Efficient GMM estimation with general missing data patterns

Chris Muris*

First version: June 2011. This version: December 2013.

Abstract

This paper considers GMM estimation from a random sample of incomplete observations. For each observation, certain components of the moment function may be unavailable. I propose an estimator for an arbitrary set of moment conditions and a general missing data pattern. The estimator is consistent and asymptotically efficient under an assumption that is weaker than missing completely at random. The estimator can be interpreted as the optimal linear combination of subsample GMM estimators. I also propose an inverse probability weighted version of the estimator that is consistent when selection is on observables. Applications to instrumental variable estimation and dynamic panel data estimation demonstrate the efficiency gain with respect to standard missing data methods.

JEL Classification: C13, C20, C23, C30.

Keywords: missing data, GMM, efficiency.

*Department of Economics, Simon Fraser University. Email: cmuris@sfu.ca. I am grateful to Ramon van den Akker, Richard Blundell, Otilia Boldea, Irene Botosaru, Pedro Duarte Bom, Katherine Carman, Miguel Atanasio Carvalho, Toru Kitagawa, Tobias Klein, Andrea Krajina, Jan Magnus, Bertrand Melenberg, Krishna Pendakur, Franco Peracchi, Pedro Santos Raposo, and Bas Werker for encouraging and insightful discussions. I also thank the seminar participants at Tilburg University, University of Bristol, Institute of Advanced Studies Vienna, Simon Fraser University, Monash University, Victoria University, and at the 2011 EEA-ESEM and the 2012 International Panel Data Conference.

1 Introduction

Missing data is prevalent in empirical research in economics. For example, Abrevaya and Donald (2010) find that missing data occurs in 40% of the publications in top economics journals. In 70% of these cases, all incomplete observations are discarded, and the analysis is then carried out with the resulting complete subsample. This is inefficient if the discarded observations contain information about the parameter of interest.

In this paper, I introduce a general GMM estimation procedure that deals with efficient estimation in the presence of incomplete observations. To the best of my knowledge, this is the first paper to address this issue.

The proposed estimation procedure combines in an efficient way the information from both complete and incomplete observations. The estimator achieves efficiency by splitting the available sample in subsamples based on the missing data pattern, and then combining optimally the efficient estimators from each subsample. No restrictions are imposed on the missing data pattern, which implies that the data can be incomplete in an arbitrary way. In terms of the missing data mechanism, it is assumed that either there is no selection or that the selection is on observables. The procedure introduced can be applied to two-step, iterative and continuous updating GMM estimators.

The procedure is shown to be consistent under an assumption that is weaker than missing completely at random (MCAR). MCAR requires that the data are statistically independent of the missing data indicator. We only require that the moment conditions hold conditional on the missing data indicator. Under this assumption, the proposed estimator attains the semiparametric efficiency bound. The extension to missing at random (MAR) is straightforward. Furthermore, the computational burden is comparable to that of the full data estimator, since the minimization problem for the missing data estimator is a linear combination of those for the full data estimators. To illustrate the problem and the proposed solution, consider the following two examples.

Example 1: Instrumental variables. First, consider instrumental variables estimation and dynamic panel data models. First, a linear instru-

mental variable model with one endogenous variable X and two instruments $Z = (Z_1, Z_2)$ is given by the relationship $y = X\beta_0 + u$ and the conditional mean assumption $\mathbb{E}(u|\, Z) = 0$. Estimation of the parameter β_0 is based on the unconditional moment condition $\mathbb{E}(Zu) = 0$. Assume that for each observation both y and X are observed. An observation with no measurement for instrument Z_2 will still be useful if the other instrument, Z_1, is observed. To see this, consider the subsample of all observations for which only (y, X, Z_1) is observed. This subsample is informative, since we can use it to estimate β_0 using the moment condition $\mathbb{E}(Z_1 u) = 0$.

In this example, we distinguish three subsamples. Observations with measurements on both instruments are placed in the first subsample; observations with only the first instrument available are placed in the second subsample; the third subsample contains the observations that only have measurements on the second instrument. Under our assumption on the missing data mechanism, β_0 can be consistently estimated using each subsample separately. Using efficient GMM in each subsample yields three consistent estimators of β_0. Any weighted average of these estimators is consistent for β_0. We show that there exist optimal weights that minimize the asymptotic variance of the estimator. Interestingly, the result does not require identification in each subsample. For example, in a setting with two endogenous variables, the information in a subsample with just one instrument will be exploited, as long as the parameter can be identified using all subsamples jointly.

Missing instruments are common in empirical research. For example, Levitt (2002) estimates the effect of police on crime in US cities. He uses the number of firefighters and the number of city workers as instruments for the number of police officers. However, not all cities provide information about both instruments in each year. Another example can be found in Rodrik et al. (2004), who investigate the effect of institutions and geography on economic growth by using trade predictions and settler mortality rates as instruments. However, for some observations it is not possible to observe both instruments at the same time.

Example 2: Dynamic panel data models. A second example with

	Missing components			
	None	$y_{i,1}$	$y_{i,4}$	$(y_{i,1}, y_{i,4})$
$y_{i,1}\Delta\epsilon_{i,3}$	X	·	X	·
$y_{i,1}\Delta\epsilon_{i,4}$	X	·	·	·
$y_{i,1}\Delta\epsilon_{i,5}$	X	·	·	·
$y_{i,2}\Delta\epsilon_{i,4}$	X	X	·	·
$y_{i,2}\Delta\epsilon_{i,5}$	X	X	·	·
$y_{i,3}\Delta\epsilon_{i,5}$	X	X	·	·

Table 1: Missing data patterns for dynamic panel data estimation using the estimator in Arellano and Bond (1991), $T = 5$.

incomplete, informative observations comes from dynamic panel data models. Interest is in the autoregressive parameter ρ in

$$y_{i,t} = \alpha_i + \rho y_{i,t-1} + u_{i,t},\ 2 \leq t \leq T.$$

Arellano and Bond (1991) propose an estimator that is based on the absence of serial correlation in the error terms, which implies the moment conditions

$$\mathbb{E}(y_{i,t-s}\Delta u_{i,t}) = 0,\ t \geq 3,\ s \geq 2.$$

Table 1 illustrates the relationship between the incompleteness of an observation and the extend to which that observation contributes to the sample moment. In Table 1, we consider the case of $T = 5$ time periods and six moment conditions. If $y_{i,1}$ is missing, observation i still contributes to three sample moments. If $y_{i,4}$ is missing, only one component of the moment function can be evaluated. More generally, the estimator proposed in this paper efficiently accommodates static and dynamic panel data models with unbalanced panels with different starting points, different endpoints, and any combination of gaps.

Related literature. There is an extensive literature on missing data settings in which each observation contributes either to all, or to none of the sample moments. This literature was initiated by Robins et al. (1994) and is

active, with recent contributions by Wang et al. (2004), Wooldridge (2007), Chen et al. (2008), Graham et al. (2010) and Graham (2010). Extending this literature to a general missing data pattern is theoretically and computationally challenging, see for example Tsiatis (2006, p. 255).

This paper does not add to this literature, or to the literature on missing data in univariate regression methods. In these settings, as soon as an observation is incomplete it will contribute to none of the sample moments and is therefore uninformative in our framework. The same holds for univariate instrumental variables case with missing dependent or endogenous variables. In the univariate regression model, efficiency gains can be obtained if one is willing to sacrifice consistency, see Dardanoni et al. (2009).

Several papers consider specific GMM settings or specific missing data patterns. The static panel data setting is investigated by Chen et al. (2010). Abowd et al. (2001) allow for attrition in a dynamic panel data model. Instrumental variables estimation with missing instruments is discussed in Abrevaya and Donald (2010) and Mogstad and Wiswall (2010). Verbeek and Nijman (1992) study a static panel data setting and exploit the existence of different missing data patterns to test for selectivity bias.

Plan of the paper. Section 2 introduces the model. Section 3 introduces the estimator and establishes its properties (consistency, asymptotic normality, efficiency). Section 4 considers the special case where the parameter is identified in each subsample. Section 5 extends the procedure to a generalized inverse probability weighting estimator to accommodate for selection on observables. Section 6 gives examples, and show that substantial efficiency gains over standard approaches are possible. Section 7 concludes.

