An equivalence result for moment equations when data are missing at random

Marian Hristache & Valentin Patilea
CREST, Ensai, France*

December 19, 2018

Abstract

We consider general statistical models defined by moment equations when data are missing at random. Using the inverse probability weighting, such a model is shown to be equivalent with a model for the observed variables only, augmented by a moment condition defined by the missing mechanism. Our framework covers a large class of parametric and semiparametric models where we allow for missing responses, missing covariates and any combination of them. The equivalence result is stated under minimal technical conditions and sheds new light on various aspects of interest in the missing data literature, as for instance the efficiency bounds and the construction of the efficient estimators, the restricted estimators and the imputation.

Key Words: Efficiency bounds; Imputation; Inverse probability weighting; Semiparametric regression; Restricted estimators.

*Address: Center for Research in Economics and Statistics (CREST), Ecole Nationale de la Statistique et de l'Analyse de l'Information (Ensai), Campus de Ker-Lann, rue Blaise Pascal, BP 37203, 35172 Bruz cedex, France.
Emails: marian.hristache@ensai.fr, valentin.patilea@ensai.fr

1 Introduction

Models defined by moment and conditional moment equations are widely used in statistics, biostatistics and econometrics. Here, we investigate general moment or conditional moment equation models with missing data. The main idea we propose is that under a missing at random assumption, the initial model with missing data is equivalent with a inverse probability weighting moment equations model for the complete observations, augmented by a moment condition defined by the missing mechanism. The equivalence, a generalization of the GMM equivalence result of Graham (2011), is stated in terms of sets of probability measures. It has numerous implications and provides valuable insight, for instance on the efficiency bound calculations and the construction of efficient estimators.

The paper is organized as follows. First, we recall some missing at random (MAR) definitions and discuss their relationship. The main equivalence result is stated in section 3. In section 4, we revisit some examples considered in the literature in the MAR setup: estimating mean functionals in parametric and nonparametric regressions; and quantile regression with missing responses and/or covariates. For these examples, our equivalence result suggests new ways for calculating efficiency bounds and constructing efficient estimators, using for instance the GMM, empirical likelihood approaches, the SMD approach of Ai & Chen (2007), or the kernel-based method of Lavergne & Patilea (2013). In section 5 we reinterpret some classes of so-called restricted estimators; see, for instance, Tsiatis (2007) and Tan (2011). Finally, in section 6 we use our general result to discuss on a common belief that the (multiple) imputation is necessary in order to capture all the information from the partially observed data.

2 MAR at a glance

In the framework of missing data, the assumption of missing at random (MAR) is presumably the most used when trying to describe an ignorable mechanism on the missingness. However, this concept, first introduced by Rubin (1976), does not have the same meaning for everyone. For simplicity, let the full observations be i.i.d. replications of a vector $L = (X, Y, Z)$ and let $R = (R_X, R_Y, R_Z) \in \{0, 1\}^3$ be a random vector such that its component takes the value 1 if we observe the corresponding component of L and 0 otherwise. For Rubin (1976) (see also, for example, Robins & Gill (1997), Little & Rubin (2002)), MAR means that missingness depends only on the observed components, denoted by $L_{(R)}$, of L :

$$\begin{gathered}\textit{the conditional law } \mathcal{L}(R|L) \textit{ of } R \textit{ given } L \\ \textit{is the same as the conditional law } \mathcal{L}(R|L_{(R)}) \textit{ of } R \textit{ given } L_{(R)}.\end{gathered} \tag{1}$$

This concept was generalized to CAR, coarsening at random, by Heitjan & Rubin (1991) (see also, for example, van der Laan & Robins (2003)) : $\mathcal{L}(C|L)$ is the same as $\mathcal{L}(C|\varphi(C, L))$

for an always observable transformation $\varphi(C, L)$ of the full data L and the censoring variable C. In the context of regression-like models, the MAR assumption is usually stated in a different and more restrictive way. A strongly ignorable selection mechanism (also called conditional independence, or selection on observables, etc.) means that, assuming some components of L are always observed,

$$\begin{array}{c} \textit{the conditional law } \mathcal{L}(R|L) \textit{ of } R \textit{ given } L \textit{ is the same} \\ \textit{as the conditional law of } R \textit{ given the always observed components of } L. \end{array} \tag{2}$$

