CIRANO
Allier savoir et décision

2016s-15

Indirect Inference with Endogenously Missing Exogenous Variables

Saraswata Chaudhuriy, David T. Frazierz, Eric Renault

Série Scientifique/Scientific Series

2016s-15

Indirect Inference with Endogenously Missing Exogenous Variables

Saraswata Chaudhuriy, David T. Frazierz, Eric Renault

Série Scientifique
Scientific Series

Montréal
Avril/April 2016

CIRANO
Allier savoir et décision

Centre interuniversitaire de recherche en analyse des organisations

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du ministère de l'Économie, de l'Innovation et des Exportations, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Quebec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the ministère de l'Économie, de l'Innovation et des Exportations, and grants and research mandates obtained by its research teams.

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l'unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.
This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 2292-0838 (en ligne)

Indirect Inference with Endogenously Missing Exogenous Variables[*]

Saraswata Chaudhuriy[†], *David T. Frazierz*[‡] *Eric Renault*[§]

Résumé/abstract

We consider consistent estimation of parameters in a structural model by Indirect Inference (II) when the exogenous variables can be missing at random (MAR) endogenously. We demonstrate that II procedures which simply discard sample units with missing observations can yield inconsistent estimates of the true structural parameters. By inverse probability weighting (IPW) the "complete case" observations, i.e., sample units with no missing variables for the observed and simulated samples, we propose a new method of II to consistently estimate the structural parameters of interest. Asymptotic properties of the new estimator are discussed. An illustration is provided based on a multinomial probit model. A small scale Monte-Carlo study in this model demonstrates the severe bias incurred by existing II estimators, and its subsequent correction by our new II estimator.

Mots clés/keywords : Indirect Inference; Missing at Random; Inverse Probability Weighting

Codes JEL/JEL Codes : L14, L62, F23, F61

[*] We thank the seminar participants at Brown, Monash and the New Zealand Econometrics Study Group for their comments, and J. Galbraith, D. Guilkey, F. Kleibergen, F. Lange, B.Werker and V. Zinde-Walsh for very useful discussions. We are grateful to R. Davidson and D. Guilkey for generously providing us with the resources for the computationally intensive Monte-Carlo study. We also thank the guest editors of this special issue and two anonymous referees for numerous helpful comments and suggestions.

[†] McGill University, Canada. Email: saraswata.chaudhuri@mcgill.ca.

[‡] Monash University, Australia. Email: david.frazier@monash.edu

[§] Brown University, United States. Email: eric renault@brown.edu

1 Introduction

Since the seminal work of Smith (1990, 1993), Gourieroux et al. (1993) and Gallant and Tauchen (1996), Indirect Inference (II) has been used for estimation in a variety of structural models where direct computation of likelihood functions is difficult but simulation from the structural model is relatively straightforward. Altonji et al. (2013) have recently remarked that in some circumstances "accommodating missing data in II is straightforward: after generating a complete set of simulated data, one simply omits observations in the same way they are omitted in the observed data." Our focus of interest in this paper is precisely a case where this argument is invalid due to the impossibility of simulating data that properly mimics the actual missing data mechanism in the Data Generating Process (DGP). As stressed by Jiang and Turnbull (2004) (Section 3.4), when data are not "Missing Completely At Random" (MCAR), the key tool of II, namely the bridge relationship (resp. binding function) in Jiang and Turnbull (resp. Gourieroux et al.) terminology, may be impossible to infer from simulations.

Generally speaking, II sets the focus on estimation of structural parameters $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ through an intermediate or auxiliary statistic that consistently estimates the true unknown value β^0 of some auxiliary parameters $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}, d_\beta \geq d_\theta$. For sake of expositional simplicity, we always define the true unknown value β^0 as the unique solution of the moment conditions

$$E[m(W, X, \beta)] = 0, \tag{1}$$

where W and X are random vectors. The vectorial function $m(., ., .)$ is known and can be assumed without loss of generality to be of dimension d_β.

Let $\{W_i, X_i\}_{i=1}^N$ stand for an i.i.d. sample drawn from the distribution of (W, X). The vector W_i (resp. X_i) could include components corresponding to different time points for the i-th sample unit, and in this sense our setup allows for panel data (large N, small T). However, for simplicity we will not pursue this aspect further.

We are interested in the case where the econometrician does not observe (W_i, X_i) for all N sample units, with the particular structure of the missingness being characterized by observing only a subsequence of $\{X_i\}_{i=1}^N$. Following common practice, define the binary random variable D_i with $D_i = 1$ when the vector X_i is observed. In other words, the econometrician knows the random subset of indices $i \in \{1, 2, ..., N\}$ for which X_i is missing, corresponding to the set of indices such that $D_i = 0$.

Throughout, we maintain the assumption that data are "Missing at Random" (MAR). Following Rubin

(1976), data are MAR when "the conditional probability of the observed pattern of missing data, given the missing data and the value of the observed data, is the same for all possible values of the missing data". With our notations, we enforce this condition by assuming that almost surely

$$\Pr[D = 1 \,|W, X] = \Pr[D = 1 \,|W] > 0. \tag{2}$$

Note that Wooldridge (2007) stresses the relevance of an extension of assumption (2) (see Wooldridge's Assumption 3.1 (iv) p. 1283) to also allow some components of the vector W_i to be unobserved whenever $D_i = 0$. In the context of II, this extension will be immaterial as long as the components of W impacted by the missing data mechanism are only endogenous variables. Generally speaking, if the above missing data mechanism only pertains to endogenous variables then the sanguine statement of Altonji et al. (2013) quoted above regarding the easy treatment of missing data in II is valid.

The focus of interest in this paper is a case where the solution put forward by Altonji et al. (2013) does not work, precisely because the missing data mechanism pertains to exogenous variables that we denote throughout by X and which we are not keen to simulate (see Section two for a more precise discussion). To emphasize our focus on missing exogenous variables, albeit endogenously missing since they are not MCAR, we will dub throughout the maintained assumption (2) the **MAR-X** property.

Under **MAR-X**, sample counterparts of the moment conditions (1) can only be deduced from the "observed" or "complete case" units $\{D_i \cdot m(W_i, X_i, \beta)\}_{i=1}^N$; i.e., when X_i is not observed we can not compute $m(W_i, X_i, \beta)$. Then, revisiting (1) as the "observed" or complete case moment conditions

$$E[Dm(W, X, \beta)] = 0 \tag{3}$$

would obviously lead to the textbook issue of selection bias. However, the use of (3) for II differs from the textbook presentation (see, e.g., Little and Rubin, 2002 and Wooldridge, 2005) in at least two respects.

First, our focus of interest is not direct estimation of β but rather indirect estimation of structural parameters θ through auxiliary parameters β. Second, unlike direct inference with missing data, the key necessary condition for validity of II is the ability of the simulated data to mimic the estimates of the auxiliary model obtained from the observed data, irrespective of what this model is. In this respect the important issue for simulation-based inference is not the difference between conditions (1) and (3), but to what extent this difference can be accounted for in our simulation-based inference procedure.

While it may be possible to accommodate the consequences due to the differences between (1) and (3), see Section two for specific details, our *main* goal is to modify the complete case moment conditions (3) into moment conditions, that are "observed" and conformable to the initial moment conditions in equation (1). To do so, we use the **MAR-X** assumption to revisit (3) as (possibly misspecified) conditional moment restrictions given W and to resort to a well chosen instrumental variable $h(W)$, leading to the "observed" moment conditions

$$E[Dh(W)m(W,X,\beta)] = 0. \tag{4}$$

Defining the true unknown propensity score as $p_0(W) = \Pr[D = 1\,|W]$, equation (4) will be conformable to the initial moment conditions of interest that define β^0 if, for all $\beta \in \mathcal{B}$,

$$E[Dh(W)m(W,X,\beta)] = E[p_0(W)h(W)m(W,X,\beta)] = E[m(W,X,\beta)],$$

where the first equality follows from the Law of Iterated Expectations (LIE) and **MAR-X**. Equivalence between moment conditions (1) and (4) then requires, by the LIE, for all $\beta \in \mathcal{B}$,

$$E\left[\{1 - p_0(W)h(W)\}\, E[m(W,X,\beta)\,|W]\right] = 0. \tag{5}$$

The identity in (5) encapsulates the two main cases of interest. In the first case, following Wooldridge (2007) (see Assumption 4.1 p. 1288), one can maintain the assumption of "exogenous selection", meaning that at β^0, the solution to (1), actually satisfies

$$E[m(W,X,\beta^0)\,|W] = 0. \tag{6}$$

Our focus of interest is not exogenous selection. In the II context, exogenous selection as in (6) amounts to a structural assumption on the auxiliary model, which is somewhat logically inconsistent with the idea of an auxiliary model:[1] II concerns indirect estimation of a structural model through a purely instrumental auxiliary model not endowed with any kind of structural belief.

In the second case, exogenous selection is not maintained and the conditional expectation computed in (6) may be any function of W.[2] As a result, obtaining the identity in equation (5) requires choosing the "instrument" $h(W)$ inversely proportional to the propensity score $p_0(W) := \Pr[D = 1|W]$; that is, to rewrite

[1] The exogenous selection assumption in (6) actually extends Wooldridge's (2007) definition for M-estimators to the case of general estimating equations, which could correspond to the first order conditions of some M-estimator.

[2] This follows since we are not willing to maintain any restrictive assumption about the probability distribution of the exogenous variables X.

our auxiliary model in (1) as the inverse probability weighted moment conditions

$$E\left[\frac{D}{p_0(W)}m(W,X,\beta)\right]=0. \tag{7}$$

Equivalence between moments (1) and (7) follows by the LIE and **MAR-X**.

While IPW has a long history in statistical inference with missing data, see, e.g., Horvitz and Thompson (1952) and Robins et al. (1994), this paper constitutes, to the best of our knowledge, the first use of IPW within simulation-based inference with endogenously missing exogenous variables. While seemingly different from its historical use, our IPW strategy is underpinned by the maintained **MAR-X** hypothesis that has found recent use in economics and econometrics. See, among others, Hirano et al. (2003), Chen et al. (2005), Chen et al. (2008), Graham et al. (2012) for cases where the missingness pattern is similar to ours, while Cattaneo (2010) and Chaudhuri and Guilkey (2014) consider more involved patterns of missingness. All of the above papers use some form of **MAR-X** to correct for selection bias in moment conditions, as is done herein.

However, unlike missing data in direct inference, because II can only conditionally simulate data given all exogenous variables, the simulation step of II induces a perverse dependence between the simulated endogenous variables and the missingness indicator, which is not present in the observed data. As a direct consequence, the standard IPW-based arguments for direct inference outlined above are not valid for our simulated counterpart. To rectify this issue, we detail a novel identification argument (see Section two) that uses IPW, along with the **MAR-X** assumption and a particular simulation design, to (jointly) identify the auxiliary parameters based on the simulated data. Together, the two IPW-based arguments allow for identification of the structural parameters.

The remainder of the paper is organized as follows. Our proposed II strategy with **MAR-X** exogenous variables as well as possible alternative strategies are discussed in Section two. Section three details implementation of this new II strategy, states the asymptotic theory of our II estimator and proposes an alternative implementation of our approach based on the generalized indirect inference (GII) approach originally proposed in Keane and Smith (2005), and elaborated on in Bruins et al. (2015), which is particularly useful when the underlying moments are non-smooth in the structural parameters. In Section four, we illustrate our new approach in a multinomial probit model similar to Section nine of Gourieroux et al. (1993) and model four of Bruins et al. (2015). However, due to the missing data problem, we carefully revisit identification of the structural parameters. Section four also contains a small scale Monte Carlo experiment illustrating the performance of our approach in a multinomial probit model. Given the non-smoothness of the binding function in the multinomial probit model, we also consider a GII implementation of our approach. The Monte Carlo results provide compelling evidence on the performance of our II strategy and its alternative GII implementation. Section five concludes, and proofs of the theoretical results are collected in the appendix.

2 II with MAR Exogenous Variables

We sketch in Section 2.1 the general problem of II in the presence of missing data. The usefulness of the **MAR-X** assumption for performing II is made explicit in Section 2.2. Since Section 2.1 simply sketches the different possibilities, precise definitions of certain terms are only provided in Section 2.2.

2.1 Indirect Inference with Missing Data

To fix ideas, we focus on the simple structural model

$$Y = r(Z, X, \varepsilon; \theta), \tag{8}$$

where $r(.)$ is a vector valued function known up to the finite dimensional parameter $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$. ε is the unobserved stochastic error whose probability distribution is assumed (without loss of generality) to be known. Y denotes the endogenous variables while the variables X and Z are independent of ε and treated as exogenous. We maintain that it is not desirable to assume the distribution of X and Z (conditional or unconditional on Y) is known. Let $\theta^0 \in \Theta$ be the true value of θ in our population of interest.

Let us first consider simulation of data from the structural model (8) when there is no missingness. Let $\tilde{\varepsilon}$ be a random variable drawn from the distribution of ε and independent of $W = (Y', Z')'$ and X. For the given $\theta \in \Theta$, consider the variables $Y(\theta)$ simulated from equation (8): $Y(\theta) = r(Z, X, \tilde{\varepsilon}; \theta)$.

For a given value $\theta \in \Theta$ used to simulate the endogenous variable $Y(\theta)$, II defines the binding function $\beta^0(\theta)$ as the solution, in β, to the simulated counterpart of moment conditions (1):

$$E\left[m(Y(\theta), Z, X, \beta)\right] = 0. \tag{9}$$

Since the conditional probability distribution of $Y(\theta_0)$, given X, Z has been devised to coincide with the conditional probability distribution of Y given X, Z, β_0 is the solution of (9) for $\theta = \theta_0$. Moreover, if

$\theta \mapsto \beta^0(\theta)$ is injective, θ^0 is the unique $\theta \in \Theta$ satisfying $\beta^0 = \beta^0(\theta)$. Implementation of II is then based on finite sample counterparts of equations (1) and (9).

It may sometimes be argued that missing data is immaterial for II, since we can purposefully omit simulated observations in the same way they are missing in the observed data. In other words, selection bias introduced when estimating the auxiliary parameters β using the observed moment conditions

$$E[Dm(Y, Z, X, \beta)] = 0$$

may be inconsequential. After all, if Y, Z, X, D can all be simulated, where with an abuse of notation θ denotes all unknown parameters governing the simulation process, we can define the binding function $\theta \mapsto \tilde{\beta}(\theta)$ as the solution of the simulated counterpart

$$E[D(\theta) \cdot m(Y(\theta), Z(\theta), X(\theta), \beta)] = 0.$$

However, our focus of interest in this paper renders the above solution infeasible for several reasons.

First, since we are unwilling to assume a distribution for the exogenous variables X, Z, obtaining simulated counterparts to X, Z is not feasible. As alluded to above, it is somewhat logically inconsistent to specify a probability distribution for exogenous variables. Moreover, as shown by Gourieroux at al. (1993), there is an efficiency gain to use for II a simulated path in which the exogenous variables are fixed at their observed values in the actual data set.

Second, when only the endogenous variables Y are missing, fixing the exogenous variables at their observed values yields a binding function for II defined as the solution to

$$E[D(\theta)m(Y(\theta), Z, X, \beta)] = 0,$$

where θ again denotes all unknown parameters governing the simulated process. Such a practice would be accurate if we were to treat both Y and D as endogenous, with $Y(\theta)$ and $D(\theta)$ simulated according to an augmented analog of the structural model in (8). However, simulation of $Y(\theta)$ and $D(\theta)$ is not feasible in our context since the missing data mechanism pertains to the exogenous variables X and so $Y(\theta)$ can not be simulated when $D = 0$.