2 Model

The starting point for our model is a standard GMM setting. The contribution of this paper is that the model is then extended to accommodate general patterns of missing data. First, consider a GMM setting in the absence of missing data. We are interested in a finite-dimensional parameter of interest

$\theta_0 \in \Theta \subset \mathbb{R}^p$, defined through the moment conditions

$$\mathbb{E}(h(X,\theta)) = 0 \Leftrightarrow \theta = \theta_0,$$

for a random vector $X \in \mathbb{R}^d$ and a known q-dimensional moment function $h : \mathbb{R}^d \times \Theta \to \mathbb{R}^q$. If a random sample $\{X_i,\, i = 1, \cdots, n\}$ is available, a GMM estimator would minimize

$$h_n(\theta)' W_n h_n(\theta),$$

where $h_n(\theta) = \frac{1}{n}\sum_{i=1}^n h(X_i,\theta)$ is the sample analog of $\mathbb{E}(h(X,\theta))$, and W_n is an appropriately chosen weight matrix.

Often, we do not observe all components of X_i for each observation i. Example include missing instruments and unbalanced panel data (see Introduction). A missing data indicator R_i describes the missing data pattern for observation i. This indicator is defined as a diagonal $q \times q$ matrix with a diagonal entry equal to 1 if the corresponding element of $h(X_i,\theta)$ can be computed, and zero otherwise. For example, let $q = 3$. For a complete observation, all components of $h(X_i,\theta)$ can be computed for all $\theta \in \Theta$, and $R_i = I_3$. An observation for which no elements of X_i are observed carries no information and has $R_i = 0_{3\times 3}$. An intermediate case of an incomplete but informative observation would occur if the first and third component of $h(X_i,\theta)$ can be computed for all $\theta \in \Theta$, but not the second. In that case,

$$R_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

For a given setting, let $\{S_1, \cdots, S_J\}$ be the set of all possible missing data patterns for a given estimation problem. Note that $J \leq 2^q$. Note that any combination of missing data patterns is allowed. For unbalanced panel data, it accommodates attrition, refreshment, arbitrary starting and ending points, as well as any combination of gaps. Initially, I will maintain the following set

of assumptions:

Assumption 1. *Assume that:*

(i) *a random sample of* $\{(R_i, R_i h(X_i, \theta), i = 1, \cdots, n)\}$ *are observed for each* θ*;*

(ii) *the model holds in each subpopulation:* $\mathbb{E}(h(X_i, \theta_0) \mid R_i = S_j) = 0$ for each j*;*

(iii) *identification:* $\theta \neq \theta_0 \Rightarrow \exists j : \mathbb{E}(R_i h(X_i, \theta) | R_i = S_j) \neq 0$*;*

(iv) *every element of* h *is observed in at least one subpopulation:* $\operatorname{rk}\left(\sum_{j=1}^{J} S_j\right) = q$*;*

(v) *the probability of observing pattern* j *is positive,* $p_j = \mathbb{P}(R_i = S_j) > 0$ *for each* j*.*

A discussion of the assumptions now follows, before discussing estimation under Assumption 1 in Section 3. Assumption 1(i) corresponds to the standard assumption of random sampling in microeconometrics. This assumption can be substantially weakened, but this would take away from the main message of this paper.

Assumption 1(ii) is a weak version of "missing completely at random" (MCAR: $X_i \perp R_i$). For a comparison between the two, see Remark 1. For an interpretation, consider partitioning the population into J subpopulations based on their missing data pattern R_i. Then, 1(ii) requires the model to be valid in every subpopulation. In Section 5, we show that the extension to a substantially weaker assumption, a mean-independence version of "missing at random" (MAR: $X_i \perp R_i | Z_i$ for some Z_i), is straightforward. Finally, 1(ii) can be weakened to require that only the observable part of the model is valid: $\mathbb{E}(R_i h(X_i, \theta_0) | R_i = S_j) = 0$.

Assumption 1(iii) establishes identification. It is slightly stronger than identification in the standard GMM setting without missing data, $\mathbb{E}(h(X_i, \theta)) = 0 \Leftrightarrow \theta = \theta_0$. It ensures that we can empirically determine the non-zero expectation of $h(X_i, \theta)$, $\theta \neq \theta_0$. To see this, assume that the $q-$th moment

condition is violated in subpopulation J, and that this subpopulation does not provide information on that moment function, $S_{J,q,q} = 0$. In that case, the model does not hold in that subpopulation, but this cannot be detected by the econometrician. A sufficient, but not necessary, condition for 1(iii) is identification in the full data GMM setting, plus the existence of a subsample for which we observe all the elements of $h(\cdot)$, i.e. $R_1 = I_q$.

The existence of such a subsample would also be sufficient for Assumption 1(iv), which requires that for each element of $h(\cdot)$, there is at least one subsample for which we can compute it. Assumption 1(iv) would also hold if we observed $J = q$ subpopulations, each of which give us information on one particular element of $h(\cdot)$. Assumption 1(v) is not restrictive. Missing data patterns that occur with zero probability can be removed from the setting, as long as this does not jeopardize Assumption 1(iv).

Remark 1. Assumption 1(ii) is weaker than "missing completely at random" (MCAR),[1] which requires $X_i \perp R_i$. Heckman et al. (1998, p. 267) make a similar distinction in the context of the estimation of the average treatment effect. MCAR requires that the missing data indicator is statistically independent of the realization. For any moment function h, MCAR implies $h(X_i, \theta) \perp R_i$ for each $\theta \in \Theta$. This, in turn, implies Assumption 1(ii), which only requires the moment conditions to hold conditional on each missing data pattern, and only requires it to hold at the true value θ_0. To demonstrate the difference between 1(ii) and MCAR, consider the univariate linear regression model, $y_i = \beta x_i + \epsilon_i$, $\mathbb{E}(\epsilon_i | X_i) = 0$. In Figure 1, we present the regression line and some simulated data. A cross represents an observation that is missing, $r_i = 0$; a dot represents an observation that is complete, $r_i = 1$. The sample can be split in two groups: those with low x_i and those with high x_i. In terms of deviation from the regression line, the data are arbitrarily missing in the sense that the estimator that uses the missing data has the same expectation as the estimator that uses the complete data. However, the situation in Figure 2.1 does not satisfy MCAR: an observation in the low group has a positive

[1] For a detailed discussion of MCAR and related concepts, see Little and Rubin (2002, Chapter 1).

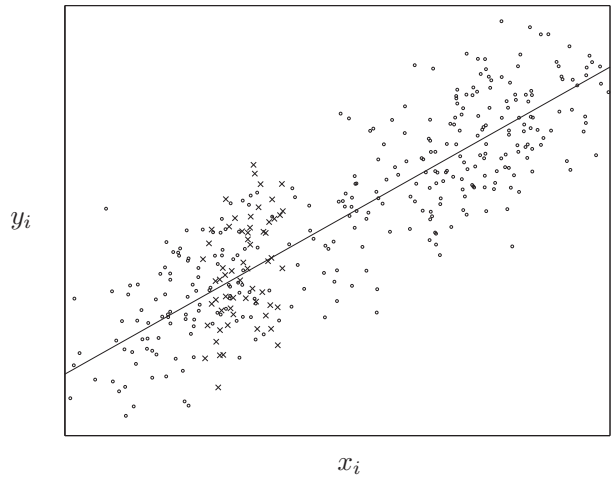

Figure 2.1: Assumption 1(ii) is weaker than MCAR. Simulated data for a univariate regression model. A cross represents a missing observation; a dot represents a complete observation.

probability of being missing, while an observation in the high group is always complete, so $\mathbb{P}(r = 1 \mid X \text{ low}) \neq \mathbb{P}(r = 1 \mid X \text{ high})$. This setting satisfies Assumption 1(ii), since $\mathbb{E}(x_i\epsilon_i | \, r = 1) = \mathbb{E}(x_i\epsilon_i | \, r = 0) = 0$. However, MCAR implies independence of the variance, and mean independence at values of the parameter other than the true value of β, which are not satisfied in this example: $\text{var}(x_i\epsilon_i | r_i = 1) > \text{var}(x_i\epsilon_i | r_i = 0)$, and $\mathbb{E}(x_i(y_i - (\beta + 1)x_i | r_i)) = \mathbb{E}(x_i\epsilon_i | r_i) - \mathbb{E}(x_i^2 | r_i) = -\mathbb{E}(x_i^2 | r_i) \neq \mathbb{E}(x_i^2)$.