This assumption was originally introduced by Rosenbaum & Rubin (1983) in the framework of randomized clinical trials, which corresponds in our simple example, with $L = (X, Y, Z)$, to the case where, for example, X is always observed, and *one and only one* of Y and Z is observed. This means that the selection vector R takes the form $R = (1, D, 1-D)$, where Y is observed iff $D = 1$ and Z is observed iff $D = 0$. In this situation, MAR means

$$\begin{aligned} P(D = 1 \mid X, Y, Z) &= P(D = 1 \mid X, Y) \\ &= 1 - P(D = 0 \mid X, Y, Z) \\ &= 1 - P(D = 0 \mid X, Z) \\ &= P(D = 1 \mid X, Z), \end{aligned}$$

or, equivalently,

$$D \perp\!\!\!\perp Z \mid X, Y \quad \text{and} \quad D \perp\!\!\!\perp Y \mid X, Z. \tag{3}$$

Meanwhile a strongly ignorable missingness mechanism writes

$$P(D = 1 \mid X, Y, Z) = P(D = 1 \mid X),$$

or, equivalently,

$$D \perp\!\!\!\perp (Y, Z) \mid X. \tag{4}$$

Clearly, the condition (4) implies the condition (3), but the reverse is not true in general. In the present work we consider the case of i.i.d. replications of a vector containing missing components for which the same subvector is missing for the incomplete replicates. In this case the MAR assumption (1) and the the strongly ignorable MAR assumption (2) coincide (and are equivalent to CAR), as is it is also the case, for example, in Cheng (1994), Tsiatis (2007), Graham (2011), among others.

Other MAR related assumptions appear in the literature. For instance, when the response Y is missing, while X and Z are observed, Wei, Ma & Carroll (2012) consider the assumption $R_Y \perp\!\!\!\perp (X, Y) \,|\, Z$ that is stronger than the MAR assumption (2), commonly used for regression models. Another assumption for the missingness mechanism is introduced in Wooldridge (2007) : $W = (X, Y)$ and $S \in \{0, 1\}$ is a random variable such that W and Z are observed whenever $S = 1$, and $S \perp\!\!\!\perp W \,|\, Z$. Wooldridge's assumption is more general than the MAR condition (2) where Z is supposed to be always observed. Indeed, Wooldridge (2007) does not suppose that W and/or Z are missing if $S = 0$.

3 Equivalent moment model

The following statement is a version of Theorems 1 and 2 in Hristache & Patilea (2017). The proof is very similar and hence will be omitted. In the following, vectors a columns matrices and for any matrix A, A' denotes its transpose.

Theorem 1 *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two models defined for random vectors $(D, W', V', U')' \in \{0,1\} \times \mathbb{R}^{d_W} \times \mathbb{R}^{d_V} \times \mathbb{R}^{d_U}$ as follows :*

$$\mathcal{M}_1 : \begin{cases} E\left[\rho_j(\gamma, W, V, U)\right] = 0, \quad \forall j \in J, \\ D \perp\!\!\!\perp \{U, V\} \mid W, \end{cases} \tag{5}$$

and

$$\mathcal{M}_2 : \begin{cases} E\left[\dfrac{D}{\pi(W)}\ \rho_j(\gamma, W, V, U)\right] = 0, \quad \forall j \in J, \\ E\left[\dfrac{D}{\pi(W)} - 1 \mid V, W\right] = 0, \end{cases} \tag{6}$$

where $\gamma \in \Gamma$ is an unknown (possibly infinite dimensional) parameter, $\rho_j : \Gamma \times \mathbb{R}^{d_W} \times \mathbb{R}^{d_V} \times \mathbb{R}^{d_U} \to \mathbb{R}$, for $j \in J$, is a collection of known measurable functions, and π is a (unknown) measurable function such that $\pi(W) > 0$ almost surely.

