Conditional Moment Restriction Models with Data Missing at Random

Marian Hristache & Valentin Patilea
CREST (Ensai)*

December 5, 2014

Abstract

We consider a general statistical model defined by moment restrictions when data are missing at random. Using the inverse probability weighting, we show that such a model is equivalent with a model for the observed variables only, augmented by a moment condition defined by the missing mechanism. In particular, our framework covers parametric and semiparametric mean regressions and quantile regressions. We allow for missing responses, missing covariates and any combination of them. The equivalence result is obtained under minimal technical conditions and sheds new light on various aspects of interest in the missing data literature, as for instance the double-robustness, imputation, the restricted estimators and the puzzling efficiency of the procedures based on the estimated, rather than known, selection probabilities. It also provides a general framework for building (efficient) estimators. Moreover, as an immediate consequence of our general theory, we derive the efficiency of the complete cases analysis in a general semiparametric regression model with responses missing at random.

Key Words: Double-robustness; Efficiency bounds; Imputation; Inverse probability weighting; MAR assumption; Moment restrictions models; Semiparametric regression; Restricted estimators.

1 Moment conditions and data missing at random

Moment and conditional moment restriction models represent a wide class of models used in statistics, biostatistics and econometrics: general regression and transformation models,

*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.
Authors emails: marian.hristache@ensai.fr, valentin.patilea@ensai.fr

linear and nonlinear quantile regression models, linear or nonlinear simultaneous equations, and econometric models of optimizing agents are only few examples. In this paper we investigate general moment or conditional moment restrictions models with missing data. Our main message could be simply stated as follows: under a missing at random assumption, the initial model with missing data is equivalent with a inverse probability weighting moment restriction model with for the observed variables, augmented by a moment condition defined by the missing mechanism. The equivalence is in terms of models, *i.e.*, in terms of sets of probability measures. This equivalence, a generalization of the GMM equivalence result of Graham (2011), has numerous implications and provides valuable insight on the efficiency bound calculations, the construction of efficient estimators, the double-robustness, the imputation, the interpretation of the paradox on the use of the true propensity score (selection probability), *etc.* In particular, the equivalence implies that in conditional moment restriction models where the regressors are non missing, dropping the incomplete observations does not affect the statistical information.

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,

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

This assumption was originally introduced by Rosenbaum & Rubin (1983) in the framework of random 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 strongly ignorable 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 'strongly ignorable' MAR condition where Z and W are supposed always observed. Indeed, Wooldridge (2007) does not suppose that W and/or Z are missing if $S = 0$.

The paper is organized as follows. The main equivalence results are provided in the sections 2 and 3. In section 4, we revisit some examples considered in the literature in the MAR setup: estimating mean functionals in parametric and nonparametric regressions; regression with missing responses and propensity score (selection probability) that depends also on auxiliary variables; quantile regression with missing responses and/or covariates. For all 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 and 6, we investigate the cases

of conditional moment restriction models. In section 5 we consider the case where the moment condition does not involve any infinite-dimensional nuisance parameter, while in section 6 we allow for such a parameter in the setup of generalized partially linear single-index regression models. In section 7 we provide a new interpretation and characterization of the double-robustness. In section 8 we reinterpret some classes of so-called restricted estimators; see, for instance, Tsiatis (2007) and Tan (2011). In section 9 we provide a clear explanation to the puzzling phenomenon of weighting using estimated selection probabilities is better than weighting using the true selection probabilities. Similar explanations have been already provided by Prokhorov & Schmidt (2009) and Graham (2011) in less general setups. Finally, in section 10 we use our general results to make clear why (multiple) imputation is not necessary to capture all the information from the partially observed data.

2 Inverse probability weighting for general models with missing data at random

Theorem 1 *Let* $(D, W', V')' \in \{0,1\} \times \mathbb{R}^{d_W} \times \mathbb{R}^{d_V}$ *be a random vector always observed such that :*

1. $0 < \pi(w) = P(D = 1|W = w) \leq 1, \quad \forall w \in suppW;$
2. $V \perp\!\!\!\perp D \mid W.$

Let $U \in \mathbb{R}^{d_U}$ *be a random vector and* J *an arbitrary set such that :*

a) DU *is always observed;*

b) $(U, V) \perp\!\!\!\perp D \mid W$ *(missing at random assumption);*

c) $E[\rho_j(\theta, W, V, U, h_j(\theta, W, V, U))] = 0, \quad \forall j \in J,$

where $\theta \in \Theta \subset \mathbb{R}^d$ *is an unknown finite dimensional parameter,* $h_j(\cdot) \in \mathbb{R}^{d_{h_j}}$ *is an unknown infinite dimensional parameter and* $\rho_j : \mathbb{R}^{d+d_W+d_V+d_U+d_{h_j}} \to \mathbb{R}^{d_{\rho_j}}$ *for* $j \in J$. *Then*

$$\begin{cases} E\left[\dfrac{D}{\pi(W)}\ \rho_j\left(\theta, W, V, \overline{U}, h_j(\theta, W, V, \overline{U})\right)\right] = 0, \quad \forall j \in J, \\ E\left[\dfrac{D}{\pi(W)} - 1 \mid W\right] = 0, \end{cases} \tag{5}$$

where $\overline{U} = U$ *or* $\overline{U} = DU$.

Proof. The last equation in the system (5) is obvious. For the first ones, with $\overline{U} = U$ or $\overline{U} = DU$, we have :

$$
\begin{aligned}
& E\left[\frac{D}{\pi(W)}\ \rho_j\left(\theta, W, V, \overline{U}, h_j(\theta, W, V, \overline{U})\right)\right] \\
& \quad = E\left[\frac{D}{\pi(W)}\ \rho_j\left(\theta, W, V, U, h_j(\theta, W, V, U)\right)\right] \\
& \quad = E\left[\frac{E\left(D \mid W, V, U\right)}{\pi(W)}\ \rho_j\left(\theta, W, V, U, h_j(\theta, W, V, U)\right)\right] \\
& \quad \overset{(\mathrm{MAR})}{=} E\left[\frac{E\left(D \mid W\right)}{\pi(W)}\ \rho_j\left(\theta, W, V, U, h_j(\theta, W, V, U)\right)\right] \\
& \quad = E\left[\frac{\pi(W)}{\pi(W)}\ \rho_j\left(\theta, W, V, U, h_j(\theta, W, V, U)\right)\right] = 0, \quad \forall j \in J. \qquad \blacksquare
\end{aligned}
$$

Remarks

1. The first equations in (5) justify the inverse probability weighting method of estimation in the missing at random framework in order to obtain consistent asymptotically normal estimators for the true value θ_0 of the parameter of interest. These estimators are seldom asymptotically efficient.

2. In Theorem 1 it is not required that the selection probability $\pi(\cdot)$ depends on all the components of the completely observed data (W, V). See, for instance, Wei, Ma & Carroll (2012) for an example of such a situation. The usual MAR assumption used in regression-like models in the literature corresponds to the case $V \equiv 1$.

3. The 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, $DU = 0$.

4. Theorem 1 remains valid in the case of a missing mechanism as considered by Wooldridge (2007) where W is observed whenever $D = 1$.

3 Equivalence with (conditional) moment equations for observed data

Theorem 1 has a reciprocal which establishes the equivalence of the stated missing at random assumption and the model defined by the moment conditions

$$E\left[\rho_j\left(\theta, W, V, U, h_j(\theta, W, V, U)\right)\right] = 0, \quad \forall j \in J, \tag{6}$$

with the model defined, at the observational level, by the equations (5).

