Econometric Theory, **32**, 2016, 917–946.
doi:10.1017/S0266466615000080

SEMIPARAMETRIC EFFICIENCY BOUNDS FOR CONDITIONAL MOMENT RESTRICTION MODELS WITH DIFFERENT CONDITIONING VARIABLES

MARIAN HRISTACHE
CREST (Ensai)

VALENTIN PATILEA
CREST (Ensai) and Bucharest University of Economic Studies

This paper addresses the problem of semiparametric efficiency bounds for conditional moment restriction models with different conditioning variables. We characterize such an efficiency bound, that in general is not explicit, as a limit of explicit efficiency bounds for a decreasing sequence of unconditional (marginal) moment restriction models. An iterative procedure for approximating the efficient score when this is not explicit is provided. Our theoretical results provide new insight for the theory of semiparametric efficiency bounds literature and open the door to new applications. In particular, we investigate a class of regression-like (mean regression, quantile regression,...) models with missing data, an example of a supply and demand simultaneous equations model and a generalized bivariate dichotomous model.

1. THE MODEL

Conditional moment restriction models represent a large class of statistical models. Seemingly unrelated nonlinear regressions, see Gallant (1975) and Müller (2009), seemingly unrelated quantile regressions, see Jun and Pinske (2009), regression models with missing data, see Robins, Rotnitzky, and Zhao (1994) and Tsiatis (2006), are only a few examples and related contributions. Many other references and examples of statistical and econometric models that could be stated as conditional moment restriction models are provided in Ai and Chen (2012) and Hansen (2008).

Valentin Patilea acknowledges financial support from the Romanian National Authority for Scientific Research, CNCS-UEFISCDI, Project PN-II-ID-PCE-2011-3-0893. Address correspondence to Marian Hristache, 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; e-mail: marian.hristache@ensai.fr.
Author's name has been corrected since original publication. An erratum notice detailing this change was also published.

© Cambridge University Press 2015

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

In this paper, we address the problem of calculating semiparametric efficiency bounds in models defined by several conditional moment restrictions with possibly different conditioning variables. More formally, the sample under study consists of independent copies of a random vector $Z \in \mathcal{Z} \subset \mathbb{R}^q$. Let J be some positive integer that is fixed in the following. For any $j \in \{1, \ldots, J\}$, let $X^{(j)}$ be a random q_j-dimensional subvector of Z, where $0 \leq q_j < q$. Let $g_j : \mathbb{R}^q \times \mathbb{R}^d \to \mathbb{R}^{p_j}$, $j \in \{1, \ldots, J\}$, denote given functions of Z and the unknown parameter $\theta \in \Theta \subset \mathbb{R}^d$. The semiparametric model we consider is defined by the conditional moment restrictions

$$\mathbb{E}\left[g_j(Z, \theta) \mid X^{(j)} \right] = 0, \quad j = 1, \ldots, J, \quad \text{almost surely (a.s.).} \tag{1}$$

Such models are common in the econometric literature, see for instance Wooldridge (2010, eqn. 14.33 and 14.36). As usual, the notation $g_j(Z, \theta)$ does not necessarily mean that the function g_j depends on all the components of Z. It is assumed that the d-dimensional parameter θ is identified by the conditional restrictions, which means there exists a unique value θ_0 such that the true law of Z satisfies equations (1). By definition, $X^{(j)}$ is a constant random variable when $q_j = 0$, and hence the conditional expectation given $X^{(j)}$ is the marginal expectation.

Particular cases of this model have been extensively studied in the literature. For $J = 1$ and $q_1 = 0$, we obtain a model defined by an unconditional set of moment equations

$$\mathbb{E}[g(Z, \theta)] = 0.$$

Hansen (1982) considered the class of GMM estimators and showed how to construct an optimal one in that class. Its asymptotic variance equals the semiparametric efficiency bound obtained by Chamberlain (1987).

The GMM method naturally extends to models defined by conditional moment equations, corresponding to the case $J = 1$ and $q_1 > 0$ in our setting, *i.e.,*

$$\mathbb{E}[g(Z, \theta) \mid X] = 0 \quad \text{(a.s.)}.$$

From a mathematical point of view, such a model is equivalent to the intersection of the models of the form

$$\mathbb{E}[a(X)\ g(Z, \theta)] = 0,$$

where $a(X)$ is an arbitrary, conformable, random matrix whose entries are square integrable. In the econometric literature, $a(X)$ is referred to as a matrix of *instruments*. The supremum of the information on θ_0 in these models yields the semiparametric Fisher information on θ_0 in the conditional equation model, obtained by Chamberlain (1992a). It is also the information on θ_0 for the unconditional moment equation

$$\mathbb{E}\left[a^*(X)\ g(Z, \theta) \right] = 0,$$

with properly chosen 'optimal' instruments $a^*(X)$.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