3 Estimation

This section introduces a class of GMM estimators that deal with general missing data patterns as modeled in Section 2. Consistency and asymptotic normality of the estimators are established. Finally, I show that the minimum-variance estimator in this class is asymptotically efficient under Assumption 1.

A GMM estimator for the parameter of interest θ_0 based on a random

sample of complete data $\{X_1, \cdots, X_n\}$ would minimize

$$h_n(\theta)' W_n h_n(\theta), \tag{3.1}$$

for some symmetric positive definite matrix W_n. This estimator is not feasible under Assumption 1, since $h(X_i, \theta)$ is only observable when $R_i = I$. Let the *$j-$th subsample* consist of the observations with $R_i = S_j$, or $G_j = \{i : R_i = S_j\}$. Let n_j be the number of observations in the $j-$th subsample. The *$j-$th subsample moment* $h_{n,j}(\theta)$ is the sample analog of the observable moment functions in the $j-$th subsample:

$$h_{n,j}(\theta) = (1/n_j) \sum_{i \in G_j} R_i h(X_i, \theta).$$

A missing data GMM estimator is defined as a minimizer of the sum of weighted squares of subsample moments:

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} \sum_{j=1}^{J} h_{n,j}(\theta)' W_{n,j} h_{n,j}(\theta). \tag{3.2}$$

The dependence of $\hat{\theta}_n$ on $W_n = (W_{n,j},\, j = 1, \cdots, J)$ is suppressed for notational convenience. Importantly, this objective function can be computed given Assumption 1.

Commonly used approaches used for dealing with missing data are nested in this class of missing data estimators. The *complete case* estimator dominates empirical work and is the default implementation in most statistical software packages. A complete case estimator uses only complete observations. Let $S_1 = I_q$, so that all components of h can be evaluated for observations with missing data pattern S_1. Then, the complete case sample estimator is based on the sample moment

$$h_{\text{cc},n}(\theta) = \frac{1}{n_1} \sum_{i \in G_1} R_i h(X_i, \theta) = \frac{1}{n_1} \sum_{i \in G_1} h(X_i, \theta).$$

which is equivalent to setting $W_{n,j} = 0_{q\times q}$, $j > 1$ and setting $W_{n,1}$ equal to the inverse of the variance of $h(X_i, \theta_0)$. This complete case estimator is clearly not efficient, because it only uses a fraction n_1/n of the observations.

The *available case* approach is a popular alternative that uses all the available data. For each component of the moment function it uses all the observations for which that component is observed. The available case estimator is based on the sample moment

$$h_{\mathrm{ac},n}(\theta) = \frac{1}{n}\bar{R}^{-1}\sum_{i=1}^{n} R_i h(X_i, \theta),$$

where $\bar{R} = \sum_{j=1}^{J} \frac{n_j}{n} S_j$ is the average missing data pattern. This corresponds to setting $W_{n,j} = S_j W_{n,\mathrm{ac}} S_j$ for each $j = 1, \ldots, J$, where $W_{n,\mathrm{ac}}$ can be chosen optimally. In the case of instrumental variable estimation, available case estimation corresponds to replacing missing instruments by zeros. The estimator uses all the available observations, but we will argue below that it does not use them efficiently.

By construction, the optimal estimator in the class of missing data GMM estimators is at least as efficient as the complete case and available case estimators, i.e. we can choose the $W_{n,j}$ to generate an estimator that is at least as efficient as those commonly used estimators. The examples in Section 6 demonstrate that the efficiency gain is substantial. Establishing the asymptotic distribution of the missing data estimator $\hat{\theta}_n$ requires some regularity conditions, stated below:

Assumption 2. *Assume that:*

(i) *Finite conditional variances: for each* $j \in \{1, \cdots, J\}$,
$\mathrm{var}\left(h(X_i, \theta_0) \mid R_i = S_j\right) = \Omega_j < \infty$;

(ii) *the moment function* $h(x, \cdot)$ *is continuously differentiable on* Θ *for all* x;

(iii) *for each pattern* j *let the* $q \times p$ *matrix* $D_j(\theta) = \mathbb{E}\left(\partial h(X, \theta)/\partial\theta' \mid R\right)$ *be uniformly bounded, in the sense that* $\sup_{\theta\in\Theta} \|D_j(\theta)\| < \infty$, *where* $\|D_j\| = \mathrm{tr}(D_j' D_j)^{1/2}$;

(iv) *for each pattern j, the derivative matrix has full rank at the true value: $D_j = D_j(\theta_0)$, then* $\mathrm{rk}(D_j) = p$*;*

(v) *the parameter space Θ is compact and θ_0 is in the interior of Θ;*

(vi) *the moment function is bounded in absolute mean:* $\sup_{\theta\in\Theta} \mathbb{E}(|h(X,\theta)|) < \infty$*;*

(vii) *for each subsample, the sequence of GMM weights $(W_{n,j},\ n \in \mathbb{N})$ satisfies $S_j W_{n,j} S_j = W_{n,j}$ and converges to a positive semidefinite matrix, W_j, with* $\mathrm{rk}(W_j) = \mathrm{rk}(S_j)$;

These assumptions are close to standard assumptions for consistency and asymptotic normality in GMM estimation as in Newey and McFadden (1994). Let me remark briefly where they differ. Assumptions 2(i) and (iii) emphasize that Assumption 1(ii) is weaker than MCAR ($X_i \perp R_i$), as they allow the conditional variance and expected derivative of the moment function to depend on the missing data pattern. They also impose the standard boundedness assumptions. Assumption 2(vii) sets the submatrix of W_j that corresponds to missing elements of the moment function equal to zero, and requires the remaining submatrix to be positive definite.

The asymptotic distribution of a missing data GMM estimator is established in the following theorem:

Theorem 1. *Under Assumptions 1 and 2, and as $n \to \infty$,*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, B^{-1}AB^{-1}\right),$$

where

$$\begin{aligned} A &= \sum_{j=1}^{J} \frac{1}{p_j} D_j' W_j \Omega_j W_j D_j \\ B &= \sum_{j=1}^{J} D_j' W_j D_j \end{aligned}$$

Proof. A proof can be found in Appendix A. It involves converting the conditional moment restrictions in Assumption 1(ii) to an augmented set of $J \times q$ unconditional moment conditions. The expression in (3.2) can be seen as a weighted sample analog of this set of unconditional moment conditions. This makes $\hat{\theta}_n$ a standard GMM estimator. □

In this class of missing data GMM estimators, the asymptotic variance can be minimized by setting each W_j equal to $W_j^* = p_j(S_j\Omega_j S_j)^+$. Note that this reduces to the familiar optimal weighting matrix if $J = 1$, $p_1 = 1$, and $S_1 = I_q$, in which case $W_1^* = \Omega_1^{-1}$. The weight given to a particular subsample increases with the size of the subpopulation $n_j \approx n \cdot p_j$, and it decreases with the variance of the the moment functions observable in the subpopulation. This extends a familiar result on optimal GMM estimation to our setting with general missing data patterns. The estimator that uses these optimal weighting matrices $W_n^* = \left(W_{n,1}^*, \ldots, W_{n,J}^*\right)$ is denoted by $\hat{\theta}_n^*$ and has limiting distribution

$$\sqrt{n}\left(\hat{\theta}_n^* - \theta_0\right) \xrightarrow{d} N\left(0, V\right). \tag{3.3}$$

where

$$V = \left(\sum_{j=1}^{J} p_j D_j'(S_j\Omega_j S_j)^+ D_j\right)^{-1}. \tag{3.4}$$

Semiparametric efficiency of $\hat{\theta}_n^*$. The missing data GMM model under consideration is a semiparametric model: we are estimating a finite-dimensional parameter θ_0 while leaving the full distribution of X_i unspecified. Consider some (smooth) parametric submodel, so that the distribution is described by a finite-dimensional parameter. The Cramer-Rao lower bound guarantees a lower bound on the variance of any regular estimator in this parametric submodel. Now consider a semiparametric estimator that is regular in every parametric submodel. The variance of this estimator must be at least as large as the supremum of the lower bounds in all parametric submodels. This supremum is called the semiparametric efficiency bound (SPEB). More information about regularity and the semiparametric efficiency bound can be found in Bickel et al.

