EFFICIENT GMM ESTIMATION WITH A GENERAL MISSING DATA PATTERN

CHRIS MURIS

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. We propose an estimator for an arbitrary set of regular 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. It can be interpreted as the optimal linear combination of subsample GMM estimators. Because of this linearity, the computational burden and the small-sample performance of the estimator are comparable to the full-data estimator. We also propose an inverse probability weighted version of the estimator that is consistent when selection is on observables. Applications to multivariate mean estimation, instrumental variable estimation, and dynamic panel data estimation demonstrate the efficiency gain with respect to existing missing data methods. We also discuss how the results can be used to optimize data collection for measuring consumer confidence.

1. Introduction

Missing data affect the majority of empirical studies in economics. In a survey of empirical research in top economics journals, Abrevaya and Donald (2010) find that missing data occurs in 40% of the publications. In 70% of these cases, a complete-case estimator is used. A complete-case estimator discards all incomplete observations. This is inefficient if the incomplete observation contain information about the parameter of interest.

The main contribution of this paper is to introduce an estimation procedure that efficiently combines information from complete and incomplete observations. Interest is in GMM estimation of a finite-dimensional parameter with a random sample. Our procedure can be applied to two-step, iterative and continuous updating GMM estimators. We do not impose restrictions on the missing data *pattern*, which means that the data can be incomplete in an arbitrary way. In terms of the missing data *mechanism*, we assume that there is no selection or that selection is on observables.

In many econometric models, observations that are incomplete can still be informative. To see this, consider instrumental variables estimation and dynamic panel data models. First, a linear instrumental variable model with one endogenous variable X and two instruments $Z = (Z_1, Z_2)$ is given by the equation $y = X\beta_0 + \epsilon$ and the conditional mean assumption $\mathbb{E}(\epsilon|\, Z) = 0$. Estimation of the parameter β_0 is based on the unconditional moment condition $\mathbb{E}(Z\epsilon) = 0$. 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 $\mathbb{E}(Z_1\epsilon) = 0$. Missing instruments are common in empirical research, see for example Levitt (2002), who uses the number of firefighters and the number of city workers as instruments to estimate the effect of police on crime. Not all cities provide information about the number of firefighters in each year, and data on the number of city workers is available for yet another subsample. 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 that are sometimes unobserved.

Key words and phrases. missing data, GMM, efficiency.

I am grateful to Ramon van den Akker, Richard Blundell, Otilia Boldea, Pedro Duarte Bom, Katherine Carman, Miguel Atanasio Carvalho, Toru Kita'gawa, Tobias Klein, Andrea Krajina, Jan Magnus, Bertrand Melenberg, Franco Peracchi, Pedro Santos Raposo, Nathanael Vellekoop, and Bas Werker for encouraging and insightful discussions.

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

	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	·	·

Dynamic panel data models provide a second example of incomplete, informative observations. Interest is in the autoregressive parameter ρ in

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

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

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

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 can efficiently accommodate static and dynamic panel data models with unbalanced panels with different starting points, endpoints, and any combination of gaps.

Standard approaches that, in contrast to the complete-case estimator, use all available information can still be inefficient. One such approach is the available-case estimator, which replaces missing moments by zeros before applying the full data estimation procedure. The available-case estimator is consistent if there is no selection. In the instrument example, available-case estimation corresponds to replacing the missing instruments by zero. For the dynamic panel data example, it corresponds to the procedure suggested in Arellano and Bond (1991, p. 281).

The key to efficient estimation is to split the random sample in subsamples based on the missing data pattern. If two instruments are available, we can 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. In the absence of selection, β_0 can be estimated using each subsample. Using efficient GMM in each subsample yields three consistent estimators of β_0. Any weighted average of these estimators is again a consistent estimator of β_0. The complete-case estimator assigns full weight to the estimator from the first subsample. The available-case estimator assigns equal weight to each estimator. We show that there exist optimal weights that minimize the asymptotic variance of the estimator.

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 fully independent of the missing data indicator, and we only require that the moment condition holds conditional on the missing data indicator. Under this assumption, the estimator is asymptotically efficient in the sense that it attains the semiparametric efficiency bound. Furthermore, the computational and small sample properties are close to those of the full data estimator, since the minimization problem for the missing data estimator is linear in full data problems.

After introducing notation in Section 2, we show in Section 3 that the procedure using subsamples can be generalized to parameter vectors and that the parameter does not need to be identifiable in each subsample. Section 4 considers the special case where the parameter is identified in each

subsample, as discussed above. In Section 5, we show that our estimation procedure can be extended to a generalized inverse probability weighting estimator in order to deal with selection on observables. Again, we will work under an assumption that is weaker than the typical missing at random (MAR) assumption. Section 6 gives some examples, and show that substantial efficiency gains over standard approaches are possible. Section 7 concludes. Proofs can be found in the Appendix A.

This paper is not concerned with univariate regression methods. 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 at the cost of unbiasedness, see Dardanoni et al. (2009).

More generally, we are not concerned with situations in which each observation contributes either to all, or to none of the sample moments. For this case there is a vast literature that addresses efficient and robust estimation under MAR. This literature was initiated by Robins et al. (1994) and is still 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).

Finally, some 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.

2. Sample moments for missing data

We introduce notation for general missing data patterns, and discuss how a missing data pattern for X implies which subset of the components of a moment function $h(X, \theta)$ can be evaluated. We introduce an assumption about the missing data mechanism, MI, which is a mean independence version of missing completely at random. MI is a necessary and sufficient condition for the complete- and available-case methods. In Section 3, we consider estimation under MI for data with a general missing data pattern.

2.1. Missing data patterns in GMM estimation.

There are 2^d ways in which the components of a random vector $X \in \mathbb{R}^d$ can be missing, since each component is either missing or not. For a given model, the number of possible patterns is J_x, which can be smaller than 2^d when some patterns are ruled out by design. We use a diagonal selection matrix $S^x \in \mathbb{R}^{d \times d}$ to describe a missing data pattern. Such a matrix has kth diagonal entry equal to 1 if and only if the kth component of X is observed, that is:

$$(S^x)_{k_1,k_2} = \begin{cases} 1 & \text{if } k_1 = k_2 \text{ and component } k_1 \text{ is observed for pattern } j, \\ 0 & \text{otherwise.} \end{cases}$$

The J_x diagonal selection matrices $S_j^x,\ j = 1, \ldots, J_x$, describe the missing data patterns. The missing data indicator $R^x \in \mathbb{R}^{d \times d}$ is a random matrix that captures which components of X are missing and takes values $S_j^x,\ 1 \leq j \leq J_x$.

In GMM estimation, a parameter of interest $\theta_0 \in \Theta \subset \mathbb{R}^p$ is defined through the moment conditions $\mathbb{E}(h(X, \theta_0)) = 0$, with moment function $h : \mathbb{R}^d \times \Theta \to \mathbb{R}^q$. If an observation is incomplete, only a subset of the components of the moment function is observable. A missing data pattern represented by S^x implies a missing moment pattern, which we describe by a diagonal selection matrix $S \in \mathbb{R}^{q \times q}$. As such, S describes a missing moment pattern for h in the same way that S^x describes a missing data pattern for X. The number of missing data patterns is greater than or equal to the number of missing moment patterns J, because different values for R^x can imply the same value for R. The missing moment indicator R takes values $S_j,\ 1 \leq j \leq J$. Let $p_j = \mathbb{P}(R = S_j)$ be the probability that missing moment pattern j occurs.

