Missing endogenous variables in conditional moment restriction models

Antonio Cosma*
Department of Management, University of Bergamo
and
Andreï Victorovitch Kostyrka
Department of Economics and Management, University of Luxembourg
and
Gautam Tripathi
Department of Economics and Management, University of Luxembourg

10th of December, 2025

Abstract

We estimate finite dimensional parameters in conditional moment restriction (CMR) models when at least one of the endogenous variables (outcomes and/or explanatory variables) in the model is missing for some individuals in the sample. We demonstrate that efficiency gains in estimation occur if and only if there is at least one endogenous variable — included in or excluded from the CMR model — that is nonmissing (observed for all individuals in the sample), which we show characterizes informative imputation. We propose a semiparametrically efficient estimator which is also "doubly robust." To illustrate the insights our estimator can provide in empirical applications with large sample sizes, we artificially induce missingness in the female labor supply model of Angrist and Evans (1998). Despite medium levels of missingness in female labor income (the outcome) and a sample size exceeding 200,000 observations, the inverse propensity score weighted generalized method of moments (GMM) estimator finds only a statistically insignificant negative effect of having a third child (the endogenous regressor) on labor income. In contrast, our efficient estimator yields point estimates of this effect that are not only comparable to the GMM estimates but are also statistically significant.

Keywords: Efficient estimation, Informative imputation, Missing at random

*We thank the editor Ivan Canay, an associate editor, and two anonymous referees for comments that greatly improved the paper. We are also grateful to Xiaohong Chen, Paul Devereux, Patrick Gagliardini, Jinyong Hahn, Valentin Patilea, and seminar participants at the University of Luxembourg, EcoSta 2023 (Waseda University), the 2023 Econometric Society Summer Meeting (Barcelona), the 2024 Asia Meeting (Ho Chi Minh City), and the 2025 World Congress (Seoul) for helpful suggestions. Andreï V. Kostyrka acknowledges financial support from FNR-Luxembourg through a PRIDE grant for the Migration and Labor (MINLAB) doctoral training unit. Simulation experiments were conducted using the University of Luxembourg HPC facilities.

1 Introduction

Applied researchers frequently estimate models using datasets where certain variables are missing for some individuals in the sample. E.g., Abrevaya and Donald (2017) note that almost 40% of the papers that appeared in the American Economic Review, the Journal of Human Resources, the Journal of Labor Economics, and the Quarterly Journal of Economics between 2006–2008 dealt with missing data, and in almost 70% of these cases the missing observations were simply dropped. However, dropping each observation with a missing variable and estimating the model only on the subsample where all variables are observed leads to selection bias. The term "selection bias" is a generic description of the problem that arises in identifying features of a "full" population from an "observed" subpopulation without taking into account the relationship between the two (the "full" vs. "observed" terminology, defined in Section 2, is from Robins, Rotnitzky, and Zhao, 1994). If not corrected, selection bias can lead to severely misleading inference. There are two mutually exclusive approaches for dealing with the selection problem: "selection on observables" and "selection on unobservables." In a selection on observables approach — selection on unobservables is briefly discussed in the supplement (cf. Appendix D.2.1) — a "missing at random" (MAR) assumption is made that, conditional on the nonmissing variables, has the effect of randomly assigning the missingness label to the "potentially missing" variables in the full population. [A random variable is said to be "potentially missing" if the probability that it is not observed for each individual lies in the open interval $(0, 1)$. In contrast, a random variable is "nonmissing" if the probability that it is not observed for each individual is zero.] This random assignment feature of the MAR assumption leads to an "inverse probability weighted" (IPW) scheme that makes the full population and the observed subpopulation statistically indistinguishable, thereby enabling identification of the full population features from the observed subpopulation.

In this paper, we consider the estimation of finite dimensional parameters in conditional moment restriction (CMR) models when some, or all, of the endogenous variables, i.e., those variables that do not appear in the conditioning set, are potentially missing. The missing endogenous variables can either be endogenous outcomes, or endogenous explanatory variables, or both. Endogeneity — pervasive in empirical research and observational studies so much so that it is almost a defining feature of microeconometrics — typically arises in the context of omitted variables, simultaneity or reverse causality, measurement error, and model interpretation. Missingness naturally occurs due to reporting issues or when researchers use multiple data sources to compile their datasets. Thus, endogeneity and missingness are both widespread in empirical applications and ignoring either leads to biased statistical inference. Nevertheless, applied researchers often choose to ignore one or the other to simplify their tasks.

To identify the parameters, we use a selection on observables approach pioneered by Graham (2011) for unconditional moment restriction (UCMR) models, and extended by Hristache and Patilea (2017), henceforth HP, for CMR models, who show that a moment condition model and the MAR assumption in the full population are equivalent to a system of sequential moment restrictions in the observed subpopulation. [Exogenous variables, i.e., variables appearing in the conditioning set, can also be missing in empirical applications. However, they have to be handled differently than missing endogenous variables because, unlike the latter, they do not lead to sequential conditional moment restrictions (HP, p. 740). Research on this topic is in progress and will be reported in a subsequent paper.] The framework of HP is very general and accommodates unconditional and conditional moment restrictions (cf. Remark D.1), infinite-dimensional parameters, missing outcomes, and missing exogenous covariates. Their focus, however, is on establishing their equivalence result. In contrast, our goal is the semiparametrically efficient estimation of the parameters of interest. Crucially, HP do not consider the role that nonmissing endogenous variables — ubiquitous in applied research — play in generating efficiency gains. For this reason, they cannot provide the necessary and sufficient conditions under which imputation is informative, whereas we do (cf. Remark 4.1 for a detailed discussion of this conceptual distinction). The terms

"informative" and "uninformative" imputation are defined in the discussion preceding Lemma 4.2. The necessary and sufficient conditions for imputation to be informative that we develop are not merely technical: they offer clear practical guidance for empirical work. As we elaborate below, CMR models with missing endogenous variables arise frequently in applications, yet the conditions under which imputing these variables yields efficiency gains are often misunderstood.

The main contributions of our paper are as follows: (i) To our knowledge, it is the first to show that in CMR models with potentially missing endogenous variables (outcomes and/or covariates), the existence of nonmissing endogenous variables is both necessary and sufficient for achieving efficiency gains in estimation from the observed sample, and this condition is equivalent to imputation being informative (Lemma 4.2). The nonmissing endogenous variables can be endogenous outcomes and/or covariates included in the model, and/or endogenous variables that are excluded from the CMR model but enter the propensity score function (Section 2). The efficiency bound for the model parameters reveals a new CMR for constructing efficient estimators that are also "doubly robust" (Section 4.1). (ii) We propose a smoothed empirical likelihood (SEL) estimator that uses the observed sample and is semiparametrically efficient (Section 4.2). For fast and reliable implementation of the SEL estimator and related inference, we have developed an open-source R package called `smoothemplik` (Kostyrka, 2025), available on the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/package=smoothemplik`. (iii) To illustrate the insights our estimator can provide in empirical applications with large sample sizes, we artificially induce missingness in the female labor supply model of Angrist and Evans (1998). We find that even with medium levels of missingness in female labor income (the outcome variable), having more than 200,000 observations is insufficient for a researcher using IPW generalized method of moments (GMM) to detect a statistically significant negative effect of having a third child (the endogenous explanatory variable) on labor income. In contrast, our efficient estimator yields point estimates of this effect that are not only comparable in sign and magnitude to the GMM estimates but are also statistically significant (Section 5). (iv) A simulation study (supplement Appendix C) reveals that the SEL estimator performs well in medium-sized samples for both point estimation and inference. The efficiency gains achieved are comparable to the maximum gains the simulation design can deliver.

The paper is organized as follows. Section 2 introduces a general CMR model with potentially missing endogenous variables. Section 3 discusses identification, and Section 4 develops efficient estimation and inference. Section 5 presents the empirical illustration, and Section 6 concludes with practical suggestions for researchers dealing with missing endogenous variables in CMR models. Implementation details, the simulation study, additional examples, and all proofs are in Appendix A–D of the online supplement.

2 Model with missing endogenous variables

Let Y_i^*, Z_i, X_i be random (column) vectors for individual $i = 1, \ldots, n$. Vector Y_i^* consists of endogenous variables (outcomes and/or explanatory variables), all of which are simultaneously not observed for some individuals in the sample. In contrast, vector Z_i consists of those endogenous variables (can be endogenous outcomes and/or endogenous covariates) that are observed for each individual in the sample. Similarly, X_i is a vector of exogenous variables which are observed for each individual in the sample. We refer to the coordinates of Y^* as being potentially missing or simply "missing" (for some individuals). Analogously, the coordinates of (Z, X) are referred to as being "nonmissing" (for all individuals).

For each i, we also observe the dummy variable $D_i := 1$ if all coordinates of Y_i^* are observed, and $D_i := 0$ if all coordinates of Y_i^* are missing. We let $Y_i := D_i Y_i^* + (1 - D_i)\mathfrak{m}$ denote the observed version of Y_i^*, where $\mathfrak{m}_{\dim(Y^*)\times 1}$ is a vector of pre-specified numbers for coding missingness, e.g., $\mathfrak{m} := (99999, \ldots, 99999)_{\dim(Y^*)\times 1}$. Following Robins, Rotnitzky, and Zhao (1994, p. 848), RRZ

hereafter, we refer to (Y_i^*, Z_i, X_i) as the "full data," and (D_i, Y_i, Z_i, X_i) as the "observed data," for individual i. Hence, $(Y_i^*, Z_i, X_i : 1 \leq i \leq n)$ is the "full sample" and $(D_i, Y_i, Z_i, X_i : 1 \leq i \leq n)$ the "observed sample." The subsample with no missing observations — obtained from the observed sample by discarding those i for which $D_i = 0$ — is called the "validation sample."

A large class of econometric models in applied economics can be written as a system of conditional moment equalities, namely, there exists $\theta^* \in \Theta \subset \mathbb{R}^{\dim(\theta^*)}$ such that

$$\mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}, \tag{2.1}$$

where $g := g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*)$ is a vector of residuals (known up to θ^*) from a set of structural equations used by the researcher to model a system of relationships between the missing and the nonmissing variables. [If there is no conditioning, then (2.1) becomes an UCMR model with some variables missing as in Chen, Hong, and Tarozzi (2008) and Graham (2011); cf. Example D.4.] The missing variables Y^* and the nonmissing variables $Z_{\text{in}} \subset Z := (Z_{\text{in}}, Z_{\text{ex}})$ are classified as endogenous (with respect to g) because they appear in g but not in the conditioning set in (2.1). We refer to Z_{in} as the included (in g) endogenous variables. The excluded (from g) nonmissing endogenous variables Z_{ex} appear neither in g nor in the conditioning set in (2.1), either due to exclusion restrictions imposed by economic theory or because of their auxiliary nature. Nonetheless, Z_{ex} (together with the remaining nonmissing variables) may be present in the probability mechanism generating the missing Y^* (i.e., the propensity score defined in Section 3), and being correlated with g can influence the probability that Y^* is observed (cf. Example D.1). The nonmissing variables $X := (X_{\text{in}}, X_{\text{ex}})$ in the conditioning set in (2.1) are classified as exogenous (with respect to g), where X_{in} are the included instrumental variables (IV) and X_{ex} the excluded IV. Included instruments are exogenous variables in g, whereas excluded instruments are those exogenous variables not in g but, based on theoretical or external considerations, appear in the conditioning set to help identify θ^*. X_{ex} also contains exogenous variables that are in the propensity score but are excluded from g. The conditional distribution of $Y^*, Z \mid X$, and the marginal distribution of X (denoted by P_X), are unknown. The objective is to use the observed sample $(D_i, Y_i, Z_i, X_i : 1 \leq i \leq n)$ to efficiently estimate θ^* in the CMR model (2.1).

Example 2.1 (IV regression with missing outcomes)**.** The canonical example of (2.1) is the linear regression $Y^* = \alpha^* + X_{\text{in}}'\beta^* + Z_{\text{in}}'\gamma^* + U$, where only the scalar outcome Y^* is missing for some observations, Z_{in} is the vector of nonmissing included endogenous regressors, and $\mathbb{E}[U \mid X] \overset{P_X\text{-a.s.}}{=} 0$ signifying that the nonmissing included and excluded regressors $X := (X_{\text{in}}, X_{\text{ex}})$ are exogenous with respect to U. Here, $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := U = Y^* - \alpha^* - X_{\text{in}}'\beta^* - Z_{\text{in}}'\gamma^*$ with $\theta^* := (\alpha^*, \beta^*, \gamma^*)$. If all regressors are endogenous, then $X_{\text{in}} = \vec{\varnothing}$, i.e., X_{in} is the empty vector, $X := X_{\text{ex}}$, and the definition of θ^* is adjusted by dropping β^*. The case where the outcome variable Y_1^* and some of the included endogenous explanatory variables Y_2^* are missing is handled by letting $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := Y_1^* - \alpha^* - X_{\text{in}}'\beta^* - Z_{\text{in}}'\gamma^* - Y_2^{*\prime}\delta^*$ with $Y^* := (Y_1^*, Y_2^*)_{1+\dim(Y_2^*)\times 1}$ and $\theta^* := (\alpha^*, \beta^*, \gamma^*, \delta^*)$. In all cases, if there are nonmissing excluded endogenous variables then $Z_{\text{ex}} \neq \vec{\varnothing}$. A classic empirical application that fits the framework of Example 2.1 is estimating the returns to schooling as in Balestra and Backes-Gellner (2017), where the outcome (earnings) is subject to missingness due to survey nonresponse and a nonmissing included endogenous regressor (educational attainment) is instrumented using an excluded IV (Swiss reforms on compulsory schooling). □

Example 2.2 (IV regression with nonmissing outcomes and missing endogenous covariates)**.** If the scalar outcome (denoted by $Z_{1,\text{in}}$) is nonmissing but endogenous explanatory variables Y^* are missing, then let $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := Z_{1,\text{in}} - \alpha^* - X_{\text{in}}'\beta^* - Z_{2,\text{in}}'\gamma^* - Y^{*\prime}\delta^*$ with $Z_{\text{in}} := (Z_{1,\text{in}}, Z_{2,\text{in}})_{1+\dim(Z_{2,\text{in}})\times 1}$ and $\theta^* := (\alpha^*, \beta^*, \gamma^*, \delta^*)$. If there are no nonmissing included endogenous covariates, then $Z_{2,\text{in}} := \vec{\varnothing}$ and the definition of θ^* is adjusted by dropping γ^*. If there are nonmissing excluded endogenous variables, then $Z_{\text{ex}} \neq \vec{\varnothing}$. Empirical studies consistent with the setup of Example 2.2

include: Bennedsen, Nielsen, Perez-Gonzalez, and Wolfenzon (2007, Section II.A), who estimate the causal effect of CEO succession on firm profitability, where the outcome is nonmissing but the endogenous dummy regressor (indicating if the incoming CEO is family) is missing due to dataset merging and instrumented by the gender of the outgoing CEO's first child; McDonough and Millimet (2017, Section 4), who instrument missing birth weight (an endogenous regressor) with nutritional program participation in a regression with math test scores as the nonmissing outcome; and Stephens and Unayama (2019, Section III), who estimate a repeated cross-section linear probability model with a nonmissing binary outcome (shared living arrangement) and a missing endogenous regressor (social security benefits), instrumented using cohort-based variation from amendments to the Social Security Act. □

3 Identification

To identify, i.e., uniquely define, θ^* without modeling the selection equation that generates the missing Y^*, we follow a selection on observables approach and assume that, conditional on all included and excluded nonmissing variables $Z := (Z_{\text{in}}, Z_{\text{ex}})$ and $X := (X_{\text{in}}, X_{\text{ex}})$, the missing observations on Y^* are missing at random, i.e.,

Assumption 3.1 (MAR)**.** *For all individuals, $D \perp\!\!\!\perp Y^* \mid Z, X$, where the symbol "$\perp\!\!\!\perp$" denotes stochastic independence.*

Let $\pi(Z,X) := \Pr(D = 1 \mid Z,X)$ denote the propensity score function. It is through the propensity score that the excluded nonmissing endogenous variables Z_{ex} enter the missing data mechanism, thereby rendering the imputation "informative," as defined in Section 4. Example D.1 in the supplement illustrates how nonmissing endogenous variables can be excluded from the CMR model and yet still appear in the propensity score. Henceforth, arguments taken by functions are suppressed when there is no danger of confusion, e.g., we write $\pi := \pi(Z,X)$ and $g_{\text{obs}} := g(Y, Z_{\text{in}}, X_{\text{in}}, \theta^*)$. The identity $Dg = Dg_{\text{obs}}$ (recall $g := g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*)$) due to the definition of Y is often used.

We can use MAR to evaluate $\mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X]$ when Y^* is missing because $\mathbb{E}[g \mid Z,X] \overset{\text{MAR}}{=} \mathbb{E}[g \mid Z,X,D=1] \overset{P_{Z,X}\text{-a.s.}}{=} \mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid Z,X]$. Therefore, under MAR,

$$\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \iff \mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \tag{3.1}$$

The right-hand-side of (3.1), which does not contain any missing Y^*, employs an IPW moment function to correct the effects of missingness. To emphasize the nonparametric nature of the propensity score function, we assume that

Assumption 3.2. *The functional form of $(Z,X) \mapsto \pi(Z,X)$ is fully unknown.*

Although π is unknown, it is nonparametrically identified and estimated as the conditional expectation of $D \mid Z,X$ from the observed sample (not just the validation sample) because (D,Z,X) are nonmissing. As shown in Proposition D.1 in the supplement, if the columns of the Jacobian matrix $J_{\dim(g)\times\dim(\theta^*)} := J(X,\theta^*) := \partial_\theta \mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X]$ are linearly independent P_X-a.s., then local identification of θ^* in the CMR (2.1) is equivalent to the local identification of θ^* in the IPW CMR (3.1). Moreover, the same condition leads to the global identification of θ^* whenever g is linear in θ^*. Since local identification of the parameters of interest in the missing data problem is not lost under MAR, and local identification is necessary for global identification, we maintain that

Assumption 3.3. *θ^* is identified.*

4 Efficient estimation and inference under MAR

Throughout the paper, the observed data $\mathscr{A}_i := (D_i, Y_i, Z_i, X_i)$, $i = 1, \ldots, n$, are assumed to be i.i.d. Unless specified otherwise, limits are taken as the sample size $n \to \infty$.

The equivalence in (3.1) reveals that, under MAR, θ^* can be estimated from the IPW CMR $\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$, which uses only the validation sample. However, in practice, estimating θ^* using the validation sample alone is not advisable due to the efficiency loss from discarding the observations on (Z, X), even though they are not missing. It is, therefore, important to know the efficiency bound for estimating θ^* in (2.1) under MAR (loosely speaking, the efficiency bound is the smallest asymptotic variance of an estimator that optimally utilizes the information from all nonmissing observations). We motivate the efficiency bound for θ^* using HP (Theorem 1), which extends the results in Graham (2011, Theorem 2.1) to CMR models.

Consider the system of $\dim(g) + 1$ equations

$$\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \tag{4.1}$$

$$\mathbb{E}[\frac{D}{\pi} - 1 \mid Z, X] \overset{P_{Z,X}\text{-a.s.}}{=} 0, \tag{4.2}$$

which do not contain any missing observations (cf. Remark D.3(i)). Eqn. (4.1) identifies θ^* in the validation sample, whereas (4.2) defines π in the observed sample. Remarkably, by Theorem 1 of HP, the CMRs in (4.1)&(4.2) are equivalent to (2.1) and MAR, i.e.,

$$(4.1)\&(4.2) \iff (2.1)\ \&\,\text{MAR}. \tag{4.3}$$

The equivalence in (4.3) reveals that, under MAR, the efficiency bound for θ^* in (2.1) is equal to the efficiency bound for estimating θ^* in (4.1)&(4.2), which is a system of sequential CMRs, i.e., CMRs with increasing conditioning sets.

To eliminate the effect of estimating π on (4.1), we follow earlier approaches (cf. Remark D.3(ii)) and transform the sequential system (4.1)&(4.2) into a system of conditional-on-X moment restrictions based on the vector of residuals from projecting Dg_{obs}/π, the moment function in (4.1), coordinatewise onto the tangent space of score functions for π, the "nuisance parameter" in (4.2). These residuals satisfy a conditional-on-X moment restriction, on which estimation of θ^* can be based.

Let $\mu_{\dim(g)\times 1} := \mu(Z, X, \theta^*) := \mathbb{E}[g \mid Z, X] \overset{\text{MAR}}{=} \mathbb{E}[g_{\text{obs}} \mid Z, X, D = 1]$ denote the nonparametric imputation of the moment function g based on (Z, X). [Nonparametric imputation of g is equivalent to nonparametric imputation of Y^* if and only if g is linear in Y^*. If g is nonlinear in Y^*, then for efficient estimation, the moment function g should be nonparametrically imputed rather than the missing variables themselves.] It is shown in Lemma D.2 that

$$\rho_{\dim(g)\times 1} := \rho(\mathscr{A}, \theta^*, \pi, \mu) := \frac{Dg_{\text{obs}}}{\pi} - \mu[\frac{D}{\pi} - 1] \tag{4.4}$$

is the vector of residuals from projecting Dg_{obs}/π coordinatewise onto the tangent space of score functions for π. Moreover (cf. Appendix D.3), ρ satisfies the conditional-on-X moment restriction

$$\mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \tag{4.5}$$

The residual vector ρ is nonparametrically estimable from the observed sample because π is nonparametrically estimable from the observed sample and, under MAR, μ is nonparametrically estimable from the validation sample (cf. Section 4.2). Therefore, estimation of θ^* can be based on (4.5).

In fact, ρ being free from the influence of estimating π suggests that (4.5) can also deliver an efficient estimator of θ^*. Indeed, as shown in (D.4), (D.5), (D.6) in the supplement, the Jacobian $\partial_{\theta^*}\mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} J$, whereas $\partial_\pi \mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ and $\partial_\mu \mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times\dim(g)}$. Hence, by Ai and Chen (2003, Theorems 4.1 and 6.1), the efficiency bound for estimating θ^* in (4.5) is given by $(\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}$, where $\Omega_\rho := \mathbb{E}[\rho\rho' \mid X] \overset{(4.5)}{=} \text{var}[\rho \mid X]$. Furthermore, as confirmed by Lemma 4.1, $(\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}$ is also the semiparametric efficiency bound for estimating θ^* in (2.1). Therefore, efficient estimation of θ^* can be based on (4.5).

Lemma 4.1. *Let Assumptions* 3.1*,* 3.2*,* 3.3 *hold. Then, under the regularity conditions specified in Assumption* D.1 *in the supplement, the semiparametric efficiency bound for estimating* θ^* *in* (2.1) *is given by* $\text{l.b.}(\theta^*) := (\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}$*. The efficiency bound does not decrease if the propensity score function is parametrically specified up to a finite dimensional parameter, or even if it is fully known.*

The abbreviation "l.b." stands for "lower bound" because the semiparametric efficiency bound is the greatest lower bound for the asymptotic variance of any $n^{1/2}$-consistent regular estimator. If there is no missingness, i.e., $Y^* \overset{\text{w.p.1}}{=} Y$, then $\rho = g$ and the bound in Lemma 4.1 becomes $(\mathbb{E}J'\Omega_g^{-1}J)^{-1}$ with $\Omega_g := \mathbb{E}[gg' \mid X]$, which is the well-known efficiency bound for estimating θ^* in the CMR model $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$.

The efficiency bound in Lemma 4.1 can be obtained by applying Ai and Chen (2012, Theorem 2.1) to (4.1)&(4.2). For completeness, Appendix D.3 contains an alternative derivation. Since $\theta \mapsto g(Y^*, Z, X, \theta)$ is not required to be differentiable, the bound is valid for non-smooth moment functions, e.g., quantile regression. The bound remains unchanged whether π is fully unknown, fully known, or known up to a finite-dimensional parameter, due to the propensity score function being ancillary to θ^* (Hahn, 1998, p. 319). This is expected, as π does not enter the moment condition (2.1) through which θ^* is defined. As noted in Chen, Hong, and Tarozzi (2008, p. 822) and Graham (2011, p. 439), ancillarity of π implies that to obtain an asymptotically efficient estimator of θ^*, the propensity score should be nonparametrically estimated, even if it is parametrically specified or fully known.

To measure the efficiency gain when all data in the observed sample — and not just those in the validation sample — are used to estimate θ^*, the efficiency bound in Lemma 4.1 is compared with $\text{l.b.}_{\text{VS}}(\theta^*)$, the efficiency bound for θ^* based on the IPW CMR $\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$. It is shown in Appendix D.3 that

$$\text{l.b.}(\theta^*) \leq_L \text{l.b.}_{\text{VS}}(\theta^*), \tag{4.6}$$

where the inequality $M_1 \leq_L M_2$ for symmetric matrices M_1, M_2 means that $M_1 - M_2$ is negative semidefinite (Löwner order).

The next result characterizes the necessary and sufficient conditions under which a semiparametrically efficient estimator of θ^*, based on the moment function ρ in (4.4) and utilizing all data in the observed sample, beats any estimator relying on the IPW moment function Dg_{obs}/π and the validation sample. Lemma 4.2, proved in Appendix D.3, is the key result of our paper and is central to understanding the conceptual difference between our work and the existing literature (cf. Remark 4.1). To facilitate the interpretation of Lemma 4.2, we introduce the following terminology: We define the nonparametric imputation of g to be "uninformative" if it is zero w.p.1, i.e., $\mu \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}$, in which case $\rho = Dg_{\text{obs}}/\pi$ and estimation using the observed sample yields no efficiency gains. In contrast, we say that the nonparametric imputation of g is "informative" if it is nonzero with positive probability ("w.p.p."), i.e., $\mu \overset{\text{w.p.p.}}{\neq} 0_{\dim(g)\times 1}$, thereby leading to maximal efficiency gains in estimation. As shown in Lemma 4.2, for efficiency gains to be realized when θ^* is estimated using the observed sample, it is necessary and sufficient that the nonparametric imputation of g be informative, which occurs if and only if there are nonmissing

endogenous variables included in or excluded from the CMR (2.1). Henceforth, keep in mind that $Z = \vec{\varnothing} \iff Z_{\text{in}} = \vec{\varnothing}$ & $Z_{\text{ex}} = \vec{\varnothing}$, whereas $Z \neq \vec{\varnothing} \iff Z_{\text{in}} \neq \vec{\varnothing}$ or $Z_{\text{ex}} \neq \vec{\varnothing}$.

Lemma 4.2. *Inequality* (4.6) *is sharp, meaning* $\text{l.b.}(\theta^*) = \text{l.b.}_{\text{VS}}(\theta^*)$ *holds if and only if* $Z = \vec{\varnothing} \iff \mu \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}$. *Consequently, efficiency gains in estimation, measured by the coordinatewise ratio* $\frac{\text{l.b.}_{\text{VS}}(\theta^*)}{\text{l.b.}(\theta^*)} \overset{(4.6)}{>} 1$, *occur if and only if* $Z \neq \vec{\varnothing} \iff \mu \overset{\text{w.p.p.}}{\neq} 0_{\dim(g)\times 1}$.

Lemma 4.2 establishes that estimation of θ^* using the validation subsample alone is asymptotically efficient if and only if there are no nonmissing endogenous variables included in or excluded from (2.1), i.e., $Z = \vec{\varnothing}$; and this is equivalent to the nonparametric imputation of g being uninformative, i.e., $\mu \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ as imputation is based solely on the nonmissing exogenous variables (included and excluded). Uninformative imputation does not yield efficiency gains in estimation. For efficiency gains to be realized when θ^* is estimated using the observed sample, it is necessary and sufficient that $Z \neq \vec{\varnothing} \iff \mu \overset{\text{w.p.p.}}{\neq} 0_{\dim(g)\times 1}$. In other words, efficiency gains in estimation from the observed sample occur if and only if there are nonmissing endogenous variables included in or excluded from the "structural" moment function g in the CMR (2.1); and this happens if and only if the nonparametric imputation of g is informative, i.e., $\mu \overset{\text{w.p.p.}}{\neq} 0_{\dim(g)\times 1}$ as it is based on the nonmissing endogenous variables (whether included or excluded) in addition to the nonmissing exogenous variables (whether included or excluded). It is informative imputation that leads to maximal efficiency gain in estimation from the observed sample. To attain maximal efficiency gains in estimation, imputation must be nonparametric, meaning that μ should be estimated nonparametrically from the validation sample. If a parametric model for μ is used for imputation, then efficient estimation is possible only if it is correctly specified (cf. Section 4.1).

Remark 4.1. The CMR model (2.1) differs *conceptually* from the models of HP, HP21 (Hristache and Patilea, 2021), and indeed the rest of the literature, in that these works never consider the possibility that nonmissing endogenous variables can generate efficiency gains in estimation. To see this, note that HP include in their propensity score auxiliary variables T that are excluded from their partially linear regression model with no endogenous regressors (HP, p. 736), without ever specifying whether T itself is exogenous or endogenous. When discussing the consequence of missing outcomes, HP (p. 740) state that *"... there is no information loss if the observations for which the outcome is missing are deleted from the sample,"* which suggests that, in models with missing outcomes, there is no scope for efficiency gains from using the observed sample. But Lemma 4.2 shows that this claim does not hold: Even in HP's model (where $Z_{\text{in}} = \vec{\varnothing}$), efficiency gains arise when $Z_{\text{ex}} = T$, i.e., when the auxiliary variables in their propensity score are endogenous. Unlike us, HP and HP21 do not consider endogenous covariates, and therefore cannot provide the necessary and sufficient conditions under which imputation is informative, i.e., when the observed sample yields maximal efficiency gains. E.g., when is imputing missing outcomes informative in linear regression models, the workhorse of the missing data literature? □

As noted earlier, Lemma 4.2 offers valuable insights into when to impute missing endogenous variables, which may be particularly appealing to applied researchers. It says that efficiency gains in estimating θ^* using all nonmissing observations in the observed sample — rather than just those in the validation sample — arise if and only if the nonparametric imputation of g is informative, for which it is necessary and sufficient that $Z \neq \vec{\varnothing} \iff Z_{\text{in}} \neq \vec{\varnothing}$ or $Z_{\text{ex}} \neq \vec{\varnothing}$. The following example demonstrates that imputing missing outcomes in linear regression models with no nonmissing endogenous regressors is uninformative and, hence, does not lead to efficiency gains in estimation.

Example 4.1 (When should missing outcomes be imputed?)**.** Consider the linear regression model $Y^* = \alpha_0^* + X_{\text{in}}'\beta_0^* + U$ with $\mathbb{E}[U \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0$, where the outcomes may be missing, there are

no nonmissing endogenous variables included and excluded ($Z_{\text{in}} = \vec{\varnothing}$ and $Z_{\text{ex}} = \vec{\varnothing}$), and no excluded instruments ($X_{\text{ex}} = \vec{\varnothing}$). Here, $g := U = Y^* - \alpha_0^* - X_{\text{in}}'\beta_0^*$; hence, $\mu := \mu(X_{\text{in}}, \alpha_0^*, \beta_0^*) = \mathbb{E}[U \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0$ implying that nonparametric imputation of g in this model is uninformative. Consequently, the validation sample alone can be used to construct a semiparametrically efficient estimator of (α_0^*, β_0^*). Indeed, Lemma 4.1 shows that the efficiency bound for estimating (α_0^*, β_0^*) is given by $(\mathbb{E}\pi(X_{\text{in}})J'\Omega_g^{-1}J)^{-1}$, where the propensity score $\pi(X_{\text{in}}) := \mathbb{E}[D \mid X_{\text{in}}]$ depends only on X_{in}, $J = -[\,1 \; X_{\text{in}}'\,]$, and $\Omega_g = \mathbb{E}[gg' \mid X_{\text{in}}]$. By Lemma 4.1, this coincides with the efficiency bound using the validation sample alone; moreover, based on the moment condition $\mathbb{E}[D(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*) \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0$, which holds only in the validation sample, the estimator proposed later in (4.12) attains the bound. In this model, imputing the missing Y^* using X_{in} (as no other nonmissing endogenous/exogenous variables are present) and employing the imputed values to estimate (α_0^*, β_0^*) does not lead to any efficiency gains. This is easily seen for the least-squares (LS) estimator, which is not semiparametrically efficient but serves to illustrate the point. Since $\mathbb{E}[Y^* \mid X_{\text{in}}, D = 1] \overset{\text{MAR}}{=} \mathbb{E}[Y^* \mid X_{\text{in}}] = \alpha_0^* + X_{\text{in}}'\beta_0^*$, the missing Y^* can be replaced by their imputed value $\hat{Y} := \hat{\alpha}_{0\text{VS}} + X_{\text{in}}'\hat{\beta}_{0\text{VS}}$, where $(\hat{\alpha}_{0\text{VS}}, \hat{\beta}_{0\text{VS}}) := \operatorname{argmin}_{\alpha,\beta} \sum_{i=1}^n D_i(Y_i - \alpha - X_{\text{in},i}'\beta)^2$ is the estimator of (α_0^*, β_0^*) from the validation sample alone. It is shown in the supplement (Appendix D.3.1) that the imputation $\hat{Y}$ does not provide any information about missing outcomes beyond what is available from the regression model itself so that $(\hat{\alpha}_{0\text{VS}}, \hat{\beta}_{0\text{VS}}) = (\hat{\alpha}_{0\text{LS}}, \hat{\beta}_{0\text{LS}})$, the LS estimator obtained from the observed sample by replacing the missing outcomes with their imputed values. Therefore, imputing the missing outcomes in linear regression models that have no nonmissing endogenous regressors does not lead to efficiency gains. □

But imputing the missing outcome when nonmissing endogenous regressors are present is informative and, hence, does lead to efficiency gains.

Example 4.2 (Example 4.1 contd.)**.** We now allow for nonmissing included endogenous regressors ($Z_{\text{in}} \neq \vec{\varnothing}$) and excluded instruments ($X_{\text{ex}} \neq \vec{\varnothing}$); nonmissing excluded endogenous variables may or may not be present, i.e., Z_{ex} may or may not be empty. The model now is $Y^* = \alpha^* + X_{\text{in}}'\beta^* + Z_{\text{in}}'\gamma^* + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] \overset{P_X\text{-a.s.}}{=} 0$. Here, $g := \varepsilon$; hence, $\mu := \mu(Z, X, \alpha^*, \beta^*, \gamma^*) = \mathbb{E}[\varepsilon \mid Z, X] \overset{\text{w.p.p.}}{\neq} 0$ implying that nonparametric imputation of g in this model is informative. Consequently, no estimator of $(\alpha^*, \beta^*, \gamma^*)$ using the validation sample alone is semiparametrically efficient. Indeed, the efficiency bound (cf. Example 4.3) — which is attained by the estimator in (4.12) with ρ defined in (4.4) — is strictly smaller than the efficiency bound for estimating $(\alpha^*, \beta^*, \gamma^*)$ from the validation sample alone because $\mu \overset{\text{w.p.p.}}{\neq} 0$. In this model, imputing the missing Y^* using (Z, X) and employing the imputed values to estimate $(\alpha^*, \beta^*, \gamma^*)$ does lead to efficiency gains. This is easily seen for the two-stage least-squares (2SLS) estimator, which is not semiparametrically efficient but illustrates the point nicely. Since $\mathbb{E}[Y^* \mid Z, X, D = 1] \overset{\text{MAR}}{=} \mathbb{E}[Y^* \mid Z, X] = \alpha^* + X_{\text{in}}'\beta^* + Z_{\text{in}}'\gamma^* + \mu$, the missing Y^* are imputed by $\hat{Y}_{\text{imp}} := \hat{\alpha}_{\text{VS}} + X_{\text{in}}'\hat{\beta}_{\text{VS}} + Z_{\text{in}}'\hat{\gamma}_{\text{VS}} + \hat{\mu}(Z, X)$, where $(\hat{\alpha}_{\text{VS}}, \hat{\beta}_{\text{VS}}, \hat{\gamma}_{\text{VS}}) := \operatorname{argmin}_{\alpha,\beta,\gamma} \sum_{i=1}^n \frac{D_i}{\hat{\pi}_i}(Y_i - \alpha - X_{\text{in},i}'\beta - \hat{Z}_{\text{in},i}'\gamma)^2$ is the 2SLS estimator in the validation sample, $\hat{Z}_{\text{in}}$ is the predicted Z_{in} from the reduced form equations for Z_{in} in the observed sample, $\hat{\mu}(Z, X) := \hat{\mu}(Z, X, \hat{\alpha}_{\text{VS}}, \hat{\beta}_{\text{VS}}, \hat{\gamma}_{\text{VS}})$ is obtained from the validation sample by nonparametrically regressing $\hat{\varepsilon}_{\text{VS}} := \hat{\alpha}_{\text{VS}} - X_{\text{in}}'\hat{\beta}_{\text{VS}} - Z_{\text{in}}'\hat{\gamma}_{\text{VS}}$ on (Z, X), and $\hat{\pi}_i := \hat{\pi}(Z_i, X_i)$ is the estimated propensity score (Section 4.2). It is shown in the supplement (Appendix D.3.2) that $\hat{Y}_{\text{imp}}$ provides information about missing outcomes that is not available from the regression model itself so that $(\hat{\alpha}_{\text{VS}}, \hat{\beta}_{\text{VS}}, \hat{\gamma}_{\text{VS}}) \neq (\hat{\alpha}_{2\text{SLS}}, \hat{\beta}_{2\text{SLS}}, \hat{\gamma}_{2\text{SLS}})$, the 2SLS estimator in the observed sample. Therefore, imputing the missing outcome when nonmissing endogenous regressors are present leads to efficiency gains. □

Example 4.3 (Example 2.1 contd.)**.** In the IV regression model with missing outcomes, $g = Y^* - \alpha^* - X_{\text{in}}'\beta^* - Z_{\text{in}}'\gamma^*$. By Lemma 4.1, the efficiency bound for θ^* is $(\mathbb{E}J'J/\Omega_\rho)^{-1}$, where $J = -[\,1 \; X_{\text{in}}' \; \mathbb{E}[Z_{\text{in}}'|X]\,]$,

$\Omega_\rho \overset{(D.11)}{=} \mathbb{E}[\pi^{-1}\operatorname{var}(Y^* \mid Z,X) \mid X] + \mathbb{E}[\mu^2 \mid X]$, and $\mu = \mathbb{E}[Y^* \mid Z,X] - \alpha^* - X_{\text{in}}'\beta^* - Z_{\text{in}}'\gamma^* \overset{\text{w.p.p.}}{\neq} 0$ because Z is endogenous with respect to g. Hence, imputation is informative and efficiency gains in estimation exist. □

Example 4.4 (Example 2.2 contd.)**.** In the IV regression model where the outcome $Z_{1,\text{in}}$ is nonmissing but endogenous explanatory variables Y^* are missing, $g := Z_{1,\text{in}} - \alpha^* - X_{\text{in}}'\beta^* - Z_{2,\text{in}}'\gamma^* - Y^{*\prime}\delta^*$. By Lemma 4.1, the efficiency bound for θ^* is $(\mathbb{E}J'J/\Omega_\rho)^{-1}$, where $J = -[\,1 \;\; X_{\text{in}}' \;\; \mathbb{E}[Z_{2,\text{in}}'|X] \;\; \mathbb{E}[Y^{*\prime}|X]\,]$, $\Omega_\rho \overset{(D.11)}{=} \delta^{*\prime}\mathbb{E}[\pi^{-1}\operatorname{var}(Y^* \mid Z,X) \mid X]\delta^* + \mathbb{E}[\mu^2 \mid X]$, and $\mu = Z_{1,\text{in}} - \alpha^* - X_{\text{in}}'\beta^* - Z_{2,\text{in}}'\gamma^* - \mathbb{E}[Y^{*\prime} \mid Z,X]\delta^* \overset{\text{w.p.p.}}{\neq} 0$ because Z is endogenous with respect to g. Hence, imputation is informative and efficiency gains in estimation exist. □

Example 4.5 (Missing completely at random (MCAR))**.** Y^* is said to be MCAR if $D \perp\!\!\!\perp (Y^*, Z, X)$. MCAR, which implies MAR, is too strong to be of much empirical interest. Nonetheless, it is worth noting that the results under MCAR can be obtained as a special case of the results under MAR. To see this, note that MCAR is equivalent to MAR plus the condition that $D \perp\!\!\!\perp (Z,X)$. But $D \perp\!\!\!\perp (Z,X)$ if and only if the propensity score function $(Z,X) \mapsto \pi(Z,X)$ is constant, i.e., there exists $\pi_{\text{MCAR}} \in (0,1)$ such that $\pi(Z,X) \overset{P_{Z,X}\text{-a.s.}}{=} \pi_{\text{MCAR}}$. Therefore, results under MCAR follow from those under MAR by simply replacing $\pi(Z,X)$ in Lemma 4.1 and Lemma 4.2 by π_{MCAR}. Efficiency gains exist under MCAR in the presence of nonmissing endogenous variables because the nonparametric imputation of g is then informative, i.e., μ is a nonzero function of (Z,X). □

4.1 Double robustness of estimators based on (4.5)

Before focusing on efficient estimation, we highlight the "double robustness" property of estimators of θ^* based on (4.5), which refers to θ^* being consistently estimable when either the selection model for D, or the model for imputing g, is correctly specified.