This inability to simulate $Y(\theta)$ when $D=0$ also implies that II based on the complete case moments

$$E[Dm(Y,Z,X,\beta)]=0$$
$$E[Dm(Y(\theta),Z,X,\beta)]=0$$

will not, in general, identify θ^0. Identification would require that the joint distributions of (D,Y,Z,X) and $(D,Y(\theta^0),Z,X)$ be equivalent. However, this can not be true in general for the following reason: the simulated error $\tilde{\varepsilon}$ used to generate $Y(\theta^0)$ is, by construction, independent of D, whereas the **MAR-X** assumption *does not* demand independence between ε (structural error in Y-equation) and D, either unconditionally or conditional on X and Z. Hence, unless D is independent of ε, which, in turn, rules out endogenous missingness of X, one can not identify θ^0 following the above approach except by happenstance. To further clarify this idea of identification failure, we refer the interested reader to Section 4.1 for a toy example that illustrates this failure.

Interestingly enough, this double data missingness hurdle may actually suggest to modify the binding function even more by considering the simulated moment conditions:

$$E[D\cdot D(\theta)\cdot m(Y(\theta),Z,X,\beta)]=0.$$

While this complicated simulated missing data mechanism may actually provide a feasible solution (see Chaudhuri et al., 2016), our focus herein is rather to use the initial moment condition in (1) by considering their IPW counterpart given in equation (7). While the **MAR-X** assumption ensures that the IPW moment conditions in (7) are equivalent to those in (1), **MAR-X** also allows us to define a binding function for II that identified the true structural parameters without simulating the missingness indicator D.

2.2 II Based on IPW Under MAR-X

The **MAR-X** assumption in (2) allows us to correct for the effects of selection bias in the identification of the auxiliary parameters β; that is, for all $\beta\in\mathcal{B}$, we have, by **MAR-X**

$$E\left[\frac{D}{p_0(W)}m(Y,Z,X,\beta)\right]=E\left[E\left(\frac{D}{p_0(W)}\middle|W,X\right)m(Y,Z,X,\beta)\right]=E\left[m(Y,Z,X,\beta)\right]. \tag{10}$$

The key property for performing II using the auxiliary model based on equation (10) is to ensure the resulting binding function will properly match. This matching requires that, for all $\beta \in \mathcal{B}$ and for all $\theta \in \Theta$,

$$E\left[\frac{D}{p_0(W)} m(Y(\theta), Z, X, \beta)\right] = E\left[m(Y(\theta), Z, X, \beta)\right]. \tag{11}$$

Demonstrating satisfaction of (11) requires a more precise study of the expectations used above. In equation (10) the notations are straightforward: expectations are computed with respect to the joint distribution of (D, Y, Z, X) given by the data generating process (DGP). The expectation operator in (11) involves jointly the DGP for the observed and simulated data. To highlight this difference, we analyze each case in turn, starting with the observed data.

The observed data $\{D_i, Y_i, Z_i, D_i X_i\}_{i=1}^N$, where $D_i X_i = 0$ if $D_i = 0$ and X_i else, can be seen as the output of the following mechanism:

(O1) Exogenous variables $\{Z_i, X_i\}_{i=1}^N$, possibly partially latent, are generated by a completely unknown DGP.

(O2) Stochastic errors $\{\varepsilon_i\}_{i=1}^N$ are drawn i.i.d. from the known probability distribution of ε, with all draws independent of $\{Z_i, X_i\}_{i=1}^N$.

(O3) Endogenous variables $\{Y_i\}_{i=1}^N$ are observed as a result of the DGP: $Y_i = r(Z_i, X_i, \varepsilon_i; \theta^0)$, with θ^0 the true unknown value of the structural parameters.

(O4) $\{D_i\}_{i=1}^N$ is drawn in the product of conditional distributions of D_i given $\{Y_i, Z_i\}_{i=1}^N$. Moreover, these conditional distributions, for $i = 1, ..., N$, are assumed (by **MAR-X**) not to depend on X_i when conditioned on W_i.

For the simulated data, a similar procedure is implicitly considered when simulating the endogenous variables. However, unlike step **(O2)** above, for some integer $S \geq 1$ we draw $s = 1, ..., S$ independent simulated samples of i.i.d. errors $\{\tilde{\varepsilon}_{is}\}_{i=1}^N$ from the known probability distribution of ε with $\{\tilde{\varepsilon}_{is}\}_{i=1}^N$ independent of $\{\varepsilon_i, Z_i, X_i, D_i\}_{i=1}^N$ by construction. Given, $\{\tilde{\varepsilon}_{is}\}_{i=1}^N$, and in accordance with **(O3)** above, we define, for all $\theta \in \Theta$: $Y_{is}(\theta) = r(Z_i, X_i, \tilde{\varepsilon}_{is}; \theta)$. The s-th simulation step produces a sequence $\{Z_i, D_i X_i, \varepsilon_i, D_i, \tilde{\varepsilon}_{is}\}_{i=1}^N$ of i.i.d. draws in a joint distribution that defines, through known transformations, the joint distribution of the variables at stake to compute the expectation in (11).

However, since $\{\tilde{\varepsilon}_{is}\}_{i=1}^N$ is independent of $\{\varepsilon_i, Z_i, X_i, D_i\}_{i=1}^N$, it is also independent of the missing data mechanism encapsulated by $\{D_i\}_{i=1}^N$. Therefore, D_i is endowed with an exogeneity status in regards to the *simulated errors* $\tilde{\varepsilon}_{is}$. In particular, because D_i is independent of $\tilde{\varepsilon}_{is}$, for each $s = 1, ..., S$, given $(\varepsilon_i, Z_i, X_i)$

since $\tilde{\varepsilon}_{is}$ jointly independent of $(\varepsilon_i, Z_i, X_i, D_i)$, we have that

$$D_i \perp Y_{is}(\theta) \;\big|\; W_i, X_i. \tag{12}$$

Note, however, that we can not compute $Y_{is}(\theta)$ when $D_i = 0$.

With the conditional independence $D \perp Y(\theta) \;\big|\; W, X$ generated through the simulation step, we can demonstrate the validity of equation (11):

$$\begin{aligned}
E\left[\frac{D}{p_0(W)} m(Y(\theta), Z, X, \beta)\right] &= E\left[E\left[\frac{D}{p_0(W)} m(Y(\theta), Z, X, \beta)\middle| W, X\right]\right] \text{ (by L.I.E.)} \\
&= E\left[E\left[\frac{D}{p_0(W)}\middle| W, X\right] E\left[m(Y(\theta), Z, X, \beta)|W, X\right]\right] \text{ (by (12))} \\
&= E\left[E\left[\frac{D}{p_0(W)}\middle| W\right] E\left[m(Y(\theta), Z, X, \beta)|W, X\right]\right] \text{ (by } \mathbf{MAR\text{-}X}\text{)} \\
&= E\left[m(Y(\theta), Z, X, \beta)\right] \text{ (by definition of } p_0(W)\text{)}.
\end{aligned} \tag{13}$$

Equation (13) is precisely what we need to ensure an II approach that is feasible and valid, when based on the auxiliary model (1), in spite of the missing data problem.[3]

More precisely, for $\beta^0 \in \mathcal{B}$ and $\theta \mapsto \beta^0(\theta)$ defined by, respectively,

$$\begin{aligned}
E\left[m(Y, Z, X, \beta^0)\right] &= 0, \\
E\left[m(Y(\theta), Z, X, \beta^0(\theta))\right] &= 0,
\end{aligned}$$

under the standard identification assumption $\beta^0 = \beta^0(\theta) \Leftrightarrow \theta = \theta^0$, comparison of (10) and (11) suggests a feasible II approach with missing data using IPW moment conditions

$$E\left[\frac{D}{p_0(W)} m(Y, Z, X, \beta^0)\right] = 0 \tag{14}$$

$$E\left[\frac{D}{p_0(W)} m(Y(\theta), Z, X, \beta^0(\theta))\right] = 0. \tag{15}$$

This is where the novelty of our approach lies. Identification of θ^0, by means of (10) and (11), which is valid by (13), does not result directly from the use of IPW and **MAR-X** in (2) but also requires the conditional independence introduced through the simulation step, i.e., (12).

Clearly, implementation of the above strategy requires consistent estimation of $p_0(W)$. The complete

[3] Note that independence between $\tilde{\varepsilon}$ and (ε, Z, X), as in a standard II context, is insufficient to yield equation (13).

asymptotic theory will be developed in Section three in the framework of a fully parametric model. This parametric model will define the set of possible DGPs according to steps **(O1), (O2), (O3)** and **(O4)** above, augmenting it by a parametric specification $p(W;\gamma)$ for $\Pr[D = 1|W]$ in step **(O4)** such that

$$p_0(W) = \Pr[D = 1\,|W] \equiv p(W;\gamma^0)$$

for a unique $\gamma^0 \in \text{Interior}(\Gamma) \subset \mathbb{R}^{d_\gamma}$.

The reader familiar with nonparametric estimation of optimal instruments knows that an alternative solution to using estimating equations like (14)-(15) would be to come up with a consistent, albeit nonparametric, estimator of $p_0(W)$. However, in the context of IPW, this would pave the way for new discussions about efficient II with missing data: Chen et al. (2008) (see also Graham, 2011 and Chaudhuri et al., 2016) show that a nonparametric estimator of $p_0(W)$ would actually lead, in general, to a more accurate estimator of β^0. This apparent paradox is easy to explain when one realizes that step **(O4)** provides a set of conditional moment restrictions

$$E[D - p_0(W)\,|W] = 0.$$

These conditional moment restrictions would in general bring relevant information about the unknown parameter β beyond what can be brought by the parametric model $p(W;\gamma)$. It must be kept in mind that maximum likelihood estimation of such a parametric model amounts to picking a subset (of dimension d_γ) of the above conditional moment restrictions,

$$E\left[\frac{D - p_0(W)}{p_0(W)(1 - p_0(W))}\frac{\partial}{\partial\gamma}p(W;\gamma^0)\right] = 0$$

that is optimal for estimation of γ^0. However, this subset of the conditional moments may not exhaust all relevant information for optimal estimation of β^0. While our focus of interest is not efficient direct estimation of β but indirect estimation of θ, obviously, the two efficiency issues are tightly related but much beyond the scope of this paper.

3 IPW Indirect Inference (IPW-II)

In this section we discuss precise implementation of our inverse probability weighted II (IPW-II) approach under **MAR-X** missing data. Asymptotic properties of the ensuing IPW-II approach are discussed in Section 3.4. Section 3.5 presents a computationally friendly implementation of this approach for non-smooth problems using the generalized II approach first proposed in Keane and Smith (2005).

3.1 Estimation of the Auxiliary Model Parameters

3.1.1 Observed Data

Following the identification strategy outlined in Section two, define the estimator $\widehat{\beta}_N$ as the solution of

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{p(W_i,\widehat{\gamma}_N)}m(Y_i,Z_i,X_i,\widehat{\beta}_N) \quad = \quad 0,$$

where $\widehat{\gamma}_N$ is the maximum likelihood estimator, the solution to $0=\sum_{i=1}^{N} l_{i,\gamma}(\gamma)$, with $l_{i,\gamma}(\gamma)=l_\gamma(D_i,W_i,\gamma)$ the score vector of the parametric model describing the missing data mechanism:

$$l_{i,\gamma}(\gamma)=\frac{\partial}{\partial\gamma}\log\left[(p(W_i,\gamma))^{D_i}\left(1-p(W_i,\gamma)\right)^{1-D_i}\right]=\frac{[D_i-p(W_i,\gamma)]}{p(W_i,\gamma)(1-p(W_i,\gamma))}\frac{\partial p(W_i,\gamma)}{\partial\gamma}.$$

Note that $(\widehat{\beta}_N',\widehat{\gamma}_N')'$ can also be seen as a joint GMM estimator provided by the just identified moment conditions

$$\begin{aligned} E\left[\frac{D_i}{p(W_i;\gamma)}m(Y_i,Z_i,,X_i,\beta)\right] &= 0 \\ E\left[l_{i,\gamma}(D_i,W_i,\gamma)\right] &= 0 \end{aligned}$$

It is well known (see, e.g., Breusch et al., 1999, Lemma 1, p93) that we can obtain an asymptotically equivalent GMM estimator by instead considering the moment conditions:

$$\begin{aligned} E\left[m_i^*(\gamma,\beta)-\Pi\left[m_i^*(\gamma,\beta)\mid l_{i,\gamma}(\gamma)\right]\right] &= 0 \\ E\left[l_\gamma(D_i,W_i,\gamma)\right] &= 0 \end{aligned} \tag{16}$$

where $m_i^*(\gamma,\beta)=\frac{D_i}{p(W_i;\gamma)}m(Y_i,Z_i,X_i,\beta)$ and, for $\beta=\beta^0$ and $\gamma=\gamma^0$, $\Pi\left[m_i^*(\gamma^0,\beta^0)\mid l_{i,\gamma}(\gamma^0)\right]$ the affine population regression of $m_i^*(\gamma^0,\beta^0)$ on $l_{i,\gamma}(\gamma^0)$:

$$\begin{aligned} \Pi\left[m_i^*(\gamma^0,\beta^0)\mid l_{i,\gamma}(\gamma^0)\right] &= \Omega_{12}\Omega_{22}^{-1}l_{i,\gamma}(\gamma^0), \\ \Omega_{12} &= \text{Cov}\left[m_i^*(\gamma^0,\beta^0),l_{i,\gamma}(\gamma^0)\right],\ \ \Omega_{22}=\text{Var}\left[l_{i,\gamma}(\gamma^0)\right]. \end{aligned}$$

Clearly, the two moments in (16) are uncorrelated at γ^0,β^0, allowing us to compute directly the asymptotic distribution of the GMM estimator $\widehat{\beta}_N$ from its asymptotic expansion[4]

$$\begin{aligned}\sqrt{N}\left(\widehat{\beta}_N-\beta^0\right) &= -\left[G_0'V_0^{-1}G_0\right]^{-1}G_0'V_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{m_i^*(\gamma^0,\beta^0)-\Pi\left[m_i^*(\gamma^0,\beta^0)\mid l_{i,\gamma}(\gamma^0)\right]\right\}+o_P(1)\\ &= -G_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{m_i^*(\gamma^0,\beta^0)-\Pi\left[m_i^*(\gamma^0,\beta^0)\mid l_{i,\gamma}(\gamma^0)\right]\right\}+o_P(1),\end{aligned}$$

where

$$G_0 = E\left[\frac{D_i}{p(W_i;\gamma^0)}\frac{\partial m(W_i,X_i,\beta^0)}{\partial\beta'}\right], \tag{17}$$

$$V_0 = \text{Var}\left[m_i^*(\gamma^0,\beta^0)\right]-\text{Var}\left[\Pi\left[m_i^*(\gamma^0,\beta^0)\mid l_{i,\gamma}(\gamma^0)\right]\right]. \tag{18}$$

Remarks:

(1) Applying again the **MAR-X** property to G_0 yields

$$G_0 = E\left[E\left[\frac{D_i}{p(W_i;\gamma^0)}\middle|W_i,X_i\right]\frac{\partial m(W_i,X_i,\beta^0)}{\partial\beta'}\right] = E\left[\frac{\partial m(W_i,X_i,\beta^0)}{\partial\beta'}\right],$$

as if we had no missing data. However, due to the missing data problem, the formula (17) provides the natural way to estimate G_0 from its sample counterpart after plugging in consistent estimators of γ^0,β^0. Similarly,

$$\text{Var}\left[m_i^*(\gamma^0,\beta^0)\right] = E\left[m_i^*(\gamma^0,\beta^0)m_i^*(\gamma^0,\beta^0)'\right] = E\left[\frac{1}{p(W_i;\gamma^0)}m(W_i,X_i,\beta^0)m'(W_i,X_i,\beta^0)\right]$$

should be estimated from the sample counterpart of (18) rather than the above equation. However, the division by $p(W_i;\gamma^0)$ in the above shows the price we pay, in terms of accuracy of $\widehat{\beta}_N$, for the missing data problem.