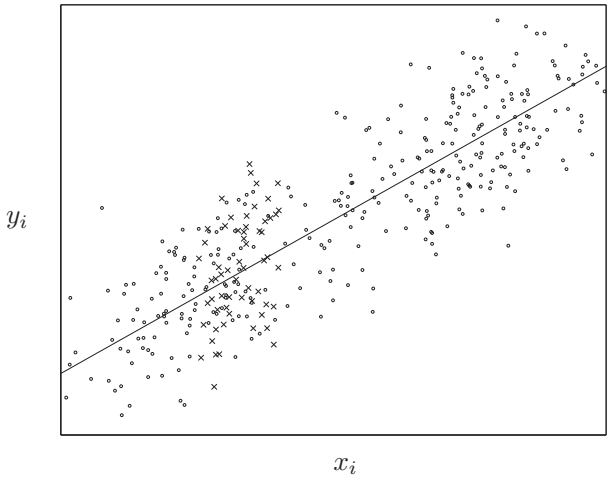

FIGURE 2.1. MI, not MCAR. Simulated data for a univariate regression model. A cross represents a missing observation; a dot represents a complete observation.

Assumption. *[FULL-RANK] The probability of observing pattern j is positive, $p_j > 0$, for each $1 \leq j \leq J$ and* $\operatorname{rk}\left(\sum_{j=1}^{J} S_j\right) = q$.

The restriction of positive probability is not restrictive, since we can eliminate patterns that occur with zero probability. The second restriction ensures that each component of the moment function is observed with positive probability.

2.2. **Missing completely at random.** Typically, three assumptions about the missing data mechanism are distinguished: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). For a detailed discussion of these concepts, see Little and Rubin (2002, Chapter 1). MCAR is the most restrictive assumption. Let $\perp$ denote statistical independence.

Assumption. *[MCAR]* $X \perp R^x$.

Assumption. *[IID1]* $(R_i^x, R_i^x X_i,\ 1 \leq i \leq n)$ *is a random sample of size n from* (R, RX).

Assumption MCAR requires that whether or not a random variable is observed is independent of the realization. For a moment function h, MCAR implies that $h(X, \theta) \perp R$ for each $\theta \in \Theta$ because $h(X, \theta)$ depends on X and not on R^x, while R is determined by R^x. This implies the following MCAR-like mean independence condition:

Assumption. *[MI]* $\mathbb{E}\left(h(X, \theta_0) \mid R = S_j\right) = 0$ *for each* $1 \leq j \leq J$.

This assumption requires the moment conditions to hold regardless of the missing data pattern. To demonstrate the difference between MCAR and MI, 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, $r_i = 0$, and 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 probability of being missing, while an observation in the high group is always complete, so $\mathbb{P}(r = 1 \mid X \text{ low}) \neq \mathbb{P}(r = 1 \mid X \text{ high})$. The data are MI, since $\mathbb{E}(x_i \epsilon_i | r = 1) = \mathbb{E}(x_i \epsilon_i | r = 0) = 0$. If we strengthened MI to include independence of the variance, or mean independence at values of the parameter other than the true value of β, MI would not be satisfied in this example: $\operatorname{var}(x_i \epsilon_i | r_i = 1) > \operatorname{var}(x_i \epsilon_i | r_i = 0)$, and $\mathbb{E}(x_i(y_i - (\beta + 1)x_i | r_i)) = \mathbb{E}(x_i \epsilon_i | r_i) - \mathbb{E}(x_i^2 | r_i) = -\mathbb{E}(x_i^2 | r_i) \neq \mathbb{E}(x_i^2)$.

The complete-case approach and the available-case approach are two popular ways to deal with missing data. Both methods are consistent under MI. The complete-case estimator is common in empirical work and is the default approach for most statistical 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 1. Then, the complete-case sample moment for $\mathbb{E}(h(X,\theta))$ is

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

where G_j is the subsample for which $R_i = S_j$ and n_j is the number of observations in subsample G_j, $1 \le j \le J$. A complete-case GMM estimator is based on the complete-case sample analog.

The available-case approach 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 sample moment is

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

where the inverse of $\hat{\bar{R}} = \sum_{j=1}^{J}(n_j/n)S_j$ is used to divide each component of the sum by the number of observations that actually contribute.

In Section 3 we consider GMM estimation under MI, and we find an estimator that is asymptotically efficient under MI. In Section 5, we consider GMM estimation under a mean independence version of MAR.

3. GMM estimation

We are interested in estimating a parameter θ_0 that is defined through the moment conditions $\mathbb{E}(h(X,\theta_0)) = 0$. Given a complete data set, we would use the optimal GMM estimator. We construct a class of estimators that are consistent under MI. We show that the asymptotic variance of an optimal estimator in this class achieves the semiparametric efficiency bound for θ_0 under MI. The results in this section are a natural generalization of the properties of the optimal full-data GMM estimator to the optimal GMM estimator with a general missing data pattern. In Section 4, we consider a special case where the parameter can be estimated using the observations for an arbitrary pattern only. In Section 5, we allow the missing data indicator to depend on observable random variables. We provide examples of the estimator in this section in Section 6. All the proofs are in Appendix A.

3.1. GMM with missing data.

We are interested in a parameter $\theta_0 \in \Theta \subset \mathbb{R}^p$ that is defined through a moment function $h : \mathbb{R}^d \times \Theta \to \mathbb{R}^q$ for which the following is assumed:

Assumption. *[MI]* $\mathbb{E}\left(h(X,\theta_0) \mid R = S_j\right) = 0$ *for each* $1 \le j \le J$.

Assumption. *[IDENTIFICATION] For any* $\theta \in \Theta$ *for which* $\theta \neq \theta_0$*, there exists at least one pattern* j *for which* $\mathbb{E}\left(h(X,\theta)\right| R = S_j) \neq 0$.

Assumption. *[IID1]* $(R_i^x, R_i^x X_i,\ 1 \le i \le n)$ *is a random sample of size* n *from* (R, RX).

A GMM estimator for θ_0 for complete data is defined as the minimizer over Θ of

$$\left(\sum_{i=1}^{n} h(X_i,\theta)\right)' W(n) \left(\sum_{i=1}^{n} h(X_i,\theta)\right), \tag{3.1}$$

for some arbitrary symmetric positive definite matrix $W(n)$. Since $h(X_i,\theta)$ is not observed for each i, this estimator is not feasible. For completeness, we restate the assumption about the available data and the missing data mechanism.

Let $h_{n,j}(\theta)$ be the sample moment for subsample $G_j = \{i : R_i = S_j\}$,

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

We define a GMM estimator for missing data as the minimizer of the modification of the full-data objective function (3.1),

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

A GMM estimator for missing data minimizes the sum of weighted subsample moments instead of weighted sample moments. Complete-case and available-case estimators can be obtained as special cases. If pattern 1 is the complete-data pattern, $S_1 = I_q$, a complete-case estimator is obtained by setting $W_1(n) = W_{\text{cc},n}$ and $W_j(n) = 0_q,\ j > 1$, where $W_{\text{cc},n}$ can be chosen optimally. The available-case estimator follows from setting $W_j(n) = S_j W_{\text{ac}}(n) S_j$ for each $j = 1, \ldots, J$, where $W_{\text{ac}}(n)$ can be chosen optimally. By construction, our estimator will be at least as efficient as the complete-case and available-case estimators. The examples in Section 6 demonstrate that the efficiency gain is substantial.

The asymptotic distribution of the estimator $\hat{\theta}_{W(n)}$ requires the assumptions stated below.