A further generalization, which can also be written in the form (1), is given by a sequential (nested) moment restrictions model, in which the $\sigma-$fields generated by the conditioning vectors satisfy the condition $\sigma\left(X^{(1)}\right) \subset \sigma\left(X^{(2)}\right) \subset \cdots \subset \sigma\left(X^{(J)}\right)$. For the expression of the semiparametric efficiency bound in the sequential case, see Chamberlain (1992b) and Ai and Chen (2012); see also Hahn (1997) and Ahn and Schmidt (1999) and references therein for examples of applications. It turns out that once again, the information on θ_0 can be obtained by taking the supremum of the information on θ_0 in the following unconditional models:

$$\mathbb{E}\left[a_j\left(X^{(j)}\right)\, g_j\left(Z,\theta\right)\right]=0, \quad j=1,\ldots,J,$$

where the number of lines of the matrices a_j is fixed and equal to the dimension of θ and the supremum is attained for a suitable choice $a_1^*\left(X^{(1)}\right),\ldots,a_J^*\left(X^{(J)}\right)$ of optimal instruments. The reason why this happens in the case with nested $\sigma-$fields is the fact that the model of interest can be written as the decreasing limit of a sequence of models for which a so-called *'spanning condition'*, similar to that considered in Newey (2004), holds and the limit of the corresponding efficient scores has an explicit solution.

In this paper, we show that the information on θ_0 in model (1) can be obtained as the limit of the information on θ_0 in a decreasing sequence of unconditional moment models of the form

$$\mathbb{E}\left[a_j^{(k)}\left(X^{(j)}\right)\, g_j\left(Z,\theta\right)\right]=0, \quad j=1,\ldots,J, \quad k=1,2,\ldots, \tag{2}$$

where the numbers of lines in the matrices $a_j^{(k)}$ increase to infinity with k. To the best of our knowledge this result is new. It provides theoretical support for a natural solution that could be used in practice: replace the model (1) by a large number of unconditional moment conditions as in equation (2) in order to approach efficiency. Herein, we also propose an alternative route for approximating the efficiency bound. More precisely, we give a general method to approximate the efficient score, which in most situations does not have an explicit form as in the aforementioned examples. In particular, our general approach for approximating the efficient score sheds new light on the functional equations used to characterize the efficient score in the regression model with unobserved explanatory variables in Robins et al. (1994); see also Tsiatis (2006) and Tan (2011).

To summarize, our theoretical results provide new insight for the theory of semiparametric efficiency bounds literature and open the door to new applications, in particular for nonlinear, simultaneous equations models and in missing data contexts. Quantile regression models with missing covariables is one such example of a new application that could be treated in our theoretical framework. This case of practical interest is different and more difficult to handle than the quantile regression with missing responses as considered by Wei, Ma, and Carroll (2012); see also Müller and van Keilegom (2012). On the other hand, our methodology allows nonlinear simultaneous equations models to be investigated with greater generality.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

The paper is organized as follows. Section 2 contains our main results. We show that under a suitable 'spanning condition' on the tangent spaces, the semiparametric Fisher information in model (1) can be obtained as the limit of the efficiency bounds for a decreasing sequence of models. In Section 3, we propose a 'backfitting' procedure, for computing the projection of the score onto the tangent space of the model. With an approximation of the efficient score in hand, we suggest a general method for constructing asymptotically efficient estimators. In Section 4, we illustrate the utility of our theoretical results for four classes of models: sequential (nested) conditional models, regression-like models with missing data, simultaneous equations models, and generalized dichotomous models. Some technical results and proofs are left to the Appendix.

2. THE MAIN RESULTS

Let us introduce some notation and definitions, see also van der Vaart (1998, Sect. 25.2 and 25.3). Given a sample space $\mathcal{Z}$ and a probability P on the sample space, we denote by $L^2(P)$ the usual Hilbert space of measurable real-valued functions that are square integrable with respect to P. For $\mathcal{H}$ a Hilbert space and $\mathcal{S} \subset \mathcal{H}$, let $\overline{S}$ denote the closure of $\mathcal{S}$ in $\mathcal{H}$. The statistical models on the sample space $\mathcal{Z}$, are denoted by $\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2...$ A statistical model is a collection of probability measures defined by their densities with respect to some fixed dominating measure on the sample space. Herein, the vectors are column matrices and $A \in \mathbb{R}^r \times \mathbb{R}^s$ means A is an $r \times s-$matrix with random elements, if not stated differently. $\mathbb{E}(A)$ denotes the expectation of A. When $A \in \mathbb{R}^r$, $\mathbb{V}(A)$ denotes its variance matrix. For $A \in \mathbb{R}^r \times \mathbb{R}^r$, $\mathbb{E}^{-1}(A)$ denotes the inverse of the square matrix $\mathbb{E}(A)$, whenever the inverse exists. Finally, for a square matrix A, let A^- denote a generalized inverse, for instance the Moore-Penrose pseudoinverse.

2.1. Efficiency Bound

The main idea we follow to derive the semiparametric efficiency bound for the parameter θ_0, is to transform the finite number of conditional moment restrictions (1) in a countable number of unconditional (marginal) moment restrictions. Next, for any finite subset of these unconditional moment restrictions, one could easily obtain the Fisher information bound. Eventually, one may expect to obtain the semiparametric efficiency bound for the model (1) as the limit of the efficiency bounds for a decreasing sequence of models defined by an increasing sequence of finite subsets of unconditional moment restrictions. Remark A.1 in the Appendix proves that in general, this intuition is not correct. However, the subsequent Lemma A.1 states that this intuition becomes correct under the additional 'spanning condition' formally stated in equation (A.2) in the Appendix.