(1993), Newey (1990), and Van der Vaart (2000, Chapter 25).

For the following theorem, the result for conditional moment restrictions for singular covariance matrices in Newey (2001), which extends a result in Chamberlain (1987), is important. The result shows that, under Assumption 1, the optimally missing data GMM estimator $\hat{\theta}_n^*$ is asymptotically efficient for θ_0.

Theorem 2. *The semiparametric efficiency bound for θ_0 is $SPEB(\theta_0) = V$, where V is as in* (3.4)*.*

4 Subsample estimation

In some situations, θ_0 can be estimated using each subsample separately. An example can be found in instrumental variable estimation with at least as many instruments as endogenous variables in each pattern. This section shows that, in such cases, an optimal linear combination of the optimal subsample GMM estimators is asymptotically efficient. Studying this special case is useful because it can be implemented using full data estimation routines. Furthermore, it shows that the missing data approach proposed in this paper can be extended without modification to other estimation frameworks, such as generalized empirical likelihood estimation.

Assume that θ_0 can be consistently estimated using each subsample separately. This corresponds to the following strengthening of Assumptions 1(ii) and 1(iii):

Assumption: Subsample identification. *Assume that, for each j, we have:* $\mathbb{E}\left(R_i h\left(X_i, \theta\right) \middle| R_i = S_j\right) = 0 \Leftrightarrow \theta = \theta_0$.

Consider the optimal GMM estimator in each subsample:

$$\hat{\theta}_{n,j} = \operatorname{argmin}_{\theta \in \Theta} h_{n,j}(\theta)' W_{n,j}^* h_{n,j}(\theta),$$

where $W_{n,j}^*$ is chosen so that it converges to the optimal weighting matrix $W_j^* = (S_j \Omega_j S_j)^+$. Assume a standard GMM setting as in the previous section.

Then, as $n \to \infty$,

$$\sqrt{n_j}(\hat{\theta}_{n,j} - \theta_0) \xrightarrow{d} N\left(0, \left(D_j'(S_j\Omega_j S_j)^+ D_j\right)^{-1}\right), \tag{4.1}$$

where the invertibility of $\left(D_j'(S_j\Omega_j S_j)^+ D_j\right)$ follows from our subsample identification assumption.

We consider weighted averages of these subsample GMM estimators, where the weights are subsample specific, $p \times p$ matrices $(A_{n,j},\ n \in \mathbb{N})$. A weighted estimator is characterized by a $J-$tuple $A_n = (A_{n,1}, \cdots, A_{n,J})$. We denote the matrix-weighted sum with matrix weights A_n by $\hat{\theta}_{A,n}$,

$$\hat{\theta}_{A,n} = \sum_{j=1}^{J} A_{n,j}\hat{\theta}_{n,j}.$$

Consistency requires $\sum_{j=1}^{J} A_j = I_p$. Under the random sampling assumption 1(i), the subsample GMM estimators are independent, so that the asymptotic variance of the weighted estimator $\hat{\theta}_{A,n}$ is given by

$$\text{Avar}\left(\hat{\theta}_{A,n}\right) = \lim_{n\to\infty} \text{var}\left(\sqrt{n}\hat{\theta}_{A,n}\right) = \sum_{j=1}^{J} \frac{1}{p_j} A_j \left(D_j'(S_j\Omega_j S_j)^+ D_j\right)^+ A_j',$$

which uses the asymptotic variance of the subsample GMM estimators in (4.1). The theorem below establishes a lower bound for the asymptotic variance of any consistent, weighted combination of independent, consistent estimators.

Theorem 3. *Let $\hat{\theta}_{j,n}$, $j = 1, \cdots, N$, be J uncorrelated estimators of a common parameter θ_0, with asymptotic distributions*

$$\sqrt{n}\left(\hat{\theta}_{j,n} - \theta_0\right) \xrightarrow{d} N\left(0, \Gamma_j\right)$$

for some invertible Γ_j. Then, for any consistent,[2] linear combination of these

[2]Consistency refers to $\lim_{n\to\infty}\left(\sum_{j=1}^{J} A_{n,j}\right) = I_p$.

estimators,

$$\hat{\theta}_{A,n} = \sum_{j=1}^{J} A_{n,j}\hat{\theta}_{n,j},$$

we have that

$$Avar\left(\hat{\theta}_{A,n}\right) - \left(\sum_{j=1}^{J} \Gamma_j^{-1}\right)^{-1}$$

is positive semidefinite. Furthermore, if the weights are set as

$$A_{j,n}^{*} = \left(\sum_{k=1}^{J} \Gamma_k^{-1}\right)^{-1} \Gamma_j^{-1}$$

then the lower bound is attained:

$$Avar\left(\hat{\theta}_{A,n}\right) = \left(\sum_{j=1}^{J} \Gamma_j^{-1}\right)^{-1}.$$

The result in Theorem 3 provides a lower bound that applies generally to uncorrelated, asymptotically Normal, estimators of a common parameter. It applies to our setting, with

$$\Gamma_j = \left(p_j D_j'(S_j\Omega_j S_j)^{+} D_j\right)^{-1}.$$

The lower bound implied by Theorem 3 is

$$\left(\sum_{j=1}^{J} p_j D_j'(S_j\Omega_j S_j)^{+} D_j\right)^{-1}$$

which equals the semiparametric efficiency bound derived in Section 3. The optimal choice of weight matrices implied by Theorem 3 is

$$A_j^{*} = V\left[p_j D_j'(S_j\Omega_j S_j)^{+} D_j\right],$$

which achieves the efficiency bound established by Theorems 2 and 3. Note that this efficient estimator requires two optimimal steps: it is the *optimal*

linear combination of the *optimal* GMM estimators in each subsample.

Remark 2. The results in this section can be applied more widely than the GMM estimation considered so far. It can be used to optimally combine estimators obtained using many estimation method under random sampling. For example, the results can be applied to generalized empirical likelihood estimation and continuous updating GMM. It is applicable to data combination methods, where different elements of the moment functions are estimated from different data sets. Another application is outlined in the following section.

5 Inverse probability weighting

For some situations, the assumption on the missing data mechanism that we have maintained until now is too strong. Remember that Assumption 1(ii) requires that the missing data indicator is mean-independent of any other random variables in our model:

Assumption 1(ii). *The model holds in each subpopulation:*

$$\mathbb{E}\left(h\left(X_i,\theta_0\right)\ \middle|\, R_i = S_j\right) = 0$$

for each $1 \leq j \leq J$.

In this Section, I introduce a weaker assumption about the missing data mechanism, and discuss consistent estimation of θ_0 in this new setting. As discussed in Section 2, Assumption 1(ii) can be seen as a less restrictive version of MCAR: $X_i \perp R_i$. A weaker assumption that is often made in the literature on missing data is *missing at random* (MAR), which requires that there exists a vector of observables Z_i such that $X_i \perp R_i | Z_i$. In the literature on program evaluation, an assumption closely related to MAR is called unconfoundedness or ignorability. The following Assumption 1(ii') stands to MAR as Assumption 1(ii) stands to MCAR:

Assumption 1(ii'). $\mathbb{E}\left(h\left(X_i,\theta_0\right)\middle|\, R_i, Z_i\right) = \mathbb{E}\left(h\left(X_i,\theta_0\right)\middle|\, Z_i\right)$.

Assumption 1(ii') allows the probability of observing pattern S_j to depend on some observables Z_i. However, it requires that, once we control for Z_i, the model again holds regardless of the observed missing data pattern. Unit i may select into subsample j, as long as this selection process only depends on covariates Z_i that are observable to the econometrician.