Assumption. *[FINITE-Ω_j] For each j, $\operatorname{var}(h(X,\theta_0) \mid R = S_j) = \Omega_j < \infty$, where $1 \le j \le J$.*

The FINITE-Ω_j assumption is not compatible with MCAR because FINITE-Ω_j allows the conditional variance of the moment function to depend on the missing data pattern.

Assumption. *[DERIVATIVE] (i) For each x, the moment function $h(x,\cdot)$ is continuously differentiable on Θ; (ii) for each pattern j let the $q \times p$ matrix $D_j(\theta) = \mathbb{E}(\partial h(X,\theta_0)/\partial\theta' \mid R)$ be uniformly bounded, in the sense that $\sup_{\theta\in\Theta} \|D_j(\theta)\| < \infty$, where $\|D_j\| = \operatorname{tr}(D_j' D_j)^{1/2}$; (iii) for each pattern j, $\operatorname{rk}(D_j) = p$.*

Assumption. *[REGULARITY] (i) The parameter space Θ is compact and θ_0 is in the interior of Θ; (ii) the moment function is bounded in absolute mean:*

$$\sup_{\theta\in\Theta} \mathbb{E}(|h(X,\theta)|) < \infty;$$

(iii) for each subsample, the sequence of GMM weights $(W_j(n),\ n \in \mathbb{N})$ satisfies $S_j W_j(n) S_j = W_j(n)$ and converges to a positive semidefinite matrix, W_j, with $\operatorname{rk}(W_j) = \operatorname{rk}(S_j)$; (iv) the distribution of X conditional on $R = S_j$, represented by density $f_j(x)$, does not depend on θ

All conditions are standard GMM assumptions, except for REGULARITY(iii), which sets the submatrix of W_j that corresponds to $S_j = 0$ equal to zero, and requires the remaining submatrix to be positive definite, and (iv) which is satisfied the parameters of the process generating the data X to be seperate from those that generate the missings R. Compare the assumption of ignorability in Little and Rubin (2002).

Theorem 3.1. *Under assumptions* MI, IDENTIFICATION, IID1, FULL-RANK, FINITE-Ω_j, DERIVATIVE, *and* REGULARITY, *we have that, as $n \to \infty$,*

$$\sqrt{n}\left(\hat{\theta}_{W(n)} - \theta_0\right) \xrightarrow{d} N\left(0, B^{-1}\left(\sum_{j=1}^{J} \frac{1}{p_j} D_j' W_j (S_j \Omega_j S_j) W_j D_j\right) B^{-1}\right),$$

where

$$B = \sum_{j=1}^{J} p_j D_j' (S_j \Omega_j S_j)^{+} D_j. \tag{3.3}$$

Proof. The proof is given in Appendix A. It involves converting the conditional moment restrictions in MIto an augmented set of Jq unconditional moment conditions. The expression in (3.2) can be seen as a weighted sample analog to this set of unconditional moment conditions. The double sum appears because we have a random sample, which implies independent subsamples. Then, we show that this is a standard GMM situation. □

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 $S_1 = I_q$. The

estimator that uses weighting matrices $W^*(n) = (W_1^*(n), \ldots, W_J^*(n))$ is denoted $\hat{\theta}_n^*$ and has limiting distribution

$$\sqrt{n}(\hat{\theta}_{W^*(n)} - \theta_0) \xrightarrow{d} N(0, B^{-1}). \tag{3.4}$$

This is an extension of the familiar result on optimal GMM: the weighting matrix for each subsample moment is proportional to the inverse of the relevant part of the variance matrix.

Remark 1. The conditional moment assumptions MIcan be viewed as a restriction on the conditional densities $f_j(x)$. Starting from a situation with no missings and $\mathbb{E}(h(X, \theta_0)) = 0$, we only allow conditional densities $f_j(x)$ that imply that the conditional expectation in the subpopulation is 0 when the parameter is equal to the true value that applies to the population. Here, we do not make explicit which conditional densities allow for. In general the set of compatible conditional densities will depend on h.

Remark 2. It is possible to formulate identification conditions that are sufficient but not necessary for IDENTIFICATION. A useful condition is that identification in one subsample implies IDENTIFICATION. If there exists a subsample j such that $\mathbb{E}(Rh(X, \theta)|R = S_j) = 0 \Leftrightarrow \theta = \theta_0)$ then IDENTIFICATIONis satisfied.

Remark 3. Replacing the variance matrices Ω_j and the derivative matrices D_j by consistent estimators leaves the asymptotic distribution of $\hat{\theta}_n^*$ unchanged.

Remark 4. The GMM estimator based on the modified objective function is computationally slightly more expensive than the full-data sample moment. The only additional computational burden comes from determining J, rather than 1, optimal matrix weights, for which an analytical expression is available, and sorting the n observations into J groups.

3.2. **Semiparametric efficiency bound.** The model defined by MI and IID1 is a semiparametric model: we are estimating a finite-dimensional parameter θ_0 and consider the infinite-dimensional η that describes the distribution of the data to be a nuisance parameter. 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 many econometric models with a random sample, we can use the methods for calculating the SPEB proposed in Newey (1990) and Severini and Tripathi (2001). For the following theorem, the result for conditional moment restrictions for singular covariance matrices in Newey (2001) that extends a result in Chamberlain (1987) is important. The result shows that the optimal GMM estimator $\hat{\theta}_{W^*(n)}$ is asymptotically efficient for θ_0 among all regular semiparametric estimators.

Theorem 3.2. *Under assumptions* MI, IID1, FULL-RANK, *and* FINITE-Ω_j, *the semiparametric efficiency bound for* θ_0 *is* $SPEB(\theta_0) = B^{-1}$, *where* B *is as in* (3.3).

Remark 5. For specific examples, it may be reasonable to assume that $\Omega_j = \Omega$ and $D_j = D$ for each j. In that case, the expression for B simplifies to $B = D' \left(\sum_{j=1}^{J} p_j (S_j \Omega S_j)^+ \right) D$. This possibly lowers the SPEB, and our estimator may no longer be efficient.

4. Subsample estimation

In some situations θ_0 can be estimated using each subsample. An example is instrumental variable estimation with, for each pattern, more instruments than endogenous variables. We show that an optimal linear combination of the optimal GMM estimators for each subsample is asymptotically efficient. We study this estimator to gain more intuition for the semiparametric efficiency bound, and because it can be implemented using only the full-data estimation routine. Moreover, this estimator

can be extended, without modification, to generalized empirical likelihood estimation. In Section 5, we generalize this approach to an optimal inverse probability weighting estimator for estimation under an assumption weaker than MI that allows for selection on observables.

Assume that θ_0 can be estimated using each subsample separately. Then the following subsample GMM estimator for θ_0 is defined for each missing data pattern j:

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

where $W_j^*(n)$ converges to the optimal weighting matrix $W_j^* = (S_j\Omega_j S_j)^+$. We look at matrix-weighted sums of these subsample GMM estimators. In particular, we are interested in the matrix weights that minimize the asymptotic variance of the sum. To find these, we need the limiting distribution of the subsample GMM estimators. Assume a standard GMM setting as in Section 3.1. Then, as $n \to \infty$,

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

A matrix-weighted sum is the matrix equivalent of a weighted average. The weights are $p \times p$ matrices that are subsample specific, $(A_j(n),\ n \in \mathbb{N})$. An estimator that is a matrix-weighted sum is characterized by a $J-$tuple $A(n) = (A_1(n), \ldots, A_J(n))$ that collects the matrix weights. We denote the matrix-weighted sum with matrix weights $A(n)$ by $\hat{\theta}_{A(n)}$, and define

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

Assuming $\sum_{j=1}^{J} A_j = I_p$, the estimator is consistent. Since we have assumed a random sample, the subsample GMM estimators are uncorrelated, so that the asymptotic variance of matrix-weighted sum $\hat{\theta}_{A(n)}$ is given by

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

which uses the asymptotic variance of the subsample GMM estimators in (4.1). From the following theorem, we can see that the choice of weight matrix A_j^*,

$$A_j^* = B^{-1}\ p_j D_j'(S_j\Omega_j S_j) D_j,$$

leads to an efficient estimator $\hat{\theta}_n^* = \hat{\theta}_{A^*(n)}$. The asymptotic variance is

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

The theorem below shows that this is a lower bound for the asymptotic variance of any matrix-weighted sum.