Since $Dg = Dg_{\text{obs}}$, we can write $\rho \overset{(4.4)}{=} \dfrac{Dg_{\text{obs}}}{\pi} - \mu[\dfrac{D}{\pi} - 1] = g + [\dfrac{D}{\pi} - 1][g - \mu]$. As π and μ are unknown functions, they have to be estimated nonparametrically. However, to avoid employing nonparametric methods, applied researchers often use "working" approximations of π and μ, denoted by $\pi_{\text{work}} := \pi_{\text{work}}(Z,X)$ and $\mu_{\text{work}} := \mu_{\text{work}}(Z,X)$, that are easier to estimate than π and μ (cf. Remark D.5). If $\pi_{\text{work}} \neq \pi$ and $\mu_{\text{work}} \neq \mu$, then (cf. Remark D.6)

$$\rho(\pi_{\text{work}}, \mu_{\text{work}}) := \frac{Dg_{\text{obs}}}{\pi_{\text{work}}} - \mu_{\text{work}}[\frac{D}{\pi_{\text{work}}} - 1] = g + [\frac{D}{\pi_{\text{work}}} - 1][g - \mu_{\text{work}}] \tag{4.7}$$

can be regarded as a noisy version of ρ with the additive term $[\dfrac{D}{\pi_{\text{work}}} - 1][g - \mu_{\text{work}}]$ capturing the error from simultaneously misspecifying π (the true propensity score function) and μ (the nonparametric imputation of g). As shown in Appendix D.3, under MAR,

$$\pi_{\text{work}} \overset{P_{Z,X}\text{-a.s.}}{=} \pi \quad \text{or} \quad \mu_{\text{work}} \overset{P_{Z,X}\text{-a.s.}}{=} \mu \implies \mathbb{E}[\rho(\pi_{\text{work}}, \mu_{\text{work}}) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \tag{4.8}$$

By (4.8), consistent estimation of θ^* can be based either on the CMR $\mathbb{E}[\rho(\pi, \mu_{\text{work}}) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ (when only the working model for D is correctly specified, i.e., $\pi_{\text{work}} = \pi$ but $\mu_{\text{work}} \neq \mu$), or on the CMR $\mathbb{E}[\rho(\pi_{\text{work}}, \mu) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ (when only the working model for imputing g is correctly specified, i.e., $\mu_{\text{work}} = \mu$ but $\pi_{\text{work}} \neq \pi$). However, neither leads to an efficient estimator of θ^* because $\rho(\pi, \mu_{\text{work}}) \neq \rho$ and $\rho(\pi_{\text{work}}, \mu) \neq \rho$. Since efficient estimation is possible only when $\pi_{\text{work}} = \pi$ and $\mu_{\text{work}} = \mu$, in which case $\rho(\pi_{\text{work}}, \mu_{\text{work}}) = \rho$, Section 4.2 focuses on constructing an efficient estimator.

As shown in Remark D.7, under MAR, $\operatorname{var}\rho(\pi,\mu) \leq_L \operatorname{var}\rho(\pi,\mu_{\text{work}})$, i.e., $\operatorname{var}\rho(\pi,\cdot)$ is minimized when $\mu_{\text{work}} = \mu$. Therefore, if the working model for D is correctly specified, then, in the spirit of RRZ (Sections 2.6 and 2.7), ρ is the "least noisy" version of $\operatorname{var}\rho(\pi,\cdot)$, on which we can base efficient estimation of θ^*.

4.2 Efficient estimation by empirical likelihood smoothing

If π and μ are fully known, then the smoothed empirical likelihood (SEL) estimator of θ^* (Kitamura, Tripathi, and Ahn, 2004, henceforth, KTA) based on (4.5) is asymptotically efficient, i.e., its asymptotic variance equals the semiparametric efficiency bound in Lemma 4.1, because $J \overset{\text{(D.4)}}{=} \partial_\theta \mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu) \mid X]\, P_X$-a.s. In fact, the asymptotic variance of the SEL estimator does not change if π and μ are replaced by their nonparametric estimators (cf. Lemma D.3). Therefore, we estimate θ^* using the SEL approach, which maximizes the empirical likelihood of the data in the observed sample subject to (4.5). Smoothing the empirical likelihood is required because (4.5) is a conditional restriction and the coordinates of X are assumed to be continuously distributed (cf. the discussion in Appendix A.1 in the supplement following Assumption A.1). Since $\mu = \mathbb{E}[g_{\text{obs}} \mid Z,X,D=1] = \dfrac{\mathbb{E}[Dg_{\text{obs}} \mid Z,X]}{\pi(Z,X)} = \dfrac{\mathbb{E}[Dg_{\text{obs}} \mid Z,X]\operatorname{pdf}_{Z,X}(Z,X)}{\mathbb{E}[D \mid Z,X]\operatorname{pdf}_{Z,X}(Z,X)}$, we estimate π and μ with the kernel estimators

$$\hat{\pi}_c(z,x) := \frac{(nc_n^{\dim(Z)+\dim(X)})^{-1}\sum_{k=1}^n D_k H_c(Z_k - z, X_k - x)}{\hat{f}_{Z,X}(z,x)}$$
$$\hat{\mu}_d(z,x,\theta) := \frac{(nd_n^{\dim(Z)+\dim(X)})^{-1}\sum_{k=1}^n D_k g(Y_k, Z_{\text{in},k}, X_{\text{in},k}, \theta) H_d(Z_k - z, X_k - x)}{\hat{f}^{\text{VS}}_{Z,X}(z,x)},$$

where $H_c(\cdot) := H(\cdot/c_n)$ and $H_d(\cdot) := H(\cdot/d_n)$ is a kernel of sufficiently high order to deal with the estimation bias in $(\hat{\pi}_c, \hat{\mu}_d)$, the subscripts $c := (c_n)$ and $d := (d_n)$ denote the bandwidths used to estimate π and μ, and

$$\hat{f}_{Z,X}(z,x) := \frac{\sum_{k=1}^n H_c(Z_k - z, X_k - x)}{nc_n^{\dim(Z)+\dim(X)}} \quad \& \quad \hat{f}^{\text{VS}}_{Z,X}(z,x) := \frac{\sum_{k=1}^n D_k H_d(Z_k - z, X_k - x)}{nd_n^{\dim(Z)+\dim(X)}}$$

estimate the joint density of (Z,X) in the observed and validation samples. Note that $\hat{\pi}_c$ is based on the observed sample, whereas $\hat{\mu}_d$ is based only on the validation sample.

For the remainder of the paper, let $\rho_j(\theta) := \rho(\mathscr{A}_j, \theta, \pi(Z_j,X_j), \mu(Z_j,X_j,\theta))$ and $\hat{\rho}_j(\theta) := \rho(\mathscr{A}_j, \theta, \hat{\pi}_c(Z_j,X_j), \hat{\mu}_d(Z_j,X_j,\theta))$. To motivate the SEL approach, for $i,j = 1,\dots,n$, let p_{ij} denote the conditional probability $\Pr(\mathscr{A} = \mathscr{A}_j \mid X = X_i)$ arising from a discrete distribution with support $(X_1,\dots,X_n) \times (\mathscr{A}_1,\dots,\mathscr{A}_n)$. Using these p_{ij} and kernel weights

$$w_{ij} := \frac{K_b(X_i - X_j)}{\sum_{k=1}^n K_b(X_i - X_k)} = \frac{(nb_n^{\dim(X)})^{-1} K_b(X_i - X_j)}{\hat{f}_X(X_i)},$$

where $K_b(\cdot) := K(\cdot/b_n)$ is a 2^{nd} order kernel, $b := (b_n)$ the bandwidth, and $\hat{f}_X(X_i) := \sum_{j=1}^n K_b(X_i - X_j)/nb_n^{\dim(X)}$, construct the smoothed loglikelihood $\sum_{i=1}^n \sum_{j=1}^n w_{ij} \log p_{ij}$. Then, given θ, solve the following optimization problem that finds the optimal discrete distribution to enforce the sample analog of (4.5), namely,

$$\max_{\substack{p_{ij}\in(0,1)\\ i,j=1,\dots,n}} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log p_{ij} \quad \text{s.t.} \quad \begin{cases} \sum_{j=1}^n p_{1j} = 1, \dots, \sum_{j=1}^n p_{nj} = 1 \\ \sum_{j=1}^n \hat{\rho}_j(\theta) p_{1j} = 0_{\dim(g)\times 1}, \ \dots, \ \sum_{j=1}^n \hat{\rho}_j(\theta) p_{nj} = 0_{\dim(g)\times 1}. \end{cases} \tag{4.9}$$

The solution to (4.9), denoted by $(\hat{p}_{ij}(\theta))_{i,j=1,\ldots,n}$, is given by (cf. Appendix D.3)

$$\hat{p}_{ij}(\theta) \overset{(D.43)}{=} \frac{w_{ij}}{1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)}, \qquad i,j=1,\ldots,n, \tag{4.10}$$

where $\hat{\lambda}_i(\theta)$, the Lagrange multipliers imposing the moment condition in (4.9), satisfy

$$\sum_{j=1}^{n} \frac{w_{ij}\hat{\rho}_j(\theta)}{1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)} = 0_{\dim(g)\times 1}, \qquad i=1,\ldots,n. \tag{4.11}$$

As $\hat{p}_{ij}(\theta) > 0$ for small enough b_n (cf. Footnote 26 in Appendix D.3), the smoothed empirical loglikelihood of θ is the value function of (4.9) given by

$$\widehat{\mathrm{SEL}}(\theta) := \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\hat{p}_{ij}(\theta) \overset{(4.10)}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(w_{ij}) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)),$$

where the $\hat{\lambda}_i(\theta)$ satisfy (4.11) (the "hat" in $(\widehat{\mathrm{SEL}}, \hat{\lambda}_i)$ emphasizes that it is based on $(\hat{\pi}, \hat{\mu})$). The estimator of θ^* is defined to be the maximizer of a trimmed version of $\widehat{\mathrm{SEL}}(\cdot)$, i.e.,

$$\hat{\theta} := \operatorname*{argmax}_{\theta\in\Theta} \widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta), \tag{4.12}$$

where — ignoring the term $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(w_{ij})$ as it does not depend on θ — the trimmed SEL objective function is given by

$$\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) := -\sum_{i=1}^{n}\hat{\mathbb{T}}_{1i}\sum_{j=1}^{n}\hat{\mathbb{T}}_{2j}w_{ij}\log(1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)), \tag{4.13}$$

and (abusing notation) the $\hat{\lambda}_i(\theta)$ in (4.13) satisfy the trimmed version of (4.11), namely,

$$\sum_{j=1}^{n} \frac{\hat{\mathbb{T}}_{2j}w_{ij}\hat{\rho}_j(\theta)}{1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)} = 0_{\dim(g)\times 1}, \qquad i=1,\ldots,n. \tag{4.14}$$

The variables $\hat{\mathbb{T}}_{1i} := \mathbb{1}(\hat{f}_X(X_i) \geq b_n^{\tau_b})$ and $\hat{\mathbb{T}}_{2j} := \mathbb{1}(\hat{f}_{Z,X}(Z_j,X_j) \geq c_n^{\tau_c}, \hat{f}^{\mathrm{VS}}_{Z,X}(Z_j,X_j) \geq d_n^{\tau_d})$, where $\tau_b, \tau_c, \tau_d \in (0,1)$, are trimming indicators included to control the instability of the local empirical likelihood $\theta \mapsto \sum_{j=1}^{n}\hat{\mathbb{T}}_{2j}w_{ij}\log(1+\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta))$ caused by the denominators of $w_{ij}, \hat{\pi}_c, \hat{\mu}_d$ becoming too small in the tails. Since $(\hat{\mathbb{T}}_{1i}, \hat{\mathbb{T}}_{2j}) \xrightarrow{\mathrm{p}} (1,1)$, this trimming scheme ensures that $\hat{\theta}$ is efficient by guaranteeing that, asymptotically, no data is lost.

Remark 4.2. (i) (SEL standard errors). The Hessian of the SEL objective function yields the SEL standard errors $\mathrm{se}_{\mathrm{SEL}}(\hat{\theta}^{(k)}) := \sqrt{[(-\nabla^2_{\theta\theta}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\hat{\theta}))^{-1}]_{kk}}$.

(ii) (Consistency and asymptotic normality). Under conditions that control the estimation error when (π, μ) is nonparametrically estimated by $(\hat{\pi}_c, \hat{\mu}_d)$, consistency and asymptotic normality of $\hat{\theta}$ can be shown by replicating KTA (Theorems 3.1 and 3.2). However, a rigorous proof of the consistency and asymptotic normality of $\hat{\theta}$ will only add length to our paper, without substantially increasing its contribution. Therefore, to minimize technical details, in Appendix D.3.3 we assume that $\hat{\theta}$ is consistent for θ^* and justify why, asymptotically, $n^{1/2}(\hat{\theta}-\theta^*)$ is distributed as a Gaussian random vector with mean zero and variance-covariance matrix equal to the efficiency bound in Lemma 4.1.

(iii) (Why SEL?). CMR models with unknown functions can also be efficiently estimated using the sieve minimum distance (SMD) approach of Ai and Chen (2003), which is suited for cases where these functions are not identifiable as conditional expectations or densities. Since the

unknown functions (π, μ) in (4.5) are conditional expectations easily handled by kernel estimators, we employ the SEL approach instead. The two methods are closely related: Footnote 4 of Ai and Chen notes that kernel estimators can be used within SMD, and Appendix D.3.3 shows that the continuous-updating SMD objective function (Ai and Chen, Eqn. 23) is a large-sample quadratic approximation of the SEL objective function. Therefore, as elaborated in Appendix A.3, both the SEL and SMD approaches can be effectively employed for efficient estimation in our setting, though other methods can also be used. Although both yield asymptotically efficient estimators, SEL inference based on the likelihood-ratio (LR) statistic tends to be more accurate in small samples than SMD's Wald-based inference (cf. Section 4.3 and Appendix A.3).

(iv) (Easing the computational burden of SEL). The SEL method can be computationally demanding due to its nonparametric smoothing. To facilitate its use, we developed the `R` package `smoothemplik` available on CRAN. All empirical and simulation results in this paper were obtained using `smoothemplik`. Cf. Appendix B.3 for details. □

4.3 Inference

The SEL estimator $\hat{\theta}$ is a nonparametric maximum likelihood estimator of θ^* under the constraint $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0$, enabling likelihood-ratio (LR)-based inference via the statistic $\mathrm{LR}(\theta) := 2[\widehat{\mathrm{SEL}}_{\mathbb{T}}(\hat{\theta}) - \widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta)]$. For hypotheses $\mathrm{H}_0 : R(\theta^*) = 0_{\dim(R)\times 1}$ with R a smooth vector function, rejection occurs for large $\mathrm{LR}(\hat{\theta}_R)$ with $\hat{\theta}_R := \underset{\theta\in\Theta: R(\theta)=0_{\dim(R)\times 1}}{\operatorname{argmax}} \widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta)$. Under H_0, $\mathrm{LR}(\hat{\theta}_R) \overset{\mathrm{d}}{\approx} \chi^2_{\dim(R)}$, so the statistic is asymptotically pivotal. Unlike the Wald statistic, the LR statistic is internally studentized and does not require variance estimation. Inverting $\mathrm{LR}(\theta)$ yields asymptotically valid, transformation-invariant confidence regions that respect parameter bounds. In small samples, LR confidence regions better capture the shape of the sampling distribution — e.g., they may be asymmetric or unbounded — whereas Wald regions (obtained by inverting the Wald statistic), being always symmetric and bounded, can yield unreliable inference. This difference is illustrated in Appendix C.3, where we describe the main findings of our simulation study. As noted in Appendix C.3.1 and C.3.2, for both simulation designs (58% missingness and 36% missingness), LR confidence intervals can be unbounded in one direction for small sample sizes (Figures C.5 and C.7). In contrast, for the same level of missingness and the same sample size, the Wald confidence intervals are always symmetric and bounded by construction, which may not accurately reflect the uncertainty in the sampling distribution.

5 Empirical illustration

Angrist and Evans (1998), hereafter AE, employ U.S. census data to establish a causal link between family size and female labor supply. Their dataset can be downloaded from `https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive`, and the variable names in this section are what AE use in their code. AE find that having more than two children has a strong negative economic impact on the labor supply of working mothers. E.g., labor market outcomes such as their labor income, their work status, the number of weeks worked, and the number of hours worked per week, are all negatively affected by the birth of a 3rd child. The number of children is potentially endogenous because fertility is likely to be simultaneously determined with the labor market outcomes. Therefore, to identify the causal effect of the number of children on a labor market outcome, AE use as excluded instruments $X_{\mathrm{ex}} := (\mathit{boys2}, \mathit{girls2})_{2\times 1}$, two dummy variables indicating whether both kids are boys or girls. These instruments are, at least intuitively, both valid and relevant: the former because the sex of a child is typically not influenced by the parents, and the latter because some parents prefer mixed-sex siblings, so having same-sex children increases the likelihood that the parents will conceive another child. [AE maintain that

sex is randomly assigned (cf. their p. 451), and the validity of their instruments is not rejected in most of their specifications. We work with their model specification with labor income as the outcome variable where the instrument validity is not rejected (pvalue $> .50$).]

To study the robustness of AE's finding to missing outcomes — because non-response to income-related questions is common in surveys — we carry out the following counterfactual exercise using the AE dataset (the "full" sample in the terminology of Section 2): We artificially induce missingness in the outcome variable (labor income) ranging from 1% to 46%. The resulting observed sample with the missing outcomes is then used to estimate a model of female labor earnings using the semiparametrically efficient SEL estimator (4.12). The validation sample is used to obtain the IPW-SEL and IPW-GMM estimators. By comparing the performance of these estimators for different levels of missingness, we can demonstrate the extent — had AE encountered missing outcomes in their data — to which the aforementioned finding in their paper is robust in the presence of missing observations. As discussed in Section 5.3, the results from this counterfactual exercise illustrate that our methodology can be used to address the issue of missing endogenous variables in applied work in a fruitful manner.

5.1 Labor earnings for working mothers

AE model the labor earnings for working mothers as $Y^* = \alpha^* + X_{\text{in}}'\beta^* + \gamma^* Z_{\text{in}} + U$, where the outcome $Y^* :=$ *incomem* (annual labor income of mother in thousands of 1995 dollars) is potentially missing, the vector of exogenous explanatory variables $X_{\text{in}} := (\textit{agem1}, \textit{agefstm}, \textit{boy1st})_{3\times 1}$ contains the current age of mother, the age of mother at first birth, and the sex of the first child, and $Z_{\text{in}} :=$ *morekids* is an endogenous dummy variable indicating that a mother has three or more kids. As in AE, we treat (*agem1*, *agefstm*) as continuously distributed and assume that $\mathbb{E}[U \mid X] \overset{P_X\text{-a.s.}}{=} 0$ with $X := (\textit{agem1}, \textit{agefstm}, \textit{boy1st}, \textit{boys2}, \textit{girls2})_{5\times 1}$. The AE moment function $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := Y^* - \alpha^* - X_{\text{in}}'\beta^* - \gamma^* Z_{\text{in}}$ with $\mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0$ and $\theta^* := (\alpha^*, \beta^*, \gamma^*)_{5\times 1}$. There are no nonmissing excluded endogenous variables, i.e., $Z_{\text{ex}} = \vec{\varnothing}$, so that $Z = Z_{\text{in}}$ throughout the empirical illustration.

AE consider six specifications, each corresponding to a different outcome variable. For each outcome, they report 12 estimators, namely, (LS, just-identified IV, over-identified 2SLS) × (1980 or 1990 data) × (all women or married women). For simplicity, we restrict our analysis to the sub-sample of white married females in 1980 (approx. 82% of all surveyed females are white in the 1980 Public Use Micro Sample), which yields a sample of size $n = 227{,}146$ and explains the minor discrepancy between the GMM estimates in Table 1 and the 2SLS estimates of Angrist and Evans (Table 7). The GMM estimates in Table 1 correspond to the 2SLS estimates in AE (Table 7, p. 465, row "Labor income", column 6), although, for convenience, we divide the labor income by 1000. Table 1 verifies that SEL estimates replicate the 2SLS estimates in AE when there are no missing outcomes. In Table 1, the model $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0$ is estimated using SEL, and the over-identified UCMR model $\mathbb{E}[\tilde{X} g] = 0_{6\times 1}$, where $\tilde{X} := (1, \textit{agem1}, \textit{agefstm}, \textit{boy1st}, \textit{boys2}, \textit{girls2})_{6\times 1}$, is estimated using iterated GMM to remove the dependence of the 2-step optimal GMM estimator on the initial estimator/weight matrix. As noted in Table 1, the hypothesis $\mathbb{E}[\tilde{X} g] = 0_{6\times 1}$ is not rejected by the J-test (pvalue $= .613$).

The GMM and SEL point estimates in Table 1 are similar, with smaller standard errors for the SEL estimates as expected. Estimates for (*agem1*, *agefstm*, *boy1st*) are virtually identical. Estimates for *morekids* may appear a bit different numerically ($\hat{\gamma}_{\text{GMM}} = -1.499, \hat{\gamma} = -2.046$), but they are statistically indistinguishable at levels of significance $\leq 1\%$. This is evident from a Hausman test of the null hypothesis that $\hat{\gamma}_{\text{GMM}}$ and $\hat{\gamma}$ estimate the same parameter (cf. Remark B.1). Hence, as their point estimates are similar, it is their standard errors that determine which estimator delivers a statistically significant estimate of the effect of having a 3$^{\text{rd}}$ child on labor income. The hypothesis that *morekids* is irrelevant for explaining labor income — testing $\gamma^* = 0$ against the alternative that

$\gamma^* \neq 0$ — is rejected at all reasonable significance levels by the SEL estimate (pvalue = .00008), but not by the GMM estimate (pvalue = .009). Thus, even with >200,000 observations, the GMM point estimate — although economically relevant — is not statistically different enough from zero to convincingly reject the irrelevance of *morekids* at the 1% level of significance. In contrast, the SEL point estimate remains both economically relevant and statistically significant at all reasonable significance levels. Therefore, as this problem with the GMM estimate only gets exacerbated when there is missingness in labor income, we carry out the counterfactual exercise outlined earlier.

5.2 Missingness in labor income

As described in Appendix B.1, we artificially induce missingness in labor income and create datasets — drawn randomly from the original AE dataset — containing missing outcomes. To ensure that our analysis is not influenced by a specific level of missingness in the labor income or a specific draw from the AE dataset, we consider 22 levels of missingness ranging from 1% to 46% and for each level of missingness we independently draw 1000 datasets with missing outcomes, and for each dataset we compute the following estimators:

- We estimate θ^* in the over-identified model $\mathbb{E}[\tilde{X}\frac{Dg_{\text{obs}}}{\pi}] = 0_{6\times 1}$ using iterated GMM, and undersmoothed bandwidth $\hat{c}_n^{\text{CV}}/3$ to estimate π, where $\hat{c}_n^{\text{CV}}$ is the cross-validated bandwidth for estimating π. We call this estimator $\hat{\theta}_{\text{GMM,IPW}}$ and compute $\text{se}_{\text{GMM}}(\hat{\theta}_{\text{GMM,IPW}})$, the GMM standard error. The $1/3$ factor and choice of smoothing bandwidths b_n, c_n, d_n are discussed in Appendix B.4.

- We estimate θ^* in the model $\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0$ using SEL and bandwidths $b_n = 1.2$ and $\hat{c}_n^{\text{CV}}/3$. We call this estimator $\hat{\theta}_{\text{SEL,IPW}}$ and compute $\text{se}_{\text{SEL}}(\hat{\theta}_{\text{SEL,IPW}})$.

- We estimate θ^* in (4.5) using the SEL estimator $\hat{\theta}$ defined in (4.12) with bandwidths $b_n = 1.2$, $\hat{c}_n^{\text{CV}}/3$, and $\hat{d}_n^{\text{CV}}/3$. We compute $\text{se}_{\text{SEL}}(\hat{\theta})$, the SEL standard error of $\hat{\theta}$.

For each level of missingness ranging from 1% to 46%, we therefore have 1000 i.i.d. copies of $(\hat{\theta}_{\text{GMM,IPW}}, \text{se}_{\text{GMM}}(\hat{\theta}_{\text{GMM,IPW}}))$, $(\hat{\theta}_{\text{SEL,IPW}}, \text{se}_{\text{SEL}}(\hat{\theta}_{\text{SEL,IPW}}))$, and $(\hat{\theta}, \text{se}_{\text{SEL}}(\hat{\theta}))$. It is useful to interpret them as 1000 independent researchers each possessing these three estimators and their standard errors for each of the 22 levels of missingness. The discussion of the results in Figure 2 and Table 2 is based on this interpretation. Additional results for higher levels of missingness are in the supplement (cf. Appendix B.2).

5.3 Results and discussion

Based on the 1000 experiments, Figure 1 plots the median standard errors of the estimated slope coefficients $(\beta^*, \gamma^*)_{4\times 1}$ as a function of the missingness. Since the size of the validation sample decreases as missingness increases, the standard errors of $(\hat{\beta}_{\text{SEL,IPW}}, \hat{\gamma}_{\text{SEL,IPW}})$ and $(\hat{\beta}_{\text{GMM,IPW}}, \hat{\gamma}_{\text{GMM,IPW}})$ are strictly increasing in the level of missingness. Moreover, the standard errors of $(\hat{\beta}_{\text{SEL,IPW}}, \hat{\gamma}_{\text{SEL,IPW}})$ are systematically smaller than the standard errors of $(\hat{\beta}_{\text{GMM,IPW}}, \hat{\gamma}_{\text{GMM,IPW}})$ because $\hat{\theta}_{\text{SEL,IPW}}$ is more efficient than the GMM estimator (SEL estimates a CMR model whereas GMM estimates an UCMR model). In turn, the standard errors of $(\hat{\beta}, \hat{\gamma})$ are smaller than the standard errors of $(\hat{\beta}_{\text{SEL,IPW}}, \hat{\gamma}_{\text{SEL,IPW}})$ because we have already shown that the SEL estimator $\hat{\theta}$ using all nonmissing observations — and not just those in the validation sample — is semiparametrically efficient. For missingness rates under 25%, the two are almost identical because efficiency gains at low missingness rates are offset by the noise from estimating the nonparametric imputation μ. Efficiency gains emerge as missingness increases, and we find that for missingness = (30%, 37%, 42%, 46%)

the standard error of $\hat{\gamma}_{\text{SEL,IPW}}$ is approximately $(1\%, 5\%, 11\%, 28\%)$ larger than that of $\hat{\gamma}$ (these numbers, difficult to read from Figure 1, are from the data used to create the figure). The one exception where the efficient estimator does not dominate its validation sample counterpart is the effect of *boy1st*, the gender of the first-born. This can be explained by the fact that *boy1st* is an uninformative predictor: Its t-statistic is ≈ 1 even in this large a sample (Table 1), and the induced missingness does not depend on *boy1st*.

To visually compare the relative performance of $\hat{\gamma}_{\text{GMM,IPW}}, \hat{\gamma}_{\text{SEL,IPW}}$, and $\hat{\gamma}$ as a function of the missingness in labor income, Figure 2 displays some features of the sampling distributions of these estimators across the 1000 researchers. For each estimator, the horizontal solid bars (the dot and the whiskers) indicate the median and the $.025^{\text{th}}$ and $.975^{\text{th}}$ quantiles of the point estimates across the researchers for different levels of missingness. The dashed lines are the medians, across all researchers, of the left- and right-endpoints of the 95% confidence intervals (CIs) for the true effect γ^*. As the missingness increases, the distribution of $\hat{\gamma}_{\text{GMM,IPW}}$ becomes more dispersed (the length of the whiskers around the median increases), and the left- and right-endpoints of the 95% CIs move steadily away from the point estimates. Compared to the GMM estimator, $\hat{\gamma}_{\text{SEL,IPW}}$ has smaller variance at 1%–20% levels of missingness; however, at high missingness levels, its dispersion increases, and a drift appears. In marked contrast, the semiparametrically efficient estimator $\hat{\gamma}$ yields 95% CIs of shorter length (the gap between the dashed lines) for high levels of missingness, lower variance of the estimator itself (the whisker width), and the median right-endpoint of the 95% CIs (the right dashed line) is always less than zero, indicating that the semiparametrically efficient estimator rejects the irrelevance of *morekids* at 5% significance in at least half the experiments for each level of missingness.

Table 2 complements Figure 2 by providing some additional information. It reports the fraction of 1000 experiments, expressed as a function of the missingness in labor income, for which the t-test fails to reject the hypothesis that *morekids* is irrelevant at the 10%, 5%, and 1% significance levels. The results in Table 2 can be interpreted as meaning that an overwhelming majority of the 1000 independent researchers using the GMM estimator would likely conclude that — irrespective of the extent of missingness in labor income — there is no statistically significant negative relationship between having more than two kids and labor income at the 1% significance level. Starting with just 20% missingness, we also see that almost 7% (resp. 23%) of the researchers using the GMM estimator would fail to reject the irrelevance of *morekids* at the 10% (resp. 5%) significance level. In contrast to the GMM estimator, with 20% missingness, all 1000 independent researchers using the SEL estimator $\hat{\gamma}$ with its smaller standard errors reject the irrelevance of *morekids* at the 10% and 5% significance levels, and only 6% of the researchers fail to reject that *morekids* is irrelevant at the 1% significance level. As the missingness in labor income increases, so does the failure to reject the irrelevance of *morekids*. With 46% missingness, almost (65%, 77%, 93%) of independent researchers using $\hat{\gamma}_{\text{GMM,IPW}}$ would find *morekids* irrelevant at the (10%, 5%, 1%) significance levels. In contrast, for the same missingness, only (28%, 43%, 76%) of independent researchers using the efficient estimator $\hat{\gamma}$ are likely to conclude irrelevance of *morekids* at the (10%, 5%, 1%) levels of significance.

In summary, the reduction in the earnings of working mothers due to having a 3^{rd} child, i.e., the negative causal effect of the endogenous explanatory variable *morekids* on labor income, can have manifold economic and social implications. However, our analysis of the AE dataset shows that if there is even medium missingness in the labor income, then having more than 200,000 observations may not be enough for the inverse propensity score weighted GMM estimator to deliver statistically significant estimates of this effect. In contrast, for the same levels of missingness, the semiparametrically efficient SEL estimator utilizes information from the nonmissing endogenous variable *morekids* to produce statistically significant point estimates of its effect on labor income that are comparable in sign and magnitude to the GMM estimates. The choice of the smoothing bandwidths (b_n, c_n, d_n) seems to have little impact on the reliability of SEL-based inference (cf. Appendix B.4). Hence, despite the lack of theory regarding how to choose the

smoothing bandwidths for the SEL approach, practitioners can use reasonably small bandwidths to smooth the empirical likelihood and nonparametrically estimate π and μ.

6 Conclusion

We believe the literature has overlooked the possibility that nonmissing endogenous variables — whether included in or excluded from CMR models — can lead to informative imputation and deliver efficiency gains in estimation when some endogenous variables (outcomes and/or covariates) are missing. Our findings are therefore highly relevant for applied researchers confronting missing endogenous variables in CMR models, and we conclude by offering the following practical recommendations:

(i) When specifying CMR models, researchers must distinguish not only between the endogenous and exogenous variables, but also whether the endogenous and exogenous variables are included in or excluded from the CMR model.

(ii) Auxiliary variables that are in the propensity score but are excluded from the CMR model must also be classified as endogenous or exogenous.

(iii) Efficiency gains in estimation from the observed sample occur if and only if there exist nonmissing endogenous variables (outcomes and/or covariates) that are included in or excluded from the CMR model.

(iv) Imputation should be based on all nonmissing variables (endogenous or exogenous) that are included in or excluded from the CMR model; i.e., in our notation, imputation should be based on $(X_{\text{in}}, X_{\text{ex}}, Z_{\text{in}}, Z_{\text{ex}})$. However, nonparametric imputation is informative, i.e., yields maximal efficiency gains in estimation, if and only if there exist nonmissing endogenous variables that are included in or excluded from the CMR model. That is, if and only if $Z_{\text{in}} \neq \vec{\varnothing}$ or $Z_{\text{ex}} \neq \vec{\varnothing}$. In particular, imputing missing outcomes in linear regression models is informative if and only if there exist nonmissing endogenous covariates that are included in or excluded from the regression.

(v) To achieve maximal efficiency gains in estimation from the observed sample, imputation must be nonparametric. If a parametric model is used for imputation, then efficient estimation is possible only if the imputation model is correctly specified.

Conflict of Interest Statement

The authors report there are no competing interests to declare.

Table 1: Estimated female labor earnings model with no missingness in labor income.

Variable	$\hat{\theta}_{\text{GMM}}$	$\text{se}_{\text{GMM}}(\hat{\theta}_{\text{GMM}})$	$\text{pvalue}(\hat{\theta}_{\text{GMM}})$	$\hat{\theta}$	$\text{se}_{\text{SEL}}(\hat{\theta})$	$\text{pvalue}(\hat{\theta})$
const.	−1.067	0.282	.00016	−0.899	0.258	.0005
agem1	0.459	0.018	3.3×10^{-143}	0.484	0.017	5.3×10^{-185}
agefstm	−0.313	0.026	2.0×10^{-34}	−0.347	0.023	1.8×10^{-50}
boy1st	0.040	0.041	.325	0.040	0.040	.31
morekids	−1.499	0.574	.009	−2.046	0.520	.00008
n					227,146	
Weak instruments F					651.7 (pvalue $= 7.8 \times 10^{-284}$)	
Endogeneity of *morekids* F					6.91 (pvalue = .009)	
Over-identification J					0.256 (pvalue = .613)	

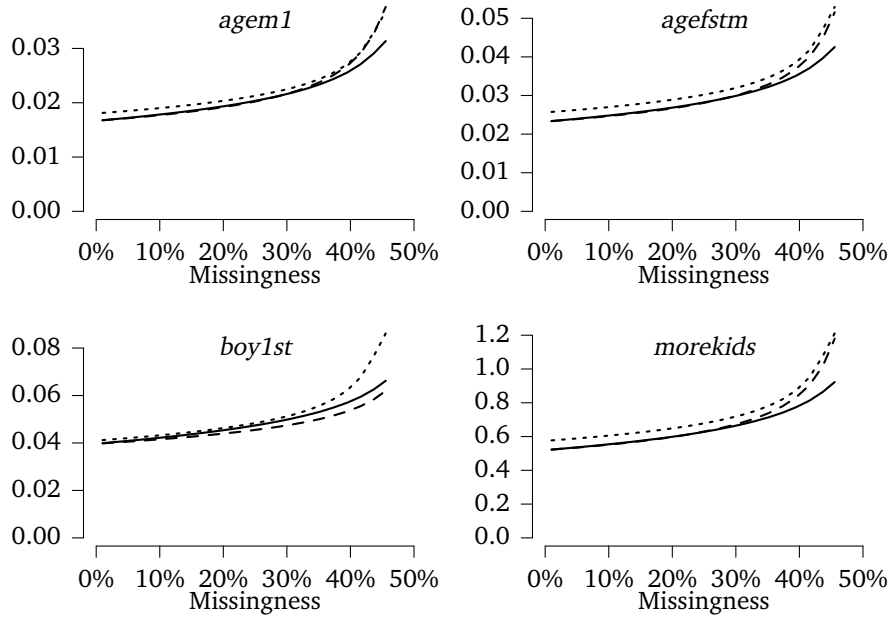
$\hat{\theta}$ is the SEL estimator of θ^* in the model $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0$ with bandwidth $b_n = 1.2$ (cf. Appendix B.4). $\hat{\theta}_{\text{GMM}}$ is the iterated GMM estimator of θ^* in the over-identified model $\mathbb{E}[\tilde{X} g] = 0_{6\times 1}$, and se_{GMM} are the GMM standard errors.

Table 2: Fraction of 1000 experiments (as a function of missingness in labor income) for which the t-test fails to reject the hypothesis that *morekids* is irrelevant.

Missingness in	size = 10%			size = 5%			size = 1%		
labor income	$\hat{\gamma}_{\text{GMM,IPW}}$	$\hat{\gamma}_{\text{SEL,IPW}}$	$\hat{\gamma}$	$\hat{\gamma}_{\text{GMM,IPW}}$	$\hat{\gamma}_{\text{SEL,IPW}}$	$\hat{\gamma}$	$\hat{\gamma}_{\text{GMM,IPW}}$	$\hat{\gamma}_{\text{SEL,IPW}}$	$\hat{\gamma}$
1%	.00	.00	.00	.00	.00	.00	.43	.00	.00
7%	.00	.00	.00	.01	.00	.00	.61	.00	.00
14%	.03	.00	.00	.12	.00	.00	.67	.01	.01
20%	.07	.00	.00	.23	.01	.00	.71	.10	.06
27%	.18	.01	.00	.34	.05	.02	.74	.27	.20
33%	.29	.11	.04	.46	.22	.09	.81	.48	.38
39%	.43	.35	.13	.61	.49	.24	.88	.76	.57
46%	.65	.72	.28	.77	.81	.43	.93	.93	.76

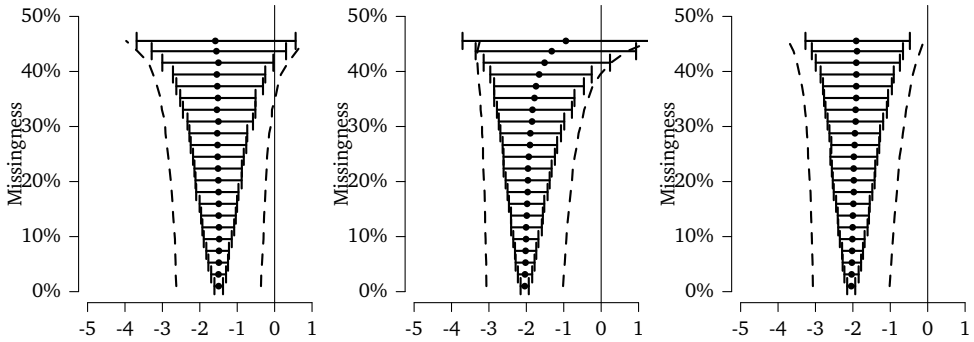
The t-test of $\gamma^* = 0$ vs. $\gamma^* \neq 0$ compares $|\frac{\hat{\gamma}_{\text{GMM,IPW}}}{\text{se}_{\text{GMM}}(\hat{\gamma}_{\text{GMM,IPW}})}|$, $|\frac{\hat{\gamma}_{\text{SEL,IPW}}}{\text{se}_{\text{SEL}}(\hat{\gamma}_{\text{SEL,IPW}})}|$, and $|\frac{\hat{\gamma}}{\text{se}_{\text{SEL}}(\hat{\gamma})}|$ with the two-sided critical values from the normal distribution.

Figure 1: Based on 1000 experiments, the standard errors of the estimated slope coefficients $(\beta^*, \gamma^*)_{4\times 1}$ as a function of missingness in labor income.



For each level of missingness: Dotted line is median of $\text{se}_{\text{GMM}}(\hat{\theta}^{(1)}_{\text{GMM,IPW}}), \ldots, \text{se}_{\text{GMM}}(\hat{\theta}^{(1000)}_{\text{GMM,IPW}})$. Dashed line is median of $\text{se}_{\text{SEL}}(\hat{\theta}^{(1)}_{\text{SEL,IPW}}), \ldots, \text{se}_{\text{SEL}}(\hat{\theta}^{(1000)}_{\text{SEL,IPW}})$. Solid line is median of $\text{se}_{\text{SEL}}(\hat{\theta}^{(1)}), \ldots, \text{se}_{\text{SEL}}(\hat{\theta}^{(1000)})$.

