{3\}}}{N_{\lambda=\{1\}} + N_{\lambda_a=\{2\}} + N_{\lambda_a=\{3\}}} = \frac{N_1 + N_3}{N} \xrightarrow{N\to\infty} P(C\in\{1,3\}).
$$

(ii) Now consider the second term of $V_{\lambda=\{1\}}^{\lambda_a=\{2,3\}}$ that is subtracted from the first term. This is the key term. It is positive definite under (A2) unless $Var(E[m(Z)|Z_1,Z_2]|Z_1) = 0$ almost surely in Z_1. Therefore, the only effect of the inclusion of units with $(C=2)$ is a decrease in the contribution of $Var(E[m(Z)|Z_1,Z_2]|Z_1)$ to the efficiency bound. The amount decreased, however, depends on the sampling and partial observability of the elements of Z reflected by the scalar multiple of $Var(E[m(Z)|Z_1,Z_2]|Z_1)$ inside the expectation. [More on this multiple after Corollary 6.]

Such gains are not always possible [Corollaries 6 and 8]. Other than that, the presentation below is similar and focuses on the second term of $V_\lambda^{\lambda_a}$ for the estimator using the richer auxiliary sample.

Corollary 5 *Let the target population in* (1) *be* $\lambda = \{2\}$*, and hence* $\lambda_p = \{2\}$*. Choices of* λ_a

considered are $\lambda_a = \{3\}$ *and* $\lambda_a = \{1,3\}$. *Under assumption A and* (2),

$$
\begin{aligned}
V_{\lambda=\{2\}}^{\lambda_a=\{3\}} &= \frac{E\left[E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2] + \frac{P(C=2|Z_1)}{P(C=3|Z_1)}Var(m(Z)|Z_1,Z_2)\middle| C=2\right]}{P(C=2|C\in\{2,3\})}, \\
V_{\lambda=\{2\}}^{\lambda_a=\{1,3\}} &= \frac{V_{\lambda=\{2\}}^{\lambda_a=\{3\}}}{P(C\in\{2,3\})} \\
&\quad - E\left[\frac{P^2(C=2|Z_1)}{P^2(C=2)}\left[\frac{1}{P(C=2|Z_1)} - \frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right].
\end{aligned}
$$

Remarks

(i) The result is similar to that in Corollary 4. While $Var(E[m(Z)|Z_1,Z_2]|Z_1)$ remains the key term, the magnitude of the efficiency gain is in general different from Corollary 4 because of a different scaling of $Var(E[m(Z)|Z_1,Z_2]|Z_1)$.

(ii) When $P(C=r|Z_1) = P(C=r)$ in addition to (2), i.e., when the auxiliary samples can be treated as validation samples [though not strictly in the sense of Carrol and Wand (1991), Sepanski and Carrol (1993), etc.], the scales of $Var(E[m(Z)|Z_1,Z_2]|Z_1)$ in Corollaries 4 and 5 are independent of Z_1. Hence the magnitude of the relative efficiency gains with the two different target populations depends only on that of (the squares of) $P(C=2)$ and $P(C=3)$.

Corollary 6 *Let the target population in* (1) *be* $\lambda = \{3\}$, *and hence* $\lambda_p = \{3\}$. *Choices of* λ_a *considered are* $\lambda_a = \{\phi\}$ *(none),* $\lambda_a = \{1\}$, $\lambda_a = \{2\}$ *and* $\lambda_a = \{1,2\}$. *Under assumption A and* (2),

$$
\begin{aligned}
V_{\lambda=\{3\}}^{\lambda_a=\{\phi\}} &= E[m(Z)m(Z)'|C=3], \\
V_{\lambda=\{3\}}^{\lambda_a=\{1\}} &= \frac{P(C\in\{1,3\})}{P(C=3)}V_{\lambda=\{3\}}^{\lambda_a=\{\phi\}}, \\
V_{\lambda=\{3\}}^{\lambda_a=\{2\}} &= \frac{P(C\in\{2,3\})}{P(C=3)}V_{\lambda=\{3\}}^{\lambda_a=\{\phi\}}, \\
V_{\lambda=\{3\}}^{\lambda_a=\{1,2\}} &= \frac{V_{\lambda=\{3\}}^{\lambda_a=\{\phi\}}}{P(C=3)} - E\left[\frac{P^2(C=3|Z_1)}{P^2(C=3)}\left[\frac{1}{P(C=3|Z_1)} - \frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right].
\end{aligned}
$$

Remarks

(i) It can be seen under the premise of the discussion in remark (i) following Corollary 4 that there is no efficiency gain from using $(\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{1\}})$ or $(\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{2\}})$ instead of $(\mathcal{S}_{\lambda_p=\{1\}}, \mathcal{S}_{\lambda_a=\{\phi\}})$, i.e., the primary sample alone without any auxiliary sample.

(ii) However, efficiency gain happens when one uses the two auxiliary samples $\mathcal{S}_{\lambda_a=\{1\}}$ and $\mathcal{S}_{\lambda_a=\{2\}}$ jointly, i.e., when estimation is based on the used sample $\mathcal{S}^{\mathcal{U}} = (\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{1,2\}}) = \mathcal{S}^{\mathcal{O}}$.

(iii) The use of $(\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{\phi\}})$ when $\lambda = \{3\}$ parallels the (natural and optimal) practice in full data standard moment restriction models. On the other hand, the representation here in terms of a super(set) population $\mathcal{D}$ is meant to accommodate for the eventual availability of auxiliary samples that may not be from the target population. ($P(C = r|Z_1) > 0$ was maintained in assumption (A2) for this super population $\mathcal{D}$ to be operative in various such cases considered in this Section.) This gives an interpretation that also parallels the well known complete case IPW estimation, but, again, with unit weights since $\lambda_p = \lambda$. Similar arguments relate the well known complete case AIPW estimation in a two-level missing data model with the use of $(\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{1\}})$ or $(\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{2\}})$. On the other hand, the use of $(\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{1,2\}})$ recognizes all the features of the three-level missing data model to deliver the efficiency gain.

In summary, any improvement in Corollaries 4 – 6 over the context-specific baseline was shown by breaking $V_\lambda^{\lambda_a}$ based on the richer auxiliary sample into two terms such that the first term matched $V_\lambda^{\lambda_a}$ of the baseline (in a sense made precise earlier) whereas the second term was negative and quantified the efficiency gain.[9] We also noted that the key information in the second term of all cases is $Var(E[m(Z)|Z_1, Z_2]|Z_1) = E\left[(E[m(Z)|Z_1, Z_2] - E[m(Z)|Z_1])\,(E[m(Z)|Z_1, Z_2] - E[m(Z)|Z_1])'\right]$, and its scalar multiples determine the magnitude of the efficiency gains under different cases.

Alternatively, one may view the term inside the square brackets of the scalar multiples as pointing to the source of the additional relevant information. In Corollaries 4 and 6, this term is proportional to $P(C = 2|Z_1)$ whereas it is proportional to $P(C = 3|Z_1)$ in Corollary 5. The intuition works by identifying the key auxiliary sample as $\mathcal{S}_{\lambda_a=\{2\}}$ in Corollaries 4 and 6, and $\mathcal{S}_{\lambda_a=\{3\}}$ in Corollary 5.

However, this intuition also leads to the following curious observations. First consider Corollary 6. In spite of being the key auxiliary sample according to this intuition, $\mathcal{S}_{\lambda_a=\{2\}}$ does not give any efficiency gain when used alone. Any gain over the so called complete case estimation ($\lambda_a = \{\phi\}$) requires the use of both auxiliary samples $\mathcal{S}_{\lambda_a=\{2\}}$ (key) and $\mathcal{S}_{\lambda_a=\{1\}}$.[10] Similarly, the result in

[9] The second term is the negative of the variance of

$$\frac{P(C=1|Z_1)}{P(C=1)}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)} - \frac{I(C=3)}{P(C=3|Z_1)}\right](E[m(Z)|Z_1, Z_2] - E[m(Z)|Z_1]) : \text{ in Corollary 4,}$$
$$\frac{P(C=2|Z_1)}{P(C=2)}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)} - \frac{I(C=2)}{P(C=2|Z_1)}\right](E[m(Z)|Z_1, Z_2] - E[m(Z)|Z_1]) : \text{ in Corollary 5,}$$
$$\frac{P(C=3|Z_1)}{P(C=3)}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)} - \frac{I(C=3)}{P(C=3|Z_1)}\right](E[m(Z)|Z_1, Z_2] - E[m(Z)|Z_1]) : \text{ in Corollary 6.}$$

This gives the precise nature of the additional relevant information. The key term is $(E[m(Z)|Z_1, Z_2] - E[m(Z)|Z_1])$ and its variability drives the efficiency gain. [See Tsiatis (2006) for an intuitive explanation in terms of projections in similar contexts.] In the discussion here we will focus on the terms inside the square brackets in the scalar multiples of $(E[m(Z)|Z_1, Z_2] - E[m(Z)|Z_1])$. The ratios of the probabilities outside the square brackets does the act of balancing to the concerned target population and are not interesting in the context of the current discussion.

[10] The apparent contradiction of this result with what would be obtained from a treatment of the topic in a manner

Corollary 5 suggests that the availability of the auxiliary sample $\mathcal{S}_{\lambda_a=\{1\}}$ frees up information in the key auxiliary sample $\mathcal{S}_{\lambda_a=\{3\}}$ and thus leads to efficiency gains beyond matching with the baseline. It is the consideration of multilevel missingness that leads to these observations. A two-level missing data model cannot give this insight as is evident from the other two cases in Corollary 6 where the used sample is $\mathcal{S}^{\mathcal{U}} = (\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{1\}})$ or $\mathcal{S}^{\mathcal{U}} = (\mathcal{S}_{\lambda_p=\{3\}}, \mathcal{S}_{\lambda_a=\{2\}})$.

To explore further under the current setup, we consider three additional examples where the primary sample itself contains the complete case, i.e., $\{3\} \in \lambda_p$. Hence auxiliary samples are not required for identification of β^0. If at all, the auxiliary samples may only help in efficiency gains.

Corollary 7 *Let the target population in* (1) *be* $\lambda = \{1,3\}$*, and hence* $\lambda_p = \{1,3\}$*. Choices of* λ_a *considered are* $\lambda_a = \{\phi\}$ *(none) and* $\lambda_a = \{2\}$*. Under assumption A and* (2)*,*

$$
\begin{aligned}
V_{\lambda=\{1,3\}}^{\lambda_a=\{\phi\}} &= E\left[E[m(Z)|Z_1]E[m(Z)'|Z_1] + Var(m(Z)|Z_1)\middle| C \in \{1,3\}\right], \\
V_{\lambda=\{1,3\}}^{\lambda_a=\{2\}} &= \frac{1}{P(C \in \{1,3\})} V_{\lambda=\{1,3\}}^{\lambda_a=\{\phi\}} \\
&\quad - E\left[\frac{P^2(C \in \{1,3\}|Z_1)}{P^2(C \in \{1,3\})}\left[\frac{1}{P(C=3|Z_1)} - \frac{1}{P(C \geq 2|Z_1)}\right] Var(E[m(Z)|Z_1, Z_2]|Z_1)\right].
\end{aligned}
$$

Remarks

The result is similar to that in Corollary 4. While $Var(E[m(Z)|Z_1, Z_2]|Z_1)$ remains the key term, the magnitude of the efficiency gain is naturally different from Corollary 4.

Corollary 8 *Let the target population in* (1) *be* $\lambda = \{2,3\}$*, and hence* $\lambda_p = \{2,3\}$*. Choices of* λ_a *considered are* $\lambda_a = \{\phi\}$ *(none) and* $\lambda_a = \{1\}$*. Under assumption A and* (2)*,*

$$
\begin{aligned}
V_{\lambda=\{2,3\}}^{\lambda_a=\{\phi\}} &= E\left[E[m(Z)|Z_1, Z_2]E[m(Z)'|Z_1, Z_2] + Var(m(Z)|Z_1, Z_2)\middle| C \in \{2,3\}\right], \\
V_{\lambda=\{2,3\}}^{\lambda_a=\{1\}} &= \frac{1}{P(C \in \{2,3\})} V_{\lambda=\{2,3\}}^{\lambda_a=\{\phi\}}.
\end{aligned}
$$

Remarks

The result is similar to Corollary 6. There is no efficiency gain from using $(\mathcal{S}_{\lambda_p=\{2,3\}}, \mathcal{S}_{\lambda_a=\{1\}})$ instead of $(\mathcal{S}_{\lambda_p=\{2,3\}}, \mathcal{S}_{\lambda_a=\{\phi\}})$, i.e., the primary sample alone without any auxiliary sample. The difference from Corollary 5 can be attributed to the difference in target population for (1).

similar to (4)–(8) is explained by the fact that $\mathcal{D}$ is always maintained as the underlying population in the latter. The latter can be easily accommodated here by working under the premise of known $P(C = r|G_1(Z))$ (similar to Proposition 3); and by taking $\phi(C = 3, G_3(Z); \beta) = \frac{I(C=3)}{P(C=3)} m(Z; \beta)$. Note that under the premise of Corollary 6, however, the multiple $\frac{I(C=3)}{P(C=3)}$ is unity when the used sample is $(\mathcal{S}_{\lambda_p=\{1\}}, \mathcal{S}_{\lambda_a=\{\phi\}})$. This is clarified in the first sentence of remark (iii) following the corollary and explains the difference in conclusion. See Appendix B.7 for details of the treatment along the line of (4)–(8) that highlights the apparent contradiction.

Corollary 9 *Let the target population in* (1) *be* $\lambda = \mathbb{C} = \{1,2,3\}$*, and hence* $\lambda_p = \{3\}$*. Choices of* λ_a *considered are* $\lambda_a = \{1\}$*,* $\lambda_a = \{2\}$ *and* $\lambda_a = \{1,2\}$*. Under assumption A and* (2)*,*[11]

$$
\begin{aligned}
V_{\lambda=\mathbb{C}}^{\lambda_a=\{1\}} &= E\left[\frac{Var(m(Z)|Z_1)}{P(C=3|Z_1)} + \frac{E[m(Z)|Z_1]E[m(Z)'|Z_1]}{P(C\in\{1,3\}|Z_1)}\right] P(C\in\{1,3\}),\\
V_{\lambda=\mathbb{C}}^{\lambda_a=\{2\}} &= E\left[\frac{Var(m(Z)|Z_1,Z_2)}{P(C=3|Z_1)} + \frac{E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2]}{P(C\in\{2,3\}|Z_1)}\right] P(C\in\{2,3\}),\\
V_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}} &= \frac{V_{\lambda=\mathbb{C}}^{\lambda_a=\{1\}}}{P(C\in\{1,3\})} - E\left[\left[\frac{1}{P(C\in\{1,3\}|Z_1)}-1\right]E[m(Z)|Z_1]E[m(Z)'|Z_1]\right]\\
&\quad -E\left[\left[\frac{1}{P(C=3|Z_1)}-\frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right],\\
&= \frac{V_{\lambda=\mathbb{C}}^{\lambda_a=\{2\}}}{P(C\in\{2,3\})} - E\left[\left[\frac{1}{P(C\geq 2|Z_1)}-1\right]E[m(Z)|Z_1]E[m(Z)'|Z_1]\right].
\end{aligned}
$$

Remarks

Efficiency gains due to the use of the auxiliary sample $\mathcal{S}_{\lambda_a=\{1,2\}}$ over the two baseline cases are self evident. There are only two differences in pattern from the rest of the corollaries:

(i) A comparison of $V_{\lambda=\mathbb{C}}^{\lambda_a=\{1\}}$ and $V_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}}$ (first expression) reveals an extra penalty for using only $\mathcal{S}_{\lambda_a=\{1\}}$ as the auxiliary sample and thereby ignoring the units with $C=2$ that actually belong to the target population when $\lambda=\mathbb{C}$. The penalty for the sub-optimal $V_{\lambda=\mathbb{C}}^{\lambda_a=\{1\}}$ is represented as an extra gain for $V_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}}$ by the second term of the latter (first expression). The key information content for this extra gain is $E[m(Z)|Z_1]E[m(Z)'|Z_1]$ [see the discussion before Proposition 10]. Its scale depends on $P(C=2|Z_1)$ and points at its source [see the discussion before Corollary 7]. The third term of $V_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}}$ represents the "usual gain" as in Corollary 7.