(2) The asymptotic variance of $\sqrt{N}(\widehat{\beta}_N-\beta^0)$, given by $G_0^{-1}V_0G_0'^{-1}$, is smaller (in terms of comparison of positive semi-definite matrices) than the asymptotic variance of a GMM estimator for β^0 obtained using the true unknown propensity score $p_0(W)=p(W;\gamma^0)$ and only

$$E\left[\frac{D_i}{p_0(W_i)}m_i(Y,Z,X,\beta)\right]=0, \tag{19}$$

for which the resulting asymptotic variance would be given by $G_0^{-1}\left[\text{Var}\left\{m_i^*(\gamma^0,\beta^0)\right\}\right]G_0'^{-1}$.

[4] Precise regularity conditions ensuring the validity of this expansion are given as Assumptions **A1**-**A5** in the appendix.

This remark is sometimes summarized by a kind of puzzling statement: "it is better to estimate the weights by a conditional MLE procedure than using known weights (if we knew them)" (Wooldridge, 2007). The explanation of this anomalous statement is simple: we take advantage of the moments provided by the score vector $l_{i,\gamma}(\gamma^0)$ to reduce the variance of $m_i^*(\gamma^0, \beta^0)$ by computing the residual of its regression on $l_{i,\gamma}(\gamma^0)$. The possible efficiency loss when using GMM based only on (19) instead of the GMM estimator $\widehat{\beta}_N$ is not due to the knowledge of γ^0 but to the omission of the second set of moment conditions $l_{i,\gamma}(\gamma^0)$.

3.1.2 Simulated Data

For a given integer $S \geq 1$, we draw $s = 1, ..., S$ independently simulated samples of i.i.d. errors $\{\tilde{\varepsilon}_{is}\}_{i=1}^N$ from the known probability distribution of ε with, for each $s = 1, ..., S$, $\{\tilde{\varepsilon}_{is}\}_{i=1}^N$ independent of $\{\varepsilon_i, Z_i, X_i, D_i\}_{i=1}^N$. We can then compute $Y_{is}(\theta) = r\left(Z_i, X_i, \tilde{\varepsilon}_{is}; \theta\right)$ and define the estimator $\tilde{\beta}_{N,s}(\theta)$ as the solution of

$$\sum_{i=1}^N \frac{D_i}{p(W_i, \widehat{\gamma}_N)} m\left(Y_{is}(\theta), Z_i, X_i, \tilde{\beta}_{N,s}(\theta)\right) = 0.$$

Following similar arguments to those developed in the previous section, when N is large, (see also (22))

$$\sqrt{N}\left(\tilde{\beta}_{N,s}(\theta^0) - \beta^0\right) = -G_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \left\{ m_{is}^*(\gamma^0, \beta^0; \theta^0) - \Pi\left[m_{is}^*(\gamma^0, \beta^0; \theta^0) \mid l_{i,\gamma}(\gamma^0)\right] \right\} + o_P(1), \tag{20}$$

where $m_{is}^*(\gamma, \beta; \theta) = \frac{D_i}{p(W_i;\gamma)} m(Y_{is}(\theta), Z_i, X_i, \beta)$,

$$\begin{aligned} \Pi\left[m_{is}^*(\gamma^0, \beta^0; \theta) \mid l_{i,\gamma}(\gamma^0)\right] &= \Omega_{12}(\theta)\Omega_{22}^{-1} l_{i,\gamma}(\gamma^0), \\ \Omega_{12}(\theta) &= \text{Cov}\left[m_{is}^*(\gamma^0, \beta^0; \theta), l_{i,\gamma}(\gamma^0)\right] \end{aligned}$$

Note that $\Omega_{12}(\theta)$ does not depend on (i, s) since all draws of $(Z, X, \tilde{\varepsilon}_{is})$ are drawn in the same distribution, which corresponds to the distribution of $(Z_i, X_i, \varepsilon_i)$. However, it is critical to note that,

$$\Omega_{12}(\theta^0) = \text{Cov}[m_{is}^*(\gamma^0, \beta^0; \theta^0), l_{i,\gamma}(\gamma^0)] \neq \text{Cov}[m_i^*(\gamma^0, \beta^0), l_{i,\gamma}(\gamma^0)] = \Omega_{12},$$

which follows from the fact that, for $\mathcal{D}\left(\varepsilon_i, Z_i, X_i, D_i\right)$ the joint probability distribution of $(\varepsilon_i, Z_i, X_i, D_i)$, in general

$$\mathcal{D}\left(\varepsilon_i, Z_i, X_i, D_i\right) \neq \mathcal{D}\left(\tilde{\varepsilon}_i, Z_i, X_i, D_i\right).$$

This discrepancy is a consequence of the missingness indicator D_i being exogenous with regards to the simulated errors $\tilde{\varepsilon}_{is}$, since $\tilde{\varepsilon}_{is}$ is, by construction, independent of (Z_i, X_i, D_i). In contrast to $\tilde{\varepsilon}_{is}$, **MAR-X** does not require that the error ε_i be independent of (Z_i, X_i, D_i) since D_i need not be independent of Y_i given (Z_i, X_i).

Following the first II estimator of Gourieroux et al. (1993) (see their Proposition 1 p. S89), an estimator of θ^0 can be obtained by calibrating the value of θ in order to minimize the distance, in some norm, between $\widehat{\beta}_N$ and the average value of the simulated auxiliary estimators

$$\bar{\beta}_{N,S}(\theta) = \frac{1}{S} \sum_{s=1}^{S} \tilde{\beta}_{N,s}(\theta).$$

From $\bar{\beta}_{N,S}(\theta)$ and (20), we deduce that, for given fixed S,

$$\sqrt{N}\left(\bar{\beta}_{N,S}(\theta^0) - \beta^0\right) = -G_0^{-1} \frac{\sqrt{N}}{N \cdot S} \sum_{i=1}^{N} \sum_{s=1}^{S} \left\{ m^*_{is}(\gamma^0, \beta^0; \theta^0) - \Pi\left[m^*_{is}(\gamma^0, \beta^0; \theta^0) \;\middle|\; l_{i,\gamma}(\gamma^0) \right] \right\} + o_P(1).$$

3.2 The Calibration Step: Wald-type IPW-II

Given auxiliary parameter estimates $\widehat{\beta}_N$, $\bar{\beta}_{N,S}(\theta)$, a Wald-type IPW-II estimator for θ^0 is obtained as

$$\widehat{\theta}^W_N(\Upsilon) := \arg\min_{\theta \in \Theta} \left[\widehat{\beta}_N - \bar{\beta}_{N,S}(\theta)\right]' \Upsilon_N^{-1} \left[\widehat{\beta}_N - \bar{\beta}_{N,S}(\theta)\right], \tag{21}$$

where Υ_N is a sequence of positive-definite weighting matrix. The notation $\widehat{\theta}^W_N(\Upsilon)$ stresses that the asymptotic distribution of this estimator depends on the choice of the weighting matrix Υ_N^{-1}.

Standard arguments for minimum distance estimators tell us that the optimal choice of weighting matrix is to take $\Upsilon_N \to_P \Upsilon(S)$, where

$$\Upsilon(S) := \lim_{N \to \infty} \mathrm{Var}\left\{ \sqrt{N}\left(\widehat{\beta}_N - \bar{\beta}_{N,S}(\theta^0)\right) \right\}.$$

To deduce the form of $\Upsilon(S)$, we first use the expansions $\sqrt{N}(\widehat{\beta}_N - \beta^0)$ and $\sqrt{N}(\bar{\beta}_{N,S}(\theta^0) - \beta^0)$, given in the previous subsections, to find

$$\sqrt{N}\left(\bar{\beta}_N - \widehat{\beta}_{N,S}(\theta^0)\right) = -G_0^{-1}\sqrt{N}\left[\bar{\xi}_{N,S} - C_0 \Omega_{22}^{-1} \sum_{i=1}^{N} l_{i,\gamma}(\gamma^0)/N\right] + o_P(1),$$

where $C_0 = \left[\Omega_{12} - \Omega_{12}(\theta^0)\right]$ and

$$\bar{\xi}_{N,S} = \frac{1}{N}\sum_{i=1}^{N} \xi_{i,S} \equiv \frac{1}{N}\sum_{i=1}^{N}\left[m_i^*(\gamma^0,\beta^0) - \frac{1}{S}\sum_{s=1}^{S} m_{is}^*(\gamma^0,\beta^0;\theta^0)\right].$$

Noting that,

$$\text{Cov}\left(\sqrt{N}\bar{\xi}_{N,S}\ ,\ C_0\Omega_{22}^{-1}\sum_{i=1}^{N} l_{i,\gamma}(\gamma^0)/\sqrt{N}\right) = C_0\Omega_{22}^{-1}C_0' = \text{Var}\left(C_0\Omega_{22}^{-1}\sum_{i=1}^{N} l_{i,\gamma}(\gamma^0)/\sqrt{N}\right)$$

and for $W_0(S) = \lim_{N\to\infty} \text{Var}\left\{\sqrt{N}\bar{\xi}_{N,S}\right\} = E\left[\xi_{i,S}\cdot\xi_{i,S}'\right]$, $\Upsilon(S)$ then has the following form:

$$\Upsilon(S) = G_0^{-1}\left[W_0(S) - C_0\Omega_{22}^{-1}C_0'\right]G_0^{-1\prime}.$$

Remarks:

(1) The term $W_0(S)$ in $\Upsilon(S)$ can further be decomposed by noting the following. One,

$$\begin{aligned}\text{Var}\left[m_{is}^*(\gamma^0,\beta^0;\theta^0)\right] &= E\left[\frac{D_i}{p^2(W_i;\gamma^0)} m(Y_{is}(\theta^0),Z_i,X_i,\beta^0)m'(Y_{is}(\theta^0),Z_i,X_i,\beta^0)\right] \\ &= E\left[\frac{1}{p(W_i;\gamma^0)} m(Y_{is}(\theta^0),Z_i,X_i,\beta^0)m'(Y_{is}(\theta^0),Z_i,X_i,\beta^0)\right] = \text{Var}\left[m_i^*(\gamma^0,\beta^0)\right],\end{aligned}$$

where the second equality comes from an argument similar to the one used to prove (13) and the third equality is implied by the fact that the joint distributions satisfy

$$\mathcal{D}\left(\varepsilon_i, Z_i, X_i\right) = \mathcal{D}\left(\tilde{\varepsilon}_{is}, Z_i, X_i\right), \tag{22}$$

with this distributional equivalence being (partly) why simulated and observed expectations can still coincide, such as, e.g., G_0 defined in (17). Two, by the same logic, for $s, s' = 1, ..., S$

$$\text{Cov}\left[m_i^*(\gamma^0,\beta^0), m_{is}^*(\gamma^0,\beta^0;\theta^0)\right] = \text{Cov}\left[m_{is}^*(\gamma^0,\beta^0;\theta^0), m_{is'}^*(\gamma^0,\beta^0;\theta^0)\right].$$

Introducing the notations,

$$I_0 = \text{Var}\left[m_i^*(\gamma^0,\beta^0)\right], \quad K_0 = \text{Cov}\left[m_i^*(\gamma^0,\beta^0), m_{is}^*(\gamma^0,\beta^0;\theta^0)\right],$$

elementary algebra then yields a familiar form (see Gourieroux et al., 1993, pS109):[5]

$$W_0(S) = \left(1 + \frac{1}{S}\right)(I_0 - K_0).$$

(2) The term K_0 can be further decomposed, by noting that, for $s, s' = 1, ..., S$, even if $s = s'$,

$$\begin{aligned} K_0 &= \text{Cov}\left\{E[m^*_{is}(\gamma^0, \beta^0; \theta^0)\,|Z_i, D_i X_i]\,, E[m^*_{is'}(\gamma^0, \beta^0; \theta^0)\,|Z_i, D_i X_i]\right\} \\ &= \text{Var}\left\{E[m^*_{is}(\gamma^0, \beta^0; \theta^0)\,|Z_i, D_i X_i]\right\} = \text{Var}\left\{E[m^*_{i}(\gamma^0, \beta^0)\,|Z_i, D_i X_i, D_i]\right\} \end{aligned}$$

which yields the following alternative specification for $I_0 - K_0$:

$$I_0 - K_0 = Var\left\{m^*_i(\gamma^0, \beta^0) - E[m^*_i(\gamma^0, \beta^0)\,|Z_i, D_i X_i, D_i]\right\}.$$

This expression makes explicit the efficiency gain due to the fact that we have not simulated (Z, X, D).

3.3 Alternative IPW-II Implementation

It is well-known that the different approaches to choosing a metric between $\widehat{\beta}_N$ and $\bar{\beta}_{N,S}(\theta)$, for the purpose of II on θ, correspond to the trinity of asymptotic tests. As summarized by Bruins et al. (2015), following a nomenclature "due to Eric Renault, the Wald and LR [likelihood ratio] approaches were first proposed in Smith (1990, 1993) and later extended by GMR [Gourieroux et al., 1993]. The LM [Lagrange multiplier] approach was first proposed in Gallant and Tauchen (1996)."

While Gourieroux et al. (1993) have stressed that the LR approach may imply some efficiency loss, they also show (see their Section 2.5) that, as far as first order asymptotics are concerned, the family of Wald-II estimators coincides with the family of LM-II estimators. However, due to the fact that our analysis depends on the matrix $C_0 = \Omega_{12} - \Omega_{12}(\theta^0) \neq 0$, in general, their results may not be directly applicable to our IPW-II estimators.

In addition, Gourieroux et al. (1993) consider the LM and LR approaches only in the context where the moment conditions used to estimate the auxiliary parameters are defined by the gradient of some objective function, like a pseudo-score. In such cases, G_0 is a symmetric Hessian matrix, and in some circumstances this Hessian matrix may coincide with the outer product matrix I_0. Since we only consider parameters defined as zeros of just-identified moment conditions, the Jacobian matrix G_0 is nonsingular but G_0 has no

[5] The term K_0 is non-zero in general because the observed and simulated samples both have in common the exogenous variables X and Z.

reason to coincide with the symmetric, positive-definite matrix I_0.