Figure 2: Based on 1000 experiments, three quantiles of $\gamma \in \{\hat{\gamma}_{\text{GMM,IPW}}, \hat{\gamma}_{\text{SEL,IPW}}, \hat{\gamma}\}$, and the medians of $\gamma \pm 1.96\,\text{se}(\gamma)$, as a function of missingness in labor income.



For each level of missingness: Dots are medians of the γ. Solid whiskers are .025 and .975 quantiles of the γ. Left dashed line is median of $\gamma - 1.96\,\text{se}(\gamma)$. Right dashed line is median of $\gamma + 1.96\,\text{se}(\gamma)$.

References

ABREVAYA, J., AND S. G. DONALD (2017): "A GMM approach for dealing with missing data on regressors," *Review of Economics and Statistics*, 99, 657–662.

AI, C., AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

——— (2012): "The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions," *Journal of Econometrics*, 170, 442–457.

ANGRIST, J. D., AND W. N. EVANS (1998): "Children and their parents' labor supply: Evidence from exogenous variation in family size," *American Economic Review*, 88, 450–477.

BALESTRA, S., AND U. BACKES-GELLNER (2017): "Heterogeneous returns to education over the wage distribution: Who profits the most?," *Labour Economics*, 44, 89–105.

BENNEDSEN, M., K. M. NIELSEN, F. PEREZ-GONZALEZ, AND D. WOLFENZON (2007): "Inside the family firm: The role of families in succession decisions and performance," *Quarterly Journal of Economics*, 122, 647–691.

CHEN, X., H. HONG, AND A. TAROZZI (2008): "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics*, 36, 343–366.

GRAHAM, B. S. (2011): "Efficiency bounds for missing data models with semiparametric restrictions," *Econometrica*, 79, 437–452.

HAHN, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

HRISTACHE, M., AND V. PATILEA (2017): "Conditional moment models with data missing at random," *Biometrika*, 104, 735–742.

——— (2021): "Equivalent models for observables under the assumption of missing at random," *Econometrics and Statistics*, 20, 153–165.

KITAMURA, Y., G. TRIPATHI, AND H. AHN (2004): "Empirical likelihood based inference in conditional moment restriction models," *Econometrica*, 72, 1667–1714.

KOSTYRKA, A. V. (2025): "smoothemplik: Smoothed empirical likelihood for efficient estimation and specification testing," `R` package version 0.0.17. Available at `https://cran.r-project.org/package=smoothemplik`.

MCDONOUGH, I. K., AND D. L. MILLIMET (2017): "Missing data, imputation, and endogeneity," *Journal of Econometrics*, 199, 141–155.

ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846–866.

STEPHENS JR., M., AND T. UNAYAMA (2019): "Estimating the impacts of program benefits: Using instrumental variables with underreported and imputed data," *Review of Economics and Statistics*, 101, 468–475.

Supplementary Material

Missing endogenous variables in conditional moment restriction models

Antonio Cosma, Andreï Victorovitch Kostyrka, and Gautam Tripathi

This supplement contains:

1. Details for SEL (Appendix A)
    - Why smooth the empirical likelihood? (Appendix A.1)
    - Implementing $\widehat{\mathrm{SEL}}_{\mathbb{T}}$ (Appendix A.2)
    - Comparing the SMD and SEL approaches (Appendix A.3)
2. Details for the empirical illustration (Appendix B)
    - Inducing missingness in labor income (Appendix B.1)
    - Additional results for higher levels of missingness (Appendix B.2)
    - Fast SEL implementation for large datasets (Appendix B.3)
3. Simulation study (Appendix C)
4. Proofs, examples, and technical details (Appendix D)

Notation and symbols not defined in the supplement are defined in the paper. Citations in the supplement not listed in its bibliography appear in the bibliography of the paper.

The title of our paper might suggest that it addresses a "2-sample" problem, where missing and nonmissing observations are available in two independent samples. However, this is not the case. The missing data problem we address is a "1-sample" problem, with all missing and nonmissing observations contained within a single sample. We thank Jinyong Hahn for bringing this to our attention.

A Details for smoothing the empirical likelihood

A.1 Why smooth the empirical likelihood?

The empirical likelihood is smoothed because we assume that

Assumption A.1. *Each coordinate of* (X, Z) *is continuously distributed with support* $\mathbb{R}$.

Assumption A.1 helps simplify the technical details when deriving the asymptotic distribution of our estimator. Each coordinate being continuously distributed simplifies the construction of the kernel estimators of π, μ, and having full support simplifies the "trimming" incorporated in the SEL objective function. Discrete coordinates of X and Z can be accommodated in the kernel estimators by using indicators in the kernel function. If all coordinates of X are discrete (as in Design 2 in the simulation study), then smoothing the empirical likelihood is not necessary and it can be shown that SEL with suitably redefined kernel weights coincides with unconditional empirical likelihood (Appendix C.6.3). The full support assumption means that the SEL objective function has to be trimmed to account only for the density of the conditioning variables becoming too small in the tails (the "denominator" problem). If some coordinates have bounded support, then additional trimming is needed to deal with the "boundary bias" of kernel estimators, which would complicate the technical details without enhancing the contribution of this paper.

A.2 Implementing $\widehat{\mathrm{SEL}}_{\mathbb{T}}$

The function $\theta \mapsto \widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta)$ is defined for those $\hat{\lambda}_i(\theta) \in \mathbb{R}^{\dim(\rho)}$ that satisfy (4.14). In practice, this constraint is incorporated in the objective function as follows. The function $\tilde{\lambda} \mapsto \sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij} \log(1+\tilde{\lambda}'\hat{\rho}_j(\theta))$ is concave on $\mathbb{R}^{\dim(\rho)}$ for large enough n. Thus, $\hat{\lambda}_i(\theta) = \operatorname{argmax}_{\tilde{\lambda}\in\mathbb{R}^{\dim(\rho)}} \sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij} \log(1+\tilde{\lambda}'\hat{\rho}_j(\theta))$, which follows by comparing its first order condition (FOC) with (4.14). Consequently,

$$\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) = -\sum_{i=1}^n \hat{\mathbb{T}}_{1i} \max_{\tilde{\lambda}_i \in \mathbb{R}^{\dim(\rho)}} \sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij} \log(1 + \tilde{\lambda}_i' \hat{\rho}_j(\theta)).$$

A.3 Comparing the SMD and SEL approaches

In the presence of missing endogenous variables, efficiency gains from the observed sample arise if and only if there exist nonmissing endogenous variables that are either included in or excluded from the structural moment function g in the CMR (2.1). In this case, estimation must be based on the CMR $\mathbb{E}[\rho \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ in the observed sample with $\rho \overset{(4.4)}{:=} \frac{D g_{\text{obs}}}{\pi} - \mu[\frac{D}{\pi} - 1]$, and μ must be estimated nonparametrically from the validation sample to attain the maximal efficiency gains. In contrast, if no such nonmissing endogenous variables exist, then $\mu \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}$, and estimation must instead rely on the CMR $\mathbb{E}[\frac{D g_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ in the validation sample. In both scenarios, it is advisable to use a semiparametrically efficient estimator, as this yields optimal inference regardless of whether efficiency gains are present. The semiparametrically efficient estimator need not be the SEL estimator; it may instead be the SMD estimator or another asymptotically equivalent estimator.[1]

[1] E.g., the optimal IV method of Newey (1993) delivers an asymptotically efficient estimator by solving a system of just-identified equations based on the semiparametrically efficient score, without requiring numerical optimization. Other approaches include those of Chen, Härdle, and Li (2003), Donald, Imbens, and Newey (2003), Smith (2007), Otsu (2011), Lavergne and Patilea (2013), and Chen, Pouzo, and Powell (2019), which combine empirical likelihood and SMD techniques to semiparametrically estimate CMR models with unknown functions.

As noted in Remark 4.2(iii), SMD and SEL approaches are closely related because Ai and Chen (2003, Footnote 4) note that kernel estimators can be used in the SMD approach. Furthermore, we show in Appendix D.3 — cf. the argument in Appendix D.3.3 leading to (D.47) — that the continuous updating version of the SMD objective function (Ai and Chen, 2003, Eqn. 23) based on kernel estimators is a quadratic approximation of the SEL objective function in large samples. Therefore, as we elaborate below, both the SEL and SMD approaches can be effectively employed for efficient estimation in our setting, though other methods can also be used.[2] In practice, applied researchers typically choose an estimation method based on its small-sample properties, the specifics of the empirical application, available computational resources, and their familiarity with the technique.

(i) As noted in Section 4.3, SEL-based inference using the empirical likelihood ratio (LR) statistic can be more accurate in small samples than SMD's Wald-based inference. The key reason is that the SEL estimator is a nonparametric maximum likelihood estimator (NPMLE) of θ^* under the constraint $\mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0$, whereas the continuous-updating SMD objective function is only a quadratic approximation to the SEL objective function. KTA (Section 5, Table III) study the size properties of the LR statistic by comparing it with Wald statistics based on several competing estimators. They find that the LR statistic's rejection probabilities remain close to the nominal level, while the Wald statistics exhibit substantial size distortions. Tripathi and Kitamura (2003) establish an optimality property of the LR statistic for specification testing in CMR models. Moreover, simulation studies in Owen (1990) and Kitamura (2001) show that the LR statistic has excellent power for testing parameter hypotheses in UCMR models, and Otsu (2010, Section 3.1) and Kitamura, Santos, and Shaikh (2012) prove optimality results for its Type I and Type II error probabilities. We conjecture that these results extend to the CMR framework. These findings provide strong theoretical support for SEL-based inference being more reliable than SMD-based inference in finite samples.

(ii) The primary computational difference between the SMD and SEL approaches can be described as follows:

- The SMD approach is a 2-step approach. To obtain an efficient SMD estimator in the 2nd step, we need a preliminary estimator of the inverse of the efficient variance-covariance matrix in the 1st step.
- The SEL approach is also 2-step. But the 1st step does not require a plug-in estimator of the efficient variance-covariance matrix as the SEL approach is internally studentized. Instead, the 1st step consists of a "inner-loop" optimization problem to obtain the vector of Lagrange multipliers enforcing the CMR in the sample (cf. Appendix A.2). The inner-loop optimization problem is a "low-dimensional" well-behaved convex optimization problem that is numerically straightforward to solve using well-defined numerical search (Kitamura, 2007, Section 8).[3] Furthermore, this numerical search is inherently parallelizable and has been fully parallelized in `smoothemplik` to leverage modern computing architectures.

Therefore, when comparing the computational cost of SMD and SEL, one must also account for the cost of obtaining a preliminary estimator of the inverse efficient variance-covariance

[2] Even though the SMD and SEL approaches are both very useful, it is important to develop and investigate complementary semiparametrically efficient procedures that may further improve on their existing properties. Brown and Newey (2002, p. 508) make the same point regarding GMM and empirical likelihood.

[3] E.g., in our empirical illustration, the inner-loop optimization is 1-dimensional and the Lagrange multiplier enforcing the 1-dimensional CMR in the observed sample is obtained rapidly by a simple line-search with guaranteed convergence.

matrix in the first step of the SMD procedure. Constructing such an estimator is not always straightforward. E.g., Ai and Chen (2003, Section 7) discuss difficulties related to the choice of sieve family and the number of terms in the sieve approximation. Moreover, as noted by Antoine, Bonnal, and Renault (2007, p. 474), using a preliminary estimator of the efficient variance-covariance matrix in the first step may affect the small-sample performance of the SMD estimator. Empirical evidence in Fisher, Hall, Jing, and Wood (1996, Section 4.1) further suggests that internal studentization can considerably improve the finite-sample behavior of test statistics. Although replacing the SEL criterion with a χ^2-distance *"... might lead to a modest saving of computational costs"* (Kitamura, 2007, p. 201), such a modification may also weaken the inferential advantages offered by the NPMLE, as discussed in (i). In short, it is not clear that one approach strictly dominates the other in terms of overall computational cost. For these reasons, we adopt the SEL approach in this paper, while remaining open to alternative methods in future work, depending on the specific problem context.

B Details for the empirical illustration

Remark B.1 (Hausman test of the hypothesis that $\hat{\gamma}_{\text{GMM}}$ and $\hat{\gamma}$ in Table 1 estimate the same parameter, against the alternative that they do not)**.** The Hausman test relies on $\text{asvar}(\hat{\gamma}_{\text{GMM}}-\hat{\gamma}) = \text{asvar}(\hat{\gamma}_{\text{GMM}}) - \text{asvar}(\hat{\gamma})$, which follows from the asymptotic efficiency of $\hat{\gamma}$. From Table 1, we have that $\hat{\gamma}_{\text{GMM}} = -1.499$, $\text{se}_{\text{GMM}}(\hat{\gamma}_{\text{GMM}}) = 0.574$, $\hat{\gamma} = -2.046$, and $\text{se}_{\text{SEL}}(\hat{\gamma}) = 0.520$. Hence, the absolute value of the t-statistic is 2.25 with pvalue $= 0.0244$, implying that $\hat{\gamma}_{\text{GMM}} - \hat{\gamma} \approx 0$ at significance levels $\leq 1\%$. Indeed, a pvalue of 2.4% in a sample of 227,146 observations — where even minimal differences should lead to rejection at all conventional significance levels — means that there is not enough evidence to convincingly claim that $\hat{\gamma}_{\text{GMM}}$ and $\hat{\gamma}$ are different. □

B.1 Inducing missingness in labor income

Recall that $Z = Z_{\text{in}}$ throughout the empirical illustration. We induce missingness in labor income through the "artificial" propensity score function[4]

$$\pi(Z, X, s) := \text{clamp}_{0.05,0.99}(0.99 - s \cdot \ell(Z, X_{\text{in}})),$$

where $s \geq 0$ controls the extent of missingness and

$$\ell(Z, X_{\text{in}}) := 0.15\, \textit{morekids} + 0.1\text{clamp}_{0,1}(\textit{agem1}) - 0.05\text{clamp}_{0,1}(\textit{agefstm}).$$

Given s, a missing outcome is created by drawing $R \overset{\text{d}}{=} \text{Unif}[0,1]$ and letting $D := \mathbb{1}(R < \pi(Z, X, s))$ and $Y := DY^* + (1-D)\mathfrak{m}$. E.g., $n^{-1}\sum_{i=1}^n \mathbb{1}(R < \pi(Z_i, X_i, 0)) = 0.99$, which implies that 1% of the outcomes are missing on average (in repeated experiments), and $n^{-1}\sum_{i=1}^n \mathbb{1}(R < \pi(Z_i, X_i, 4.2)) = 0.54$ implying 46% missingness on average.

Each (s, R) pair leads to a dataset — drawn randomly from the original AE dataset — containing missing outcomes with the extent of missingness determined by s. To ensure that our analysis is not influenced by a specific level of missingness in the labor income, or a specific draw from the AE dataset, we consider 22 levels of missingness ranging from 1% to 46%, and for each level of missingness we randomly draw 1000 datasets with missing outcomes. This is done by independently repeating the following experiment 1000 times:

[4]For $a \in [a_{\min}, a_{\max}]$, the "clamping" function $\text{clamp}_{l,u}(a) := l + (u-l)\dfrac{a - a_{\min}}{a_{\max} - a_{\min}}$ maps $a \mapsto [l, u]$ so that the propensity score is bounded away from zero and one.

1. Generate $R \overset{\text{d}}{=} \text{Unif}[0,1]$.

2. For $s \in \{0, 0.2, 0.4, \ldots, 4.2\}$, let $D := \mathbb{1}[R < \pi(Z,X,s)]$ and $Y := DY^* + (1-D)\mathfrak{m}$. This creates 22 datasets with missing outcomes — each dataset drawn randomly from the original AE dataset — with missingness ranging from 1% to 46%.

3. For each of the 22 datasets, compute:

   (i) $\hat{\theta}_{\text{GMM,IPW}}$ and $\text{se}_{\text{GMM}}(\hat{\theta}_{\text{GMM,IPW}})$;

   (ii) $\hat{\theta}_{\text{SEL,IPW}}$ and $\text{se}_{\text{SEL}}(\hat{\theta}_{\text{SEL,IPW}})$;

   (iii) $\hat{\theta}$ and $\text{se}_{\text{SEL}}(\hat{\theta})$.

For each level of missingness ranging from 1% to 46%, these 1000 experiments yield 1000 i.i.d. copies of $(\hat{\theta}_{\text{GMM,IPW}}, \text{se}_{\text{GMM}}(\hat{\theta}_{\text{GMM,IPW}}))$, $(\hat{\theta}_{\text{SEL,IPW}}, \text{se}_{\text{SEL}}(\hat{\theta}_{\text{SEL,IPW}}))$, and $(\hat{\theta}, \text{se}_{\text{SEL}}(\hat{\theta}))$.

B.2 Additional results for higher levels of missingness

In this section we report some additional results for the estimated marginal effect of *morekids* on labor income for levels of missingness in labor income up to 70%.[5] Based on 1000 simulation experiments, we report the following results for the semiparametrically efficient SEL estimator $\hat{\gamma}$ and the GMM estimator $\hat{\gamma}_{\text{GMM,IPW}}$:

(i) The MSE of $\hat{\gamma}$ and $\hat{\gamma}_{\text{GMM,IPW}}$ averaged across the simulations are reported in Table B.1. To compute the MSE, we take as the true marginal effect the value reported in Table 1 in the paper when there is no missingness in labor income. Specifically, the "true value" for $\hat{\gamma}_{\text{GMM,IPW}}$ is the value of $\hat{\gamma}_{\text{GMM,IPW}}$ without missingness (-1.499 as reported in Table 1); the "true value" for $\hat{\gamma}$ is the value of $\hat{\gamma}$ without missingness (-2.046 in Table 1).

(ii) The average lengths of the 95% confidence intervals (CIs) based on $\hat{\gamma}$ and $\hat{\gamma}_{\text{GMM,IPW}}$, as a function of the missingness in labor income, are displayed in Figure B.1. For each level of missingness, the average CI length was obtained by finding the medians of the upper and lower endpoints of the 95% CIs across the 1000 simulations, and taking their difference.

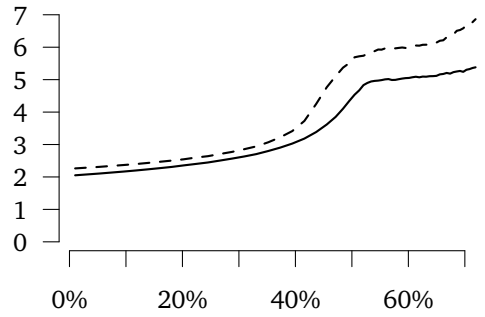
Table 1 reveals that the results for missingness between 50%–70% are in line with those reported in the paper for missingness up to 50%. As the missingness in labor income increases, both $\hat{\gamma}$ and $\hat{\gamma}_{\text{GMM,IPW}}$ get less accurate. However, the MSE of $\hat{\gamma}_{\text{GMM,IPW}}$ is always greater than the MSE of $\hat{\gamma}$, so that even at 70% missingness in labor income the MSE of $\hat{\gamma}_{\text{GMM,IPW}}$ is 62.6% more than the MSE of $\hat{\gamma}$, implying that $\hat{\gamma}$ remains more accurate than $\hat{\gamma}_{\text{GMM,IPW}}$ even when there is 70% missingness in labor income. This is also reflected in the average lengths of the 95% CIs for the two estimators displayed in Figure B.1: For each level of missingness in the labor income, the 95% CI based on $\hat{\gamma}_{\text{GMM,IPW}}$ is longer than the 95% CI based on $\hat{\gamma}$. At 70% missingness the average 95% CI based on $\hat{\gamma}_{\text{GMM,IPW}}$ is about 20% longer than the 95% CI based on $\hat{\gamma}$, suggesting that even at this level of missingness tests of hypothesis based on $\hat{\gamma}$ have higher power than those based on $\hat{\gamma}_{\text{GMM,IPW}}$.

[5] Numerical findings not reported here suggest that $\text{se}_{\text{SEL}}(\hat{\gamma})$ is more robust to the choice of c_n than $\text{se}_{\text{SEL}}(\hat{\gamma}_{\text{SEL,IPW}})$, which can be explained by doubly robust estimators being free from the influence of estimating π. The practical implication is that applied researchers should use $\hat{\gamma}$ not only when missingness is high (because of its smaller standard errors), but also for lower levels of missingness (because of its robustness with respect to c_n).

Table B.1: MSE of $\hat{\gamma}$ and $\hat{\gamma}_{\text{GMM,IPW}}$ for levels of missingness up to 70%.

Missingness in labor income		25%	40%	55%	70%
MSE	$\hat{\gamma}_{\text{GMM,IPW}}$	0.123	0.408	1.948	2.626
	$\hat{\gamma}$	0.107	0.290	1.460	1.615
Gains	$\frac{\text{MSE}(\hat{\gamma}_{\text{GMM,IPW}})}{\text{MSE}(\hat{\gamma})} - 1$	14.9%	40.6%	33.4%	62.6%

Figure B.1: Average length of 95% CI for $\hat{\gamma}$ (solid) and $\hat{\gamma}_{\text{GMM,IPW}}$ (dashed) as a function of the missingness in labor income.



B.3 Fast SEL implementation for large datasets

Even though the sample size in the AE dataset is $n = 227{,}146$ and the SEL approach requires nonparametric smoothing, implementing the efficient estimator and obtaining the SEL standard errors (Remark 4.2(i)) is relatively fast. Using 4$^{\text{th}}$-order-accurate numerical gradients for Hessian estimation on a single thread of an AMD Epyc ROME 7H12 2.6-GHz processor takes only 6–12 minutes on average. With full parallelization, one SEL evaluation takes 0.8 seconds, its numerical gradient takes 3.3 seconds, and the entire workflow with estimation and standard error computation terminates successfully in under 2 minutes.

In practice, given the enormous sample size in this application, numerical routines are much faster and memory-efficient if very small SEL kernel weights w_{ij} are replaced by 0. A viable alternative is using bounded-support kernels. Therefore, in this application, we use the Bartlett kernel $k_b(x) := (1-|x|/b)\mathbb{1}(|x|/b \le 1)$ with bandwidth $b_n = 1.2$ to create a sparse smoothing matrix $[w_{ij}] := [K_{ij}/\sum_{k=1}^{n} K_{ik}]$, where the product kernel

$$K_{ij} := k_b(agem1_i - agem1_j)k_b(agefstm_i - agefstm_j) \\ \mathbb{1}(boy1st_i = boy1st_j)\mathbb{1}(boys2_i = boys2_j)\mathbb{1}(girls2_i = girls2_j).$$

Choice of bandwidth is discussed in Appendix B.4.

The following points highlight how computation time can be reduced:

- If an observation (Y_j, X_j, Z_j) appears in the data set more than once ($m_j > 1$ times), then, following Owen (2017), the search for $\hat{\lambda}_i(\theta)$ in the "inner loop" of the SEL objective function (Appendix A.2) can be simplified by counting all full duplicates, discarding the repetitions, and replacing the corresponding SEL weight w_{ij} with $m_j w_{ij}$.

- If the conditioning variable values coincide for two observations ($X_i = X_j$), then $\hat{\lambda}_i(\theta) = \hat{\lambda}_j(\theta)$, and the extra weighted EL maximization can be skipped.

- Since conditioning on discrete variables is achieved through product kernels that are non-zero only if the values taken by all discrete variables are identical, the smoothing matrix

$[w_{ij}]$ is a block matrix that can be stored as a sparse matrix or in a list (the latter enables parallel search for $\hat{\lambda}_i(\theta)$ in the blocks).

In the AE dataset, each observation (Y_j, X_j, Z_j) has average multiplicity $m_j \approx 2.9$. After these duplicates are accounted for in w_{ij}, each unique exogenous X_i has average multiplicity ≈ 108. The discrete instruments partition the $[w_{ij}]$ matrix into 4 blocks, each block being ≈95% sparse. This brings down the total number of arithmetic operations per one SEL evaluation by a factor of $\approx$ 81,000 (almost hundred thousand times).

Obtaining $\hat{\theta}$ is only 1.9 times slower that $\hat{\theta}_{\text{SEL,IPW}}$; this ratio is remarkably stable for all considered levels of missingness. For higher levels of missingness, finding $\hat{\theta}_{\text{SEL,IPW}}$ may even take longer due to the deteriorating behaviour of the SEL objective function based on $\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$.

Since the SEL function based on (4.5) depends on the nonparametrically estimated $\hat{\mu}_d(Z_j, X_j, \theta)$ that has to be re-calculated for each new SEL evaluation (during maximization and numerical differentiation), computation of $\hat{\mu}_d(Z_j, X_j, \theta)$ is done via optimized `C++` routines through the `Rcpp` interface and `RcppArmadillo` functions (Eddelbuettel and François, 2011; Eddelbuettel and Sanderson, 2014).

B.4 Bandwidth choice

We use the shrinkage factor 1/3 as an ad hoc, computationally simple method of reducing the bias of the kernel estimators of π and μ (Calonico, Cattaneo, and Farrell, 2018, discuss this issue comprehensively). The bandwidths c_n and d_n are obtained by cross-validation; the median $(\hat{c}_n^{\text{CV}}, \hat{d}_n^{\text{CV}})$ is (4.0, 4.3).

We estimated the model with SEL for various bandwidths b_n from 1.1 to 2.5, and the point estimates of all coefficients and SEL standard errors remained virtually unchanged (±4%). Any $b_n \leq 1$ with a finite-support kernel in this application implies no smoothing, i.e., conditioning on all unique combinations of discrete instrument values. This results in a highly fragmented conditioning set and violation of the spanning condition in (4.9) for certain combinations of exogenous variable values. The efficient estimator is also pretty robust to the smoothing bandwidths c_n and d_n. Under 40% missingness, keeping $d_n = 2.2$ and varying $c_n \in \{1.3, 2.0, 2.5, 4.0\}$, we observed little (±8%) relative variation of $\hat{\theta}$. More notably, keeping $c_n = 2.5$ and varying $d_n \in \{1.5, 2.3, 3.0\}$, we observed changes only in the 3rd or 4th significant digit of $\hat{\theta}$.

C Simulation study

In the simulation study, all included regressors are endogenous and there are no excluded endogenous variables ($Z_{\text{ex}} = \vec{\varnothing}$) so that $Z = Z_{\text{in}}$, and the only instruments are excluded ($X_{\text{in}} = \vec{\varnothing}$) so that $X = X_{\text{ex}}$. We compare the small sample behavior of $\hat{\theta}$ with the estimator constructed using only the validation sample when, in addition to MAR, it is assumed that $\pi(Z,X)$ does not depend on Z, i.e., $\pi(Z,X) \overset{P_{Z,X}\text{-a.s.}}{=} \tilde{\pi}(X)$ for some $\tilde{\pi}(X) \in (0,1)$. This condition, which is equivalent to assuming that $D \perp\!\!\!\perp Z \mid X$, is also maintained in HP (p. 736), cf. their discussion of the partially linear single-index model. It is sometimes implicitly used to justify estimation in the validation sample without inverse propensity score weighting because $\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \overset{\pi(Z,X)=\tilde{\pi}(X)}{\iff} \mathbb{E}[Dg_{\text{obs}} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$.[6] MAR and $\pi(Z,X) \overset{P_{Z,X}\text{-a.s.}}{=} \tilde{\pi}(X)$ imply that $\mathscr{J}_{\varpi_*} = J$ and $\Sigma = \Omega_g/\tilde{\pi}$, where $\mathscr{J}_{\varpi_*}$ is defined in the proof of (4.6) and $\Omega_g := \mathbb{E}[gg' \mid X]$. Hence, $\text{l.b.}_{\text{VS}}(\theta^*) \overset{\text{(D.25)}}{=} (\mathbb{E}\tilde{\pi}J'\Omega_g^{-1}J)^{-1}$. In the simulation study, the SEL estimator of θ^* based on the moment condition $\mathbb{E}[Dg_{\text{obs}} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ is denoted by $\hat{\theta}_{\text{VS}}$.[7] Note that even though the propensity score depends only on X in the simulation study, the presence of nonmissing included endogenous variables in the structural model ensures that efficiency gains still occur because imputation of g is done using both Z and X so that $\mu \neq 0$, i.e., imputation is informative (cf. Lemma 4.2).

C.1 Designs

We consider two designs with the same structural model, namely, a simplified version of the linear IV regression in Example 2.1 given by $Y^* := \alpha^* + \gamma^* Z + U\sigma(X)$, where the outcome Y^* is missing for some individuals, the single regressor Z is endogenous, and X is the sole excluded IV for Z, i.e., X satisfies $\mathbb{E}[U \mid X] = 0$ P_X-a.s. The difference between the designs is in how Z and X are modeled. In Design 1, Z and X are both continuously distributed, and the reduced form equation for Z is given by $Z := \zeta_0 + \zeta_1 X + V$. In contrast, in Design 2, Z and X are both dummy variables, and the reduced form equation for Z is given by $Z := \mathbb{1}(\zeta_0 + \zeta_1 X + V > 0)$.[8] In both designs, $\left[\begin{smallmatrix} U \\ V \end{smallmatrix}\right] \Big| X \overset{\text{d}}{=} \text{N}(\left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{smallmatrix}\right])$ with $\sigma_U^2 = 1$, $\sigma_V^2 = 2$, $\sigma_{UV} = 1$, and $\alpha^* = \gamma^* = \zeta_1 = 1$. The reduced form intercept ζ_0 differs across the designs — $\zeta_0 = 1$ in Design 1, and $\zeta_0 = 0$ in Design 2 — to ensure that $\mathbb{E}Z$ is close to $\mathbb{E}X$.

Throughout this section, Φ denotes the cumulative distribution function, and ϕ the probability density function, of a $\text{N}(0,1)$ random variable. The results reported in this section are based on 5000 Monte Carlo replications, and $n = 500, 1000, 2000, 4000$, corresponding to "relatively small," "small," "medium," and "large" sample sizes. We call $n = 500$ to be a relatively small sample because it includes the missing observations. Indeed, if $n = 500$ then a validation sample of approx. 200 or fewer observations in some draws can be reasonably considered to be relatively small in the semiparametric context.

[6] Similar assumptions can also be made to reduce the dimensionality of the propensity score. E.g., if $D \perp\!\!\!\perp X \mid Z$ is assumed in addition to MAR, then π only contains Z; and if $D \perp\!\!\!\perp (Z,X)$ is assumed in addition to MAR, then Y^* is MCAR and π contains neither Z nor X, i.e., π is a constant equal to π_{MCAR} (cf. Example 4.5). Efficiency bounds for these special cases follow from Lemma 4.1 by simply replacing $\pi(Z,X)$ with $\tilde{\pi}(Z)$ and π_{MCAR}, respectively.

[7] The condition that $\pi(Z,X)$ does not depend on Z is imposed only in the simulation study and is used nowhere else in the paper. In particular, in the empirical illustration in Section 5, the SEL estimator of θ^* from the validation sample, denoted by $\hat{\theta}_{\text{SEL,IPW}}$, is based on the CMR $\mathbb{E}[\frac{Dg(Y,Z,X,\theta^*)}{\pi(Z,X)} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$.

[8] Designs where all regressors and instruments are discrete are not uncommon in microeconometric applications. As in RRZ (Section 2.5), efficiency gains in these designs are much more apparent because no smoothing is required to implement the efficient estimator.

C.1.1 Design 1

In this design, $X \overset{\text{d}}{=} \text{Unif}[0,1]$. The skedastic function $\sigma^2(x) := |x-r|^\nu + 1/15$, with $r = -1/3$ and $\nu = 2$, due to Cragg (1983), is popular with researchers to model conditional heteroskedasticity (and is used by us as well). The parameter ν determines the degree of heteroskedasticity (conditional homoskedasticity follows if $\nu = 0$). The regressor Z can be classified as being strongly endogenous in the heteroskedastic case because $\text{corr}(Z, U\sigma(X)) \approx 0.66$ when $\nu = 2$. This poses a serious problem because the bias of the slope coefficient, relative to its true value, when Y is regressed on Z in the validation sample is $\approx 42.1\%$ when averaged across the simulations. There is no issue with weak IV because $\text{corr}(Z,X) \approx 0.20$ and in those Monte Carlo replications where the first-stage F-statistics are < 10 new data were re-generated until the first-stage F-statistics became ≥ 10 ($\approx 57\%$ of all replications for $n = 500$ and 0.4% for $n = 2000$). The nonmissingness indicator D is drawn from a Bernoulli distribution with success probability $\tilde{\pi}(X) := l + (u-l)\Phi((X - r_{\tilde{\pi}})/\sigma_{\tilde{\pi}})$, where $l = 0.25$, $u = 0.95$, $r_{\tilde{\pi}} = 0.1$, and $\sigma_{\tilde{\pi}} = 0.5$, are chosen to make $\tilde{\pi}$ be bounded away from 0 and 1. As explained below, the parameter $r_{\tilde{\pi}}$, which controls the horizontal shift of the propensity score function, turns out to be more important than the degree of heteroskedasticity in determining the maximum efficiency gain — as measured by the variance of an efficient estimator based only on the subsample with no missing observations divided by the variance of our semiparametrically efficient estimator based on the observed sample, namely, the ratio $\dfrac{\text{l.b.}_{\text{VS}}(\gamma^*)}{\text{l.b.}(\gamma^*)} \overset{(4.6)}{>} 1$ — that this simulation design can deliver.

Figure C.1 in Appendix C.5 plots $\text{l.b.}_{\text{VS}}(\gamma^*)/\text{l.b.}(\gamma^*)$ as a function of the propensity score shift and the heteroskedasticity parameter.[9] It can be seen from Figure C.1 that the shift of the propensity score function determines how many values of Y^* are lost in the sample. The percentage of nonmissing observations as a function of $r_{\tilde{\pi}}$ is shown in black. If $r_{\tilde{\pi}}$ is too large or too small, then $\tilde{\pi}$ becomes almost constant on the support of X, which resembles missing completely at random instead of MAR. In the simulations $r_{\tilde{\pi}} = 0.1$, which yields a retention rate of $\approx 42\%$ (i.e., in $\approx 58\%$ of observations the outcome Y^* is missing). The degree of heteroskedasticity does not appear to have a major impact on the efficiency gains, which is not surprising because $\text{l.b.}(\gamma^*)$ and $\text{l.b.}_{\text{VS}}(\gamma^*)$ are both robust to the form of the skedastic function. Indeed, comparing $\nu = 0$ (conditional homoskedasticity) with $\nu = 2$, Figure C.1 reveals that the maximum efficiency gains are roughly the same ($\approx 42\%$). Therefore, for both Design 1 and Design 2 (described next), we generate data — and report simulation results — only for heteroskedastic errors since that is the empirically relevant case.

C.1.2 Design 2

In this design, $\Pr(X = 1) = 0.6$. Consequently, (Z,X) takes as values the vertices of the unit rectangle $\{(0,0),(0,1),(1,0),(1,1)\}$ with probability $(0.200, 0.144, 0.200, 0.456)$. The nonmissingness indicator D is drawn from a Bernoulli distribution with success probability $\tilde{\pi}(X) := 0.9X + 0.25(1-X)$, which yields an average retention rate of $\mathbb{E}D = 64\%$. With this choice of $\tilde{\pi}$, observations corresponding to $X = 0$ are more likely to be missing than observations corresponding to $X = 1$. The skedastic function $\sigma^2(X) := X + 16(1-X)$ creates higher dispersion, hence, more uncertainty, when there is less data, which strengthens the case for using the efficient estimator. The maximum efficiency gain $\text{l.b.}_{\text{VS}}(\gamma^*)/\text{l.b.}(\gamma^*)$ that Design 2 can deliver is $\approx 31\%$. Compared to Design 1, endogeneity of Z is even more of a problem in Design 2 because the relative bias of the slope coefficient when Y is regressed on Z in the validation sample is $\approx 179\%$ when averaged across the simulations. In this design, X is not a weak instrument because $\text{corr}(Z,X) \approx 0.27$ and the average first-stage F statistic is ≈ 16.3 when $n = 500$.

[9] In both designs, the ratio $\text{l.b.}_{\text{VS}}(\gamma^*)/\text{l.b.}(\gamma^*)$ was obtained by numerical integration on the simplified expressions given in Appendix C.6.

C.2 Implementation

The SEL estimator $\hat{\theta} := (\hat{\alpha}, \hat{\gamma})$ is implemented by maximizing the SEL objective function with $\hat{\mathbb{T}}_{1i} := 1$ and $\hat{\mathbb{T}}_{2j} := 1$ (Appendix A.2). Similarly, $\hat{\theta}_{\text{VS}} := (\hat{\alpha}_{\text{VS}}, \hat{\gamma}_{\text{VS}})$ is implemented by maximizing the SEL objective function with $\hat{\mathbb{T}}_{1i} := 1$, $\hat{\mathbb{T}}_{2j} := 1$, and $\hat{\rho} := Dg$, where $g := Y^* - \alpha^* - \gamma^* Z$. The LR confidence region for the slope coefficient is obtained by treating the intercept as a nuisance parameter. Specifically, let $\hat{\alpha}(\gamma) := \text{argmax}_{\alpha \in \mathbb{R}} \widehat{\text{SEL}}_{\mathbb{T}}(\alpha, \gamma)$ and denote by $\text{LR}^{\text{p}}(\gamma) := 2[\widehat{\text{SEL}}_{\mathbb{T}}(\hat{\alpha}, \hat{\gamma}) - \widehat{\text{SEL}}_{\mathbb{T}}(\hat{\alpha}(\gamma), \gamma)]$ the profile LR statistic obtained by concentrating out the intercept. Then, the lower level set $\{\gamma \in \mathbb{R} : \text{LR}^{\text{p}}(\gamma) \leq k_\tau\}$ is a $(1-\tau)100\%$ confidence region for γ^*. Whether this confidence region is an interval or not depends on the shape of $\gamma \mapsto \text{LR}^{\text{p}}(\gamma)$. If $\gamma \mapsto \text{LR}^{\text{p}}(\gamma)$ is quasiconvex, which appears to be the case in our simulation study because in both designs $\gamma \mapsto \widehat{\text{SEL}}_{\mathbb{T}}(\hat{\alpha}(\gamma), \gamma)$ seems close to being concave for the sample sizes we consider (Figures C.5 and C.7), then the confidence region is an interval.[10] The endpoints of the LR CI are obtained by numerically finding the roots of the equation $\text{LR}^{\text{p}}(\gamma) = k_\tau$ using Brent's method (the initializing points for the root-finding algorithm are chosen to be the endpoints of the Wald CI). The same approach is used to obtain the LR CI based on $\hat{\theta}_{\text{VS}}$.

C.2.1 Design 1

Here, w_{ij} is constructed using Gaussian kernels, and $\hat{\tilde{\pi}}_c, \hat{\mu}_d$ are Nadaraya-Watson estimators with bandwidths c, d and Gaussian kernels. Before constructing $\hat{\tilde{\pi}}_{c_n}$ and $\hat{\mu}_{d_n}$, an injective transformation is applied to map distinct elements of $(Z_1, \ldots, Z_n)$ and $(X_1, \ldots, X_n)$ into the interval $(0, 1)$ such that the transformed observations become more equispaced and do not fall into the boundary region. This procedure, motivated by the discussion in Hall (1990, Section 3), is helpful in dealing with the bandwidth and edge effects issues; e.g., equispacing the observations is a simple device for improving the performance of kernel estimators of conditional expectation functions because the bandwidth does not have to be adaptive if the observations on the conditioning variables are relatively equispaced. It is described in detail in Appendix C.5 as it may be useful to other applied researchers.