(ii) Unlike in other cases, a comparison of $V_{\lambda=\mathbb{C}}^{\lambda_a=\{2\}}$ and $V_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}}$ (second expression) reveals that the key term in the efficiency gain is not $Var(E[m(Z)|Z_1,Z_2]|Z_1)$ but $E[m(Z)|Z_1]E[m(Z)'|Z_1]$. Its scale depends on $P(C=1|Z_1)$ and points at the source of this additional information. A closer inspection reveals that this gain actually corresponds to the extra penalty for the sub-optimal $V_{\lambda=\mathbb{C}}^{\lambda_a=\{2\}}$ similar to the last remark. There is no "usual gain" in efficiency (in the sense of the last line of remark (i)), and hence this does not contradict the result in Corollary 8.

4 Efficiency gains: Simple non-monotone missing/coarsening model

In this section we consider a simple extension of the monotone missing/coarsening model studied so far to a case of non-monotone missing/coarsening and show the efficiency gains due to the use of

[11] The choice $\lambda_a = \{\phi\}$ is not considered here because this relates to the complete case IPW estimation [see Corollary 6]. Efficiency of AIPW (e.g. $\lambda_a = \{1\}$ or $\lambda_a = \{2\}$) over IPW in this context is well understood in the literature.

auxiliary samples. As in Section 2, take $Z = (Z_1, Z_2, Z_3)$. The observed data is $(C, G_C(Z))$ for $C \in \mathbb{C} := 1, 2, 3, \infty$; and $G_1(Z) = Z_1$, $G_2(Z) = (Z_1, Z_2)$, $G_3(Z) = (Z_1, Z_3)$ and $G_\infty(Z) = (Z_1, Z_2, Z_3)$. The subscript "$\infty$" denotes the complete data. Notice that while $G_1(Z) \subset G_2(Z) \subset G_\infty(Z)$ and $G_1(Z) \subset G_3(Z) \subset G_\infty(Z)$, there is no overall monotonicity in the transformation $G_C(Z)$. This is the incremental non-monotonicity over the setup in Section 3, and it allows us to focus on the incremental efficiency gains. See Chaudhuri and Guilkey (2013) for an example of such missingness.

Since (2) ensures that the cases where $\lambda \neq \mathbb{C}$ are straightforward variations of the case where $\lambda = \mathbb{C}$, here we focus on the latter to show how auxiliary samples can help in efficiency gains. Therefore, we take the baseline object of comparison as $V^{\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}}$ from Corollary 9, Section 3. Since Z_2 and Z_3 are treated symmetrically, this does not entail any loss of generality.

Like Section 3, let the observed sample $\mathcal{S}^{\mathcal{O}} := \{C_i, G_{C_i}(Z_i)\}_{i=1}^N$ be a collection of N units from the full population $\mathcal{D} := (C, Z : C \in \mathbb{C})$ such that N_r units are of the type $(C = r, G_r(Z))$ for $r = 1, 2, 3, \infty$. The used sample for the baseline is $\mathcal{S}^{\mathcal{U}}_{\mathcal{B}} = (\mathcal{S}_{\lambda_p=\{\infty\}}, \mathcal{S}_{\lambda_a=\{1,2\}})$ and $\mathcal{S}^{\mathcal{U}}_{\mathcal{B}} \subset \mathcal{S}^{\mathcal{O}}$ since the latter also contains units with $C = 3$ for which $G_3(Z) = (Z_1, Z_3)$ is observed. Therefore, a minor modification (similar to $V^{\lambda_a=\{1\}}_{\lambda=\mathbb{C}}$ or $V^{\lambda_a=\{2\}}_{\lambda=\mathbb{C}}$ in Corollary 9) gives $V^{\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}} \equiv V^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}}$ as:[12]

$$
\begin{aligned}
& V^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}} \\
= \; & E\left[\frac{Var(m(Z)|Z_1, Z_2)}{P(C=\infty|Z_1)} + \frac{Var(E[m(Z)|Z_1, Z_2]|Z_1)}{P(C \in \{2,\infty\}|Z_1)} + \frac{E[m(Z)|Z_1]E[m(Z)'|Z_1]}{P(C \neq 3|Z_1)}\right] P(C \neq 3). \quad (11)
\end{aligned}
$$

Now we show how the incremental Z-information available, i.e., $G_3(Z)$, that also introduces non-monotone missingness in the observed data, can be used for efficiency gains. The general theory for non-monotone missing/coarsening model was developed by Robins and co-authors [see, for example, Gill and Robins (1997) and Vansteelandt et al. (2007)]. However, often the estimators are computationally difficult and closed form expressions for the efficient influence functions are in general not available [see Tsiatis (2006)]. Instead, to keep the exposition of efficiency gains from utilizing $G_3(Z)$ simple, we consider two alternative routes — (i) direct augmentation of the moment vector, and (ii) simplifying assumptions — that have not received much attention in this literature.

[12] $\varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}$, such that $V^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}} := E\left[\varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}} \varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}'}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}} | C \neq 3\right]$, is given by

$$
\frac{P(C \neq 3)}{P(C \neq 3|Z_1)}\left[\frac{I(C=\infty)\,(m(Z) - E[m(Z)|Z_1, Z_2))}{P(C=\infty|Z_1, C \neq 3)} + \frac{I(C \in \{2,\infty\})\,(E[m(Z)|Z_1, Z_2) - E[m(Z)|Z_1])}{P(C \in \{2,\infty\}|Z_1, C \neq 3)} + I(C \neq 3)E[m(Z)|Z_1]\right].
$$

We maintain the spirit that additional auxiliary samples involve additional sample units, as described before in Section 3. This is reflected in the expression of $\varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}$ by the ratio of probabilities outside the square bracket and the indicator $I(C \neq 3)$ for the last term; both adjust for the fact that the observed data is from the population $(C, Z : C = 1, 2, \infty)$ whereas the moment restriction in (1) is defined in the full population $(C, Z : C = 1, 2, 3, \infty)$. Simple changes can alternatively accommodate the second type of second round coarsening described in footnote 8.

The idea of direct augmentation of the moment vector is similar to Chaudhuri and Guilkey (2013). For example, consider the moment vector

$$\varphi^{\text{augment}}_{\lambda=\mathbb{C}}(C, G_C(Z); \beta) := \left[\varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}(\beta)', \varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,3\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}(\beta)'\right]'. \tag{12}$$

$\varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}(\beta)$ is defined in footnote 12 where the expression is for $\varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}(\beta)$ evaluated at $\beta = \beta^0$; the occurrences of β are in the function $m(Z;\beta)$. $\varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,3\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}(\beta)$ is its counterpart based on the used sample $(\mathcal{S}_{\lambda_p=\{\infty\}}, \mathcal{S}_{\lambda_a=\{1,3\}}) \subset \mathcal{S}^{\mathcal{O}}$, i.e., by interchanging the roles of the units with $(C = 2, G_2(Z))$ and $(C = 3, G_3(Z))$. The asymptotic variance of the efficient GMM estimator based on the augmented moment vector (12) cannot exceed the baseline (11). Other augmenting moment vectors can also be considered. Moreover, often a weak notion of efficiency can be attached to such estimators [see Remark (2) following Proposition 1 in Chaudhuri and Guilkey (2013)].

Now consider the second route, i.e., simplifying assumptions. In particular, let us assume that

$$Z_2 \perp Z_3 | Z_1. \tag{13}$$

This allows us to work with closed form expressions of "locally" efficient influence functions.[13] See Anderson and Perlman (1991) and the references therein for more on the scope of this assumption. Considerable generality is lost, and hence we keep the following discussion short and informal.

Consider the class of estimators $\widehat{\beta}$ whose asymptotically linear representation is

$$\sqrt{N}\left(\widehat{\beta} - \beta^0\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(C_i, G_{C_i}(Z_i)) + o_p(1) \tag{14}$$

where $\psi(C, G_C(Z)) := -\Omega M' V^{-1} \varphi(C, G_C(Z))$, $M := E\left[\frac{\partial}{\partial \beta'} m(Z; \beta^0)\right]$, $\Omega := \left(M' V^{-1} M\right)^{-1}$,

$$\begin{aligned} \varphi(C, G_C(Z); \beta) &:= \varphi_{[1,\infty]}(C, G_C(Z); \beta) + \varphi_{[2]}(C, G_C(Z); \beta) + \varphi_{[3]}(C, G_C(Z); \beta), \\ \text{and } V &:= E\left[\varphi(C, G_C(Z); \beta^0)\varphi(C, G_C(Z); \beta^0)'\right]. \end{aligned}$$

The representation of $\varphi(C, G_C(Z); \beta)$ as the sum of three components is to highlight the contribution

[13] Cattaneo (2010) makes another simplifying assumption for closed form expressions. Taken in the present context, this avoids the main problem of working through the recursive steps of approximating an inverse operator to obtain the efficient influence function by ruling out the case $C = \infty$. It compensates for the absence of $C = \infty$ on identification by taking $m(Z; \beta) = [m_2(Z_1, Z_2; \alpha_2)', m_3(Z_1, Z_2; \alpha_3)']'$ where $\beta = (\alpha_2', \alpha_3')'$ so that $C = 2$ and $C = 3$ can be treated as the respective complete cases for the two sets of rows in $m(Z; \beta)$ giving the two sets of moment restrictions.

of the various types of units $(C, G_C(Z))$. These three components are, respectively,

$$\begin{aligned}
\varphi_{[1,\infty]}(C, G_C(Z); \beta) &:= \frac{I(C=\infty)}{P(C=\infty|G_1(Z))}\left(m(Z;\beta) - E[m(Z;\beta)|G_1(Z)]\right) + E[m(Z;\beta)|G_1(Z)], \\
\text{and } \varphi_{[r]}(C, G_C(Z); \beta) &:= \frac{P(C=r|G_1(Z))}{P(C\in\{r,\infty\}|G_1(Z))}\left(E[m(Z;\beta)|G_r(Z)] - E[m(Z;\beta)|G_1(Z)]\right) \\
&\quad \times \left[\frac{I(C=r)}{P(C=r|G_1(Z))} - \frac{I(C=\infty)}{P(C=\infty|G_1(Z))}\right] \text{ for } r=2,3.
\end{aligned}$$

The expression of $\varphi_{[1,\infty]}(C, G_C(Z); \beta)$ relates to the standard two-level missing data model where $G_1(Z) = G_2(Z) = G_3(Z) = Z_1$. Under the premise of Corollary 9 in Section 2, i.e., when $G_2(Z)$ and $G_3(Z)$ do not exist, its final term $E[m(Z;\beta)|G_1(Z)]$ would require a non-unity weight $P(C \in \{1,\infty\})/P(C \in \{1,\infty\}|G_1(Z))$. The contributions of the units with $(C = \infty, G_\infty(Z))$ and $(C = 1, G_1(Z))$ are well understood. The contribution of the units with $(C = 2, G_2(Z))$ in this context can be understood by considering the role played by $\varphi_{[2]}(C, G_C(Z); \beta)$ that is evident from comparing

$$\begin{aligned}
&\varphi_{[1,\infty]}(C, G_C(Z); \beta) + \varphi_{[2]}(C, G_C(Z); \beta) \\
= \;& \frac{I(C=\infty)\left(m(Z) - E[m(Z)|Z_1, Z_2)\right)}{P(C=\infty|Z_1)} + \frac{I(C\in\{2,\infty\})\left(E[m(Z)|Z_1, Z_2) - E[m(Z)|Z_1]\right)}{P(C\in\{2,\infty\}|Z_1)} + E[m(Z)|Z_1]
\end{aligned}$$

with the expression of $\varphi^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}(\beta)$ in footnote 12. Therefore, the key efficiency gain over the baseline has to come from the third term $\varphi_{[3]}(C, G_C(Z); \beta)$. This is seen by noting that (13) gives [14]

$$\begin{aligned}
V &= \frac{V^{\lambda_p=\{\infty\},\lambda_a=\{1,2\}}_{\lambda=\mathbb{C}=\{1,2,3,\infty\}}}{P(C\neq 3)} - E\left[\left[\frac{1}{P(C\neq 3|Z_1)} - 1\right] E[m(Z)|Z_1]E[m(Z)'|Z_1]\right] \\
&\quad -E\left[\frac{P(C=3|Z_1)Var(E[m(Z)|Z_1, Z_3]|Z_1)}{P(C\in\{3,\infty\}|Z_1)P(C=\infty|Z_1)}\right].
\end{aligned}$$

As before, the first term on the righthand side adjusts for the effect of different sample sizes on the sampling variance. This is not interesting for our purpose. The second term is the extra gain in efficiency similar to Corollary 9. Equivalently, this is the extra penalty for using the sub-optimal $\mathcal{S}^{\mathcal{U}}_{\mathcal{B}} = (S_{\lambda_p=\{\infty\}}, S_{\lambda_a=\{1,2\}})$ that ignores the Z_1-information contained in $G_3(Z)$. The third term

[14] For notational convenience suppress the arguments Z and β^0, and take m as a scalar. Therefore,

$$\begin{aligned}
V &= E\left[\frac{Var(m|Z_1, Z_2)}{P(C=\infty|Z_1)} + \frac{Var(E[m|Z_1, Z_2]|Z_1)}{P(C\in\{2,\infty\}|Z_1)} + E[m^2|Z_1]\right] + E\left[\frac{P(C=3|Z_1)Var(E[m|Z_1, Z_3]|Z_1)}{P(C\in\{3,\infty\}|Z_1)P(C=\infty|Z_1)}\right] \\
&\quad -2E\left[\frac{P(C=3|Z_1)(m - E[m|Z_1, Z_2])(E[m|Z_1, Z_3] - E[m|Z_1])}{P(C\in\{3,\infty\}|Z_1)P(C=\infty|Z_1)}\right] \\
&\quad -2E\left[\frac{P(C=3|Z_1)(E[m|Z_1, Z_2] - E[m|Z_1])(E[m|Z_1, Z_3] - E[m|Z_1])}{P(C\in\{2,\infty\}|Z_1)P(C\in\{3,\infty\}|Z_1)}\right].
\end{aligned}$$

Under (13), the last term is zero, and the term in the second line is twice the second term in the first line.

captures the key efficiency gain due to the Z-information contained in $G_3(Z)$ that is accomplished by the optimal use of the observed sample $\mathcal{S}^{\mathcal{O}}$ under (13). Under (13), $\widehat{\beta}$ is asymptotically at least as efficient as the efficient GMM estimator based on (12) [see Appendix B.8 for details]. In fact, as we formally show in Proposition 10, $\widehat{\beta}$ is locally efficient under (13). (Appendix A contains the proof.)