For a sequence of positive-definite weighting matrices $\Psi_N \to_p \Psi$, with Ψ positive-definite, a general Wald IPW-II estimator $\widehat{\theta}_N^W(\Psi)$ of θ can be defined as the solution to the minimization program:

$$\widehat{\theta}_N^W(\Psi) = \arg\min_{\theta\in\Theta} \left[\widehat{\beta}_N - \bar{\beta}_{N,S}(\theta)\right]' \Psi_N \left[\widehat{\beta}_N - \bar{\beta}_{N,S}(\theta)\right].$$

From Section 3.2, and standard argument for minimum distance estimation, the optimal choice Ψ^* of Ψ is given by $\Psi^* = \Upsilon^{-1}(S) \equiv G_0' H_0^{-1}(S) G_0$, where

$$H_0(S) = W_0(S) - C_0\Omega_{22}^{-1}C_0', \;\; W_0(S) = \left(1+\frac{1}{S}\right)[I_0 - K_0]$$

In the case where $C_0 = 0$, and G_0 a Hessian matrix, Gourieroux et al. (1993) demonstrate that LR-type II estimators are asymptotically equivalent to $\widehat{\theta}_N^W(G_0)$ and not efficient in general since

$$G_0 \neq G_0 \left[I_0 - K_0\right]^{-1} G_0'.$$

The Wald-type II estimator $\widehat{\theta}_N^W(\Psi)$ can be computationally expensive when $\bar{\beta}_{N,S}(\theta)$ is not known in closed form. Even though the moment conditions of the auxiliary model are not necessarily defined as a gradient vector, we can extend the original LM-type II approach by defining a LM-type IPW-II estimator $\widehat{\theta}_N^{LM}(A)$ as the solution of the minimization program:

$$\widehat{\theta}_N^{LM}(A) = \arg\min_{\theta\in\Theta} \left[M_{N,S}\left(\widehat{\beta}_N, \widehat{\gamma}_N, \theta\right)\right]' A_N \left[M_{N,S}\left(\widehat{\beta}_N, \widehat{\gamma}_N, \theta\right)\right],$$

where A_N is a sequence of positive-definite matrices with probability limit A and

$$M_{N,S}\left(\widehat{\beta}_N, \widehat{\gamma}_N, \theta\right) = \frac{1}{N\cdot S}\sum_{i=1}^{N}\sum_{s=1}^{S}\frac{D_i}{p(W_i;\widehat{\gamma}_N)} m(Y_{is}(\theta), Z_i, X_i, \widehat{\beta}_N).$$

When $C_0 = 0$, Gourieroux et al. (1993) have shown that the estimator $\widehat{\theta}_N^{LM}(A)$ is asymptotically equivalent to $\widehat{\theta}_N^W(G_0 A G_0')$. The extension to our more general case is straightforward. In particular, the optimal choice A^* of A is $$A^* = H_0^{-1}(S).$$

Exact implementation of the LM-type IPW-II approach can be carried out using the following algorithm, which deals with potential non-smoothness, in θ, of $M_{N,S}\left(\widehat{\beta}_N, \widehat{\gamma}_N, \theta\right)$.[6]

[6] Such situations arise in simulation-based estimation of discrete choice models because the simulated dependent variable, as

Algorithm for implementing of IPW-II

- **Step 0:** Using the observed $\{W_i, D_i\}_{i=1}^N$ estimate $\widehat{p}_0(W_i) := p(W_i; \widehat{\gamma}_N)$ for each $i = 1, \ldots, N$ where $\widehat{\gamma}_N$ is a the maximum likelihood estimator based on any *given* parametric specification $p(W; \gamma)$ for $p_0(W)$, and where γ is some $d_\gamma \times 1$ unknown parameter.

- **Step 1:** Using the observed sample $\{W_i, D_i, D_i X_i\}_{i=1}^N$, obtain $\widehat{\beta}_N$ as:
$\widehat{\beta}_N \ := \ \arg_{\beta \in \mathcal{B}} \left\{ \frac{1}{N} \sum_{i=1}^N \frac{D_i}{p(W_i;\gamma)} m(Y_i, Z_i, X_i, \beta) = 0 \right\},$

- **Step 2a:** Sort the observed sample so that the first $N_1 = \sum_{i=1}^N D_i$ units have $D_i = 1$, i.e., have X_i observed. For any given $\theta \in \Theta$, and for each $i = 1, \ldots, N_1$, generate:

$$\widetilde{\varepsilon}_{is} \overset{\text{i.i.d.}}{\sim} F_\varepsilon^0, \ \ Y_{is}(\theta) = r(Z_i, X_i, \theta, \widetilde{\varepsilon}_{is}) \text{ for } s = 1, \ldots, S$$

  where S is the pre-specified number of simulations. Set $Y_{is}(\theta) = 0$ for $s = 1, \ldots, S$ and $i = N_1 + 1, \ldots, N$. (This is inconsequential because we will not use these remaining i's.)

- **Step 2b:** For any given positive definite matrix A_N, obtain the II estimator $\widehat{\theta}_N^{LM}(A)$ as:

$$\left\| M_{N,S} \left(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N^{LM}(A) \right) \right\|_{A_N} \ \leq \ o_P \left(N^{-1/2} \right) + \inf_{\theta \in \Theta} \left\| M_{N,S} \left(\widehat{\beta}_N, \widehat{\gamma}_N, \theta \right) \right\|_{A_N} . \tag{23}$$

  We call $\widehat{\theta}_N^{LM}(A)$ a LM-type IPW-II estimator of θ^0.

Remarks:

(1) The IPW-II procedure models $p_0(W)$ parametrically and is susceptible to misspecification. Adverse consequences of parametric misspecification of $p_0(W)$ in Step 0, and remedy thereof by using doubly robust estimating functions for β or by nonparametric estimation of $p_0(W)$ have been studied for general direct IPW estimators [e.g., Scharfstein et al. (1999), Hirano et al. (2003), Chen et al. (2008)]. Extensions of these results to indirect estimators is considered in Chaudhuri et al. (2016).

(2) The optimal choice of $A = \text{plim} A_N$ follows from Gourieroux et al. (1993) with an additional modification due to the fact that the nuisance parameter $p_0(W)$ is estimated. Even with the optimal A, the relative efficiency of the II estimator of θ with respect to the full information maximum likelihood estimator ultimately depends on the "richness" of the auxiliary model. Bruins et al. (2015) provide an illuminating demonstration with simulations.

a function of θ, i.e., $Y(\theta)$, can change discretely (e.g. from 0 to 1) with an infinitesimal change in θ.

3.4 Asymptotic Distribution of the IPW-II Estimator

In this subsection, we provide precise results for consistency and asymptotic normality of the IPW-II estimator $\widehat{\theta}_N^{LM}(A)$ in (23).[7] We deviate from the standard II treatment and present results that accommodate non-smoothness with respect to θ in the moment vector $m(Y(\theta), Z, X, \beta)$, as in, e.g., discrete choice models. The required technical assumptions **A1-A7**, along with the proofs of the stated results, which are similar in spirit to and based on Pakes and Pollard (1989), are collected in the Appendix.

Proposition 3.1 *Let **A1**-**A6**(1) in the Appendix hold. Let S be fixed and $A_N \xrightarrow{P} A$ as $N \to \infty$ where A is positive definite. Then the IPW-II estimator in* (23) *satisfies:* $\widehat{\theta}_N^{LM}(A) \xrightarrow{P} \theta^0$.

Proposition 3.2 *Let Assumptions **A1**-**A7** in the Appendix hold. Let S be fixed and $A_N \xrightarrow{P} A$ as $N \to \infty$ where A is symmetric and positive definite. Let $\frac{\partial}{\partial\theta'}\beta^0(\theta^0)$ be full column rank. Then the IPW-II estimator in* (23) *satisfies:* $\sqrt{N}(\widehat{\theta}_N^{LM}(A) - \theta^0) \xrightarrow{d} N(0, \Sigma(A))$ *where:*

$$\begin{aligned}
\Sigma(A) &:= \left[\frac{\partial\beta^0(\theta^0)'}{\partial\theta} G_0' A G_0 \frac{\partial\beta^0(\theta^0)}{\partial\theta'}\right]^{-1} \frac{\partial\beta^0(\theta^0)'}{\partial\theta} G_0' A H_0(S) A G_0 \frac{\partial\beta^0(\theta^0)}{\partial\theta'} \left[\frac{\partial\beta^0(\theta^0)'}{\partial\theta} G_0' A G_0 \frac{\partial\beta^0(\theta^0)}{\partial\theta'}\right]^{-1}, \\
H_0 &:= W_0(S) - C_0\Omega_{22}^{-1}C_0' \equiv \left(1 + \frac{1}{S}\right)[I_0 - K_0] - C_0\Omega_{22}^{-1}C_0'
\end{aligned}$$

Remarks: (1) The optimal A is $A^* = H_0^{-1}(S)$. Hence, the optimal asymptotic variance given the auxiliary model is: $\Sigma(A^*) = \left[\frac{\partial\beta^0(\theta^0)'}{\partial\theta} G_0' H_0^{-1}(S) G_0 \frac{\partial\beta^0(\theta^0)}{\partial\theta'}\right]^{-1}$. The missing X and the estimation of the nuisance parameters γ to model this missingness make this optimal asymptotic variance different from that given in Proposition 4 of Gourieroux et al. (1993). Without the former, the term $W_0(S)$ in $H_0(S)$ would reduce to the standard definitions given in Gourieroux et al. (1993); i.e., $\xi_{i,S}$, defining the asymptotic expansion of $\sqrt{N}(\widehat{\beta}_N - \widehat{\beta}_{N,S}(\theta^0))$, and hence, $\sqrt{N}(\widehat{\theta}_N^{LM}(A) - \theta^0)$, would reduce to $m(Y_i, Z_i, X_i, \beta^0) - \frac{1}{S}\sum_{s=1}^{S} m(Y_{is}(\theta^0), Z_i, X_i, \beta^0)$. Without the latter, $C_0\Omega_{22}^{-1}C_0'$ would not appear.

(2) The matrix $H_0(S)$ can be written in the equivalent form

$$\begin{aligned}
H_0(S) &:= E\left[\left(\xi_{i,S} - \Pi[\xi_{i,S}|l_{i,\gamma}(\gamma^0)]\right)\left(\xi_{i,S} - \Pi[\xi_{i,S}|l_{i,\gamma}(\gamma^0)]\right)'\right], \\
\xi_{i,S} &:= \frac{D_i}{p_0(W_i)}\left[m(Y_i, Z_i, X_i, \beta^0) - \frac{1}{S}\sum_{s=1}^{S} m(Y_{is}(\theta^0), Z_i, X_i, \beta^0)\right],
\end{aligned}$$

where $l_{i,\gamma}(\gamma)$ is the score of the missingness likelihood, and $\Pi[\xi_{i,S}|l_{i,\gamma}(\gamma^0)]$, stands for the affine regression of $\xi_{i,S}$ on $l_{i,\gamma}(\gamma^0)$. Using this formula the optimal asymptotic variance of the IPW-II estimator can be stated

[7] Equivalent results can be obtained for the Wald-type IPW-II estimator. However, the arguments mirror those in Gourieroux et al. (1993) and so are not presented for brevity.

as

$$\left[\frac{\partial\beta^0(\theta^0)'}{\partial\theta}G_0'\left\{E\left[\left(\xi_{i,S}-\Pi[\xi_{i,S}|l_{i,\gamma}(\gamma^0)]\right)\left(\xi_{i,S}-\Pi[\xi_{i,S}|l_{i,\gamma}(\gamma^0)]\right)'\right]\right\}^{-1}G_0\frac{\partial\beta^0(\theta^0)}{\partial\theta'}\right]^{-1}.$$

The above formula is similar to existing formulas describing the asymptotic variance of estimators in the presence of missing data, see, e.g., Wooldridge (2007).

(3) The IPW-II estimator is based on inverse probability weighting the so called "complete cases", i.e., sample units with no missing variables, to correct for the endogenous missingness/selection. This makes it widely applicable to scenarios where the pattern of missingness is more complex [see Little and Rubin (2002)]. For example, let $X=(X_1',X_2')'$ and suppose we observe $(Y',Z')'$ for some sample units, $(Y',Z',X_1')'$ for some and $(Y',Z',X')'$ for the rest. This is a scenario of monotonic pattern in missingness. If there is another subset of the sample units where we observe $(Y',Z',X_2')'$, then this is a scenario of non-monotonic pattern in missingness. The above algorithm can be directly applied under both scenarios since it works with the "complete cases" only, i.e, sample units for which we observe $(Y',Z',X')'$. However, the estimator will not be semiparameterically efficient in the sense of Robins et al. (1994) and Robins and Rotnitzky (1995).

Since the driving force behind the potential loss in efficiency related to Remarks (2) and (3) above are well understood now, we abstract from such efficiency considerations to keep this paper short.

3.5 Smoothed Implementation: IPW-GII Estimator

Implementation of the IPW-II estimator when $M_{N,S}(\beta,\gamma,\theta)$ is non-smooth in θ can be computationally burdensome. Following Keane and Smith (2005) and Bruins et al. (2015), we consider an alternative estimator that simplifies estimation via smoothing. The smoothed estimator is obtained in the same manner as $\widehat{\theta}_N^{LM}(A)$, except that $Y_{is}(\theta)$ in the original algorithm is replaced by a transformation $Y_{is}(\theta,h_N)$ that is smooth (continuously differentiable) in θ for $h_N>0$, where

$$\lim_{h_N\to 0}Y_{is}(\theta,h_N)=Y_{is}(\theta)\text{ for all }s=1,\dots,S\text{ and }i=1,\dots,N. \tag{24}$$

The term h_N controls the smoothness of the transformation – larger (smaller) h_N leads to a more (less) smooth transformation but increases (decreases) estimation bias – and needs to be specified by the user taking into consideration the sample size N and the simulation size S.

Such transformations are widely used in simulation-based estimation of discrete choice models to avoid computational difficulties arising from the non-differentiability of the concerned estimating equations with respect to θ (see Train, 2009). To our knowledge, Keane and Smith (2005) were first to propose its use in the context of II. They named the ensuing II procedure Generalized Indirect Inference (GII). Bruins et al. (2015) present a thorough theoretical exposition of GII.

We formally define the GII (smoothed) estimator $\widetilde{\theta}_N^h(A)$ as a solution of:

$$\left\| M_{N,S}^h\left(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h(A_N)\right)\right\|_{A_N} \;\leq\; o_P\left(N^{-1/2}\right) + \inf_{\theta\in\Theta}\left\| M_{N,S}^h\left(\widehat{\beta}_N, \widehat{\gamma}_N, \theta\right)\right\|_{A_N}, \tag{25}$$

where $M_{N,S}^h(\beta,\gamma,\theta) := \frac{1}{NS}\sum_{i=1}^N \frac{D_i}{p(W_i;\gamma)}\sum_{s=1}^S m(Y_{is}(\theta,h_N), Z_i, X_i^{\text{obs}}, \beta,)$ and refer to $\widetilde{\theta}_N^h(A)$ as the IPW-GII estimator of θ^0.

The proposed smoothing approach in Keane and Smith (2005) and Bruins et al. (2015) is more sophisticated than (25) and involves choosing the appropriate smoothing parameter h_N in two steps, which is not fully reflected in the definition (25). In our Monte Carlo experiment involving estimation of structural parameters in a multinomial probit model, however, a naive one-step choice of h_N for the IPW-GII estimator provides significant improvements over the IPW-II estimator. In particular, not only does it reduce the computational cost substantially but it also improves the asymptotic normality approximation for the distribution of the II estimator.[8]

Asymptotic equivalence between $\widetilde{\theta}_N^h(A)$ and $\widehat{\theta}_N^{LM}(A)$ is ensured by letting $h_N \to 0$ at a controlled rate ($\sqrt{N}h_N = o(1)$) and under additional, but standard, technical conditions on the quantities depending on h_N. We collect these conditions as Assumption **A8** in the Appendix.

Proposition 3.3 *Under **Assumptions A1-A8** in the Appendix, for some sequence of non-negative real numbers h_N satisfying $\sqrt{N}h_N = o(1)$,*

$$\sqrt{N}\left(\widehat{\theta}_N^{LM}(A) - \widetilde{\theta}_N^h(A)\right) = o_P(1).$$

4 Illustrative Example: Multinomial Probit Model

Herein, we consider a multinomial probit model similar to Section 9 in Gourieroux et al. (1993). However, our choice of the auxiliary model is different. It is based on moment conditions (14)-(15) and leads to ordinary least squares computations, which has similarities with the auxiliary models in Keane and Smith (2005), Li (2010) and Bruins et al. (2015). In particular, Keane and Smith (2005) and Bruins et al. (2015) use this auxiliary model to estimate the parameters of a multinomial probit model. Section 4.1 specifies the auxiliary

[8] With minor modifications to the assumptions and the theoretical results presented in this paper one can also accommodate the two-step procedure for the choice of h_N, if needed, following the results in Bruins et al (2015).

model for II and establishes the identification conditions **A2** and **A3** (in the appendix) without explicit consideration of the missing variables. However, missing variables under **MAR-X** can be accommodated by simply replacing the moment vector for the auxiliary model by its inverse probability weighted version. The satisfaction of **A2** and **A3** ensure the adequacy of the auxiliary model for use in II. Section 4.2 presents a simulation study demonstrating the effectiveness in finite samples of the IPW-II and IPW-GII estimators in this model when the exogenous variable X is missing endogenously following **MAR-X** in (2).