Let us introduce some more notation. For a vector a, we denote by a' its transpose. If $\zeta : \mathcal{Z} \times \Theta \to \mathbb{R}^m$, $m \geq 1$, is some given function of Z and θ and X is

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

some subvector of Z, we denote

$$\partial_{\theta'}\mathbb{E}[\zeta \mid X] = \partial_{\theta'}\mathbb{E}[\zeta(Z,\theta_0) \mid X] = \frac{\partial}{\partial\theta'}\mathbb{E}[\zeta(Z,\theta) \mid X]\Big|_{\theta=\theta_0} \in \mathbb{R}^d \times \mathbb{R}^m, \tag{3}$$

when such derivatives of the map $\theta \mapsto \mathbb{E}[\zeta(Z,\theta) \mid X]$ exist. A similar notation will be used with the conditional expectation $\mathbb{E}(\cdot \mid X)$ replaced by the marginal (unconditional) expectation with respect to the law of Z. Let us point out that the maps $\theta \mapsto \zeta(z,\theta)$ may not be everywhere differentiable. Next, let us define

$$\underline{g} = \left(g_1', \ldots, g_J'\right)' \in \mathbb{R}^p, \quad \text{where } p = p_1 + \cdots + p_J,$$

and let $\underline{X}$ denote the vector of all components of Z contained in the subvectors $X^{(j)}$, $j = 1, \ldots, J$. Let P_0 be the true law of the vector Z and $P_{X^{(j)}}$ the law of $X^{(j)}$, $j = 1, \ldots, J$.

For the purpose of transforming conditional moments in unconditional versions, let us consider J countable sets of square integrable functions $\mathcal{W}^{(j)} = \left\{w_k^{(j)} : k \in \mathbb{N}^*\right\} \subset L^2\left(P_{X^{(j)}}\right)$ such that $\overline{\text{lin}}\mathcal{W}^{(j)} = L^2\left(P_{X^{(j)}}\right)$, that is the linear span of $\mathcal{W}^{(j)}$ is dense in $L^2\left(P_{X^{(j)}}\right)$, for $1 \le j \le J$. For any $k \in \mathbb{N}^*$, let

$$\underline{g}_k^w(Z,\theta) = \begin{pmatrix} w_{(k)}^{(1)}(X^{(1)}) \otimes g_1(Z,\theta) \\ w_{(k)}^{(2)}(X^{(2)}) \otimes g_2(Z,\theta) \\ \vdots \\ w_{(k)}^{(J)}(X^{(J)}) \otimes g_J(Z,\theta) \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product and

$$w_{(k)}^{(j)}\left(X^{(j)}\right) = \left(w_1^{(j)}\left(X^{(j)}\right), \ldots, w_k^{(j)}\left(X^{(j)}\right)\right)' \in \mathbb{R}^k.$$

Let $I_{\theta_0}^{(k)}$ be the Fisher information on θ_0 in the model

$$\mathbb{E}\left[\underline{g}_k^w(Z,\theta)\right] = 0, \tag{4}$$

that is

$$I_{\theta_0}^{(k)} = \left[\partial_{\theta'}\mathbb{E}\left(\underline{g}_k^w(Z,\theta_0)\right)\right]' \mathbb{V}^- \left[\underline{g}_k^w(Z,\theta_0)\right] \partial_{\theta'}\mathbb{E}\left(\underline{g}_k^w(Z,\theta_0)\right).$$

See Chamberlain (1987) and Newey (2001), see also Chen and Pouzo (2009) for the nonsmooth case.

We can now state the main result of the paper.

THEOREM 1. *Under the Assumptions T and SP in Section 2.2, the information bound I_{θ_0} on θ_0 at P_0 in model (1) is given by*

$$I_{\theta_0} = \lim_{k\to\infty} I_{\theta_0}^{(k)},$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

where, for any $k \in \mathbb{N}^*$, $I_{\theta_0}^{(k)}$ *is the Fisher information on* θ_0 *in the model defined as in (4).*

In the general theory of efficiency bounds, the semiparametric Fisher information on a finite-dimensional parameter in a semiparametric model is the infimum of the Fisher information over all its parametric submodels; see for instance Newey (1990). For models defined by conditional moment equations, Theorem 1 shows that the same semiparametric Fisher information can be obtained alternatively as the lower limit of the semiparametric Fisher information in a sequence of decreasing supra-models. The main reason for this is that with such a decreasing sequence of supra-models, the 'spanning condition' (A.2) holds true. Moreover, since $L^2(P_0)$ is a separable Hilbert space, Theorem 1 can be restated under the following equivalent form.

Let

$$\mathcal{B} = \left\{ \left(\omega_1 \left(X^{(1)} \right), \ldots, \omega_J \left(X^{(J)} \right) \right) : \ \omega_{j,lk} \in L^2 \left(P_{X^{(j)}} \right), \right. \\ \left. 1 \le l \le d, \ 1 \le k \le p_j, \ 1 \le j \le J \right\};$$

that is, for any j, $\omega_j\left(X^{(j)}\right)$ is a $d \times p_j$ matrix of functions whose entries $\omega_{j,lk}\left(X^{(j)}\right)$ are square integrable, for $1 \le l \le d$ and $1 \le k \le p_j$. For a $d \times p-$ matrix $\omega = \omega(\underline{X}) = \left(\omega_1\left(X^{(1)}\right), \ldots, \omega_J\left(X^{(J)}\right)\right) \in \mathcal{B}$, let $I_{\theta_0}(\omega)$ be the Fisher information on θ_0 in the model defined by the marginal moment restrictions

$$\mathbb{E}\left[\omega_j \left(X^{(j)} \right) \ g_j \left(Z, \theta \right) \right] = 0, \quad j \in \{1, \ldots, J\}, \tag{5}$$

model which can also be written in the compact form $\mathbb{E}\left[\omega(\underline{X}) \ \underline{g}(Z,\theta)\right] = 0$.