The models $\mathcal{M}_1$ and $\mathcal{M}_2$ are equivalent if restricted to the laws of $(D, W', V', DU')'$; more precisely,

1. $(D, W', V', U')' \in \mathcal{M}_1 \quad \Rightarrow \quad (D, W', V', U')' \in \mathcal{M}_2$,
2. $(D, W', V', U')' \in \mathcal{M}_2 \quad \Rightarrow \quad \exists\ (\widetilde{D}, \widetilde{W}', \widetilde{V}', \widetilde{U}')' \in \mathcal{M}_1$ *such that* $(\widetilde{D}, \widetilde{W}', \widetilde{V}', \widetilde{D}\widetilde{U}')'$ *and* $(D, W', V', DU')'$ *have the same distribution.*

Remarks

1. The link of this theorem with models where data are missing at random is made if we consider that the vector U is observed if and only if $D = 1$. The theorem then basically says that at the observational level, which means for the laws of the observed vector $(D, W', V', DU')'$, the two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are equivalent. As a consequence, inference for the law of $(D, W', V', U')'$ in the model $\mathcal{M}_1$, a moment conditions model under an assumption of data missing at random, could be done based on the model $\mathcal{M}_2$, which is defined using only the observed part $(D, W', V', DU')'$ of the vector vector $(D, W', V', U')'$. In particular, efficiency bound calculations and efficient estimator constructions could be done in the model $\mathcal{M}_2$, which in many cases could be much easier.

2. The underlying condition "DU is always observed" includes the usual case

$$\begin{cases} D = 0 & \text{if } U \text{ is not observed,} \\ D = 1 & \text{if } U \text{ is observed,} \end{cases}$$

   but it is more general. When $D = 1$ one observes the value of U. Meanwhile, one should read that when $D = 0$, U could be observed or not since whatever the value of U is, $DU = 0$.

3. No identification assumption for γ_0, the true value of the parameter γ, is involved in the statements of Theorem 1. In other words, γ_0 is identifiable in the complete data model given in the equation (5), under the MAR assumption, if and only if it is identifiable in the model (6), at the observational level.

4 Some examples revisited

In this section we present two examples of models already studied in the literature for which our approach gives new insights and sometimes allows for simpler methods for obtaining efficiency bounds and asymptotically efficient estimators. The guiding principle is to use Theorem 1 and put the model of interest, in the presence of a MAR mechanism, under an equivalent form

$$\begin{cases} E[g_1(\theta, \alpha, X, Y, Z) \mid X] = 0 \\ E[g_2(\alpha, X, Y, Z) \mid X, Z] = 0, \end{cases} \tag{7}$$

where the two sets of equations are orthogonal, meaning that

$$E[g_1(\theta, \alpha, X, Y, Z)\, g_2'(\alpha, X, Y, Z) \mid X, Z] = 0.$$

The equivalent model (4) has a sequential moment structure that allows to compute the efficiency bound; see Ai & Chen (2012). Moreover, the finite dimensional interest parameter θ can be efficiently estimated from the first equations, with the (possibly infinite dimensional) nuisance parameter α known or suitably estimated from the last equations. A similar statement on the efficient estimation of θ, in the particular case of a finite dimensional α and without conditioning on X and X, Z, can be found in Theorem 2.2, point 8, of Prokhorov & Schmidt (2009).

4.1 Mean functionals with data missing at random

Consider the problem of estimating the mean of functionals of the variables in a parametric regression model with missing responses :

$$\begin{cases} E[h(X,Y)-\theta]=0 \\ \\ E[Y-r(X,\alpha)\,|\,X]=0. \end{cases} \tag{8}$$

The parameter of interest here is $\theta = E[h(X,Y)]$, where $h(\cdot,\cdot)$ is some given squared-integrable function; see Müller (2009), see also Hristache & Patilea (2017). Some usual examples are the mean of the response variable ($h(x,y)=y$), the second order moment of the response ($h(x,y)=vec(yy')$), the cross-product of the response and the covariate vector ($h(x,y)=vec(yx')$). (Here, $vec(\cdot)$ is the vectorization operator that transforms a matrix in a column vector by stacking the columns of the matrix.) For simplicity, we take Y with real values in the following of this section.