4.1 Indirect Inference: Multinomial Probit Model

Consider a $(J+1)$-alternative multinomial probit model with the alternative 0 as the baseline:

$$\begin{aligned} Y_j &= 1\left(U_j > \max(0, U_k : k = 1, \ldots, J \text{ and } k \neq j)\right), \text{ for } j = 1, ..., J \\ U_j &= Z_j'\alpha + X'\lambda_j + e_j, \\ \text{and } (e_1, \ldots, e_k)' &= \Omega^{1/2}(\varepsilon_1, \ldots, \varepsilon_k)' \text{ with } \Omega^{1/2} \text{ lower triangular such that } \Omega^{1/2}\Omega^{1/2'} = \Omega. \end{aligned} \tag{26}$$

Let $(\varepsilon_1, \ldots, \varepsilon_k)' \sim N(0, I_k)$ be independent of $Z = (Z_1', \ldots, Z_J')'$, i.e., say the alternative dependent variables, and X, i.e., say the purely individual specific variables.[9] This corresponds to the structural model (8). Let the structural parameters be $\theta = (\alpha', h_1', \ldots, h_J', \omega')'$ where ω are the unique unrestricted elements of Ω. $\theta = \theta^0$ in our population of interest.

Our implementation of II in this multinomial probit model follows the same steps described in Section 3.1-3.3. One possible choice for $m(.)$, which we follow in the Monte-Carlo experiment in Section 4.2, is to take:

$$m(R, Z, X, \beta) = \left[\begin{array}{c} \begin{pmatrix} \zeta(R_1 - \zeta'\beta_1) \\ \vdots \\ \zeta(R_J - \zeta'\beta_J) \end{pmatrix} \\ \text{vech}\left[\begin{pmatrix} R_1 - \zeta'\beta_1 \\ \vdots \\ R_J - \zeta'\beta_J \end{pmatrix}\begin{pmatrix} R_1 - \zeta'\beta_1 \\ \vdots \\ R_J - \zeta'\beta_J \end{pmatrix}' - \begin{pmatrix} \beta_{11} & \ldots & \beta_{1J} \\ \vdots & \vdots & \vdots \\ \beta_{1J} & \ldots & \beta_{JJ} \end{pmatrix}\right] \end{array}\right] \tag{27}$$

where R (stands for response) is either Y or $Y(\theta)$, as appropriate. $R_j = 1(R = j)$ for $j = 1, \ldots, J$ and $\beta = (\beta_1', \ldots, \beta_J', \beta_{11}, \ldots, \beta_{1J}, \beta_{22}, \ldots, \beta_{2J}, \ldots, \beta_{JJ})'$. ζ is some vector valued function of Z and X; for

[9] Normality of ε rules out ties in U_j's almost surely in Z and X. Also assume that the usual restrictions for identification, such as standardizing α, λ_j's and Ω with respect to the $(1,1)$-th element of Ω, and/or any other context specific restrictions are imposed. We abstract from all such issues that are peripheral to the message of our paper.

example, $\zeta = (1, Z', X')'$. A "richer" function ζ, for example, that also includes quadratic terms in Z and X, increases the "richness" of the auxiliary model and generally leads to higher efficiency of II; see Bruins et al. (2015) for a careful demonstration.

This choice of $m(.)$ has the benefit of only requiring simple computations of equation-by-equation ordinary least squares in a seemingly unrelated regression (SUR) model with J response variables $1(Y = j)$ or $1(Y(\theta) = j)$ for $j = 1, \ldots, J$; *same set* of regressors ζ for all regressions; and regression errors with unknown variance-covariance matrix. In particular, this choice of $m(.)$ leads to the first order conditions (that are efficient given ζ) for the SUR model regression coefficients, augmented by the estimating equations for the unique elements in the variance-covariance matrix of the SUR regression errors. Hence, for a given ζ, the computation and efficiency consideration involved with this choice of $m(.)$ are the same as that due to the quasi maximum likelihood estimation of the parameters of the auxiliary model in Bruins et al. (2015).

Lemma 4.1 below shows that when no variables are missing, standard least squares identification conditions are sufficient for the key identification conditions **A2** and **A3** to hold in II based on the auxiliary model induced by the choice of $m(.)$ in (27). The proof is trivial and hence omitted.

Lemma 4.1 *Define $Y_j := 1(Y = j)$ and $Y_j(\theta) := 1(Y(\theta) = j)$. Then Assumption **A2** in the Appendix holds if $E[\zeta\zeta']$ is non-singular, while Assumption **A3** in the Appendix holds under the additional orthogonality restriction $E[\zeta(Y_j(\theta) - Y_j(\theta^0)] = 0$ or, equivalently, $E[\zeta(Y_j(\theta) - Y_j)] = 0$ for $j = 1, \ldots, J$ if and only if $\theta = \theta^0$.*

Remarks:

(1) The lemma also applies to other discrete response models as long as the non-singularity and orthogonality conditions hold. This does not contradict the well known results that, typically such orthogonality (or even mean independence) conditions are not sufficient for non-parametric identification of the structural parameters in discrete response models. While apparently no other distributional assumption has been made in its statement, the lemma is highly parametric and could not possibly be used without knowing the distribution of $Y_j(\theta)$ conditional on Z, X.

(2) Section 4.2 takes $\zeta = (1, Z', X')'$ and, therefore, according to Lemma 4.1 it implicitly requires for identification of θ^0 the following high level orthogonality conditions:

(a) $P(Y_j(\theta) = 1) = P(Y_j(\theta^0) = 1)$ for all $j = 1, \ldots, J$ if and only if $\theta = \theta^0$.

(b) $E[Z(P(Y_j(\theta) = 1|Z, X) - P(Y_j(\theta^0) = 1|Z, X))] = 0$ for all $j = 1, \ldots, J$ if and only if $\theta = \theta^0$.

(c) $E[X(P(Y_j(\theta)=1|Z,X)-P(Y_j(\theta^0)=1|Z,X))]=0$ for all $j=1,\dots,J$ if and only if $\theta=\theta^0$.

(3) A "richer" ζ, for example, that also includes quadratic terms in Z and X, would impose additional such orthogonality conditions and thereby would lead to higher precision of the II estimates.

(4) The result directly applies to our framework of endogenously missing exogenous variables X by replacing $m(R,Z,X,\beta)$ in (27) by $\frac{D}{p(W;\gamma^0)}m(R,Z,X,\beta)$ and appealing to **MAR-X** if $R=Y$, or by following similar arguments to those in equation (13) if $R=Y(\theta)$.

Finally, Lemma 4.1 can also be used to identify the pseudo-true θ (call it θ^*) estimated by II when the exogenous variables X are missing endogenously following **MAR-X** in (2) and the missingness is simply ignored. Hereafter, we will refer to an II procedure that simply ignores the missingness as standard II.

Consider the following toy example where, for simplicity of demonstration, we take $J=1$, ignore Z, and make specific and convenient distributional assumptions that are covered by our maintained assumptions.

Toy Example: Let the structural model and the missingness mechanism be characterized by:

$$Y=1(X\lambda^0+\varepsilon\geq 0)\quad\text{and}\quad D=1(Y\gamma^0+v\geq 0)$$

where the scalar random variable X, the structural error ε and the missingness error v are assumed to be independent. Let $\theta^0=\lambda^0$. Following (27), define $m(R,X,\beta)=X(R-X\beta)$ for $R=Y$ or $R=Y(\theta)$. We ignore the overidentifying (second moment) restrictions from (27) for simplicity.

Therefore, standard II defines $\tilde{\beta}^0$ and $\tilde{\beta}^0(\theta)$ as follows:

$$\tilde{\beta}^0 \text{ solves } E[DX(Y-X\beta)]=0, \text{ and } \tilde{\beta}^0(\theta) \text{ solves } E[DX(Y(\theta)-X\beta)]=0.$$

These are essentially the population version of the first two steps of standard II. The final step obtains θ^* by the matching exercise $\tilde{\beta}^0=\tilde{\beta}^0(\theta^*)$, which by Lemma 4.1 holds if and only if $E[DXY(\theta^*)]=E[DXY]$. Letting F_T denote the distribution function of any variable T, we know:

$$\begin{aligned} E[DXY(\theta^*)] &= E\left[((1-F_v(-\gamma^0))(1-F_\epsilon(-X\theta^0))+(1-F_v(0))F_\epsilon(-X\theta^0))(1-F_\epsilon(-X\theta^*))X\right], \\ E[DXY] &= E\left[(1-F_v(-\gamma^0))(1-F_\epsilon(-X\theta^0))X\right]. \end{aligned}$$

The above equalities follow from using **MAR-X** in (2), the conditional (on X) independence between Y and $Y(\theta)$, and the fact that $F_\varepsilon=F_{\tilde{\varepsilon}}$. For simplicity, assume the specific and convenient distributions: $\varepsilon\sim N(0,1), v\sim N(0,1)$ and $X\sim$ Bernoulli(q). Denote the distribution function of $N(0,1)$ by $\Phi(.)$ and its inverse by $\Phi^{-1}(.)$. Equating $E[DXY(\theta^*)] = E[DXY]$, standard II yields

$$\text{(pseudo-true value)} \quad \theta^* = \Phi^{-1}\left(\frac{\Phi(\gamma^0)\Phi(\theta^0)}{\Phi(\gamma^0)\Phi(\theta^0) + \Phi(0)(1-\Phi(\theta^0))}\right) \neq \theta^0 \quad \text{(true value)},$$

unless $\gamma^0 = 0$, i.e., unless the missingness is exogenous. Hence, it is the endogeneity of the missingness that causes the problem of identification with standard II. Our proposed IPW-II estimator solves this problem.

4.2 Simulation Study: Three Alternative ($J = 2$) Probit Model

The simulation design considered here is similar to Model 4 in Keane and Smith (2005) and Bruins et al. (2015). In particular, we consider the multinomial probit model in (26) with $J = 2$ for simplicity. For each $i = 1, \ldots, N$, we generate the exogenous regressors as: $Z_{ji} \overset{\text{i.i.d.}}{\sim} \chi_1^2 - 1$ for $j = 1, 2$ and $X_i \overset{\text{i.i.d.}}{\sim} N(1, 2)$ independent of each other. Normalizing all the parameters in the model by the (1,1)-th element of Ω, i.e., equivalently, by fixing $\omega_{11} = 1$ (not to be estimated), we take $\theta^0 = (\alpha^0 = 1, \lambda_1^0 = 1, \lambda_2^0 = 2, \omega_{12}^0 = .5, \omega_{22}^0 = 1)'$. We generate the structural errors $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, I_2)$ and $e_i = \Omega^{0^{1/2}} \varepsilon_i$ independent of the regressors Z_{1i}, Z_{2i}, X_i and, finally, we generate the outcome Y_i following (26) for each $i = 1, \ldots, N$.

We consider the following missingness mechanism that determines the observability of X. Generate

$$D_i = 1(\gamma_1^0 \times 1(Y_i = 1) + \gamma_2^0 \times 1(Y_i = 2) + \gamma_3^0 \times Z_{2i} \geq v_i)$$

for each $i = 1, \ldots, N$ with $v_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$ independent of the structural errors e_i and the exogenous variables $Z_i = (Z_{1i}, Z_{2i})'$ and X_i. Hence, **MAR-X** in (2) holds. Take $\gamma_1^0 = -.5, \gamma_2^0 = .5$ and $\gamma_3^0 = 1$. This leads to roughly 50% of sample units with missing X.

We consider the auxiliary model and parameters as defined by the moments $m(.)$ in (27) with $\zeta = (1, Z_1, Z_2, X)'$. Four different LM-type II estimators are considered: the standard II estimator, an infeasible II estimator, the IPW-II and IPW-GII estimators introduced in Section 3. The standard II estimator works with the complete case data $\{D_iY_i, D_iZ_i, D_iX_i\}_{i=1}^N$, i.e., sample units without any missing variables. Standard II ignores the endogenous missingness of X and thus can be biased, gauging the magnitude and consequences of this bias is the first purpose of the simulation study. The infeasible II estimator works with the infeasible full data set $\{Y_i, Z_i, X_i\}_{i=1}^N$, which is only available because we have generated the data, and is not available in practice. The infeasible II is the II estimator that one would use if there were no missingness in the data. Its finite-sample behavior provides an infeasible benchmark for the performance of II in this context. The IPW-II and IPW-GII estimators work with the observed data $\{D_i, Y_i, Z_i, D_iX_i\}_{i=1}^N$

but account for the endogeneity in the missingness of X. These estimators are designed to correct the bias of the standard II estimator, and demonstrating this is the second purpose of the simulation study. The third purpose of the study is to add a word of caution by demonstrating that the asymptotic normal approximation for the IPW-II estimator in **Proposition** 3.2 may be inadequate even in reasonably large samples. Lastly, the simulation study demonstrates that the IPW-GII estimator, thanks to the smoothing proposed by Keane and Smith (2005) and Bruins et al. (2015), does not suffer from this issue and is much faster to implement than the other IPW-II estimator.

We compute the mean bias (MBIAS), mean absolute bias (ABIAS), standard deviation (STD), interquartile range (IQR) and the coverage of a 95% Wald-confidence interval (COV95) for all the estimators of $(\alpha^0, \lambda_1^0, \lambda_2^0, \omega_{12}^0, \omega_{22}^0)$ for sample sizes $N = 200, 500, 1000$ and 5000. We take $S = 10$ for all estimators. The standard II, infeasible II and IPW-II estimators are computed by the *patternsearch* routine in Matlab. On the other hand, the smoothness of the optimization problem for the IPW-GII estimator allows the use of the gradient-based Matlab routine *fminunc*. Following Bruins et al. (2015), the initial value is set at the true parameter value for all four estimation procedures. All four estimators use the (estimator specific) optimal weighting matrix (see **Proposition** 3.2), and in effect are continuously updated GMM estimators. All results are based on $10,000$ Monte-Carlo trials.

To abstract from biases due to small sample sizes and instead focus on the bias that arises because the standard II estimator deliberately ignores the endogenous missingness, we only report the results for the standard II estimator based on $N = 5000$ in Table 1.[10]

θ	MBIAS	ABIAS	STD	IQR	COV95
α	0.0331	0.0472	0.0647	0.0727	94.05
λ_1	0.0259	0.0487	0.0648	0.0698	91.73
λ_2	0.4905	0.4905	0.1013	0.5008	1.03
ω_{12}	-0.1297	0.2053	0.2216	0.2568	91.82
ω_{22}	1.2701	1.2707	0.4538	1.3488	17.54

Table 1: Monte-Carlo results for the Multinomial probit ($J = 2$) model. MBIAS, ABIAS, STD, IQR and COV95 are the mean bias, absolute bias, (Monte-Carlo) standard deviation, interquartile range and coverage of a 95% Wald-type confidence interval for the standard II estimator for the different elements of θ when $N = 5000$.

This estimator is badly biased (see MBIAS) and as a consequence, the coverage of the 95% confidence intervals for the unknown parameters can be extremely low: indeed as low as 1%.

Table 2 reports the results for the other three estimators. As expected from the results in Section 3, the IPW-II corrects the bias of the standard II estimator. Its bias (MBIAS) decreases considerably as the sample

[10] Results for other sample sizes are available from the authors.