Theorem 4.1. *For each $j = 1, \ldots, J$, let A_j be a $p \times p$ matrix such that $\sum_{j=1}^{J} A_j = I_p$. Then*

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

is positive semidefinite.

Therefore, the estimator is the optimal linear combination of the optimal GMM estimators for each subsample. As such, it does not contain any additional nonlinear or nonparametric ingredients, which suggests that the higher-order asymptotic properties and small-sample performance of the efficient estimator under MI are of the same order as those of the full-data optimal GMM estimator.

Remark 6. The discussion in this section suggests the following procedure to obtain an efficient estimator: (1) estimate $B = \sum_{j=1}^{J} p_j D_j'(S_j\Omega_j S_j)^+ D_j$; (2) estimate $A_{j,n}^* = B^{-1}\ p_j D_j' \left(S_j\Omega_j S_j\right)^+ D_j$; (3) set $\hat{\theta}_{A^*(n)} = \sum_{j=1}^{J} A_{j,n}^* \hat{\theta}_{j,n}$.

Remark 7. The results in this section can be used to optimally combine estimators obtained using any estimation method, provided that the data used for different estimators is independent. For example, the results can be applied to generalized empirical likelihood estimation. Another example is a combination of estimators applied to different data sets.

5. Inverse probability weighting

In the previous section we derived an optimal estimator under MI and IID1. For some applications, the MI assumption is too strong. In this section, we introduce a weaker assumption about the missing data mechanism, CMI, that allows the missing data indicator to depend on some observed random variables. We generalize the inverse probability weighting (IPW) estimator to a class of estimators that are consistent under CMI. Then, we use techniques from Sections 3 and 4 to derive the efficient IPW estimator.

5.1. **Missing at random.** For many situations, both MCAR and MI are too strong. A significantly weaker assumption that can be used is missing at random, MAR. Organize the data into two groups, (X, Z), where $X \in \mathbb{R}^d$, $Z \in \mathbb{R}^{d_z}$. The random vector X enters the moment function, but the random vector Z does not; it is a vector of auxiliary variables. The missing data pattern for X is captured by $R^x \in \mathbb{R}^{d\times d}$, a random matrix that takes values $\{S_1^x, \ldots, S_{J_x}^x\}$. The following assumption is a typical version of MAR, although different versions are possible:

Assumption. *[MAR] For each pattern j, $X \perp R^x \mid Z$.*

Assumption. *[IID2] $(R_i^x, R_i^x X_i, Z_i,\ 1 \le i \le n)$ is a random sample of size n from $(R^x, R^x X, Z)$.*

The MAR assumption allows the process that generates the missing data to depend on that data. It requires that there exists an auxiliary random vector Z that is always observed and that removes the dependence between R^x and X. This is a significantly weaker assumption than MCAR, especially when many relevant variables are included in Z.

We will formulate an assumption that relaxes MAR in the way that MI relaxes MCAR. As in the MCAR case, the missing data indicator R^x implies a missing data indicator R that describes which components of $h(X, \theta)$ can be evaluated when $R^x X$ is observed instead of X. There are J such patterns for h, denoted $\{S_1, \ldots, S_J\}$.

Consider pattern j, and let r_j be an indicator function that equals 1 if and only if the missing data follow pattern j. Let $V_j \in \mathbb{R}^{d_j}$ be a random vector that consists of a subset of the components of (X, Z). We assume that there exists a function p_j that determines the probability of observing pattern j : $p_j(V_j) = \mathbb{P}(r_j = 1 \mid V_j)$. Let $V = \cup_{j=1}^J V_j$.

Assumption. *[CMI] (i) $\mathbb{E}\left(h(X, \theta_0) \mid r_j,\ V_j\right) = \mathbb{E}\left(h(X, \theta_0) \mid V_j\right)$; (ii) $p_j(V_j)$ is observed if $r_j = 1$; (iii) $\mathbb{P}(r_j = 1|V) = \mathbb{P}(r_j = 1|V_j)$; (iv) there exists $\delta > 0$ such that $p_j(V_j) \ge \delta$ for each V_j.*

The first assumption captures the essence of MAR, and assumptions (ii)–(iv) are necessary for the construction of an inverse probability weighted estimator in Section 5.2. We are not interested in the function $p_j(V_j)$ and assume that the function is known or can be $\sqrt{n_j}-$consistently estimated, which under CMI is not very restrictive given the results in Hirano et al. (2003). Notice that elements of X can be included in V_j if they are observed whenever $r_j = 1$. Also, missing data indicators $r_k,\ k \ne j$, can be included, provided the resulting p_j obeys CMI (iv).

5.2. **Optimal IPW.** A standard tool for missing data with a binary missing data pattern that satisfies MAR is inverse probability weighting (IPW); see for example Wooldridge (2007). In this section we consider a generalization of IPW estimators to the case of general missing data patterns. The assumption of CMI ensures the consistency of such an IPW estimator. First, note that we can rewrite $R = \sum_{j=1}^J r_j S_j$. If we have a function $h(X, \theta_0)$ for which $\mathbb{E}(h(X, \theta_0)) = 0$ then, in general, $\mathbb{E}(Rh(X, \theta_0)) \ne 0$. Now let $\tilde{R}(V) = \sum_{j=1}^J \frac{r_j}{p_j(V_j)} S_j$. Then

$$\mathbb{E}\left(\tilde{R}(V)\middle| V\right) = \sum_{j=1} \frac{\mathbb{E}(r_j|V_j)}{p_j(V_j)} S_j = \sum_{j=1}^J S_j.$$

and, using iterated expectations, $\mathbb{E}\left(\tilde{R}(V)h(X,\theta_0)\right) = 0$.

This motivates the use of the adjusted subsample moment $\tilde{h}_{n,j}$,

$$\tilde{h}_{n,j} = \frac{1}{n_j}\sum_{i\in G_j}\frac{1}{p_j(V_j)}R_i h(X_i,\theta_0).$$

An IPW version of the complete-case estimator minimizes $\tilde{h}'_{n,1}W^*_{\text{cc}}(n)\tilde{h}_{n,1}$, and an IPW version of the available-case estimator minimizes

$$\left(\sum_{j=1}^{J}\tilde{h}_{n,j}\right)' W^*_{\text{ac}}(n)\left(\sum_{j=1}^{J}\tilde{h}_{n,j}\right),$$

where the respective W^* can be chosen optimally.

This suggests an extension of the method in Section 4. Assume that θ_0 can be estimated using each subsample separately. Furthermore, the assumptions for asymptotic normality of the optimal GMM estimator and CMI hold. Then, the parameter θ_0 is identifiable within subsample G_j. Denote the optimal subsample IPW estimator $\hat{\tilde{\theta}}_{n,j}$:

$$\hat{\tilde{\theta}}_{n,j} = \text{argmin}_{\theta\in\Theta}\,\tilde{h}_{n,j}(\theta)'W^*_j(n)\tilde{h}_{n,j}(\theta), \tag{5.1}$$

with W^* equal to the optimal weighting matrix for this problem. The limiting distribution of $\hat{\tilde{\theta}}_{n,j}$ is that of a standard GMM estimator: as $n_j \to \infty$,

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

We do not impose any structure on Λ_j, since we have not specified whether the function is known, or whether a parametric or nonparametric estimator was used.