The regression function $r(x,\alpha)$ has a known (parametric) form, X is always observed, Y is only observed when $D=1$ and a MAR assumption holds : $D \perp\!\!\!\perp Y\,|\,X$. With $\pi(x)=P(D=1\,|\,X=x)$, the model can be written, at the observational level, under the following equivalent form :

$$\begin{cases} E\left\{\dfrac{D}{\pi(X)}\,[h(X,Y)-\theta]\right\}=0 \\ \\ E\left\{D\,[Y-r(X,\alpha)]\,|\,X\right\}=0 \\ \\ E\left[\dfrac{D}{\pi(X)}-1\,|\,X\right]=0. \end{cases} \tag{9}$$

The last two equations being orthogonal, since

$$E\left\{\left[\frac{D}{\pi(X)}-1\right]D\,[Y-r(X,\alpha)]\,|\,X\right\}=\left[\frac{1}{\pi(X)}-1\right]E\left\{D\,[Y-r(X,\alpha)]\,|\,X\right\}=0,$$

it is also equivalent to the model defined by the following system of orthogonal equations, where $\sigma^2(X)$ stands for the conditional variance $V(Y|X)$:

$$\begin{cases} E\left\{ \dfrac{D}{\pi(X)} \left[h(X,Y) - \theta\right] \right. \\ \qquad - \dfrac{1}{\sigma^2(X)\,\pi(X)}\, E\left[\dfrac{D}{\pi(X)}\, h(X,Y)\,(Y - r(X,\alpha))\,|\,X \right] D\left[Y - r(X,\alpha)\right] \\ \qquad \left. - \, E\left[\dfrac{D}{\pi(X)}\,(h(X,Y) - \theta)\,|\,X \right] \left[\dfrac{D}{\pi(X)} - 1 \right] \right\} = 0 \\ E\left\{ D\ [Y - r(X,\alpha)]\,|\,X \right\} = 0 \\ E\left[\dfrac{D}{\pi(X)} - 1\,|\,X \right] = 0. \end{cases} \tag{10}$$

Solving for θ, we get

$$\theta = E\left[\Phi(Y,X,D;\alpha,\sigma^2,\pi,\eta_1,\eta_2)\right]$$

where

$$\begin{aligned} \Phi(Y,X,D;\alpha,\sigma^2,\pi,\eta_1,\eta_2) &= \frac{D}{\pi(X)}\,h(X,Y) - E\left[\frac{D}{\pi(X)}\,h(X,Y)\,|\,X\right] \left[\frac{D}{\pi(X)} - 1\right] \\ &\quad - \frac{1}{\sigma^2(X)\,\pi(X)}\,E\left[\frac{D}{\pi(X)}\,h(X,Y)\,(Y - r(X,\alpha))\,|\,X\right] D\left[Y - r(X,\alpha)\right], \end{aligned}$$

$$\eta_1(X) = E\left[D\,h(X,Y)\,|\,X\right]$$

and

$$\eta_2(X) = \eta_2(X;\alpha) = E\left[D\,h(X,Y)\,(Y - r(X,\alpha))\,|\,X\right].$$

With the variance $\sigma^2(\cdot)$ and the functions $\eta_1(\cdot)$ and $\eta_2(\cdot;\cdot)$ estimated nonparametrically, the plug-in estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \Phi(Y_i, X_i, D_i; \widehat{\alpha}, \widehat{\sigma^2}, \widehat{\pi}, \widehat{\eta_1}, \widehat{\eta_2})$$

would be efficient. Since the first equation in the system (10) is orthogonalized with respect to the last one, for the propensity score $\pi(\cdot)$, one could use a parametric model without affecting the efficiency bound.

4.2 Quantile regression with data missing at random

A particular setting of quantile regression with missing data at random is considered in Wei, Ma & Carroll (2012). For $0 < \tau < 1$, the conditional quantile $Q_\tau(Y\,|\,X,Z)$ of the always observed response Y given the regressor vectors Z (always observed) and X (observed iff $D = 1$) is assumed to be linear,

$$Q_\tau(Y \mid X, Z) = X'\beta_{1,\tau} + Z'\beta_{2,\tau}, \tag{11}$$

and the missingness mechanism is defined by the strong missing at random condition

$$D \perp\!\!\!\perp (X, Y) \mid Z. \tag{12}$$

Taking in (6) $U = X$, $V = Y$, $W = Z$, $\rho_j(\beta_\tau, W, V, U) = (X', Z')' [\mathbb{1}_{Y - X'\beta_{1,\tau} - Z'\beta_{2,\tau} \leq 0} - \tau] \times a_j(U, W) \triangleq \rho(X, Y, Z, \beta_\tau) \times a_j(X, Z)$, $j \in \mathbb{N}$, where the family of functions $\{a_j\}_{j\in\mathbb{N}}$ spans $L^2(X, Z)$, the model defined by (11) and (12) can be written under the following equivalent form :

$$\begin{cases} E[D\,\rho(Y, X, Z, \beta_\tau) \mid X, Z] = 0 \\ \\ E\left[\dfrac{D}{\pi(Z)} - 1 \mid Z\right] = 0. \end{cases} \tag{13}$$