Theorem 2 *Let $(D, W', V')' \in \{0,1\} \times \mathbb{R}^{d_W} \times \mathbb{R}^{d_V}$ be a random vector always observed such that :*

1. $0 < \pi(w) = P\left(D = 1 | W = w\right) \leq 1, \quad \forall w \in suppW$;
2. $V \perp\!\!\!\perp D \mid W$.

Let $\overline{U} \in \mathbb{R}^{d_U}$ be an observed random vector such that the system of equations (5) is verified, where $\theta \in \Theta \subset \mathbb{R}^d$ is an unknown finite dimensional parameter. Let J be some arbitrary set, $h_j(\cdot) \in \mathbb{R}^{d_{h_j}}$ be an unknown infinite dimensional parameter and $\rho_j : \mathbb{R}^{d+d_W+d_V+d_U+d_{h_j}} \to \mathbb{R}^{d_{\rho_j}}$ for $j \in J$.

There exists a random vector $\widetilde{U} \in \mathbb{R}^{d_U}$ such that :

a) $D\widetilde{U} = D\overline{U}$ is always observed;

b) $(\widetilde{U}, V) \perp\!\!\!\perp D \mid W$ (missing at random assumption);

c) $E\left[\rho_j\left(\theta, W, V, \widetilde{U}, h_j(\theta, W, V, \widetilde{U})\right)\right] = 0, \quad \forall j \in J.$

Proof. Consider a new variable $\widetilde{U}$ such that the conditional law $\mathcal{L}(\widetilde{U}, V \mid W, D)$ of $(\widetilde{U}, V)$ on W and D is the same as the *observed* conditional law $\mathcal{L}(U, V \mid W, D = 1)$:

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

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

Then, by construction,

$$\widetilde{U} \perp\!\!\!\perp D \mid W, V,$$

since $\mathcal{L}(\widetilde{U} \mid W, V, D = 0) \;=\; \mathcal{L}(U \mid W, V, D = 1) \;=\; \mathcal{L}(\widetilde{U} \mid W, V, D = 1)$, which together with $V \perp\!\!\!\perp D \mid W$ gives

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

We also have

$$D\ \rho_j\left(\theta, W, V, U, h_j(\theta, W, V, U)\right) = D\ \rho_j\left(\theta, W, V, \widetilde{U}, h_j(\theta, W, V, \widetilde{U})\right), \quad \forall j \in J. \qquad (7)$$

We then get

$$\begin{aligned}
E\left[\rho_j\left(\theta, W, V, \widetilde{U}, h_j(\theta, W, V, \widetilde{U})\right)\right] &= E\left\{E\left[\rho_j\left(\theta, W, V, \widetilde{U}, h_j(\theta, W, V, \widetilde{U})\right) \mid W, V\right]\right\} \\
&\overset{(1.)+(2.)}{=} E\left\{\frac{1}{\pi(W)}\ E\left(D \mid W, V\right)\ E\left[\rho_j\left(\theta, W, V, \widetilde{U}, h_j(\theta, W, V, \widetilde{U})\right) \mid W, V\right]\right\} \\
&\overset{\widetilde{U} \perp D \mid W}{=} E\left\{\frac{1}{\pi(W)}\ E\left[D\ \rho_j\left(\theta, W, V, \widetilde{U}, h_j(\theta, W, V, \widetilde{U})\right) \mid W, V\right]\right\} \\
&\overset{(7)}{=} E\left\{\frac{1}{\pi(W)}\ D\ \rho_j\left(\theta, W, V, U, h_j(\theta, W, V, U)\right)\right\} \\
&\overset{(5)}{=} 0, \quad \forall j \in J.
\end{aligned}$$

In conclusion, we obtain :

- $D\widetilde{U} = D\overline{U} = DU$ which is always observed,
- $(\widetilde{U}, V) \perp\!\!\!\perp D \mid W$ (missing at random),
- the complete data $\left(D, W, V, \widetilde{U}\right)$ satisfy

$$E\left[\rho_j\left(\theta, W, V, \widetilde{U}, h_j(\theta, W, V, \widetilde{U})\right)\right] = 0, \quad \forall j \in J. \qquad \blacksquare$$

Remarks.

1. The construction of $\widetilde{U}$ in the general missing data framework defined above shows that, except maybe some special cases, the missing at random assumption can not be tested from the observed data, since we can not distinguish at the observational level between the distributions of U and $\widetilde{U}$. This is true in particular for the models defined as in (6), that is, for all statistical models defined by a family of distributions indexed by an arbitrary parameter. It generalizes a result of Robins & Gill (1997), in the sense that it applies for full data with continuous components, but it is obtained under a stronger MAR assumption.

2. No identification assumption for θ_0, the true value of the parameter θ, is involved in the statements of Theorems 1 and 2. In other words, θ_0 is identifiable in the complete data model given in the equation (6), under the MAR assumption, if and only if it is identifiable in the model (5) at the observational level.

4 Some examples revisited

In this section we present some 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 put the models under investigation under the form

$$\begin{cases} E[\rho_1(X, Y, Z, \theta, \alpha) \,|\, X] = 0 \\ \\ E[\rho_2(X, Y, Z, \alpha) \,|\, X, Z] = 0 \end{cases} \tag{8}$$

where the two sets of equations are orthogonal, meaning that

$$E[\rho_1(X, Y, Z, \theta, \alpha)\, \rho_2'(X, Y, Z, \alpha) \,|\, X, Z] = 0.$$

Then, θ can be efficiently estimated from the first equations, with α known or $\sqrt{n}-$consistently estimated from the last equations. A similar statement on the efficient estimation of θ, in the particular case without conditioning on X and X, Z, can be found in Theorem 2.2, point 8, of Prokhorov & Schmidt (2009).

4.1 Estimating mean functionals with missing data

Müller (2009) considered the following generalization of estimating the mean of a response variable in a parametric regression model with missing responses :

$$\begin{cases} E[h(X, Y) - \theta_0] = 0 \\ \\ E[Y - r(X, \alpha_0) \,|\, X] = 0. \end{cases} \tag{9}$$

The parameter of interest here is $\theta_0 = E[h(X, Y)]$. 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 :

$$\begin{cases} E\left\{ \dfrac{D}{\pi(X)} \, [h(X, Y) - \theta_0] \right\} = 0 \\ \\ E\left\{ D \, [Y - r(X, \alpha_0)] \,|\, X \right\} = 0 \\ \\ E\left[\dfrac{D}{\pi(X)} - 1 \,|\, X \right] = 0. \end{cases} \tag{10}$$

The last two equations being orthogonal, since

$$E\left\{\left[\frac{D}{\pi(X)}-1\right]D\left[Y-r(X,\alpha_0)\right]\mid X\right\}=\left[\frac{1}{\pi(X)}-1\right]E\left\{D\left[Y-r(X,\alpha_0)\right]\mid X\right\}=0,$$