Proposition 10 *Consider the coarsening setup of this section where* $\mathbb{C} := \{1, 2, 3, \infty\}$. *Let* β^0 *be defined by* (1) *with* $\lambda = \mathbb{C}$. *Let assumption A (with the complete case denoted by* ∞ *instead of* $R = 3$*) and* (2) *hold. Let* V *be a* $d_m \times d_m$ *finite positive definite matrix. Then any regular estimator* $\widehat{\beta}$ *whose asymptotically linear representation is given by* (14) *is (locally) efficient for* β^0 *when* (13) *holds.*

As in the discussion following Proposition 1 (remark (iii)), an alternative way to delineate the contribution of each level of coarsened data is to work with the sequential projection similar to the Frisch-Waugh-Lovell theorem on the following set of moment restrictions:

$$
\begin{aligned}
E\left[\phi(C=\infty, G_\infty(Z); \beta)\right] &= 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0, && (15)\\
E\left[\phi_3(C=\{3,\infty\}, G_3(Z))|G_3(Z)\right] &= 0 \text{ almost surely } G_3(Z), && (16)\\
E\left[\phi_2(C=\{2,\infty\}, G_2(Z))|G_2(Z)\right] &= 0 \text{ almost surely } G_2(Z), && (17)\\
E\left[\phi_1(C, G_C(Z))|G_1(Z)\right] &= 0 \text{ almost surely } G_1(Z), && (18)
\end{aligned}
$$

where, the moment vectors are, as before,

$$
\begin{aligned}
&\phi(C=\infty, G_\infty(Z); \beta) := \frac{I(C=\infty)}{P(C=\infty|G_1(Z))} m(Z;\beta),\\
&\phi_r(C=\{r,\infty\}, G_r(Z)) := I(C=\{r,\infty\})\left[I(C=\infty) - P(C=\infty|C\in\{r,\infty\}, G_r(Z))\right], \text{ for } r=2,3, \text{ and}\\
&\phi_1(C, G_C(Z)) := \left[I(C=\infty) - P(C=\infty|G_1(Z)), I(C=3) - P(C=3|G_1(Z)), I(C=2) - P(C=2|G_1(Z))\right]'.
\end{aligned}
$$

To appreciate the role played by the conditional independence assumption (13), let us first work without imposing it. Now, continuing as before, it is straightforward to see that the following set of working moment restrictions *successively* lead to greater efficiency gains:

$$
\begin{aligned}
E\left[\widetilde{\phi}(C=\{3,\infty\}, G_3(Z), G_\infty(Z); \beta)\right] &= 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0 \text{ [based on (15)–(16)]},\\
E\left[\widetilde{\widetilde{\phi}}(C=\{2,3,\infty\}, G_2(Z), G_3(Z), G_\infty(Z); \beta)\right] &= 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0 \text{ [based on (15)–(17)]},\\
E\left[\widetilde{\widetilde{\phi}}(C, G_C(Z); \beta)\right] &= 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta^0 \text{ [based on (15)–(18)]}.
\end{aligned}
$$

The presentation would possibly be more informative if the successive moment restrictions considered were in the order [(15), (18)], [(15), (18), (17)] and [(15), (18), (17), (16)]. However, the current form maintains uniformity with the discussion following Proposition 1. More on the order and its relation with the conditional independence assumption in (13) is discussed below.

The moment vectors $\widetilde{\phi}(C=\{3,\infty\}, G_3(Z), G_\infty(Z);\beta)$, $\widetilde{\widetilde{\phi}}(C=\{2,3,\infty\}, G_2(Z), G_3(Z), G_\infty(Z);\beta)$ and $\widetilde{\widetilde{\widetilde{\phi}}}(C, G_C(Z);\beta)$ are residuals from successive projections and are defined as:

$$
\begin{aligned}
& \widetilde{\phi}(C=\{3,\infty\}, G_3(Z), G_\infty(Z);\beta) \\
:= \;& \phi(.;.) - E\left[\phi(.;.)\phi_3(.)'|G_3(Z)\right]\left(E\left[\phi_3(.)\phi_3(.)'|G_3(Z)\right]\right)^{-1}\phi_3(.) \\
= \;& \frac{I(C=\infty)}{P(C=\infty|G_1(Z))}\left(m(Z;\beta) - E[m(Z;\beta)|G_3(Z)]\right) + \frac{I(C\in\{3,\infty\})}{P(C\in\{3,\infty\}|G_1(Z))}E[m(Z;\beta)|G_3(Z)].
\end{aligned}
$$

$$
\begin{aligned}
& \widetilde{\widetilde{\phi}}(C=\{2,3,\infty\}, G_2(Z), G_3(Z), G_\infty(Z);\beta) \\
:= \;& \widetilde{\phi}(.;\beta) - E\left[\widetilde{\phi}(.;\beta)\phi_2(.)'|G_2(Z)\right]\left(E\left[\phi_2(.)\phi_2(.)'|G_2(Z)\right]\right)^{-1}\phi_2(.) \\
= \;& \frac{I(C=\infty)}{P(C=\infty|G_1(Z))}\left(m(Z;\beta) - E[m(Z;\beta)|G_3(Z)] - E[m(Z;\beta)|G_2(Z)]\right) \\
& + \frac{I(C\in\{3,\infty\})}{P(C\in\{3,\infty\}|G_1(Z))}E[m(Z;\beta)|G_3(Z)] + \frac{I(C\in\{2,\infty\})}{P(C\in\{2,\infty\}|G_1(Z))}E[m(Z;\beta)|G_2(Z)] \\
& + \left[\frac{I(C\in\{2,\infty\})}{P(C\in\{2,\infty\}|G_1(Z))} - \frac{I(C=\infty)}{P(C=\infty|G_1(Z))}\right]\frac{P(C=3|G_1(Z))}{P(C\in\{3,\infty\}|G_1(Z))}E\left[[m(Z;\beta)|G_3(Z)]\,|G_2(Z)\right].
\end{aligned}
$$

$$
\begin{aligned}
& \widetilde{\widetilde{\widetilde{\phi}}}(C, G_C(Z);\beta) \\
:= \;& \widetilde{\widetilde{\phi}}(.;\beta) - E\left[\widetilde{\widetilde{\phi}}(.;\beta)\phi_2(.)'|G_1(Z)\right]\left(E\left[\phi_1(.)\phi_1(.)'|G_1(Z)\right]\right)^{-1}\phi_1(.) \\
= \;& \frac{I(C=\infty)}{P(C=\infty|G_1(Z))}\left(m(Z;\beta) - E[m(Z;\beta)|G_3(Z)] - E[m(Z;\beta)|G_2(Z)] + E[m(Z;\beta)|G_1(Z)]\right) \\
& + \frac{I(C\in\{3,\infty\})}{P(C\in\{3,\infty\}|G_1(Z))}E[m(Z;\beta)|G_3(Z)] + \frac{I(C\in\{2,\infty\})}{P(C\in\{2,\infty\}|G_1(Z))}E[m(Z;\beta)|G_2(Z)] \\
& + E[m(Z;\beta)|G_1(Z)] + \left[\frac{I(C\in\{2,\infty\})}{P(C\in\{2,\infty\}|G_1(Z))} - \frac{I(C=\infty)}{P(C=\infty|G_1(Z))}\right]\frac{P(C=3|G_1(Z))}{P(C\in\{3,\infty\}|G_1(Z))} \\
& \qquad \times \left(E\left[[m(Z;\beta)|G_3(Z)]\,|G_2(Z)\right] - E[m(Z;\beta)|G_1(Z)]\right).
\end{aligned}
$$

The conditional independence assumption implies $E\left[[m(Z;\beta)|G_3(Z)]\,|G_2(Z)\right] = E[m(Z;\beta)|G_1(Z)]$. Therefore, the last term of $\widetilde{\widetilde{\widetilde{\phi}}}(C, G_C(Z);\beta)$ is zero. A simple rearrangement of the remaining terms shows that $\widetilde{\widetilde{\widetilde{\phi}}}(C, G_C(Z);\beta) = \varphi(C, G_C(Z);\beta)$, i.e., the relevant component in the asymptotically linear representation of $\widehat{\beta}$ in (14). Thus $\widetilde{\widetilde{\widetilde{\phi}}}(C, G_C(Z);\beta)$ inherits the efficiency property in Proposition

10. The approach of sequential projection presents an alternative way of appreciating the contribution of each level of coarsened data or, in the terminology used in our paper, each type of auxiliary samples.

A remark on the role played by the conditional independence assumption is in order. In the monotone missing example following Proposition 1 it is straightforward to see that the (sequential) order in which the projections are performed does not matter. The conditional independence assumption in (13) ensures the same in the non-monotone missing case. Without (13), the approach of sequential projection by extending the insight of Brown and Newey (1998) and Graham (2011), however, does not lead to a unique moment vector $\overset{\approx}{\phi}(C, G_C(Z); \beta)$ (in the case above).

Case by case modifications of our presentation similar to Sections 3 and 4 are straightforward extensions. In applications where the additional assumptions can be justified and the extra computations are not prohibitive, the results in this paper should encourage the optimal use of the observed data whereby one extracts all the relevant information from the auxiliary samples for efficiency gains.

References

Anderson, S. A. and Perlman, M. D. (1991). Lattice-ordered conditional independence models for missing data. *Statistics and Probability Letters*, 12: 465–486.

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

Busso, M., DiNardo, J., and McCrary, J. (2009). Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects. Mimeo.

Carrol, R. and Wand, M. (1991). Semiparametric Estimation in Logistic Measurement Error Models. *Journal of Royal Statistical Society, Series B*, 53: 573–585.

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

Chaudhuri, S. and Guilkey, D. K. (2013). GMM with Multiple Missing Variables. 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. (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *Annals of Statistics*, 36: 808–843.

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. and Robins, J. M. (1997). Sequential Models for Coarsening and Missingness. In Lin, D. Y. and Fleming, T. R., editors, *Proceedings of The First Seattle Symposium in Biostatistics: Survival Analysis*, Lecture Notes in Statitsics, pages 295–305. New York: Springer-Verlag.

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.

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

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

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

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

Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores. *Econometrica*, 71: 1161–1189.

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

Lunceford, J. and Davidian, M. (2004). Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects : A Comparative Study. *Statistics in Medicine*, 23: 2937–2960.

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.

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.

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.

Sepanski, J. and Carrol, R. (1993). Semiparametric Quasi-likelihood and Variance Estimation in Measurement Error Models. *Journal of Econometrics*, 58: 223–256.

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

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for mean of repeated outcomes under nonignorable nonmotonone nonresponse. *Biometrika*, 94: 841–860.

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

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

A Appendix: Proof of main results

We use f and F to denote the density and distribution function, 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:

There are $2^R - 1$ separate statements in the proposition, each corresponding to only one $\lambda \in \Lambda$ used to define β^0 via the moment restrictions in (1). The proofs of all statements are similar. Hence to avoid repetition, we work with a generic λ. We need to use the relationship (20) 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 semiparameteric 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:

$$\mathcal{T} := a(Z_1) + \sum_{r=1}^{R} I(C = r) \frac{b_r(Z_1)}{a_r(Z_1)} + \sum_{r=2}^{R} I(C \geq r) a(Z_1, \ldots, Z_r), \tag{19}$$

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 the verify-out-of-sample cases and also the verify-in-sample case. 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:

$$\begin{aligned} & s(Z_1) + I(C \in \lambda)\frac{\dot{P}(C \in \lambda|Z_1)}{P(C \in \lambda|Z_1)} + I(C \notin \lambda)\frac{\dot{P}(C \notin \lambda|Z_1)}{P(C \notin \lambda|Z_1)} \\ = \; & I(C \in \lambda)\left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(Z_1|C \in \lambda)\right] + I(C \notin \lambda)\left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(Z_1|C \notin \lambda)\right]. \end{aligned} \tag{20}$$

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 hold

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

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 $\psi(A, C, G_C(Z)) \in \mathcal{T}$ (or more precisely, in the mean square closure of $\{B_{d_\beta \times d_\theta} t(C, G_C(Z)) : t(C, G_C(Z)) \in \mathcal{T}$ and any constant matrix $B\}$) such that

$$E[\psi(A, C, G_C(Z))S(C, G_C(Z))'] = \frac{\partial \beta^0(\theta_0)}{\partial \theta'}. \tag{21}$$

We do this by verifying (21) after conjecturing that $\psi(A, C, G_C(Z)) = -(AM_\lambda)^{-1}A\varphi_\lambda(C, G_C(Z))$. Verification of (21) is equivalent to showing that

$$E[\varphi_\lambda(C, G_C(Z))S(C, G_C(Z))'] = E\left[m(Z;\beta^0)\left\{s(Z_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_r|Z_1,\ldots,Z_{r-1})'\right\}\middle| C \in \lambda\right]. \tag{22}$$

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 (19), 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 (20) 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

$$\begin{aligned}
& E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}\varphi_{(1)}(C,G_C(Z))S(C,G_C(Z))'\right] \\
= \;& E\left[E[m(Z;\beta^0)|Z_1]|C\in\lambda\right]\frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)} + E\left[E[m(Z;\beta^0)|Z_1]s(Z_1|C\in\lambda)'|C\in\lambda\right] \\
& - E\left[E[m(Z;\beta^0)|Z_1]\frac{\dot{P}(C\in\lambda|Z_1)'}{P(C\in\lambda)}\right] + E\left[E[m(Z;\beta^0)|Z_1]\frac{\dot{P}(C\in\lambda|Z_1)'}{P(C\in\lambda)}\right] \\
= \;& E\left[m(Z;\beta^0)|C\in\lambda\right]\frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)} + E\left[E[m(Z;\beta^0)|Z_1]s(Z_1|C\in\lambda)'|C\in\lambda\right] + 0 \\
= \;& 0 + E[m(Z;\beta^0)s(Z_1|C\in\lambda)'|C\in\lambda] + 0 \qquad (23)
\end{aligned}$$

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

$$\begin{aligned}
& E\left[\frac{P(C\in\lambda|G_1(Z))}{P(C\in\lambda)}\varphi_{(r)}(C,G_C(Z))S(C,G_C(Z))'\right] \\
= \;& E\left[\frac{P(C\in\lambda|Z_1)}{P(C\in\lambda)}\left(E[m(Z;\beta^0)|Z_1,\ldots,Z_r] - E[m(Z;\beta^0)|Z_1,\ldots,Z_{r-1}]\right)\sum_{s=r}^{R} s(Z_s|Z_1,\ldots,Z_{s-1})\right] \\
= \;& E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m(Z;\beta^0)|Z_1,\ldots,Z_r]s(Z_r|Z_1,\ldots,Z_{r-1})'\right] \\
= \;& E\left[m(Z;\beta^0)s(Z_r|Z_1,\ldots,Z_{r-1})'|C\in\lambda\right] \qquad (24)
\end{aligned}$$

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 (19), along with (2).