COROLLARY 1. *Under the conditions of Theorem 1, we have*

$$I_{\theta_0} = \sup_{\omega \in \mathcal{B}} I_{\theta_0}(\omega).$$

The proof of Theorem 1 is based on the fact that the tangent space for the model (1) is the intersection of the tangent spaces corresponding to the J conditional moment models defined by each of the vector functions $g_1, \ldots, g_J$. (See the Appendix for the formal definition of a tangent space.) This result is of independent interest and hence we state it below. Moreover, it will be used in Section 3 to propose an alternative strategy for approximating the efficient score.

PROPOSITION 1. *The tangent space (of the nonparametric part) of the model (1) defined by* $\mathbb{E}\left[g_j(Z,\theta) \mid X^{(j)}\right] = 0 \ (a.s.), \quad \forall j = 1, \ldots, J$, *is the intersection of the tangent spaces (of the nonparametric parts) of the models defined, for each* $j \in \{1, \ldots, J\}$, *by*

$$\mathbb{E}\left[g_j \left(Z, \theta \right) \mid X^{(j)} \right] = 0 \ (a.s.).$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

2.2. Assumptions

Assumption T. There exist bounded, square integrable vector functions $b_1, \ldots, b_J$ (with the same dimensions as $g_1, \ldots, g_J$, respectively) such that:

1. for any $i, j \in \{1, \ldots, J\}$ with $i \neq j$, $\mathbb{E}\big(g_i(Z, \theta_0)\, b_j'(Z) \mid X^{(i)}, X^{(j)}\big) = 0$ (a.s.);
2. for any $i \in \{1, \ldots, J\}$, $\mathbb{E}\big(b_i(Z)\, g_i'(Z, \theta_0) \mid X^{(i)}\big)$ is invertible (a.s.) and

$$c_b = \max_{1 \leq i \leq J} \left\| \mathbb{E}^{-1}\left(b_i(Z)\, g_i'(Z, \theta_0) \mid X^{(i)}\right) \right\|_\infty < \infty;$$

3. $\max_{1 \leq i \leq J} \left\| \mathbb{E}\big(g_i'(Z, \theta_0)\, g_i(Z, \theta_0) \mid X^{(i)}\big) \mathbb{E}\big(\mathbf{1}\{b_i(Z) \neq 0\} \mid X^{(i)}\big) \right\|_\infty < \infty$ (a.s.).

Assumption SP.

1. The models $\mathcal{P}$ defined by (1) and $\mathcal{P}_k$ defined by (4), with $k \in \mathbb{N}^*$, can be written in the semiparametric form

$$\mathcal{P} = \big\{P_{\theta,\eta} : \theta \in \Theta,\ \eta \in H\big\}, \quad \mathcal{P}_k = \big\{P_{\theta,\eta} : \theta \in \Theta,\ \eta \in H_k\big\},\ k \in \mathbb{N}^*,$$

   and satisfy the assumptions of van der Vaart (1998, Lemma 25.25).
2. The Fisher information matrices I_{θ_0} and $I_{\theta_0}^{(k)}$ on θ_0 in models $\mathcal{P}$ and $\mathcal{P}_k$ respectively, for any $k \in \mathbb{N}^*$, are well defined and nonsingular.

Primitive conditions guaranteeing Assumption T are provided in the technical Assumption T' in the Appendix. Let us comment on the role of Assumption T. It ensures that $\mathcal{M} \cap \mathcal{T}'$ is dense in $\mathcal{T}'$ where $\mathcal{M}$ is the subspace of bounded functions of Z. This density condition is commonly used in the literature to show that the candidate tangent space $\mathcal{T}'$ of the nonparametric part of the model is indeed that tangent space. However, contrary to what is sometimes implicitly supposed in the literature, there is no general mathematical result that guarantees that the subspace of bounded functions is dense in any subspace of the square integrable functions of Z.[1] Our proofs also rely on the fact that $\mathcal{M} \cap \mathcal{T}'$ is dense in $\mathcal{T}'$ and Assumption T provides a general sufficient condition guaranteeing this property with $J \geq 1$. In the case of bounded and orthogonal g_j's, one could simply take $b_j = g_j$, $1 \leq j \leq J$.