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

$$\left\{\begin{array}{l} E\left\{\dfrac{D}{\pi(X)}\left[h(X,Y)-\theta\right]\right. \\ \qquad -\dfrac{1}{\sigma^2(X)\,\pi(X)}E\left[\dfrac{D}{\pi(X)}h(X,Y)\left(Y-r(X,\alpha_0)\right)\mid X\right]D\left[Y-r(X,\alpha)\right] \\ \qquad \left.-E\left[\dfrac{D}{\pi(X)}\left(h(X,Y)-\theta_0\right)\mid X\right]\left[\dfrac{D}{\pi(X)}-1\right]\right\}=0 \\ E\left\{D\left[Y-r(X,\alpha)\right]\mid X\right\}=0 \\ E\left[\dfrac{D}{\pi(X)}-1\mid X\right]=0. \end{array}\right. \tag{11}$$

If the selection probability, or propensity score, $\pi(X)$ has a parametric form $\pi(X)=p(X,\gamma)$, the information on θ_0 can be deduced from the result of Chamberlain (1992); see also Hristache & Patilea (2011) where the efficient score for $(\theta_0,\alpha_0,\gamma_0)$ is derived in this context of sequential moments. In the general case of the missing data model (9) with unknown propensity score $\pi(X)$, an efficient ad-hoc estimator was obtained by Müller (2009), after calculating the efficiency bound for θ_0 in this model. Since it is equivalent to the moment equations model (10), the expression of the efficiency bound is a particular case of the result obtained by Ai & Chen (2012), if we accept that their variational independence assumption on the parameters of the model seems to be satisfied in this situation (see Assumption A and the comments before this assumption on page 452 in Ai & Chen (2012)). A GMM or SMD type estimator can then be used to efficiently estimate θ_0.

A related framework is obtained when considering the problem of estimating the mean $\theta_0=E(Y)$ of Y in the presence of some covariate vector X when some of the Y's are missing according to a MAR assumption. Here $h(X,Y)=Y$ but the regression function $r(X)=E(Y|X)$ of Y on X can be nonparametric, as in Cheng (1994). We thus have :

$$\left\{\begin{array}{l} E(Y-\theta_0)=0 \\ E[Y-r(X)\mid X]=0, \end{array}\right. \tag{12}$$

and the MAR assumption $Y \perp\!\!\!\perp D \,|\, X$. At the observational level this can be written equivalently under the form

$$\begin{cases} E\left[\dfrac{D}{\pi(X)}\,(Y-\theta_0)\right]=0 \\ E\{D\,[Y-r(X)]\,|\,X\}=0 \\ E\left[\dfrac{D}{\pi(X)}-1\,|\,X\right]=0. \end{cases} \tag{13}$$

Orthogonalizing the first equation with respect to the two others (which are already orthogonal) leads to :

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

which further gives

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

that is,

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

In this way, from the first equation of this last system, we obtain an efficient estimator of

θ_0 by taking

$$\widetilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{r}(X_i) = \frac{\displaystyle\sum_{j=1}^{n} D_j\, Y_j\, K_h(X_i - X_j)}{\displaystyle\sum_{j=1}^{n} D_j\, K_h(X_i - X_j)}.$$

The asymptotic law of this estimator was obtained by Cheng (1994); its efficiency can also be deduced from the efficiency bound calculations in Chen, Hong & Tarozzi (2008). The estimator $\widehat{r}(x)$ of $r(x)$ comes from the second equation in (13):

$$E\{D\,[Y - r(X)]\,|\,X\} = 0 \quad \Leftrightarrow \quad E(DY\,|\,X) = r(X)\,E(D\,|\,X) \quad \Rightarrow \quad r(X) = \frac{E(DY\,|\,X)}{E(D\,|\,X)}$$

$$\Rightarrow \quad \widehat{r}(X) = \frac{\dfrac{\displaystyle\sum_{j=1}^{n} D_j\, Y_j\, K_h(X - X_j)}{\displaystyle\sum_{j=1}^{n} K_h(X - X_j)}}{\dfrac{\displaystyle\sum_{j=1}^{n} D_j\, K_h(X - X_j)}{\displaystyle\sum_{j=1}^{n} K_h(X - X_j)}} = \frac{\displaystyle\sum_{j=1}^{n} D_j\, Y_j\, K_h(X - X_j)}{\displaystyle\sum_{j=1}^{n} D_j\, K_h(X - X_j)},$$

where K is a kernel function, h a bandwidth and $K_h(\cdot) = K(\cdot/h)/h$.

Estimating the mean $\theta_0 = E(Y)$ from model (12) while some Y's are missing is equivalent to consider only the first equation $(Y - \theta_0)$ and the MAR assumption, since the second equation only defines the regression function $r(X) = E(Y\,|\,X)$. In other words, the equivalent moment equations model (13) becomes

$$\begin{cases} E\left[\dfrac{D}{\pi(X)}\,(Y - \theta_0)\right] = 0 \\[2ex] E\left[\dfrac{D}{\pi(X)} - 1\,|\,X\right] = 0. \end{cases} \tag{14}$$

After orthogonalization, we get

$$\begin{cases} E\left\{\dfrac{D}{\pi(X)}\,[Y - \theta] - [r(X) - \theta]\,\left[\dfrac{D}{\pi(X)} - 1\right]\right\} = 0 \\[2ex] E\left[\dfrac{D}{\pi(X)} - 1\,|\,X\right] = 0, \end{cases}$$

that is

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

The corresponding efficient estimator of θ_0 is

$$\widetilde{\theta}_n=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i}{\widehat{\pi}(X_i)}\left[Y-\widehat{r}(X_i)\right]+\widehat{r}(X_i)\right],$$

for suitable estimators $\widehat{r}$ of r and $\widehat{\pi}$ of π. If a parametric model is specified for r and π, this is a well known doubly robust estimator of the mean; see Rotnitzky, Lei, Sued & Robins (2012) for references and recent developments.

If in the model defined by the equations (10) or (13) one does not want to use non-parametric estimators for $r(X)$ and/or $\pi(X)$ (due to the curse of dimensionality, for example), but one still wants some protection against a misspecified parametric form of these functions, one can use an encompassing model of the form

$$\begin{cases} E\left\{\dfrac{D}{\pi(X,\gamma)}\left[h(X,Y)-\theta\right]\right\}=0 \\ \\ E\left\{D\left[Y-r(X,\alpha)\right]a(X,\alpha,\gamma)\right\}=0 \\ \\ E\left\{\left[D-\pi(X,\gamma)\right]b(X,\gamma)\right\}=0. \end{cases} \tag{15}$$

The functions $a(\cdot)$ and $b(\cdot)$ can be taken such that the parameters α and γ are obtained as solutions of the first order conditions from suitable minimum distance procedures. For example, if one assumes that $\pi(X)$ is known, one can take

$$a(X,\alpha,\gamma)=a(X,\alpha)=\partial_\alpha r(X,\alpha)\,\frac{1-\pi(X)}{\pi^2(X)}$$

as in Cao, Tsiatis & Davidian (2009), which corresponds to a parameter α which minimizes the weighted least squares criterion $E\left\{D\,\dfrac{1-\pi(X)}{\pi^2(X)}\left[Y-r(X,\alpha)\right]^2\right\}$. With unknown γ, one can take $b(X,\gamma)=\partial_\gamma\pi(X,\gamma)$ in system (15), corresponding to the minimization of $E\left\{\left[D-\pi(X,\gamma)\right]^2\right\}$ with respect to γ. Such ad-hoc choices for $a(\cdot)$ and $b(\cdot)$ may induce a loss in the asymptotic efficiency of the moment estimator. However, this asymptotic efficiency loss could be rewarded by other interesting features, as for instance better finite sample properties, due to avoiding the use of nonparametric estimators or double robustness property (see equations (40)).