		N = 200					N = 500				
Estimator	θ	MBIAS	ABIAS	STD	IQR	COV95	MBIAS	ABIAS	STD	IQR	COV95
Infeasible II	α	0.0576	0.1063	0.1540	0.1645	93.36	0.0243	0.0617	0.0938	0.0969	91.84
	λ_1	0.0330	0.1000	0.1463	0.1500	93.45	0.0239	0.0672	0.0984	0.1013	92.82
	λ_2	0.0924	0.1119	0.1653	0.1894	91.94	0.0959	0.3368	0.4767	0.5154	94.86
	ω_{12}	-0.0513	0.2740	0.3399	0.3438	95.54	0.0519	0.0705	0.1022	0.1146	90.69
	ω_{22}	0.0852	0.4624	0.6028	0.6088	95.29	-0.0047	0.1596	0.2003	0.2003	94.64
IPW-II	α	0.1237	0.2100	0.3758	0.3957	94.81	0.0648	0.1121	0.1794	0.1908	93.16
	λ_1	0.0875	0.1723	0.3133	0.3253	95.55	0.0437	0.1126	0.1803	0.1856	93.49
	λ_2	0.2425	0.2683	0.5282	0.5812	95.53	0.1021	0.1201	0.2084	0.2321	93.71
	ω_{12}	-0.0634	0.4215	0.5195	0.5233	94.07	-0.0152	0.3018	0.3690	0.3694	96.23
	ω_{22}	0.3961	0.8948	3.3841	3.4072	99.17	0.1016	0.5535	0.9864	0.9916	98.20
IPW-GII	α	0.3230	0.3253	0.7685	0.7146	92.30	0.0802	0.1829	0.3974	0.3674	95.05
	λ_1	0.3190	0.2753	0.6837	0.6283	91.62	0.0746	0.1498	0.3196	0.3038	94.45
	λ_2	0.6686	0.5356	1.3933	1.1999	91.39	0.1585	0.1861	0.6308	0.3838	94.22
	ω_{12}	0.1954	0.3952	0.7685	0.8270	92.98	0.1071	0.2464	0.4486	0.5040	93.42
	ω_{22}	1.5954	0.7944	3.5950	2.2757	93.56	0.4260	0.4769	1.3016	1.0343	94.45

		N = 1000					N = 5000				
Estimator	θ	MBIAS	ABIAS	STD	IQR	COV95	MBIAS	ABIAS	STD	IQR	COV95
Infeasible II	α	0.0156	0.0463	0.0664	0.0682	94.63	0.0070	0.0222	0.0310	0.0318	91.16
	λ_1	0.0141	0.0493	0.0727	0.0741	92.95	0.0072	0.0268	0.0372	0.0379	93.12
	λ_2	0.0302	0.0427	0.0645	0.0712	91.51	0.0146	0.0225	0.0320	0.0351	91.29
	ω_{12}	-0.0031	0.1162	0.1465	0.1465	95.21	-0.0021	0.0563	0.0709	0.0709	94.83
	ω_{22}	0.0343	0.1928	0.2452	0.2476	94.56	0.0137	0.0901	0.1135	0.1143	94.56
IPW-II	α	0.0387	0.0866	0.1279	0.1336	94.01	0.0106	0.0354	0.0513	0.0523	93.37
	λ_1	0.0269	0.0809	0.1228	0.1257	91.88	0.0148	0.0421	0.0608	0.0626	92.05
	λ_2	0.0632	0.0744	0.1150	0.1312	90.23	0.0248	0.0315	0.0458	0.0522	91.96
	ω_{12}	-0.0086	0.2403	0.2960	0.2962	96.81	-0.0002	0.1155	0.1451	0.1451	94.97
	ω_{22}	0.0684	0.4245	0.5475	0.5517	95.72	0.0250	0.1987	0.2515	0.2527	94.72
IPW-GII	α	0.0071	0.1037	0.1916	0.2073	94.28	0.0128	0.0800	0.1212	0.1617	94.62
	λ_1	0.0040	0.0829	0.1334	0.1653	94.86	0.0099	0.0690	0.1019	0.1386	94.55
	λ_2	0.0161	0.063	0.2137	0.1279	93.53	0.0236	0.1307	0.1967	0.2638	94.35
	ω_{12}	0.0618	0.1302	0.2788	0.2682	93.06	0.0063	0.0873	0.1332	0.1747	94.94
	ω_{22}	0.1367	0.2015	0.5596	0.4254	94.01	0.0386	0.1954	0.3036	0.3973	94.86

Table 2: Monte-Carlo results for the Multinomial probit ($J = 2$) model. MBIAS, ABIAS, STD, IQR and COV95 are the mean bias, absolute bias, standard deviation, interquartile range and coverage of a 95% Wald-type confidence interval for the concerned estimator for the different elements of the parameter vector θ. STD is not based on the asymptotic variance formula in Proposition 3.2 but on the Monte-Carlo. Results are obtained by 10000 Monte-Carlo trials.

size increases. ABIAS, STD and IQR also display similar pattern with the increase in sample size. The coverage (COV95) is good. Overall, keeping in mind that X is missing for roughly 50% sample units, the finite-sample behavior of the IPW-II estimator does not deviate much from that of the infeasible benchmark provided by the infeasible II estimator, especially when the sample size is not too low.

Similar phenomenon of bias correction is observed for the IPW-GII estimator. However, its bias (MBIAS) is larger than that of the IPW-II estimator. Its ABIAS, STD and IQR are also generally larger than that of the IPW-II estimator. These features are possibly due to the naive one-step choice for the smoothing parameter h_N in the implementation of the IPW-GII estimator.[11]

Nevertheless, the IPW-GII estimator indeed serves the dual purpose stated in Section 3. The IPW-GII estimator is much faster than the IPW-II estimator, and more importantly, while the studentized IPW-II estimator is far from from being normally distributed, even for sample size $N = 5000$, no such problem arises for the IPW-GII estimator;[12] Figure 1 gives precise details.

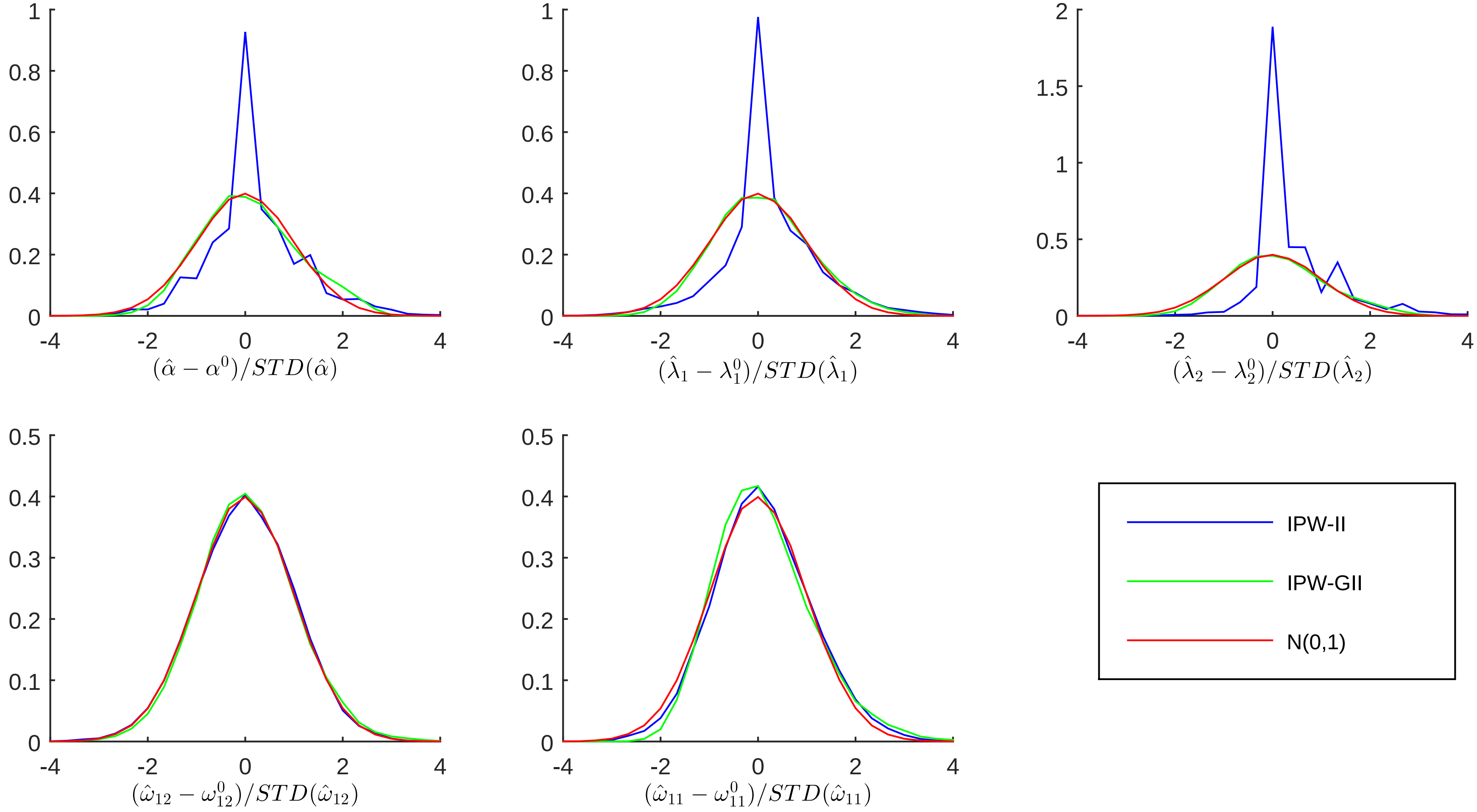
5 Conclusion

In this paper we have demonstrated the problems with identification and consistent estimation of the structural parameters by II when the exogenous variables can be endogenously missing following the **MAR-X** assumption, which can arise in empirical work for reasons such as survey non-response, survey revisions, cost-effective survey design, etc. Our proposed solution can be implemented as either the IPW-II or the IPW-GII estimator, with the smoothed IPW-GII approach being particularly useful in non-smooth problems. This novel estimation method corrects for the sample selection bias in the estimation of the auxiliary parameters with the observed data and the simulated data using the method of inverse probability weighting. The desirable performance of the proposed II approach was demonstrated theoretically and via simulations. The extremely poor performance of standard II estimators that simply discard sample units with missingness was also demonstrated via simulations. We conclude by noting that the selection due to the missing data is handled by our proposed method in one step that only involves the estimation of the missingness (conditional) probabilities using a binary choice model, such as logit or probit, and hence the proposed method retains the computational attractiveness of II procedures.

[11] The smoothing parameter h_N is .078, .0571, .0458, .0284 respectively for $N = 200, 500, 1000, 5000$. This is in rough accordance to the requirements of Proposition 3.3 but with a slight tilt toward zero for the smaller sample sizes $N = 200, 500$ to reduce the bias due to smoothing.

[12] The same issue is also present in the infeasible II estimator. However, for both the infeasible II and IPW-II, the quality of the normal approximation is better if we use a richer auxiliary model by augmenting $\zeta = (1, Z', X')'$ with quadratic terms in Z and X. This removes some wiggliness in the corresponding kernel density plots. These figures are not included for brevity but can be found in the previous version of the paper and are available from the authors.

Figure 1: Kernel density plots for the studentized IPW-II and IPW-GII estimators of the different elements of the parameter vector θ when sample size $N = 5000$. Results are reported based on 10000 Monte Carlo trials. It is clear that IPW-II can be far away from the standard normal approximation (red line) while IPW-GII does not suffer from this problem.



References

Altonji, J. J., Smith, A. A., and Vidangos, I. (2013). Modelling Earning Dynamics. *Econometrica*, 81: 1395–1454.

Breusch, T., Qian, H., Schmidt, P., and Wyhowski, D. (1999). Redundancy of Moment Conditions. *Journal of Econometrics*, 91:89 – 111.

Bruins, M., Duffy, J., Keane, M., and Smith, A. (2015). Generalized Indirect Inference for Discrete Choice Models. Mimeo.

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

Chaudhuri, S., Frazier, D. T., and Renaut, E. (2016). Efficiency of Indirect Inference with Missing Data. mimeo.

Chaudhuri, S. and Guilkey, D. K. (2014). GMM with Multiple Missing Variables. Forthcoming: Jorunal of Applied Econometrics.

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

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

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

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

Gallant, R. and Tauchen, G. (1996). Which Moments to Match? *Econometric Theory*, 12: 657–681.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect Inference. *Journal of Applied Econometrics*, 8: S85–S118.

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

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

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

Horvitz, D. and Thompson, D. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of American Statistical Association*, 47: 663–685.

Jiang, W. and Turnbull, B. (2004). The Indirect Method: Inference Based on Intermediate Statistics - A Synthesis and Examples. *Statistical Science*, 19: 239–263.

Keane, M. and Smith, A. A. (2005). Generalized Indirect Inference for Discrete Choice Models. Technical report, Yale University.

Kleibergen, F. (2005). Testing Parameters In GMM Without Assuming That They Are Identified. *Econometrica*, 73: 1103–1123.

Li, T. (2010). Indirect inference in structural econometric models. *Journal of Econometrics*, 157: 120–128.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data.* Wiley, Hoboken, NJ.

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

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):122–129.

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

Rubin, D. (1976). Inference and Missing Data. *Biometrika*, 63: 581–592.

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

Smith, A. A. (1990). *Three Essays on the Solution and Estimation of Dynamic Macroeconomic Models,.* PhD thesis, Duke University.

Smith, A. A. (1993). Estimating Nonlinear Time-series Models using Simulated Vector Autoregressions. *Journal of Applied Econometrics*, 8: S63–S84.

Train, K. E. (2009). *Discrete Choice Methods with Simulation.* Cambridge university press.

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

A Appendix: Technical assumptions and proofs

A.1 Technical Assumptions

The following notations are used. For a $d \times d$ matrix A and a $c \times d$ matrix B, define $\|B\|_A := \sqrt{\text{Trace}(BAB')}$ and $\|B\| := \|B\|_{A=I_d}$. Define $\mathcal{N}_\delta(\theta^0) \subset \Theta$, $\mathcal{N}_\delta(\beta^0) \subset \mathcal{B}$ and $\mathcal{N}_\delta(\gamma^0) \subset \Gamma$ as some generic open neighborhoods of radius δ for θ^0, β^0 and γ^0 respectively. Finally, define

$$M(\beta,\gamma) := E\left[\frac{D}{p(W;\gamma)}m(Y,Z,X,\beta)\right] \text{ and } M(\beta,\gamma,\theta) := E\left[\frac{D}{p(W;\gamma)}m(Y(\theta),Z,X,\beta)\right].$$

By the definition of β^0 in (1), the definition of the binding function in (15), and the above Lemmas,

$$M(\beta^0,\gamma^0) = M(\beta^0,\gamma^0,\theta^0) = 0. \tag{28}$$

Assumption A1:

(a) Structural Model in (8): ε has a known distribution $F_\varepsilon = F_\varepsilon^0$ and is independent of Z and X whose unknown distribution is $F_{(Z,X)} = F_{(Z,X)}^0$.

(b) Strict overlap: For MAR in (2), $p_0(W) := P(D = 1|W) \in [p, 1)$ for a constant $p > 0$.

(c) Observed sample: $\{W_i, D_i, D_iX_i\}_{i=1}^N$ are i.i.d. copies of W, D, and DX.

Assumption A2: β^0 is the unique solution to equation (1).
Assumption A3: For all $\theta \in \Theta$, the binding function $\beta(\theta)$ defined in equation (15) satisfies $\beta^0 = \beta(\theta)$ if and only if $\theta = \theta^0$.
Assumption A4 : There exists a unique $\gamma^0 \in \Gamma$ and a function $p(w, \gamma) : \text{Support}(W) \times \Gamma \mapsto (0, 1)$ such that $p_0(w) = p(w; \gamma^0)$ for all $w \in \text{Support}(W)$. $\Gamma \subset \mathbb{R}^{d_\gamma}$ is compact and d_γ is finite.
Assumption A5:

(a) $\Theta \subset \mathbb{R}^{d_\theta}$ and $\mathcal{B} \subset \mathbb{R}^{d_\beta}$ are compact with $\theta^0 \in \text{interior}(\Theta)$ and $\beta^0 \in \text{interior}(\mathcal{B})$.