To guarantee Assumption SP.2, it suffices to suppose that for any $1 \leq j \leq J$: (i) $\left\| \mathbb{V}\big(g_j(Z, \theta_0) \mid X^{(j)}\big) \right\|_\infty < \infty$; (ii) the maps $\theta \mapsto \mathbb{E}\big(g_j(Z, \theta_0) \mid X^{(j)} = x^{(j)}\big)$ are differentiable for $P_{X^{(j)}}-$almost all $x^{(j)}$; and (iii) the information matrix

$$\mathbb{E}\left\{\left[\partial_{\theta'} \mathbb{E}\left(g_j(Z, \theta_0) \mid X^{(j)}\right)\right]' \mathbb{V}^{-}\left[g_j(Z, \theta_0) \mid X^{(j)}\right] \partial_{\theta'} \mathbb{E}\left(g_j(Z, \theta_0) \mid X^{(j)}\right)\right\}$$

is nonsingular.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

A consequence of Assumption SP (see van der Vaart, 1998, Lemma 25.25) is that the parameter defined by $\psi(P_{\theta,\eta}) = \theta$ is differentiable at $P_0 = P_{\theta_0,\eta_0}$ with respect to the tangent space $\mathcal{T} = \mathcal{T}(\mathcal{P}, P_0)$. It also ensures that the tangent space $\mathcal{T}$ can be written as the sum of the finite dimensional subspace spanned by the components of the parametric score S_{θ_0} and the tangent space $\mathcal{T}'$ corresponding to the nonparametric part $\mathcal{P}' = \left\{P_{\theta_0,\eta} : \ \eta \in H\right\}$ of the model $\mathcal{P}$:

$$\mathcal{T} = \text{lin} S_{\theta_0} + \mathcal{T}'.$$

3. APPROXIMATING THE EFFICIENT SCORE

To simplify the presentation, in this section let us consider only the case $J = 2$. To obtain an efficient estimator, a common way is to solve θ from the efficient score equations; see van der Vaart (1998, Sect. 25.8). By definition, the efficient score is the componentwise projection of the score S_{θ_0} on the orthogonal complement of the tangent space $\mathcal{T} = \mathcal{T}(\mathcal{P}, P_0)$ defined in equation (A.11). In the projection of S_{θ_0} on $\mathcal{T}^{\perp}$, only the nonparametric part of the tangent space matters. Moreover, the projection of S_{θ_0} is componentwise. It is then common practice in the literature to identify $\mathcal{T}$ with the subspace of $\left\{L^2(P_0)\right\}^d = \oplus_{k=1}^d L^2(P_0)$ obtained as the $d-$fold cartesian product of the nonparametric part of $\mathcal{T}$. Here, the direct sum of Hilbert spaces is considered with the usual inner product $\langle(\phi_1, \cdots, \phi_d), (\psi_1, \cdots, \psi_d)\rangle = \langle\phi_1, \psi_1\rangle + \cdots + \langle\phi_d, \psi_d\rangle$. Therefore, we will slightly change our notation for the tangent spaces. More precisely, let us define

$$\mathcal{T} = \left\{s \in \left\{L^2(P_0)\right\}^d : \ \mathbb{E}(s) = 0, \ \mathbb{E}\left(g_i(Z, \theta_0)s'(Z) \mid X^{(i)}\right) = 0, \ i = 1, 2\right\}$$

$$= \mathcal{T}_1 \cap \mathcal{T}_2,$$

where, for $i = 1, 2$,

$$\mathcal{T}_i = \left\{s \in \left\{L^2(P_0)\right\}^d : \ \mathbb{E}(s) = 0, \ \mathbb{E}\left(g_i(Z, \theta_0)s'(Z) \mid X^{(i)}\right) = 0\right\},$$

so that

$$\mathcal{T}_i^{\perp} = \left\{s \in \left\{L^2(P_0)\right\}^d : \ s(Z) = a_i\left(X^{(i)}\right) \ g_i(Z, \theta_0)\right\},$$

where a_i is an arbitrary matrix-valued function (of compatible dimension $d \times p_i$) whose entries are square integrable functions of $X^{(i)}$, $i = 1, 2$. Clearly, $\mathcal{T}^{\perp} = \overline{\mathcal{T}_1^{\perp} + \mathcal{T}_2^{\perp}}$.

In general, the projection of S_{θ_0} on $\mathcal{T}^{\perp}$ is not explicit. To approximate this projection and to further build an asymptotically efficient estimator for model (1), two strategies can be developed from the results obtained in this paper. The first

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

one is based on Theorem 1. With the same notations as in Section 2.1, let $\mathcal{P}_k$ be the model defined by (4), *i.e.,*

$$\mathbb{E}\left[w_{(k)}^{(j)}\left(X^{(j)}\right)\otimes g_j(Z,\theta)\right]=0,\quad j=1,\ldots,J.$$

Then $\{\mathcal{P}_k\}_{k\in\mathbb{N}^*}$ is a decreasing sequence of models and the corresponding tangent spaces satisfy (see the proof of Theorem 1)

$$\mathcal{T}_1\supset\mathcal{T}_2\supset\cdots\supset\mathcal{T}_k\supset\mathcal{T}_{k+1}\supset\cdots\supset\bigcap_{k=1}^{\infty}\mathcal{T}_k=\mathcal{T},$$

where $\mathcal{T}$ is the tangent space of model (1). Taking the sequence of the efficient scores $\overline{S}_{\theta_0}^{(k)}$ in models $\mathcal{P}_k$, Hansen and Sargent (1991, Thm. 4.5) ensures the convergence (in $L^2(P_0)$) of $\overline{S}_{\theta_0}^{(k)}$ to the efficient score $\overline{S}_{\theta_0}$ of model (1). In conclusion, using sufficiently numerous suitable instruments in the original model (1) delivers an approximate score $\overline{S}_{\theta_0}^{(k)}$ arbitrarily close to $\overline{S}_{\theta_0}$. Results of this type have already been discussed in the literature in some particular cases of model (1); see, among others, Chamberlain (1992b), Newey (1993), Hahn (1997), and Han and Phillips (2006). How many such instruments to choose, in particular how to define the dependence of k on the sample size n, is an open question in our general setting and will be the subject of future work.