The two sets of equations being already orthogonal (with respect to the $\sigma-$field $\sigma(\{X, Z\})$), in this situation we can efficiently estimate the parameter $\beta_\tau = (\beta'_{1,\tau}, \beta'_{2,\tau})'$ from the complete data only, that is from the model defined by (11) keeping for the statistical analysis only the observations for which all the components of the vector $(Y, X', Z')'$ are observed. The gain in efficiency observed in the simulation experiment of Wei, Ma & Carroll (2012) for their multiple imputation improved estimator comes, in our opinion, from the supplementary parametric assumption on the form of the conditional density of X given Z (see their Assumption 4).

A more general linear quantile regression model defined by (11) with missing data at random is considered in Chen, Wan & Zhou (2014). With their notations, we have

$$Y = Z'\,\theta(\tau) + \varepsilon, \qquad P(\varepsilon \leq 0 \mid Z) = \tau, \quad 0 < \tau < 1, \tag{14}$$

for the full data model. They also denote by X the always observed components of the vector $(Y, Z')'$ and with X^c the components of the same vector that are observed iff the binary variable D takes the value 1 and use the "standard" missing at random assumption $P(D = 1 \mid Y, Z) = P(D = 1 \mid X, X^c) = P(D = 1 \mid X) = \pi(X)$. This fits our framework by taking $U = X$, $V = 1$, $W = X^c$ and

$$\rho_j(\theta(\tau), W, V, U) = Z\,[\mathbb{1}_{Y - Z'\theta(\tau) \leq 0} - \tau] \times a_j(U, W) \triangleq \rho(Y, Z, \theta(\tau)) \times a_j(Z), \quad j \in \mathbb{N},$$

where the family of functions $\{a_j\}_{j\in\mathbb{N}}$ spans $L^2(Z)$. The equivalent moment equations model, at the observational level, can be written as

$$\begin{cases} E\left\{\dfrac{D}{\pi(X)}\, Z\,[\mathbb{1}_{Y - Z'\theta(\tau) \leq 0} - \tau] \mid Z\right\} = 0 \\ \\ E\left[\dfrac{D}{\pi(X)} - 1 \mid X\right] = 0. \end{cases} \tag{15}$$

The information bound for this model is given in Hristache & Patilea (2016). It can not be calculated explicitly, except some special cases, which includes the missing responses as before or the case where X or/and Z are discrete. It is different from the information bound given in Chen, Hong & Tarozzi (2008) which corresponds to a model defined by an *unconditional* quantile moment and a MAR assumption and could be represented equivalently under the form

$$\begin{cases} E\left\{\dfrac{D}{\pi(X)}\, Z\left[\mathbb{1}_{Y-Z'\theta(\tau)\leq 0}-\tau\right]\right\}=0 \\ E\left[\dfrac{D}{\pi(X)}-1\,|\,X\right]=0. \end{cases} \tag{16}$$

The models (15) and (16) are quite different and so are the corresponding efficiency bounds, so that no estimation procedure given in Chen, Wan & Zhou (2014) could be efficient in their linear quantile regression model (14) with missing data at random.

5 Restricted estimators for quantile regressions and general conditional moment models with data missing at random

The model defined by the regression-like equation

$$E\left[\rho\left(\theta,Y,X,V\right)\ \mid X,V\right]=0,$$

and the MAR selection mechanism

$$P(D=1\,|\,Y,X,V,W)=P(D=1\,|\,W)=\pi(W)$$

is equivalent, at the observational level, to the following model defined by conditional moment equations :

$$\mathcal{P}:\ \begin{cases} E\left[\dfrac{D}{\pi\left(W\right)}\,\rho\left(\theta,Y,X,V\right)\ \mid X,V\right]=0, \\ E\left[\dfrac{D}{\pi\left(W\right)}-1\ \mid\ W\right]=0. \end{cases}$$

This framework includes many situations. For instance, taking $W'=(Y',V',Z')$ we obtain the case in which some regressors (conditioning variables) X are missing, while with $W'=(X',V',Z')$ we cover the case of missing responses. Splitting Y in an observed subvector Y_o and a not always observed subvector Y_u, with $W' = (Y_o', V', Z')$ this corresponds to the case where both some responses and some covariates are missing. In all these examples, U is the vector of not always observed components of the data vector.