Analogously to Section 4, we introduce the class of estimators

$$\hat{\tilde{\theta}}_{A(n)} = \sum_{j=1}^{J} A_j(n)\hat{\theta}_{n,j}, \tag{5.2}$$

for any $J-$tuple of $p\times p$ matrices $A(n) = (A_1(n),\ldots,A_J(n))$ that satisfies $\sum_{j=1}^{J}A_j(n) = I_p$. For each sequence $A(n)$ that converges to some A, the asymptotic variance is given by

$$\lim_{n\to\infty}\text{var}(\sqrt{n}\hat{\tilde{\theta}}_{A(n)}) = \sum_{j=1}^{J}\frac{1}{p_j}A_j\Gamma_j A_j'.$$

A straightforward modification of Theorem 4.1 shows that the lower bound on the asymptotic variance for any estimator in the class of matrix-weighted sums is given by

$$\tilde{B}^{-1} = \left(\sum_{j=1}^{J}p_j\Gamma_j\right)^{-1}.$$

Setting $A^*_j = \tilde{B}^{-1}\,p_j\Gamma_j$ achieves that bound.

6. Examples

This section contains four examples that illustrate the methods in this paper and demonstrate the efficiency gains with respect to a complete-case and an available-case analysis. The first example concerns a multivariate mean estimation problem that corresponds to a two-period panel data model with attrition. In the second example, we discuss an instrumental variable model where the instruments are partially observed. The third example is the estimator proposed by Arellano and Bond (1991) for dynamic panel data models. In the fourth example, we use our results to optimally design a data set to measure the change in consumer confidence when nonresponse is expected. The derivations are available upon request.

TABLE 2. Comparison of asymptotic variances.

Estimator	avar $(\hat{\mu}_2)$	avar $(\hat{\mu}_2 - \hat{\mu}_1)$
full data	1	$2\ (1-\rho)$
complete case	$1/p_1$	$2\ (1-\rho)/p_1$
available case	$1/p_1$	$(1-2\rho) + 1/p_1$
optimal	$1/p_1(1-\rho^2(1-p_1))$	$(1-2\rho) + 1/p_1\ (1-\rho^2(1-p_1))$

6.1. **Attrition in two periods.** We study a two-period panel data model with attrition as an example of multivariate mean estimation with missing data. We present analytical results for the asymptotic variance of the estimators.

A health club is interested in measuring the change in the weight of new members after they join. New members are weighed upon registration, and a random sample of new members is selected to come back for a reweighing after six months. Let $X_{i,1}$ be the weight of member i upon registration and let $X_{i,2}$ be the weight of that member after six months.

An error component model can be used to model $X_i = (X_{i,1}, X_{i,2})$: $X_{i,t} = \mu_t + \alpha_i + \epsilon_{it},\ t = 1, 2$, where $\mathbb{E}(\alpha_i) = 0,\ \text{var}(\alpha_i) = \sigma_a^2$ and $\mathbb{E}(\epsilon_{it}) = 0,\ \text{var}(\epsilon_{it}) = \sigma_e^2$ for each $t = 1, 2$. We normalize $\sigma_a^2 + \sigma_e^2$ and denote $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. As a result, $\mathbb{E}(X_i) = (\mu_1, \mu_2)$ and $\Omega = \text{var}(X_i) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

There are two missing data patterns, corresponding to two groups. For an observation i in the first group we observe both $X_{i,1}$ and $X_{i,2}$. For an observation in group 2 we observe only $X_{i,1}$. In other words, $d = 2,\ q = 2,\ J = 2, S_1 = I_2$, and $S_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Assuming that all members who are called for a reweighing show up, the health center has full control over the randomization mechanism, so we assume MI and $\Omega_1 = \Omega_2 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Finally, define $p_1 = \mathbb{P}(R = S_1)$.

The estimation is focused on μ_2 and $(\mu_2 - \mu_1)$ and based on the moment conditions $E(h(X, \mu)) = \mathbb{E}(X - \mu) = 0$. We consider four estimators. The first is the full-data estimator, which equals the sample mean using all n observations. This estimator is not feasible because it uses observations that are missing. We include this estimator to quantify the amount of information that is lost because of the missing data. The second estimator is the complete-case estimator and uses only the complete observations in group 1. The third estimator is the available-case estimator. This estimator uses the maximum number of observations per component: $n_1 + n_2$ for μ_1 and n_1 for μ_2. Finally, we consider the optimal sample mean.

The asymptotic variances of the estimators in this example for $\hat{\mu}_2$ and $(\hat{\mu}_2 - \hat{\mu}_1)$ are given in Table 2. In Figures 6.1 and 6.2 we compare the variances as a function of ρ.

The key element of this example is the individual effect, which introduces correlation between the components of X_i. The optimal estimator efficiently exploits this correlation. An interesting finding is that including observations for members who are observed only upon registration increases the precision for the average weight after six months and for the average change in weight.

The first column of Table 2 and Figure 6.1 show that, for estimating μ_2, the complete-case and the available-case estimators do not recover any of the information that is lost because of the missing data, even when the components are highly correlated. The optimal estimator efficient exploits the correlation. As the individual effect becomes more important, the performance of the optimal estimator relative to the full-data estimator improves. In particular, if $\rho = 1$, observing $X_{i,2}$ does not give any additional information, and the optimal estimator is as efficient as the full-data estimator.

The second column of Table 2 and Figure 6.2 describe the relative performance of the estimator of $\mu_2 - \mu_1$. All estimators benefit from the correlation between X_{i1} and X_{i2}. In the absence of correlation, the optimal estimator coincides with the available-case estimator. If the components are perfectly correlated, both the optimal estimator and the complete-case estimator retrieve all the information.

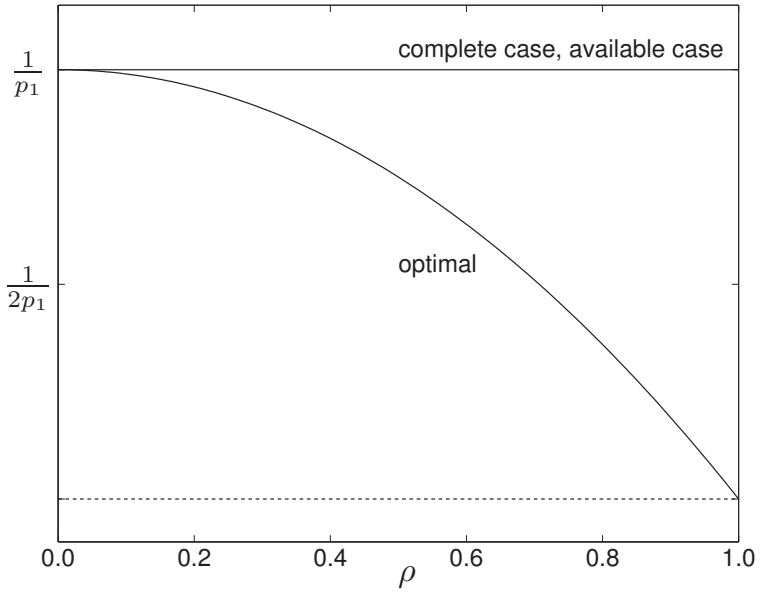

FIGURE 6.1. Asymptotic variances of $\hat{\mu}_2$ as a function of ρ.