To define the estimators that are consistent under Assumption 1(ii'), additional notation is required. First, let Z_i be a random vector of covariates, which can contain some elements from X_i. Second, let $r_{i,j}$ be a missing data indicator function that equals 1 if and only if the missing data follow pattern j. In other words, for each $j \in \{1, \cdots, J\}$,

$$r_{i,j} = \begin{cases} 1 & \text{if } R_i = S_j, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, let $p_j(Z_i)$ be the conditional probability of observing pattern j given Z_i:

$$p_j(Z_i) = \mathbb{P}(r_{i,j} = 1 | Z_i) = \mathbb{P}(R_i = S_j | Z_i).$$

It is straightforward to allowing the set of covariates Z_i to differ across patterns j. Under Assumption 1(ii'), the following inversely weighted moment conditions hold for each j:

$$\begin{aligned} E\left[\frac{r_{i,j}}{p_j(Z_i)} h(X_i, \theta_0)\right] &= E_{Z,R}\left[E_{X|Z,R}\left[\frac{r_{i,j}}{p_j(Z_i)} h(X_i, \theta_0) \middle| Z_i, R_i\right]\right] \\ &= E_{Z,R}\left[\frac{r_{i,j}}{p_j(Z_i)} E_{X|Z,R}\left[h(X_i, \theta_0) | Z_i, R_i\right]\right] \\ &= E_Z\left[\frac{E_{R|Z}\left[r_{i,j} | Z_i\right]}{p_j(Z_i)} E_{X|Z}\left[h(X_i, \theta_0) | Z_i\right]\right] \\ &= E_Z\left[E_{X|Z}\left[h(X_i, \theta_0) | Z_i\right]\right] \\ &= 0. \end{aligned}$$

This suggests that a GMM estimator based on the inversely weighted moment conditions is consistent for θ_0. Estimation of θ_0 is therefore a straightforward application of the methods discussed in Sections 3 and 4 to the inversely

weighted moment conditions.

To see this, consider first estimation under assumption *Subsample identification* in Section 4 that the parameter θ_0 can be consistently estimated using each subsample separately.[3] In this case, estimating θ_0 in each subsample reduces to the standard missing data setting with only two possible missing data patterns (an observation is either complete or it is completely missing). In this setting, $p_j(Z_i)$ would be called the *propensity score* for subpopulation j. Consider the following procedure.

1. Estimate $p_j(Z_i)$ nonparametrically by the method described in Hirano et al. (2003). Call this estimator $\hat{p}_j(Z_i)$. Second, estimate θ_0 by optimal GMM applied to
$$E\left[\frac{r_{i,j}}{\hat{p}_j(Z_i)}h(X_i,\theta_0)\right]=0.$$
Results in Chen et al. (2008) establish the efficiency of this estimator;

2. Repeat this for each j, so that we have J efficient, independent, subsample estimators;

3. Define $\hat{\theta}_n$ as the optimal linear combination of these J optimal subsample estimators, as in Section 4, Theorem 4.

During the development of this paper, Chaudhuri (2012) has established some useful results on the efficiency bounds for this setting.

6 Examples

This section demonstrates the efficiency gains of the proposed procedures with respect to standard procedures to deal with missing data. In the first example, I discuss an instrumental variable model where the instruments are partially

[3] If there are subsamples in which θ_0 is not identifiable, estimation of θ_0 can still take into account the information from this subsample, by extending the approach in Section 2. Estimation is by GMM based on the moment functions $k(X_i,\theta,p(Z_i)) = \left(\frac{r_{i1}}{p_1(Z_i)}h(X_i,\theta_0),\cdots,\frac{r_{iJ}}{p_J(Z_i)}h(X_i,\theta_0)\right)$ with a (nonparametric) plug-in estimator as in Step 1 below.

observed. The second example is the dynamic panel data estimator proposed by Arellano and Bond (1991).

6.1 Instrumental variables

The first example is a linear instrumental variables model where the dependent and explanatory variables are always observed, but we do not always observe both of the instruments. Either instrument can be missing for a subsample. The approach is easily generalized to multiple explanatory variables, multiple instruments, and nonlinear models, but the stylized setup in this section has the advantage that it allows us to derive analytical results. The problem of partially missing instruments is common; a recent example can be found in Angrist et al. (2006).

The dependent variable y is linearly related to an explanatory variable x, $y = x\theta_0 + u$. Two instruments are available, w_1 and w_2, which motivates the following unconditional moment conditions to estimate θ_0:

$$0 = \mathbb{E}\left(\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}(y - \theta_0 x)\right) = \mathbb{E}\left(\begin{pmatrix} w_1 u \\ w_2 u \end{pmatrix}\right).$$

There are three groups of observations, $J = 3$. For the first group we observe both instruments. For the second group we observe only w_1, and for the third group we observe only w_2, so that the missing data patterns are:

$$S_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ S_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ S_3 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

We assume that the instruments are similar: they are equally likely to be missing: $p_2 = p_3 = (1 - p_1)/2$; they have the same correlation with the explanatory variable: $\mathbb{E}(w_1 x) = \mathbb{E}(w_2 x) = \lambda$; and they have the same first two moments: $\mathbb{E}(w_j) = 0$ and $\mathbb{E}(w_j^2) = 1$, $j = 1, 2$. The instruments have correlation $\rho = \mathrm{cov}(w_1, w_2)$.

We assume that the variance matrices are the same for all groups:

$$\Omega_1 = \Omega_2 = \Omega_3 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where the form of Ω could result from the additional assumptions $\mathbb{E}(w_j^2\epsilon^2) = \mathbb{E}(w_j^2)\mathbb{E}(\epsilon^2) = 1,\ j = 1, 2$. Furthermore, we normalize the variance of the explanatory variable, $\text{var}(x) = 1$. We have

$$\text{var}(x, w_1, w_2) = \begin{pmatrix} 1 & \lambda & \lambda \\ \lambda & 1 & \rho \\ \lambda & \rho & 1 \end{pmatrix},$$

$$|\text{var}(x, w_1, w_2)| = (-1)\rho^2 + (2\lambda^2)\rho + (1 - 2\lambda^2),$$

and since $\text{var}(x, w_1, w_2)$ must be positive semidefinite, it follows that $\rho > 2\lambda^2 - 1$. We fix $\lambda = \frac{1}{\sqrt{2}}$ so that the lower bound for ρ is 0. The choice of λ does not affect the relative efficiency of the estimators.

We consider five estimators. The *full data* estimator is an infeasible estimator that uses both instruments, even when they are not observed. The *complete case* estimator discards observations for which either one of the instruments is not observed. The *available case* estimator replaces missing instruments by zeros. This amount to estimating each of the moment functions using all the observations for which that moment function is observed. The *complete moment* estimator uses only one instrument. Finally, the *optimal* estimator is the estimator proposed in Section 3.

In Figure 6.1 we plot the asymptotic variance of our estimators as a function of ρ for $p_1 = 0.5$. The key aspect of this example is that the two instruments provide similar sources of information. Therefore, as ρ (the correlation between w_1 and w_2) increases, we expect two effects. First, the total amount of information for θ_0 decreases, so we expect the variance of all estimators to increase. Second, the amount of information on the instrument that is missing increases. Since the optimal estimator is constructed such that it efficiently

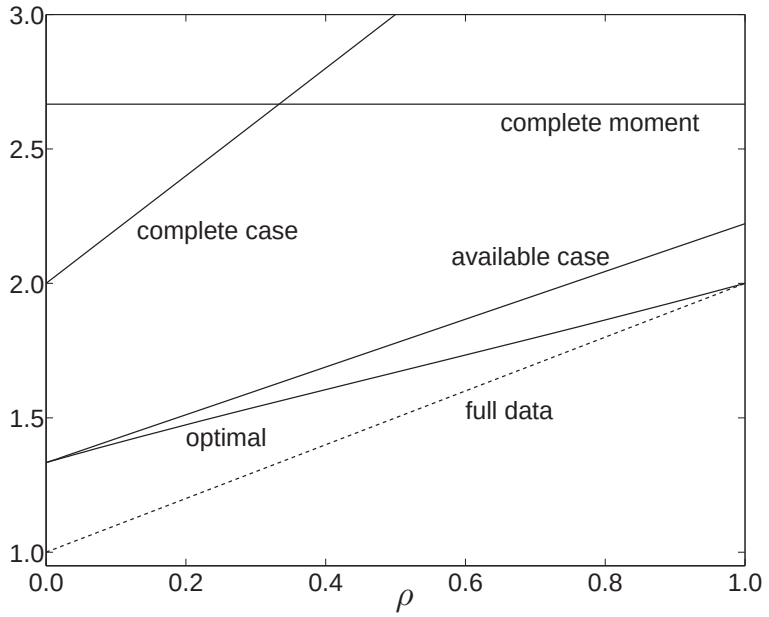

Figure 6.1: Asymptotic variance for various estimators of β as a function of ρ, $p_1 = 0.5$.

exploits the correlation between the components of the moment conditions, we expect the relative performance of the optimal estimator to increase.