For the model

$$\mathcal{P}_{(1)}: \quad E\left[\frac{D}{\pi(W)}\, \rho(\theta, Y, X, V) \mid X, V\right] = 0,$$

denoting by P_0 the true law of $(Y', X', V', Z')'$, the tangent space is

$$\begin{aligned}\mathcal{T}_{(1)} &= \Big\{ s \in \left\{L^2(P_0)\right\}^{\oplus d} : \ E(s) = 0, \\ &\qquad E\left[\frac{D}{\pi(W)}\, \rho(\theta, Y, X, V)\ \ s'(Y, X, V, Z) \mid X, V\right] = 0 \Big\}.\end{aligned}$$

For the model

$$\mathcal{P}_{(2)}: \quad E\left[\frac{D}{\pi(W)} - 1 \mid W\right] = 0,$$

the tangent space is

$$\begin{aligned}\mathcal{T}_{(2)} &= \Big\{ s \in \left\{L^2(P_0)\right\}^{\oplus d} : \ E(s) = 0, \\ &\qquad E\left[\left(\frac{D}{\pi(W)} - 1\right)\ s'(Y, X, V, Z) \mid W\right] = 0 \Big\}.\end{aligned}$$

The tangent space $\mathcal{T}$ of $\mathcal{P} = \mathcal{P}_{(1)} \cap \mathcal{P}_{(2)}$ is (see Hristache & Patilea (2016))

$$\mathcal{T} = \mathcal{T}_{(1)} \cap \mathcal{T}_{(2)}.$$

We obtain the efficient score $\overline{S}_\theta$ by projecting the score S_θ on $\mathcal{T}^\perp$,

$$\overline{S}_\theta = \Pi\left(S_\theta \mid \mathcal{T}^\perp\right) = \Pi\left(S_\theta \mid \overline{\mathcal{T}_{(1)}^\perp + \mathcal{T}_{(2)}^\perp}\right),$$

which gives the following solution :

$$\begin{aligned}\overline{S}_\theta &= a_1^*(X, V)\ \frac{D}{\pi(W)}\ \rho(\theta, Y, X, V) + a_2^*(W)\ \left(\frac{D}{\pi(W)} - 1\right) \\ &\in \mathcal{T}_{(1)}^\perp + \mathcal{T}_{(2)}^\perp,\end{aligned}$$

where

$$\begin{aligned}a_1^*(X, V) &= \left\{ -E\left(\partial_\theta \rho' | X, V\right) + E\left[E\left(a_1^* \rho | W\right) \frac{1-\pi}{\pi} \rho' | X, V\right]\right\} \\ &\qquad \times E^{-1}\left(\frac{1}{\pi(W)}\ \rho\ \rho' \mid X, V\right), \\ a_2^*(W) &= -E\left[a_1^*(X, V)\ \ \rho \mid W\right].\end{aligned}$$

Remark. $\overline{S}_\theta$ is also the efficient score in the model

$$\mathcal{P}: \begin{cases} E\left[a_1^*(X,V)\ \dfrac{D}{\pi(W)}\ \rho(\theta,Y,X,V)\right] = 0 \\ \\ E\left[a_2^*(W)\ \left(\dfrac{D}{\pi(W)} - 1\right)\right] = 0, \end{cases},$$

or in the model defined by the moment condition

$$E\left[a_1^*(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V) + a_2^*(W)\ \left(\frac{D}{\pi(W)} - 1\right)\right] = 0.$$

As shown in Hristache & Patilea (2016), a_1^* satisfies an equation of the form

$$a_1^*(X,V) = \gamma(X,V) + T\left(a_1^*(X,V)\right),$$

with T a contraction operator. The solution of this equation is unique, but in order to obtain it one needs to use nonparametric estimators at each step of the iterative procedure. An alternative approach would be to consider finite dimensional subspaces $\mathcal{S}_1 \subset \mathcal{T}_{(1)}^{\perp}$ and $\mathcal{S}_2 \subset \mathcal{T}_{(2)}^{\perp}$ when calculating the "efficient score", leading to an approximately efficient score. We obtain in this way what is known in the literature as *restricted estimators*. We can write :

$$\mathcal{T}_{(1)}^{\perp} = \left\{s = a_1(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V):\ a_1 \in L^2(P_0)\right\}$$