To the best of our knowledge, how to choose an optimal data-driven bandwidth when smoothing the empirical likelihood remains an open problem. Our goal in this simulation study is to show that, with appropriately chosen bandwidths, there is room for substantial efficiency gains. To quantify these gains, we acted as the oracle to choose the optimal b_n by repeating the simulation experiment on a grid of b_n and picking the bandwidth that minimized the average (across the simulation replications) RMSE of the estimator of γ^*. In particular, since b_n is the only bandwidth required to smooth the empirical likelihood for implementing $\hat{\gamma}_{\text{VS}}$, for every sample size, we estimated $\hat{\gamma}_{\text{VS}}$ on a coarse grid of bandwidths, and the oracle SEL bandwidth b_n^* was chosen to minimise the RMSE of $\hat{\gamma}_{\text{VS}}$. The bandwidth b_n^* was also used to implement the efficient estimator $\hat{\gamma}$. With this optimal b_n^*, we chose (c_n^*, d_n^*) via least-squares cross-validation on each individual simulated data set to implement $\tilde{\pi}_c$ and μ_d. The oracle bandwidth b_n^*, and the median of the cross-validated bandwidths (c_n^*, d_n^*), are reported in Table C.1, which contains the summary statistics for the estimated slope coefficients $\hat{\gamma}$ and $\hat{\gamma}_{\text{VS}}$ averaged across the simulations. The manner in which (b_n^*, c_n^*, d_n^*) were chosen highlights the following points: (i) Substantial efficiency gains are possible if the bandwidths are chosen appropriately; (ii) the gains in efficiency are not too sensitive to the choice of bandwidth; (iii) the bandwidths for estimating the propensity score $\tilde{\pi}$ and the function μ required for nonparametric imputation can be chosen by cross-validation.[11]

[10] Quasiconvexity of $\gamma \mapsto \text{LR}^{\text{p}}(\gamma)$ implies that its lower level sets are convex. Since convex sets are connected, and the only connected sets in $\mathbb{R}$ are intervals, it follows that if $\gamma \mapsto \text{LR}^{\text{p}}(\gamma)$ is quasiconvex then its lower level sets are intervals.

[11] In a separate set of simulations, we also acted as the oracle for choosing (c_n, d_n) along with b_n. The efficiency gains in these simulations were marginally higher, e.g., 6% instead of 1% for $n = 500$, and 49% instead of 44% for

C.2.2 Design 2

If $X \in \{0,1\}$, then $\mathbb{E}[Y^* - \alpha^* - \gamma^* Z \mid X] \overset{P_X\text{-a.s.}}{=} 0 \iff \mathbb{E}\tilde{X}[Y^* - \alpha^* - \gamma^* Z] = 0$, where $\tilde{X} := [\frac{1}{X}]$; i.e., discreteness of the conditioning variable exactly identifies θ^*. Hence, as shown in Appendix C.6.3, if $\hat{\theta}$ solves $\sum_{j=1}^n \tilde{X}_j \hat{\rho}(\mathscr{A}_j, \hat{\theta}) = 0$, then it also maximizes the SEL objective function with weights $w_{ij} := \mathbb{1}(X_i = X_j)/\sum_{k=1}^n \mathbb{1}(X_i = X_k)$. The same argument reveals that $\hat{\theta}_{\text{VS}} = (\sum_{j=1}^n D_j \tilde{X}_j \tilde{Z}_j')^{-1} \sum_{j=1}^n D_j \tilde{X}_j Y_j$ is the IV estimator obtained using the validation sample.

C.3 Results and discussion

We now describe the main findings of our simulation study, which follow our theoretical results fairly closely.

C.3.1 Design 1

The distribution of the estimators appears to be centred around the true value, and is close to being normal (Figure C.2). Since the mean and median biases are close to zero (Table C.1), the efficiency gains (whether measured by the ratio of the Monte Carlo variances, or the ratio of the Monte Carlo mean squared errors) range from about 1.3% (when $n = 500$) to about 45% (when $n = 4000$). In comparison, as noted in Section C.1.1, the maximum efficiency gain the simulation design can deliver is about 42%. Figures C.3 and C.4 show that the RMSE of $\hat{\gamma}$ is relatively insensitive to the bandwidths b_n and (c_n, d_n) over a large enough interval.

Table C.2 contains the coverage probabilities for LR CIs and their median lengths (when the intervals are bounded) in the Monte Carlo replications. This table emphasizes the following key findings. Firstly, for small sample sizes, the LR CIs can be unbounded in one direction (Figure C.5). E.g., the last column of Table C.2 shows that when $n = 500$ and nominal coverage probability is 90%, the LR CIs based on $\hat{\gamma}_{\text{VS}}$ are unbounded in 6.1% of the Monte Carlo replications, and those based on $\hat{\gamma}$, only in 0.4%. For samples of size 1000 or more, the fraction of unbounded CIs was less than 0.1% for all confidence levels, whilst 0.9% of the intervals (with nominal coverage $= 90\%$) based on $\hat{\gamma}_{\text{VS}}$ were unbounded. Secondly, although their coverage probabilities are close to nominal, the LR CIs based on $\hat{\gamma}$ are much shorter than those based on $\hat{\gamma}_{\text{VS}}$. The difference in the lengths of the CIs is clear evidence of the efficiency gains from $\hat{\gamma}$. In large samples, the ratio of their lengths is close to the square root of relative efficiency gains, as it should be, and in small samples, the gains are even larger.

C.3.2 Design 2

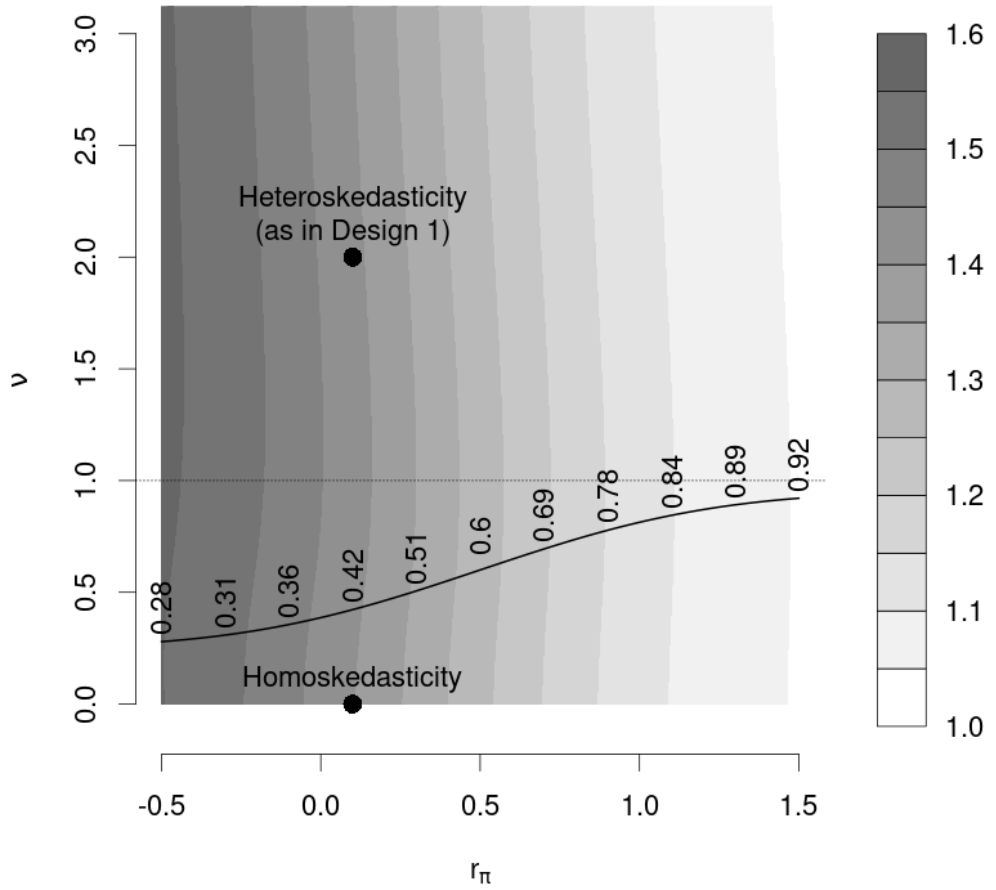
The simulation results for Design 2 are summarized in Table C.3. The increase in $\text{MSE}(\hat{\gamma}_{\text{VS}})$ compared to the $\text{MSE}(\hat{\gamma})$, i.e., $[\text{MSE}(\hat{\gamma}_{\text{VS}}) - \text{MSE}(\hat{\gamma})]/\text{MSE}(\hat{\gamma})$, can be very large for small sample sizes, e.g., 533% when $n = 500$. This, however, is a sample size effect reflecting how $\mathbb{E}\tilde{X}\tilde{Z}'$, required in the implementation of $\hat{\theta}_{\text{VS}}$, is estimated. Indeed, in simulation results not reported here, we replaced $\hat{\theta}_{\text{VS}}$ with $(\sum_{j=1}^n \tilde{X}_j \tilde{Z}_j')^{-1} \sum_{j=1}^n D_j \tilde{X}_j Y_j$, which estimates $\mathbb{E}\tilde{X}\tilde{Z}'$ using the entire sample (because Z and X are never missing), and found that this led to significant improvement in the performance of $\hat{\gamma}_{\text{VS}}$, namely, its average bias (resp. standard deviation) reduced by more than $1/4^{\text{th}}$ (resp. $1/2$) when $n = 500$. The efficiency gains stabilize as the sample size increases. For $n = 4000$ they are approximately 39%, which is close to the maximum that Design 2 can deliver. The smoothed densities of $\hat{\gamma} - \gamma^*$ and $\hat{\gamma}_{\text{VS}} - \gamma^*$ are in Figure C.6. Both estimators appear to be Gaussian, with a larger dispersion for $\hat{\gamma}_{\text{VS}}$ as expected. In small samples, the efficiency gains for Design 2 are higher than those for Design 1 because, unlike Design 1, no nonparametric smoothing is required in Design 2.

$n = 2000$. However, we only report the results for the cross-validated bandwidths (c_n, d_n).

Table C.4 contains the coverage probabilities for LR CIs, and their lengths (when the intervals are bounded), averaged across the Monte Carlo replications. As with Design 1, we find that: (i) For small sample sizes, the LR CIs based on $\hat{\gamma}_{\text{VS}}$ can be unbounded in one direction (Figure C.7). E.g., the last column of Table C.4 reveals that when $n = 500$ and nominal coverage probability is 90%, the LR CIs based on $\hat{\gamma}_{\text{VS}}$ are unbounded in 3.1% of the Monte Carlo replications. In contrast, the LR CIs based on $\hat{\gamma}$ are bounded even when $n = 500$ (except in one simulation sample when nominal coverage probability is 99%). (ii) The LR CIs based on $\hat{\gamma}$ are much shorter than those based on $\hat{\gamma}_{\text{VS}}$. Moreover, the empirical coverage probabilities are very close to nominal owing to the fact that both estimators are empirical-likelihood-based. For small sample sizes, the high accuracy of the empirical coverage probability in Design 2 is due to the absence of any nonparametric smoothing, whereas the slightly more conservative behaviour of CIs in Design 1 is likely caused by nonparametric smoothing and the fact that the estimation-optimal bandwidths used to implement the CIs need not be testing-optimal.

C.4 Tables and figures for the simulation study

Figure C.1: Heat map of l.b.$_{\text{VS}}(\gamma^*)$/l.b.(γ^*) as a function the propensity score shift ($r_{\tilde{\pi}}$) and the degree of heteroskedasticity (ν) in Design 1.



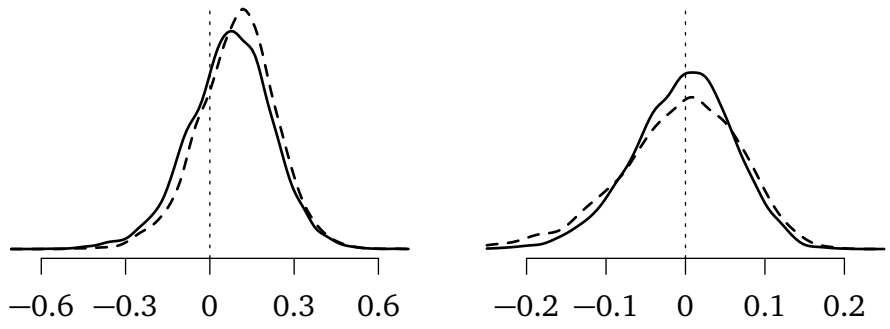
The darker the shade, the larger the efficiency gain l.b.$_{\text{VS}}(\gamma^*)$/l.b.(γ^*). The solid line and the numbers above it show the proportion of nonmissing observations.

Table C.1: Simulation summary for the estimated γ^* in Design 1.

n	Est.	b_n^*	c_n^*	d_n^*	Med. Bias	Mean Bias	Std. Dev.	$\frac{\text{Med. AD}(\cdot)}{\text{Med. AD}(\hat{\gamma})}$	$\frac{\text{MeanAD}(\cdot)}{\text{MeanAD}(\hat{\gamma})}$	$\frac{\text{var}(\cdot)}{\text{var}(\hat{\gamma})}$	$\frac{\text{MSE}(\cdot)}{\text{MSE}(\hat{\gamma})}$
500	$\hat{\gamma}$	0.150	0.144	0.321	0.0871	0.0792	0.1473	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.150	–	–	0.1040	0.0965	0.1379	1.0364	1.0138	0.8766	1.0130
1000	$\hat{\gamma}$	0.114	0.121	0.258	0.0198	0.0100	0.1269	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.114	–	–	0.0249	0.0112	0.1355	1.0646	1.0652	1.1406	1.1412
2000	$\hat{\gamma}$	0.086	0.102	0.220	−0.0024	−0.0092	0.0967	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.086	–	–	−0.0014	−0.0168	0.1155	1.1301	1.1710	1.4277	1.4449
4000	$\hat{\gamma}$	0.065	0.086	0.189	−0.0003	−0.0041	0.0642	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.065	–	–	−0.0009	−0.0085	0.0771	1.1567	1.1831	1.4415	1.4529

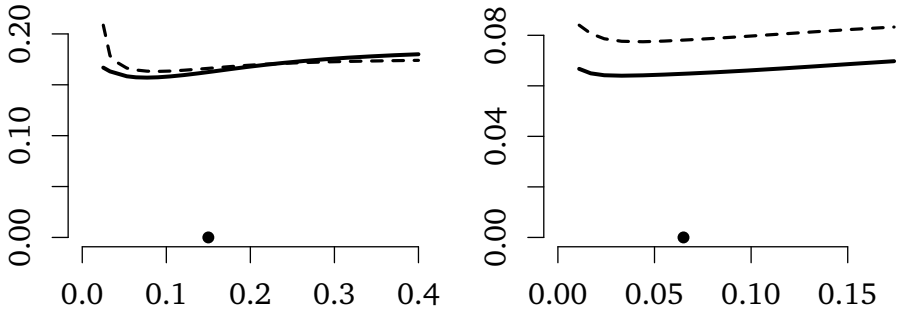
The oracle bandwidth b_n^* is chosen using the procedure described in Section C.2.1. The c_n^*, d_n^* reported here are the medians (across all simulations) of the bandwidths chosen via cross-validation. A "–" indicates that $\hat{\gamma}_{\text{VS}}$ does not depend on c_n^*, d_n^*. AD is short for Absolute Deviation.

Figure C.2: Smoothed density of $\hat{\gamma} - \gamma^*$ (solid) and $\hat{\gamma}_{\text{VS}} - \gamma^*$ (dashed) in Design 1.



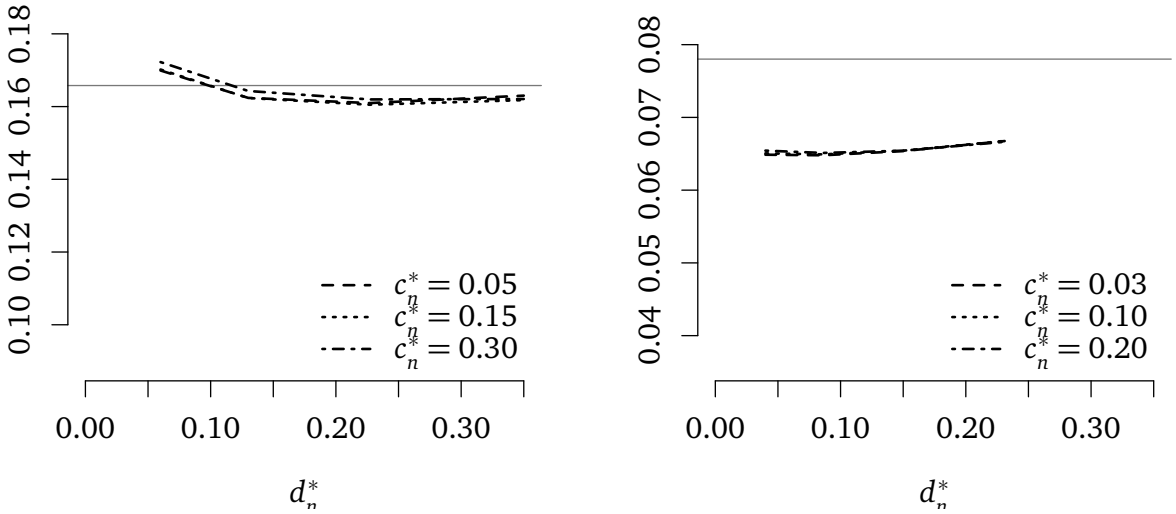
The left panel is for $n = 500$, and the right panel for $n = 4000$.

Figure C.3: RMSE of $\hat{\gamma}$ (solid) and $\hat{\gamma}_{\text{VS}}$ (dashed) as a function of b_n in Design 1.



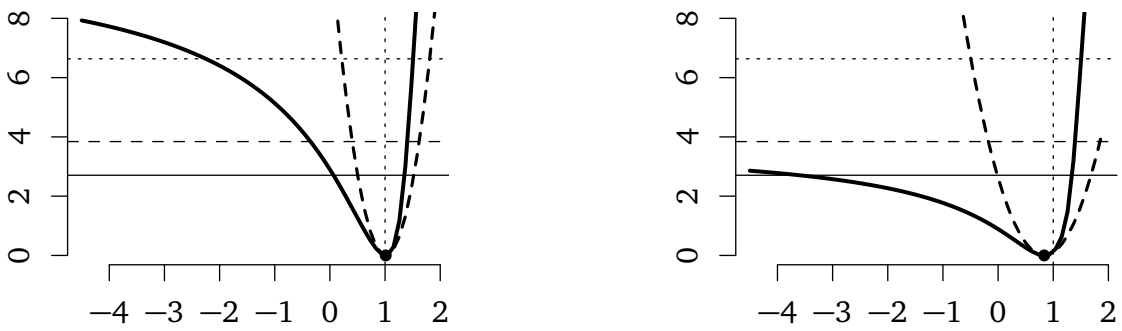
The left panel is for $n = 500$, and the right panel for $n = 4000$. The black dot is b_n^*.

Figure C.4: RMSE($\hat{\gamma}$) as a function of (c_n^*, d_n^*) in Design 1.



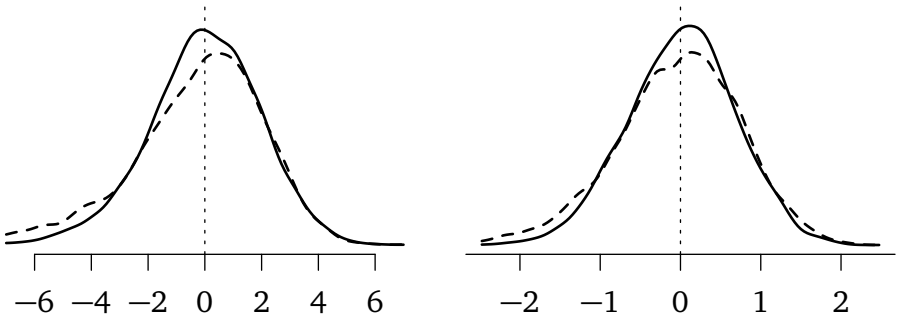
The left panel is for $n = 500$, and the right panel for $n = 4000$. The horizontal line is RMSE($\hat{\gamma}_{\text{VS}}$).

Figure C.5: Shape of $\gamma \mapsto \mathrm{LR}^{\mathrm{p}}(\gamma)$ in Design 1.



In the left plot, the solid curve is $\mathrm{LR}^{\mathrm{p}}(\gamma)$, whereas the dashed curve is the Wald statistic $\mathrm{W}(\gamma) := |(\hat{\gamma} - \gamma)/\,\mathrm{se}(\hat{\gamma})|^2$. The right plot shows the LR statistic based on $\hat{\gamma}_{\mathrm{VS}}$ (solid), and the corresponding Wald statistic (dashed). The vertical line is the location of the true $\gamma(= 1)$, whereas the black point shows the location of $\hat{\gamma}$ in the left plot, and $\hat{\gamma}_{\mathrm{VS}}$ in the right plot. The horizontal lines are the $\{.9, .95, .99\}$-quantiles of a χ^2_1 random variable. The above plots were obtained using one simulated dataset with $n = 500$ (220 observations in the validation sample). In this dataset, the 95% — hence, the 99% — LR confidence interval based on $\hat{\gamma}_{\mathrm{VS}}$ is unbounded from the left. [Numerical evaluations reveal that the line $y = 3.81$ is a horizontal asymptote to the graph of the $\hat{\gamma}_{\mathrm{VS}}$-based LR statistic at $-\infty$. Therefore, the left branch of the graph of the LR statistic based on $\hat{\gamma}_{\mathrm{VS}}$ never exceeds the .95 quantile (3.84) — hence, the .99 quantile (6.63) — of a χ^2_1 random variable.]

Figure C.6: Smoothed density of $\hat{\gamma} - \gamma^*$ (solid) and $\hat{\gamma}_{\mathrm{VS}} - \gamma^*$ (dashed) in Design 2.



The left panel is for $n = 500$, and the right panel for $n = 4000$.

Table C.2: LR confidence intervals for γ^* in Design 1.

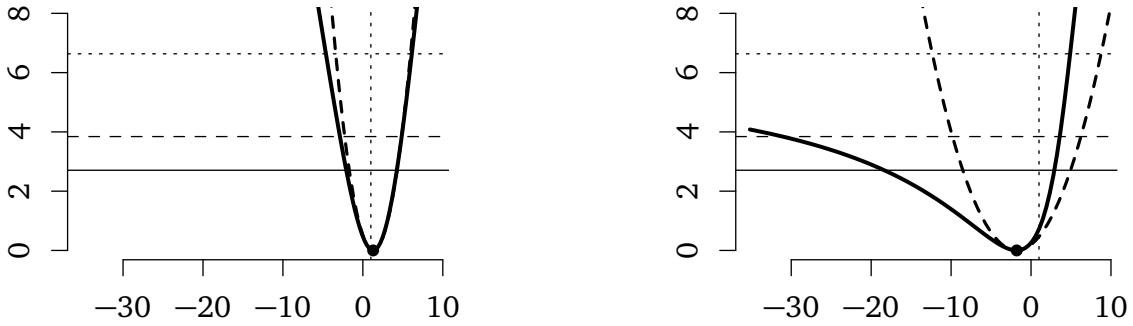
		Coverage Probability			
n	Estimator	Nominal	Empirical	Median length	% bounded
500	$\hat{\gamma}$	0.90	0.916	0.646	100
		0.95	0.968	0.828	100
		0.99	0.995	1.354	99.6
	$\hat{\gamma}_{\text{VS}}$	0.90	0.920	0.698	100
		0.95	0.970	0.943	100.0
		0.99	0.995	1.820	93.9
1000	$\hat{\gamma}$	0.90	0.933	0.489	100
		0.95	0.971	0.605	100
		0.99	0.996	0.879	100.0
	$\hat{\gamma}_{\text{VS}}$	0.90	0.942	0.584	100
		0.95	0.973	0.753	100.0
		0.99	0.996	1.246	99.1
2000	$\hat{\gamma}$	0.90	0.914	0.331	100
		0.95	0.958	0.401	100
		0.99	0.993	0.551	100
	$\hat{\gamma}_{\text{VS}}$	0.90	0.911	0.398	100
		0.95	0.957	0.491	100
		0.99	0.994	0.710	100
4000	$\hat{\gamma}$	0.90	0.913	0.219	100
		0.95	0.957	0.264	100
		0.99	0.993	0.353	100
	$\hat{\gamma}_{\text{VS}}$	0.90	0.912	0.258	100
		0.95	0.955	0.313	100
		0.99	0.994	0.429	100

A "100.0" in the last column (due to roundoff rules) indicates that there are 99.95% or more bounded intervals. The bandwidths used here are the same used to estimate $\hat{\gamma}$ and $\hat{\gamma}_{\text{VS}}$ (cf. Section C.2.1).

Table C.3: Simulation summary for the estimated γ^* in Design 2.

n	Estimator	Median Bias	Mean Bias	Std. Dev.	$\frac{\text{MedianAD}(\cdot)}{\text{MedianAD}(\hat{\gamma})}$	$\frac{\text{MeanAD}(\cdot)}{\text{MeanAD}(\hat{\gamma})}$	$\frac{\text{var}(\cdot)}{\text{var}(\hat{\gamma})}$	$\frac{\text{MSE}(\cdot)}{\text{MSE}(\hat{\gamma})}$
500	$\hat{\gamma}$	0.0418	−0.0316	2.0204	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	−0.0252	−0.6041	5.0498	1.1716	1.4167	6.2470	6.3349
1000	$\hat{\gamma}$	0.0269	−0.0266	1.3979	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.0042	−0.2288	1.8047	1.1067	1.1989	1.6668	1.6930
2000	$\hat{\gamma}$	0.0407	0.0193	0.9634	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.0150	−0.0808	1.1751	1.1572	1.1845	1.4877	1.4942
4000	$\hat{\gamma}$	0.0338	0.0136	0.6693	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.0224	−0.0356	0.7884	1.1519	1.1608	1.3879	1.3901

Figure C.7: Shape of $\gamma \mapsto \mathrm{LR}^{\mathrm{p}}(\gamma)$ in Design 2.



In the left plot, the solid curve is the LR statistic $\mathrm{LR}^{\mathrm{p}}(\gamma)$, whereas the dashed curve is the Wald statistic $\mathrm{W}(\gamma) := |(\hat{\gamma} - \gamma)/\,\mathrm{se}(\hat{\gamma})|^2$. The right plot shows the LR statistic based on $\hat{\gamma}_{\mathrm{VS}}$ (solid), and the corresponding Wald statistic (dashed). The vertical line is the location of the true $\gamma(= 1)$, whereas the black point shows the location of $\hat{\gamma}$ in the left plot, and $\hat{\gamma}_{\mathrm{VS}}$ in the right plot. The horizontal lines are the $\{.9, .95, .99\}$-quantiles of a χ^2_1 random variable. The above plots were obtained using one simulated dataset with $n = 500$ (320 observations in the validation sample). In this dataset, the 99% LR confidence interval based on $\hat{\gamma}_{\mathrm{VS}}$ is unbounded from the left. [Numerical evaluations reveal that the line $y = 6.47$ is a horizontal asymptote to the graph of the $\hat{\gamma}_{\mathrm{VS}}$-based LR statistic at $-\infty$. Therefore, the left branch of the graph of the LR statistic based on $\hat{\gamma}_{\mathrm{VS}}$ never exceeds the .99 quantile (6.63) of a χ^2_1 random variable.]

Table C.4: LR confidence intervals for γ^* in Design 2.

		Coverage Probability			
n	Estimator	Nominal	Empirical	Median length	% bounded
500	$\hat{\gamma}$	.90	.905	6.66	100
		.95	.952	8.17	100
		.99	.991	11.51	100.0
	$\hat{\gamma}_{\text{VS}}$	.90	.897	8.43	96.9
		.95	.949	10.77	94.1
		.99	.990	16.67	84.2
1000	$\hat{\gamma}$	.90	.903	4.59	100
		.95	.953	5.54	100
		.99	.993	7.54	100
	$\hat{\gamma}_{\text{VS}}$	.90	.900	5.53	100.0
		.95	.952	6.83	99.8
		.99	.992	9.91	99.2
2000	$\hat{\gamma}$	.90	.898	3.19	100
		.95	.952	3.83	100
		.99	.990	5.12	100
	$\hat{\gamma}_{\text{VS}}$	.90	.897	3.73	100
		.95	.947	4.53	100
		.99	.991	6.23	100
4000	$\hat{\gamma}$	.90	.904	2.24	100
		.95	.957	2.68	100
		.99	.991	3.55	100
	$\hat{\gamma}_{\text{VS}}$	.90	.903	2.59	100
		.95	.948	3.11	100
		.99	.991	4.18	100

C.5 Implementation details for Design 1

Motivated by the discussion in Hall (1990, Section 3), before estimating $\tilde{\pi}$ and μ we apply an injective transformation to $(Z_1,\ldots,Z_n)$ and $(X_1,\ldots,X_n)$ to map their distinct elements into $(0,1)$ in order to simplify the problem of bandwidth choice and deal with edge effects in kernel estimators. We have found that doing so improves the performance of our kernel estimators. We describe this procedure in detail as it may be useful to other applied researchers as well.

Let the random variable A denote Z or X (if Z or X are vectors, the procedure is applied coordinatewise). There are no missing observations in $\mathscr{A} := (A_1,\ldots,A_n)$ because Z and X are observed for each i. Let $M_n := \sum_{i=1}^n D_i$ be the size of the validation sample. Since the validation sample only contains those i for which $D_i = 1$, we have $\mathscr{A} = \mathscr{V} \cup \mathscr{N}$, where $\mathscr{V}$ is the ordered array (keeping ties preserved) of observations in the validation sample, and $\mathscr{N}$ is the ordered array (keeping ties preserved) of observations not in the validation sample.[12] Let $A^{\text{VS}}_{(1)} \leq \ldots \leq A^{\text{VS}}_{(M_n)}$ denote the ordered observations in $\mathscr{V}$, and define

$$\begin{aligned}
\mathscr{N}_1 &:= \text{ordered array of elements of } \mathscr{N} \text{ in } (-\infty, A^{\text{VS}}_{(1)}) \\
\mathscr{N}_j &:= \text{ordered array of elements of } \mathscr{N} \text{ in } (A^{\text{VS}}_{(j-1)}, A^{\text{VS}}_{(j)}) \qquad (j = 2,\ldots,M_n) \\
\mathscr{N}_{M_n+1} &:= \text{ordered array of elements of } \mathscr{N} \text{ in } (A^{\text{VS}}_{(M_n)}, \infty).
\end{aligned}$$

If $(A^{\text{VS}}_{(j-1)}, A^{\text{VS}}_{(j)})$ is empty (e.g., when there are ties in $\mathscr{V}$), then $\mathscr{N}_j$ is the empty array. Let $\hat{F}_{\mathscr{V}}(a) := M_n^{-1} \sum_{j=1}^{M_n} \mathbb{1}(A^{\text{VS}}_{(j)} \leq a)$, $a \in \mathbb{R}$, be the empirical cumulative distribution function (cdf) of the observations in $\mathscr{V}$, and define

$$\begin{aligned}
\mathscr{T}_1 &:= \text{set of tick marks in an equispaced grid on } (0, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) - \frac{0.5}{M_n}) \\
&\qquad \text{with as many ticks as the number of distinct elements in } \mathscr{N}_1, \\
\mathscr{T}_j &:= \text{set of tick marks in an equispaced grid on } (\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(j-1)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(j)}) - \frac{0.5}{M_n}) \\
&\qquad \text{with as many ticks as the number of distinct elements in } \mathscr{N}_j,\ j = 2,\ldots,M_n, \\
\mathscr{T}_{M_n+1} &:= \text{set of tick marks in an equispaced grid on } (\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(M_n)}) - \frac{0.5}{M_n}, 1) \\
&\qquad \text{with as many ticks as the number of distinct elements in } \mathscr{N}_{M_n}.
\end{aligned}$$

Note that $\mathscr{T}_j$ is empty if $\mathscr{N}_j$ is the empty array.

Now, for $i = 1,\ldots,n$, map $A_i \to (0,1)$ as follows:

$$\Psi_n(A_i) := \begin{cases} \hat{F}_{\mathscr{V}}(A_i) - \dfrac{0.5}{M_n} & \text{if } A_i \in \{A^{\text{VS}}_{(1)},\ldots,A^{\text{VS}}_{(M_n)}\} \\ \text{tick in } \mathscr{T}_j, \text{ repeated as many times as the multiplicity of } A_i \in \mathscr{N}_j, \text{ such that the order in which } A_i \text{ appears in } \mathscr{N}_j \text{ is preserved} & \text{if } A_i \in \mathscr{N}_j \ (j = 1,\ldots,M_n+1). \end{cases}$$

In words, Ψ_n makes distinct elements of $(A_1,\ldots,A_n)$ equally spaced in the validation and non-validation subsamples by placing observations from the validation sample $0.5/M_n$ units below their values under $\hat{F}_{\mathscr{V}}$, whereas observations not in the validation sample are placed equally apart in $0 < \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) - 0.5/M_n \leq \ldots \leq \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(M_n)}) - 0.5/M_n < 1$, taking ties into account. Spacing the observations equally in each subsample mitigates the problem of bandwidth selection in low-density regions of the conditioning variable, and ensuring that the observations stay away from the boundary improves the small-sample properties of the kernel estimators. As Ψ_n is injective by construction, the information set used to estimate the conditional expectations remains unchanged.

The following numerical example illustrates how Ψ_n works.

[12] Note that $\mathscr{V}$ and $\mathscr{N}$ may have elements in common (corresponding to different i).

Example C.1. Let $n = 11$, $M_n = 5$, $\mathscr{V} = (1,1,3,4,6)$ and $\mathscr{N} = (0,2,2,5.9,7,8)$. In this dataset, $A^{\text{VS}}_{(1)} = 1$, $A^{\text{VS}}_{(2)} = 1$, $A^{\text{VS}}_{(3)} = 3$, $A^{\text{VS}}_{(4)} = 4$, $A^{\text{VS}}_{(5)} = 6$. Hence, $\mathscr{N}_1 = (0)$, $\mathscr{N}_2 = \vec{\varnothing}$, $\mathscr{N}_3 = (2,2) = (2^{\text{multiplicity=2}})$, $\mathscr{N}_4 = \vec{\varnothing}$, $\mathscr{N}_5 = (5.9)$, and $\mathscr{N}_6 = (7,8)$. Next, as $\hat{F}_{\mathscr{V}}(a) = [2\mathbb{1}(1 \le a) + \mathbb{1}(3 \le a) + \mathbb{1}(4 \le a) + \mathbb{1}(6 \le a)]/5$, we have

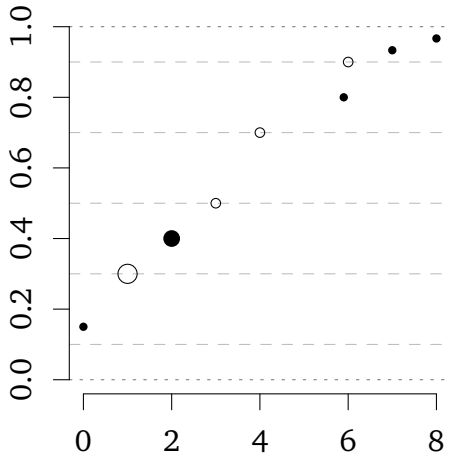
$$
\begin{aligned}
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) = \frac{2}{5} &\implies (0, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) - \frac{0.5}{M_n}) = (0, 0.3) \overset{\mathscr{N}_1=(0)}{\Longrightarrow} \mathscr{T}_1 = \{0.15\} \\
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(2)}) = \frac{2}{5} &\implies (\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(2)}) - \frac{0.5}{M_n}) = (0.3, 0.3) \overset{\mathscr{N}_2=\vec{\varnothing}}{\Longrightarrow} \mathscr{T}_2 = \vec{\varnothing} \\
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(3)}) = \frac{3}{5} &\implies (\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(2)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(3)}) - \frac{0.5}{M_n}) = (0.3, 0.5) \overset{\mathscr{N}_3=(2^{\text{multiplicity=2}})}{\Longrightarrow} \mathscr{T}_3 = \{0.4\} \\
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(4)}) = \frac{4}{5} &\implies (\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(3)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(4)}) - \frac{0.5}{M_n}) = (0.5, 0.7) \overset{\mathscr{N}_4=\vec{\varnothing}}{\Longrightarrow} \mathscr{T}_4 = \vec{\varnothing} \\
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(5)}) = 1 &\implies (\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(4)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(5)}) - \frac{0.5}{M_n}) = (0.7, 0.9) \overset{\mathscr{N}_5=(5.9)}{\Longrightarrow} \mathscr{T}_5 = \{0.8\} \\
&(\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(5)}) - \frac{0.5}{M_n}, 1) = (0.9, 1) \overset{\mathscr{N}_6=(7,8)}{\Longrightarrow} \mathscr{T}_6 = \{\frac{28}{30}, \frac{29}{30}\}.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\Psi_{11}(\mathscr{V}) &= (\hat{F}_{\mathscr{V}}(1) - \frac{0.5}{5}, \hat{F}_{\mathscr{V}}(1) - \frac{0.5}{5}, \hat{F}_{\mathscr{V}}(3) - \frac{0.5}{5}, \hat{F}_{\mathscr{V}}(4) - \frac{0.5}{5}, \hat{F}_{\mathscr{V}}(6) - \frac{0.5}{5}) \\
&= (0.3, 0.3, 0.5, 0.7, 0.9); \\
\Psi_{11}(\mathscr{N}) &= (\text{ticks in } \mathscr{T}_1, \ldots, \mathscr{T}_6 \text{ preserving the multiplicity and order in } \mathscr{N}_1, \ldots, \mathscr{N}_6) \\
&= (0.15, 0.4, 0.4, 0.8, \frac{28}{30}, \frac{29}{30}).
\end{aligned}
$$

The graph of Ψ_{11} is shown in Figure C.8. □

Figure C.8: The graph of Ψ_{11}.



The empty circles denote points in $\mathscr{V}$, and the filled circles denote points in $\mathscr{N}$. The larger circles correspond to observations with multiplicity 2.

C.6 Efficiency gains in the simulation designs

Let $\tilde{Z} := (1, Z)_{2\times 1}$. The efficiency bound for estimating $\theta^* = (\alpha^*, \gamma^*)_{2\times 1}$ in the structural model $Y^* = \tilde{Z}'\theta^* + \sigma(X)U$ is given by $\text{l.b.}(\theta^*) = (\mathbb{E}J'J/\Omega_\rho)^{-1}$, where, cf. Example 4.3,

$$\begin{aligned} J &= -\begin{bmatrix} 1 & \mathbb{E}[Z \mid X] \end{bmatrix} \\ \Omega_\rho &= \tilde{\pi}^{-1}\mathbb{E}[\text{var}(Y^* \mid Z, X) \mid X] + \mathbb{E}[\mu^2 \mid X] \\ \mu &= \mathbb{E}[Y^* \mid Z, X] - \tilde{Z}'\theta^*. \end{aligned}$$

As noted at the beginning of Section C, MAR and the condition $D \perp\!\!\!\perp Z \mid X$ imply that $\text{l.b.}_{\text{VS}}(\theta^*) = (\mathbb{E}\tilde{\pi}J'J/\Omega_g)^{-1}$, where $\Omega_g = \mathbb{E}[(Y^* - \tilde{Z}'\theta^*)^2 \mid X] = \mathbb{E}[\sigma^2(X)U^2 \mid X] = \sigma^2(X)\sigma_U^2$, because $U \perp\!\!\!\perp X$. Hence, $\text{l.b.}_{\text{VS}}(\theta^*) = \sigma_U^2(\mathbb{E}\tilde{\pi}J'J/\sigma^2(X))^{-1}$. We now obtain the efficiency gain $\text{l.b.}_{\text{VS}}(\gamma^*)/\text{l.b.}(\gamma^*)$ by simplifying the expressions for J and Ω_ρ in the two designs. Recall that $\tilde{X} := (1, X)_{2\times 1}$ and let $\zeta := (\zeta_0, \zeta_1)_{2\times 1}$, so that $\zeta_0 + \zeta_1 X = \tilde{X}'\zeta$.