4.2 Regression with missing responses

Consider now the problem of estimating the parameter θ of a regression function $r(X,\theta) = E(Y \mid X)$ when the response Y is not always observed. If the indicator D of missing Y's satisfies a MAR assumptions

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

where the regressors X and the auxiliary vector Z are always observed. Taking in (5) $U=Y$, $W=(X,Z)$, $V=1$, $\rho_j(W,V,U,\theta) = [Y - r(X,\theta)]\, a_j(X)$, $j \in \mathbb{N}$, where the family of functions $\{a_j\}_{j\in\mathbb{N}}$ spans $L^2(X)$, we can write this regression model with missing responses under the following equivalent form :

$$\begin{cases} E\left\{ \dfrac{D}{\pi(X,Z)} \left[Y - r(X,\theta)\right] \mid X \right\} = 0 \\ \\ E\left[\dfrac{D}{\pi(X,Z)} - 1 \mid X, Z \right] = 0. \end{cases} \tag{16}$$

Following the strategy for obtaining the efficient score in a sequential conditional moment equations model described in section 2.1 of Hristache & Patilea (2011), we have

$$\begin{aligned} &\frac{D}{\pi(X,Z)}\left[Y - r(X,\theta)\right] - E\left\{ \frac{D}{\pi(X,Z)}\left[Y - r(X,\theta)\right] \left[\frac{D}{\pi(X,Z)} - 1\right] \mid X, Z \right\} \\ &\qquad \times Var^{-1}\left[\frac{D}{\pi(X,Z)} - 1 \mid X, Z \right] \left[\frac{D}{\pi(X,Z)} - 1\right] \\ &= \frac{D}{\pi(X,Z)}\left[Y - r(X,\theta)\right] - \left[E(Y \mid X,Z) - r(X,\theta)\right] \frac{1-\pi(X,Z)}{\pi(X,Z)} \\ &\qquad \times \frac{\pi^2(X,Z)}{\pi(X,Z)\left[1-\pi(X,Z)\right]} \left[\frac{D}{\pi(X,Z)} - 1\right] \\ &= \frac{D}{\pi(X,Z)}\left[Y - r(X,\theta)\right] - \left[E(Y \mid X,Z) - r(X,\theta)\right] \left[\frac{D}{\pi(X,Z)} - 1\right] \\ &\triangleq \rho(Y,X,Z,\theta). \end{aligned}$$

Model (16) then becomes

$$\begin{cases} E[\rho(Y,X,Z,\theta) \mid X] = 0 \\ \\ E\left[\dfrac{D}{\pi(X,Z)} - 1 \mid X, Z \right] = 0, \end{cases} \tag{17}$$

where θ can be efficiently estimated from the first conditional moment equations, since it contains all the information on θ in model (17) and is the same regardless a parametric or nonparametric form of $\pi(X, Z)$, as explained in section 4.2 of Hristache & Patilea (2011). The form of the efficient score and hence of the information on θ in this model can also be found in Tan (2011), Tsiatis (2007), Chen & Breslow (2004), Holcroft, Rotnitzky & Robins (1997).

In the particular case of no auxiliary variables Z, the orthogonalized equations model (17) is identical to the initial model (16), which can be interpreted as saying that a complete case analysis achieves efficiency in this situation.

4.3 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 $Q_\tau(Y \mid X, Z)$ of the always observed response Y given the regressor vectors Z (always observed) and X (observed iff $D = 1$) is assumed to be linear,

$$Q_\tau(Y \mid X, Z) = X'\beta_{1,\tau} + Z'\beta_{2,\tau}, \tag{18}$$

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

$$D \perp\!\!\!\perp (X, Y) \mid Z. \tag{19}$$

Taking in (5) $U = X$, $V = Y$, $W = Z$, $\rho_j(W, V, U, \beta_\tau) = (X', Z')' \left[\mathbb{1}_{Y - X'\beta_{1,\tau} - Z'\beta_{2,\tau} \le 0} - \tau\right]$ $\times a_j(U, W) \triangleq \rho(X, Y, Z, \beta_\tau) \times a_j(X, Z)$, $j \in \mathbb{N}$, where the family of functions $\{a_j\}_{j\in\mathbb{N}}$ spans $L^2(X, Z)$, the model defined by (18) and (19) can be written under the following equivalent form :

$$\begin{cases} E[D\,\rho(Y, X, Z, \beta_\tau) \mid X, Z] = 0 \\ \\ E\left[\dfrac{D}{\pi(Z)} - 1 \mid Z\right] = 0, \end{cases} \tag{20}$$

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 (18) 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), even if at a first sight it is not obvious that this additional assumptions changes the information bound of θ. To see why, suppose that only the parametric

form of the conditional expectation of X given Z is known and not the entire conditional law :

$$E(X \mid Z) = m(Z, \eta). \tag{21}$$

At the observational level, the model defined by (18), (19) and (21) is equivalently given by

$$\begin{cases} E[D\,\rho(Y, X, Z, \beta_\tau) \mid X, Z] = 0 \\ \\ E\left[\dfrac{D}{\pi(Z)} - 1 \mid Z\right] = 0 \\ \\ E\{D\,[X - m(Z, \eta)] \mid Z\} = 0. \end{cases} \tag{22}$$

The conclusion is the same : an efficient estimator for β_τ can be obtain from the first set of equations (meaning the use of complete data only) since we still have the orthogonality condition

$$E\{D\,\rho(Y, X, Z, \beta_\tau)\,D\,[X - m(Z, \eta)] \mid X, Z\} = 0.$$

It will be rather surprising if the parametric form of the conditional law of X given Z changes the efficiency bound on β_τ, since the only effect on the model defined by the equations (22) would be the replacement of $X - m(Z, \eta)$ with a countable set of functions of X and Z, known up to the ancillary parameter η.

A more general linear quantile regression model defined by (18) with missing data at random is considered in Chen, Wan & Zhou (2014). With their notations, we have

$$Y = Z'\,\theta(\tau) + \varepsilon, \qquad P(\varepsilon \leq 0 \mid Z) = \tau, \quad 0 < \tau < 1 \tag{23}$$

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

$$\begin{cases} E\left\{\dfrac{D}{\pi(X)}\,Z\,[\mathbb{1}_{Y - Z'\theta(\tau) \leq 0} - \tau] \mid Z\right\} = 0 \\ \\ E\left[\dfrac{D}{\pi(X)} - 1 \mid X\right] = 0. \end{cases} \tag{24}$$

The information bound for this model is given in Hristache & Patilea (2011). 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

$$\begin{cases} E\left\{\dfrac{D}{\pi(X)}\, Z\left[\mathbb{1}_{Y-Z'\theta(\tau)\leq 0}-\tau\right]\right\}=0 \\ \\ E\left[\dfrac{D}{\pi(X)}-1\,|\,X\right]=0. \end{cases} \tag{25}$$

The models (24) and (25) 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 (23) with missing data at random.