(b) For $l = (l_1, l_2, l_3)$ where $l_1 \in \text{Support}(Y \text{ or } Y(\theta))$ (as appropriate) and $(l_2, l_3) \in \text{Support}(Z, X)$: $m(l, \beta)$ is continuous in β for all l, and $\|m(l, \beta)\|^2 \leq g(l)$ for all l and $E[g(l)] < \infty$.

(c) For $\delta > 0$: $\displaystyle\sup_{\theta\in\Theta,\beta\in\mathcal{N}_\delta(\beta^0),\gamma\in\mathcal{N}_\delta(\gamma^0)} \frac{\|M_{N,S}(\beta,\gamma,\theta) - M(\beta,\gamma,\theta)\|}{1 + \|M_{N,S}(\beta,\gamma,\theta)\| + \|M(\beta,\gamma,\theta)\|} = o_P(1).$

Assumption A6:

(a) $p(w; \gamma)$ is continuous in $\gamma \in \Gamma$ for all $w \in \text{Support}(W)$.

(b) For some $\delta > 0$: $p(w; \gamma)$ is twice continuously differentiable in $\gamma \in \mathcal{N}_\delta(\gamma^0)$ for all $w \in \text{Support}(W)$, and the derivatives $p_\gamma(w; \gamma) := \frac{\partial}{\partial \gamma'} p(w; \gamma)$ and $p_{\gamma\gamma}(w; \gamma) := \frac{\partial}{\partial \gamma'} p'_\gamma(w; \gamma)$ satisfy: $\sup_{\gamma\in\mathcal{N}_\delta(\gamma^0)} \|p_\gamma(w; \gamma)\|^2 + \sup_{\gamma\in\mathcal{N}(\gamma^0)} \|p_{\gamma\gamma}(w; \gamma)\| < b(w)$ for all $w \in \text{Support}(W)$ where $b(w) \geq 0$ and $E[b(w)] < \infty$.

(c) The score $l_\gamma(D, W; \gamma) := (D - p(W; \gamma))p'_\gamma(W; \gamma)/[p(W; \gamma)(1 - p(W; \gamma))]$ is such that $B_0 := E\left[l_\gamma(D, W; \gamma^0)l'_\gamma(D, W; \gamma^0)\right]$ is nonsingular.

Assumption A7:

(a) For each $l = (l_1, l_2, l_3)$ where $l_1 \in \text{Support}(Y \text{ or } Y(\theta))$ (as appropriate) and $(l_2, l_3) \in \text{Support}(Z, X)$, $m(l, \beta)$ is continuously differentiable in $\beta \in \mathcal{N}_\delta(\beta^0)$ for some $\delta > 0$. Allow for changing the order of differentiation and integration, i.e., let $E\left[\sup_{\beta\in\mathcal{N}_\delta(\beta^0)} \|\partial m(l, \beta)/\partial \beta'\|\right] < \infty$.

(b) $G_0 := E\left[\frac{\partial}{\partial \beta'} m(Y, Z, X, \beta^0)\right] \equiv E\left[\frac{\partial}{\partial \beta'} m(Y(\theta^0), Z, X, \beta^0)\right]$ is nonsingular.

(c) $\sqrt{N}\bar{\xi}_{N,S} \xrightarrow{d} N(0, W_0(S))$ where $W_0(S) = \left(1 + \frac{1}{S}\right)(I_0 - K_0)$, $\bar{\xi}_{N,S} := \sum_{i=1}^N \xi_{i,S}/N$,

$$\xi_{i,S} := \frac{D}{p(W; \gamma^0)} m(Y, Z, X, \beta^0) - \frac{1}{S}\sum_{s=1}^S \frac{D}{p(W; \gamma^0)} m(Y(\theta^0), Z, X, \beta^0).$$

(d) For $\theta = \theta^0$: $(\partial/\partial\theta')M(\beta^0, \gamma^0, \theta)$ has rank d_θ and is continuously differentiable in θ.

(e) For every positive sequences $\{\delta_N\}$ and $\delta_N = o(1)$

$$\sup_{\theta\in\mathcal{N}_{\delta_N}(\theta^0),\beta\in\mathcal{N}_{\delta_N}(\beta^0),\gamma\in\mathcal{N}_{\delta_N}(\gamma^0)} \frac{\sqrt{N}\|M_{N,S}(\beta,\gamma,\theta) - M(\beta,\gamma,\theta) - M_{N,S}(\beta^0,\gamma^0,\theta^0)\|}{1 + \sqrt{N}\|M_{N,S}(\beta,\gamma,\theta)\| + \sqrt{N}\|M(\beta,\gamma,\theta)\|} = o_P(1).$$

To establish the asymptotic properties of the GII estimator, additionally define for each h:

$$M^h(\beta, \gamma, \theta) := E\left[\frac{D}{p(W; \gamma)} m(Y(\theta, h), Z, X, \beta)\right].$$

As before like (28) and further using (24),

$$M(\beta^0, \gamma^0) = M(\beta^0, \gamma^0, \theta^0) = M^{h=0}(\beta^0, \gamma^0, \theta^0) = 0. \tag{29}$$

The following assumptions on $M_N^h(\beta,\gamma,\theta)$ $M^h(\beta,\gamma,\theta)$ and $M(\beta,\gamma,\theta)$ are additionally maintained for the asymptotic equivalence of the GII and II estimators.

Assumption A8: For some $\delta > 0$ and a finite $b > 0$, let the following hold for $M_{N,S}^h(.)$ and its limit counterpart $M^h(.)$:[13]

(a) $\sup\limits_{\theta\in\Theta,\beta\in\mathcal{N}_\delta(\beta^0),\gamma\in\mathcal{N}_\delta(\gamma^0)} \|M^h(\beta,\gamma,\theta) - M(\beta,\gamma,\theta)\| \leq b\times h$ for $h \in [0,\delta)$.

(b) $\sup\limits_{h\in[0,\delta)} \sup\limits_{\theta\in\Theta,\beta\in\mathcal{N}_\delta(\beta^0),\gamma\in\mathcal{N}_\delta(\gamma^0)} \dfrac{\|M_N^h(\beta,\gamma,\theta) - M^h(\beta,\gamma,\theta)\|}{1+\|M_N^h(\beta,\gamma,\theta)\|+\|M^h(\beta,\gamma,\theta)\|} = o_P(1).$

(c) (i) $\sup\limits_{h\in(0,\delta)} \sup\limits_{\theta\in\mathcal{N}_\delta(\theta^0),\beta\in\mathcal{N}_\delta(\beta^0),\gamma\in\mathcal{N}_\delta(\gamma^0)} \left\|\dfrac{\partial}{\partial(\beta',\gamma')}\left(M_N^h(\beta,\gamma,\theta) - M^h(\beta,\gamma,\theta)\right)\right\| = o_P(1).$

(c) (ii) $\sup\limits_{h\in(0,\delta)} \sup\limits_{\theta\in\mathcal{N}_\delta(\theta^0),\beta\in\mathcal{N}_\delta(\beta^0),\gamma\in\mathcal{N}_\delta(\gamma^0)} \left\|\dfrac{\partial}{\partial\theta'}\left(M_N^h(\beta,\gamma,\theta) - M^h(\beta,\gamma,\theta)\right)\right\| = O_P(N^{-1/2}).$

(d) $\frac{\partial}{\partial(\beta',\gamma',\theta')}M^h(\beta,\gamma,\theta)$ is continuous in β,γ,θ, h for $(\beta,\gamma,\theta)\in\mathcal{N}_\delta(\beta^0,\gamma^0,\theta^0)$ and $h\in[0,\delta)$.

(e) $\sup\limits_{h\in(0,\delta)} \dfrac{\sqrt{N}\|M_N^h(\beta^0,\gamma^0,\theta^0) - M^h(\beta^0,\gamma^0,\theta^0) - M_N(\beta^0,\gamma^0,\theta^0)\|}{1+\sqrt{N}\|M_N^h(\beta^0,\gamma^0,\theta^0)\|+\sqrt{N}\|M^h(\beta^0,\gamma^0,\theta^0)\|} = o_P(1).$

Remark 1: It is well known that by Assumptions **A4** and **A6**, the maximum likelihood estimator $\widehat{\gamma}_N$ that gives $\widehat{p}_0(w) = p(w,\widehat{\gamma}_N)$ for Step 0 of modified II satisfies:

$$\sqrt{N}(\widehat{\gamma}_N - \gamma^0) = B_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^N l_\gamma(D_i,W_i;\gamma^0) + o_P(1). \tag{30}$$

Also, (2), Assumptions **A1**(b)-(c), **A5**(a)-(b) and **A7**(a)-(b) and (30) give for $\widehat{\beta}_N$ from Step 1:

$$\begin{aligned}\sqrt{N}(\widehat{\beta}_N - \beta^0) &= -G_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^N\left\{m_i^*(\gamma^0,\beta^0) - \Omega_{12}\Omega_{22}^{-1}l_{i,\gamma}(D_i,W_i;\gamma^0)\right\} + o_P(1), \qquad (31)\\ m_i^*(\gamma^0,\beta^0) &= \frac{D_i}{p_0(W_i;\gamma^0)}m(Y_i,Z_i,X_i,\beta^0),\ \ \Omega_{12} = E\left[m_i^*(\gamma^0,\beta^0)l_{i,\gamma}(D_i,W_i;\gamma^0)'\right]\end{aligned}$$

See. e.g., Chaudhuri and Min (2012) for (30) and (31). Similar steps and (13) give for $\widehat{\beta}_N(\theta^0)$, defined as

$$\widehat{\beta}_N(\theta^0) \ := \ \arg_{\beta\in\mathcal{B}}\left\{M_{N,S}\left(\beta,\widehat{\gamma}_N,\theta^0\right) = 0\right\}, \tag{32}$$

the asymptotically linear representation

$$\begin{aligned}\sqrt{N}(\widehat{\beta}_N(\theta^0) - \beta^0) &= -G_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{S}\sum_{s=1}^S\left\{m_{is}^*(\gamma^0,\beta^0;\theta^0) - \Omega_{12}(\theta^0)\Omega_{22}^{-1}l_{i,\gamma}(D_i,W_i;\gamma^0)\right\} + o_P(1), \quad (33)\\ m_{is}^*(\gamma^0,\beta^0;\theta^0) &= \frac{D_i}{p_0(W_i;\gamma^0)}m(Y_{is}(\theta^0),Z_i,X_i,\beta^0),\ \ \Omega_{12}(\theta^0) = E\left[m_{is}^*(\gamma^0,\beta^0;\theta^0))l_{i,\gamma}(D_i,W_i;\gamma^0)'\right]\end{aligned}$$

Therefore, under Assumption **A7**(c) and for a fixed S, using (30), (31) and (33) jointly give:

$$\begin{aligned}\sqrt{N}(\widehat{\beta}_N - \widehat{\beta}_N(\theta^0)) &= -G_0^{-1}\sqrt{N}\left[\bar{\xi}_{N,S} - C_0(\widehat{\gamma}_N - \gamma^0)\right] \ \xrightarrow{d}\ N\left(0, G_0^{-1}H_0(S)G_0^{-1'}\right) \qquad (34)\\ H_0(S) &= (1+\frac{1}{S})(I_0 - K_0) - C_0B_0^{-1}C_0',\ \ C_0 = \Omega_{12} - \Omega_{12}(\theta^0)\end{aligned}$$

[13] See Remark 5 for an explanation of these assumptions.

Remark 2: It is important to point out that the equivalence of the G_0 terms in the expansions (31) and (33), follows since, under the maintained assumptions, for α being any one of β, γ, θ,

$$\frac{\partial M(\gamma^0, \beta^0, \theta^0)}{\partial \alpha'} \quad = \quad \left\{ \frac{\partial}{\partial \alpha'} \frac{1}{S} \sum_{s=1}^{S} E\left[\frac{D}{p(W;\gamma)} m(Y_s(\theta), Z, X, \beta) \right] \right\} \Big|_{\beta^0, \theta^0, \gamma^0} = \left\{ \frac{\partial}{\partial \alpha'} E\left[m(Y_s(\theta), Z, X, \beta) \right] \right\} \Big|_{\beta^0, \theta^0, \gamma^0},$$

which follows by **MAR-x** and the knowledge that the S simulated samples are iid copies. More generally then, the partial derivatives $\partial M / \partial \theta'$, $\partial M / \partial \beta'$, $\partial M / \partial \gamma'$ (evaluated at the truth) do not depend on S. Likewise, it is important to note the role that S does play in the asymptotic variance of $\bar{\xi}_{N,S}$.

Remark 3: Assumption **A5**(c) is an uniform convergence condition for which the strict overlap assumption in **A1**(b) plays a crucial role. The same holds for the stochastic equicontinuity condition Assumption **A7**(e) that is similar to condition (iii) in Theorem 3.3 of Pakes and Pollard (1989), but additionally it allows for nuisance parameters close to their true values. Such high-level assumptions need to be verified on a case-by-case basis [see, e.g., Cattaneo (2010) or Chaudhuri and Guilkey (2014)].

Remark 4: Since it is also well known how to account for random weighting matrix [see Lemmas 3.4 and 3.5 of Pakes and Pollard (1989)], we abstract from it in all the proofs below and instead directly assume in the concerned propositions that the weighting matrix A_N is possibly based on some preliminary consistent estimators of the concerned parameters such that $A_N \xrightarrow{P} A$ where A is a positive definite matrix. Hence in what follows let $\widehat{\theta}_N := \widehat{\theta}_N^{LM}(A)$.

Remark 5: Assumption **A8** is a high-level condition restricting the choice of kernels (e.g. logistic or normal) used within the smoothing step of the generalized II procedure. Essentially it imposes sufficient smoothness condition on $M_N^h(.)$ and $M^h(.)$ to facilitate simple proofs of the desired asymptotic properties of the GII estimator. The denominators in **A8**(b) and (d) add slightly more generality (similar to those in **A5**(d) and **A7**(e)). The asymmetric treatment with respect to (β, γ) and θ in **A8**(c) (i) and (ii) respectively is due to the fact that we do not formally establish $\sqrt{N}$-consistency of $\widetilde{\theta}_N^h$ prior to demonstrating its asymptotic normality. The stronger condition in (ii) bears resemblance with the assumptions on suitable central limit theorem for Jacobians in the weak identification literature (see Kleibergen (2005)).

A.2 Proofs

Proof of Proposition 3.1: For notational simplicity, in what follows we drop the S subscript from the definition of $M_{N,S}(\cdot)$ since S is assumed fixed.

The proof proceeds by showing that $\|M(\beta^0, \gamma^0, \widehat{\theta}_N)\| = o_P(1)$. Under Assumptions **A2** and **A3**, this condition is sufficient for $\widehat{\theta}_N \xrightarrow{P} \theta^0$ by virtue of (28), (30), (31) [where the last two give: $\widehat{\gamma}_N \in \mathcal{N}_\delta(\gamma^0)$ and $\widehat{\beta}_N \in \mathcal{N}_\delta(\beta^0)$ respectively with probability approaching 1], and the continuity implied by Assumptions **A1**(b), **A6**(a) and **A5**(b). Note that by the triangle inequality:

$$\begin{aligned}\|M(\beta^0, \gamma^0, \widehat{\theta}_N)\| \leq & \|M(\beta^0, \gamma^0, \widehat{\theta}_N) - M(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N)\| + \|M(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N) - M_N(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N)\| \\ & + \|M_N(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N)\|.\end{aligned}$$

By (30), (31), and the continuity implied by Assumptions **A1**(b), **A6**(a) and **A5**(b), the first term on the right hand side, i.e., $\|M(\beta^0, \gamma^0, \widehat{\theta}_N) - M(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N)\|$ is $o_P(1)$. (30), (31) and Assumption **A5**(c) imply that the second term $\|M(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N) - M_N(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N)\|$ is $o_P(1)$. The definition in (23) implies that the third term $\|M_N(\widehat{\beta}_N, \widehat{\gamma}_N, \widehat{\theta}_N)\| \leq \|M_N(\widehat{\beta}_N, \widehat{\gamma}_N, \theta^0)\| = \|M(\widehat{\beta}_N, \widehat{\gamma}_N, \theta^0)\| + o_P(1)$ where the equality follows from (30), (31) and Assumption **A5**(c). Since (30), (31) and, as before, the continuity of $M(\beta, \gamma, \theta^0)$ in β and γ imply that $\|M(\widehat{\beta}_N, \widehat{\gamma}_N, \theta^0)\| = \|M(\beta^0, \gamma^0, \theta^0)\| + o_P(1)$, it follows by (28) that the third term is also $o_P(1)$. ■

Proof of Proposition 3.2: For notational simplicity, again, in what follows we drop the S subscript from the definition of $M_{N,S}(\cdot)$.