To understand why the relative performance of the complete-case and the available-case estimators depends on the correlation, consider that the complete-case estimator corresponds to first calculating $X_{i,2} - X_{i,1}$ and then averaging, while the available-case estimator averages the $X_{i,1}$ and the $X_{i,2}$ and then takes the difference. For the complete-case estimator the individual effects drop out, so that high values of σ_a^2 (ρ) are not reflected in the variance of the estimator. For the variance of the available-case estimator, σ_a^2 does play a role, because this estimator includes observations for which only one period is available. An increase in σ_α^2 therefore increases the variance of the available-case estimator.

6.2. **Instrumental variables.** We study a simple linear instrumental variable model where the dependent and explanatory variables are always observed, but instruments can be incomplete. We consider the linear case with one explanatory variable and two instruments. Either instrument can be missing for a subsample. The approach is easily generalized to multiple explanatory variables, multiple instruments, and nonlinear models. The 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, $y = \beta x + \epsilon$. Two instruments, w_1 and w_2, are available, which motivates the following unconditional moment conditions to estimate β:

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

We assume that the dependent variable and the explanatory variable are always observed. There are three groups of observations, $J_x = 3$. 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. As a result, $J = 3$ and

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

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 are both standardized so that $\mathbb{E}(w_j) = 0,\ j = 1, 2$ and $\mathbb{E}(w_j^2) = 1,\ j = 1, 2$. The instruments have correlation $\rho = \text{cov}(w_1, w_2)$.

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

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

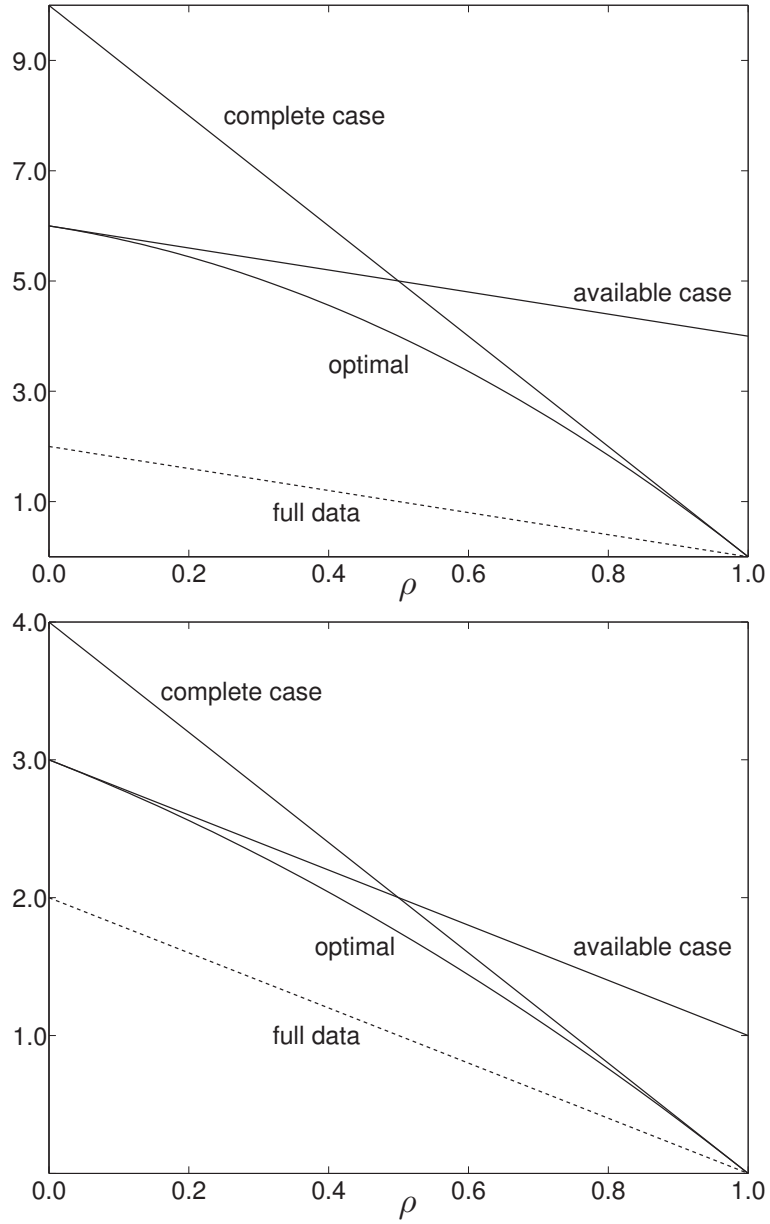

FIGURE 6.2. Asymptotic variances of $\hat{\mu}_2 - \hat{\mu}_1$ as a function of ρ. Top panel: $p_1 = 0.2$. Bottom panel: $p_1 = 0.5$.

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

We consider five estimators. The first four (full data, complete case, available case, and optimal) have been discussed in the text and in Example 6.1. The fifth, which we call the complete-moment estimator, uses one moment only. Because the instruments are similar, the two complete-moment estimators have the same asymptotic variance.

In Figure 6.3 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 act as similar sources of information for

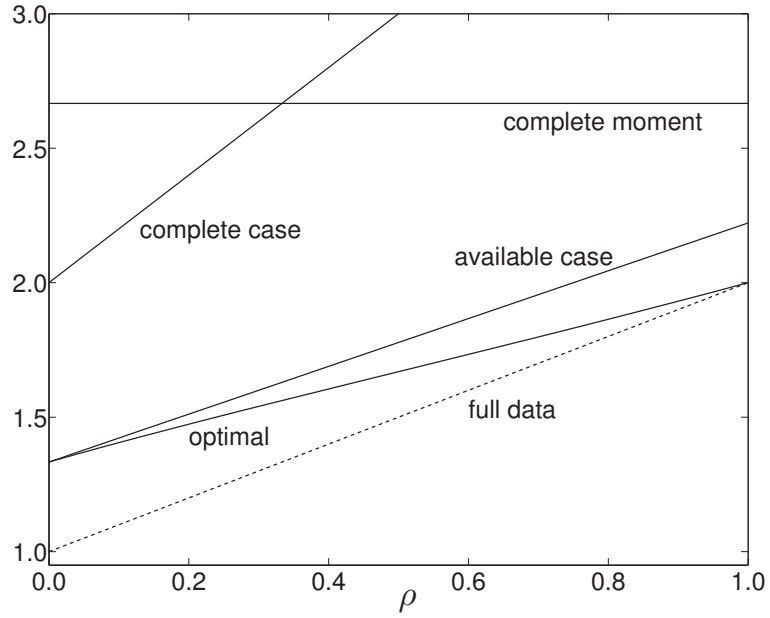

FIGURE 6.3. Asymptotic variance for various estimators of β as a function of ρ, $p_1 = 0.5$.

estimating β. Therefore, as the correlation between w_1 and w_2 increases, we expect two effects. First, the total amount of information for β decreases, so we expect all estimators to be worse. Secondly, the amount of information on the instrument that is missing increases. Since the optimal estimator is constructed such that it efficiently 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 optimal sample mean is able to recover all the information. The complete-case and complete-moment estimators are always outperformed by the available-case estimator and the optimal sample mean.

6.3. **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 it.