The optimal estimator is efficient among the feasible estimators. Except for $\rho = 0$, it outperforms the available case estimator. As ρ increases, the relative performance of the optimal estimator with respect to the available estimator increases: the available case estimator uses all the available data but does not efficiently use the correlation between the instruments. As ρ approaches 1, the variance of the optimal missing data estimator approach that of the infeasible estimator. The complete case and complete moment estimators are always outperformed by the available case estimator and the optimal sample mean.

6.2 Dynamic panel data

The goal of this setting is to demonstrate the performance of our method in a more complicated model and to provide an example where the variance matrix is not known. In particular, we look at a dynamic panel data model, and use

continuous updating GMM to estimate the parameters of the model. The parameter of interest ρ describes the relationship between current and lagged values of a random variable $y_{i,t}$:

$$y_{i,t} = \alpha_i + \rho y_{i,t-1} + u_{i,t},\ 2 \leq t \leq T.$$

We normalize $\mathbb{E}(\alpha_i) = 0,\ \text{var}(\alpha_i) = \sigma_a^2$, and $\mathbb{E}(u_{it}) = 0,\ \text{var}(u_{it}) = 1$. Furthermore, we assume no autocorrelation in the error terms: $\mathbb{E}(u_{i,t}u_{i,s}) = 0$ whenever $s \neq t$. Arellano and Bond (1991) propose an estimator that is widely used: the optimal GMM estimator based on the $(T-2)(T-1)/2$ moment conditions $\mathbb{E}(y_{i,t-s}\Delta\epsilon_{i,t}) = 0,\ t \geq 3, s \geq 2$.

For any observation i, if $y_{i,t}$ is not observed, then several components of the moment function are not observed. For an example with $T = 5$, see Table 1 in the introduction. For the purposes of this simulation, we consider the case $T = 9$, which corresponds to the example in Blundell and Bond (1998). This gives 28 moment conditions for 1 parameter. If any of the $y_{i,t}$ are missing, the moment function is incompletely observed: if $y_{i,1}$ is not observed, 7 components of the moment function are not observed; if $y_{i,4}$ is not observed, 12 components of the moment function are not observed.

We perform a Monte Carlo analysis to compare the relative performance of the estimator introduced in this paper to the full data, complete case, and available case estimators. We use a continuous updating version of the Arellano-Bond estimator to estimate ρ. When estimating the variance matrix, we assume that $\Omega_j = \Omega$ for each j.

We consider different values for the variance of the individual effect $\sigma_\alpha^2 \in \{0.1, 1\}$ and the parameter of interest $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. We set the number of observations to $n = 10000$ and perform $s = 1000$ simulations per parameter combination. There are 10 missing data patterns. Patterns $j = 1, \ldots, 9$ have $y_{i,j}$ missing and the other variables observed. Pattern 10 corresponds to the subsample with all variables observed. The severity of the missing data problem is determined by a parameter $p = \mathbb{P}(R = S_j)$. In the population, a proportion $\mathbb{P}(R = S_{10}) = 1 - 9p$ of the observations are complete. We con-

σ_α^2	ρ	p	cc	ac	opt
0.1	0.1	0.02	1.19	1.12	1.08
		0.06	2.29	1.46	1.41
	0.2	0.02	1.29	1.23	1.18
		0.06	2.37	1.34	1.27
	0.5	0.02	1.82	1.77	1.69
		0.06	3.35	2.50	2.25
	0.8	0.02	8.61	8.11	7.74
		0.06	15.95	11.76	10.45
1	0.1	0.02	1.71	1.47	1.46
		0.06	3.04	1.89	1.84
	0.2	0.02	1.91	1.70	1.68
		0.06	3.75	2.35	2.21
	0.5	0.02	5.10	4.75	4.59
		0.06	8.61	5.85	5.33
	0.8	0.02	2.04	2.20	1.92
		0.06	3.47	3.30	2.62

Table 2: Simulation results for a continuous updating version of the Arellano-Bond estimator in a dynamic panel data context with $T = 9$ time periods. Reported results are variances relative to the full data estimator, for the (i) complete case estimator (cc); (ii) available case estimator (ac); and (iii) optimal (opt) estimator. Results of a Monte Carlo study with $n = 1000$ observations, and $s = 1000$ replications. of a continuous updating Arellano-Bond estimator. Details of the simulation design are in Section 6.2.

sider $p \in \{0.02, 0.06\}$ so that 82% (respectively 46%) of the observations are complete.

Table 2 reports the variance of the complete case, available case, and optimal estimator divided by the variance of the full-data estimator. The complete case estimator is always outperformed by the available-case estimator, except for $(\sigma_\alpha^2, \rho, p) = (1, 0.8, 0.02)$. The optimal estimator always outperforms the other two estimators, which provides evidence in favor of the missing data GMM estimator proposed in this paper that complements the asymptotic efficiency result in Section 3. The optimal estimator seems to gain more when p

is larger. For some parameter configurations, the efficiency gain is substantial.

7 Conclusion

This paper considered efficient GMM estimation from a random sample of complete and incomplete observations. The semiparametric efficiency bound is derived for a model with an assumption that is weaker than missing completely at random. An efficient estimator is obtained by assigning observations to subsamples on the basis of their missing data pattern, and by optimally combining the information in these subsamples. This approach allows us to extend the estimator to a setting where selection is on unobservables. Examples demonstrated the flexibility of the approach and the efficiency gains that can be obtained over standard approaches.

Some aspects of the paper could be further investigated. First, the framework that we constructed to deal with a general missing data pattern suggests some tests for sample selection. In particular, if the parameter is identifiable in each subsample, a test of equality of the subsample estimators can be used to detect sample selection. Second, the mathematical result underlying Section 4 may be of independent interest. We will explore extensions and further applications in future work.

A Proofs

Proof. [Theorem 1] Theorem 1 establishes the consistency and asymptotic normality of $\hat{\theta}_n$. First, I show that $\hat{\theta}_n$ can be regarded as a GMM estimator, and what the corresponding moment functions are. Second, I verify that Assumptions 1 and 2 imply the conditions in Newey and McFadden (1994), Theorems 2.6 and 3.4.

Note that

$$\begin{aligned}\hat{\theta}_n &= \text{argmin}_{\theta\in\Theta}\sum_{j=1}^{J} h_{n,j}(\theta)' W_{n,j} h_{n,j}(\theta). \\ &= \text{argmin}_{\theta\in\Theta}\begin{bmatrix} h_{n,1}(\theta) & \cdots & h_{n,J}(\theta)\end{bmatrix}\begin{bmatrix} W_{n,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & W_{n,J}\end{bmatrix}\begin{bmatrix} h_{n,1}(\theta) \\ \vdots \\ h_{n,J}(\theta)\end{bmatrix} \quad \text{(A.1)}\end{aligned}$$

The probability limit for each $h_{n,j}(\theta) = \frac{1}{n_j}\sum_{i\in G_j} R_i h(X_i,\theta)$, as $n_j \sim n \to \infty$, is

$$\begin{aligned}\text{plim}\,\frac{1}{n_j}\sum_{i\in G_j} R_i h(X_i,\theta) &= \text{plim}\,\frac{1}{n}\frac{n}{n_j}\sum_{i\in G_j} 1_{\{R_i=S_j\}} R_i h(X_i,\theta) \\ &= \mathbb{E}\left[\frac{1}{p_j} 1_{\{R_i=S_j\}} R_i h(X_i,\theta)\right] \\ &= \frac{1}{p_j} S_j \mathbb{E}\left[1_{\{R_i=S_j\}} h(X_i,\theta)\right].\end{aligned}$$

Consider the function

$$\begin{aligned}k(X_i,R_i,\theta) &= \begin{bmatrix} \frac{1}{p_1} 1_{\{R_i=S_1\}} S_1 \\ \vdots \\ \frac{1}{p_J} 1_{\{R_i=S_J\}} S_J\end{bmatrix} h(X_i,\theta). \\ &= a(R_i)\, h(X_i,\theta)\end{aligned}$$

suppressing the dependence of $a(\cdot)$ on $(p_1,\cdots,p_J)$. Then, $\hat{\theta}_n$ is a GMM estimator based on the unconditional moment conditions

$$E\left[k(X_i,\theta_0)\right] = 0,$$

using a block-diagonal weight matrix. All that remains is to check the conditions in Theorems 2.6 and 3.4 in NM94.