Therefore, (23) and (24) verify (22), and hence (21). That $\varphi_\lambda(C, G_C(Z))$ belongs to $\mathcal{T}$ in (19) 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 (20) 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 (19) and also (20), are zeros. Rest of the terms in $\mathcal{T}$ (including the additional 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 $-(AM_1)^{-1}Am(Z;\beta^0)$. 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 (22) 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:

$$\mathcal{T} := a(Z_1) + \sum_{r=2}^{R} I(C \geq r) a(Z_1, \ldots, Z_r), \tag{25}$$

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 (20) 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 (22), pathwise differentiability is established by verifying that $\varphi_{\lambda[k]}(C, G_C(Z)) \in \mathcal{T}$ in (25) 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: [$P(C = r|G_1(Z)) = P(C = r|G_1(Z); \gamma^0)$ for $r = 1, 2, \infty$. $\gamma^0 \in \Gamma \subset \mathbb{R}^{d_\gamma}$ 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

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

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:

$$\mathcal{T} := a(Z_1) + B' S_\gamma(C|Z_1) + \sum_{r=2}^{R} I(C \geq r) a(Z_1, \ldots, G_r), \tag{26}$$

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 (20) as follows

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

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

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

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 (20) above. The last line uses (2).

As in (22), pathwise differentiability is established by verifying that $\varphi_{\lambda[pk]}(C, G_C(Z)) \in \mathcal{T}$ in (26)

satisfies

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

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

$$\begin{aligned} & E\left[\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m(Z;\beta^0)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right)S(C,G_C(Z))'\right] \\ = \; & E\left[E[m(Z;\beta^0)|Z_1]\frac{\frac{\partial}{\partial\gamma'}P(C\in\lambda|Z_1;\gamma^0)}{P(C\in\lambda)}\frac{\partial\gamma^0}{\partial\theta'}\right]. \end{aligned} \tag{27}$$

Consider the LHS of (27). 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 (27) becomes

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

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 (27) is equal to the RHS. This completes the proof.

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

For all the corollaries below we use a restricted version of Proposition 1 to obtain $\varphi_{\lambda_p}^{\lambda_a}(C,Z)$. In particular the underlying population of the observed data in the proposition is treated as $\mathcal{D}_{\lambda_p\cup\lambda_a}:=\mathcal{D}_{\lambda_p}\cup\mathcal{D}_{\lambda_a}$. We provide some details in the first part of the first proof, but omit them from the rest.

Also, it is important to remember the following relations that are used extensively throughout. Consider any two functions $h(Z)$ and $h(C,Z)$. If the expectations exist then

$$\begin{aligned} E_{\mathcal{D}_{\lambda_p}}[h(Z)|Z_1] &= E_{\mathcal{D}_{\lambda_p\cup\lambda_a}}[h(Z)|Z_1]=E[h(Z)|Z_1] \text{ (by virtue of (2))}, \\ E_{\mathcal{D}_{\lambda_p}}[h(C,Z)] &= E_{\mathcal{D}_{\lambda_p\cup\lambda_a}}[h(C,Z)|C\in\lambda_p]=E[h(C,Z)|C\in\lambda_p], \\ E_{\mathcal{D}_{\lambda_p}}[h(C,Z)|Z_1] &= E_{\mathcal{D}_{\lambda_p\cup\lambda_a}}[h(C,Z)|Z_1,C\in\lambda_p]=E[h(C,Z)|Z_1,C\in\lambda_p]. \end{aligned}$$

The subscripts in the probabilities and the expectations represent the underlying population on which they operate. As in the main text, no subscript is used if the underlying population is $\mathcal{D}$.

Taking $h(C,Z) = I(C=r)$ in the last two relations give $P_{\mathcal{D}_{\lambda_p}}(C=r) = P_{\mathcal{D}_{\lambda_p\cup\lambda_a}}(C=r|C\in\lambda_p) = P_{\mathcal{D}_{\lambda_p\cup\lambda_a}}(C=r)/P_{\mathcal{D}_{\lambda_p\cup\lambda_a}}(C\in\lambda_p) = P(C=r|C\in\lambda_p) = P(C=r)/P(C\in\lambda_p)$ and $P_{\mathcal{D}_{\lambda_p}}(C=r|Z_1) = P_{\mathcal{D}_{\lambda_p\cup\lambda_a}}(C=r|Z_1, C\in\lambda_p) = P_{\mathcal{D}_{\lambda_p\cup\lambda_a}}(C=r|Z_1)/P_{\mathcal{D}_{\lambda_p\cup\lambda_a}}(C\in\lambda_p|Z_1) = P(C=r|Z_1, C\in\lambda_p) = P(C=r|Z_1)/P(C\in\lambda_p|Z_1)$ respectively.

Proof of Corollary 4:

First consider the scenario where $\lambda=\lambda_p=\{1\}$ and $\lambda_a=\{3\}$. Apply Proposition 1 with $\lambda=\{1\}$ but with the distribution of C conditional on $C\in\lambda_p\cup\lambda_a=\{1,3\}$ instead of the unconditional distribution. (Alternatively, one can view this as a two-level missing data model with $Y=(Z_2,Z_3)$ and $X=Z_1$ as in Chen et al. (2008) and apply their Theorem 1 (case 1).) Therefore, we obtain

$$\begin{aligned}\varphi^{\lambda_a=\{3\}}_{\lambda=\{1\}}(C,Z) &= g(C,Z;P_*=P_{\mathcal{D}_{\{1\}\cup\{3\}}}) \text{ where}\\ g(C,Z;P_*) &:= \frac{I(C=1)}{P_*(C=1)}E[m(Z)|Z_1] + \frac{P_*(C=1|Z_1)}{P_*(C=1)}\frac{I(C=3)}{P_*(C=3|Z_1)}(m(Z)-E[m(Z)|Z_1]).\end{aligned}$$

Hence, we obtain $V^{\lambda_a=\{3\}}_{\lambda=\{1\}} := E_{\mathcal{D}_{\{1\}\cup\{3\}}}\left[\varphi^{\lambda_a=\{3\}}_{\lambda=\{1\}}(C,Z)\varphi^{\lambda_a=\{3\}}_{\lambda=\{1\}}(C,Z)'\right]$ as

$$\begin{aligned}& V^{\lambda_a=\{3\}}_{\lambda=\{1\}}\\ =\;& E_{\mathcal{D}_{\{1\}\cup\{3\}}}\left[g(C,Z;P_*=P_{\mathcal{D}_{\{1\}\cup\{3\}}})g(C,Z;P_*=P_{\mathcal{D}_{\{1\}\cup\{3\}}})'\right]\\ =\;& E_{\mathcal{D}_{\{1\}\cup\{3\}}}\left[\frac{I(C=1)}{P^2_{\mathcal{D}_{\{1\}\cup\{3\}}}(C=1)}E[m(Z)|Z_1]E[m(Z)|Z_1]' + \frac{P^2_{\mathcal{D}_{\{1\}\cup\{3\}}}(C=1|Z_1)}{P^2_{\mathcal{D}_{\{1\}\cup\{3\}}}(C=1)}\frac{Var(m(Z)|Z_1)}{P_{\mathcal{D}_{\{1\}\cup\{3\}}}(C=3|Z_1)}\right]\\ =\;& \frac{1}{P_{\mathcal{D}_{\{1\}\cup\{3\}}}(C=1)}E_{\mathcal{D}_{\{1\}\cup\{3\}}}\left[E[m(Z)|Z_1]E[m(Z)|Z_1]' + \frac{P_{\mathcal{D}_{\{1\}\cup\{3\}}}(C=1|Z_1)}{P_{\mathcal{D}_{\{1\}\cup\{3\}}}(C=3|Z_1)}Var(m(Z)|Z_1)\middle| C=1\right]\\ =\;& \frac{1}{P(C=1|C\in\{1,3\})}E\left[E[m(Z)|Z_1]E[m(Z)|Z_1]' + \frac{P(C=1|Z_1)}{P(C=3|Z_1)}Var(m(Z)|Z_1)\middle| C=1\right].\end{aligned}$$

Now consider the scenario where $\lambda=\lambda_p=\{1\}$ and $\lambda_a=\{2,3\}$. Applying Proposition 1 directly for $\lambda=1$ we obtain, by suitably rearranging terms, that

$$\begin{aligned}\varphi^{\lambda_a=\{2,3\}}_{\lambda=\{1\}}(C,Z) &= g(C,Z;P_*=P)\\ &\quad + \frac{P(C=1|Z_1)}{P(C=1)}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)} - \frac{I(C=3)}{P(C=3|Z_1)}\right](E[m(Z)|Z_1,Z_2]-E[m(Z)|Z_1]).\end{aligned}$$

Using variance computations similar to those presented in Appendix B, i.e., in particular noting that the covariance between the two terms is equal to negative of the variance of the last term, we obtain

$$\begin{aligned}& V^{\lambda_a=\{2,3\}}_{\lambda=\{1\}} := E\left[\varphi^{\lambda_a=\{2,3\}}_{\lambda=\{1\}}(C,Z)\varphi^{\lambda_a=\{2,3\}}_{\lambda=\{1\}}(C,Z)'\right]\\ =\;& \frac{1}{P(C=1)}E\left[E[m(Z)|Z_1]E[m(Z)|Z_1]' + \frac{P(C=1|Z_1)}{P(C=3|Z_1)}Var(m(Z)|Z_1)\middle| C=1\right]\\ & -Var\left(\frac{P(C=1|Z_1)}{P(C=1)}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)} - \frac{I(C=3)}{P(C=3|Z_1)}\right](E[m(Z)|Z_1,Z_2]-E[m(Z)|Z_1])\right)\\ =\;& \frac{V^{\lambda_a=\{3\}}_{\lambda=\{1\}}}{P(C\in\{1,3\})} - E\left[\frac{P^2(C=1|Z_1)}{P^2(C=1)}\left[\frac{1}{P(C=3|Z_1)} - \frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right]. \blacksquare\end{aligned}$$

Remark: We use $g(C,Z;P_*)$ in the proofs of all the corollaries, but its definition changes.

Proof of Corollary 5:

First consider the scenario where $\lambda = \lambda_p = \{2\}$ and $\lambda_a = \{3\}$. Following similar steps as in the proof of Proposition 1 we obtain[15]

$$\begin{aligned}\varphi_{\lambda=\{2\}}^{\lambda_a=\{3\}}(C,Z) &= g(C,Z;P_* = P_{\mathcal{D}_{\{2\}\cup\{3\}}}) \text{ where} \\ g(C,Z;P_*) &:= \frac{I(C=2)}{P_*(C=2)}E[m(Z)|Z_1,Z_2] + \frac{P_*(C=2|Z_1)}{P_*(C=2)}\frac{I(C=3)}{P_*(C=3|Z_1)}(m(Z) - E[m(Z)|Z_1,Z_2]).\end{aligned}$$

Hence, we obtain $V_{\lambda=\{2\}}^{\lambda_a=\{3\}} := E_{\mathcal{D}_{\{2\}\cup\{3\}}}\left[\varphi_{\lambda=\{2\}}^{\lambda_a=\{3\}}(C,Z)\varphi_{\lambda=\{2\}}^{\lambda_a=\{3\}}(C,Z)'\right]$ as

$$\begin{aligned}& V_{\lambda=\{2\}}^{\lambda_a=\{3\}} \\ =\ & E_{\mathcal{D}_{\{2\}\cup\{3\}}}\left[g(C,Z;P_* = P_{\mathcal{D}_{\{2\}\cup\{3\}}})g(C,Z;P_* = P_{\mathcal{D}_{\{2\}\cup\{3\}}})'\right] \\ =\ & \frac{1}{P(C=2|C\in\{2,3\})}E\left[E[m(Z)|Z_1,Z_2]E[m(Z)|Z_1,Z_2]' + \frac{P(C=2|Z_1)}{P(C=3|Z_1)}Var(m(Z)|Z_1,Z_2)\middle| C=2\right].\end{aligned}$$

Now consider the scenario where $\lambda = \lambda_p = \{2\}$ and $\lambda_a = \{1,3\}$. Apply Proposition 1 directly as before and after suitably rearranging terms we obtain $\varphi_{\lambda=\{2\}}^{\lambda_a=\{1,3\}}(C,Z)$ and its variance as

$$\begin{aligned}\varphi_{\lambda=\{2\}}^{\lambda_a=\{1,3\}}(C,Z) &= g(C,Z;P_* = P) \\ &\quad + \frac{P(C=2|Z_1)}{P(C=2)}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)} - \frac{I(C=2)}{P(C=2|Z_1)}\right](E[m(Z)|Z_1,Z_2] - E[m(Z)|Z_1]), \\ \Rightarrow V_{\lambda=\{1\}}^{\lambda_a=\{2,3\}} &= \frac{V_{\lambda=\{2\}}^{\lambda_a=\{3\}}}{P(C\in\{2,3\})} \\ &\quad - E\left[\frac{P^2(C=2|Z_1)}{P^2(C=2)}\left[\frac{1}{P(C=2|Z_1)} - \frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right]. \ \blacksquare\end{aligned}$$

Proof of Corollary 6:

First consider the scenario where $\lambda = \lambda_p = \{3\}$ and $\lambda_a = \{\phi\}$. Following similar steps we obtain

$$\varphi_{\lambda=\{3\}}^{\lambda_a=\{\phi\}}(C,Z) = \frac{I(C=3)}{P_{\mathcal{D}_{\{3\}}}(C=3)}m(Z) = m(Z)$$

keeping in mind that $I(C=3) = P_{\mathcal{D}_{\{3\}}}(C=3) = 1$ identically under this scenario. Hence $V_{\lambda=\{3\}}^{\lambda_a=\{\phi\}} = E_{\mathcal{D}_{\{3\}}}[m(Z)m(Z)'] = E[m(Z)m(Z)'|C=3]$.