Here, we will investigate an alternative approach, based on the form of the tangent space given in Proposition 1, or, to be more precise, based on the aforementioned writing of the orthogonal tangent space $\mathcal{T}^{\perp}=\overline{\mathcal{T}_1^{\perp}+\mathcal{T}_2^{\perp}}$. For this purpose, we will use the iterative ('backfitting' or successive approximation) procedure considered in Bickel, Klaassen, Ritov, and Wellner (1993, Thm. A.4.2, p. 438); BKRW hereafter. Let $H_i=\mathcal{T}_i^{\perp}$, $g_i=g(Z,\theta_0)$, $i=1,2$, and let $\partial_\theta\mathbb{E}(g_i')$ be the transpose of the matrix $\partial_{\theta'}\mathbb{E}(g_i)$ defined in equation (3). The steps of the procedure we propose are the following:

1. Set $m=0$. Take $a_1^{(0)}=0$.
2. Put $m=m+1$. Calculate

$$\overline{S}_{\theta_0}^{(m)}=a_1^{(m)}\left(X^{(1)}\right)\,g_1+a_2^{(m)}\left(X^{(2)}\right)\,g_2,$$

where

$$\begin{aligned}a_1^{(m)}\left(X^{(1)}\right)=a_1^{(m)}\left(X^{(1)},\theta_0\right)&=-\partial_\theta\mathbb{E}\left(g_1'\mid X^{(1)}\right)\,\mathbb{V}^-\left(g_1\mid X^{(1)}\right)\\&+\mathbb{E}\left[\partial_\theta\mathbb{E}\left(g_2'\mid X^{(2)}\right)\,\mathbb{V}^-\left(g_2\mid X^{(2)}\right)\,g_2\,g_1'\mid X^{(1)}\right]\mathbb{V}^-\left(g_1|X^{(1)}\right)\\&+\mathbb{E}\left[\mathbb{E}\left[a_1^{(m-1)}\left(X^{(1)}\right)\,g_1\,g_2'\mid X^{(2)}\right]\mathbb{V}^-\left(g_2|X^{(2)}\right)\,g_2\,g_1'\mid X^{(1)}\right]\\&\times\mathbb{V}^-\left(g_1|X^{(1)}\right)\end{aligned}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

and

$$a_2^{(m)}\left(X^{(2)}\right)=a_2^{(m)}\left(X^{(2)},\theta_0\right)=-\partial_\theta\mathbb{E}\left(g_2'\mid X^{(2)}\right)\,\mathbb{V}^-\left(g_2\mid X^{(2)}\right)$$
$$-\mathbb{E}\left[a_1^{(m)}\left(X^{(1)}\right)\,g_1\,g_2'\mid X^{(2)}\right]\,\mathbb{V}^-\left(g_2|X^{(2)}\right).$$

3. Repeat from step 2 till the convergence of $\overline{S}_{\theta_0}^{(m)}$.

If $\mathcal{S}\subset\mathcal{H}$ is a linear subspace and $h\in\mathcal{H}$, let $\Pi(h|\mathcal{S})$ be the projection of h on $\overline{S}$. For subspaces $\mathcal{S}\subset\oplus_{k=1}^d L^2(P_0)$, $\Pi(s|\mathcal{S})$ denotes the (componentwise) projection of a vector $s\in\oplus_{k=1}^d L^2(P_0)$ on $\overline{S}$. Theorem A.4.2 (A) from BKRW directly yields the following result.

LEMMA 1. *Assume that the conditions of Theorem 1 hold true. When* $m\to\infty$,

$$\overline{S}_{\theta_0}^{(m)}=a_1^{(m)}(X^{(1)})g_1+a_2^{(m)}(X^{(2)})g_2\longrightarrow\overline{S}_{\theta_0}=\Pi\left(S_{\theta_0}|\mathcal{T}^\perp\right)=\Pi\left(S_{\theta_0}|\overline{H_1+H_2}\right)$$

in $\oplus_{k=1}^d L^2(P_0)$, *where* $g_i=g(Z,\theta_0)$, $i=1,2$.

Let us point out that even if Lemma 1 guarantees the convergence of the iterations $\overline{S}_{\theta_0}^{(m)}$, it is not necessarily true that the sequences $a_1^{(m)}\left(X^{(1)}\right)g_1$ and $a_2^{(m)}\left(X^{(2)}\right)g_2$ converge. Sufficient mild conditions are provided in Theorem A.4.2 (C) of BKRW, that are

$$\overline{S}_{\theta_0}=\Pi\left(S_{\theta_0}|\mathcal{T}^\perp\right)=a_1^*\left(X^{(1)}\right)\cdot g_1+a_2^*\left(X^{(2)}\right)\cdot g_2\in\mathcal{T}_1^\perp+\mathcal{T}_2^\perp \tag{6}$$

with $a_1^*\left(X^{(1)}\right)\cdot g_1\in\mathcal{T}_1^\perp\cap\left(\mathcal{T}_1^\perp\cap\mathcal{T}_2^\perp\right)^\perp\subset\mathcal{T}_1^\perp$. Moreover, by Proposition A.4.1 of BKRW, condition (6) is equivalent to the existence of a solution $a_1^*g_1$ and $a_2^*g_2$ for the system