C.6.1 Design 1

In this design, $Z = \tilde{X}'\zeta + V$. Hence,

$$\begin{aligned} \mathbb{E}[U \mid Z, X] &= \mathbb{E}[U \mid X, \tilde{X}'\zeta + V] && \text{(defn. of } Z\text{)} \\ &= \mathbb{E}[U \mid X, V] && ((X, V) \mapsto (X, \tilde{X}'\zeta + V) \text{ is injective}) \\ &= \mathbb{E}[U \mid V]. && ((U, V) \perp\!\!\!\perp X) \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}[Y^* \mid Z, X] &= \tilde{Z}'\theta^* + \sigma(X)\mathbb{E}[U \mid Z, X] \\ &= \tilde{Z}'\theta^* + \sigma(X)\mathbb{E}[U \mid V] \\ &= \tilde{Z}'\theta^* + \sigma(X)\frac{\sigma_{UV}}{\sigma_V^2}V. && (U, V \text{ jointly normal}) \end{aligned}$$

Consequently,

$$\mu = \sigma(X)\frac{\sigma_{UV}}{\sigma_V^2}V \implies \mathbb{E}[\mu^2 \mid X] = \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}.$$

Next, as $\text{var}[U \mid X] = \mathbb{E}[\text{var}[U \mid Z, X] \mid X] + \text{var}[\mathbb{E}[U \mid Z, X] \mid X]$ by variance decomposition,

$$\begin{aligned} \mathbb{E}[\text{var}[U \mid Z, X] \mid X] &= \text{var}[U \mid X] - \text{var}[\mathbb{E}[U \mid Z, X] \mid X] \\ &= \text{var}[U \mid X] - \text{var}[\mathbb{E}[U \mid V] \mid X] \\ &= \text{var}[U \mid X] - \text{var}[\frac{\sigma_{UV}}{\sigma_V^2}V \mid X] \\ &= \text{var}[U \mid X] - \frac{\sigma_{UV}^2}{\sigma_V^4}\text{var}[V \mid X] \\ &= \sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2}. && ((U, V) \perp\!\!\!\perp X) \end{aligned}$$

Consequently,

$$\mathbb{E}[\text{var}[Y^* \mid X, Z] \mid X] = \sigma^2(X)\mathbb{E}[\text{var}[U \mid Z, X] \mid X] = \sigma^2(X)[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2}].$$

Combining these results, we get that

$$\Omega_\rho = \frac{1}{\tilde{\pi}(X)}\sigma^2(X)[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2}] + \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}.$$

Therefore, the efficiency bound for θ^* in design 1 is

$$\text{l.b.}(\theta^*) = \Big(\mathbb{E}\frac{J'J}{\Omega_\rho}\Big)^{-1} = \Big(\mathbb{E}\frac{\left[\begin{smallmatrix} 1 & \tilde{X}'\zeta \\ \tilde{X}'\zeta & (\tilde{X}'\zeta)^2 \end{smallmatrix}\right]}{\frac{\sigma^2(X)}{\tilde{\pi}(X)}[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2}] + \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}}\Big)^{-1}.$$

Furthermore, the efficiency bound for estimating θ^* using only the validation sample is

$$\text{l.b.}_{\text{VS}}(\theta^*) = \sigma_U^2(\mathbb{E}\frac{\tilde{\pi}(X)}{\sigma^2(X)}J'J)^{-1} = \sigma_U^2\Big(\mathbb{E}\frac{\tilde{\pi}(X)}{\sigma^2(X)}\left[\begin{smallmatrix} 1 & \tilde{X}'\zeta \\ \tilde{X}'\zeta & (\tilde{X}'\zeta)^2 \end{smallmatrix}\right]\Big)^{-1}.$$

Hence, the efficiency gain $\text{l.b.}_{\text{VS}}(\gamma^*)/\text{l.b.}(\gamma^*)$ can be obtained from the expressions for the 2×2 matrices $\text{l.b.}_{\text{VS}}(\theta^*)$ and $\text{l.b.}(\theta^*)$ by extracting their $(2,2)$ elements.

C.6.2 Design 2

In this design, $Z = \mathbb{1}(\tilde{X}'\zeta + V > 0)$. Hence,

$$J = -[1 \quad \mathbb{E}[Z \mid X]] = -[1 \quad \Pr(Z=1 \mid X)] = -[1 \quad \Phi(\tfrac{\tilde{X}'\zeta}{\sigma_V})].$$

For $d \in \mathbb{R}$, joint normality of (U,V) implies that

$$\mathbb{E}[U\mathbb{1}(V \le d)] = \mathbb{E}[\mathbb{E}[U \mid V]\mathbb{1}(V \le d)] = \frac{\sigma_{UV}}{\sigma_V^2}\mathbb{E}[V\mathbb{1}(V \le d)] = -\frac{\sigma_{UV}}{\sigma_V}\phi(\frac{\tilde{X}'\zeta}{\sigma_V}).$$

Hence, $\mathbb{E}[U\mathbb{1}(V > d)] = \frac{\sigma_{UV}}{\sigma_V}\phi(\frac{\tilde{X}'\zeta}{\sigma_V})$ because $\mathbb{E}U = 0$. Consequently,

$$\begin{aligned}
\mathbb{E}[U \mid Z,X] &= (1-Z)\mathbb{E}[U \mid Z=0,X] + Z\mathbb{E}[U \mid Z=1,X] && (Z \in \{0,1\}) \\
&= (1-Z)\frac{\mathbb{E}[U\mathbb{1}(Z=0) \mid X]}{\Pr(Z=0 \mid X)} + Z\frac{\mathbb{E}[U\mathbb{1}(Z=1) \mid X]}{\Pr(Z=1 \mid X)} \\
&= (1-Z)\frac{\mathbb{E}[U\mathbb{1}(V \le -\tilde{X}'\zeta) \mid X]}{\Pr(V \le -\tilde{X}'\zeta \mid X)} + Z\frac{\mathbb{E}[U\mathbb{1}(V > -\tilde{X}'\zeta) \mid X]}{\Pr(V > -\tilde{X}'\zeta \mid X)} \\
&= -(1-Z)\frac{\sigma_{UV}}{\sigma_V}\frac{\phi(\frac{\tilde{X}'\zeta}{\sigma_V})}{\Phi(\frac{-\tilde{X}'\zeta}{\sigma_V})} + Z\frac{\sigma_{UV}}{\sigma_V}\frac{\phi(\frac{\tilde{X}'\zeta}{\sigma_V})}{\Phi(\frac{\tilde{X}'\zeta}{\sigma_V})} && ((U,V) \text{ normal and indep. of } X) \\
&= \frac{\sigma_{UV}}{\sigma_V}[Z - \Phi(\frac{\tilde{X}'\zeta}{\sigma_V})]G(\frac{\tilde{X}'\zeta}{\sigma_V}),
\end{aligned}$$

where $G(t) := \phi(t)/[\Phi(t)\Phi(-t)]$, $t \in \mathbb{R}$, is the probit weight function (Schumann and Tripathi, 2018). Therefore,

$$\mathbb{E}[Y^* \mid Z,X] = \tilde{Z}'\theta^* + \sigma(X)\mathbb{E}[U \mid Z,X] = \tilde{Z}'\theta^* + \sigma(X)\frac{\sigma_{UV}}{\sigma_V}[Z - \Phi(\frac{\tilde{X}'\zeta}{\sigma_V})]G(\frac{\tilde{X}'\zeta}{\sigma_V}).$$

Consequently, $\mu = \sigma(X)\frac{\sigma_{UV}}{\sigma_V}[Z - \Phi(\frac{\tilde{X}'\zeta}{\sigma_V})]G(\frac{\tilde{X}'\zeta}{\sigma_V})$. Therefore, since $\mathbb{E}[\mu \mid X] = 0$,

$$\begin{aligned}
\mathbb{E}[\mu^2 \mid X] = \text{var}[\mu \mid X] &= \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}G^2(\frac{\tilde{X}'\zeta}{\sigma_V})\,\text{var}[Z \mid X] \\
&= \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}G^2(\frac{\tilde{X}'\zeta}{\sigma_V})\Phi(\frac{\tilde{X}'\zeta}{\sigma_V})\Phi(-\frac{\tilde{X}'\zeta}{\sigma_V}) \\
&= \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}G(\frac{\tilde{X}'\zeta}{\sigma_V})\phi(\frac{\tilde{X}'\zeta}{\sigma_V}).
\end{aligned}$$

Next, as $\text{var}[U \mid X] = \mathbb{E}[\text{var}[U \mid Z,X] \mid X] + \text{var}[\mathbb{E}[U \mid Z,X] \mid X]$ by variance decomposition,

$$\begin{aligned}
\mathbb{E}[\text{var}[U \mid Z,X] \mid X] &= \text{var}[U \mid X] - \text{var}[\mathbb{E}[U \mid Z,X] \mid X] \\
&= \sigma_U^2 - \text{var}[\frac{\sigma_{UV}}{\sigma_V}[Z - \Phi(\frac{\tilde{X}'\zeta}{\sigma_V})]G(\frac{\tilde{X}'\zeta}{\sigma_V}) \mid X] \qquad (U \perp\!\!\!\perp X) \\
&= \sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G^2(\frac{\tilde{X}'\zeta}{\sigma_V}) \text{var}[Z \mid X] \\
&= \sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G^2(\frac{\tilde{X}'\zeta}{\sigma_V}) \Phi(-\frac{\tilde{X}'\zeta}{\sigma_V}) \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) \\
&= \sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V}) \phi(\frac{\tilde{X}'\zeta}{\sigma_V}).
\end{aligned}$$

Consequently,

$$\mathbb{E}[\text{var}[Y^* \mid X,Z] \mid X] = \sigma^2(X)\mathbb{E}[\text{var}[U \mid Z,X] \mid X] = \sigma^2(X)[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V}) \phi(\frac{\tilde{X}'\zeta}{\sigma_V})].$$

Combining these results, we get that

$$\Omega_\rho = \frac{1}{\tilde{\pi}(X)} \sigma^2(X)[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V}) \phi(\frac{\tilde{X}'\zeta}{\sigma_V})] + \sigma^2(X) \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V}) \phi(\frac{\tilde{X}'\zeta}{\sigma_V}).$$

Therefore, the efficiency bound for θ^* in design 2 is

$$\begin{aligned}
\text{l.b.}(\theta^*) &= \Big(\mathbb{E}\frac{J'J}{\Omega_\rho}\Big)^{-1} \\
&= \Big(\mathbb{E}\frac{\begin{bmatrix} 1 & \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) \\ \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) & \Phi^2(\frac{\tilde{X}'\zeta}{\sigma_V}) \end{bmatrix}}{\frac{\sigma^2(X)}{\tilde{\pi}(X)}[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V}) \phi(\frac{\tilde{X}'\zeta}{\sigma_V})] + \sigma^2(X) \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V}) \phi(\frac{\tilde{X}'\zeta}{\sigma_V})}\Big)^{-1} \\
&= \Big(\mathbb{E}\frac{\begin{bmatrix} 1 & \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) \\ \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) & \Phi^2(\frac{\tilde{X}'\zeta}{\sigma_V}) \end{bmatrix}}{\frac{(1-X)\sigma^2(0)+X\sigma^2(1)}{(1-X)\tilde{\pi}(0)+X\tilde{\pi}(1)}[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V}) \phi(\frac{\tilde{X}'\zeta}{\sigma_V})] + [(1-X)\sigma^2(0) + X\sigma^2(1)] \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V}) \phi(\frac{\tilde{X}'\zeta}{\sigma_V})}\Big)^{-1},
\end{aligned}$$

where the last equality follows because $X \in \{0,1\}$.

Furthermore, the efficiency bound for estimating θ^* using only the validation sample is

$$\text{l.b.}_{\text{VS}}(\theta^*) = \sigma_U^2 (\mathbb{E}\frac{\tilde{\pi}(X)}{\sigma^2(X)} J'J)^{-1} = \sigma_U^2 \Big(\mathbb{E}\frac{(1-X)\tilde{\pi}(0) + X\tilde{\pi}(1)}{(1-X)\sigma^2(0) + X\sigma^2(1)} \begin{bmatrix} 1 & \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) \\ \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) & \Phi^2(\frac{\tilde{X}'\zeta}{\sigma_V}) \end{bmatrix}\Big)^{-1}.$$

Hence, the efficiency gain $\text{l.b.}_{\text{VS}}(\gamma^*)/\text{l.b.}(\gamma^*)$ can be obtained from the expressions for the 2×2 matrices $\text{l.b.}_{\text{VS}}(\theta^*)$ and $\text{l.b.}(\theta^*)$ by extracting their $(2,2)$ elements.

C.6.3 $\widehat{\text{SEL}}$ in Design 2

Here, $\hat{\rho}_j(\theta)$ is scalar and $\hat{\theta}$ maximizes the version of $\widehat{\text{SEL}}_{\mathbb{T}}$ with $\hat{\mathbb{T}}_{1i} := 1$, $\hat{\mathbb{T}}_{2j} := 1$, and the redefined weights

$$w_{ij} := \frac{\mathbb{1}(X_i = X_j)}{\sum_{k=1}^n \mathbb{1}(X_i = X_k)} = \mathbb{1}(X_j = X_i)[\frac{\mathbb{1}(X_i = 0)}{n(1-\bar{X})} + \frac{\mathbb{1}(X_i = 1)}{n\bar{X}}] \tag{C.1}$$

with $\bar{X} := \sum_{j=1}^n X_j/n$. The maximizers of the inner optimization problems in

$$\widehat{\mathrm{SEL}}(\theta) := -\sum_{i=1}^n \max_{\tilde{\lambda}_i \in \mathbb{R}} \sum_{j=1}^n w_{ij} \log(1 + \tilde{\lambda}_i \hat{\rho}_j(\theta)),$$

denoted by $\hat{\lambda}_i(\theta)$, $i = 1, \ldots, n$, satisfy the FOC

$$\begin{aligned}
0 &= \sum_{j=1}^n \frac{w_{ij}\hat{\rho}_j(\theta)}{1 + \hat{\lambda}_i(\theta)\hat{\rho}_j(\theta)} && (i = 1, \ldots, n) \\
&\overset{(C.1)}{=} \frac{\mathbb{1}(X_i = 0)}{n(1-\bar{X})} \sum_{j=1}^n \frac{1(X_j = X_i)\hat{\rho}_j(\theta)}{1 + \hat{\lambda}_i(\theta)\hat{\rho}_j(\theta)} + \frac{\mathbb{1}(X_i = 1)}{n\bar{X}} \sum_{j=1}^n \frac{1(X_j = X_i)\hat{\rho}_j(\theta)}{1 + \hat{\lambda}_i(\theta)\hat{\rho}_j(\theta)} \\
&= \begin{cases} \displaystyle\sum_{j=1}^n \frac{\mathbb{1}(X_j = 0)\hat{\rho}_j(\theta)}{1 + \hat{l}_0(\theta)\hat{\rho}_j(\theta)} & \text{if } X_i = 0 \\ \displaystyle\sum_{j=1}^n \frac{\mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)}{1 + \hat{l}_1(\theta)\hat{\rho}_j(\theta)} & \text{if } X_i = 1, \end{cases} && \text{(C.2)}
\end{aligned}$$

where $\hat{l}_0(\theta), \hat{l}_1(\theta) \in \mathbb{R}$ solve (C.2). Hence, $\hat{\lambda}_i(\theta) = \hat{l}_0(\theta)\mathbb{1}(X_i = 0) + \hat{l}_1(\theta)\mathbb{1}(X_i = 1)$ and

$$\begin{aligned}
\widehat{\mathrm{SEL}}(\theta) &= -\sum_{i=1}^n \sum_{j=1}^n w_{ij} \log(1 + \hat{\lambda}_i(\theta)\hat{\rho}_j(\theta)) \\
&= -\sum_{i=1}^n \sum_{j=1}^n \frac{\mathbb{1}(X_j = X_i)}{\sum_{k=1}^n \mathbb{1}(X_i = X_k)} \log(1 + \hat{l}_0(\theta)\mathbb{1}(X_i = 0)\hat{\rho}_j(\theta) + \hat{l}_1(\theta)\mathbb{1}(X_i = 1)\hat{\rho}_j(\theta)) \\
&= -\sum_{i=1}^n \sum_{j=1}^n \frac{\mathbb{1}(X_j = X_i)}{\sum_{k=1}^n \mathbb{1}(X_i = X_k)} \log(1 + \hat{l}_0(\theta)\mathbb{1}(X_j = 0)\hat{\rho}_j(\theta) + \hat{l}_1(\theta)\mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)) \\
&= -\sum_{j=1}^n \log(1 + \hat{l}_0(\theta)\mathbb{1}(X_j = 0)\hat{\rho}_j(\theta) + \hat{l}_1(\theta)\mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)) \sum_{i=1}^n \frac{\mathbb{1}(X_j = X_i)}{\sum_{k=1}^n \mathbb{1}(X_i = X_k)} \\
&= -\sum_{j=1}^n \log(1 + \hat{l}_0(\theta)\mathbb{1}(X_j = 0)\hat{\rho}_j(\theta) + \hat{l}_1(\theta)\mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)) && \text{(C.3)} \\
&\overset{(C.2)}{=} -\max_{l_0, l_1 \in \mathbb{R}} \sum_{j=1}^n \log(1 + l_0\mathbb{1}(X_j = 0)\hat{\rho}_j(\theta) + l_1\mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)),
\end{aligned}$$

where (C.3) follows from the fact that

$$\begin{aligned}
\sum_{i=1}^n \frac{\mathbb{1}(X_j = X_i)}{\sum_{k=1}^n \mathbb{1}(X_i = X_k)} &= \sum_{i=1}^n \frac{\mathbb{1}(X_j = X_i)}{\mathbb{1}(X_i = 0)n(1-\bar{X}) + \mathbb{1}(X_i = 1)n\bar{X}} \\
&= \sum_{i=1}^n \frac{\mathbb{1}(X_j = X_i)[\mathbb{1}(X_j = 0) + \mathbb{1}(X_j = 1)]}{\mathbb{1}(X_i = 0)n(1-\bar{X}) + \mathbb{1}(X_i = 1)n\bar{X}} \\
&= \sum_{i=1}^n \frac{\mathbb{1}(X_i = 0)\mathbb{1}(X_j = 0) + \mathbb{1}(X_i = 1)\mathbb{1}(X_j = 1)}{\mathbb{1}(X_i = 0)n(1-\bar{X}) + \mathbb{1}(X_i = 1)n\bar{X}} \\
&= \frac{\mathbb{1}(X_j = 0)}{n(1-\bar{X})} \sum_{i=1}^n \mathbb{1}(X_i = 0) + \frac{\mathbb{1}(X_j = 1)}{n\bar{X}} \sum_{i=1}^n \mathbb{1}(X_i = 1) \\
&= \mathbb{1}(X_j = 0) + \mathbb{1}(X_j = 1) \\
&= 1.
\end{aligned}$$

Therefore, $\widehat{\mathrm{SEL}}(\cdot)$ coincides with unconditional empirical likelihood because

$$\begin{aligned}
\widehat{\mathrm{SEL}}(\theta) &= -\max_{l_0,l_1\in\mathbb{R}} \sum_{j=1}^{n} \log(1 + \mathbb{1}(X_j = 0) l_0 \hat{\rho}_j(\theta) + \mathbb{1}(X_j = 1) l_1 \hat{\rho}_j(\theta)) \\
&= -\max_{l_0,l_1\in\mathbb{R}} \sum_{j=1}^{n} \log(1 + l_0 \hat{\rho}_j(\theta) + \mathbb{1}(X_j = 1)(l_1 - l_0) \hat{\rho}_j(\theta)) \\
&= -\max_{l\in\mathbb{R}^2} \sum_{j=1}^{n} \log(1 + l' \tilde{X}_j \hat{\rho}_j(\theta)). \qquad (X \in \{0,1\} \implies \mathbb{1}(X = 1) = X)
\end{aligned}$$

Consequently, if $\hat{\theta}$ solves $\sum_{j=1}^{n} \tilde{X}_j \hat{\rho}_j(\hat{\theta}) = 0$, then it also maximizes $\widehat{\mathrm{SEL}}(\cdot)$.

D Proofs, examples, and technical details

Remark D.1 (HP vs. Ackerberg, Chen, Hahn, and Liao (2014))**.** HP (p. 736) consider an arbitrary family of moment conditions by interacting their moment functions with basis functions that span a space of square-integrable functions. Therefore, Eqns. 4 and 6 in HP are conditional (not unconditional) moment equalities with the conditioning set determined by variables in the basis functions. Consequently, the model in HP is not encompassed within Ackerberg et al. □

D.1 Illustrating the nonmissing excluded endogenous variables Z_{ex} introduced in Section 2

The following example illustrates how nonmissing endogenous variables can be excluded from a CMR model with missing endogenous variables, and yet still appear in the propensity score function. Accompanying the example is a small simulation experiment demonstrating that efficiency gains in estimation from the observed sample are still achievable.

Example D.1. A university placement office conducts an online survey of professional achievements of its recent alumni. Their records already contain the number of years of education (*educ*) and the years of labor market experience (*exper*). They also ask about hourly wage ($wage^*$) and hours worked (*hours*) per week. Respondents report their hours worked, but not everyone is willing to disclose their wage, so $wage^*$ is missing for some respondents in the sample.[13]

Using this data, a researcher wants to estimate the wage regression

$$\log(wage^*) = \beta_0^* + \beta_1^* educ + \beta_2^* exper + U, \tag{D.1}$$

under the assumption that the regressors *educ* and *exper* are both exogenous with respect to U.[14] Note that *hours* is excluded from the wage regression because the outcome variable in (D.1) is hourly wage (in log), which does not depend on the hours worked per week, as these are fixed in employment contracts. However, as wages and *hours* are jointly determined in equilibrium, *hours* is correlated with $\log(wage^*)$, making it endogenous with respect to U. Therefore, *hours* is a nonmissing endogenous variable excluded from the wage regression (D.1). Nevertheless, *hours* appears in the propensity score function as it is informative about the probability of survey completion (cf. Footnote 13 for an explanation).

The CMR model in this example is

$$\mathbb{E}[\log(wage^*) - \beta_0^* - \beta_1^* educ - \beta_2^* exper \mid educ, exper] \overset{\text{w.p.1}}{=} 0. \tag{D.2}$$

In our notation, (D.2) $\iff$ $\mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0$, where the moment function $g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta^*) := \log(wage^*) - \beta_0 - \beta_1 educ - \beta_2 exper$ with $Y^* := \log(wage^*)$; no nonmissing included endogenous variables ($Z_{\text{in}} := \vec{\varnothing}$) and *hours* the only nonmissing excluded endogenous variable ($Z_{\text{ex}} := hours$), implying that $Z = Z_{\text{ex}} = hours$; $X_{\text{in}} := (educ, exper)_{2\times 1}$ and $X_{\text{ex}} := \vec{\varnothing}$, implying that $X = X_{\text{in}} = (educ, exper)$; and $\theta^* := (\beta_0^*, \beta_1^*, \beta_2^*)_{3\times 1}$. The propensity score function (the probability that $wage^*$ is reported) is $\pi(hours, educ, exper)$, which depends on the nonmissing endogenous variable *hours* excluded from the CMR (D.2).

[13] For simplicity, assume that higher-paying, prestigious jobs require longer working hours (the actual relationship between hours worked and wages can be highly nonlinear, cf., e.g., Bick, Blandin, and Rogerson, 2022). Individuals in such positions may be less inclined to report their wages due to security concerns. Moreover, individuals working longer hours may have limited time to complete the survey, leading to premature survey termination and incomplete data (in most real surveys, questions on wages typically appear after those on experience and hours worked). If, conditional on individuals' educational attainment, labor market experience, and hours worked, the likelihood of not reporting wages is the same for everyone, then the missingness in $wage^*$ can appropriately be modeled as MAR.

[14] Exogeneity of *educ* is assumed here for convenience, as our goal is to illustrate how a nonmissing endogenous variable can be excluded from a CMR model with missing outcomes.

Simulation design. We now present the results from a small simulation experiment to demonstrate how the single nonmissing endogenous variable *hours*, which is excluded from the wage regression (D.1) with missing outcomes, can still deliver efficiency gains in estimation from the observed sample.

The simulated data are generated as follows.[15] Summary statistics for the simulated dataset are in Table D.1 (sample size $n = 1000$).

1. Generate *educ* as a random integer between 8 and 20 with probability mass function $p(i) = .06 + .02(i-8)$ if $8 \leq i \leq 12$ and $.115 - .015(i-13)$ if $13 \leq i \leq 20$.

2. Generate *exper* as a random integer in $\{0,1,2,3\}$ with probabilities $(.1,.5,.3,.1)$.

3. Generate $U \overset{\mathrm{d}}{=} \mathrm{N}(0, \sigma_U^2(X))$, where $\sigma_U^2(X) := (1 + \mathit{educ} + \mathit{exper})^2$.

4. Obtain the dependent variable in (D.1) as $\log(\mathit{wage}^*) := 5 + \mathit{educ} + 2\mathit{exper} + U$.

5. Generate $W \overset{\mathrm{d}}{=} \mathrm{N}(0, \sigma_W^2(X))$, where $\sigma_W^2(X) := 1 + \mathit{educ} + \mathit{exper}$.

6. Obtain the nonmissing excluded endogenous variable as

$$\mathit{hours} := \max\{1, \mathrm{round}(35 + 2\mathrm{clamp}_{-10,15}(\log(\mathit{wage}^*)) + W\sigma_W(W))\},$$

   where the clamping function $\mathrm{clamp}_{l,u}(\cdot)$ is defined in Footnote 4, and the function $\mathrm{round}(\cdot)$ rounds off to the nearest integer. This specification makes *hours* strongly endogenous ($\mathrm{corr}(U, \mathit{hours}) \approx 0.66$).

7. Compute the propensity score function as $\pi(Z,X) := \max\{0.05, \varpi(Z,X)\}$, where $\varpi(Z,X)$ is defined to be

   - $\mathrm{clamp}_{0,0.179}(\mathit{hours}) + \mathrm{clamp}_{0,0.13}(\mathit{educ}) + \mathrm{clamp}_{0,0.12}(\mathit{exper})$ for 25% missingness in wage^*;
   - $\mathrm{clamp}_{0,0.405}(\mathit{hours}) + \mathrm{clamp}_{0,0.29}(\mathit{educ}) + \mathrm{clamp}_{0,0.27}(\mathit{exper})$ for 50% missingness in wage^*;
   - $\mathrm{clamp}_{0,0.658}(\mathit{hours}) + \mathrm{clamp}_{0,0.485}(\mathit{educ}) + \mathrm{clamp}_{0,0.435}(\mathit{exper})$ for 75% missingness in wage^*.

   This design yields three levels of missingness in wage^* (namely, 25%, 50%, 75%), and makes the propensity score function $\pi(\mathit{hours}, \mathit{educ}, \mathit{exper})$ a decreasing function of its arguments, leading to higher missingness of wage^* for individuals with more hours worked, education, and experience.

8. Finally, let $D := \mathbb{1}(\pi(Z,X) > R)$ with $R \overset{\mathrm{d}}{=} \mathrm{Unif}[0,1] \perp\!\!\!\perp (Z,X,U)$ ensuring that MAR holds, and $\mathit{wage} := D\mathit{wage}^* + (1-D)\mathfrak{m}$.

As in the empirical illustration (cf. Section 5.2), for each level of missingness in wage^* we estimate θ^* using the semiparametrically efficient SEL estimator $\hat{\theta}$ based on the observed sample, and the SEL-IPW estimator $\hat{\theta}_{\mathrm{SEL,IPW}}$ based on the validation sample.

The propensity score π is estimated nonparametrically with a mixed kernel smoother by regressing D on (*educ*, *exper*, *hours*), treating *hours* as continuously distributed and applying the injective transformation in Appendix C.5. Bandwidth for estimating π was chosen by cross-validation. The function μ is estimated nonparametrically with a mixed kernel smoother by regressing Dg on (*educ*, *exper*, *hours*) in the validation sample, treating *hours* as continuously

[15] Here, "generate" means "draw a pseudo-random variable independent of previous draws."

distributed and applying the injective transformation in Appendix C.5. Bandwidth for estimating μ was chosen by cross-validation.

Simulation results and discussion. The simulation results for $\hat{\theta}$ and $\hat{\theta}_{\text{SEL,IPW}}$ reported in Table D.2 and Figure D.1 are based on 10,000 simulations. They clearly reveal that $\hat{\theta}$ is working as desired and delivering meaningful efficiency gains. For each level of missingness, $\hat{\theta}$ is properly centered at the true value with very little bias — much less compared to the bias of $\hat{\theta}_{\text{SEL,IPW}}$, and its MSE for the coefficients on *educ* and *exper* is lower than that of $\hat{\theta}_{\text{SEL,IPW}}$. As the missingness in *wage** increases, the bias and variance of $\hat{\theta}$ worsen less rapidly than those of $\hat{\theta}_{\text{SEL,IPW}}$. Hence, in terms of the MSE, the efficiency of $\hat{\theta}$ relative to $\hat{\theta}_{\text{SEL,IPW}}$ increases with the missingness in *wage**. The much smaller bias of $\hat{\theta}$ is due to the double robustness property of ρ, which makes it insensitive to local perturbations in π and μ. Under low missingness rates, the bias is virtually zero for both estimators, while the MSE of $\hat{\theta}_{\text{SEL,IPW}}$ is 9.2–13.2% higher, suggesting that there is no penalty to using the efficient SEL estimator $\hat{\theta}$ even when the missingness problem is relatively mild. □

Table D.1: (Example D.1) Summary statistics for the simulated dataset.

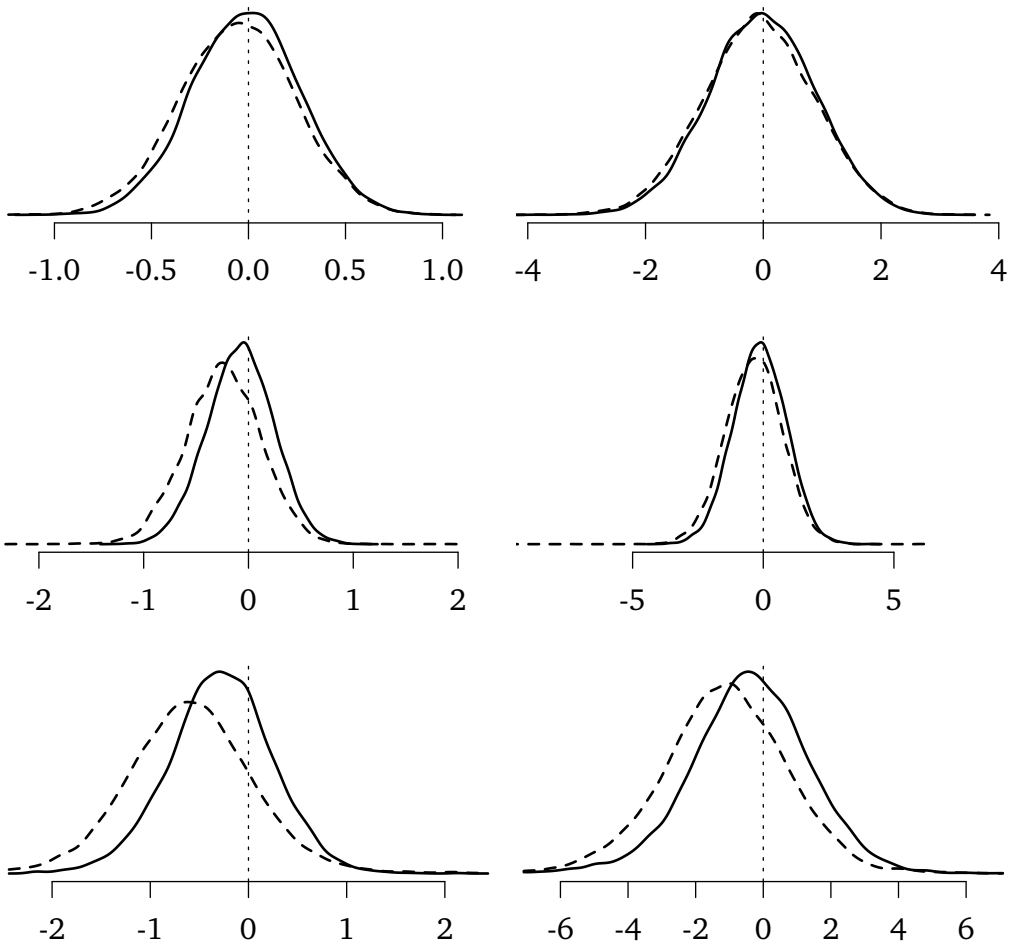
Variable	Mean	Std. Dev.	1st quartile	median	3rd quartile
$\log(wage^*)$	20.6	22.2	6.0	20.0	34.6
educ	12.8	2.93	11.0	12.0	15.0
exper	1.4	0.80	1.0	1.0	2.0
hours	39.8	5.32	36.0	40.0	43.0

Sample size $n = 1000$.

Table D.2: (Example D.1) Simulation summary based on 10,000 simulations.

	Estimator	Missingness in $wage^*$	(Intercept)	*educ*	*exper*
Average Bias	$\hat{\theta}_{\text{SEL,IPW}}$	25%	0.0763	−0.0526	−0.0383
		50%	0.4753	−0.2404	−0.1766
		75%	1.1608	−0.5969	−0.5583
	$\hat{\theta}$	25%	0.0080	−0.0108	−0.0082
		50%	0.1525	−0.0751	−0.0654
		75%	0.5114	−0.2777	−0.1868
MSE	$\hat{\theta}_{\text{SEL,IPW}}$	25%	14.4073	0.0896	0.8915
		50%	26.4917	0.1896	1.3143
		75%	45.2471	0.4167	2.1280
	$\hat{\theta}$	25%	12.7285	0.0769	0.8168
		50%	16.7797	0.1064	1.0388
		75%	20.3621	0.1593	1.2788
$\frac{\text{MSE}(\hat{\theta}_{\text{SEL,IPW}})}{\text{MSE}(\hat{\theta})} - 1$		25%	13.2%	16.5%	9.2%
		50%	57.9%	78.1%	26.5%
		75%	122.2%	161.5%	66.4%

Figure D.1: (Example D.1) Smoothed density of $\hat{\beta}_{\text{educ}} - \beta^*_{\text{educ}}$ (left) and $\hat{\beta}_{\text{exper}} - \beta^*_{\text{exper}}$ (right). The solid line is the SEL estimator based on the observed sample, and the dashed line is the SEL-IPW estimator based on the validation sample. The 1st row displays the smoothed densities when there is 25% missingness in *wage**; the 2nd row for 50% missingness; and the 3rd row for 75% missingness.



D.2 Proofs, examples, and technical details for Section 3

D.2.1 Selection on observables vs. selection on unobservables

A selection on unobservables approach for identifying parameters is one where the MAR assumption does not hold, and identification is achieved by imposing distributional and exclusion restrictions on variables in the full population. This explains why a selection on unobservables approach is commonly employed in a parametric framework, whereas selection on observables is typically used in semiparametric settings. The following example, motivated by the discussion in Fitzgerald, Gottschalk, and Moffitt (1998, Section III.A), illustrates why MAR is called selection on observables, and why its negation is termed selection on unobservables.

Example D.2. Consider the IV regression in Example 2.1 with potentially missing outcomes. Then, letting $R \mid Z, X \overset{\mathrm{d}}{=} \mathrm{Unif}(0,1)$, we have the following system of equations: $D := \mathbb{1}(\pi(Z,X) > R)$ (the selection model) and $Y^* = \alpha^* + X_{\mathrm{in}}'\beta^* + Z_{\mathrm{in}}'\gamma^* + U$ (the structural model) with $\mathbb{E}[U \mid X] \overset{P_X\text{-a.s.}}{=} 0$. Since the structural model specifies Y^* to be a linear function of $(X_{\mathrm{in}}, Z_{\mathrm{in}}, U)$, it is clear that $D \perp\!\!\!\perp Y^* \mid Z,X \iff \mathbb{1}(\pi(Z,X) > R) \perp\!\!\!\perp U \mid Z,X$. Therefore, if $R \perp\!\!\!\perp U \mid Z,X$ (i.e., the model error in the selection equation cannot affect the potential outcomes in the structural equation), then $D \perp\!\!\!\perp Y^* \mid Z,X$ (i.e., MAR holds); hence, MAR is called selection on observables. Moreover, $D \not\perp\!\!\!\perp Y^* \mid Z,X \implies R \not\perp\!\!\!\perp U \mid Z,X$, i.e., if MAR is false, then R and U cannot be conditionally independent given Z,X. In other words, if MAR does not hold, then the unobservable model error in the selection equation can affect the potential outcomes in the structural equation; consequently, as noted in Fitzgerald et al. (p. 257), the negation of MAR is called selection on unobservables. □

D.2.2 Local identification in conditional moment models

Let the $\dim(g) \times \dim(\theta^*)$ matrix $J(X,\theta) := \partial_\theta \mathbb{E}[g(Y^*, Z_{\mathrm{in}}, X_{\mathrm{in}}, \theta) \mid X]$, and recall that $J := J(X,\theta^*)$ and $\|\cdot\|$ is the Euclidean norm.

Definition D.1 (Linear independence P_X-a.s.)**.** The columns of J are said to be linearly independent P_X-a.s. if, for all $\alpha \in \mathbb{R}^{\dim(\theta^*)}$, $P_X(J\alpha = 0_{\dim(g)\times 1}) = 1 \implies \alpha = 0_{\dim(\theta^*)\times 1}$.

In this section, we extend Rothenberg (1971) to show that θ^* in (2.1) is locally identified if the columns of J are linearly independent P_X-a.s. We begin by defining the notion of observational equivalence for conditional moment equalities.

Definition D.2 (Observational equivalence of parameters)**.** Parameters $\theta^*, \theta^\dagger \in \Theta$ are said to be observationally equivalent for (2.1) if

$$\mathbb{E}[g(Y^*, Z_{\mathrm{in}}, X_{\mathrm{in}}, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \overset{P_X\text{-a.s.}}{=} \mathbb{E}[g(Y^*, Z_{\mathrm{in}}, X_{\mathrm{in}}, \theta^\dagger) \mid X].$$

In other words, θ^* and $\theta^\dagger$ are observationally equivalent if they satisfy the same conditional moment equality. Next, we define what it means for θ^* to be locally identified.

Definition D.3 (Local identification)**.** Parameter $\theta^* \in \Theta$ is said to be locally identified if there exists an open ball centered at θ^*, say $\mathcal{N}^* \subset \Theta$, such that the punctured open ball $\mathcal{N}^* \setminus \{\theta^*\}$ does not contain any element observationally equivalent to θ^*.

We now prove the local identification result stated in Section 3. Although Lemma D.1 below looks as if it should be well known, we have been unable to find it in the literature.[16]

[16] Identification of parameters defined via unconditional moment equalities is discussed in Newey and McFadden (1994, Section 2.2.3).

Lemma D.1. *Assume that* (P_X*-a.s.*): (a) $\theta \mapsto J(X,\theta)$ *is well defined on an open* (*in* Θ) *ball centered at* θ^*; *and* (b) $\theta \mapsto J(X,\theta)$ *is continuous at* θ^*. *If the columns of* J *are linearly independent* P_X*-a.s., then* θ^* *in* (2.1) *is locally identified.*

Proof of Lemma D.1. Suppose to the contrary that θ^* is not locally identified. Then, by Definition D.3, each punctured open ball centered at θ^* contains at least one element different from θ^* that is observationally equivalent to θ^*. This yields a sequence $(\theta_j)_{j\in\mathbb{N}} \subset \Theta$ such that (i) $\lim_{j\to\infty} \theta_j = \theta^*$, (ii) $\theta_j \neq \theta^*$ for each $j \in \mathbb{N}$, and (iii) θ_j is observationally equivalent to θ^* for each $j \in \mathbb{N}$. Letting $m(X,\theta) := \mathbb{E}[g(Y^*, Z_{\text{in}}, X_{\text{in}}, \theta) \mid X]$ and $q := \dim(g)$, an element by element mean value expansion of $m(X,\theta_j)$ about θ^* reveals that

$$m(X,\theta_j) \overset{\text{(a)}}{=} m(X,\theta^*) + \begin{bmatrix} d_1'(X, \theta^* + \lambda_1(\theta_j - \theta^*)) \\ \vdots \\ d_q'(X, \theta^* + \lambda_q(\theta_j - \theta^*)) \end{bmatrix} (\theta_j - \theta^*) \qquad P_X\text{-a.s.}, \tag{D.3}$$

where d_k' denotes the k^{th} row of J, and each $\lambda_k \in (0,1)$. Hence, by (iii) and Definition D.2,

$$\begin{bmatrix} d_1'(X, \theta^* + \lambda_1(\theta_j - \theta^*)) \\ \vdots \\ d_q'(X, \theta^* + \lambda_q(\theta_j - \theta^*)) \end{bmatrix} (\theta_j - \theta^*) \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \qquad (j \in \mathbb{N})$$

By (ii), $r_j := (\theta_j - \theta^*)/\|\theta_j - \theta^*\|$ is well defined for each j. Hence, we can write the previous displayed equation as

$$\begin{bmatrix} d_1'(X, \theta^* + \lambda_1 r_j \|\theta_j - \theta^*\|) \\ \vdots \\ d_q'(X, \theta^* + \lambda_q r_j \|\theta_j - \theta^*\|) \end{bmatrix} r_j \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \qquad (j \in \mathbb{N})$$

Now, (r_j) is a bounded sequence in $\mathbb{R}^{\dim(\theta^*)}$ because $\|r_j\| = 1$ for each j. Hence, by the Bolzano-Weierstrass theorem, there exists a subsequence $(s_j) \subset (r_j)$, and $r^* \in \mathbb{R}^{\dim(\theta^*)}$ with $\|r^*\| = 1$, such that $\lim_{j\to\infty} s_j = r^*$. In particular, since (s_j) is a subsequence of (r_j), we have that

$$\begin{bmatrix} d_1'(X, \theta^* + \lambda_1 s_j \|\theta_j - \theta^*\|) \\ \vdots \\ d_q'(X, \theta^* + \lambda_q s_j \|\theta_j - \theta^*\|) \end{bmatrix} s_j \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}. \qquad (j \in \mathbb{N})$$

Since each row of $J(X,\theta)$ is continuous at θ^* (P_X-a.s.) if and only if $J(X,\theta)$ is continuous at θ^* (P_X-a.s.), letting $j \to \infty$ in the previous displayed equation, (b) and (i) imply that

$$\begin{bmatrix} d_1'(X, \theta^*) \\ \vdots \\ d_q'(X, \theta^*) \end{bmatrix} r^* \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \iff J r^* \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}.$$

But, as $r^* \neq 0_{\dim(\theta^*)\times 1}$, this contradicts the assumption that the columns of J are linearly independent P_X-a.s. The desired result follows. □

Remark D.2. (i). For (a) in the statement of Lemma D.1 to hold, it is necessary that Θ has a non-empty interior.