Now consider the scenario where $\lambda = \lambda_p = \{3\}$ and $\lambda_a = \{1\}$. Applying Proposition 1 we obtain

$$\begin{aligned}\varphi_{\lambda=\{3\}}^{\lambda_a=\{1\}}(C,Z) &= \frac{I(C=3)E[m(Z)|Z_1]}{P_{\mathcal{D}_{\{3\}\cup\{1\}}}(C=3)} + \frac{P_{\mathcal{D}_{\{3\}\cup\{1\}}}(C=3|Z_1)}{P_{\mathcal{D}_{\{3\}\cup\{1\}}}(C=3)}\frac{I(C=3)(m(Z) - E[m(Z)|Z_1])}{P_{\mathcal{D}_{\{3\}\cup\mathcal{D}_{\{1\}}}}(C=3|Z_1)} \\ &= \frac{I(C=3)}{P_{\mathcal{D}_{\{3\}\cup\{1\}}}(C=3)}m(Z).\end{aligned}$$

[15]The two terms below respectively matches with $I(C=2)s(Z_1,Z_2|C=2)$ and $I(C=3)s(Z_2,Z_3|Z_1)$ in the score function of the parametric submodel (using Chen et al. (2008)'s Theorem 1 (case 1)-factorization of likelihood for convenience). This is verified by noting that the terms excluding the indicators are respectively $L_0^2(F(Z_2,Z_1|C=2)$ (applying (1) and (2)) and $L_0^2(F(Z_2,Z_3|Z_1)$. Similar steps are used for cases where the results do not follow by a direct application of Proposition 1.

Hence, we obtain $V^{\lambda_a=\{1\}}_{\lambda=\{3\}} := E_{\mathcal{D}_{\{3\}\cup\{1\}}}\left[\varphi^{\lambda_a=\{1\}}_{\lambda=\{3\}}(C,Z)\varphi^{\lambda_a=\{1\}}_{\lambda=\{3\}}(C,Z)'\right]$ as

$$\begin{aligned} V^{\lambda_a=\{1\}}_{\lambda=\{3\}} &= E_{\mathcal{D}_{\{3\}\cup\{1\}}}\left[\frac{I(C=3)}{P^2_{\mathcal{D}_{\{3\}\cup\{1\}}}(C=3)}m(Z)m(Z)'\right] \\ &= E_{\mathcal{D}_{\{3\}\cup\{1\}}}\left[\frac{m(Z)m(Z)'}{P_{\mathcal{D}_{\{3\}\cup\{1\}}}(C=3)}|C=3\right] \\ &= \frac{E[m(Z)m(Z)'|C=3]}{P_{\mathcal{D}_{\{3\}\cup\{1\}}}(C=3)} = \frac{P(C\in\{1,3\})}{P(C=3)}V^{\lambda_a=\{\phi\}}_{\lambda=\{3\}}. \end{aligned}$$

Exactly same steps give the result for the scenario where $\lambda=\lambda_p=\{3\}$ and $\lambda_a=\{2\}$.

Finally consider the scenario where $\lambda=\lambda_p=\{3\}$ and $\lambda_a=\{1,2\}$. Applying Proposition 1 and after suitably rearranging terms we obtain $\varphi^{\lambda_a=\{1,2\}}_{\lambda=\{3\}}(C,Z)$ and its variance as

$$\begin{aligned} \varphi^{\lambda_a=\{1,2\}}_{\lambda=\{3\}}(C,Z) &= \frac{I(C=3)}{P(C=3)}m(Z) \\ &\quad +\frac{P(C=3|Z_1)}{P(C=3)}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)}-\frac{I(C=3)}{P(C=3|Z_1)}\right](E[m(Z)|Z_1,Z_2]-E[m(Z)|Z_1]), \\ \Rightarrow V^{\lambda_a=\{1,2\}}_{\lambda=\{3\}} &= \frac{V^{\lambda_a=\{\phi\}}_{\lambda=\{3\}}}{P(C=3)} \\ &\quad -E\left[\frac{P^2(C=3|Z_1)}{P^2(C=3)}\left[\frac{1}{P(C=3|Z_1)}-\frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right]. \blacksquare \end{aligned}$$

Proof of Corollary 7:

Consider the scenario where $\lambda=\lambda_p=\{1,3\}$ and $\lambda_a=\{\phi\}$. Applying Proposition 1 we obtain

$$\begin{aligned} \varphi^{\lambda_a=\{\phi\}}_{\lambda=\{1,3\}}(C,Z) &= g(C,Z;P_*=P_{\mathcal{D}_{\{1\}\cup\{3\}}}) \text{ where} \\ g(C,Z;P_*) &:= \frac{I(C\in\{1,3\})}{P_*(C\in\{1,3\})}E[m(Z)|Z_1]+\frac{P_*(C\in\{1,3\}|Z_1)}{P_*(C=\{1,3\})}\frac{I(C=3)}{P_*(C=3|Z_1)}(m(Z)-E[m(Z)|Z_1]). \end{aligned}$$

Note that $P_{\mathcal{D}_{\{1\}\cup\{3\}}}(C\in\{1,3\})=P_{\mathcal{D}_{\{1\}\cup\{3\}}}(C\in\{1,3\}|Z_1)=1$ identically, and so is $I(C\in\{1,3\})$. Hence, we obtain $V^{\lambda_a=\{\phi\}}_{\lambda=\{1,3\}} := E_{\mathcal{D}_{\{1\}\cup\{3\}}}\left[\varphi^{\lambda_a=\{\phi\}}_{\lambda=\{1,3\}}(C,Z)\varphi^{\lambda_a=\{\phi\}}_{\lambda=\{1,3\}}(C,Z)'\right]$ as

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

Now consider the scenario where $\lambda=\lambda_p=\{1,3\}$ and $\lambda_a=\{2\}$. Applying Proposition 1 and after suitably rearranging terms we obtain $\varphi^{\lambda_a=\{2\}}_{\lambda=\{1,3\}}(C,Z)$ and its variance as

$$\begin{aligned} \varphi^{\lambda_a=\{2\}}_{\lambda=\{1,3\}}(C,Z) &= g(C,Z;P_*=P) \\ &\quad +\frac{P(C\in\{1,3\}|Z_1)}{P(C\in\{1,3\})}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)}-\frac{I(C=3)}{P(C=3|Z_1)}\right](E[m(Z)|Z_1,Z_2]-E[m(Z)|Z_1]), \\ \Rightarrow V^{\lambda_a=\{2\}}_{\lambda=\{1,3\}} &= \frac{V^{\lambda_a=\{\phi\}}_{\lambda=\{1,3\}}}{P(C\in\{1,3\})} \\ &\quad -E\left[\frac{P^2(C\in\{1,3\}|Z_1)}{P^2(C\in\{1,3\})}\left[\frac{1}{P(C=3|Z_1)}-\frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right]. \blacksquare \end{aligned}$$

Proof of Corollary 8:

First consider the scenario where $\lambda = \lambda_p = \{2,3\}$ and $\lambda_a = \{\phi\}$. Following similar steps as in the proof of Proposition 1 we obtain

$$\varphi^{\lambda_a=\{\phi\}}_{\lambda=\{2,3\}}(C,Z) = g(C,Z;P_* = P_{\mathcal{D}_{\{2\}\cup\{3\}}}) \text{ where}$$
$$g(C,Z;P_*) := \frac{I(C\in\{2,3\})}{P_*(C\in\{1,3\})}E[m(Z)|Z_1,Z_2] + \frac{P_*(C\in\{2,3\}|Z_1)}{P_*(C=\{2,3\})}\frac{I(C=3)}{P_*(C=3|Z_1)}(m(Z)-E[m(Z)|Z_1,Z_2]).$$

Note that $P_{\mathcal{D}_{\{2\}\cup\{3\}}}(C\in\{2,3\}) = P_{\mathcal{D}_{\{2\}\cup\{3\}}}(C\in\{2,3\}|Z_1) = 1$ identically, and so is $I(C\in\{2,3\})$. Hence, we obtain $V^{\lambda_a=\{\phi\}}_{\lambda=\{2,3\}} := E_{\mathcal{D}_{\{2\}\cup\{3\}}}\left[\varphi^{\lambda_a=\{\phi\}}_{\lambda=\{2,3\}}(C,Z)\varphi^{\lambda_a=\{\phi\}}_{\lambda=\{2,3\}}(C,Z)'\right]$ as

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

Now consider the scenario where $\lambda = \lambda_p = \{2,3\}$ and $\lambda_a = \{1\}$. Applying Proposition 1 and after suitably rearranging terms we obtain $\varphi^{\lambda_a=\{1\}}_{\lambda=\{2,3\}}(C,Z)$ and its variance as

$$\begin{aligned}
\varphi^{\lambda_a=\{1\}}_{\lambda=\{2,3\}}(C,Z) &= g(C,Z;P_*=P) \\
&\quad + \frac{P(C\in\{2,3\}|Z_1)}{P(C\in\{2,3\})}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)} - \frac{I(C\in\{2,3\})}{P(C\in\{2,3\}|Z_1)}\right](E[m(Z)|Z_1,Z_2] - E[m(Z)|Z_1]) \\
&= g(C,Z;P_*=P) \\
&\quad + \frac{P(C\in\{2,3\}|Z_1)}{P(C\in\{2,3\})}\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)} - \frac{I(C\geq 2)}{P(C\geq 2|Z_1)}\right](E[m(Z)|Z_1,Z_2] - E[m(Z)|Z_1]) \\
&= g(C,Z;P_*=P), \\
\Rightarrow V^{\lambda_a=\{1\}}_{\lambda=\{2,3\}} &= \frac{V^{\lambda_a=\{\phi\}}_{\lambda=\{2,3\}}}{P(C\in\{2,3\})}. \ \blacksquare
\end{aligned}$$

Proof of Corollary 9:

The proof here needs a little more care than in Corollary 6 when applying Proposition 1 because the target population is $\mathcal{D}$ and hence may be a superset of $\mathcal{D}_{\lambda_p\cup\lambda_a}$. This means repeated use of relationships such as:

$$\begin{aligned}
E[m(Z;\beta)] &= E_{\mathcal{D}_{\lambda_p\cup\lambda_a}}\left[\frac{P(C\in\lambda_p\cup\lambda_a)}{P(C\in\lambda_p\cup\lambda_a|Z_1)}\frac{I(C=3)}{P_{\mathcal{D}_{\lambda_p\cup\lambda_a}}(C=3|Z_1)}m(Z;\beta)\right] \\
&= E_{\mathcal{D}_{\lambda_p\cup\lambda_a}}\left[\frac{I(C=3)}{P(C=3|Z_1)}m(Z;\beta)\right]P(C\in\lambda_p\cup\lambda_a).
\end{aligned}$$

First consider the scenario where $\lambda_p = \{3\}$ and $\lambda_a = \{1\}$. Using the above relationship to verify pathwise differentiability in Proposition 1 we obtain

$$\begin{aligned}
\varphi^{\lambda_a=\{1\}}_{\lambda=\mathbb{C}}(C,Z) &= \frac{P(C\in\{1,3\})}{P(C\in\{1,3\}|Z_1)}\left[\frac{I(C=3)\,(m(Z)-E[m(Z)|Z_1])}{P_{\mathcal{D}_{\{3\}\cup\{1\}}}(C=3|Z_1)} + I(C\in\{1,3\})E[m(Z)|Z_1]\right], \\
\Rightarrow V^{\lambda_a=\{1\}}_{\lambda=\mathbb{C}} &= E_{\mathcal{D}_{\{3\}\cup\{1\}}}\left[\varphi^{\lambda_a=\{1\}}_{\lambda=\mathbb{C}}(C,Z)\varphi^{\lambda_a=\{1\}}_{\lambda=\mathbb{C}}(C,Z)'\right] \\
&= E\left[\frac{Var(m(Z)|Z_1)}{P(C=3|Z_1)} + \frac{E[m(Z)|Z_1]E[m(Z)'|Z_1]}{P(C\in\{1,3\}|Z_1)}\right]P(C\in\{1,3\}).
\end{aligned}$$

The denominator of the second term in the last line is the penalty in terms of inflated variance that occurs because the original moment restriction is defined in the population $\mathcal{D}$ and not in $\mathcal{D}_{\{3\}\cup\{1\}}$. So the intuition from the previous corollaries do not hold here.

Exactly similarly, in the scenario where $\lambda_p = \{3\}$ and $\lambda_a = \{2\}$ we obtain

$$\begin{aligned}
\varphi_{\lambda=\mathbb{C}}^{\lambda_a=\{2\}}(C,Z) &= \frac{P(C\in\{2,3\})}{P(C\in\{2,3\}|Z_1)}\left[\frac{I(C=3)\left(m(Z)-E[m(Z)|Z_1,Z_2]\right)}{P_{\mathcal{D}_{\{3\}\cup\{2\}}}(C=3|Z_1)}+I(C\in\{2,3\})E[m(Z)|Z_1,Z_2]\right], \\
\Rightarrow V_{\lambda=\mathbb{C}}^{\lambda_a=\{2\}} &= E\left[\frac{Var(m(Z)|Z_1,Z_2)}{P(C=3|Z_1)}+\frac{E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2]}{P(C\in\{2,3\}|Z_1)}\right]P(C\in\{2,3\}).
\end{aligned}$$

Finally in the scenario where $\lambda_p = \{3\}$ and $\lambda_a = \{1,2\}$ we can directly apply Proposition 1 and suitably rearrange the terms to obtain

$$\begin{aligned}
\varphi_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}}(C,Z) &= \left[\frac{I(C=3)}{P(C=3|Z_1)}\left(m(Z)-E[m(Z)|Z_1]\right)+E[m(Z)|Z_1]\right] \\
&\quad+\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)}-\frac{I(C=3)}{P(C=3|Z_1)}\right]\left(E[m(Z)|Z_1,Z_2]-E[m(Z)|Z_1]\right), \\
\Rightarrow V_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}} &= \frac{V_{\lambda=\mathbb{C}}^{\lambda_a=\{1\}}}{P(C\in\{1,3\})}-E\left[\left[\frac{1}{P(C\in\{1,3\}|Z_1)}-1\right]E[m(Z)|Z_1]E[m(Z)'|Z_1]\right] \\
&\quad-E\left[\left[\frac{1}{P(C=3|Z_1)}-\frac{1}{P(C\geq 2|Z_1)}\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right].
\end{aligned}$$

Rearranging the terms of $\varphi_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}}(C,Z)$ in an alternate way we obtain[16]

$$\begin{aligned}
\varphi_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}}(C,Z) &= \left[\frac{I(C=3)}{P(C=3|Z_1)}\left(m(Z)-E[m(Z)|Z_1,Z_2]\right)+E[m(Z)|Z_1,Z_2]\right] \\
&\quad+\left[\frac{I(C\geq 2)}{P(C\geq 2|Z_1)}-1\right]\left(E[m(Z)|Z_1,Z_2]-E[m(Z)|Z_1]\right), \\
\Rightarrow V_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}} &= E\left[\frac{Var(m(Z)|Z_1,Z_2)}{P(C=3|Z_1)}+E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2]\right] \\
&\quad+E\left[\left[\frac{1}{P(C\geq 2|Z_1)}-1\right]Var(E[m(Z)|Z_1,Z_2]|Z_1)\right] \\
&= \frac{V_{\lambda=\mathbb{C}}^{\lambda_a=\{2\}}}{P(C\in\{2,3\})}-E\left[\left[\frac{1}{P(C\geq 2|Z_1)}-1\right]E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2]\right] \\
&\quad+E\left[\left[\frac{1}{P(C\geq 2|Z_1)}-1\right]E\left[E[m(Z)|Z_1,Z_2]E[m(Z)'|Z_1,Z_2]|Z_1\right]\right] \\
&\quad-E\left[\left[\frac{1}{P(C\geq 2|Z_1)}-1\right]E[m(Z)|Z_1]E[m(Z)'|Z_1]\right] \\
&= \frac{V_{\lambda=\mathbb{C}}^{\lambda_a=\{2\}}}{P(C\in\{2,3\})}-E\left[\left[\frac{1}{P(C\geq 2|Z_1)}-1\right]E[m(Z)|Z_1]E[m(Z)'|Z_1]\right]. \blacksquare
\end{aligned}$$

Proof of Proposition 10:

STEP - 1: Consider a regular parametric sub-model indexed by a finite-dimensional parameter θ for the joint distribution of the observed data $(C, G_C(Z))$. So the joint density of the observed

[16]Here the covariance between the terms in the first and second lines in the expression for $\varphi_{\lambda=\mathbb{C}}^{\lambda_a=\{1,2\}}(C,Z)$ is zero. In the other corollaries it was the negative of the variance of the corresponding second (i.e., the adjustment) term.