5 Conditional moment models with missing data at random

Consider a general semiparametric model defined by the conditional moment equations

$$E\left[\rho(\theta, Y, X, Z, V, Y^*, X^*)\mid X, X^*\right]=0, \tag{26}$$

where $\rho(\cdot)$ is a given vector-valued function. The random vectors Y, X, Z and V are always observed, while Y^* and X^* are observed only when $D=1$ and

$$(Y^*, X^*, V) \perp\!\!\!\perp D \mid Y, X, Z.$$

The equation (26) is a particular case of the system of equations in condition c) in Theorem 1 considered with an infinite set J, constant functions $h_j(\cdot)$, $W'=(Y', X', Z')$, $U'=(Y^{*\prime}, X^{*\prime})$. Moreover,

$$\pi(W)=P(D=1\mid W)=P(D=1\mid Y, X, Z, V, Y^*, X^*).$$

By Theorem 1 and 2, one can write the equivalent model

$$\begin{cases} E\left\{\dfrac{D}{\pi(W)}\,\rho(\theta, Y, X, Z, V, Y^*, X^*)\mid X, X^*\right\}=0 \\ \\ E\left[\dfrac{D}{\pi(W)}-1\,|\,W\right]=0. \end{cases} \tag{27}$$

This remark generalizes the GMM equivalence stated in Theorem 2.1 of Graham (2011) who considered the particular case of constant X, X^*.

Let us point out that for what follows in this section, one could consider the more general situation with several moment equations as in (26) and different conditioning variables. Such a situation was considered, for instance, in Hristache & Patilea (2011). For the sake of readability we stay with the present framework.

At our best knowledge, there is no general strategy to derive an asymptotically efficient estimator of θ in the model defined by (27). Except some particular cases, a result on the efficiency bound in this model is not available. In the case where all the conditioning variables are observed, *i.e.*, X^* is a constant random variable, one obtains a sequential moment restrictions model

$$\begin{cases} E\left\{\dfrac{D}{\pi(Y,X,Z)}\,\rho(\theta,Y,X,Z,V,Y^*) \mid X\right\} = 0 \\ \\ E\left[\dfrac{D}{\pi(Y,X,Z)} - 1 \,|\, Y,X,Z\right] = 0 \end{cases} \tag{28}$$

which is a particular case of the models considered by Ai & Chen (2012). Under some technical assumptions, especially on the separability (independence) of finite and infinite dimensional parameters, one can obtain the efficiency bound of model (28) and an asymptotically efficient estimator of θ. This strategy is also applicable for the linear quantile regression model with missing data of Wei, Ma & Carroll (2012) given by the equations (18) and (19), equivalently written under the form (20). This means that in the quantile framework of Wei, Ma & Carroll (2012), one could efficiently estimate the parameter $\theta = (\beta'_{1,\tau}, \beta'_{2,\tau})'$, even when $\pi(\cdot)$ is unknown and has to be estimated nonparametrically. On the other hand, let us point out that one could no longer use the general results of Ai & Chen (2012) in the linear quantile regression model (24) with missingness affecting also the covariates and unknown propensity score $\pi(\cdot)$. More details on the reason are provided below. In this case one only disposes of root$-n$ consistent estimators for $\theta = \theta(\tau)$. See Chen, Wan & Zhou (2014) and Ai & Chen (2007) for such estimators.

Let us detail the more complicated case of missing conditioning variables. Results on efficiency bounds for models such as (27) defined by conditional moments with different conditioning variables, but not sequential moments as in Ai & Chen (2012), are available only if the unknown parameters entering the equations are finite dimensional. See Hristache & Patilea (2011). Here this corresponds to a parametric form $\pi(w) = p(w,\alpha) = p(y,x,z,\alpha)$ of the propensity score, where $\alpha \in \mathbb{R}^{d_\alpha}$ and p has a known form. One then obtains from (27)

$$\begin{cases} E\left\{\dfrac{D}{p(Y,X,Z,\alpha)}\,\rho(\theta,Y,X,Z,V,Y^*,X^*) \mid X,X^*\right\} = 0 \\ \\ E\left[\dfrac{D}{p(Y,X,Z,\alpha)} - 1 \,|\, Y,X,Z\right] = 0. \end{cases} \tag{29}$$

As it is explained in Hristache & Patilea (2011), a possible way of deriving the efficiency bound and constructing approximately efficient estimators is to consider the sequence of decreasing models

$$\begin{cases} E\left\{ a_{(k)}(X, X^*) \otimes \dfrac{D}{p(Y, X, Z, \alpha)} \rho(\theta, Y, X, Z, V, Y^*, X^*) \right\} = 0 \\ E\left\{ b_{(k)}(Y, X, Z) \otimes \left[\dfrac{D}{p(Y, X, Z, \alpha)} - 1 \right] \right\} = 0, \end{cases} \tag{30}$$

where $\{a_k : k \in \mathbb{N}^*\} \subset L^2(X, X^*)$ and $\{b_k : k \in \mathbb{N}^*\} \subset L^2(Y, X, Z)$ are countable sets of squared integrable functions such that their linear spans are dense in $L^2(X, X^*)$ and $L^2(Y, X, Z)$ respectively,

$$\begin{aligned} a_{(k)}(X, X^*) &= (a_1(X, X^*), \ldots, a_k(X, X^*))' \in \mathbb{R}^k, \\ b_{(k)}(Y, X, Z) &= (b_1(Y, X, Z), \ldots, b_k(Y, X, Z))' \in \mathbb{R}^k, \qquad k \in \mathbb{N}^*, \end{aligned}$$

and $\otimes$ denotes the Kronecker product.

6 Semiparametric regression models with responses missing at random

Semiparametric regression models, particularly those aiming to handle the curse of dimensionality arising in nonparametric regression, have been recently studied at a great extend in the framework of missing data. See, for instance, Wang, Linton & Härdle (2004), Liang, Wang, Robins & Carroll (2004), Wang & Sun (2007), Liang (2008), Yang, Xue & Cheng (2009), Wang (2009), Wang, Shen, He & Wang (2010), Lai & Wang (2011), Fan, Härdle, Wang & Zhu (2013), Xue (2013), Chen & Van Keilegom (2013), Wang, Zhang & Härdle (2014).

Consider a generalized partially linear single-index regression model defined by

$$E\left[Y - \mu\left(X'\beta + g(Z'\gamma) \right) \mid X, Z \right] = 0, \tag{31}$$

where $(Y, X, Z) \in \mathbb{R} \times \mathbb{R}^{d_\beta} \times \mathbb{R}^{d_\gamma}$, (β, γ) is an unknown vector in $\mathbb{R}^{d_\beta} \times \mathbb{R}^{d_\gamma}$ with the first component equal to 1 and $g(\cdot)$ is an unknown univariate function and $\mu(\cdot)$ is a known link function. Assume first that the responses Y are missing at random :

$$Y \perp\!\!\!\perp D \mid X, Z, \tag{32}$$

and Y is observed whenever $D = 1$. The model defined by (31) and the MAR assumption (32) can then be equivalently written, at the observational level, under the form

$$\begin{cases} E\left\{D\left[Y - \mu\left(X'\beta + g(Z'\gamma)\right)\right] \mid X, Z\right\} = 0 \\ \\ E\left[\dfrac{D}{\pi(X,Z)} - 1 \mid X, Z\right] = 0. \end{cases} \tag{33}$$

Note that the two equations of the last system are orthogonal in the sense that

$$E\left\{D\left[Y - \mu\left(X'\beta + g(Z'\gamma)\right)\right]\left[\frac{D}{\pi(X,Z)} - 1\right] \mid X, Z\right\} = 0.$$

Then, the efficiency bound for (β, γ) in model (33), which coincides with the one in the model defined by the equations (31) and (32), is equal to the inverse of the information on (β, γ) in the model defined only by the conditional moment

$$E\left\{D\left[Y - \mu\left(X'\beta + g(Z'\gamma)\right)\right] \mid X, Z\right\} = 0.$$

In other words, the complete observations contain all the information on (β, γ), there is no information loss if the observations (Y, X, Z) for which Y is missing are deleted from the sample. This remark opens the door for alternative semiparametric estimation strategies in several models considered in the literature, see for instance Wang & Sun (2007), Lai & Wang (2011), Wang, Zhang & Härdle (2014). One could directly use semiparametric multiple-index estimators, such as proposed by Ichimura & Lee (1991), Hristache, Juditsky, Polzehl & Spokoiny (2001), Dalalyan, Juditsky & Spokoiny (2008).