(ii). The condition that the columns of the Jacobian matrix J are linearly independent P_X-a.s. also leads to the global identification of θ^* whenever g is linear in θ^*, because the mean value expansion in (D.3) is exact whenever g is linear in θ^*. □

The following useful result is a direct consequence of Lemma D.1.

Proposition D.1. *Local identification of θ^* in the CMR* (2.1) *is equivalent to the local identification of θ^* in the IPW CMR* (3.1).

Proof of Proposition D.1. Assume that the columns of the Jacobian matrix J are linearly independent P_X-a.s. By Lemma D.1, this is sufficient to ensure that θ^* is locally identified. Hence, as π does not depend on θ (Assumption 3.2), $J = \partial_{\theta^*}\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X]$ P_X-a.s. Therefore, the columns of J are linearly independent P_X-a.s. if and only if the columns of $\partial_{\theta^*}\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X]$ are linearly independent P_X-a.s. □

As noted in Remark D.2, Lemma D.1 implies the global identification of θ^* whenever g is linear in θ^*. This is easily verified for linear regression models.

Example D.3. In Example 2.1, $g := Y^* - \alpha^* - X'_{\text{in}}\beta^* - Z'_{\text{in}}\gamma^*$. Hence, $\mathbb{E}[g \mid X] = \mathbb{E}[Y^* \mid X] - \alpha^* - X'_{\text{in}}\beta^* - \mathbb{E}[Z'_{\text{in}} \mid X]\gamma^*$. Consequently, $\theta^* := (\alpha^*, \beta^*, \gamma^*)$ is globally identified if and only if the columns of $\begin{bmatrix} 1 & X'_{\text{in}} & \mathbb{E}[Z'_{\text{in}} \mid X] \end{bmatrix}$ are linearly independent P_X-a.s. The connection with Lemma D.1 is apparent because, in this example, $J = -\begin{bmatrix} 1 & X'_{\text{in}} & \mathbb{E}[Z'_{\text{in}} \mid X] \end{bmatrix}$. □

D.3 Proofs, examples, and technical details for Section 4

Remark D.3. (i) The system of equations (4.1)&(4.2) differs from the one in Ackerberg et al. (2014) because (4.1), unlike Eqn. 1 of Ackerberg et al., is a CMR.

(ii) The approach of combining moment conditions we employ is similar to those used earlier by Newey (1994, p. 17), Ai and Chen (2012, Section 2), Ackerberg et al. (Eqn. 8), and Hristache and Patilea (2016, Section 4.1). □

Additional notation. Henceforth, $L_2(Z,X)^\perp$ denotes the orthogonal complement of $L_2(Z,X)$, the set of real-valued functions of (Z,X) with finite second moments. For a generic random vector W, let $L_{2,0}(W) := \{\psi \in L_2(W) : \mathbb{E}\psi(W) = 0\}$ be the set of real-valued functions of W with finite second moments whose expectation is zero. If $S \subset L_2(W)$, then $S^\perp$ is the orthogonal complement of S in $L_2(W)$, and $\mathscr{P}_S(a)$ denotes the orthogonal projection of $a \in L_2(W)$ onto S using the inner product $\langle a_1, a_2\rangle := \mathbb{E}[a_1 a_2]$. The norm induced by the inner product is $\|a\|_2 := \sqrt{\mathbb{E}a^2}$. Let $\sigma_g^2 := \sigma_g^2(X) := \mathbb{E}[g'g \mid X]$, and $\|\cdot\|_\infty$ denote the supremum norm, e.g., $\|\sigma_g^2\|_\infty := \sup_{\text{supp}(X)} \sigma_g^2$. The $\dim(\theta^*)$-fold cartesian product of $L_2(Z,X)$ with itself is denoted by $L_2(Z,X)^{\dim(\theta^*)}$;

Lemma D.2. *If* (4.1)&(4.2) *hold, then $\mu(\frac{D}{\pi} - 1)$ is the coordinatewise projection of $\frac{Dg_{\text{obs}}}{\pi}$ onto $L_2(D,Z,X) \cap L_2(Z,X)^\perp$. Hence, ρ defined in* (4.4) *is the residual from projecting $\frac{Dg_{\text{obs}}}{\pi}$ coordinatewise onto the tangent space of score functions for π.*

Proof Lemma D.2. The tangent space of score functions for π is equal to $L_2(D,Z,X) \cap L_2(Z,X)^\perp$, the tangent space of score functions for the density of $D \mid Z,X$ (cf. the proof of Lemma 4.1), because π determines the density of $D \mid Z,X$. Therefore, it is enough to show that $\mu(\frac{D}{\pi} - 1)$ is the coordinatewise projection of $\frac{Dg_{\text{obs}}}{\pi}$ onto $L_2(D,Z,X) \cap L_2(Z,X)^\perp$. So let (4.1)&(4.2) hold, which, by the equivalence in (4.3), implies that MAR holds. Let $\alpha \in \mathbb{R}^{\dim(g)}$ be such that $\|\alpha\| = 1$. Then,

$\alpha'\mu[D/\pi-1] \in L_2(D,Z,X)$ because

$$\begin{aligned}
(\alpha'\mu[\frac{D}{\pi}-1])^2 &\leq \|\mu\|^2|\frac{D}{\pi}-1|^2 && \text{(Cauchy-Schwarz)}\\
&\leq 2\|\mu\|^2(\frac{1}{\pi^2}+1) && (|a-b|^2 \leq 2(a^2+b^2))\\
&\leq 2(\frac{1}{(\inf\pi)^2}+1)\|\mu\|^2, && \text{(Ass. D.1(i))}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\|\mu\|^2 &= \mathbb{E}\|\mathbb{E}[g \mid Z,X]\|^2 && \text{(defn. } \mu)\\
&= \mathbb{E}[(\mathbb{E}[g^{(1)} \mid Z,X])^2 + \ldots + (\mathbb{E}[g^{(\dim(g))} \mid Z,X])^2]\\
&\leq \mathbb{E}[\mathbb{E}[(g^{(1)})^2 \mid Z,X] + \ldots + \mathbb{E}[(g^{(\dim(g))})^2 \mid Z,X]] && \text{(cond. Jensen)}\\
&= \mathbb{E}[g^{(1)}]^2 + \ldots + \mathbb{E}[g^{(\dim(g))}]^2 && \text{(iterated expectations)}\\
&= \mathbb{E}\|g\|^2\\
&< \infty. && \text{(Ass. D.1(ii))}
\end{aligned}$$

Moreover, $\alpha'\mu(D/\pi-1) \in L_2(Z,X)^\perp$ because

$$\mathbb{E}[\mu(\frac{D}{\pi}-1) \mid Z,X] \overset{P_{Z,X}\text{-a.s.}}{=} \mu\mathbb{E}[\frac{D}{\pi}-1 \mid Z,X] \overset{(4.2)}{=} 0_{\dim(g)\times 1}.$$

Hence, $\alpha'\mu[D/\pi-1] \in L_2(D,Z,X) \cap L_2(Z,X)^\perp$. Therefore, to prove that $\alpha'\mu(D/\pi-1)$ is the projection of $\alpha' Dg_{\text{obs}}/\pi$ onto $L_2(D,Z,X) \cap L_2(Z,X)^\perp$, it remains to show that the residual $\alpha' Dg_{\text{obs}}/\pi - \alpha'\mu(D/\pi-1) \overset{(4.4)}{=} \alpha'\rho$ is orthogonal to $L_2(D,Z,X) \cap L_2(Z,X)^\perp$, i.e., we have to show that $\alpha'\rho \in (L_2(D,Z,X) \cap L_2(Z,X)^\perp)^\perp$, where[17]

$$(L_2(D,Z,X) \cap L_2(Z,X)^\perp)^\perp = \overline{L_2(D,Z,X)^\perp + L_2(Z,X)} = L_2(D,Z,X)^\perp + L_2(Z,X).$$

But,

$$\begin{aligned}
\alpha'\rho \overset{(4.4)}{=} \frac{D\alpha' g_{\text{obs}}}{\pi} - \alpha'\mu[\frac{D}{\pi}-1] &= \frac{D\alpha' g}{\pi} - \alpha'\mu[\frac{D}{\pi}-1] && (Dg = Dg_{\text{obs}})\\
&= \frac{D}{\pi}\alpha'(g-\mu) + \alpha'\mu \in L_2(D,Z,X)^\perp + L_2(Z,X)
\end{aligned}$$

because $\alpha'\mu \in L_2(Z,X)$ as shown earlier, and $\frac{D}{\pi}\alpha'(g-\mu) \in L_2(D,Z,X)^\perp$ since

$$\begin{aligned}
\mathbb{E}[\frac{D}{\pi}\alpha'(g-\mu) \mid D,Z,X] &\overset{P_{D,Z,X}\text{-a.s.}}{=} \frac{D}{\pi}\alpha'\mathbb{E}[(g-\mu) \mid D,Z,X]\\
&= \frac{D}{\pi}\alpha'(\mathbb{E}[g \mid D,Z,X]-\mu)\\
&\overset{\text{MAR}}{=} \frac{D}{\pi}\alpha'(\mathbb{E}[g \mid Z,X]-\mu)\\
&\overset{\text{defn. }\mu}{=} 0. \qquad \square
\end{aligned}$$

[17]The first equality follows from Bickel, Klassen, Ritov, and Wellner (1993, Appendix A.2, Eqn. 4, p. 425) because $L_2(D,Z,X)$ and $L_2(Z,X)^\perp$ are closed subspaces (completeness of Hilbert spaces), and the second because $L_2(D,Z,X)^\perp + L_2(Z,X)$ is closed (a consequence of the orthogonality of $L_2(D,Z,X)^\perp$ and $L_2(Z,X)$, cf. Halmos, 1951, Theorem 6, p. 25).

Lemma D.3. *Under* (4.1)&(4.2)*, the Jacobian of* $\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid X]$ *with respect to* θ^* *is equal to* J*, whereas the Jacobians with respect to* π *and* μ *vanish, i.e.,*

$$\partial_{\theta^*}\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid X] \overset{P_X\text{-a.s.}}{=} J \tag{D.4}$$

$$\partial_{\pi}\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \tag{D.5}$$

$$\partial_{\mu}\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times\dim(g)}. \tag{D.6}$$

The effect of estimating a parameter is captured through its Jacobian, and the Jacobians with respect to π and μ vanish by (D.5) and (D.6). This explains why the asymptotic variance of the SEL estimator does not inflate if π and μ are replaced by their nonparametric estimators. For details, cf. the discussion following Ai and Chen (2003, Eqn. 15).

Proof of Lemma D.3. Let (4.1)&(4.2) hold, which, by the equivalence in (4.3), implies that MAR holds. Assume that derivatives with respect to (θ^*,π,μ) can be exchanged with conditional (on Z,X) expectations. Then, since $\pi := \mathbb{E}[D\mid Z,X]$ does not depend on θ^* (Assumption 3.2), we have

$$\begin{aligned}
&\partial_{\theta^*}\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid Z,X]\\
&\overset{(4.4)}{=} \partial_{\theta^*}\mathbb{E}[\frac{Dg(Y,Z_{\text{in}},X_{\text{in}},\theta^*)}{\pi(Z,X)}\mid Z,X] - \mathbb{E}[\frac{D}{\pi(Z,X)}-1\mid Z,X]\partial_{\theta^*}\mu(Z,X,\theta^*)\\
&\overset{(4.2)}{=} \frac{1}{\pi(Z,X)}\partial_{\theta^*}\mathbb{E}[Dg(Y,Z_{\text{in}},X_{\text{in}},\theta^*)\mid Z,X]\\
&= \frac{1}{\pi(Z,X)}\partial_{\theta^*}\mathbb{E}[Dg(Y^*,Z_{\text{in}},X_{\text{in}},\theta^*)\mid Z,X] \qquad (Dg(Y^*,Z_{\text{in}},X_{\text{in}},\theta^*) = Dg(Y,Z_{\text{in}},X_{\text{in}},\theta^*))\\
&\overset{\text{MAR}}{=} \frac{1}{\pi(Z,X)}\mathbb{E}[D\mid Z,X]\partial_{\theta^*}\mathbb{E}[g(Y^*,Z_{\text{in}},X_{\text{in}},\theta^*)\mid Z,X] \quad P_{Z,X}\text{-a.s.}\\
&= \partial_{\theta^*}\mathbb{E}[g(Y^*,Z_{\text{in}},X_{\text{in}},\theta^*)\mid Z,X].
\end{aligned}$$

Consequently, conditioning on X, and recalling that the tower property of conditional expectations holds almost surely and $J := \partial_\theta\mathbb{E}[g(Y^*,Z_{\text{in}},X_{\text{in}},\theta^*)\mid X]$, we have that

$$\partial_{\theta^*}\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid X] \overset{P_X\text{-a.s.}}{=} \partial_{\theta^*}\mathbb{E}[g(Y^*,Z_{\text{in}},X_{\text{in}},\theta^*)\mid X] = J;$$

i.e., (D.4) holds.

Next,

$$\begin{aligned}
\partial_{\pi}\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid Z,X] &\overset{(4.4)}{=} -\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi^2}\mid Z,X] + \mu\mathbb{E}[\frac{D}{\pi^2}\mid Z,X]\\
&\overset{(4.2)}{=} -\mathbb{E}[\frac{Dg_{\text{obs}}}{\pi^2}\mid Z,X] + \frac{\mu}{\pi}\\
&= -\frac{1}{\pi^2}\mathbb{E}[Dg\mid Z,X] + \frac{\mu}{\pi} \qquad (Dg = Dg_{\text{obs}})\\
&\overset{\text{MAR}}{=} -\frac{1}{\pi}\mathbb{E}[g\mid Z,X] + \frac{\mu}{\pi} \quad P_{Z,X}\text{-a.s.}\\
&\overset{\text{defn. }\mu}{=} 0_{\dim(g)\times 1}.
\end{aligned}$$

Consequently, conditioning on X, we get $\partial_\pi\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid X] \overset{P_X\text{-a.s.}}{=} 0$; i.e., (D.5) holds.

Finally,

$$\partial_{\mu}\mathbb{E}[\rho(\mathscr{A},\theta^*,\pi,\mu)\mid Z,X] \overset{(4.4)}{=} -I_{\dim(g)\times\dim(g)}\mathbb{E}[\frac{D}{\pi}-1\mid Z,X] \overset{(4.2)}{=} 0_{\dim(g)\times\dim(g)} \quad P_{Z,X}\text{-a.s.},$$

which implies that (D.6) also holds. □

Lemma D.4. *The following result shows that, under MAR, $\mathbb{E}[gg' \mid Z,X]$ can be written as a convex combination of $\mathbb{E}[\rho\rho' \mid Z,X]$ and $\mu\mu'$ with weights π and $1-\pi$, namely,*

$$\pi\mathbb{E}[\rho\rho' \mid Z,X] + (1-\pi)\mu\mu' \overset{\text{MAR}}{=} \mathbb{E}[gg' \mid Z,X] \quad P_{Z,X}\text{-a.s.} \tag{D.7}$$

Implications of (D.7) *are used several times in the paper (cf. Remark D.4).*

Proof of Lemma D.4. Since $Dg = Dg_{\text{obs}}$,

$$\begin{aligned}
\mathbb{E}[\rho\rho' \mid Z,X] &\overset{(4.4)}{=} \mathbb{E}[(\frac{Dg}{\pi} - \mu[\frac{D}{\pi} - 1])(\frac{Dg}{\pi} - \mu[\frac{D}{\pi} - 1])' \mid Z,X] \\
&\overset{P_{Z,X}\text{-a.s.}}{=} \frac{1}{\pi^2}\mathbb{E}[Dgg' \mid Z,X] - \mathbb{E}[\frac{Dg}{\pi}\mu'[\frac{D}{\pi} - 1] \mid Z,X] \\
&\quad - \mathbb{E}[\mu[\frac{D}{\pi} - 1]\frac{Dg'}{\pi} \mid Z,X] + \mu\mu'\mathbb{E}[(\frac{D}{\pi} - 1)^2 \mid Z,X].
\end{aligned} \tag{D.8}$$

Now,

$$\mathbb{E}[Dgg' \mid Z,X] \overset{\text{MAR}}{=} \mathbb{E}[D \mid Z,X]\mathbb{E}[gg' \mid Z,X] = \pi\mathbb{E}[gg' \mid Z,X] \quad P_{Z,X}\text{-a.s.}$$

Hence,

$$\frac{1}{\pi^2}\mathbb{E}[Dgg' \mid Z,X] \overset{\text{MAR}}{=} \frac{1}{\pi}\mathbb{E}[gg' \mid Z,X] \quad P_{Z,X}\text{-a.s.} \tag{D.9}$$

Next,

$$\begin{aligned}
\mathbb{E}[\frac{Dg}{\pi}\mu'[\frac{D}{\pi} - 1] \mid Z,X] &\overset{P_{Z,X}\text{-a.s.}}{=} \mathbb{E}[\frac{Dg}{\pi}[\frac{D}{\pi} - 1] \mid Z,X]\mu' && (\mu := \mathbb{E}[g \mid Z,X]) \\
&= \mathbb{E}[\frac{g}{\pi}[\frac{D^2}{\pi} - D] \mid Z,X]\mu' \\
&= \mathbb{E}[\frac{Dg}{\pi}[\frac{1}{\pi} - 1] \mid Z,X]\mu' && (D^2 = D) \\
&= \mathbb{E}[Dg \mid Z,X]\mu'\frac{1}{\pi}[\frac{1}{\pi} - 1] && (\pi := \mathbb{E}[D \mid Z,X]) \\
&\overset{\text{MAR}}{=} \mathbb{E}[D \mid Z,X]\mathbb{E}[g \mid Z,X]\mu'\frac{1}{\pi}[\frac{1}{\pi} - 1] \quad P_{Z,X}\text{-a.s.} \\
&= \mu\mu'[\frac{1}{\pi} - 1]. && (\pi := \mathbb{E}[D \mid Z,X], \mu := \mathbb{E}[g \mid Z,X])
\end{aligned}$$

Moreover, since $\mathbb{E}[\frac{D}{\pi} - 1 \mid Z,X] \overset{P_{Z,X}\text{-a.s.}}{=} 0$,

$$\begin{aligned}
\mathbb{E}[(\frac{D}{\pi} - 1)^2 \mid Z,X] &\overset{P_{Z,X}\text{-a.s.}}{=} \operatorname{var}[\frac{D}{\pi} - 1 \mid Z,X] \\
&\overset{P_{Z,X}\text{-a.s.}}{=} \frac{1}{\pi^2}\operatorname{var}[D \mid Z,X] \\
&= \frac{\pi(1-\pi)}{\pi^2} = \frac{1}{\pi} - 1.
\end{aligned} \tag{D.10}$$

Hence, $P_{Z,X}$-a.s.,

$$\mathbb{E}[\rho\rho' \mid Z,X] \overset{(D.8)}{=} \frac{1}{\pi}\mathbb{E}[gg' \mid Z,X] - 2\mu\mu'[\frac{1}{\pi} - 1] + \mu\mu'[\frac{1}{\pi} - 1] = \frac{1}{\pi}\mathbb{E}[gg' \mid Z,X] - \mu\mu'[\frac{1}{\pi} - 1],$$

which implies that $\pi\mathbb{E}[\rho\rho' \mid Z,X] + (1-\pi)\mu\mu' \overset{P_{Z,X}\text{-a.s.}}{=} \mathbb{E}[gg' \mid Z,X]$. □

Remark D.4. Eqn. (D.7) has several useful consequences.

(i) It implies that, under MAR,

$$\Omega_\rho \overset{P_X\text{-a.s.}}{=} \mathbb{E}[\pi^{-1}\,\mathrm{var}(g \mid Z,X) \mid X] + \mathbb{E}[\mu\mu' \mid X]. \tag{D.11}$$

This expression for Ω_ρ is used in Examples 4.3 and 4.4. To show (D.11), note that

$$\begin{aligned}(\text{D.7}) &\iff \mathbb{E}[\rho\rho' \mid Z,X] \overset{\text{MAR}}{=} \pi^{-1}\mathbb{E}[gg' \mid Z,X] - \frac{1-\pi}{\pi}\mu\mu' \quad P_{Z,X}\text{-a.s.} \\ &\iff \mathbb{E}[\rho\rho' \mid Z,X] \overset{\text{MAR}}{=} \pi^{-1}\,\mathrm{var}[g \mid Z,X] + \mu\mu' \quad P_{Z,X}\text{-a.s.},\end{aligned}$$

because $\mathrm{var}[g \mid Z,X] \overset{P_{Z,X}\text{-a.s.}}{=} \mathbb{E}[gg' \mid Z,X] - \mu\mu'$. Hence, (D.11) follows because $\Omega_\rho := \mathbb{E}[\rho\rho' \mid X]$.

(ii) An alternative expression for Ω_ρ under MAR is given by

$$\Omega_\rho \overset{P_X\text{-a.s.}}{=} \mathbb{E}[\frac{Dgg'}{\pi^2} \mid X] - \mathbb{E}[\frac{1-\pi}{\pi}\mu\mu' \mid X]. \tag{D.12}$$

This implies (D.26), which is used in the proof of (4.6). To show (D.12), note that as $\pi := \mathbb{E}[D \mid Z,X]$, we have

$$\begin{aligned}(\text{D.7}) &\iff \mathbb{E}[\rho\rho' \mid Z,X] \overset{\text{MAR}}{=} \mathbb{E}[\frac{gg'}{\pi} \mid Z,X] - \frac{1-\pi}{\pi}\mu\mu' \quad P_{Z,X}\text{-a.s.} \\ &\overset{(\text{D.9})}{\iff} \mathbb{E}[\rho\rho' \mid Z,X] \overset{\text{MAR}}{=} \mathbb{E}[\frac{Dgg'}{\pi^2} \mid Z,X] - \frac{1-\pi}{\pi}\mu\mu' \quad P_{Z,X}\text{-a.s.}\end{aligned}$$

Hence, (D.12) follows because $\Omega_\rho := \mathbb{E}[\rho\rho' \mid X]$.

(iii) Applying the trace operator to both sides of (D.7), we get that

$$\pi\mathbb{E}[\rho'\rho \mid Z,X] + (1-\pi)\mu'\mu \overset{P_{Z,X}\text{-a.s.}}{=} \mathbb{E}[g'g \mid Z,X].$$

Hence, $P_{Z,X}$-a.s.,

$$\mathbb{E}[\rho'\rho \mid Z,X] = \frac{1}{\pi}\mathbb{E}[g'g \mid Z,X] - \frac{(1-\pi)}{\pi}\mu'\mu \le \frac{1}{\inf\pi}\mathbb{E}[g'g \mid Z,X].$$

Consequently, P_X-a.s.,

$$\mathbb{E}[\rho'\rho \mid X] \le \frac{1}{\inf\pi}\mathbb{E}[g'g \mid X] \le \frac{\|\sigma_g^2\|_\infty}{\inf\pi} \overset{\text{Ass. D.1(i,ii)}}{<} \infty. \tag{D.13}$$

This bound is used in the proof of Lemma D.5. □

Proof of (4.5). Follows from

$$\mathbb{E}[\rho \mid X] \overset{(4.4)}{=} \mathbb{E}[\frac{Dg_{\text{obs}}}{\pi} \mid X] - \mathbb{E}[\mu(\frac{D}{\pi} - 1) \mid X] \overset{(4.1)}{=} -\mathbb{E}[\mu(\frac{D}{\pi} - 1) \mid X] \qquad (P_X\text{-a.s.})$$

and the fact that $\mathbb{E}[\mu(\frac{D}{\pi} - 1) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ because

$$\mathbb{E}[\mu(\frac{D}{\pi} - 1) \mid Z,X] \overset{P_{Z,X}\text{-a.s.}}{=} \mu\mathbb{E}[\frac{D}{\pi} - 1 \mid Z,X] \overset{(4.2)}{=} 0_{\dim(g)\times 1}.$$ □

The efficiency bound in Lemma 4.1 can be obtained by applying Ai and Chen (2012, Theorem 2.1) to (4.1)&(4.2). Here, we present an alternative derivation under the following conditions.

Assumption D.1. (i) $\inf_{\text{supp}(X,Z)} \pi > 0$; (ii) $\mathbb{E}\operatorname{tr} J'J < \infty$ *and* $\|\sigma_g^2\|_\infty < \infty$; (iii) *The matrix* $\mathbb{E}J'\Omega_\rho^{-1}J$ *exists and is nonsingular, and the matrix* $\mathbb{E}[J'\Omega_\rho^{-1}\frac{1-\pi}{\pi}\mu\mu'\Omega_\rho^{-1}J]$ *exists.*

(i) is necessary for θ^* to be $n^{1/2}$-estimable. In (ii), $\mathbb{E}\operatorname{tr} J'J < \infty$ implies that each element of J has finite second moment. Consequently, span(J), which denotes the set of all linear combinations of the column vectors of J, i.e., the column space of J, is closed in $L_2(X)^{\dim(g)}$. The condition $\|\sigma_g^2\|_\infty < \infty$ uniformly bounds the skedastic function for each coordinate of g. (i) and (ii) are used in the proof of Lemma D.5. (iii), which implies that the efficiency bound in Lemma 4.1 is well defined, is also necessary for θ^* to be $n^{1/2}$-estimable.

Proof of Lemma 4.1. We use the approach of Severini and Tripathi (2001, 2013, Section 12) to derive the efficiency bound for θ^*. Let $I_0 := [0, t_0]$ for some $t_0 > 0$. With respect to an appropriate dominating measure, define the probability density function $v^2 := \text{pdf}_{Y^*,Z|X}$. Let $t \mapsto v_t$ be a real-valued function defined on I_0 such that $v_t\big|_{t=0} = v$, and, suppressing the dominating measure, $\int v_t^2(y,z \mid x) = 1$ for all $(t,x) \in I_0 \times \text{supp}(X)$. The score corresponding to $\dot{v}$, the tangent to v_t at $t = 0$, is $S_{\dot{v}} := 2\dot{v}/v \in L_2(Y^*,Z,X) \cap L_2(X)^\perp$.

Let $t \mapsto \theta_t^*$ denote a $\mathbb{R}^{\dim(\theta^*)}$-valued function defined on I_0 such that $\theta_t^*\big|_{t=0} = \theta^*$ and $\int_{\text{supp}(Y^*)\times\text{supp}(Z)} g(y,z,x,\theta_t^*)v_t^2(y,z \mid x) = 0_{\dim(g)\times 1}$ for all $(t,x) \in I_0 \times \text{supp}(X)$. Then, differentiating with respect to t and evaluating at $t = 0$, we have

$$J\dot{\theta}^* + \mathbb{E}[gS_{\dot{v}} \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}, \tag{D.14}$$

where the vector $\dot{\theta}^*$ is the tangent to θ_t^* at $t = 0$. Note that (D.14) restricts the tangent vector $S_{\dot{v}}$ to be such that

$$\mathbb{E}[gS_{\dot{v}} \mid X] \overset{P_X\text{-a.s.}}{\in} \text{span}(J). \tag{D.15}$$

Let $A := A(X)$ be a $r \times \dim(g)$ matrix with $r \geq \dim(g)$ such that $B := \mathbb{E}[AJ]$ has column rank $\dim(\theta^*)$,[18] and let B^+ denote the generalized inverse of B. Then,

$$\text{(D.14)} \implies \dot{\theta}^* = -B^+\mathbb{E}[A\mathbb{E}(gS_{\dot{v}} \mid X)]. \tag{D.16}$$

Since observations on Y^* can be missing, θ^* has to be identified as a feature of $q^2 := \text{pdf}_{D,Y^*,Z,X}$, the joint density of D, Y^*, Z, X. In particular, suppose that we want the efficiency bound for estimating the functional $\eta(\log q^2) := c'\theta^*$, where $c \in \mathbb{R}^{\dim(\theta^*)}$ is arbitrary. Now,

$$\begin{aligned} q^2 := \text{pdf}_{D,Y^*,Z,X} &= \text{pdf}_{D,Y^*|Z,X}\,\text{pdf}_{Z,X} \\ &\overset{\text{MAR}}{=} \text{pdf}_{D|Z,X}\,\text{pdf}_{Y^*|Z,X}\,\text{pdf}_{Z,X} \\ &= \text{pdf}_{D|Z,X}\,\text{pdf}_{Y^*,Z,X} \\ &= \text{pdf}_{D|Z,X}\,\text{pdf}_{Y^*,Z|X}\,\text{pdf}_X \\ &= p^2v^2f^2, \end{aligned}$$

where $p^2 := \text{pdf}_{D|Z,X}$ and $f^2 := \text{pdf}_X$. Hence, $\eta(\log q^2) \overset{\text{MAR}}{=} \eta(\log p^2 + \log v^2 + \log f^2)$.

Let $t \mapsto p_t$ and $t \mapsto f_t$ be real-valued functions defined on I_0 such that $p_t\big|_{t=0} = p$, $f_t\big|_{t=0} = f$, and (suppressing the dominating measures), $\int p_t^2(d \mid z,x) = 1$ for all $(t,z,x) \in I_0 \times \text{supp}(Z) \times \text{supp}(X)$ and $\int f_t^2(x) = 1$ for all $(t,x) \in I_0 \times \text{supp}(X)$ for all $t \in I_0$. Therefore, since $\log q_t^2 \overset{\text{MAR}}{=} \log p_t^2 + \log v_t^2 + \log f_t^2$ and θ_t^* are related via the requirement that $\eta(\log q_t^2) = c'\theta_t^*$ for all $t \in I_0$, it follows

[18] For instance, let $A(X) := J'w(X)$, where $w(X)$ is a $\dim(g) \times \dim(g)$ matrix that is positive definite P_X-a.s. Since the columns of J are linearly independent P_X-a.s., so are the columns of the $\dim(\theta^*) \times \dim(\theta^*)$ matrix $J'w(X)J$. Hence, $\mathbb{E}[A(X)J]$ has column rank $\dim(\theta^*)$.

that $\nabla\eta(S_{\dot{q}}) = c'\dot{\theta}^*$, where $\nabla\eta$ is the pathwise derivative of η and $S_{\dot{q}} := S_{\dot{p}} + S_{\dot{\nu}} + S_{\dot{f}}$, with $S_{\dot{p}} := 2\dot{p}/p \in L_2(D,Z,X) \cap L_2(Z,X)^\perp$ and $S_{\dot{f}} := 2\dot{f}/f \in L_{2,0}(X)$. Hence, by (D.16),

$$\nabla\eta(S_{\dot{q}}) = -c'B^+\mathbb{E}[A\mathbb{E}(gS_{\dot{\nu}} \mid X)]. \tag{D.17}$$

To show that $\nabla\eta$ is a linear functional of $S_{\dot{q}}$, we also have to write the right-hand-side of (D.17) in terms of $S_{\dot{q}}$. To do so, we now obtain an expression for $\mathbb{E}[gS_{\dot{\nu}} \mid X]$ in terms of $S_{\dot{q}}$. Let $\mu := \mathbb{E}[g \mid Z,X]$ and[19]

$$\rho := \rho(D,Y^*,Z,X) := \frac{Dg}{\pi} - \mu[\frac{D}{\pi} - 1]. \tag{D.18}$$

Then, as shown after the proof of this lemma,

$$\begin{aligned}
&\mathbb{E}[\rho S_{\dot{p}} \mid Z,X] = 0_{\dim(g)\times 1}, \qquad \forall S_{\dot{p}} \in L_2(D,Z,X) \cap L_2(Z,X)^\perp, \\
&\mathbb{E}[\rho S_{\dot{\nu}} \mid Z,X] = \mathbb{E}[gS_{\dot{\nu}} \mid Z,X], \qquad \forall S_{\dot{\nu}} \in L_2(Y^*,Z,X) \cap L_2(X)^\perp, \\
&\mathbb{E}[\rho S_{\dot{f}} \mid Z,X] = \mu S_{\dot{f}}, \qquad \forall S_{\dot{f}} \in L_{2,0}(X).
\end{aligned} \tag{D.19}$$

Hence, since $S_{\dot{q}} := S_{\dot{p}} + S_{\dot{\nu}} + S_{\dot{f}}$, we have that

$$\mathbb{E}[\rho S_{\dot{q}} \mid Z,X] \overset{(D.19)}{=} \mathbb{E}[gS_{\dot{\nu}} \mid Z,X] + \mu S_{\dot{f}}.$$

Consequently, as $S_{\dot{f}} \in L_{2,0}(X)$ and $\mathbb{E}[\mu \mid X] = \mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$,

$$\mathbb{E}[\rho S_{\dot{q}} \mid X] = \mathbb{E}[gS_{\dot{\nu}} \mid X] + S_{\dot{f}}\mathbb{E}[\mu \mid X] = \mathbb{E}[gS_{\dot{\nu}} \mid X] \overset{(D.15)}{\in} \operatorname{span}(J) \quad P_X\text{-a.s.} \tag{D.20}$$

Collecting the restrictions on $S_{\dot{q}}$, it follows that

$$\nabla\eta(S_{\dot{q}}) \overset{(D.17)}{=} -c'B^+\mathbb{E}[A\mathbb{E}(gS_{\dot{\nu}} \mid X)] \overset{(D.20)}{=} -c'B^+\mathbb{E}[A\mathbb{E}(\rho S_{\dot{q}} \mid X)] \tag{D.21}$$

is a linear functional defined on the tangent space of score functions

$$\begin{aligned}
\dot{\mathscr{M}} := \{S_{\dot{q}} \in L_2(D,Y^*,Z,X) : S_{\dot{q}} = S_{\dot{p}} + S_{\dot{\nu}} + S_{\dot{f}}, \text{ where } S_{\dot{p}} \in L_2(D,Z,X) \cap L_2(Z,X)^\perp, \\
S_{\dot{\nu}} \in L_2(Y^*,Z,X) \cap L_2(X)^\perp,\ S_{\dot{f}} \in L_{2,0}(X), \\
\text{and } \mathbb{E}[\rho S_{\dot{q}} \mid X] \overset{P_X\text{-a.s.}}{\in} \operatorname{span}(J).\}
\end{aligned} \tag{D.22}$$

The tangent space is closed in the norm induced by the inner product $\langle\cdot,\cdot\rangle$ (Lemma D.5).

Note that

$$\begin{aligned}
\nabla\eta(S_{\dot{q}}) \overset{(D.21)}{=} -c'B^+\mathbb{E}[A\mathbb{E}(\rho S_{\dot{q}} \mid X)] &= -c'B^+\mathbb{E}[A\rho S_{\dot{q}}] && (S_{\dot{q}} \in \dot{\mathscr{M}}) \\
&= \langle -c'B^+A\rho, S_{\dot{q}}\rangle \\
&= \langle -c'B^+A\rho, \mathscr{P}_{\dot{\mathscr{M}}}(S_{\dot{q}})\rangle \\
&= \langle -\mathscr{P}_{\dot{\mathscr{M}}}(c'B^+A\rho), S_{\dot{q}}\rangle, && \text{(D.23)}
\end{aligned}$$

where the last equality is due to the fact that projection operators are self-adjoint. Since $-\mathscr{P}_{\dot{\mathscr{M}}}(c'B^+A\rho) \in \dot{\mathscr{M}}$, it follows by (D.23) and the Riesz-Fréchet theorem (Luenberger, 1969, Theorem 2, p. 109) that if $\mathbb{E}[\mathscr{P}_{\dot{\mathscr{M}}}(c'B^+A\rho)]^2 < \infty$, then $\nabla\eta$ is a bounded linear functional on the tangent space with representer $-\mathscr{P}_{\dot{\mathscr{M}}}(c'B^+A\rho)$. This implies that η is a pathwise differentiable functional, and the efficiency bound for estimating η is given by the squared operator norm of $\nabla\eta$, namely, $\mathbb{E}[\mathscr{P}_{\dot{\mathscr{M}}}(c'B^+A\rho)]^2$.

[19] Since $Dg = Dg(Y,Z,X,\theta^*)$, the definitions of ρ in (4.4) and (D.18) are equivalent.

To obtain $\mathscr{P}_{\dot{\mathscr{M}}}(c'B^+A\rho)$, we proceed as follows. Let

$$\dot{\mathscr{S}} := \{\dot{S} \in L_{2,0}(D, Y^*, Z, X) : \mathbb{E}[\rho\dot{S} \mid X] \overset{P_X\text{-a.s.}}{\in} \operatorname{span}(J)\}.$$

Then, $\dot{\mathscr{S}}$ is closed in the norm topology (same proof as for Lemma D.5), and $\dot{\mathscr{M}} \subset \dot{\mathscr{S}}$.[20] Letting $V := \mathbb{E}J'\Omega_\rho^{-1}J$, where $\Omega_\rho := \mathbb{E}[\rho\rho' \mid X]$, we have that

$$\mathscr{P}_{\dot{\mathscr{S}}}(c'B^+A\rho) \overset{\text{Lemma D.6}}{=} \rho'\Omega_\rho^{-1}JV^{-1}c.$$

But, as shown towards the end of this proof, $\rho'\Omega_\rho^{-1}JV^{-1}c \in \dot{\mathscr{M}}$. Therefore,

$$\mathscr{P}_{\dot{\mathscr{M}}}(c'B^+A\rho) = \rho'\Omega_\rho^{-1}JV^{-1}c. \qquad (\dot{\mathscr{M}} \subset \dot{\mathscr{S}})$$

Consequently, the efficiency bound for estimating η is given by[21]

$$\begin{aligned}
\mathbb{E}[\mathscr{P}_{\dot{\mathscr{M}}}(-c'B^+A\rho)]^2 &= c'V^{-1}\mathbb{E}[J\Omega_\rho^{-1}\rho\rho'\Omega_\rho^{-1}J]V^{-1}c \\
&= c'V^{-1}\mathbb{E}[J\Omega_\rho^{-1}\mathbb{E}(\rho\rho' \mid X)\Omega_\rho^{-1}J]V^{-1}c \\
&= c'V^{-1}c \qquad \text{(D.24)} \\
&< \infty. \qquad \text{(Ass. D.1(iii))}
\end{aligned}$$

The desired result follows because c is arbitrary.