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

$$\begin{aligned} f_\theta(C, G_C(Z)) &= \prod_{r=1,2,3,\infty} (P_\theta(C=r|G_r(Z)) f_\theta(G_r(Z)))^{I(C=r)} \\ &= (P_\theta(C=1|Z_1) f_\theta(Z_1))^{I(C=1)} (P_\theta(C=2|Z_1) f_\theta(Z_1) f_\theta(Z_2|Z_1))^{I(C=2)} \\ &\quad \times (P_\theta(C=3|Z_1) f_\theta(Z_1) f_\theta(Z_3|Z_1))^{I(C=3)} (P_\theta(C=\infty|Z_1) f_\theta(Z_1) f_\theta(Z_3, Z_2|Z_1))^{I(C=\infty)} \end{aligned}$$

where (2) is used in the last two lines. Note, however, that (13) has not been used and hence $f_\theta(Z_3, Z_2|Z_1)$ has *not been factorized* as $f_\theta(Z_2|Z_1) f_\theta(Z_3|Z_1)$ in the last term of the last line.

This gives the log of the joint density as

$$\begin{aligned} \log f_\theta(C, G_C(Z)) &= \log f_\theta(Z_1) + \sum_{r=1,2,3,\infty} I(C=r) \log P_\theta(C=r|Z_1) \\ &\quad + \sum_{r=2,3} I(C=r) \log f_\theta(Z_r|Z_1) + I(C=\infty) \log f_\theta(Z_3, Z_2|Z_1). \end{aligned}$$

The score function with respect to θ is

$$\begin{aligned} S_\theta(C, G_C(Z)) &= s_\theta(Z_1) + \sum_{r=1,2,3,\infty} I(C=r) \frac{\dot{P}_\theta(C=r|Z_1)}{P_\theta(C=r|Z_1)} \\ &\quad + \sum_{r=2,3} I(C=r) s_\theta(Z_r|Z_1) + I(C=\infty) s_\theta(Z_3, Z_2|Z_1) \end{aligned}$$

where $\dot{P}_\theta(C=r|Z_1) := \frac{\partial}{\partial\theta} P_\theta(C=r|Z_1)$; $s_\theta(Z_1) := \frac{\partial}{\partial\theta} \log f_\theta(Z_1)$; $s_\theta(Z_r|Z_1) := \frac{\partial}{\partial\theta} \log f_\theta(Z_r|Z_1)$ for $r = 2, 3$; and $s_\theta(Z_3, Z_2|Z_1) := \frac{\partial}{\partial\theta} \log f_\theta(Z_3, Z_2|Z_1)$. Henceforth, we omit the subscript θ from the quantities evaluated at $\theta = \theta_0$.

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

$$\mathcal{T} := a(Z_1) + \sum_{r=1,2,3,\infty} I(C=r) \frac{b_r(Z_1)}{a_r(Z_1)} + \sum_{r=2,3} I(C=r) a(Z_1, Z_r) + I(C=\infty) a(Z_1, Z_2, Z_3), \quad (28)$$

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 $\sum_{r=1,2,3,\infty} I(C=r) \frac{b_r(Z_1)}{a_r(Z_1)} \in L_0^2(F(C|Z_1))$; $a(Z_1, Z_r) \in L_0^2(F(Z_r|Z_1))$ for $r = 2, 3$; and $a(Z_1, Z_2, Z_3) \in L_0^2(F(Z_3, Z_2|Z_1))$. Once again note that we have not imposed (13) which would additionally restrict $\mathcal{T}$ by requiring that $a(Z_1, Z_2, Z_3) = a(Z_1, Z_2) + a(Z_1, Z_3)$.

STEP - 2: $(d_m = d_\beta)$ Differentiating $E[m(Z; \beta^0)] = 0$ under the integral we obtain

$$\frac{\partial \beta^0(\theta_0)}{\partial \theta'} = -M^{-1} E\left[m(Z; \beta^0) \left\{s(Z_1)' + s(Z_3, Z_2|Z_1)'\right\}\right].$$

Now we impose the restriction (13) and show that $\psi(C, G_C(Z))$ satisfies $E[\psi(C, G_C(Z)) S(C, G_C(Z))'] = \frac{\partial \beta^0(\theta_0)}{\partial \theta'}$, i.e., the path-derivative condition by verifying:

$$E[\varphi(C, G_C(Z)) S(C, G_C(Z))'] = E\left[m(Z; \beta^0) \left\{s(Z_1)' + s(Z_3, Z_2|Z_1)'\right\}\right].$$

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

Consider the first term $\varphi_{[1,\infty]}(C, G_C(Z))$ of $\varphi(C, G_C(Z))$. Taking expectation conditional on Z_1 and using (2) we obtain

$$\begin{aligned} E\left[\varphi_{[1,\infty]}(C, G_C(Z)) S(C, G_C(Z))'\right] &= E\left[(m(Z; \beta^0) - E[m(Z; \beta^0)|Z_1]) s(Z_3, Z_2|Z_1)'\right] \\ &\quad + E\left[E[m(Z; \beta^0)|Z_1] S(C, G_C(Z))'\right]. \end{aligned}$$

Using $s(Z_3, Z_2|Z_1) \in L_0^2(F(Z_3, Z_2|Z_1))$, the first term on the RHS becomes $E\left[m(Z;\beta^0)s(Z_3, Z_2|Z_1)'\right]$. Recalling that $\sum_{r=1,2,3,\infty} \dot{P}(C = r|Z_1) = 0$, $s(Z_r|Z_1) \in L_0^2(F(Z_r|Z_1))$ for $r = 2, 3$; and $s(Z_3, Z_2|Z_1) \in L_0^2(F(Z_3, Z_2|Z_1))$, the second term on the RHS becomes $E\left[m(Z;\beta^0)s(Z_1)'\right]$ using (2). Therefore,

$$E\left[\varphi_{[1,\infty]}(C, G_C(Z))S(C, G_C(Z))'\right] = E\left[m(Z;\beta^0)\left\{s(Z_1)' + s(Z_3, Z_2|Z_1)'\right\}\right]. \tag{29}$$

Now consider the second term of $\varphi(C, G_C(Z))$, i.e., $\varphi_{[2]}(C, G_C(Z))$. Again, taking expectation conditional on Z_1 and using (2) eliminate the terms corresponding to $s(Z_1)$, $\sum_{r=1,2,3,\infty} I(C = r)\frac{\dot{P}(C=r|Z_1)}{P(C=r|Z_1)}$ and $I(C = 3)s(Z_3|Z_1)$ in $E\left[\varphi_{[2]}(C, G_C(Z))S(C, G_C(Z))'\right]$. Then, by noting that $s(Z_2|Z_1) \in L_0^2(F(Z_2|Z_1))$ and $s(Z_3, Z_2|Z_1) \in L_0^2(F(Z_3, Z_2|Z_1))$, to eliminate terms with $E[m(Z;\beta^0)|Z_1]$ we obtain

$$\begin{aligned}
& E\left[\varphi_{[2]}(C, G_C(Z))S(C, G_C(Z))'\right] \\
= \; & E\left[\frac{P(C = 2|Z_1)}{P(C \in \{2,\infty\}|Z_1)}\left[\frac{I(C = 2)}{P(C = 2|Z_1)} - \frac{I(C = \infty)}{P(C = \infty|Z_1)}\right]\right. \\
& \left.\times\, E[m(Z;\beta^0)|Z_2, Z_1]\left\{I(C = 2)s(Z_2|Z_1) + I(C = \infty)s(Z_3, Z_2|Z_1)\right\}'\right] \\
= \; & E\left[\frac{P(C = 2|Z_1)}{P(C \in \{2,\infty\}|Z_1)}E[m(Z;\beta^0)|Z_2, Z_1]\left\{s(Z_2|Z_1)' - s(Z_3, Z_2|Z_1)'\right\}\right] \\
= \; & E\left[\frac{P(C = 2|Z_1)}{P(C \in \{2,\infty\}|Z_1)}E[m(Z;\beta^0)|Z_2, Z_1]s(Z_3|Z_1)'\right] = 0
\end{aligned}$$

where the 0 follows by noting $s(Z_3|Z_1) \in L_0^2(F(Z_3|Z_1))$ and then using (13). The equality before that follows because $s(Z_3, Z_2|Z_1) = s(Z_3|Z_2, Z_1) + s(Z_2|Z_1) = s(Z_3|Z_1) + s(Z_2|Z_1)$ under (13).

Exactly the same steps give zero for the third term, $\varphi_{[3]}(C, G_C(Z))$, of $\varphi(C, G_C(Z))$. Therefore, taken together with (29), these verify that $\psi(C, G_C(Z))$ satisfies the path-derivative condition.

To show that $\varphi(C, G_C(Z))$ belongs to the tangent set $\mathcal{T}$ in (28), rearrange its terms in a fashion similar to Section 2, and for brevity (in space) express in terms of (C, Z) as follows:

$$\begin{aligned}
& \varphi(C, G_C(Z)) \\
= \; & E[m(Z;\beta)|Z_1] + \sum_{r=2}^{3}\frac{I(C = r)}{P(C \in \{r,\infty\}|Z_1)}\left(E[m(Z;\beta)|Z_r, Z_1] - E[m(Z;\beta)|Z_1]\right) \\
& + \frac{I(C = \infty)}{P(C = \infty|Z_1)}\left\{(m(Z) - E[m(Z)|Z_1]) - \sum_{r=2}^{3}\frac{P(C = r|Z_1)\left(E[m(Z;\beta)|Z_r, Z_1] - E[m(Z;\beta)|Z_1]\right)}{P(C \in \{r,\infty\}|Z_1)}\right\}.
\end{aligned}$$

The first term in the second line is a function of only Z_1 and has expectation zero. So it corresponds to the term $a(Z_1)$ in $\mathcal{T}$. The r-th term (without the multiplier $I(C = r)$) in the second line is a function of Z_r and Z_1. Also, conditional on Z_1 it has expectation zero. So it corresponds to the term $a(Z_r, Z_1)$ term in $\mathcal{T}$ for $r = 2, 3$. The term on the last line (without the multiplier $I(C = \infty)$) is a function of Z_3, Z_2 and Z_1. Also, conditional on Z_1 it has expectation 0. So it corresponds to the term $a(Z_3, Z_2, Z_1)$ in $\mathcal{T}$. All other terms in $\mathcal{T}$ are corresponded identically by zeros. This shows that $\varphi(C, G_C(Z))$ is in $\mathcal{T}$ and hence $\Psi(C, G_C(Z))$ is the efficient influence function. ■

Remark: (29) implies that $\psi_{[1,\infty]}(C, G_C(Z))$ also satisfies the path-derivative condition, even when (13) does not hold. So it is an influence function. However, $\varphi_{[1,\infty]}(C, G_C(Z))$ is typically not in $\mathcal{T}$ because that would require $s(Z_2|Z_1) = s(Z_3|Z_1) = 0$ but $s(Z_3, Z_2|Z_1) \neq 0$ [or equivalently, for our purpose, $s(Z_3|Z_2, Z_1) = s(Z_2|Z_3, Z_1) \neq 0$] that cannot be true in general. Hence $\psi_{[1,\infty]}(C, G_C(Z))$ is not the efficient influence function.

B Appendix: Demonstration of auxiliary results

B.1 Expression for V_λ in Proposition 1:

Under (2) and assumption A(2) we know that $E\left[\varphi_\lambda(C, G_C(Z))\right] = E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} m(Z)\right] = 0$ by (1). Therefore, by taking expectation conditional on $G_{r-1}(Z)$ to eliminate the expectation of the cross product of the first term with the r-th $(r = 2, \ldots, R)$ remaining term, we obtain V_λ is

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

Now taking expectation conditional on $G_{\min(r,s)-1}(Z)$, each term in the last line is 0. Similarly, taking expectation conditional on $G_{r-1}(Z)$, each term on the second line becomes

$$
E\left[\frac{P^2(C\in\lambda|G_1(Z))}{P^2(C\in\lambda)}\sum_{r=2}^{R}\frac{1}{P(C\geq r|G_1(Z))} Var(E[m(Z)|G_r(Z)]|G_{r-1}(Z))\right].
$$

Therefore, it follows directly that
$V_\lambda = E\left[\frac{P(C\in\lambda|G_1(Z))}{P^2(C\in\lambda)} E[m(Z)|G_1(Z)]E[m(Z)|G_1(Z)]'\right] + E\left[\frac{P^2(C\in\lambda|G_1(Z))}{P^2(C\in\lambda)}\sum_{r=2}^{R}\frac{Var(E[m(Z)|G_r(Z)]|G_{r-1}(Z))}{P(C\geq r|G_1(Z))}\right]$. ■

B.2 Expression for $V_{\lambda[k]}$ in Proposition 2:

Note that $\varphi_{\lambda[k]}(C, G_C(Z)) = \varphi_\lambda(C, G_C(Z)) - \frac{I(C\in\lambda) - P(C\in\lambda|G_1(Z))}{P(C\in\lambda)} E[m(Z)|G_1(Z)]$. Also, as before, we can obtain

$$
\begin{aligned}
& E\left[\varphi_\lambda(C, G_C(Z))\frac{I(C\in\lambda) - P(C\in\lambda|G_1(Z))}{P(C\in\lambda)} E[m(Z)|G_1(Z)]'\right] \\
= \;& E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}\frac{I(C\in\lambda) - P(C\in\lambda|G_1(Z))}{P(C\in\lambda)} E[m(Z)|G_1(Z)]E[m(Z)|G_1(Z)]'\right] \\
= \;& E\left[\frac{P(C\in\lambda|G_1(Z))(1 - P(C\in\lambda|G_1(Z)))}{P^2(C\in\lambda)} E[m(Z)|G_1(Z)]E[m(Z)|G_1(Z)]'\right] \\
= \;& E\left[\left\{\frac{I(C\in\lambda) - P(C\in\lambda|G_1(Z))}{P(C\in\lambda)} E[m(Z)|G_1(Z)]\right\}\{.\}'\right]. \qquad (30)
\end{aligned}
$$