Consistency. *Identification* (NM94, 2.6, (i)) in this setting corresponds

to

$$E\left[k\left(X_i, R_i, \theta\right)\right] = 0 \Leftrightarrow \theta = \theta_0.$$

$\Leftarrow$: Using Assumption 1(ii), we have

$$\begin{aligned} E\left[k\left(X_i, R_i, \theta\right)\right] &= E\left[a\left(R_i\right) h\left(X_i, \theta\right)\right] \\ &= E\left[a\left(R_i\right) E\left[h\left(X_i, \theta\right) \middle| R_i\right]\right] \\ &= E\left[a\left(R_i\right) 0\right] = 0. \end{aligned}$$

$\Rightarrow$: Consider a $\theta \neq \theta_0$. By Assumption 1(iii), there exists a j such that $\mathbb{E}\left(R_i h\left(X_i, \theta\right) \middle| R_i = S_j\right) \neq 0$. Consider the expectation of the j-th block of $k(X_i, R_i, \theta)$, conditional on $R_i = S_j$:

$$\mathbb{E}\left[\frac{1}{p_j} 1_{\{R_i = S_j\}} S_j h\left(X_i, \theta\right) \middle| R_i = S_j\right] = \frac{1}{p_j} \mathbb{E}\left[R_i h\left(X_i, \theta\right) \middle| R_i = S_j\right] \neq 0$$

where the inequality follows from 1(iii). A non-zero block of k implies a non-zero k.

Continuity of

$$k\left(X_i, \theta\right) = a\left(R_i\right) h\left(X_i, \theta\right)$$

follows immediately from the continuity of h (Assumption 2(ii)) and the observation that $a\left(\cdot\right)$ does not depend on θ. *Compactness* is imposed by Assumption 2(v).

Assumption 2(vi) requires that $h\left(\cdot\right)$ is *uniformly bounded* in absolute mean. Multiplication by $a(R)$ to obtain $k\left(\cdot\right)$ preserves this boundedness, because $p_j > 0$ by Assumption 1(v). Continuity, compactness, and uniform convergence are (NM94, 2.6, (ii)-(iv)). Therefore, conditions (i)-(iv) of NM94, Theorem 2.6 are satisfied, and hence $\hat{\theta}_n \to \theta_0$.

Asymptotic normality. Similar to the proof of consistency, this proof is straightforward because we have established that $\hat{\theta}_n$ is the GMM estimator based on $E\left[k\left(X_i, R_i, \theta_0\right)\right] = 0$. Given that

$$k\left(X_i, R_i, \theta\right) = a\left(R_i\right) h\left(X_i, \theta\right),$$

it is straightforward to show that conditions imposed in Assumption 2 imply the conditions of NM94, Theorem 3.4. A discussion is available upon request. Here, I derive the expression for the asymptotic variance. From NM94, Th. 3.4, the asymptotic variance of a GMM estimator is

$$\left(G'WG\right)^{-1}\left(G'W\Omega WG\right)\left(G'WG\right)^{-1}$$

where G is the expected derivative of the moment function evaluated at the true value θ_0, and Ω is the variance of the moment function at the true value θ_0. To arrive at the expression for the asymptotic variance in the text, it is useful to determine the derivative and variance for each block. Consider the expected derivative of the j-th block of the moment function,

$$k_j(\theta) = \frac{1}{p_j} 1_{\{R_i=S_j\}} R_i h(X_i, \theta),$$

evaluated at θ_0:

$$\begin{aligned}
\mathbb{E}\left[\left.\frac{\partial\left[\frac{1}{p_j}1_{\{R_i=S_j\}}R_i h(X_i,\theta)\right]}{\partial\theta}\right|_{\theta=\theta_0}\right] &= \mathbb{E}\left[\left.\frac{1}{p_j}1_{\{R_i=S_j\}}R_i\,\frac{\partial h(X_i,\theta)}{\partial\theta}\right|_{\theta=\theta_0}\right] \\
&= p_j\frac{1}{p_j}S_j\mathbb{E}\left[\left.\frac{\partial h(X_i,\theta_0)}{\partial\theta}\right| R_i = S_j\right] \\
&= S_j D_j,
\end{aligned}$$

and consider its variance:

$$\begin{aligned}
\mathbb{E}\left[k_j k_j'\right] &= \mathbb{E}\left[\frac{1}{p_j^2}1_{\{R_i=S_j\}}R_i h(X_i,\theta_0)\, h(X_i,\theta_0)' R_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left.\frac{1}{p_j^2}1_{\{R_i=S_j\}}R_i h(X_i,\theta_0)\, h(X_i,\theta_0)' R_i\right| R_i\right]\right] \\
&= \mathbb{E}\left[\frac{1}{p_j^2}\mathbb{E}\left[\left.1_{\{R_i=S_j\}}R_i h(X_i,\theta_0)\, h(X_i,\theta_0)' R_i\right| R_i\right]\right] \\
&= \frac{1}{p_j}S_j\Omega_j S_j.
\end{aligned}$$

In our setup, the weight matrix W is block-diagonal (see the expression in (A.1)) and so

$$\begin{aligned} G'WG &= \begin{bmatrix} D_1'S_1 & \cdots & D_J'S_J \end{bmatrix} \begin{bmatrix} W_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & W_J \end{bmatrix} \begin{bmatrix} S_1D_1 \\ \vdots \\ S_JD_J \end{bmatrix} \\ &= \sum_{j=1}^{J} D_j'S_jW_jS_jD_j \\ &= \sum_{j=1}^{J} D_j'W_jD_j \end{aligned}$$

where the last equality follows from Assumption 2(vii). Similarly,

$$\begin{aligned} \left(G'W\Omega WG\right) &= \sum_{j=1}^{J} \frac{1}{p_j} D_j'S_jW_jS_j\Omega_jS_jW_jS_jD_j \\ &= \sum_{j=1}^{J} \frac{1}{p_j} D_j'W_j\Omega_jW_jD_j. \end{aligned}$$

It follows that

$$Avar\left(\hat{\theta}_n\right) = \left(\sum_{j=1}^{J} D_j'W_jD_j\right)^{-1} \left(\sum_{j=1}^{J} \frac{1}{p_j} D_j'W_j\Omega_jW_jD_j\right) \left(\sum_{j=1}^{J} D_j'W_jD_j\right)^{-1},$$

where the invertibility of $\left(\sum_{j=1}^{J} D_j'W_jD_j\right)$ follows from Assumption 1(iv) and Assumptions 2(iv) and 2(vii). □

Proof. [Theorem 2]

Each observation provides two random objects that can be used for estimation: a missing data indicator, R_i, and the observable elements of the moment function, $R_ih(X_i, \theta)$. One set of moment conditions that restricts the joint distribution of these two objects comes from Assumption 1(ii):

$$\mathbb{E}(\, R_ih(X_i, \theta_0)|\, R_i)) = 0,$$

which provides us with information on θ_0. A second set of population parameters is $(p_1, \cdots, p_J)$. We can summarize the information on these parameters by a set of moment conditions for the marginal distribution of R_i:

$$\mathbb{E}(R_i) = \sum_{j=1}^{J} p_j S_j.$$

Note: Under the typical MCAR assumption, we would have more information about the distribution of R_i conditional on X_i, which we can exploit as additional moment conditions. For a discussion in the binary missing data case, see Graham (2010).