$$\begin{cases} a_1^*\left(X^{(1)}\right)\,g_1=\rho_1-\mathbb{E}\left[a_2^*\left(X^{(2)}\right)\,g_2\,g_1'\mid X^{(1)}\right]\,\mathbb{V}^-\left(g_1|X^{(1)}\right)\,g_1\\ a_2^*\left(X^{(2)}\right)\,g_2=\rho_2-\mathbb{E}\left[a_1^*\left(X^{(1)}\right)\,g_1\,g_2'\mid X^{(2)}\right]\,\mathbb{V}^-\left(g_2|X^{(2)}\right)\,g_2,\end{cases} \tag{7}$$

where

$$\rho_i=\rho_i(Z,\theta_0):=\Pi\left(S_{\theta_0}|\mathcal{T}_i^\perp\right)=\mathbb{E}\left(S_{\theta_0}g_i'\mid X^{(i)}\right)\,\mathbb{V}^-\left(g_i\mid X^{(i)}\right)\,g_i$$
$$=-\partial_\theta\mathbb{E}\left(g_i'\mid X^{(i)}\right)\,\mathbb{V}^-\left(g_i\mid X^{(i)}\right)\,g_i.$$

(A careful inspection of the proof of Proposition A.4.1 of BKRW shows that the condition $H_1+H_2=\mathcal{T}_1^\perp+\mathcal{T}_2^\perp$ is a closed subspace, is not necessary for deriving that result, since what is really used in their proof is the relation $H_1^\perp\cap H_2^\perp=(H_1+H_2)^\perp$). If in addition, the system (7) has a unique solution, the backfitting algorithm above is nothing but a convergent iterative procedure for finding it.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

In applications, a convenient way to check uniqueness is to prove a contraction property. This is the case for instance if $\mathcal{T}_1^{\perp} \cap \mathcal{T}_2^{\perp} = \{0\}$, which in our framework holds if

$$\mathbb{E}\left(g_1 \; g_2' \mid X^{(1)}, X^{(2)}\right) = 0$$

(in the sequential case, this can be achieved by writing the initial system in an equivalent form satisfying the orthogonal condition above; see Section 4.1).

In the general case where $\mathcal{T}_1^{\perp} \cap \mathcal{T}_2^{\perp} \neq \{0\}$ the system (7), rewritten as in Proposition A.4.1 of BKRW under the form

$$\begin{cases} h_1^* = \Pi\left(S_{\theta_0} - h_2^* | \mathcal{T}_1^{\perp}\right) \\ h_2^* = \Pi\left(S_{\theta_0} - h_1^* | \mathcal{T}_2^{\perp}\right), \end{cases}$$

does not necessarily have the contraction property. In our problem $h_1^* = a_1^* g_1$ and $h_2^* = a_2^* g_2$ with g_1 and g_2 given. Hence, it suffices to check a contraction property for $a_1^* g_1$ and $a_2^* g_2$ or some given transformations of them. We will see in Section 4.2 that in the regression-like models with missing data framework, see Robins et al. (1994), the equations (7) lead to a contraction property for some given transformations of $a_1^* g_1$ and $a_2^* g_2$.

The 'backfitting' algorithm we propose above, involves θ_0 that is unknown. In practice one can use the following steps: (i) build $\widetilde{\theta}_n$, a $\sqrt{n}-$consistent estimator of θ_0, for instance the *smooth minimum distance estimator (SMD)* as in Lavergne and Patilea (2013); (ii) estimate nonparametrically $a_1^{(m^\star)}$ and $a_2^{(m^\star)}$, the solution of the 'backfitting' algorithm obtained after, say, $m^\star$ iterations using $\widetilde{\theta}_n$ instead of θ_0; and (iii) use classical GMM to construct an efficient estimator $\widehat{\theta}^{(m^\star)}$, based on the approximate efficient score equations $\mathbb{E}\left(\widehat{\widetilde{S}}_\theta\right) = 0$, where

$$\widehat{\widetilde{S}}_\theta = \widehat{a}_1^{(m^\star)}\left(X^{(1)}, \widetilde{\theta}_n\right) \; g_1(Z,\theta) + \widehat{a}_2^{(m^\star)}\left(X^{(2)}, \widetilde{\theta}_n\right) \; g_2(Z,\theta),$$

and $\widehat{a}_i^{(m^\star)}\left(X^{(i)}, \widetilde{\theta}_n\right)$ are nonparametric estimates of $a_i^{(m^\star)}\left(X^{(i)}, \theta_0\right)$, $i = 1, 2$.

4. APPLICATIONS

In this section, we illustrate the utility of our theoretical results for four classes of models: sequential (nested) conditional models, regression-like models with missing data, simultaneous equation models, and bivariate binary choice models. The general results in Sections 2 and 3 above allow us: (a) to complete a semiparametric efficiency bound result of Chamberlain (1992b); (b) to generalize the mean regression with missing data setting of Robins et al. (1994) and Tan (2011) to more general moment conditions, which includes for example, quantile regressions; (c) to consider a simultaneous equation model with perhaps more appealing orthogonality conditions on the error terms; and (d) to generalize the bivariate probit models.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