Therefore, $V_{\lambda[k]} = V_\lambda - \frac{P(C\in\lambda|G_1(Z))(1-P(C\in\lambda|G_1(Z)))}{P^2(C\in\lambda)} E[m(Z)|G_1(Z)]E[m(Z)|G_1(Z)]'$. ■

B.3 Expression for $V_{\lambda[pk]}$ in Proposition 3:

For the first expression of $V_{\lambda[pk]}$ note that

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

By using the fact that $E[S_\gamma(C|G_1(Z))|G_1(Z)] = 0$, we know that

$$E\left[\frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} E[m(Z)|G_1(Z)] S_\gamma(C|G_1(Z))'\right] = 0. \tag{31}$$

Therefore, in $E[\varphi_{\lambda[pk]}(C, G_C(Z))\varphi_{\lambda[pk]}(C, G_C(Z))']$, taking expectations conditional on $G_{r-1}(Z)$ for the cross product corresponding to the r-term in the second set of terms of $\varphi_{\lambda[k]}(C, G_C(Z))$, gives

$$E\left[\varphi_{\lambda[k]}(C, G_C(Z))\Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)].\middle| S_\gamma(C|G_1(Z))\right)'\right] = 0$$

The first expression of $V_{\lambda[pk]}$ follows since

$$Var\left(\Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right)\right) = B\left(E\left[S_\gamma(C|Z_1)S_\gamma(C|Z_1)'\right]\right)^{-1} B'.$$

Now consider the second expression of $V_{\lambda[pk]}$. Note that $\varphi_{\lambda[pk]}(C, G_C(Z))$ can also be written as

$$\begin{aligned}\varphi_{\lambda[pk]}(C, G_C(Z)) &= \varphi_\lambda(C, G_C(Z)) - \frac{I(C \in \lambda) - P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} E[m(Z)|G_1(Z)] \\ &\quad + \Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right)\end{aligned}$$

Using (30) and, then, (1) we obtain

$$E\left[\varphi_\lambda(C, G_C(Z))\frac{I(C \in \lambda) - P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} E[m(Z)|G_1(Z)]'\right] = Var\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|G_1(Z)]\right).$$

From (31) and the lines below it we know that

$$\begin{aligned}&E\left[\left(\varphi_\lambda(C, G_C(Z)) - \frac{I(C \in \lambda) - P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} E[m(Z)|G_1(Z)]\right)\right. \\ &\quad \left.\Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right)'\right] \\ = \; &E\left[\frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} E[m(Z)|G_1(Z)]\Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right)'\right] = 0.\end{aligned}$$

Finally noting that $B = E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m(Z;\beta)|G_1(Z)] S_\gamma(C|G_1(Z))'\right]$ in the expression for $Var\left(\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m(Z;\beta)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right)\right)$, it follows that

$$\begin{aligned}&Var\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|G_1(Z)]\right) - Var\left(\Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right)\right) \\ = \; &Var\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z)|G_1(Z)] - \Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)]\middle| S_\gamma(C|G_1(Z))\right)\right).\end{aligned}$$

Hence we get the second expression for $V_{\lambda[pk]}$. ■

B.4 Expression for $\varphi(C, G_C(Z))$ in footnote 4 of remark (iii) following Proposition 1:

The proof is same as the rest showing the efficiency bounds. We briefly show the first and the second step with a just-identified model.

STEP - 1: In the regular parametric sub-model indexed by θ, the score function of the joint density of the observed data $(C, G_C(Z))$ in terms of the full data (C, Z) is

$$\begin{aligned} S_\theta(C, G_C(Z)) &= \sum_{r=1}^{3} I(C=r)\frac{\dot{P}_\theta(C=r|G_r(Z))}{P_\theta(C=r|G_r(Z))} + s_\theta(Z_1) \\ &\quad + I(C \in \{2,3\}) s_\theta(Z_2|Z_1) + I(C=3) s_\theta(Z_3|Z_2, Z_1) \end{aligned}$$

where the notations carry the same meaning as before. The tangent set for the model can be characterized by functions of the form:

$$\mathcal{T} := \sum_{r=1}^{3} I(C=r)\frac{b_r(G_r(Z))}{a_r(G_r(Z))} + a(Z_1) + I(C \in \{2,3\}) a(Z_1, Z_2) + I(C=3) a(Z_1, Z_2, Z_3), \tag{32}$$

where $\sum_{r=1}^{r}(a_r(G_r(Z)), b_r(G_r(Z))) = (1,0)$; $a(Z_1) \in L_0^2(F(Z_1))$; $a(Z_1, Z_2) \in L_0^2(F(Z_2|Z_1))$; and $a(Z_1, Z_2, Z_3) \in L_0^2(F(Z_3|Z_2, Z_1))$.

STEP - 2: $(d_m = d_\beta)$ Differentiating $E[m(Z;\beta^0)] = 0$ under the integral we obtain

$$\frac{\partial \beta^0(\theta_0)}{\partial \theta'} = -M^{-1} E\left[m(Z;\beta^0)\left\{s(Z_1)' + s(Z_2|Z_1)' + s(Z_3|Z_2, Z_1)'\right\}\right].$$

We show that $E[\varphi(C, G_C(Z)) S(C, G_C(Z))'] = E\left[m(Z;\beta^0)\left\{s(Z_1)' + s(Z_2|Z_1)' + s(Z_3|Z_2, Z_1)'\right\}\right]$ to complete the proof as before. We show this term by term for the three terms in $\varphi_{\mathbb{C}}(C, G_C(Z))$. As before, the argument β is omitted below to signify that terms are evaluated at $\beta = \beta^0$. To maintain uniformity in presentation we work with the expression

$$\begin{aligned} \varphi_{\mathbb{C}}(C, G_C(Z)) &= E[m(Z)|G_1(Z)] \\ &\quad + \frac{I(C=2) + I(C=3)}{P(C=2|G_2(Z)) + P(C=3|G_3(Z))}\left(E[m(Z)|G_2(Z)] - E[m(Z;\beta)|G_1(Z)]\right) \\ &\quad + \frac{I(C=3)}{P(C=3|G_3(Z))}\left(m(Z) - E[m(Z)|G_2(Z)]\right). \end{aligned}$$

Consider the first term of $\varphi(C, G_C(Z))$. Taking suitable conditional expectations and using the identity $\sum_{r=1}^{3} \dot{P}(C=r|G_r(Z)) = 0$, it follows that

$$E\left[E[m(Z)|G_1(Z)]\left\{\sum_{r=1}^{3} I(C=r)\frac{\dot{P}(C=r|G_r(Z))}{P(C=r|G_r(Z))} + s(Z_1)\right\}'\right] = E\left[m(Z)s(Z_1)'\right].$$

Noting the identity $1 = \sum_{r=1}^{3} I(C=r)$ and subsequently using $s(Z_2|Z_1) \in L_0^2(F(Z_2|Z_1))$ along with (2), it follows that $E\left[E[m(Z)|G_1(Z)] I(C \in \{2,3\}) s(Z_2|Z_1)'\right] = 0$. Similarly, first taking expectation conditional on $G_3(Z)$ and using (2), then replacing $P(C=3|G_3(Z))$ by $1 - P(C=2|G_2(Z)) - P(C=1|G_1(Z))$, and finally using $s(Z_3|Z_2, Z_1) \in L_0^2(F(Z_2|Z_2, Z_1))$, it follows that $E\left[E[m(Z)|G_1(Z)] I(C=3) s(Z_3|Z_2, Z_1)'\right] = 0$.

Now consider the second term of $\varphi(C, G_C(Z))$. Taking suitable conditional expectations and

using the identities $\sum_{r=1}^{3} P(C=r|G_r(Z))=1$ and $\sum_{r=1}^{3} \dot{P}(C=r|G_r(Z))=0$, it follows that

$$
\begin{aligned}
E\left[\frac{I(C=2)+I(C=3)}{P(C=2|G_2(Z))+P(C=3|G_3(Z))}\left(E[m(Z)|G_2(Z)]-E[m(Z;\beta)|G_1(Z)]\right)\sum_{r=1}^{3} I(C=r)\frac{\dot{P}(C=r|G_r(Z))'}{P(C=r|G_r(Z))}\right] \\
= \quad E\left[\left(E[m(Z)|G_2(Z)]-E[m(Z;\beta)|G_1(Z)]\right)\frac{\sum_{r=2}^{3}\dot{P}(C=r|G_r(Z))'}{1-P(C=1|G_1(Z))}\right] \\
= \quad -E\left[\left(E[m(Z)|G_2(Z)]-E[m(Z;\beta)|G_1(Z)]\right)\frac{\dot{P}(C=1|G_1(Z))'}{1-P(C=1|G_1(Z))}\right]=0
\end{aligned}
$$

where the last equality follows by taking expectation conditional on $G_1(Z)$. Similar arguments give the expectation of the cross product of the second term with $s(Z_1)$ is 0, and that with $I(C \in \{2,3\})s(Z_2|Z_1)$ is $E\left[m(Z)s(Z_2|Z_1)'\right]$ where, additionally, $s(Z_2|Z_1) \in L_0^2(F(Z_2|Z_1))$ is used to get rid of the term $E[m(Z)|G_1(Z)]$. Exactly same arguments give the expectation of the cross product of the second term with $I(C=3)s(Z_3|Z_2,Z_1)$ is 0 once we use $s(Z_3|Z_2,Z_1) \in L_0^2(F(Z_3|Z_2,Z_1))$ is used to get rid of the entire term $(E[m(Z)|G_2(Z)]-E[m(Z)|G_1(Z)])$.

Similar steps give that the expectation of the cross product of the third term of $\varphi(C,G_C(Z))$ with $S(C,G_C(Z))$ is $E[m(Z)s(Z_3|Z_2,Z_1)']$ where, instead of taking expectation conditional on $G_1(Z)$ (as for the second term above) one obviously needs to condition on $G_2(Z)$. ■

B.5 More on the forms of the influence function from remark (iii) following Proposition 1:

It may be useful to consider the simplified setup of the verify-in-sample case of Chen et al. (2008) to explore the structure of the influence function further for a better understanding of the contribution of the various parts of the influence function broken down to characterize the incremental information available due to lower levels of coarsenings. Let $\lambda = \mathbb{C}$. Let us follow a convention whereby a notation $G_0(Z)$ is used such that $E[m(Z;\beta)|G_0(Z)] := 0$; and let $R+1$ be such that $I(C \geq R+1)=0$ identically. One can express $\varphi(C,G_C(Z);\beta)$ in two equivalent forms:

$$
\begin{aligned}
\varphi(C,G_C(Z);\beta) &= \sum_{r=1}^{R}\frac{I(C\geq r)}{P(C\geq r|G_1(Z))}\left(E[m(Z;\beta)|G_r(Z)]-E[m(Z;\beta)|G_{r-1}(Z)]\right) \qquad (33)\\
&= \sum_{r=1}^{R}\left\{\frac{I(C\geq r)}{P(C\geq r|G_1(Z))}-\frac{I(C\geq r+1)}{P(C\geq r+1|G_1(Z))}\right\}E[m(Z;\beta)|G_r(Z)] \qquad (34)
\end{aligned}
$$

The expression in (33) characterizes the incremental relevant (for estimation of β^0) information in $G_r(Z)$ versus $G_{r-1}(Z)$. For each r, the term inside the parenthesis is the incremental information and is an innovation with respect to $G_{r-1}(Z)$ by virtue of the properties of conditional expectations. The weight given to it in the influence function increases with how often we observe this piece of information. Its contribution to the variance of the influence function also has the same weight.[17] Now consider the second expression, i.e., (34). For a given r, this works with the total relevant information contained in $G_r(Z)$ itself, i.e., $E[m(Z;\beta)|G_r(Z)]$. The weight given to it eliminates double-counting of information because $G_r(Z)$ not only contributes when $C=r$, but also does for all $C>r$ since $G_r(Z) \subset G_{r+s}(Z)$ for all $s=1,2,\ldots$.[18]

[17] If Z_1 is discrete then for a given value z_1 of Z_1 the weight is inversely related to $P(C >= r)$ inside the z_1-cell. Since the weights do not numerically add up to 1, actual estimation of β^0 can be performed after re-weighting the incremental information by $\frac{I(C\geq r)}{P(C\geq r|G_1(Z)=z_1)}/\sum_{j=1}^{R}\frac{I(C\geq j)}{P(C\geq j|G_1(Z)=z_1)}$ for better stability of the estimator [see Lunceford and Davidian (2004), Busso et al. (2009), Graham et al. (2012)].

[18] While the weights add up to 1, they may be negative and can have adverse consequences on estimation.

B.6 The Doubly robust property to parametric misspecifications mentioned in the discussion following Proposition 1 (Reference: footnote 5):

Denote the parametric models by $P(C \geq r|G_1(Z);\gamma)$ and $E[m(Z;\beta)|G_r(Z);\nu]$. Let $\tilde{\varphi}^*_\lambda(C, G_C(Z);\beta)$ denote the plugged-in moment vector at γ^* and ν^*. First consider the case where the parametric model $E[m(Z;\beta)|G_r(Z);\nu]$ is correct. Characterize it by $E[m(Z;\beta)|G_r(Z);\nu^*] = E[m(Z;\beta)|G_r(Z)]$. Using the expression of $\varphi_\lambda(C, G_C(Z);\beta)$ in the statement of Proposition 1, note that for any β,

$$\begin{aligned}
& E[\tilde{\varphi}^*_\lambda(C, G_C(Z);\beta)] \\
= \; & E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)]\right] \\
& + E\left[\frac{P(C \in \lambda|G_1(Z);\gamma^*)}{P(C \in \lambda)} \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|G_1(Z);\gamma^*)} \left(E[m(Z;\beta)|G_r(Z)] - E[m(Z;\beta)|G_{r-1}(Z)]\right)\right] \\
= \; & E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta)|G_1(Z)]\right] + 0 = E[m(Z;\beta)|C \in \lambda]
\end{aligned}$$

by (2). Hence $E[\tilde{\varphi}^*_\lambda(C, G_C(Z);\beta)] = 0 \Longleftrightarrow \beta = \beta^0$ identifies the same β^0 as (1). Now consider the case where the parametric model $P(C \geq r|G_1(Z);\gamma)$ is correct. Characterize it by $P(C \geq r|G_1(Z);\gamma^*) = P(C \geq r|G_1(Z))$. Using the expression in (3), note that for any β,

$$\begin{aligned}
& E[\tilde{\varphi}^*_\lambda(C, G_C(Z);\beta)] \\
= \; & E\left[\frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \frac{I(C = R)}{P(C = R|G_1(Z))} m(Z;\beta)\right] \\
& + E\left[\frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \sum_{r=2}^{R-1} \left\{\frac{I(C \geq r)}{P(C \geq r|G_1(Z))} - \frac{I(C \geq r+1)}{P(C \geq r+1|G_1(Z))}\right\} E[m(Z;\beta)|G_r(Z);\nu^*]\right] \\
& + E\left[\left\{\frac{I(C \in \lambda)}{P(C \in \lambda)} - \frac{P(C \in \lambda|G_1(Z))}{P(C \in \lambda)} \frac{I(C \geq 2)}{P(C \geq 2|G_1(Z))}\right\} E[m(Z;\beta)|G_1(Z);\nu^*]\right] \\
= \; & E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} m(Z;\beta)\right] + 0 + 0 = E[m(Z;\beta)|C \in \lambda]
\end{aligned}$$

by (2). Hence $E[\tilde{\varphi}^*_\lambda(C, G_C(Z);\beta)] = 0 \Longleftrightarrow \beta = \beta^0$ identifies the same β^0 as (1). ■

A discussion of the implications of the double robustness of the influence function on the (higher-order) asymptotic properties of semiparametric estimators can be found in Rothe and Firpo (2012).