In the case where missingness affects the covariates and the selection probability depends on Y, one could suitably rewrite the system (33) with different conditioning variables for the two equations. See section 5 for a related setup. The efficiency bound calculation in such situations remains an open problem.

7 A new look at double-robustness

Consider that the true value of a parameter $\theta \in \Theta \subset \mathbb{R}^d$ is the unique solution of the equations

$$E\left[\rho(\theta, Y, X)\right] = 0 \tag{34}$$

where $\rho(\cdot)$ takes values in $\mathbb{R}^d$. Moreover, assume the MAR condition

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

In the spirit of our equivalent model approach, we could write this model and the MAR condition under the form

$$\begin{cases} E\left[\dfrac{D}{\pi(X)}\ \rho(\theta, Y, X)\right] = 0 \\ \\ E\left[\dfrac{D}{\pi(X)} - 1 \mid X\right] = 0. \end{cases} \tag{35}$$

If one does not want to use nonparametric estimators for $\pi(\cdot)$, one has to specify a parametric missingness model $\pi(\cdot) = p(\cdot, \alpha)$, for some parameter $\alpha \in \mathcal{A} \subset \mathbb{R}^{d_\alpha}$. The logistic regression is a common choice for modeling $\pi(\cdot)$. Alternatively, θ and α can be jointly estimated from the following generalization (larger model) of the equations (35) :

$$\begin{cases} E\left[\dfrac{D}{p(X, \alpha)}\ \rho(\theta, Y, X)\right] = 0 \\ \\ E\left[\dfrac{D}{p(X, \alpha)} - 1 \mid X\right] = 0. \end{cases} \tag{36}$$

If one does not trust the model for the propensity score $\pi(\cdot)$, one could look for a warranty against misspecification. The usual way is to consider an unconditional moment restriction weaker than the second equation in (36). Such a moment restriction should preserve θ_0 as unique solution of the first equation in (36). A natural candidate for the unconditional moment restriction is the equation

$$E\left[\left(\frac{D}{p(X, \alpha)} - 1 \mid X\right) E[\rho(\theta, Y, X) \mid X]\right] = 0.$$

However, in general $E[\rho(\theta, Y, X) \mid X]$ has an unknown form and has to be estimated nonparametrically. In order to avoid once again nonparametric estimation, one could consider a 'working regression model' for the conditional expectation of $\rho(\theta, \cdot)$ given X. Let $m(\cdot, \theta, \eta)$, $\eta \in \mathbb{R}^{d_\eta}$ be such a regression model and let

$$E\left\{\frac{D}{p(X, \alpha)}\left[\rho(\theta, Y, X) - m(X, \theta, \eta)\right] \mid X\right\} = 0 \tag{37}$$

be the equation that takes into account the 'working regression model' and the missingness. Thus this equation has to be added to the model (36). If one does not trust the 'working regression model', again one has to replace the equation (37) by a weaker unconditional restriction, as for instance

$$E\left\{\frac{D}{p(X, \alpha)}\left[\rho(\theta, Y, X) - m(X, \theta, \eta)\right]\right\} = 0. \tag{38}$$

We can summarize everything as follows. One has the following equations based on correctly modeling $\pi(\cdot)$ and $E[\rho(\theta, Y, X) \mid X]$:

$$\begin{cases} E\left[\dfrac{D}{p(X,\alpha)}\ \rho(\theta, Y, X)\right] = 0 \\ E\left[\dfrac{D}{p(X,\alpha)} - 1 \mid X\right] = 0 \\ E\left\{\dfrac{D}{p(X,\alpha)}\left[\rho(\theta, Y, X) - m(X, \theta, \eta)\right] \mid X\right\} = 0. \end{cases} \tag{39}$$

If one does not trust the two parametric models for $\pi(\cdot)$ and $E[\rho(\theta, Y, X) \mid X]$, one could alternatively use the equations

$$\begin{cases} E\left[\dfrac{D}{p(X,\alpha)}\ \rho(\theta, Y, X)\right] = 0 \\ E\left[\left(\dfrac{D}{p(X,\alpha)} - 1\right)\ A(X)\right] = 0 \\ E\left\{\dfrac{D}{p(X,\alpha)}\left[\rho(\theta, Y, X) - m(X, \theta, \eta)\right]\ B(X)\right\} = 0. \end{cases} \tag{40}$$

The solution $(\theta^*, \alpha^*, \eta^*)$ obtained from this larger model could give a different value θ^* from the true value θ_0 of the parameter of interest θ if the model (40) is wrong. Would it be possible to choose the "instruments" $A(X)$ and $B(X)$ (A and B can also depend on the parameters θ, α and η) such that θ can be consistently estimated even if both models for $\pi(\cdot)$ and $E[\rho(\theta, Y, X) \mid X]$ are wrong ? The answer at this question seems to be rather negative, but several estimation procedures exist in the literature protecting against the misspecification of either $\pi(\cdot)$ or $E[\rho(\theta, Y, X) \mid X]$ when the other assumption holds true, that is, estimate θ in an intermediate model between (39) and (40). For this, we have to look for instruments $A(X)$ and $B(X)$ such that estimating θ from (40) leads to a consistent estimator of θ_0 if at least one of the two last sets of conditions in (40) holds true conditionally on X, *i.e.*, if at least one of the following models is correct :

$$\begin{cases} E\left[\dfrac{D}{p(X,\alpha)}\ \rho(\theta, Y, X)\right] = 0 \\ E\left[\left(\dfrac{D}{p(X,\alpha)} - 1\right)\ A(X)\ \mid\ X\right] = 0 \\ E\left\{\dfrac{D}{p(X,\alpha)}\left[\rho(\theta, Y, X) - m(X, \theta, \eta)\right]\ B(X)\right\} = 0 \end{cases} \tag{41}$$

or

$$\begin{cases} E\left[\dfrac{D}{p(X,\alpha)}\ \rho(\theta,Y,X)\right]=0 \\ \\ E\left[\left(\dfrac{D}{p(X,\alpha)}-1\right)\ A(X)\right]=0 \\ \\ E\left\{\dfrac{D}{p(X,\alpha)}\left[\rho(\theta,Y,X)-m(X,\theta,\eta)\right]\ B(X)\ \mid\ X\right\}=0. \end{cases} \tag{42}$$

In other words, we look for an estimator of θ_0 in the model defined as the union of the models (41) and (42). Therefore, $B(X)$ should be determined such that θ_0 is the unique solution of the system (41), since this model is the same irrespective of the expression of $A(X)$, while $A(X)$ is determined such that θ_0 is the unique solution of the system (42). The "classical" choice, at least in the regression setup (and in particular for estimating the mean of Y), corresponds to $B(X)=1$ and $A(X)=m(X,\theta,\eta)$ and a rewriting of the first equation in (41) or (42) as a combination of the first two sets of equations under the form

$$E\left\{\frac{D}{p(X,\alpha)}\ \rho(\theta,Y,X)-\left[\frac{D}{p(X,\alpha)}-1\right]\ m(X,\theta,\eta)\right\}=0. \tag{43}$$

In fact, the doubly robust estimators in the literature are, in most of the cases, defined as solutions of equations of the form (43), where a preliminary (conditional) maximum likelihood estimator for α is used. One exception at this type of construction of doubly robust estimators can be found in Graham, Campos De Xavier Pinto & Egel (2012), where θ and α are jointly estimated by a method of moments.