$\mathcal{S}_1 \subset \mathcal{T}_{(1)}^{\perp}$ finite dimensional $\quad\Rightarrow\quad \exists\ a_1^{(1)}, \ldots, a_1^{(k)} \in L^2(P_0)$ s.t.

$$\mathcal{S}_1 = lin\left\{a_1^{(i)}(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V):\ 1 \le i \le k\right\}$$

$$\Leftrightarrow \quad \mathcal{S}_1^{\perp} = \left\{s \in \left\{L^2(P_0)\right\}^{\oplus d} : E\left(a_1^{(i)}\ \frac{D}{\pi}\ \rho\ s'\right) = 0,\ 1 \le i \le k\right\}.$$

Compare to

$$\mathcal{T}_{(1)} = \left\{s \in \left\{L^2(P_0)\right\}^{\oplus d} : E\left(\frac{D}{\pi}\ \rho\ s' \mid X,V\right) = 0\right\}.$$

Similarly for $\mathcal{S}_2 \subset \mathcal{T}_{(2)}^{\perp}$:

$$\mathcal{T}_{(2)}^{\perp} = \left\{s = a_2(W)\ \left(\frac{D}{\pi(W)} - 1\right):\ a_2 \in L^2(P_0)\right\}$$

$\mathcal{S}_2 \subset \mathcal{T}_{(2)}^{\perp}$ finite dimensional $\quad \Rightarrow \quad \exists\, a_2^{(1)}, \ldots, a_2^{(l)} \in L^2(P_0)$ s.t.

$$\mathcal{S}_2 = lin\left\{a_2^{(j)}(W)\ \left(\frac{D}{\pi(W)} - 1\right):\ 1 \le j \le l\right\}$$

$$\Leftrightarrow \qquad \mathcal{S}_2^{\perp} = \left\{s \in \left\{L_0^2(P_0)\right\}^{\oplus d} : E\left[a_2^{(j)}\ \left(\frac{D}{\pi(W)} - 1\right)\ s'\right] = 0, \quad 1 \le j \le k\right\}.$$

An optimal class 1 restricted estimator (see Tsiatis (2007), Tan (2011)) is solution of the approximated efficient score equation

$$E\left\{\overline{a}_1^{(1)}(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V) + \overline{a}_2^{(1)}(W)\ \left(\frac{D}{\pi(W)} - 1\right)\right\} = 0,$$

where $\overline{a}_1^{(1)}$ and $\overline{a}_2^{(2)}$ are given by

$$\begin{aligned}
\overline{S}_\theta &= \Pi\left(S_\theta \mid \mathcal{S}_1 + \mathcal{S}_2\right) \\
&= \overline{a}_1^{(1)}(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V) + \ \overline{a}_2^{(1)}(W)\ \left(\frac{D}{\pi(W)} - 1\right).
\end{aligned}$$

In fact, $\overline{S}_\theta$ is the efficient score in the following moment equations model :

$$\mathcal{P}' : \begin{cases}
E\left[a_1^{(1)}(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V)\right] = 0 \\
\vdots \\
E\left[a_1^{(k)}(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V)\right] = 0 \\
E\left[a_2^{(1)}(W)\ \left(\frac{D}{\pi(W)} - 1\right)\right] = 0 \\
\vdots \\
E\left[a_2^{(l)}(W)\ \left(\frac{D}{\pi(W)} - 1\right)\right] = 0
\end{cases}$$

This allows for a new, simple and intuitive interpretation of the optimal class 1 restricted estimators as efficient estimators in a larger model, obtained from the initial one by using appropriate "instruments" to transform the conditional moment equations in a (growing) number of unconditional moment conditions. Another advantage of this new perspective is the access to the most commonly used methods of obtaining efficient estimators in moment equations models such as GMM, SMD (see Lavergne & Patilea (2013)) or empirical likelihood estimators.