B.7 A treatment of the result in Corollary 6 along the line of (4)–(8) (Reference: footnote 10):

As mentioned before, when $\lambda = \lambda_p = \{3\}$, an inverse probability weighted moment vector based on the complete data $(C = 3, G_3(Z))$ (the primary sample in this case) that is unbiased for $E[m(Z;\beta)|\lambda = \{3\}]$ under (2) is

$$\phi(C = 3, G_3(Z);\beta) := \frac{I(C = 3)}{P(C = 3)} m(Z;\beta).$$

The difference from Corollary 6 is the treatment of the multiple $\frac{I(C=3)}{P(C=3)}$ as non-unity. Now we demonstrate the respective contributions of $G_2(Z)$ and $G_1(Z)$ following the same arguments as in the discussion based on the moment restrictions (4) — (6). However, it is important to remember that the final efficiency result here based on the corresponding moment restrictions in (8) matches with that in Proposition 2 and not in Proposition 1. This compromise is unavoidable since we work under the moment restrictions (4) – (6) to match with the presentation in the main text.

The new $\phi(C=3,G_3(Z);\beta)$ requires modifying $\widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)$ as

$$\begin{aligned}
&\widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)\\
:= \;& \phi(C=3,G_3(Z);\beta) - E\left[\phi(C=3,G_3(Z);\beta)\phi_2(C,G(Z))|C\geq 2,G_2(Z)\right]\\
&\times\left(E\left[\phi_2^2(C,G(Z))|C\geq 2,G_2(Z)\right]\right)^{-1}\phi_2(C,G(Z))\\
= \;& \frac{I(C=3)}{P(C=3)}\left(m(Z;\beta)-E\left[m(Z;\beta)|G_2(Z)\right]\right)+\frac{I(C\geq 2)}{P(C\geq 2|G_1(Z))}\frac{P(C=3|G_1(Z))}{P(C=3)}E\left[m(Z;\beta)|G_2(Z)\right].
\end{aligned}$$

This follows because

$$\begin{aligned}
&E\left[\phi_2^2(C,G(Z))|C\geq 2,G_2(Z)\right]=\frac{P(C=2|G_1(Z))P(C=3|G_1(Z))}{P^2(C\geq 2|G_1(Z))},\text{ and}\\
&E\left[\phi(C=3,G_3(Z);\beta)\phi_2(C,G(Z))|C\geq 2,G_2(Z)\right]=\frac{P(C=2|G_1(Z))P(C=3|G_1(Z))}{P^2(C\geq 2|G_1(Z))}\frac{E[m(Z;\beta)|G_2(Z)]}{P(C=3)}.
\end{aligned}$$

Note that under the premise of Corollary 6, $\frac{I(C\geq 2)}{P(C\geq 2|G_1(Z))}$ is unity when $\lambda_p=\{3\}$ and $\lambda_a=\{2\}$. Because of such differences, here $E\left[\phi(.;\beta^0)\phi(.;\beta^0)'\right]-E\left[\widetilde{\phi}(.;\beta^0)\widetilde{\phi}(.;\beta^0)'\right]$ is positive semi-definite while no efficiency gain was possible in Corollary 6.[19]

Similarly, the modification of $\phi(C=3,G_3(Z);\beta)$ (and thus $\widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)$) requires modifying $\widetilde{\widetilde{\phi}}(C,G(Z);\beta)$ as

$$\begin{aligned}
\widetilde{\widetilde{\phi}}(C,G(Z);\beta) \;:=\; & \widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)-E\left[\widetilde{\phi}(C\geq 2,G_2(Z),G_3(Z);\beta)\phi_1(C,G(Z))'|G_1(Z)\right]\\
&\times\left(E\left[\phi_1(C,G(Z))\phi_1(C,G(Z))'|G_1(Z)\right]\right)^{-1}\phi_1(C,G(Z))\\
=\; & \frac{I(C=3)}{P(C=3|)}\left(m(Z;\beta)-E\left[m(Z;\beta)|G_2(Z)\right]\right)\\
&+\frac{I(C\geq 2)}{P(C\geq 2|G_1(Z))}\frac{P(C=3|G_1(Z))}{P(C=3)}\left(E\left[m(Z;\beta)|G_2(Z)\right]-E\left[m(Z;\beta)|G_1(Z)\right]\right)\\
&+\frac{P(C=3|G_1(Z))}{P(C=3)}E\left[m(Z;\beta)|G_1(Z)\right].
\end{aligned}$$

The above follows because

$$\left(E\left[\phi_1(C,G(Z))\phi_1(C,G(Z))'|G_1(Z)\right]\right)^{-1}=\frac{1}{P(C=1|G_1(Z))}\begin{bmatrix}\frac{1-P(C=2|G_1(Z))}{P(C=3|G_1(Z))} & 1\\ 1 & \frac{1-P(C=3|G_1(Z))}{P(C=2|G_1(Z))}\end{bmatrix},$$

and $E\left[\phi(C=3,G_3(Z);\beta)\phi_2(C,G(Z))|C\geq 2,G_2(Z)\right]$

$$=\frac{P(C=3|G_1(Z))}{P(C=3)}\frac{P(C=1|G_1(Z))}{P(C\geq 2|G_1(Z))}E[m(Z;\beta^0)|G_1(Z)]\left[P(C=3|G_1(Z)),P(C=2|G_1(Z))\right].$$

In fact, $\widetilde{\widetilde{\phi}}(C,G(Z);\beta)=\varphi_{\lambda=\{3\},[k]}(C,G_C(Z);\beta)$ defined in the statement of Proposition 2 for the scenario where $R=3$. This improvement here is in the same spirit as that in Corollary 6 when $\lambda_a=\{1,2\}$. Exactly same improvement follows by altering the order of projection, i.e., the one in footnote 19 first and then the one corresponding to $\phi_2(C,G(Z))$.

[19] Similarly, the difference from the result in Corollary 6 with $\lambda_a=\{1\}$ can be seen by defining

$$\begin{aligned}
\widetilde{\phi}^{\dagger}(C,G(Z);\beta) \;:=\; & \phi(.;.)-E\left[\phi(.;.)\phi_1(.)'|G_1(Z)\right]\left(E\left[\phi_1(.)\phi_1(.)'|G_1(Z)\right]\right)^{-1}\phi_1(.)\\
=\; & \frac{I(C=3)}{P(C=3)}\left(m(Z;\beta)-E[m(Z;\beta)|G_1(Z)]\right)+\frac{P(C=3|G_1(Z))}{P(C=3)}E[m(Z;\beta)|G_1(Z)].
\end{aligned}$$

For the same reason as above, here $E\left[\phi(.;\beta^0)\phi(.;\beta^0)'\right]-E\left[\widetilde{\phi}^{\dagger}(.;\beta^0)\widetilde{\phi}^{\dagger}(.;\beta^0)'\right]$ is positive semi-definite.

B.8 Efficiency gain under (13) due to the use of $\widehat{\beta}$ defined in (14) over the efficient GMM estimator based on the augmented moment vector in (12) (Reference: discussion just before Proposition 10):

We provide a stronger result here. Note that the two moment vectors in (12) are taken by ignoring the sample units $(C = 3, G_3(Z))$ and $(C = 2, G_2(Z))$ respectively. This was done to maintain uniformity in presentation that treats the second round of coarsening [see footnote 8] as complete deletion the "irrelevant" sample units. Instead, consider the alternative interpretation as in footnote 8 whereby we retain the Z_1-information of the relevant units [similar to Chaudhuri and Guilkey (2013)]. Thus the extra penalty in efficiency, such as in Corollary 9 and Proposition 10, is no longer a concern. Accordingly, consider the efficient GMM estimator, denoted by $\widetilde{\beta}$, based on the following variant of the augmented moment vector in (12):

$$\begin{aligned}
\tilde{\varphi}(C, G_C(Z); \beta) &= \left[\tilde{\varphi}_2(C, G_C(Z); \beta)', \tilde{\varphi}_3(C, G_C(Z); \beta)'\right]', \text{ where for } r = 2, 3, \\
\tilde{\varphi}_r(C, G_C(Z); \beta) &:= \frac{I(C = \infty)}{P(C = \infty | Z_1)}(m(Z; \beta) - E[m(Z; \beta)|Z_1]) + E[m(Z; \beta)|Z_1] \\
&\quad + \left[\frac{I(C \in \{r, \infty\})}{P(C \in \{r, \infty\}|Z_1)} - \frac{I(C = \infty)}{P(C = \infty|Z_1)}\right](E[m(Z; \beta)|Z_1, Z_2] - E[m(Z; \beta)|Z_1]).
\end{aligned}$$

For brevity of notation, here we use Z_1, (Z_1, Z_2) and (Z_1, Z_3) instead of $G_1(Z)$, $G_2(Z)$ and $G_3(Z)$ respectively. Also, the dependence on β is suppressed in the following for quantities evaluated at $\beta = \beta^0$. Now, when (13) holds, under standard assumptions it can be shown that

$$\begin{aligned}
\sqrt{N}\left(\widetilde{\beta} - \beta^0\right) &\xrightarrow{d} N\left(0, \tilde{\Omega} := \left(\begin{bmatrix} M \\ M \end{bmatrix}' \begin{pmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{pmatrix}^{-1} \begin{bmatrix} M \\ M \end{bmatrix}\right)^{-1}\right), \text{ where} \\
\Sigma_{rr} &:= J_{[1,\infty]} - J_{[r]} \text{ for } r = 2, 3, \\
\Sigma_{23} &= \Sigma_{32} := J_{[1,\infty]} - J_{[2]} - J_{[3]}, \\
J_{[1,\infty]} &:= \left[\frac{1}{P(C = \infty|Z_1)} Var(m(Z)|Z_1) + E[m(Z)|Z_1]E[m(Z)'|Z_1]\right], \text{ and} \\
J_{[r]} &:= E\left[\frac{P(C = r|Z_1)Var(E[m(Z)|Z_1, Z_r]|Z_1)}{P(C = \infty|Z_1)P(C \in \{r, \infty\}|Z_1)}\right] \text{ for } r = 2, 3.
\end{aligned}$$

On the other hand, we know from the discussion around (14) that

$$\sqrt{N}\left(\widehat{\beta} - \beta^0\right) \xrightarrow{d} N\left(0, \Omega := \left(M'\left(J_{[1,\infty]} - J_{[2]} - J_{[3]}\right)^{-1} M\right)^{-1}\right).$$

In the following we show that $\tilde{\Omega} - \Omega$ is positive semi-definite (psd). By using the formula of partitioned inverse, this is equivalent to showing that

$$\left[\Sigma_{22}^{-1} + (I - \Sigma_{22}^{-1}\Sigma_{23})\Sigma_{33.2}^{-1}(I - \Sigma_{22}^{-1}\Sigma_{23})'\right]^{-1} - \left(J_{[1,\infty]} - J_{[2]} - J_{[3]}\right)$$

is psd where $\Sigma_{33.2} := \Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{23}$. By using Woodbury's identity (first equality below), we obtain

$$\begin{aligned}
&\left[\Sigma_{22}^{-1} + (I - \Sigma_{22}^{-1}\Sigma_{23})\Sigma_{33.2}^{-1}(I - \Sigma_{22}^{-1}\Sigma_{23})'\right]^{-1} \\
= \;& \Sigma_{22} - \Sigma_{22}(I - \Sigma_{22}^{-1}\Sigma_{23})\left(\Sigma_{33.2} + (I - \Sigma_{22}^{-1}\Sigma_{23})'\Sigma_{22}(I - \Sigma_{22}^{-1}\Sigma_{23})\right)^{-1}(I - \Sigma_{22}^{-1}\Sigma_{23})'\Sigma_{22} \\
= \;& \Sigma_{22} - (\Sigma_{22} - \Sigma_{23})\left(\Sigma_{33} + \Sigma_{22} - 2\Sigma_{32}\right)^{-1}(\Sigma_{22} - \Sigma_{23}).
\end{aligned}$$

From the definitions, we know that $\Sigma_{22} = J_{[1,\infty]} - J_{[2]}$, $\Sigma_{22} - \Sigma_{23} = J_{[3]}$ and $\Sigma_{33} + \Sigma_{22} - 2\Sigma_{32} = J_{[2]} + J_{[3]}$. Therefore,

$$\begin{aligned}
& \left[\Sigma_{22}^{-1} + (I - \Sigma_{22}^{-1}\Sigma_{23})\Sigma_{33.2}^{-1}(I - \Sigma_{22}^{-1}\Sigma_{23})'\right]^{-1} - \left(J_{[1,\infty]} - J_{[2]} - J_{[3]}\right) \\
= \; & \Sigma_{22} - (\Sigma_{22} - \Sigma_{23})\left(\Sigma_{33} + \Sigma_{22} - 2\Sigma_{32}\right)^{-1}(\Sigma_{22} - \Sigma_{23}) - \left(J_{[1,\infty]} - J_{[2]} - J_{[3]}\right) \\
= \; & \left(J_{[1,\infty]} - J_{[2]}\right) - J_{[3]}\left(J_{[2]} + J_{[3]}\right)^{-1} J_{[3]} - \left(J_{[1,\infty]} - J_{[2]} - J_{[3]}\right) \\
= \; & J_{[3]} - J_{[3]}\left(J_{[2]} + J_{[3]}\right)^{-1} J_{[3]}.
\end{aligned}$$

Hence $\tilde{\Omega} - \Omega$ is psd if and only if $J_{[3]} - J_{[3]}\left(J_{[2]} + J_{[3]}\right)^{-1} J_{[3]}$ is psd. However, since the relevant matrices are positive definite (in particular, nonsingular), this holds if and only if

$$J_{[3]}^{-1}\left(J_{[2]} + J_{[3]}\right) J_{[3]}^{-1} - J_{[3]}^{-1} = J_{[3]}^{-1} J_{[2]} J_{[3]}^{-1}$$

is psd. This is true under the maintained assumptions (see the definition of $J_{[2]}$ above). ■