To see that with $A(X)=m(X,\theta,\eta)$ and $B(X)=1$ we obtain indeed a doubly robust estimator, first note that in the system (41), for an arbitrary choice of $A(X)$ the validity of the missingness mechanism $P(D=1\,|\,X)=p(X,\alpha)$ and the subsequent MAR assumption implied by the second equation leads to

$$E\left[\frac{D}{p(X,\alpha)}\ \rho(\theta,Y,X)\right]=E\left[\frac{E(D|X)}{p(X,\alpha)}\ \rho(\theta,Y,X)\right]=E\left[\rho(\theta,Y,X)\right]=0\quad\text{iff}\quad\theta=\theta_0.$$

This is true whether or not $m(X,\theta,\eta)$ correctly specifies $E[\rho(\theta,Y,X)\mid X]$ and for any $B(X)$, so that $\theta=\theta_0$ is the unique solution (in θ) for the system (41) if the missingness model is correct. For the second part of the doubly robustness, note that with $A(X)=m(X,\theta,\eta)$ if we take the difference of the first two (sets of) equations in the system (42)

we obtain

$$E\left\{\frac{D}{p(X,\alpha)}\,\rho(\theta,Y,X)-\left[\frac{D}{p(X,\alpha)}-1\right]\,m(X,\theta,\eta)\right\}$$

$$=E\left\{\frac{D}{p(X,\alpha)}\,\left[\rho(\theta,Y,X)-m(X,\theta,\eta)\right]+m(X,\theta,\eta)\right\}$$

$$=E\left[0+m(X,\theta,\eta)\right]=E\left\{E[\rho(\theta,Y,X)\mid X]\right\}$$

$$=E\left[\rho(\theta,Y,X)\right]=0\quad\text{iff}\quad\theta=\theta_0,$$

the second equality being obtained under the correct assumption of the working regression model $E[\rho(\theta,Y,X)\mid X]=m(X,\theta,\eta)$ stated by the third set of equations of the system (42).

One conclusion of the previous paragraph is that the double robustness is obtained for any choice of $B(X)$. Here is another way to explain why this is true. From any of the models defined by (40), (41) or (42) we obtain the following general form for an estimating equation for θ at the observational level (with A_1 and B_1 of right dimensions) :

$$E\left\{\frac{D}{p(X,\alpha)}\,\rho(\theta,Y,X)-\left[\frac{D}{p(X,\alpha)}-1\right]\,A_1(X,\theta,\alpha,\eta)\right.$$
$$\left.+\frac{D}{p(X,\alpha)}\left[\rho(\theta,Y,X)-m(X,\theta,\eta)\right]\,B_1(X,\theta,\alpha,\eta)\right\}=0. \qquad (44)$$

Rearranging the terms, it can be written as :

$$E\left\{\rho(\theta,Y,X)+\left[\frac{D}{p(X,\alpha)}-1\right]\,\left[\rho(\theta,Y,X)-A_1(X,\theta,\alpha,\eta)\right]\right.$$
$$\left.+\frac{D}{p(X,\alpha)}\left[\rho(\theta,Y,X)-m(X,\theta,\eta)\right]\,B_1(X,\theta,\alpha,\eta)\right\}=0. \qquad (45)$$

The model for complete data being defined as

$$E\left[\rho(\theta,Y,X)\right]=0\quad\text{iff}\quad\theta=\theta_0,$$

doubly robustness will be obtained if we have, for any θ,

$$E\left\{\left[\frac{D}{p(X,\alpha)}-1\right]\,\left[\rho(\theta,Y,X)-A_1(X,\theta,\alpha,\eta)\right]\right.$$
$$\left.+\frac{D}{p(X,\alpha)}\left[\rho(\theta,Y,X)-m(X,\theta,\eta)\right]\,B_1(X,\theta,\alpha,\eta)\right\}=0, \qquad (46)$$

whenever $E(D|X) = p(X,\alpha)$ or $E\left[\rho(\theta,Y,X)\,|\,X\right] = m(X,\theta,\eta)$. In the case where $E(D|X) = p(X,\alpha)$, θ is part of the solution (θ,α,η) of the equations (46) for any choice of $A_1(X,\theta,\alpha,\eta)$, if $B_1(X,\theta,\alpha,\eta)$ is such that the third set of equations in (40) is satisfied. Such a condition on B_1 essentially requires only the right number of equations for estimating η for any fixed value of θ and α. On the other hand, in the case $E\left[\rho(\theta,Y,X)\,|\,X\right] = m(X,\theta,\eta)$, one should have, for any θ,

$$E\left\{\left[\frac{E(D|X)}{p(X,\alpha)} - 1\right]\,\left[\rho(\theta,Y,X) - A_1(X,\theta,\alpha,\eta)\right]\right\} = 0,$$

which entails

$$A_1(X,\theta,\alpha,\eta) = m(X,\theta,\eta). \tag{47}$$

In conclusion, in order to obtain a doubly robust estimator for θ_0 in the model defined by (39) it is necessary and sufficient to estimate θ_0 in the model (40) with $A(X)$ having the first components given by (47) and $A(X)$ and $B(X)$ chosen such that the parameter (θ,α,η) is identifiable in the model defined by (40).

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

The model defined by the regression-like equation

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

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 :

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

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.

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 (2011))

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

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 (2011), 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.

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

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

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

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

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

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

9 Why weighting using estimated selection probabilities is better than weighting using the true selection probabilities

Assume that the selection probability $\pi(\cdot)$ has a parametric form : $\pi(\cdot) = p(\cdot,\gamma)$, with γ an unknown parameter (not necessarily finite dimensional !). It is known that, in many situations if not always, estimating γ leads to a more efficient estimator for θ than estimating θ from

$$E\left[\frac{D}{p(W,\gamma)}\ \rho_j(\theta,W,U,h_j(\theta,W,U))\right] = 0, \quad \forall j \in J,$$

with known γ. See, among others, Prokhorov & Schmidt (2009), Wooldridge (2002) or Hitomi, Nishiyama & Okui (2008) for examples, discussions and further references on this "puzzling phenomenon".