It remains to verify that $\rho'\Omega_\rho^{-1}JV^{-1}c \in \dot{\mathscr{M}}$. Observe that

$$\dot{m} := \rho'\Omega_\rho^{-1}JV^{-1}c \overset{\text{(D.18)}}{=} (\frac{Dg}{\pi} - \mu[\frac{D}{\pi} - 1])'\Omega_\rho^{-1}JV^{-1}c =: \dot{m}_1 + \dot{m}_2 + \dot{m}_3,$$

where $\dot{m}_1 := -\mu'\Omega_\rho^{-1}JV^{-1}(D/\pi - 1)c$, $\dot{m}_2 := Dg'\Omega_\rho^{-1}JV^{-1}c/\pi$, and $\dot{m}_3 := 0$. Now,

$$\begin{aligned}
\mathbb{E}\dot{m}_1^2 &= c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}(\frac{D}{\pi} - 1)^2\mu\mu'\Omega_\rho^{-1}J]V^{-1}c \\
&= c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\mathbb{E}[(\frac{D}{\pi} - 1)^2 \mid Z, X]\mu\mu'\Omega_\rho^{-1}J]V^{-1}c \\
&\overset{\text{(D.10)}}{=} c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\frac{1-\pi}{\pi}\mu\mu'\Omega_\rho^{-1}J]V^{-1}c \\
&< \infty \qquad \text{(Ass. D.1(iii))}
\end{aligned}$$

and

$$\mathbb{E}[\dot{m}_1 \mid Z, X] = -\mu'\Omega_\rho^{-1}JV^{-1}c\mathbb{E}[\frac{D}{\pi} - 1 \mid Z, X] = 0 \implies \dot{m}_1 \in L_2(Z, X)^\perp.$$

[20] Let $\dot{m} \in \dot{\mathscr{M}}$. By (D.22), $\mathbb{E}[\rho\dot{m} \mid X] \overset{P_X\text{-a.s.}}{\in} \operatorname{span}(J)$ and $\dot{m} = S_{\dot{p}} + S_{\dot{\nu}} + S_{\dot{f}}$, where $S_{\dot{p}} \in L_2(D, Z, X) \cap L_2(Z, X)^\perp$, $S_{\dot{\nu}} \in L_2(Y^*, Z, X) \cap L_2(X)^\perp$, $S_{\dot{f}} \in L_{2,0}(X)$, which imply that $\mathbb{E}\dot{m} = 0$. Hence, $\dot{m} \in \dot{\mathscr{S}}$.

[21] Since $\mathbb{E}[\rho S_{\dot{p}} \mid Z, X] = 0_{\dim(g)\times 1}$ in (D.19) holds irrespective of whether π is fully known or known up to a finite dimensional parameter, and the argument leading to (D.24) does not depend on the form of $S_{\dot{p}}$, it follows that the efficiency bound for θ^* does not decrease if π is fully known, or known up to a finite dimensional parameter. Hitomi, Nishiyama, and Okui (2008) provide a comprehensive treatment of this puzzling phenomenon. A similar issue arises in estimating models with stratified samples when the stratum shares are known (Tripathi, 2011, Section 2).

Thus, $\dot{m}_1 \in L_2(D,Z,X) \cap L_2(X)^\perp$. In addition,

$$\begin{aligned}
\mathbb{E}\dot{m}_2^2 &= c'V^{-1}\mathbb{E}[\frac{DJ'\Omega_\rho^{-1}gg'\Omega_\rho^{-1}J}{\pi^2}]V^{-1}c \\
&= c'V^{-1}\mathbb{E}[\frac{J'\Omega_\rho^{-1}}{\pi^2}\mathbb{E}[Dgg' \mid Z,X]\Omega_\rho^{-1}J]V^{-1}c \\
&\overset{\text{MAR}}{=} c'V^{-1}\mathbb{E}[\frac{J'\Omega_\rho^{-1}}{\pi^2}\mathbb{E}[D \mid Z,X]\mathbb{E}[gg' \mid Z,X]\Omega_\rho^{-1}J]V^{-1}c \\
&= c'V^{-1}\mathbb{E}[\frac{J'\Omega_\rho^{-1}\mathbb{E}(gg' \mid Z,X)\Omega_\rho^{-1}J}{\pi}]V^{-1}c \\
&\overset{\text{(D.7)}}{=} c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}(\mathbb{E}[\rho\rho' \mid Z,X] + \frac{1-\pi}{\pi}\mu\mu')\Omega_\rho^{-1}J]V^{-1}c \\
&= c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\rho\rho'\Omega_\rho^{-1}J]V^{-1}c + c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\frac{1-\pi}{\pi}\mu\mu'\Omega_\rho^{-1}J]V^{-1}c \\
&= c'V^{-1}c + c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\frac{1-\pi}{\pi}\mu\mu'\Omega_\rho^{-1}J]V^{-1}c \qquad (\Omega_\rho := \mathbb{E}[\rho\rho' \mid X]) \\
&< \infty. \qquad \text{(Ass. D.1(iii))}
\end{aligned}$$

Moreover, since $\pi(Z,X) := \mathbb{E}[D \mid Z,X]$ and $\mu := \mathbb{E}[g \mid Z,X]$,

$$\mathbb{E}[\dot{m}_2 \mid Z,X] = \mathbb{E}[\frac{Dg'\Omega_\rho^{-1}JV^{-1}c}{\pi} \mid Z,X] = \frac{1}{\pi}\mathbb{E}[D \mid Z,X]\mathbb{E}[g' \mid Z,X]\Omega_\rho^{-1}JV^{-1}c = \mu'\Omega_\rho^{-1}JV^{-1}c.$$

Hence, as $\mathbb{E}[\mu \mid X] = \mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$,

$$\mathbb{E}[\dot{m}_2 \mid X] = \mathbb{E}[\mu' \mid X]\Omega_\rho^{-1}JV^{-1}c \overset{P_X\text{-a.s.}}{=} 0 \implies \dot{m}_2 \in L_2(X)^\perp.$$

Therefore, $\dot{m}_2 \in L_2(Y^*,Z,X) \cap L_2(X)^\perp$. Furthermore,

$$\mathbb{E}[\rho\dot{m} \mid Z,X] = \mathbb{E}[\rho\rho'\Omega_\rho^{-1}JV^{-1}c \mid Z,X] = \mathbb{E}[\rho\rho' \mid Z,X]\Omega_\rho^{-1}JV^{-1}c.$$

Hence, recalling that $\Omega_\rho := \mathbb{E}[\rho\rho' \mid X]$, we obtain that $\mathbb{E}[\rho\dot{m} \mid X] = JV^{-1}c \in \text{span}(J)$. It follows that $\rho'\Omega_\rho^{-1}JV^{-1}c \in \dot{\mathscr{M}}$. □

Lemma D.5. *Under Assumptions* 3.1 *and* D.1(i, ii)*, $\dot{\mathscr{M}}$ is closed.*

Proof of Lemma D.5. The tangent score vectors $S_{\dot{p}} \in L_2(D,Z,X) \cap L_2(Z,X)^\perp =: \dot{\mathscr{M}}_1$, $S_{\dot{\nu}} \in L_2(Y^*,Z,X) \cap L_2(X)^\perp =: \dot{\mathscr{M}}_2$, and $S_{\dot{f}} \in L_{2,0}(X) =: \dot{\mathscr{M}}_3$ are pairwise orthogonal. Indeed,

$$\begin{aligned}
\mathbb{E}[S_{\dot{p}}S_{\dot{\nu}} \mid Z,X] &\overset{\text{MAR}}{=} \mathbb{E}[S_{\dot{p}} \mid Z,X]\mathbb{E}[S_{\dot{\nu}} \mid Z,X] \overset{S_{\dot{p}}\in L_2(Z,X)^\perp}{=} 0 \implies S_{\dot{p}} \perp S_{\dot{\nu}} \\
\mathbb{E}[S_{\dot{p}}S_{\dot{f}} \mid Z,X] &= S_{\dot{f}}\mathbb{E}[S_{\dot{p}} \mid Z,X] \overset{S_{\dot{p}}\in L_2(Z,X)^\perp}{=} 0 \implies S_{\dot{p}} \perp S_{\dot{f}} \\
\mathbb{E}[S_{\dot{\nu}}S_{\dot{f}} \mid X] &= S_{\dot{f}}\mathbb{E}[S_{\dot{\nu}} \mid X] \overset{S_{\dot{\nu}}\in L_2(X)^\perp}{=} 0 \implies S_{\dot{\nu}} \perp S_{\dot{f}}.
\end{aligned}$$

Pairwise orthogonality of $\dot{\mathscr{M}}_1, \dot{\mathscr{M}}_2, \dot{\mathscr{M}}_3$ is used to show that $\text{cl}(\dot{\mathscr{M}}) \subset \dot{\mathscr{M}}$. Let $\dot{m} \in \text{cl}(\dot{\mathscr{M}})$. Then, there exists a sequence $(\dot{m}_j)_{j\in\mathbb{N}} \subset \dot{\mathscr{M}}$ such that $\lim_{j\to\infty}\|\dot{m}_j - \dot{m}\|_2 = 0$. It remains to prove that $\dot{m} \in \dot{\mathscr{M}}$. Since $\dot{m}_j \in \dot{\mathscr{M}}$ for each j, by (D.22) we have that $\dot{m}_j = \dot{m}_{j1} + \dot{m}_{j2} + \dot{m}_{j3}$, where $(\dot{m}_{j1}, \dot{m}_{j2}, \dot{m}_{j3}) \in \dot{\mathscr{M}}_1 \times \dot{\mathscr{M}}_2 \times \dot{\mathscr{M}}_3$ and $\mathbb{E}[\rho\dot{m}_j \mid X] \overset{P_X\text{-a.s.}}{\in} \text{span}(J)$. Following the argument in Halmos (1951, Theorem 6, p. 25), pairwise orthogonality of $\dot{\mathscr{M}}_1, \dot{\mathscr{M}}_2, \dot{\mathscr{M}}_3$ implies that the sequences $(\dot{m}_{j1})_{j\in\mathbb{N}} \subset \dot{\mathscr{M}}_1$, $(\dot{m}_{j2})_{j\in\mathbb{N}} \subset \dot{\mathscr{M}}_2$, and $(\dot{m}_{j3})_{j\in\mathbb{N}} \subset \dot{\mathscr{M}}_3$ are Cauchy. Therefore, as $\dot{\mathscr{M}}_1, \dot{\mathscr{M}}_2, \dot{\mathscr{M}}_3$ are

$\|\cdot\|_2$-closed — because L_2-spaces and the orthogonal complements of their linear subspaces are closed — we have that $\dot{m} = \dot{m}_1 + \dot{m}_2 + \dot{m}_3$ for some $(\dot{m}_1, \dot{m}_2, \dot{m}_3) \in \dot{\mathscr{M}}_1 \times \dot{\mathscr{M}}_2 \times \dot{\mathscr{M}}_3$. Hence, $\dot{m} \in \dot{\mathscr{M}}$ follows if $\mathbb{E}[\rho\dot{m} \mid X] \overset{P_X\text{-a.s.}}{\in} \operatorname{span}(J)$.

We show that $\mathbb{E}[\rho\dot{m} \mid X] \overset{P_X\text{-a.s.}}{\in} \operatorname{span}(J)$ by demonstrating that $\lim_{j\to\infty} \|\mathbb{E}[\rho\dot{m}_j \mid X] - \mathbb{E}[\rho\dot{m} \mid X]\|_2 = 0$ (recall that $\|b\|_2 := \sqrt{\mathbb{E}b'b}$ if b is a random vector). Indeed, $\lim_{j\to\infty} \|\mathbb{E}[\rho\dot{m}_j \mid X] - \mathbb{E}[\rho\dot{m} \mid X]\|_2 = 0$ implies that, P_X-a.s., $(\mathbb{E}[\rho\dot{m}_j \mid X])_{j\in\mathbb{N}}$ is a convergent sequence in $\operatorname{span}(J)$ because $\mathbb{E}[\rho\dot{m}_j \mid X] \overset{P_X\text{-a.s.}}{\in} \operatorname{span}(J)$ for each j. Therefore, its limit $\mathbb{E}[\rho\dot{m} \mid X] \overset{P_X\text{-a.s.}}{\in} \operatorname{span}(J)$ because $\operatorname{span}(J)$ is $\|\cdot\|_2$-closed (cf. the discussion after Assumption D.1).

To show that $\lim_{j\to\infty} \|\mathbb{E}[\rho\dot{m}_j \mid X] - \mathbb{E}[\rho\dot{m} \mid X]\|_2 = 0$, we proceed as follows. Let $\rho^{(k)}$ denote the k^{th} coordinate of ρ. Then,

$$
\begin{aligned}
\|\mathbb{E}[\rho(\dot{m}_j - \dot{m}) \mid X]\|_2^2 &= \mathbb{E}\sum_{k=1}^{\dim(g)} (\mathbb{E}[\rho^{(k)}(\dot{m}_j - \dot{m}) \mid X])^2 && (\|b\|_2 := \sqrt{\mathbb{E}b'b}, \dim(\rho) = \dim(g)) \\
&\le \mathbb{E}\sum_{k=1}^{\dim(g)} \mathbb{E}[(\rho^{(k)})^2 \mid X]\mathbb{E}[(\dot{m}_j - \dot{m})^2 \mid X] && \text{(cond. Cauchy-Schwarz)} \\
&= \mathbb{E}(\mathbb{E}[\rho'\rho \mid X]\mathbb{E}[(\dot{m}_j - \dot{m})^2 \mid X]) \\
&\overset{\text{(D.13)}}{\le} \frac{\|\sigma_g^2\|_\infty}{\inf \pi}\mathbb{E}(\dot{m}_j - \dot{m})^2 && \text{(Ass. D.1(i,ii))} \\
&\xrightarrow[j\to\infty]{} 0
\end{aligned}
$$

because $\lim_{j\to\infty} \|\dot{m}_j - \dot{m}\|_2 = 0$. The desired result follows. □

Proof of (D.19). Observe that

$$
\begin{aligned}
\mathbb{E}[\rho S_{\dot{p}} \mid Z,X] &= \mathbb{E}[\frac{Dg}{\pi}S_{\dot{p}} \mid Z,X] - \mathbb{E}[\mu(\frac{D}{\pi} - 1)S_{\dot{p}} \mid Z,X] && (S_{\dot{p}} \in L_2(D,Z,X) \cap L_2(Z,X)^\perp) \\
&= \frac{1}{\pi}\mathbb{E}[DgS_{\dot{p}} \mid Z,X] - \mu\mathbb{E}[(\frac{D}{\pi} - 1)S_{\dot{p}} \mid Z,X] \\
&\overset{\text{MAR}}{=} \frac{1}{\pi}\mathbb{E}[DS_{\dot{p}} \mid Z,X]\mathbb{E}[g \mid Z,X] - \mu(\frac{1}{\pi}\mathbb{E}[DS_{\dot{p}} \mid Z,X] - \mathbb{E}[S_{\dot{p}} \mid Z,X]) \\
&= \frac{\mu}{\pi}\mathbb{E}[DS_{\dot{p}} \mid Z,X] - \frac{\mu}{\pi}\mathbb{E}[DS_{\dot{p}} \mid Z,X] && (\mu := \mathbb{E}[g \mid Z,X], S_{\dot{p}} \in L_2(Z,X)^\perp) \\
&= 0_{\dim(g)\times 1}.
\end{aligned}
$$

Next,

$$
\begin{aligned}
\mathbb{E}[\rho S_{\dot{\nu}} \mid Z,X] &= \mathbb{E}[\frac{Dg}{\pi}S_{\dot{\nu}} \mid Z,X] - \mathbb{E}[\mu(\frac{D}{\pi} - 1)S_{\dot{\nu}} \mid Z,X] && (S_{\dot{\nu}} \in L_2(Y^*,Z,X) \cap L_2(X)^\perp) \\
&= \frac{1}{\pi}\mathbb{E}[DgS_{\dot{\nu}} \mid Z,X] - \mu\mathbb{E}[(\frac{D}{\pi} - 1)S_{\dot{\nu}} \mid Z,X] \\
&\overset{\text{MAR}}{=} \frac{1}{\pi}\mathbb{E}[D \mid Z,X]\mathbb{E}[gS_{\dot{\nu}} \mid Z,X] - \mu\mathbb{E}[\frac{D}{\pi} - 1 \mid Z,X]\mathbb{E}[S_{\dot{\nu}} \mid Z,X] \\
&= \mathbb{E}[gS_{\dot{\nu}} \mid Z,X]. && (\mathbb{E}[\tfrac{D}{\pi} - 1 \mid Z,X] = 0)
\end{aligned}
$$

Finally,

$$\begin{aligned}
\mathbb{E}[\rho S_{\dot f} \mid Z,X] &= \mathbb{E}[\frac{Dg}{\pi} S_{\dot f} \mid Z,X] - \mathbb{E}[\mu(\frac{D}{\pi}-1) S_{\dot f} \mid Z,X] && (S_{\dot f} \in L_{2,0}(X)) \\
&= \frac{S_{\dot f}}{\pi}\mathbb{E}[Dg \mid Z,X] - \mu S_{\dot f}\mathbb{E}[\frac{D}{\pi}-1 \mid Z,X] \\
&\overset{\text{MAR}}{=} \frac{S_{\dot f}}{\pi}\mathbb{E}[D \mid Z,X]\mathbb{E}[g \mid Z,X] && (\mathbb{E}[\tfrac{D}{\pi}-1 \mid Z,X]=0) \\
&= S_{\dot f}\mu. && (\mu := \mathbb{E}[g \mid Z,X])
\end{aligned}$$

The desired result follows. □

Lemma D.6. $\mathscr{P}_{\dot{\mathscr{S}}}(c'B^+A\rho) = \rho'\Omega_\rho^{-1}JV^{-1}c$.

Proof of Lemma D.6. Observe that $\dot m := \dot m(D,Y^*,Z,X) := \rho'\Omega_\rho^{-1}JV^{-1}c \in \dot{\mathscr{S}}$ because

$$\begin{aligned}
\mathbb{E}\dot m^2 = c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\rho\rho'\Omega_\rho^{-1}J]V^{-1}c = c'V^{-1}c &\overset{\text{Ass. D.1(iii)}}{<} \infty, \\
\mathbb{E}[\dot m \mid X] = \mathbb{E}[\rho' \mid X]\Omega_\rho^{-1}JV^{-1}c &\overset{(4.5)}{=} 0 \implies \mathbb{E}\dot m = 0, \\
\mathbb{E}[\rho\dot m \mid X] = \mathbb{E}[\rho\rho' \mid X]\Omega_\rho^{-1}JV^{-1}c &= JV^{-1}c \in \operatorname{span}(J).
\end{aligned}$$

Next, let $\dot S \in \dot{\mathscr{S}}$. Then, since $\mathbb{E}[\rho\dot S \mid X] \in \operatorname{span}(J)$, which implies that $\mathbb{E}[\rho\dot S \mid X] = J\alpha$ for some $\alpha \in \mathbb{R}^{\dim(\theta^*)}$, we have that

$$\begin{aligned}
\langle c'B^+A\rho - \dot m, \dot S\rangle &= c'B^+\mathbb{E}[A\rho\dot S] - \mathbb{E}[\rho'\Omega_\rho^{-1}JV^{-1}c\dot S] && (\text{defn. } \dot m) \\
&= c'B^+\mathbb{E}[A\rho\dot S] - c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\rho\dot S] \\
&= c'B^+\mathbb{E}[A\mathbb{E}(\rho\dot S \mid X)] - c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\mathbb{E}(\rho\dot S \mid X)] \\
&= c'B^+\mathbb{E}[AJ]\alpha - c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}J]\alpha \\
&= c'B^+\mathbb{E}[AJ]\alpha - c'\alpha && (V := \mathbb{E}[J\Omega_\rho^{-1}J]) \\
&= c'B^+B\alpha - c'\alpha && (B := \mathbb{E}[AJ]) \\
&= c'\alpha - c'\alpha && (B \text{ full column rank}) \\
&= 0.
\end{aligned}$$

The desired result follows because $\dot S$ was arbitrary. □

Proof of (4.6). For a vector of real-valued functions $\varpi := (\varpi^{(1)},\ldots,\varpi^{(\dim(\theta^*))}) \in L_2(Z,X)^{\dim(\theta^*)}$, define the $\dim(g)\times\dim(\theta^*)$ matrix $\mathscr{J}_\varpi$ whose k^{th} column is $J_k + \mathbb{E}[\frac{Dg_{\text{obs}}}{\pi^2}\varpi^{(k)} \mid X]$, where J_k is the k^{th} column of J. Then, by Ai and Chen (2003, Theorems 4.1 and 6.1),

$$\text{l.b.}_{\text{VS}}(\theta^*) = (\mathbb{E}\mathscr{J}_{\varpi_*}'\Sigma^{-1}\mathscr{J}_{\varpi_*})^{-1}, \tag{D.25}$$

where $\Sigma := \mathbb{E}[\frac{Dg_{\text{obs}}g_{\text{obs}}'}{\pi^2} \mid X]$ and $\varpi_* \in L_2(Z,X)^{\dim(\theta^*)}$ is such that $\mathbb{E}\mathscr{J}_{\varpi_*}'\Sigma^{-1}\mathscr{J}_{\varpi_*} \le_L \mathbb{E}\mathscr{J}_\varpi'\Sigma^{-1}\mathscr{J}_\varpi$ for all $\varpi \in L_2(Z,X)^{\dim(\theta^*)}$. Since $Dg = Dg_{\text{obs}} \implies \Sigma = \mathbb{E}[\frac{Dgg'}{\pi^2} \mid X]$, it follows from (D.12) that

$$\Sigma \overset{P_X\text{-a.s.}}{=} \Omega_\rho + \mathbb{E}[\frac{1-\pi}{\pi}\mu\mu' \mid X]. \tag{D.26}$$

Consequently, $\Sigma \overset{P_X\text{-a.s.}}{\ge_L} \Omega_\rho \iff \Omega_\rho^{-1} \overset{P_X\text{-a.s.}}{\ge_L} \Sigma^{-1}$ implying that

$$\mathbb{E}J'\Omega_\rho^{-1}J \ge_L \mathbb{E}J'\Sigma^{-1}J \ge_L \mathbb{E}\mathscr{J}_{\varpi_*}'\Sigma^{-1}\mathscr{J}_{\varpi_*}, \tag{D.27}$$

where the 2nd inequality in (D.27) is due to the definition of ϖ_*. Therefore, we have that

$$(\mathbb{E}J'\Omega_\rho^{-1}J)^{-1} \leq_L (\mathbb{E}\mathscr{J}'_{\varpi_*}\Sigma^{-1}\mathscr{J}_{\varpi_*})^{-1} \iff \text{l.b.}(\theta^*) \leq_L \text{l.b.}_{\text{VS}}(\theta^*). \qquad \square$$

Proof of Lemma 4.2. If $Z = \vec{\varnothing}$, then the propensity score is a function of X alone so that $\varpi_*\big|_{Z=\vec{\varnothing}} \in L_2(X)^{\dim(\theta^*)}$. Hence, $\mathscr{J}_{\varpi_*} \overset{Z=\vec{\varnothing}}{=} J$ (from the definition of $\mathscr{J}_\varpi$ and (3.1)), and $\mu := \mathbb{E}[g \mid Z,X] \overset{Z=\vec{\varnothing}}{=} \mathbb{E}[g \mid X] \overset{(2.1)}{=} 0_{\dim(g)\times 1}$ P_X-a.s., which implies that $\Sigma \overset{(D.26)}{=} \Omega_\rho$ P_X-a.s. Consequently, $Z = \vec{\varnothing} \implies \mu \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \implies \text{l.b.}(\theta^*) = \text{l.b.}_{\text{VS}}(\theta^*)$, i.e., we have shown sufficiency.

Next, to prove that these conditions are also necessary, we want to show that

$$Z \neq \vec{\varnothing} \implies P_{Z,X}(\mu \neq 0_{\dim(g)}) > 0 \implies \text{l.b.}(\theta^*) <_L \text{l.b.}_{\text{VS}}(\theta^*).$$

To show this, begin by observing that

$$\begin{aligned}
\Sigma \overset{P_X\text{-a.s.}}{=} \Omega_\rho &\overset{(D.26)}{\iff} \mathbb{E}[\frac{1-\pi}{\pi}\mu\mu' \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \\
&\iff \mu \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1} \\
&\overset{\mu:=\mathbb{E}[g|Z,X]}{\iff} \mathbb{E}[g \mid Z,X] \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1} \\
&\iff Z = \vec{\varnothing},
\end{aligned} \tag{D.28}$$

where the second equivalence is true because $P_{Z,X}(\pi \neq 0) = 1$ by Assumption D.1(i), and the last equivalence is true because $Z = \vec{\varnothing} \implies \mathbb{E}[g \mid Z = \vec{\varnothing}, X] \overset{(2.1)}{=} 0_{\dim(g)\times 1}$ (P_X-a.s.), and

$$\begin{aligned}
\mathbb{E}[g \mid Z,X] \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1} &\implies \mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1} \quad \& \quad \mathbb{E}[g \mid Z] \overset{P_Z\text{-a.s.}}{=} 0_{\dim(g)\times 1} \\
&\implies Z = \vec{\varnothing} \text{ because otherwise } Z \text{ would be exogenous w.r.t. } g.
\end{aligned}$$

Hence,

$$\text{(D.26)\&(D.28)} \implies Z \neq \vec{\varnothing} \implies P_{Z,X}(\mu \neq 0_{\dim(g)}) > 0 \implies \Sigma \overset{P_X\text{-a.s.}}{>}_L \Omega_\rho.$$

But this implies that the 1st inequality in (D.27) is strict, i.e., we have that

$$Z \neq \vec{\varnothing} \implies P_{Z,X}(\mu \neq 0_{\dim(g)}) > 0 \implies \mathbb{E}J'\Omega_\rho^{-1}J >_L \mathbb{E}J'\Sigma^{-1}J \geq_L \mathbb{E}\mathscr{J}'_{\varpi_*}\Sigma^{-1}\mathscr{J}_{\varpi_*},$$

from which it follows that

$$\begin{aligned}
Z \neq \vec{\varnothing} \implies P_{Z,X}(\mu \neq 0_{\dim(g)}) > 0 &\implies (\mathbb{E}J'\Omega_\rho^{-1}J)^{-1} <_L (\mathbb{E}\mathscr{J}'_{\varpi_*}\Sigma^{-1}\mathscr{J}_{\varpi_*})^{-1} \\
&\implies \text{l.b.}(\theta^*) <_L \text{l.b.}_{\text{VS}}(\theta^*).
\end{aligned}$$

Hence, these conditions are also necessary. $\square$

D.3.1 Details for Example 4.1

In this example, we consider a regression model where the outcomes may be missing and there are no nonmissing endogenous variables (included or excluded) and no excluded instruments, i.e., here $Y^* = \alpha_0^* + X_{\text{in}}'\beta_0^* + U$ with $\mathbb{E}[U \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0$. We explain in detail how to use LS to estimate the structural parameters $\theta_0^* := (\alpha_0^*, \beta_0^*)_{p_0\times 1}$, where $p_0 := 1 + \dim(X_{\text{in}})$, when the missing Y^* are imputed using X_{in}, i.e., in this example, the propensity score is a function of X_{in} alone, namely, $\pi := \pi(X_{\text{in}})$.

1. To motivate $\hat{\theta}_{0\text{VS}} := (\hat{\alpha}_{0\text{VS}}, \hat{\beta}_{0\text{VS}})$, the LS estimator in the validation sample, note that if Y^* were nonmissing then the LS estimator of $(\alpha_0^*, \beta_0^*,)$ solves the optimization problem $\min_{\alpha,\beta} n^{-1}\sum_{i=1}^n (Y_i^* - \alpha - X_{\text{in},i}'\beta)^2$, whose FOC at the population level is the UCMR

$$\mathbb{E}[\begin{bmatrix}1\\ X_{\text{in}}\end{bmatrix}(Y^* - \alpha_0^* - X_{\text{in}}'\beta_0^*)] = 0_{p_0\times 1}. \tag{D.29}$$

However, as Y^* is missing, (D.29) cannot be used to estimate (α_0^*, β_0^*). Instead, if estimation is to be done using the validation sample alone, then, as justified subsequently, the appropriate UCMR to use is the IPW version of (D.29), namely,[22]

$$\mathbb{E}[D\begin{bmatrix}1\\ X_{\text{in}}\end{bmatrix}(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*)] = 0_{p_0\times 1}. \tag{D.30}$$

But (D.30) is the population level FOC for

$$(\hat{\alpha}_{0\text{VS}}, \hat{\beta}_{0\text{VS}}) := \operatorname*{argmin}_{\alpha,\beta} n^{-1}\sum_{i=1}^n D_i(Y_i - \alpha - X_{\text{in},i}'\beta)^2. \tag{D.31}$$

To justify why (D.30) is true, note that

$$\begin{aligned}
\mathbb{E}[U \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0 &\overset{\text{MAR}}{\implies} \mathbb{E}[\frac{D}{\pi}(Y^* - \alpha_0^* - X_{\text{in}}'\beta_0^*) \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0 \\
&\iff \mathbb{E}[\frac{D}{\pi}(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*) \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0 \qquad (DY = DY^*) \\
&\iff \mathbb{E}[D(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*) \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0 \qquad (\pi := \pi(X_{\text{in}})) \\
&\implies \mathbb{E}[D\begin{bmatrix}1\\ X_{\text{in}}\end{bmatrix}(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*)] = 0_{p_0\times 1}.
\end{aligned}$$

Therefore, (D.30) is true.

2. To motivate $\hat{\theta}_{0\text{LS}} := (\hat{\alpha}_{0\text{LS}}, \hat{\beta}_{0\text{LS}})$, the LS estimator in the observed sample when the missing outcomes are imputed, recall that if Y^* were nonmissing then the LS estimator of (α_0^*, β_0^*) solves the population level FOC (D.29). As justified subsequently, under MAR, (D.29) is equivalent to the following UCMR based on the observed data with imputed outcome $Y_{\text{imp}} := \mathbb{E}[Y^* \mid X_{\text{in}}, D = 1] \overset{\text{MAR}}{=} \mathbb{E}[Y^* \mid X_{\text{in}}] = \alpha_0^* + X_{\text{in}}'\beta_0^*$:[23]

$$\mathbb{E}[D\begin{bmatrix}1\\ X_{\text{in}}\end{bmatrix}(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*)] = 0_{p_0\times 1}. \tag{D.32}$$

But (D.32) is the population level FOC for

$$(\hat{\alpha}_{0\text{LS}}, \hat{\beta}_{0\text{LS}}) := \operatorname*{argmin}_{\alpha,\beta} n^{-1}\sum_{i=1}^n D_i(Y_i - \alpha - X_{\text{in},i}'\beta)^2. \tag{D.33}$$

To justify (D.32), note that by iterated expectations,

$$\text{LHS of (D.29)} = \mathbb{E}[\begin{bmatrix}1\\ X_{\text{in}}\end{bmatrix}\mathbb{E}(Y^* - \alpha_0^* - X_{\text{in}}'\beta_0^* \mid X_{\text{in}})].$$

[22] As shown in the justification for (D.30), the propensity score does not appear in (D.30) because in this example it only depends on X_{in}.

[23] The imputed outcome does not appear in (D.32) because the imputation $Y_{\text{imp}} = \alpha_0^* + X_{\text{in}}'\beta_0^*$ does not provide any information about missing outcomes beyond what is available from the regression model itself.

Now,

$$\begin{aligned}\mathbb{E}(Y^* - \alpha_0^* - X_{\text{in}}'\beta_0^* \mid X_{\text{in}}) &= \mathbb{E}(Y^* \mid X_{\text{in}}) - \alpha_0^* - X_{\text{in}}'\beta_0^* \\ &\overset{\text{MAR}}{=} \mathbb{E}(Y^* \mid X_{\text{in}}, D) - \alpha_0^* - X_{\text{in}}'\beta_0^* \\ &= D[\mathbb{E}(Y^* \mid X_{\text{in}}, D=1) - \alpha_0^* - X_{\text{in}}'\beta_0^*] \\ &\quad + (1-D)[\mathbb{E}(Y^* \mid X_{\text{in}}, D=0) - \alpha_0^* - X_{\text{in}}'\beta_0^*] \\ &= D\mathbb{E}(Y - \alpha_0^* - X_{\text{in}}'\beta_0^* \mid X_{\text{in}}, D=1) \\ &\quad + (1-D)[Y_{\text{imp}} - \alpha_0^* - X_{\text{in}}'\beta_0^*],\end{aligned}$$

because $\mathbb{E}(Y^* \mid X_{\text{in}}, D=0) \overset{\text{MAR}}{=} \mathbb{E}(Y^* \mid X_{\text{in}}, D=1) =: Y_{\text{imp}}$. Therefore, as $Y_{\text{imp}} = \alpha_0^* + X_{\text{in}}'\beta_0^*$, we have that

$$\begin{aligned}\mathbb{E}(Y^* - \alpha_0^* - X_{\text{in}}'\beta_0^* \mid X_{\text{in}}) &= D\mathbb{E}(Y - \alpha_0^* - X_{\text{in}}'\beta_0^* \mid X_{\text{in}}, D=1) \\ &= D\mathbb{E}\left(\frac{D(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*)}{\pi} \mid X_{\text{in}}\right).\end{aligned}$$

It follows that

$$\begin{aligned}\text{LHS of (D.29)} &\overset{\text{MAR}}{=} \mathbb{E}[D\begin{bmatrix}1\\X_{\text{in}}\end{bmatrix}\mathbb{E}\left(\frac{D(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*)}{\pi} \mid X_{\text{in}}\right)] \\ &= \mathbb{E}[\mathbb{E}(D \mid X_{\text{in}})\begin{bmatrix}1\\X_{\text{in}}\end{bmatrix}\mathbb{E}\left(\frac{D(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*)}{\pi} \mid X_{\text{in}}\right)] \\ &= \mathbb{E}[\pi\begin{bmatrix}1\\X_{\text{in}}\end{bmatrix}\mathbb{E}\left(\frac{D(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*)}{\pi} \mid X_{\text{in}}\right)] \qquad (\pi = \mathbb{E}(D \mid X_{\text{in}})) \\ &= \mathbb{E}[D\begin{bmatrix}1\\X_{\text{in}}\end{bmatrix}(Y - \alpha_0^* - X_{\text{in}}'\beta_0^*)] \\ &= \text{LHS of (D.32)}.\end{aligned}$$

But, $\mathbb{E}[U \mid X_{\text{in}}] \overset{P_{X_{\text{in}}}\text{-a.s.}}{=} 0$ implies that (D.29) holds. Therefore, (D.32) is also true.

3. Since (D.31) and (D.33) are identical, it follows from their FOC that

$$\hat{\theta}_{\text{0VS}} = \hat{\theta}_{\text{0LS}} = \Big(\sum_{i=1}^n D_i \tilde{X}_{\text{in},i}\tilde{X}_{\text{in},i}'\Big)^{-1}\Big(\sum_{i=1}^n D_i \tilde{X}_{\text{in},i} Y_i\Big). \qquad (\tilde{X}_{\text{in},i} := (1 X_{\text{in},i})_{p_0 \times 1})$$

D.3.2 Details for Example 4.2

In this example, we consider a regression model where the outcomes may be missing but we allow for nonmissing (included and excluded) endogenous variables and excluded instruments, i.e., now $Y^* = \alpha^* + X_{\text{in}}'\beta^* + Z_{\text{in}}'\gamma^* + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] \overset{P_X\text{-a.s.}}{=} 0$. We explain in detail how to use 2SLS to estimate the structural parameters $\theta^* := (\alpha^*, \beta^*, \gamma^*)_{p\times 1}$, where $p := 1 + \dim(X_{\text{in}}) + \dim(Z_{\text{in}})$, when the missing outcomes are imputed using the nonmissing (included and excluded) endogenous variables and instruments, i.e., unlike Example 4.1, the propensity score is now a function of (Z, X).

1. To motivate $\hat{\theta}_{\text{VS}} := (\hat{\alpha}_{\text{VS}}, \hat{\beta}_{\text{VS}}, \hat{\gamma}_{\text{VS}})$, the 2SLS estimator in the validation sample, note that if Y^* were nonmissing then the 2SLS estimator of $(\alpha^*, \beta^*, \gamma^*)$ solves the optimization problem $\min_{\alpha,\beta,\gamma} n^{-1}\sum_{i=1}^n (Y_i^* - \alpha - X_{\text{in},i}'\beta - \hat{Z}_{\text{in},i}'\gamma)^2$, whose FOC at the population level is the UCMR

$$\mathbb{E}[\begin{bmatrix}1\\X_{\text{in}}\\\text{BLP}(Z_{\text{in}} \mid X)\end{bmatrix}(Y^* - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^*)] = 0_{p\times 1}. \qquad \text{(D.34)}$$

However, as Y^* is missing, (D.34) cannot be used to estimate $(\alpha^*, \beta^*, \gamma^*)$. Instead, if estimation is to be done using the validation sample alone, then, as justified subsequently, the appropriate UCMR to use is the IPW version of (D.34), namely,

$$\mathbb{E}[\frac{D}{\pi}\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}} \mid X)\end{bmatrix}(Y-\alpha^*-X_{\text{in}}'\beta^*-\text{BLP}(Z_{\text{in}} \mid X)'\gamma^*)]=0_{p\times 1}. \tag{D.35}$$

But (D.35) is the population level FOC for the IPW 2SLS estimator in the validation sample, i.e.,

$$(\hat{\alpha}_{\text{VS}}, \hat{\beta}_{\text{VS}}, \hat{\gamma}_{\text{VS}}) := \underset{\alpha,\beta,\gamma}{\text{argmin}}\, n^{-1}\sum_{i=1}^{n}\frac{D_i}{\hat{\pi}_i}(Y_i-\alpha-X_{\text{in},i}'\beta-\hat{Z}_{\text{in},i}'\gamma)^2. \tag{D.36}$$

To justify why (D.35) is true, let $V := Z_{\text{in}} - \text{BLP}(Z_{\text{in}} \mid X)$ and note that

$$\begin{aligned}
&\mathbb{E}[\frac{D}{\pi}\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}} \mid X)\end{bmatrix}(Y-\alpha^*-X_{\text{in}}'\beta^*-\text{BLP}(Z_{\text{in}} \mid X)'\gamma^*)]\\
&=\mathbb{E}[\frac{D}{\pi}\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}} \mid X)\end{bmatrix}(Y-\alpha^*-X_{\text{in}}'\beta^*-Z_{\text{in}}'\gamma^*)]+\mathbb{E}[\frac{D}{\pi}\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}} \mid X)\end{bmatrix}V'\gamma^*]\\
&=\mathbb{E}[\frac{D}{\pi}\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}} \mid X)\end{bmatrix}(Y-\alpha^*-X_{\text{in}}'\beta^*-Z_{\text{in}}'\gamma^*)],
\end{aligned}$$

because[24]

$$\begin{aligned}
\mathbb{E}[\frac{D}{\pi}V] &= 0_{\dim(Z_{\text{in}})\times 1}\\
\mathbb{E}[\frac{D}{\pi}X_{\text{in}}V'] &= 0_{\dim(X_{\text{in}})\times\dim(Z_{\text{in}})}\\
\mathbb{E}[\frac{D}{\pi}\text{BLP}(Z_{\text{in}} \mid X)V'] &= 0_{\dim(Z_{\text{in}})\times\dim(Z_{\text{in}})}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{E}[\varepsilon \mid X] \overset{P_X\text{-a.s.}}{=} 0 &\overset{\text{MAR}}{\Longrightarrow} \mathbb{E}[\frac{D}{\pi}(Y^*-\alpha^*-X_{\text{in}}'\beta^*-Z_{\text{in}}'\gamma^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0\\
&\iff \mathbb{E}[\frac{D}{\pi}(Y-\alpha^*-X_{\text{in}}'\beta^*-Z_{\text{in}}'\gamma^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0 \qquad (DY=DY^*)\\
&\Longrightarrow \mathbb{E}[\frac{D}{\pi}\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}} \mid X)\end{bmatrix}(Y-\alpha^*-X_{\text{in}}'\beta^*-Z_{\text{in}}'\gamma^*)]=0_{p\times 1}.
\end{aligned}$$

Therefore, (D.35) is true.

[24] Since V is a function of (Z_{in}, X) alone, $\mathbb{E}[\frac{D}{\pi}X_{\text{in}}V' \mid Z, X] \overset{P_{Z,X}\text{-a.s.}}{=} \frac{X_{\text{in}}V'}{\pi}\mathbb{E}[D \mid Z, X] = X_{\text{in}}V'$. The orthogonality property of BLP residuals V then implies that $\mathbb{E}[\frac{D}{\pi}X_{\text{in}}V'] = \mathbb{E}[X_{\text{in}}V'] = 0_{\dim(X_{\text{in}})\times\dim(Z_{\text{in}})}$. Similarly, it can be shown that $\mathbb{E}[\frac{D}{\pi}V] = 0_{\dim(Z_{\text{in}})\times 1}$ and $\mathbb{E}[\frac{D}{\pi}\text{BLP}(Z_{\text{in}} \mid X)V'] = 0_{\dim(Z_{\text{in}})\times\dim(Z_{\text{in}})}$.