Since $\widehat{\theta}_N \xrightarrow{P} \theta^0$, it follows by (28) and Assumption **A7**(d) that $\|\widehat{\theta}_N - \theta^0\| = O_P\left(\|M(\beta^0, \gamma^0, \widehat{\theta}_N)\|\right)$. Under our maintained assumptions and (30) and (31), it can then be shown that $\|M(\beta^0, \gamma^0, \widehat{\theta}_N)\|$ and hence $\|\widehat{\theta}_N - \theta^0\|$ is $O_P(N^{-1/2})$. Details are available from the authors. Given this, and that our assumptions are essentially same as that in Theorem 3.5 of Pakes and Pollard (1989), the rest of the proof is also similar. Hence we only provide a sketch of the proof below, and highlight the differences that appear only to the end of the proof.

For now let $d_\theta = d_\beta$. Justifying by virtue of (30), (31) and the $\sqrt{N}$-consistency of $\widehat{\theta}_N$, linearize $M_N(\widehat{\zeta}_N, \theta)$ in a $\sqrt{N}$-neighborhood of θ^0 by the function [see, for example, Chen et al. (2003)]:

$$L_N(\theta) := M_N(\beta^0, \gamma^0, \theta^0) + \frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \beta'}(\widehat{\beta}_N - \beta^0) + \frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \gamma'}(\widehat{\gamma}_N - \gamma^0) + \frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \theta'}(\theta - \theta^0).$$

Define $\theta_N^* = \arg\min_\theta \|L_N(\theta)\|$. For the application of Assumption **A7**(e) in the remainder of the proof choose δ_N such that $\widehat{\beta}_N \in \mathcal{N}_{\delta_N}(\beta^0)$, $\widehat{\gamma}_N \in \mathcal{N}_{\delta_N}(\gamma^0)$, and both $\widehat{\theta}_N, \theta_N^* \in \mathcal{N}_{\delta_N}(\theta^0)$. It can now be shown (details available from the authors) by (30), (31), Assumption **A7**(e) and the $\sqrt{N}$-consistency of $\widehat{\theta}_N$ that $\|M_N(\widehat{\beta}_N, \widehat{\gamma}_N, \theta) - L_N(\theta)\| = o_P(N^{-1/2})$ for both $\theta = \widehat{\theta}_N$ and $\theta = \theta_N^*$, and thus, subsequently, by Assumption **A7**(d) that

$$\sqrt{N}(\widehat{\theta}_N - \theta^0) = \sqrt{N}(\theta_N^* - \theta^0) = o_p(1). \tag{35}$$

Now note by (32): $\widehat{\beta}_N(\theta^0)$ satisfies $0 = M_N(\widehat{\beta}_N(\theta^0), \widehat{\gamma}_N, \theta^0)$. Expanding the right hand side gives:

$$0 = M_N(\beta^0, \gamma^0, \theta^0) + \frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \beta'}(\widehat{\beta}_N(\theta^0) - \beta^0) + \frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \gamma'}(\widehat{\gamma}_N - \gamma^0) + o_P(N^{-1/2}). \tag{36}$$

On the other hand, since $\theta_N^* = \arg\min_\theta \|L_N(\theta)\|$, it follows that $o_P(N^{-1/2}) = L_N(\theta_N^*)$. Hence by the definition of $L_N(\theta_N^*)$ and using $\sqrt{N}$-consistency of $\widehat{\beta}_N$, $\widehat{\gamma}_N$ and θ_N^* it follows that:

$$o_P(N^{-1/2}) = M_N(\beta^0, \gamma^0, \theta^0) + \frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial(\beta', \gamma', \theta')}\left[(\widehat{\beta}_N - \beta^0)', (\widehat{\gamma}_N - \gamma^0)', (\theta_N^* - \theta^0)'\right]'. \tag{37}$$

Therefore, equating (36) and (37) gives:

$$\frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \theta'}\sqrt{N}(\theta_N^* - \theta^0) = -\frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \beta'}\sqrt{N}(\widehat{\beta}_N - \widehat{\beta}_N(\theta^0)) + o_P(1).$$

Until now in this proof we have disregarded the over-identifying nature of the system with respect to θ. However, when $d_\theta < d_\beta$, and $A_N \xrightarrow{P} A$ (positive definite), under Assumption **A7**(d), standard methods modify the above relation as, up to an $o_P(1)$ term:

$$\frac{\partial M'(\beta^0, \gamma^0, \theta^0)}{\partial \theta} A \frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \theta'}\sqrt{N}(\theta_N^* - \theta^0) = -\frac{\partial M'(\beta^0, \gamma^0, \theta^0)}{\partial \theta} A \frac{\partial M(\beta^0, \gamma^0, \theta^0)}{\partial \beta'}\sqrt{N}(\widehat{\beta}_N - \widehat{\beta}_N(\theta^0)).$$

From the above expression, and **Remark 3**, we see that θ_N^*, and hence $\widehat{\theta}_N$, depends on S only through the dependence of $\bar{\xi}_{N,S}$ on S.

Differentiating $M(\beta^0(\theta), \gamma^0, \theta^0)$ with respect to θ at $\theta = \theta^0$ and using Assumption **A7**(d):

$$\frac{\partial}{\partial \theta'}M(\beta^0(\theta^0), \gamma^0, \theta^0) = \frac{\partial}{\partial \beta'}M(\beta^0(\theta^0), \gamma^0, \theta^0)\frac{\partial}{\partial \theta'}\beta^0(\theta^0) = G_0 \frac{\partial}{\partial \theta'}\beta^0(\theta^0)$$

where the last equality follows by Assumption **A3**, **A4**, **A7**(a), (b) and **MAR-X** in (2). Combining the above and using (34) and (35) we obtain:

$$\sqrt{N}(\widehat{\theta}_N - \theta^0) = \left[\frac{\partial \beta^0(\theta^0)'}{\partial \theta} G_0' A G_0 \frac{\partial \beta^0(\theta^0)}{\partial \theta'}\right]^{-1} \frac{\partial \beta^0(\theta^0)'}{\partial \theta} G_0' A \sqrt{N}\left[\bar{\xi}_{N,S} - C_0(\widehat{\gamma}_N - \gamma^0)\right] + o_P(1)$$

where $C_0 = \Omega_{12} - \Omega_{12}(\theta^0)$ and $\bar{\xi}_{N,S} = \frac{1}{N}\sum_{i=1}^N \xi_{i,S} \equiv \frac{1}{N}\sum_{i=1}^N \left[m_i^*(\gamma^0, \beta^0) - \frac{1}{S}\sum_{s=1}^S m_{is}^*(\gamma^0, \beta^0; \theta^0)\right]$. Again, by similar arguments to those in **Remark 3**, $\sqrt{N}(\widehat{\theta}_N - \theta^0)$ depends on S only through the dependence of $\bar{\xi}_{N,S}$ on S, which under the maintained assumptions yields

$$\sqrt{N}(\widehat{\theta}_N - \theta^0) \rightarrow_d N(0, \Sigma(A)). \blacksquare$$

Proof of Proposition 3.3: For notational simplicity, we will drop the N subscript from h (with the understanding that for any given N, $h > 0$ but $h = o(N^{-1/2})$) and the S subscript from the definition of $M_{N,S}^h(\cdot)$. Also, since the weighting matrix A_N can be handled in the same manner as in **Propositions** 3.2, we only consider the just-identified case ($d_\theta = d_\beta$) and take $A_N = A = I_{d_\beta}$. The proof now proceeds in two steps, first we demonstrate consistency of $\widetilde{\theta}_N^h$ for θ^0, and we then demonstrate $\|\widetilde{\theta}_N^h - \widehat{\theta}_N\| = o_P(N^{-1/2})$. The entire proof closely follows that of **Propositions** 3.1 and 3.2 except that having established consistency we slightly deviate to emphasize the fact that $M_N^h(\beta, \gamma, \theta)$ is indeed differentiable with respect to θ for $h > 0$.

Consistency: Following **Proposition** 3.1, by continuity of $M(\beta, \gamma, \theta)$ in θ, the result follows if $\|M(\beta^0, \gamma^0, \widetilde{\theta}_N^h)\| = o_P(1)$ as $h \to 0$. This condition will be sufficient for $\widetilde{\theta}_N^h \xrightarrow[h\to 0]{P} \theta^0$ by the same arguments as **Proposition** 3.1. By the triangle inequality:

$$\begin{aligned}\|M(\beta^0, \gamma^0, \widetilde{\theta}_N^h)\| \leq & \|M(\beta^0, \gamma^0, \widetilde{\theta}_N^h) - M(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h)\| + \|M(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h) - M^h(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h)\| \\ & + \|M^h(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h) - M_N^h(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h)\| + \|M_N^h(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h)\|. \qquad (38)\end{aligned}$$

As before, by Assumptions **A1**(b), **A6**(a) and **A5**(b), $\|M(\beta^0, \gamma^0, \widetilde{\theta}_N^h) - M(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h)\|$ is $o_P(1)$. For the second term on the RHS of (38) note that, due to (30) and (31), $\widehat{\gamma}$ and $\widehat{\beta}$ belong respectively in $\mathcal{N}_\delta(\gamma^0)$ and $\mathcal{N}_\delta(\beta^0)$ with probability approaching one. Hence the second term is $o_P(1)$ by Assumption **A8**(a) and the condition that $h \to 0$. Similar arguments give the third term on the RHS to be $o_P(1)$ by virtue of Assumption **A8**(b). Finally consider the fourth term and note that: $\|M_N^h(\widehat{\beta}_N, \widehat{\gamma}_N, \widetilde{\theta}_N^h)\| \leq \|M_N^h(\widehat{\beta}_N, \widehat{\gamma}_N, \theta^0)\| + o_P(1) = \|M^h(\widehat{\beta}_N, \widehat{\gamma}_N, \theta^0)\| + o_P(1)$ where the first inequality follows from (25) and the second by Assumption **A8**(b). Now, (i) the Lipschitz continuity of M^h in **A8**(a), (ii) continuity of $M(.)$ with respect to β and γ that is implied by Assumptions **A1**(b), **A5**(b) and **A6**(a), along with (iii) (30) and (31) give for $h \to 0$, $\|M^h(\widehat{\beta}_N, \widehat{\gamma}_N, \theta^0)\| = \|M(\widehat{\beta}_N, \widehat{\gamma}_N, \theta^0)\| + o_P(1) = \|M(\beta^0, \gamma^0, \theta^0)\| + o_P(1)$, and this is $o_P(1)$ by (29). Hence the fourth term is also $o_P(1)$ and thus it follows that $\widetilde{\theta}_N^h \xrightarrow[h\to 0]{P} \theta^0$.

Asymptotic equivalence: In a just-identified model, $\widetilde{\theta}_N^h$ satisfies the definition in (25) if

$$o_P(1) = \sqrt{N} M_N^h(\widehat{\beta}, \widehat{\gamma}_N, \widetilde{\theta}_N^h).$$

Denoting $\zeta = (\beta', \gamma', \theta')'$ for simplicity, and expanding the RHS we obtain:

$$\begin{aligned} o_P(1) \quad = \quad & \sqrt{N} M_N^h(\zeta^0) + \frac{\partial}{\partial \beta'} M_N^h(\bar{\zeta}_{\beta,N})\sqrt{N}(\widehat{\beta}_N - \beta^0) + \frac{\partial}{\partial \gamma'} M_N^h(\bar{\zeta}_{\gamma,N})\sqrt{N}(\widehat{\gamma}_N - \gamma^0) \\ & + \frac{\partial}{\partial \theta'} M_N^h(\bar{\zeta}_{\theta,N})\sqrt{N}(\widetilde{\theta}_N^h - \theta^0)\end{aligned}$$

for some (row-by-row) mean-values $\bar{\zeta}_{\beta,N}$, $\bar{\zeta}_{\gamma,N}$ and $\bar{\zeta}_{\theta,N}$. Therefore, by $\sqrt{N}$-consistency of $\widehat{\beta}_N$ and $\widehat{\gamma}_N$ from (31) and (30), consistency of $\widetilde{\theta}_N^h$ (just established above), uniform convergence in Assumptions **A8**(c)(i) (applied to the second and third terms on RHS) and **A8**(c)(ii) (applied to the last term on RHS), the continuity assumption in **A8**(d), it follows that

$$o_P(1) = \sqrt{N} M_N^h(\zeta^0) + \frac{\partial M(\zeta^0)}{\partial \zeta'} \sqrt{N} \left[(\widehat{\beta}_N - \beta^0)', (\widehat{\gamma}_N - \gamma^0)', (\widetilde{\theta}_N^h - \theta^0)' \right]'.$$

Finally take $\delta_N > 0$ and $\delta_N = o(N^{-1/2})$, and note that:

$$\begin{aligned} \sup_{h \in (0,\delta_N)} \sqrt{N} \| M_N^h(\zeta^0) - M_N(\zeta^0) \| &\leq \sup_{h \in (0,\delta_N)} \sqrt{N} \| (M_N^h(\zeta^0) - M^h(\zeta^0)) - (M_N(\zeta^0) - M(\zeta^0)) \| \\ &\quad + \sup_{h \in (0,\delta_N)} \sqrt{N} \| M^h(\zeta^0) - M(\zeta^0) \| \\ &\leq o_P(1) + \sqrt{N} b \times \delta_N \end{aligned}$$

with probability approaching 1, respectively by Assumptions **A8** (d) (along with the fact that $M(\zeta^0) = 0$) and (a). Since $\delta_N = o(N^{-1/2})$ as dictated by the statement of the Proposition, it now follows that $\sup_{h \in (0,\delta)} \sqrt{N} \| M_N^h(\zeta^0) - M_N(\zeta^0) \| = o_P(1)$ and hence

$$o_P(1) = \sqrt{N} M_N(\zeta^0) + \frac{\partial M(\zeta^0)}{\partial \zeta'} \sqrt{N} \left[(\widehat{\beta}_N - \beta^0)', (\widehat{\gamma}_N - \gamma^0)', (\widetilde{\theta}_N^h - \theta^0)' \right]' = \sqrt{N} L_N(\widetilde{\theta}_N^h)$$

for $L_N(\theta)$ defined in the proof of **Proposition** 3.2. Therefore, $\| L_N(\widetilde{\theta}_N^h) \| = o_P(N^{-1/2})$. Now by following the same steps as in that proof we obtain $\sqrt{N} \| \widetilde{\theta}_N^h - \widehat{\theta}_N \| = o_P(1)$. ■

1130, rue Sherbrooke Ouest, bureau 1400, Montréal (Québec) H3A 2M8
Tél. : 514-985-4000 • Téléc. : 514-985-4039
www.cirano.qc.ca • info@cirano.qc.ca

Centre interuniversitaire de recherche en analyse des organisations
Center for Interuniversity Research and Analysis on Organizations