The parameter of interest ρ describes the relationship between current and lagged values of a random variable $y_{i,t}$, $y_{i,t} = \alpha_i + \rho y_{i,t-1} + \epsilon_{i,t}$, $2 \leq t \leq T$. We assume that $\mathbb{E}(\alpha_i) = 0$, $\text{var}(\alpha_i) = \sigma_a^2$, and $\mathbb{E}(\epsilon_{it}) = 0$, $\text{var}(\epsilon_{it}) = \sigma_e^2$, and $\mathbb{E}(\epsilon_{i,t}\epsilon_{i,s}) = 0$ 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 do not assume the variance matrix to be known, and 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 normalize $\sigma_\epsilon^2 = 1$. 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 $n = 10000$ and perform $s = 1000$

TABLE 3. Relative variance of the complete-case (cc), available-case (ac), and optimal (opt) estimator in a Monte Carlo study of a continuous updating Arellano-Bond estimator, with $n = 10000,\ s = 1000$, and $T = 9$. The missing data patterns are described in the text.

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

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. This missing data pattern is determined by a parameter p such that $p = \mathbb{P}(R = S_j)$ for each $j = 1, \ldots, 9$, and $\mathbb{P}(R = S_{10}) = 1 - 9p$. We consider $p \in \{0.02, 0.06\}$ so that 82% (respectively 46%) of the observations are complete.

Table 3 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 worse than 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. In contrast to the case where the Ω_j are known, this is not true by construction. The optimal estimator seems to gain more when p is larger. For some parameter configurations, the efficiency gain is substantial.

6.4. **Panel design.** We have considered optimal estimation for given missing data patterns. This analysis is useful for many applications in economics, where the researcher has no control over the data-collection process. For the data collector the relative performance of estimators under different missing data patterns is of importance. Assuming that the researcher uses efficient methods to deal with missing data, what is the best way to collect the data? We discuss data collection for a variable that varies over individuals and over time. We are interested in estimating the change in the population average of the variable over time. We consider three ways to collect the data: repeated cross-sections, a panel, and a rotating panel.

A researcher wants to measure the change in consumer confidence over a period of three years. Denote the confidence of consumer i at time t by $X_{i,t}$, where $1 \leq t \leq 3$, which can be modeled using error components: $X_{i,t} = \alpha_i + \mu_t + \epsilon_{i,t}$. The level of consumer confidence at time t is μ_t. Some consumers may have, across all periods, a more optimistic or pessimistic outlook on the economy, and this is captured by α_i, $\mathbb{E}(\alpha_i) = 0$, and $\text{var}(\alpha_i) = \sigma_a^2$. The idiosyncratic error term $\epsilon_{i,t}$ captures random errors in the elicitation process, and we assume that $\mathbb{E}(\epsilon_{i,t}) = 0$ and $\text{var}(\epsilon_{i,t}) = \sigma_e^2$. It follows that

$$\text{var}(X_{i,t}) = (\sigma_a^2 + \sigma_e^2) \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

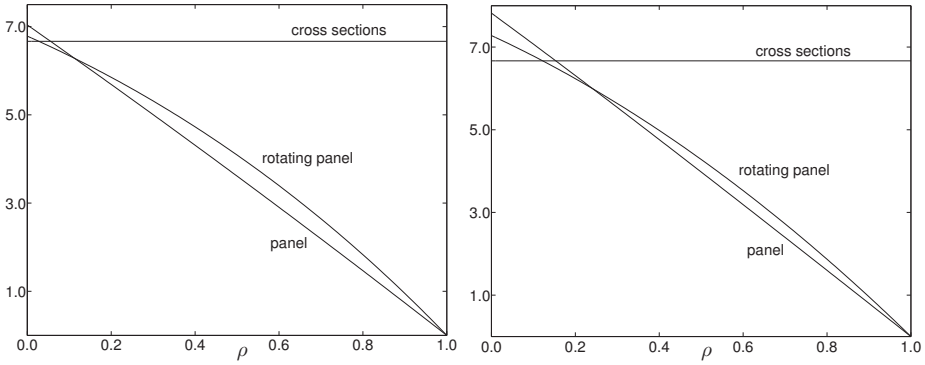

FIGURE 6.4. Asymptotic variances of optimal estimators of the change in consumer confidence using different data collection methods; $p = 0.1$. Left panel: δ_1. Right panel: δ_2.

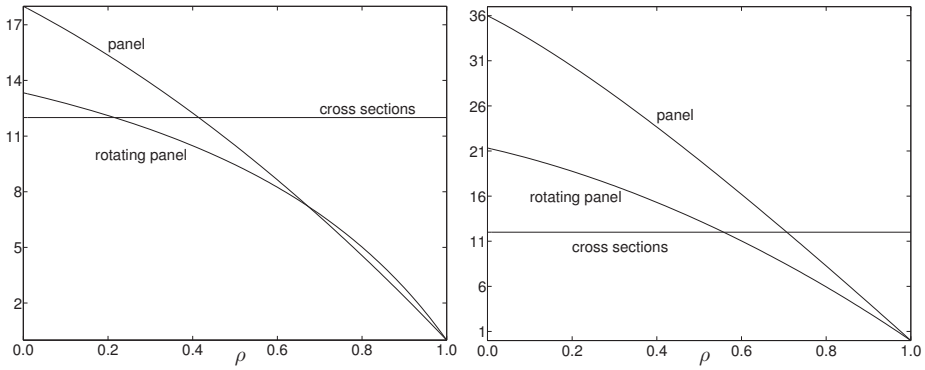

FIGURE 6.5. Asymptotic variances of optimal estimators of the change in consumer confidence using different data collection methods; $p = 0.5$. Left panel: δ_1. Right panel: δ_2.

where $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. The level of consumer confidence does not have an interpretation, so we normalize $\sigma_a^2 + \sigma_e^2 = 1$. The parameters of interest are the changes in consumer confidence, $\mathbb{E}(X_{i,t} - X_{i,t-1}) = \mu_t - \mu_{t-1} = \delta_{t-1}$, for $t = 2, 3$.

The researcher has a budget of $\$M$. Surveying a person once costs \$1, so the researcher can obtain at most M consumer confidence measurements. She considers three ways of collecting the data. The first is a repeated cross-section: for each period, survey a random sample of $M/3$ consumers from the population. The second is a panel: randomly select $M/3$ consumers and survey them in each period. The third is a rotating panel: randomly select $M/4$ consumers to survey in periods 1 and 2, and randomly sample $M/4$ consumers for periods 2 and 3. All these methods exhaust the research budget.

Not all the surveys are completed, which leads to missing data. The missing data mechanism is assumed to be MI. The probability that a consumer does not respond, or stops responding, is p. The research budget allocated to this consumer is lost. Once the data are collected, the researcher will use the methods in this paper to estimate δ_1 and δ_2 optimally. Figures 6.4 and 6.5 show the asymptotic variance of $\hat{\delta}_1$ and $\hat{\delta}_2$ for each of the approaches for $p = 0.1$ and $p = 0.5$ respectively.

The relative performance of the cross-section method increases as the probability of nonresponse increases: a panel member is lost forever, so the effect of nonresponse for the (rotating) panel is stronger than for the cross-section. As ρ increases, the relative performance of the cross-section method decreases, since there is no information available on the missing data, whereas the panel methods can extract some information through the individual effect. The variance of the panel methods is similar for $p = 0.1$, but the rotating panel leads to more substantially more efficient estimators for $p = 0.5$.

7. Conclusion

This paper considered efficient GMM estimation from a random sample of complete and incomplete observations. We derived the semiparametric efficiency bound under an assumption that is weaker than missing completely at random. We introduced an efficient estimator by assigning observations to subsamples on the basis of their missing data pattern. 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.