Similar procedures can be used for class 2 restricted estimators, based on

$$\Pi\left(S_\theta \mid \mathcal{S}_1 + \mathcal{T}_{(2)}^{\perp}\right) = \overline{a}_1^{(2)}(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V) + \overline{a}_2^{(2)}(W)\ \left(\frac{D}{\pi(W)} - 1\right)$$

and class 3 restricted estimators (Tan (2011)), based on

$$\Pi\left(S_\theta \mid \mathcal{T}_{(1)}^{\perp} + \mathcal{S}_2\right) = \overline{a}_1^{(3)}(X,V)\ \frac{D}{\pi(W)}\ \rho(\theta,Y,X,V) + \overline{a}_2^{(3)}(W)\ \left(\frac{D}{\pi(W)} - 1\right).$$

6 Is imputation informative ?

Multiple imputation is a widely used method to generate substitute values when data are missing. However, under the MAR assumption, the interest of multiple imputation in the context of conditional moment restriction models is at least questionable, as discussed in the following.

Consider that $(D, W', V', DU')'$ is always observed and consider the MAR assumption

$$(U,V) \perp\!\!\!\perp D \mid W. \tag{17}$$

Then, any substitute observation generated from the law of $\widetilde{U}$ is adequate to replace a missing U, where the law of $\widetilde{U}$ should be such that

$$\mathcal{L}(\widetilde{U} \mid \widetilde{W}, \widetilde{V}, \widetilde{D} = 0) = \mathcal{L}(U \mid W, V, D = 1) = \mathcal{L}(\widetilde{U} \mid \widetilde{W}, \widetilde{V}, \widetilde{D} = 1).$$

(Here, $\mathcal{L}(V_1 \mid V_2)$ denotes the conditional law of V_1 given V_2.) Since, in general, the law $\mathcal{L}(U \mid W, V, D = 1)$ is unknown, one can estimate it, parametrically or nonparametrically, and generate substitute observations from this estimate. This is the so-called parametric or nonparametric imputation. See, for instance, Wang & Chen (2009), Wei, Ma & Carroll (2012), Chen & Van Keilegom (2013) for some nonparametric imputation applications.

The equivalence established by Theorem 1 for models defined by moment restrictions, implies that *all* the information on the parameter θ in the initial model under the MAR assumption (17) is contained in the model defined by the equations (6). Let us point out that the last equation of the model (6) includes the information contained in the incomplete observations. Indeed, to estimate $\pi(\cdot)$, parametrically or nonparametrically, one uses *all* the observations of W. This remark opens new perspectives for defining estimators of θ without using substitute observations. Moreover, this remark sheds some new light on a common justification used in the literature, namely that imputation is necessary in order to capture the information contained in the partially observed data.

References

Ai, C. & Chen, X. (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics* **141**, 5–43.

Ai, C. & Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics* **170**, 442–457. Thirtieth Anniversary of Generalized Method of Moments.

Chen, S. X. & Van Keilegom, I. (2013). Estimation in semiparametric models with missing data. *Annals of the Institute of Statistical Mathematics* **65**, 785–805.

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

Chen, X., Wan, A. T. & Zhou, Y. (2014). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association* , accepted manuscript.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association* **89**, 81–87.

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

Heitjan, D. F. & Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics* **19**, 2244–2253.

Hristache, M. & Patilea, V. (2016). Semiparametric efficiency bounds for conditional moment restriction models with different conditioning variables. *Econometric Theory* **32**, 917–946.

Hristache, M. & Patilea, V. (2017). Conditional moment models with data missing at random. *Biometrika* **104**, 735–742.

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

Little, R. & Rubin, D. (2002). *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.

Müller, U. U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *The Annals of Statistics* **37**, 2245–2277.

Prokhorov, A. & Schmidt, P. (2009). GMM redundancy results for general missing data problems. *Journal of Econometrics* **151**, 47 – 55.

Robins, J. M. & Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in medicine* **16**, 39–56.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Tan, Z. (2011). Efficient restricted estimators for conditional mean models with missing data. *Biometrika* **98**, 663–684.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer.

van der Laan, M. J. & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer.

Wang, D. & Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics* **37**, pp. 490–517.

Wei, Y., Ma, Y. & Carroll, R. J. (2012). Multiple imputation in quantile regression. *Biometrika* **99**, 423–438.

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