Denote

$$\mathbb{E}(\,\psi_1(R_i, X_i; \theta_0)|\, R_i) \equiv \mathbb{E}(\,R_i h(X_i, \theta_0)|\, R_i))$$

and $\mathbb{E}(\rho_2(R_i; p)) \equiv \mathbb{E}(R_i - \sum_{j=1}^{J} p_j S_j)$, where $p = (p_1, \ldots, p_J)$. Since R_i has finite support, there exists a function $M(\cdot)$ such that the unconditional moment restrictions

$$\mathbb{E}(M(R_i)\psi_1(R_i, X_i; \theta_0)) = \mathbb{E}(\rho_1(R_i, X_i; \theta_0)) = 0$$

contain the same information as the conditional moment restrictions $\mathbb{E}(\,\psi_1(R_i, X_i; \theta_0)|\, R_i) = 0$.

Consider the complete set of parameters $\beta_0 = (\theta_0, p)$. The asymptotic efficiency bound for β_0 based on the unconditional moment restrictions

$$\mathbb{E}(\rho(R_i, X_i; \beta_0)) = \begin{pmatrix} \rho_1(R_i, X_i; \theta_0) \\ \rho_2(R_i; p) \end{pmatrix} = 0$$

is $\Lambda_0 = \left(D_0' \Sigma_0^{-1} D_0\right)^{-1}$, where $D_0 = \mathbb{E}\left(\frac{\partial \rho(R, X; \beta_0)}{\partial(\theta)}\right)$ and $\Sigma_0 = \mathbb{E}\left(\rho(R, X; \beta_0)\rho'(R, X; \beta_0)\right)$, following Chamberlain (1987). D_0 can be partitioned as $D_0 = \begin{pmatrix} \mathbb{E}(\frac{\partial \rho_1(\beta_0)}{\partial \theta}) & 0 \\ 0 & \mathbb{E}(\frac{\partial \rho_2(\beta_0)}{\partial p}) \end{pmatrix}$. The off-diagonal blocks of D_0 are zero, since θ_0 only features in ρ_1 and p only features in ρ_2. Therefore, the bound for θ_0 under $\mathbb{E}(\rho_1) = 0$ equals the bound for θ_0 under $\mathbb{E}(\rho) = 0$, and we conclude that ρ_2 is not informative for θ_0.

We can find the semiparametric efficiency bound for θ_0 given the conditional moment conditions

$$\mathbb{E}(\, R_i h(X_i, \theta_0) |\, R_i) = \mathbb{E}(\, \rho(R_i, X_i, \theta_0) |\, R_i) = 0$$

by applying the result in Newey (2001, Theorem 5.2) that extends the well-known result in Chamberlain (1987). Let $D_\rho(R_i) = \frac{\partial \mathbb{E}(\rho(X_i, R_i, \theta_0)|R_i)}{\partial \theta}$ and

$$\Sigma_\rho(R_i) = \mathbb{E}(\, \rho(X_i, R_i, \theta_0)\rho(X_i, R_i, \theta_0)' |\, R_i).$$

The semiparametric efficiency bound is equal to

$$\text{SPEB}(\theta_0) = \left(\mathbb{E}\left(D_\rho(R_i)'\Sigma_\rho(R_i)^+ D_\rho(R_i)\right)\right)^{-1},$$

It is easily verified that, in our case,

$$D_\rho(S_j) = S_j D_j = S_j \mathbb{E}\left(\frac{\partial h(X_i, \theta_0)}{\partial \theta} \middle| R_i = S_j \right)$$

and $\Sigma_\rho(S_j) = S_j \Omega_j S_j$. Then,

$$\begin{aligned} \text{SPEB}(\theta_0) &= \left(\sum_{j=1}^{J} p_j D_j' S_j (S_j \Omega_j S_j)^+ S_j D_j \right)^{-1} \\ &= \left(\sum_{j=1}^{J} p_j D_j' (S_j \Omega_j S_j)^+ D_j \right)^{-1}. \end{aligned}$$

□

Proof. [Theorem 3.] The result is closely related to Abadir and Magnus (2005, Exercise 12.18): "For any two matrices K_1 and K_2satisfying $K_1^{'} K_2 = I$, we have that

$$K_1^{'} K_1 - \left(K_2^{'} K_2 \right)^{-1}$$

is positive semidefinite."

Our result follows from choosing K_1 and K_2 appropriately. First, let

$$K_1' = \begin{bmatrix} A_1\Gamma_1^{-1/2} & \cdots & A_J\Gamma_J^{-(1/2)} \end{bmatrix}$$

where (1/2) refers to the Cholesky decomposition. Also define

$$K_2' = \begin{bmatrix} \Gamma_1^{1/2} & \cdots & \Gamma_J^{1/2} \end{bmatrix}.$$

Then

$$\begin{aligned} K_1'K_1 &= \sum_{j=1}^{J} A_j\Gamma_j^{-1}A_j', \\ (K_2'K_2)^{-1} &= \left(\sum_{j=1}^{J}\Gamma_j\right)^{-1}, \\ K_1'K_2 &= \sum_{j=1}^{J} A_j\Gamma_j^{-1/2}\Gamma_j^{1/2} \\ &= I, \end{aligned}$$

and the result on the lower bound follows. The optimal weights can easily be verified. □

References

Abadir, K. and J. R. Magnus (2005): *Matrix Algebra*, Cambridge University Press.

Abowd, J. M., B. Crépon, and F. Kramarz (2001): "Moment estimation with attrition: an application to economic models," *Journal of the American Statistical Association*, 96, 1223–1231.

Abrevaya, J. and S. G. Donald (2010): "A GMM approach for dealing with missing data on regressors and instruments," Manuscript, March 2010.

Angrist, J., V. Lavy, and A. Schlosser (2006): "Multiple experiments

for the causal link between the quantity and quality of children," *MIT Department of Economics Working Paper Series No. 06-26*.

Arellano, M. and S. Bond (1991): "Some tests of specification for panel data: monte carlo evidence and an application to employment equations," *The Review of Economic Studies*, 58, 277.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993): *Efficient and adaptive estimation for semiparametric models*, Baltimore: Johns Hopkins University Press.

Blundell, R. and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 87, 115–143.

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

Chaudhuri, S. (2012): "GMM with Missing Moment Restrictions," *Mimeo*.

Chen, B., G. Yi, and R. Cook (2010): "Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random," *Journal of the American Statistical Association*, 105, 336–353.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics*, 36, 808–843.

Dardanoni, V., S. Modica, and F. Peracchi (2009): "Regression with imputed covariates: a generalized missing indicator approach," CEIS Research Paper 150, Tor Vergata University, CEIS.

Graham, B. (2010): "Efficiency bounds for missing data models with semiparametric restrictions," *Econometrica*, Forthcoming.

Graham, B., C. de Xavier Pinto, and D. Egel (2010): "Inverse probability tilting for moment condition models with missing data," Manuscript, August 2010.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Levitt, S. D. (2002): "Using electoral cycles in police hiring to estimate the effects of police on crime: reply," *The American Economic Review*, 92, pp. 1244–1250.

Little, R. J. A. and D. B. Rubin (2002): *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, New York: Wiley, 2nd ed.

Mogstad, M. and M. Wiswall (2010): "Instrumental variables estimation with partially missing instruments," Manuscript, May 2010.

Newey, W. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (2001): "Conditional moment restrictions in censored and truncated regression models," *Econometric Theory*, 17, 863–888.

Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by D. McFadden and R. Engle, Elsevier, vol. 4, 2111–2245.

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

Rodrik, D., A. Subramanian, and F. Trebbi (2004): "Institutions rule: the primacy of institutions over geography and integration in economic development," *Journal of Economic Growth*, 9, 131–165.

Tsiatis, A. (2006): *Semiparametric theory and missing data*, Springer Verlag.

Van der Vaart, A. (2000): *Asymptotic statistics*, Cambridge University Press.

Verbeek, M. and T. Nijman (1992): "Testing for selectivity bias in panel data models," *International Economic Review*, 33, 681–703.

Wang, Q., O. Linton, and W. Härdle (2004): "Semiparametric regression analysis with missing response at random," *Journal of the American Statistical Association*, 99, 334–345.

Wooldridge, J. (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141, 1281–1301.