Appendix A. Proofs

Proof. [Proof of Theorem 3.1]

Abbreviate Newey and McFadden (1994) to NM94. We are going to construct a function Q_0 such that conditions (i)-(iv) in (NM94, Theorem 2.1) are satisfied with respect to Q_0 and Q_n. We defined Q_n in 3.2.

Construction of Q_0. Note that

$$\begin{aligned}\mathbb{E}\left(Rh(X,\theta)|\, R=S_j\right) &= S_j \int h(x,\theta) f_{x|j}(x)dx \\ &= \frac{1}{p_j} S_j \int h(x,\theta) f_{x,r}(x,r)dx \\ &= \frac{1}{p_j} S_j \int h(x,\theta) f_j(x) f_x(x)dx \\ &= \frac{1}{p_j} S_j \mathbb{E}\left(h(x,\theta) f_j(x)\right).\end{aligned}$$

Consider the function

$$k(x,\theta) = \begin{pmatrix} \frac{1}{p_1} S_1 h(x,\theta) f_1(x) \\ \vdots \\ \frac{1}{p_J} S_J h(x,\theta) f_J(x) \end{pmatrix}.$$

Form the blockdiagonal matrix W_n from the blocks $(W_{1,n}, \ldots, W_{j,n})$, and let $W_n \to W$. Define $Q_0(\theta) = k(\theta)' W k(\theta)$. This function can be seen as a GMM criterion function for the Jq moment conditions implied by the conditional moment restrictions $\mathbb{E}(Rh(X,\theta_0)|R) = 0$.

Identification (and compactness) - i and ii. Because of MI, IDENTIFICATION, and REGULARITY(iii), $Q_0(\theta)$ has a unique minimum at θ_0. Therefore, condition (i) for (NM94, Theorem 2.1) is satisfied. Condition (ii) is automatically satisfied by REGULARITY(i).

Continuity - iii. Continuity of k follows immediately from the continuity of h and the requirement that f_j does not depend on θ. This implies that $\frac{1}{p_J} S_j h(x,\theta) f_j(x)$ is continuous in θ if h is.

Uniform convergence - iv. The sample average for subsample j, $\tilde{h}_j(\theta)$, converges uniformly to the j-th conditional expectation. First, we show why the subsample average would converge to the conditional expectation if the inner function were bounded. Then we show that convergence is uniform.

Consider the sample average $\frac{1}{n}\sum_i 1\{R_i = S_j\} R_i h(X_i,\theta)$. By the law of large numbers, this converges to

$$\begin{aligned}\mathbb{E}(1\{R=S_j\}Rh(X,\theta)) &= \sum_k p_k \mathbb{E}(1\{R=S_j\}Rh(X,\theta)|\, R=S_k) \\ &= p_j S_j \mathbb{E}(\, h(X,\theta)|\, R=S_j)\end{aligned}$$

This implies that the subsample average $\tilde{h}_j \to S_j \mathbb{E}(h(X,\theta)|\, R = S_j)$. This shows convergence for each component of k, and therefore for k, and therefore, with condition 6, for Q.

For each component j of the function k, equals $\frac{1}{p_j} S_j h(x,\theta) f_j(x)$. Since $f_j \in [0,1]$ and $p_j \in (0,1]$, the boundedness of h translates to boundedness of k. Continuity of k follows from continuity of h. Therefore, convergence is uniform. See (NM94, Lemma 2.4).

Conditions (i)-(iv) of NM94 are satisfied, and hence $\hat{\theta}_n \to \theta_0$. □

Proof. [Proof of Theorem 3.2]

Each observation provides two random objects that we can use for estimation: a missing moment indicator R_i and the observed elements of the moment function $R_i h(X_i, \dot{})$. The moment conditions are provided by MI, which states that $\mathbb{E}(Rh(X,\theta_0)\ |R)) = 0$. Furthermore, we have that $\mathbb{E}(R) = \sum_{j=1}^{J} p_j S_j$. Under the typical MCAR assumption, we have more information about R, which we can exploit as additional moment conditions, see Graham (2010). However, MI does not provide conditional moment conditions of R on X, or some function of X.

Therefore, the model implies the following moment restrictions on our data: (i) $\mathbb{E}(R) = \sum_{j=1}^{J} p_j S_j$, and (ii) $\mathbb{E}(Rh(X,\theta_0)\ |R)) = 0$. First, we show that the unconditional moment restrictions (i) are not informative for θ_0. Then we derive SPEB(θ_0) using the conditional moment restrictions (ii).

First, denote

$$\mathbb{E}(Rh(X,\theta_0)\ |R)) = \mathbb{E}(\psi_1(R,X;\theta_0)\ |R)$$

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

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

contain the same information as $\mathbb{E}(\psi_1(R,X;\theta_0)\ |R) = 0$. Let $\beta_0 = (\theta_0, p)$. The asymptotic efficiency bound for β_0 based on the unconditional moment restrictions $\mathbb{E}(\rho(R,X;\beta_0)) = \begin{pmatrix} \rho_1(R,X;\theta_0) \\ \rho_2(R;p) \end{pmatrix} = 0$ 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.

Next. we can find the semiparametric efficiency bound for θ_0 given the conditional moment conditions

$$\mathbb{E}(Rh(X,\theta_0)\ |R) = \mathbb{E}(\rho(R,X,\theta_0)\ |R) = 0$$

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

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

The semiparametric efficiency bound is equal to

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

In our case, $D_\rho(S_j) = S_j D_j = S_j \mathbb{E}\left(\frac{\partial h(X,\theta_0)}{\partial \theta}\middle|\, R = S_j\right)$ 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. [Proof of Theorem 4.1]

Let $\Gamma_j = D_j'(S_j\Omega_j S_j)^+ D_j$. $\Gamma_j = \Gamma_j'$ and, because of IDENTIFICATION+, Γ_j is invertible. We need to show that, for any $J-$tuple of weighting matrices ($A_j \in \mathbb{R}^{p\times p},\ j = 1, \ldots, J$),

$$\sum_{j=1}^{J} \frac{1}{p_j} A_j \Gamma_j^{-1} A_j' - (\sum_{j=1}^{J} p_j \Gamma_j)^{-1}$$

is positive semidefinite. Let $K_1' = \left[1/\sqrt{p_1} A_1 \Gamma_1^{-1/2} \quad \ldots \quad 1/\sqrt{p_J} A_J \Gamma_J^{-1/2}\right]$, so that $K_1'K_1 = \sum_{j=1}^{J} \frac{1}{p_j} A_j \Gamma_j^{-1} A_j'$. Similarly, let

$$K_2' = \left[\sqrt{p_1}\Gamma_1^{1/2} \quad \ldots \quad \sqrt{p_J}\Gamma_J^{1/2}\right],$$

so that $(K_2'K_2)^{-1} = (\sum_{j=1}^{J} p_j D_j'(S_j\Omega_j S_j)^+ D_j)^{-1}$.

Furthermore, $K_1'K_2 = \sum_{j=1}^{J} \sqrt{p_j}/\sqrt{p_j} A_j \Gamma_j^{-1/2}\Gamma_j^{1/2} = \sum_{j=1}^{J} A_j = I_p$. Then, by Abadir and Magnus (2005, Exercise 12.18), $K_1'K_1 - (K_2'K_2)^{-1}$ is positive semidefinite, which completes the proof. □

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.

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.

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.

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.

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

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.

Department of Economics, Simon Fraser University
E-mail address: cmuris@sfu.ca
URL: http://www.sfu.ca/~cmuris