The problem does not really come from estimating γ or not, but from the fact that for known γ one forgets that the equation

$$E\left[\frac{D}{p(W,\gamma)} - 1 \mid W\right] = 0$$

is still part of the (equivalent) definition of the model. It does not play any role in estimating γ when it is supposed to be known, but gives different estimating equations for θ to be estimated efficiently. Namely, $\dfrac{D}{\pi}\ \rho_j$ must be replaced by the "orthogonalized" version

$$\frac{D}{\pi}\ \rho_j - E\left[\frac{D}{\pi}\ \rho_j \cdot \left(\frac{D}{\pi} - 1\right)\ \mid W\right]\ V^{-1}\left(\frac{D}{\pi} - 1 \mid W\right) \cdot \left(\frac{D}{\pi} - 1\right)$$

in the second set of equations (5). Writing the model in the form (5) implies that, without loss of generality (and replacing if necessary U with $\widetilde{U}$ which are not distinguishable at the observational level), the MAR assumption

$$P(D=1 \mid U,W) = \pi(W) = p(W,\gamma)$$

is satisfied. We then obtain :

$$\begin{aligned} E\left[\frac{D}{\pi}\,\rho_j \cdot \left(\frac{D}{\pi}-1\right) \mid W\right] &= E\left\{E\left[\frac{D}{\pi}\,\left(\frac{D}{\pi}-1\right) \mid W,U\right]\cdot \rho_j \mid W\right\} \\ &= E\left[E\left(\frac{D^2}{\pi^2}-\frac{D}{\pi} \mid W,U\right)\cdot \rho_j \mid W\right] = E\left[\left(\frac{1}{\pi}-1\right)\cdot \rho_j \mid W\right] \\ &= \left(\frac{1}{\pi}-1\right)\cdot E\left(\rho_j \mid W\right) = \frac{1-\pi}{\pi}\,E\left(\rho_j \mid W\right), \end{aligned}$$

$$V\left(\frac{D}{\pi}-1 \mid W\right) = V\left(\frac{D}{\pi} \mid W\right) = \frac{1}{\pi^2}\,V(D|W) = \frac{1}{\pi^2}\,\pi(1-\pi) = \frac{1-\pi}{\pi},$$

which gives

$$\begin{aligned} &\frac{D}{\pi}\,\rho_j - E\left[\frac{D}{\pi}\,\rho_j \cdot \left(\frac{D}{\pi}-1\right) \mid W\right]\; V^{-1}\left(\frac{D}{\pi}-1 \mid W\right)\cdot\left(\frac{D}{\pi}-1\right) \\ &\quad = \frac{D}{\pi}\,\rho_j - \frac{1-\pi}{\pi}\,E\left(\rho_j \mid W\right)\cdot\frac{\pi}{1-\pi}\cdot\left(\frac{D}{\pi}-1\right) \\ &\quad = \frac{D}{\pi}\,\rho_j - E\left(\rho_j \mid W\right)\,\left(\frac{D}{\pi}-1\right). \end{aligned}$$

The system (5) then becomes

$$\begin{cases} E\left[\dfrac{D}{\pi(W)}-1 \mid W\right] = 0, \\ E\left[\dfrac{D}{\pi}\,\rho_j - E\left(\rho_j \mid W\right)\,\left(\dfrac{D}{\pi}-1\right)\right] = 0, \quad \forall j \in J, \end{cases} \tag{48}$$

with $\pi = \pi(W) = p(W,\gamma)$. Estimating θ efficiently from the second set of these equations, i.e.

$$E\left[\frac{D}{\pi}\,\rho_j - E\left(\frac{D}{\pi}\,\rho_j \mid W\right)\,\left(\frac{D}{\pi}-1\right)\right] = 0, \quad \forall j \in J,$$

gives now the same result irrespective of knowing or not knowing γ. This is clearly true when J is finite. When J is not finite, one expects that the Fisher information in the model (48) is the supremum of the Fisher information over all its submodels obtained by replacing J by any finite set $J' \subset J$ in (48). This would be true under a so-called *spanning condition*, see Newey (2004) and Hristache & Patilea (2011). The validity of such a spanning condition in the present framework will be investigated elsewhere.

Note that on the basis of model (5), for a finite number of equations and where only finite dimensional parameters enter in the defining equations of the model, this interpretation of the fact that it could be better to estimate auxiliary parameters instead of using their known value was already given in Prokhorov & Schmidt (2009). Here, using our general equivalence result, we show that this explanation is valid for the initial moment restrictions models with missing data under the MAR assumption.

10 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 shown in the following.

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

$$(U, V) \perp\!\!\!\perp D \mid W. \tag{49}$$

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

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

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. One could argue that generating substitute observations from a parametric or nonparametric estimate of the law $\mathcal{L}(\widetilde{U} \mid W, V, D = 0)$ allows to use the additional information contained in the observations of $(W', V')'$ when U is not observed.

The equivalence established by Theorem 1 and 2 for models defined by moment restrictions, implies that *all* the information on the parameter θ in the initial model under the MAR assumption (49) is contained in the model defined by the equations (5). Let us point out that the last equation of the model (5) 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.

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.

Cao, W., Tsiatis, A. A. & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–734.

Chamberlain, G. (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics* **10**, 20–26.

Chen, J. & Breslow, N. E. (2004). Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **32**, pp. 359–372.

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.

Dalalyan, A. S., Juditsky, A. & Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. *J. Mach. Learn. Res.* **9**, 1647–1678.

Fan, Y., Härdle, W. K., Wang, W. & Zhu, L. (2013). Composite quantile regression for the single-index model. Tech. rep., Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät.

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

Graham, B. S., Campos De Xavier Pinto, C. & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies* **79**, 1053–1079.

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

Hitomi, K., Nishiyama, Y. & Okui, R. (2008). A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters. *Econometric Theory* **24**, 1717–1728.

Holcroft, C. A., Rotnitzky, A. & Robins, J. M. (1997). Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference* **65**, 349 – 374.

Hristache, M., Juditsky, A., Polzehl, J. & Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics* **29**, 1537–1566.

Hristache, M. & Patilea, V. (2011). Semiparametric efficiency bounds for seemingly unrelated conditional moment restrictions. ArXiv:1111.6428.

Ichimura, H. & Lee, L.-F. (1991). Semiparametric least squares estimation of multiple index models: single equation estimation. In *Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge*.

Lai, P. & Wang, Q. (2011). Partially linear single-index model with missing responses at random. *Journal of Statistical Planning and Inference* **141**, 1047–1058.

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.

Liang, H. (2008). Generalized partially linear models with missing covariates. *Journal of multivariate analysis* **99**, 880–895.

Liang, H., Wang, S., Robins, J. M. & Carroll, R. J. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association* **99**.

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.

Newey, W. K. (2004). Efficient semiparametric estimation via moment restrictions. *Econometrica* **72**, 1877–1897.

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.

Rotnitzky, A., Lei, Q., Sued, M. & Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–456.

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.

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

Wang, Q. & Sun, Z. (2007). Estimation in partially linear models with missing responses at random. *Journal of Multivariate Analysis* **98**, 1470–1493.

Wang, Q., Zhang, T. & Härdle, W. K. (2014). An extended single index model with missing response at random. Tech. rep., Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.

Wang, Q.-H. (2009). Statistical estimation in partial linear models with covariate data missing at random. *Annals of the Institute of Statistical Mathematics* **61**, 47–84.

Wang, Y., Shen, J., He, S. & Wang, Q. (2010). Estimation of single index model with missing response at random. *Journal of Statistical Planning and Inference* **140**, 1671–1690.

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

Wooldridge, J. (2002). Inverse probability weighted $M-$estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal* **1**, 117–139.

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

Xue, L. (2013). Estimation and empirical likelihood for single-index models with missing data in the covariates. *Computational Statistics & Data Analysis* **68**, 82–97.

Yang, Y., Xue, L. & Cheng, W. (2009). Empirical likelihood for a partially linear model with covariate data missing at random. *Journal of Statistical Planning and Inference* **139**, 4143–4153.