2. To motivate $\hat{\theta}_{2SLS} := (\hat{\alpha}_{2SLS}, \hat{\beta}_{2SLS}, \hat{\gamma}_{2SLS})$, the 2SLS estimator in the observed sample when the missing outcomes are imputed, recall that if Y^* were nonmissing then the 2SLS estimator of $(\alpha^*, \beta^*, \gamma^*)$ solves the population level FOC (D.34). As justified subsequently, under MAR, (D.34) is equivalent to the following UCMR based on the observed data with imputed outcome $Y_{\text{imp}} := \mathbb{E}[Y^* \mid Z, X, D = 1] \overset{\text{MAR}}{=} \mathbb{E}[Y^* \mid Z, X] = \alpha^* + X_{\text{in}}'\beta^* + Z_{\text{in}}'\gamma^* + \mu$:[25]

$$\begin{aligned}
&\mathbb{E}[D \begin{bmatrix} 1 \\ X_{\text{in}} \\ \text{BLP}(Z_{\text{in}} \mid X) \end{bmatrix} (Y - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^*)] \\
&+ \mathbb{E}[(1-D) \begin{bmatrix} 1 \\ X_{\text{in}} \\ \text{BLP}(Z_{\text{in}} \mid X) \end{bmatrix} (Y_{\text{imp}} - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^*)] \\
&= 0_{p\times 1}.
\end{aligned} \tag{D.37}$$

But (D.37) is the population level FOC for the 2SLS estimator in the observed sample with imputed missing outcomes, namely,

$$\begin{aligned}
(\hat{\alpha}_{2SLS}, \hat{\beta}_{2SLS}, \hat{\gamma}_{2SLS}) := \underset{\alpha,\beta,\gamma}{\text{argmin}}\, n^{-1} \sum_{i=1}^{n} [D_i(Y_i - \alpha - X_{\text{in},i}'\beta - \hat{Z}_{\text{in},i}'\gamma)^2 \\
+ (1 - D_i)(\hat{Y}_{\text{imp},i} - \alpha - X_{\text{in},i}'\beta - \hat{Z}_{\text{in},i}'\gamma)^2].
\end{aligned} \tag{D.38}$$

To justify (D.37), note that by iterated expectations,

$$\text{LHS of (D.34)} = \mathbb{E}[\begin{bmatrix} 1 \\ X_{\text{in}} \\ \text{BLP}(Z_{\text{in}} \mid X) \end{bmatrix} \mathbb{E}(Y^* - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^* \mid Z, X)].$$

Now,

$$\begin{aligned}
&\mathbb{E}(Y^* - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^* \mid Z, X) \\
&= \mathbb{E}(Y^* \mid Z, X) - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^* \\
&\overset{\text{MAR}}{=} \mathbb{E}(Y^* \mid Z, X, D) - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^* \\
&= D[\mathbb{E}(Y^* \mid Z, X, D = 1) - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^*] \\
&\quad + (1 - D)[\mathbb{E}(Y^* \mid Z, X, D = 0) - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^*] \\
&= D\mathbb{E}(Y - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^* \mid Z, X, D = 1) \\
&\quad + (1 - D)[Y_{\text{imp}} - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^*],
\end{aligned}$$

because $\mathbb{E}(Y^* \mid Z, X, D = 0) \overset{\text{MAR}}{=} \mathbb{E}(Y^* \mid Z, X, D = 1) =: Y_{\text{imp}}$. Therefore,

$$\text{LHS of (D.34)} \overset{\text{MAR}}{=} T_1 + T_2,$$

where

$$\begin{aligned}
T_1 &:= \mathbb{E}[\begin{bmatrix} 1 \\ X_{\text{in}} \\ \text{BLP}(Z_{\text{in}} \mid X) \end{bmatrix} D\mathbb{E}(Y - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^* \mid Z, X, D = 1)] \\
T_2 &:= \mathbb{E}[\begin{bmatrix} 1 \\ X_{\text{in}} \\ \text{BLP}(Z_{\text{in}} \mid X) \end{bmatrix} (1 - D)(Y_{\text{imp}} - \alpha^* - X_{\text{in}}'\beta^* - \text{BLP}(Z_{\text{in}} \mid X)'\gamma^*)].
\end{aligned}$$

[25] Unlike Example 4.1, $\mu \neq 0$ here. Consequently, the imputation $Y_{\text{imp}} = \alpha^* + X_{\text{in}}'\beta^* + Z_{\text{in}}'\gamma^* + \mu$ provides information about missing outcomes that is not available from the regression model itself.

Note that T_2 is the second term of (D.37). To demonstrate that (D.34) and (D.37) are identical, we now show that T_1 coincides with the first term of (D.37). Now,

$$\mathbb{E}(Y-\alpha^*-X_{\text{in}}'\beta^*-\text{BLP}(Z_{\text{in}}\mid X)'\gamma^*\mid Z,X,D=1) \\ =\mathbb{E}(\frac{D(Y-\alpha^*-X_{\text{in}}'\beta^*-\text{BLP}(Z_{\text{in}}\mid X)'\gamma^*)}{\pi}\mid Z,X).$$

Hence, recalling that $\pi:=\mathbb{E}[D\mid Z,X]$, we have that

$$\begin{aligned} T_1 &= \mathbb{E}[D\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}}\mid X)\end{bmatrix}\mathbb{E}(\frac{D(Y-\alpha^*-X_{\text{in}}'\beta^*-\text{BLP}(Z_{\text{in}}\mid X)'\gamma^*)}{\pi}\mid Z,X)] \\ &= \mathbb{E}[\mathbb{E}(D\mid Z,X)\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}}\mid X)\end{bmatrix}\mathbb{E}(\frac{D(Y-\alpha^*-X_{\text{in}}'\beta^*-\text{BLP}(Z_{\text{in}}\mid X)'\gamma^*)}{\pi}\mid Z,X)] \\ &= \mathbb{E}[D\begin{bmatrix}1\\ X_{\text{in}}\\ \text{BLP}(Z_{\text{in}}\mid X)\end{bmatrix}(Y-\alpha^*-X_{\text{in}}'\beta^*-\text{BLP}(Z_{\text{in}}\mid X)'\gamma^*)]. \end{aligned}$$

Therefore, we have shown that

$$\text{LHS of (D.34)} \overset{\text{MAR}}{=} \text{LHS of (D.37)}.$$

But, $\mathbb{E}[\varepsilon\mid X]\overset{P_X\text{-a.s.}}{=}0$ and the orthogonality property of BLP residuals implies that (D.34) holds. Hence, (D.37) is also true.

3. In this example, $\hat{\theta}_{\text{VS}}\neq\hat{\theta}_{\text{2SLS}}$ because the FOC of (D.36) and (D.38) yield

$$\hat{\theta}_{\text{VS}}=(\sum_{i=1}^n\frac{D_i}{\hat{\pi}_i}W_{\text{in},i}W_{\text{in},i}')^{-1}(\sum_{i=1}^n\frac{D_i}{\hat{\pi}_i}W_{\text{in},i}Y_i) \qquad (W_{\text{in},i}:=(1,X_{\text{in},i},\hat{Z}_{\text{in},i})_{p\times 1})$$

$$\hat{\theta}_{\text{2SLS}}=(\sum_{i=1}^n W_{\text{in},i}W_{\text{in},i}')^{-1}(\sum_{i=1}^n W_{\text{in},i}[D_iY_i+(1-D_i)\hat{Y}_{\text{imp},i}]).$$

Example D.4 (UCMR models)**.** If there is no conditioning in (2.1), then the efficiency bound in Lemma 4.1 reduces to the one obtained by Chen, Hong, and Tarozzi (2008, Theorem 1), and Graham (2011, p. 439), for estimating parameters in UCMR models with some variables missing. Muris (2020) extends their results to allow for incomplete data across several strata. □

Remark D.5 (Working approximations)**.** In applications, working approximations are often parametric or semiparametric models. A popular choice is the best linear predictor (BLP) $\pi_{\text{work}}=\text{BLP}[D\mid Z,X]$ and $\mu_{\text{work}}=\text{BLP}[g_{\text{obs}}\mid Z,X,D=1]$, estimated by regressing D on (Z,X) in the observed sample, and regressing g_{obs} coordinatewise on (Z,X) in the validation sample, respectively. □

Remark D.6 (Why do all doubly robust moment functions — but not all doubly robust estimators — look the same?)**.** The form of doubly robust moment function $\rho(\pi_{\text{work}},\mu_{\text{work}})$ in (4.7) is very similar to the moment function in Hristache and Patilea (2021, Eqn. 20). Indeed, as Hristache and Patilea themselves note right after their Eqn. 20, all doubly robust moment functions look the same! This is not surprising because each doubly robust moment function is obtained as the residual from projecting the structural moment function g on the tangent space of score functions for the propensity score π. That is why, e.g., the doubly robust moment functions in Hristache

and Patilea (2021, Eqn. 20), Graham, de Xavier Pinto, and Egel (2012, p. 1058, obtained by subtracting their Eqn. 8 from their Eqn. 7), and Graham (2011, p. 441), all look the same as the doubly robust moment function in Scharfstein, Rotnitzky, and Robins (1999, their example in Section 3.2.3, p. 1141), who first developed the notion of double robustness.

The difference lies in the moment conditions used to construct the doubly robust estimators of the parameters of interest:

- Hristache and Patilea (2021, Section 4) consider estimating θ in an UCMR model (cf. their Eqn. 11 on page 157 of their paper where they define their UCMR model). Therefore, the doubly robust estimator of θ in their Eqn. 20 (on p. 159 of their paper) solves a UCMR based on their doubly robust moment function.

- In contrast, we estimate parameters in CMR models and our parameter θ^* is defined via the CMR in (2.1). Therefore, as discussed in Section 4.1, we can base a doubly robust estimator of θ^* on the CMR $\mathbb{E}[\rho(\pi_{\text{work}}, \mu_{\text{work}}) \mid X] \overset{P_X\text{-a.s.}}{=} 0_{\dim(g)\times 1}$ if either the working model for D or the working model for imputing g is correctly specified.

Clearly, the doubly robust estimator of Hristache and Patilea (2021, Eqn. 20) is based on a UCMR, whereas our doubly robust estimator is based on a CMR. Therefore, the two doubly robust estimators are not the same! In particular, as evident from Lemma 4.1, our doubly robust estimator is semiparametrically efficient if the working model for μ (used for imputing g) and the working model for D (used for estimating the propensity score π) are both correctly specified. In contrast, the doubly robust estimator of Hristache and Patilea cannot achieve the semiparametric efficiency bound even if the working models for μ and π are both correctly specified. □

Proof of (4.8). Under MAR, the "noise" $[\frac{D}{\pi_{\text{work}}} - 1][g - \mu_{\text{work}}]$ in (4.7) is mean independent of (Z, X) when the misspecification is not simultaneous (cf. (D.41) below). Indeed, as

$$\begin{aligned}\mathbb{E}[[\frac{D}{\pi_{\text{work}}} - 1][g - \mu_{\text{work}}] \mid Y^*, Z, X] &\overset{P_{Y^*,Z,X}\text{-a.s.}}{=} [g - \mu_{\text{work}}][\frac{\mathbb{E}[D \mid Y^*, Z, X]}{\pi_{\text{work}}} - 1] \\ &\overset{\text{MAR}}{=} [g - \mu_{\text{work}}][\frac{\pi}{\pi_{\text{work}}} - 1], \end{aligned} \tag{D.39}$$

it follows from the tower property of conditional expectations that

$$\mathbb{E}[[\frac{D}{\pi_{\text{work}}} - 1][g - \mu_{\text{work}}] \mid Z, X] \overset{P_{Z,X}\text{-a.s.}}{=} [\mu - \mu_{\text{work}}][\frac{\pi}{\pi_{\text{work}}} - 1]. \tag{D.40}$$

Hence, under MAR,

$$\begin{aligned}\mathbb{E}[[\frac{D}{\pi_{\text{work}}} - 1][g - \mu_{\text{work}}] \mid Z, X] &\overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1} \\ &\overset{(D.40)}{\iff} \pi_{\text{work}} \overset{P_{Z,X}\text{-a.s.}}{=} \pi \quad \text{or} \quad \mu_{\text{work}} \overset{P_{Z,X}\text{-a.s.}}{=} \mu. \end{aligned} \tag{D.41}$$

Therefore, as $\mathbb{E}[g \mid X] \overset{(2.1)}{=} 0_{\dim(g)\times 1}$ P_X-a.s., we have that

$$\pi_{\text{work}} \overset{P_{Z,X}\text{-a.s.}}{=} \pi \text{ or } \mu_{\text{work}} \overset{P_{Z,X}\text{-a.s.}}{=} \mu \implies \mathbb{E}[\rho(\pi_{\text{work}}, \mu_{\text{work}}) \mid X] \overset{(4.7),(D.41)}{=} 0_{\dim(g)\times 1} \; P_X\text{-a.s.} \qquad \square$$

Remark D.7 (Showing that ρ is the "least noisy" version of $\text{var}\,\rho(\pi, \cdot)$)**.** Note that

$$\text{cov}(g, [\frac{D}{\pi} - 1][g - \mu_{\text{work}}]) = 0_{\dim(g)\times\dim(g)},$$

because, by (D.39), $\mathbb{E}[[\frac{D}{\pi}-1][g-\mu_{\text{work}}] \mid Y^*,Z,X] \overset{P_{Y^*,Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}$. Therefore,

$$\operatorname{var}\rho(\pi,\mu_{\text{work}}) \overset{(4.7)}{=} \operatorname{var} g + \operatorname{var}[\frac{D}{\pi}-1][g-\mu_{\text{work}}].$$

As shown below,

$$\operatorname{var}[\frac{D}{\pi}-1][g-\mu_{\text{work}}] \overset{\text{MAR}}{=} \mathbb{E}[\frac{1-\pi}{\pi}\operatorname{var}[g \mid Z,X]] + \mathbb{E}[\frac{1-\pi}{\pi}[\mu-\mu_{\text{work}}][\mu-\mu_{\text{work}}]']. \tag{D.42}$$

Hence,

$$\operatorname{var}\rho(\pi,\mu_{\text{work}}) = \operatorname{var} g + \mathbb{E}[\frac{1-\pi}{\pi}\operatorname{var}[g \mid Z,X]] + \mathbb{E}[\frac{1-\pi}{\pi}[\mu-\mu_{\text{work}}][\mu-\mu_{\text{work}}]'],$$

which implies that $\operatorname{var}\rho(\pi,\mu) \overset{\text{MAR}}{\leq_L} \operatorname{var}\rho(\pi,\mu_{\text{work}})$. Finally, to show (D.42), observe that

$$\begin{aligned}
&\operatorname{var}[\frac{D}{\pi}-1][g-\mu_{\text{work}}] \\
&\quad= \mathbb{E}[\frac{D}{\pi}-1]^2[g-\mu_{\text{work}}][g-\mu_{\text{work}}]' \\
&\quad= \mathbb{E}[[g-\mu_{\text{work}}][g-\mu_{\text{work}}]'\mathbb{E}[(\frac{D}{\pi}-1)^2 \mid Y^*,Z,X]] && \text{(tower property)} \\
&\quad= \mathbb{E}[[g-\mu_{\text{work}}][g-\mu_{\text{work}}]'\mathbb{E}[(\frac{D}{\pi^2}+1-\frac{2D}{\pi}) \mid Y^*,Z,X]] && (D^2=D) \\
&\quad= \mathbb{E}[[g-\mu_{\text{work}}][g-\mu_{\text{work}}]'(\frac{\mathbb{E}[D \mid Y^*,Z,X]}{\pi^2}+1-\frac{2\mathbb{E}[D \mid Y^*,Z,X]}{\pi})] \\
&\quad\overset{\text{MAR}}{=} \mathbb{E}[\frac{1-\pi}{\pi}[g-\mu_{\text{work}}][g-\mu_{\text{work}}]'] && (\pi := \mathbb{E}[D \mid Z,X]) \\
&\quad= \mathbb{E}[\frac{1-\pi}{\pi}(\varsigma\varsigma' + \varsigma[\mu-\mu_{\text{work}}]' + [\mu-\mu_{\text{work}}]\varsigma' + [\mu-\mu_{\text{work}}][\mu-\mu_{\text{work}}]')] && (\varsigma := g-\mu) \\
&\quad= \mathbb{E}[\frac{1-\pi}{\pi}\varsigma\varsigma'] + \mathbb{E}[\frac{1-\pi}{\pi}[\mu-\mu_{\text{work}}][\mu-\mu_{\text{work}}]'] && (\mathbb{E}[\varsigma \mid Z,X] \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}) \\
&\quad= \mathbb{E}[\frac{1-\pi}{\pi}\mathbb{E}[\varsigma\varsigma' \mid Z,X]] + \mathbb{E}[\frac{1-\pi}{\pi}[\mu-\mu_{\text{work}}][\mu-\mu_{\text{work}}]'] && \text{(tower property)} \\
&\quad= \mathbb{E}[\frac{1-\pi}{\pi}\operatorname{var}[\varsigma \mid Z,X]] + \mathbb{E}[\frac{1-\pi}{\pi}[\mu-\mu_{\text{work}}][\mu-\mu_{\text{work}}]'] && (\mathbb{E}[\varsigma \mid Z,X] \overset{P_{Z,X}\text{-a.s.}}{=} 0_{\dim(g)\times 1}) \\
&\quad\overset{\varsigma:=g-\mu}{=} \mathbb{E}[\frac{1-\pi}{\pi}\operatorname{var}[g \mid Z,X]] + \mathbb{E}[\frac{1-\pi}{\pi}[\mu-\mu_{\text{work}}][\mu-\mu_{\text{work}}]']. && \square
\end{aligned}$$

Proof of (4.10) and (4.11). The Lagrangian for (4.9) is

$$\mathscr{L}(\theta) := \sum_{i=1}^n\sum_{j=1}^n w_{ij}\log p_{ij} - \sum_{i=1}^n \xi_i(\theta)(\sum_{j=1}^n p_{ij}-1) - \sum_{i=1}^n\sum_{j=1}^n \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)p_{ij},$$

where $(\xi_i(\theta))_{i=1,\ldots,n}$ and $(\hat{\lambda}_i(\theta))_{i=1,\ldots,n}$ are the multipliers in (4.9) that impose the adding up constraint and the moment condition, respectively. The solution $(\hat{p}_{ij}(\theta))_{i,j=1,\ldots,n}$ to the constrained optimization problem in (4.9) is positive,[26] and solves the FOC

$$\frac{w_{ij}}{\hat{p}_{ij}(\theta)} - \xi_i(\theta) - \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta) = 0, \qquad i,j=1,\ldots,n. \tag{D.43}$$

[26]Since w_{ij} is based on a 2nd order kernel, $w_{ij}>0$ for small enough b_n. Hence, if b_n is small enough, then $\hat{p}_{ij}(\theta)=0$ cannot maximize the objective function in (4.9) because $w_{ij}\lim_{\hat{p}_{ij}(\theta)\to 0+}\log\hat{p}_{ij}(\theta) = -\infty$. On the other hand, if $\hat{p}_{ij}(\theta)<0$, then the objective function in (4.9) is undefined.

As $\sum_{j=1}^n w_{ij} = 1$, $\sum_{j=1}^n \hat{p}_{ij}(\theta) \overset{(4.9)}{=} 1$, and $\sum_{j=1}^n \hat{\rho}_j(\theta)\hat{p}_{ij}(\theta) \overset{(4.9)}{=} 0$, we have

$$w_{ij} \overset{(D.43)}{=} \xi_i(\theta)\hat{p}_{ij}(\theta) + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)\hat{p}_{ij}(\theta) \quad \forall i,j \implies 1 = \xi_i(\theta) \quad \forall i.$$

Therefore,

$$\hat{p}_{ij}(\theta) \overset{(D.43)}{=} \frac{w_{ij}}{1 + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)}, \qquad i,j = 1,\dots,n.$$

Moreover, as the $\hat{p}_{ij}(\theta)$ have to satisfy the constraints in (4.9), the $\hat{\lambda}_i(\theta)$ satisfy

$$\sum_{j=1}^n \frac{w_{ij}\hat{\rho}_j(\theta)}{1 + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)} = 0_{\dim(g)\times 1}, \qquad i = 1,\dots,n. \qquad \square$$

D.3.3 Asymptotic normality of the SEL estimator

Throughout this section, $\hat{m}_i(\theta) := \sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij}\hat{\rho}_j(\theta)$, $\hat{\Omega}_i(\theta) := \sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij}\hat{\rho}_j(\theta)\hat{\rho}_j'(\theta)$, $\|\cdot\|$ is the Euclidean norm, $\mathscr{B}_\epsilon := \{\theta \in \Theta : \|\theta - \theta^*\| < \epsilon\}$ is an open ball of radius $\epsilon > 0$ centered at θ^* (hence, consistency of $\hat{\theta}$ implies that $\Pr(\hat{\theta} \in \mathscr{B}_\epsilon) \to 1$), and the symbol $\lesssim$ indicates that the left-hand side of an inequality is bounded by a positive constant times the right-hand-side, where the constant does not depend on i,j,n,θ,ϵ. As $\hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}\log(1 + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)) = \log(1 + \hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta))$ and $|x| < 1/2 \implies |\log(1+x) - x| \le 2x^2$, we have that

$$\forall (i,j,\theta) \in \mathbb{N}\times\mathbb{N}\times\mathscr{B}_\epsilon, \quad \hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)| < \frac{1}{2} \implies$$
$$|\hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}\log(1 + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)) - \hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)| \lesssim \hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)|^2.$$

Hence, by (4.13),

$$\forall (i,j,\theta) \in \mathbb{N}\times\mathbb{N}\times\mathscr{B}_\epsilon, \quad \hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)| < \frac{1}{2} \implies$$
$$|-\frac{1}{n}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) - \frac{1}{n}\sum_{i=1}^n \hat{\mathbb{T}}_{1i}\hat{\lambda}_i'(\theta)\hat{m}_i(\theta)| \lesssim \frac{1}{n}\sum_{i=1}^n \hat{\mathbb{T}}_{1i}\sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)|^2. \quad \text{(D.44)}$$

Furthermore, letting $\hat{r}_i(\theta) := \sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij}\hat{\rho}_j(\theta)\dfrac{(\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta))^2}{1 + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)}$, we have

$$0_{\dim(g)\times 1} \overset{(4.14)}{=} \sum_{j=1}^n \frac{\hat{\mathbb{T}}_{2j} w_{ij}\hat{\rho}_j(\theta)}{1 + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)} = \sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij}\hat{\rho}_j(\theta)[1 - \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta) + \frac{(\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta))^2}{1 + \hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)}]$$
$$= \hat{m}_i(\theta) - \hat{\Omega}_i(\theta)\hat{\lambda}_i(\theta) + \hat{r}_i(\theta),$$

which implies that

$$\hat{\mathbb{T}}_{1i}\hat{\lambda}_i(\theta) = \hat{\mathbb{T}}_{1i}\hat{\Omega}_i^{-1}(\theta)\hat{m}_i(\theta) + \hat{\mathbb{T}}_{1i}\hat{\Omega}_i^{-1}(\theta)\hat{r}_i(\theta). \quad \text{(D.45)}$$

Therefore, letting $\widehat{\mathrm{SMD}}(\theta) := n^{-1}\sum_{i=1}^n \hat{\mathbb{T}}_{1i}\hat{m}_i'(\theta)\hat{\Omega}_i^{-1}(\theta)\hat{m}_i(\theta)$ denote the kernel-based continuous updating version of the SMD objective function (Ai and Chen, 2003, Eqn. 23), we have

$$\forall (i,j,\theta) \in \mathbb{N}\times\mathbb{N}\times\mathscr{B}_\epsilon, \quad \hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)| < \frac{1}{2} \overset{(D.44)\&(D.45)}{\implies}$$
$$|-\frac{1}{n}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) - \widehat{\mathrm{SMD}}(\theta)| \lesssim \frac{1}{n}\sum_{i=1}^n \hat{\mathbb{T}}_{1i}\sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)|^2$$
$$+ |\frac{1}{n}\sum_{i=1}^n \hat{\mathbb{T}}_{1i}\hat{r}_i'(\theta)\hat{\Omega}_i^{-1}(\theta)\hat{m}_i(\theta)|. \quad \text{(D.46)}$$

The following "high-level" conditions are used to simplify the right-hand-side of (D.46).

Assumption D.2. *θ^* lies in the interior of Θ, and $\mathscr{B}_\epsilon$ is such that*:

(i) *The eigenvalues of the matrix $\Omega_i(\theta) := \mathbb{E}[\rho(\mathscr{A},\theta)\rho'(\mathscr{A},\theta) \mid X_i]$ are, P_X-a.s., uniformly bounded away from zero on $\mathscr{B}_\epsilon$;*

(ii) $\lim_{\epsilon\to 0+} \sup_{\theta\in\mathscr{B}_\epsilon} \|\mathbb{E}[\rho(\mathscr{A},\theta) \mid X]\| \overset{P_X\text{-a.s.}}{=} 0$;

(iii) $\max_{1\le i,j\le n} \sup_{\theta\in\mathscr{B}_\epsilon} \hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)| = o_{\mathrm{p}}(1)$;

(iv) $\sup_{\theta\in\mathscr{B}_\epsilon} n^{-1}\sum_{i=1}^n \hat{\mathbb{T}}_{1i}\sum_{j=1}^n \hat{\mathbb{T}}_{2j} w_{ij}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)|^2 = o_{\mathrm{p}}(1)$;

(v) $\sup_{\theta\in\mathscr{B}_\epsilon} |n^{-1}\sum_{i=1}^n \hat{\mathbb{T}}_{1i}\hat{r}_i'(\theta)\hat{\Omega}_i^{-1}(\theta)\hat{m}_i(\theta)| = o_{\mathrm{p}}(1)$.

(i) is a standard assumption when variance-covariance matrices that depend on parameters are estimated. (ii) makes sense because $\mathbb{E}[\rho(\mathscr{A},\theta^*) \mid X] \overset{(4.5)}{=} 0_{\dim(g)\times 1}$ P_X-a.s. (iii), (iv), and (v) can be justified as follows. Under regularity conditions that make kernel estimators uniformly consistent, it can be shown that, for $k \in \{1,\dots,5\}$, there exist sequences of positive real numbers $(\delta_{k,n})_{n\in\mathbb{N}} \to 0+$ such that:

(a) $\max_{1\le j\le n} \sup_{\theta\in\mathscr{B}_\epsilon} \hat{\mathbb{T}}_{2j}\|\hat{\rho}_j(\theta) - \rho_j(\theta)\| = O_{\mathrm{p}}(\delta_{1,n})$;

(b) $\max_{1\le i\le n} \sup_{\theta\in\mathscr{B}_\epsilon} \hat{\mathbb{T}}_{1i}\|\hat{m}_i(\theta) - \mathbb{E}[\rho(\mathscr{A},\theta) \mid X_i]\| = O_{\mathrm{p}}(\delta_{2,n})$;

(c) $\max_{1\le i\le n} \sup_{\theta\in\mathscr{B}_\epsilon} \hat{\mathbb{T}}_{1i}\|\hat{\Omega}_i(\theta) - \Omega_i(\theta)\| = O_{\mathrm{p}}(\delta_{3,n})$.

Using (a), (b), (c), and following the argument in KTA (Lemma B.1), it can be shown that:

(d) $\max_{1\le i\le n} \sup_{\theta\in\mathscr{B}_\epsilon} \hat{\mathbb{T}}_{1i}\|\hat{r}_i(\theta)\| = O_{\mathrm{p}}(\delta_{4,n})$, where $\hat{r}_i(\theta)$ is the remainder in (D.45);

(e) $\max_{1\le i\le n} \sup_{\theta\in\mathscr{B}_\epsilon} \hat{\mathbb{T}}_{1i}\|\hat{\lambda}_i(\theta) - \lambda_{i,n}(\theta)\| = O_{\mathrm{p}}(\delta_{5,n})$, where $\lambda_{i,n}(\theta)$ solves

$$\sum_{j=1}^n \frac{w_{ij}\rho_j(\theta)}{1+\lambda_{i,n}'(\theta)\rho_j(\theta)} = 0_{\dim(g)\times 1}.$$

Therefore, (iii) follows from (a) and (e) provided $\max_{1\le i,j\le n} \sup_{\theta\in\mathscr{B}_\epsilon} |\lambda_{i,n}'(\theta)\rho_j(\theta)| = o_{\mathrm{p}}(1)$, which can be justified as for KTA (Eqn. 3.1).[27] Finally, (iv) is a direct consequence of (iii); and (v) follows from (i), (ii), (b), (c), and (d).

The right-hand-side of (D.46) is asymptotically negligible under Assumption D.2(iv, v). Hence, as $\Pr(\max_{1\le i,j\le n} \sup_{\theta\in\mathscr{B}_\epsilon} \hat{\mathbb{T}}_{1i}\hat{\mathbb{T}}_{2j}|\hat{\lambda}_i'(\theta)\hat{\rho}_j(\theta)| < \frac{1}{2}) \xrightarrow{\text{Ass. D.2(iii)}} 1$, we have that

$$\sup_{\theta\in\mathscr{B}_\epsilon} |-\frac{1}{n}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) - \widehat{\mathrm{SMD}}(\theta)| = o_{\mathrm{p}}(1). \tag{D.47}$$

It is clear from (D.47) that the kernel-based continuous updating version of the SMD objective function is a quadratic approximation of the SEL objective function in large samples. Hence, it is not surprising that (D.47) implies that in large samples the SEL estimator $\hat{\theta}$ is also an approximate local minimizer of $\widehat{\mathrm{SMD}}$, i.e.,

$$\widehat{\mathrm{SMD}}(\hat{\theta}) \le \inf_{\theta\in\mathscr{B}_\epsilon} \widehat{\mathrm{SMD}}(\theta) + o_{\mathrm{p}}(1). \tag{D.48}$$

[27] KTA justify this condition by assuming that the random variable $\sup_{\theta\in\Theta}\|\rho(\mathscr{A},\theta)\|$ has enough moments, and that $\lambda_{i,n}(\theta)$ is asymptotically negligible uniformly on $\mathscr{B}_\epsilon$. The intuition as to why the Lagrange multiplier $\lambda_{i,n}(\theta)$ should be asymptotically negligible comes from the fact that it is enforcing a correctly specified moment condition in the sample.

Therefore, as $\hat{\theta}$ is consistent for θ^*, the result $n^{1/2}(\hat{\theta} - \theta^*) \xrightarrow{\mathrm{d}} \mathrm{N}(0_{\dim(\theta^*)\times 1}, (\mathbb{E}J'\Omega_\rho^{-1}J)^{-1})$ can be shown by applying the arguments in the proof of Ai and Chen (Theorem 4.1) to $\widehat{\mathrm{SMD}}(\hat{\theta})$ — keeping in mind the comment after their Eqn. 23, and translating their conditions on sieves and the number of approximating functions into conditions on the order of the kernel H, the bandwidths (b_n, c_n, d_n), and the trimming parameters (τ_b, τ_c, τ_d), such that the remainder terms in the expansion of $\widehat{\mathrm{SMD}}(\hat{\theta})$ about $\widehat{\mathrm{SMD}}(\theta^*)$ decay sufficiently fast.

Proof of (D.48). Let $(R_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative random variables such that $R_n = o_{\mathrm{p}}(1)$. Then, as $-n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\hat{\theta}) \overset{(4.12)}{\leq} \inf_{\theta\in\mathscr{B}_\epsilon} -n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta)$ and $R_n \geq 0$, we have that

$$\begin{aligned}
-n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\hat{\theta}) &\leq \inf_{\theta\in\mathscr{B}_\epsilon} -n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) + R_n \\
&\leq |\inf_{\theta\in\mathscr{B}_\epsilon} -n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) - \inf_{\theta\in\mathscr{B}_\epsilon} \widehat{\mathrm{SMD}}(\theta)| + \inf_{\theta\in\mathscr{B}_\epsilon} \widehat{\mathrm{SMD}}(\theta) + R_n \qquad (\widehat{\mathrm{SMD}} \geq 0) \\
&\leq \sup_{\theta\in\mathscr{B}_\epsilon} |-n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) - \widehat{\mathrm{SMD}}(\theta)| + \inf_{\theta\in\mathscr{B}_\epsilon} \widehat{\mathrm{SMD}}(\theta) + R_n,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
\widehat{\mathrm{SMD}}(\hat{\theta}) &\leq \sup_{\theta\in\mathscr{B}_\epsilon} |-n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) - \widehat{\mathrm{SMD}}(\theta)| + \inf_{\theta\in\mathscr{B}_\epsilon} \widehat{\mathrm{SMD}}(\theta) + R_n + \widehat{\mathrm{SMD}}(\hat{\theta}) + n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\hat{\theta}) \\
&\overset{\text{w.p.a.1}}{\leq} 2\sup_{\theta\in\mathscr{B}_\epsilon} |-n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) - \widehat{\mathrm{SMD}}(\theta)| + \inf_{\theta\in\mathscr{B}_\epsilon} \widehat{\mathrm{SMD}}(\theta) + R_n,
\end{aligned}$$

because

$$\hat{\theta} \in \mathscr{B}_\epsilon \implies \widehat{\mathrm{SMD}}(\hat{\theta}) + n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\hat{\theta}) \leq \sup_{\theta\in\mathscr{B}_\epsilon} |-n^{-1}\widehat{\mathrm{SEL}}_{\mathbb{T}}(\theta) - \widehat{\mathrm{SMD}}(\theta)|$$

and $\Pr(\hat{\theta} \in \mathscr{B}_\epsilon) \to 1$. The desired result follows from (D.47) and that $R_n = o_{\mathrm{p}}(1)$. □

References

ACKERBERG, D., X. CHEN, J. HAHN, AND Z. LIAO (2014): "Asymptotic efficiency of semiparametric two-step GMM," *Review of Economic Studies*, 81, 919–943.

ANTOINE, B., H. BONNAL, AND E. RENAULT (2007): "On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood," *Journal of Econometrics*, 138, 461–487.

BICK, A., A. BLANDIN, AND R. ROGERSON (2022): "Hours and wages," *NBER Working Paper No. 26722*, `https://www.nber.org/papers/w26722`.

BICKEL, P. J., C. A. J. KLASSEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore, MD, USA.

BROWN, B. W., AND W. K. NEWEY (2002): "GMM, efficient bootstrapping, and improved inference," *Journal of Business and Economic Statistics*, 20, 507–517.

CALONICO, S., M. D. CATTANEO, AND M. H. FARRELL (2018): "On the effect of bias estimation on coverage accuracy in nonparametric inference," *Journal of the American Statistical Association*, 113, 767–779.

CHEN, S. X., W. HÄRDLE, AND M. LI (2003): "An empirical likelihood goodness-of-fit test for time series," *Journal of the Royal Statistical Society, Series B*, 65, 663–678.

CHEN, X., D. POUZO, AND J. L. POWELL (2019): "Penalized sieve GEL for weighted average derivatives of nonparametric quantile IV regressions," *Journal of Econometrics*, 213, 30–53.

CRAGG, J. G. (1983): "More efficient estimation in the presence of heteroscedasticity of unknown form," *Econometrica*, 49, 751–764.

DONALD, S. G., G. W. IMBENS, AND W. K. NEWEY (2003): "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics*, 117, 55–93.

EDDELBUETTEL, D., AND R. FRANÇOIS (2011): "Rcpp: Seamless R and C++ Integration," *Journal of Statistical Software*, 40(8), 1–18.

EDDELBUETTEL, D., AND C. SANDERSON (2014): "RcppArmadillo: Accelerating R with high-performance C++ linear algebra," *Computational Statistics and Data Analysis*, 71, 1054–1063.

FISHER, N. I., P. HALL, B.-Y. JING, AND A. T. A. WOOD (1996): "Improved pivotal methods for constructing confidence regions with directional data," *Journal of the American Statistical Association*, 91, 1062–1070.

FITZGERALD, J., P. GOTTSCHALK, AND R. MOFFITT (1998): "An analysis of sample attrition in panel data," *Journal of Human Resources*, 33, 251–299.

GRAHAM, B. S., C. C. DE XAVIER PINTO, AND D. EGEL (2012): "Inverse probability tilting for moment condition models with missing data," *Review of Economic Studies*, 79, 1053–1079.

HALL, P. (1990): "Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems," *Journal of Multivariate Analysis*, 32, 177–203.

HALMOS, P. R. (1951): *Introduction to Hilbert space*. Chelsea, New York, NY, USA, 2nd edn.

HITOMI, K., Y. NISHIYAMA, AND R. OKUI (2008): "A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters," *Econometric Theory*, 24, 1717–1728.

HRISTACHE, M., AND V. PATILEA (2016): "Semiparametric efficiency bounds for conditional moment restriction models with different conditioning variables," *Econometric Theory*, 32, 917–946.

KITAMURA, Y. (2001): "Asymptotic optimality of empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661–1672.

KITAMURA, Y. (2007): "Empirical likelihood methods in econometrics: Theory and practice," in *Advances in Economics and Econometrics*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 3, pp. 174–237. Cambridge University Press, Cambridge, UK.

KITAMURA, Y., A. SANTOS, AND A. M. SHAIKH (2012): "On the asymptotic optimality of empirical likelihood for testing moment restrictions," *Econometrica*, 80, 413–423.

LAVERGNE, P., AND V. PATILEA (2013): "Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory," *Journal of Econometrics*, 177, 47–59.

LUENBERGER, D. G. (1969): *Optimization by vector space methods*. John Wiley and Sons.

MURIS, C. (2020): "Efficient GMM estimation with incomplete data," *Review of Economics and Statistics*, 102, 518–530.

NEWEY, W. K. (1993): "Efficient estimation of models with conditional moment restrictions," in *Handbook of Statistics, vol. 11*, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, pp. 2111–2245. Elsevier, The Netherlands.

NEWEY, W. K. (1994): "Series estimation of regression functionals," *Econometric Theory*, 10, 1–28.

NEWEY, W. K., AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics, vol. IV*, ed. by R. F. Engle, and D. L. McFadden, pp. 2111–2245. Elsevier, The Netherlands.

OTSU, T. (2010): "On Bahadur efficiency of empirical likelihood," *Journal of Econometrics*, 157, 248–256.

OTSU, T. (2011): "Empirical likelihood estimation of conditional moment restriction models with unknown functions," *Econometric Theory*, 27, 8–46.

OWEN, A. (1990): "Empirical likelihood and small samples," in *Computing Science and Statistics: Proceedings of the Symposium on the Interface*, pp. 79–88. Springer-Verlag, Berlin.

OWEN, A. B. (2017): "A weighted self-concordant optimization for empirical likelihood," Manuscript, `https://artowen.su.domains/empirical/countnotes.pdf`.

ROTHENBERG, T. J. (1971): "Identification in parametric models," *Econometrica*, 39, 577–591.

SCHARFSTEIN, D. O., A. ROTNITZKY, AND J. M. ROBINS (1999): "Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion)," *Journal of the American Statistical Association*, 94, 1096–1146.

SCHUMANN, M., AND G. TRIPATHI (2018): "Convexity of probit weights," *Statistics and Probability Letters*, 143, 81–85.

SEVERINI, T. A., AND G. TRIPATHI (2001): "A simplified approach to computing efficiency bounds in semiparametric models," *Journal of Econometrics*, 102, 23–66.

——— (2013): "Semiparametric efficiency bounds for microeconometric models: A survey," *Foundations and Trends in Econometrics*, 6, 163–397.

SMITH, R. J. (2007): "Efficient information theoretic inference for conditional moment restrictions," *Journal of Econometrics*, 138, 430–460.

TRIPATHI, G., AND Y. KITAMURA (2003): "Testing conditional moment restrictions," *Annals of Statistics*, 31, 2059–2095.

TRIPATHI, G. (2011): "Generalized method of moments (GMM) based inference with stratified samples when the aggregate shares are known," *Journal of Econometrics*, 165, 258–265.