4.1. Sequential Conditional Moments

Important cases where equations (7) have an explicit solution are the cases where $\sigma\left(X^{(1)}\right) \subset \sigma\left(X^{(2)}\right)$ holds true. In the case $J = 2$, the model $\mathbb{E}\left(g_j(Z,\theta) \mid X^{(j)}\right) = 0$, $j = 1, 2$, defined in (1) can be equivalently written in the form

$$\begin{cases} \mathbb{E}\left(\widetilde{g}_1(Z,\theta) \mid X^{(1)}\right) = 0 \\ \mathbb{E}\left(g_2(Z,\theta) \mid X^{(2)}\right) = 0, \end{cases} \tag{8}$$

where

$$\widetilde{g}_1(Z,\theta) = g_1(Z,\theta) - \mathbb{E}\left(g_1(Z,\theta_0)\ g_2'(Z,\theta_0) \mid X^{(2)}\right)\ \mathbb{V}^{-1}\left(g_2(Z,\theta_0) \mid X^{(2)}\right)\ g_2(Z,\theta).$$

Here, we suppose that $\mathbb{V}\left(g_1(Z,\theta_0) \mid X^{(1)}\right)$ and $\mathbb{V}\left(g_2(Z,\theta_0) \mid X^{(2)}\right)$ are invertible and this guarantees that θ_0 is also identified by the equations (8). Recall that g_i is a short notation for $g_i(Z,\theta_0)$ and similarly let $\widetilde{g}_i$ replace $\widetilde{g}_i(Z,\theta_0)$.

Notice that $\widetilde{g}_1$ is the residual of the projection of g_1 on g_2 with respect to $\sigma\left(X^{(2)}\right)$ and $\mathbb{E}\left(\widetilde{g}_1\ g_2' \mid X^{(2)}\right) = 0$. Let $\widetilde{\mathcal{T}}_1$ be the tangent space of the model defined by the first equation in (8). By the definition of $\widetilde{g}_1$, it is quite clear that condition $\widetilde{\mathcal{T}}_1^{\perp} \cap \mathcal{T}_2^{\perp} = \{0\}$ holds true. Next, multiplying the ith equation in (7) by g_i, taking conditional expectation given $X^{(i)}$ and finally multiplying by $\mathbb{V}^{-1}\left(g_i \mid X^{(i)}\right)$, $i = 1, 2$, the system (7) corresponding to model (8) becomes:

$$\begin{cases} \widetilde{a}_1^*\left(X^{(1)}\right) = -\partial_\theta \mathbb{E}\left(\widetilde{g}_1' \mid X^{(1)}\right)\ \mathbb{V}^{-1}\left(\widetilde{g}_1 \mid X^{(1)}\right) \\ \qquad\qquad -\mathbb{E}\left(\widetilde{a}_2^*\left(X^{(2)}\right) \cdot g_2\ \widetilde{g}_1' \mid X^{(1)}\right)\ \mathbb{V}^{-1}\left(\widetilde{g}_1 \mid X^{(1)}\right) \\ \widetilde{a}_2^*\left(X^{(2)}\right) = -\partial_\theta \mathbb{E}\left(g_2' \mid X^{(2)}\right)\ \mathbb{V}^{-1}\left(g_2 \mid X^{(2)}\right) \\ \qquad\qquad -\mathbb{E}\left(\widetilde{a}_1^*\left(X^{(1)}\right) \cdot \widetilde{g}_1\ g_2' \mid X^{(2)}\right)\ \mathbb{V}^{-1}\left(g_2 \mid X^{(2)}\right). \end{cases} \tag{9}$$

Since by definition $\mathbb{E}\left(\widetilde{a}_2^*\left(X^{(2)}\right) g_2\ \widetilde{g}_1' \mid X^{(1)}\right) = \mathbb{E}\left[\widetilde{a}_2^*\left(X^{(2)}\right)\mathbb{E}\left(g_2\ \widetilde{g}_1' \mid X^{(2)}\right) \mid X^{(1)}\right] = 0$ and $\mathbb{E}\left(\widetilde{a}_1^*\left(X^{(1)}\right)\widetilde{g}_1\ g_2' \mid X^{(2)}\right) = \widetilde{a}_1^*\left(X^{(1)}\right)\mathbb{E}\left(\widetilde{g}_1\ g_2' \mid X^{(2)}\right) = 0$, we obtain:

$$\begin{cases} \widetilde{a}_1^*\left(X^{(1)}\right) = -\partial_\theta \mathbb{E}\left(\widetilde{g}_1' \mid X^{(1)}\right)\ \mathbb{V}^{-1}\left(\widetilde{g}_1 \mid X^{(1)}\right) \\ \widetilde{a}_2^*\left(X^{(2)}\right) = -\partial_\theta \mathbb{E}\left(g_2' \mid X^{(2)}\right) \mathbb{V}^{-1}\left(g_2 \mid X^{(2)}\right). \end{cases} \tag{10}$$

($\partial_\theta \mathbb{E}(\widetilde{g}_i')$ denotes the transpose of the matrix $\partial_{\theta'} \mathbb{E}(\widetilde{g}_i)$.) The efficient score $\overline{S}_{\theta_0}$ can then be written as:

$$\begin{aligned} \overline{S}_{\theta_0} &= \widetilde{a}_1^*(X) \cdot \widetilde{g}_1 + \widetilde{a}_2^*(X) \cdot g_2 \\ &= -\partial_\theta \mathbb{E}\left(\widetilde{g}_1' \mid X^{(1)}\right)\ \mathbb{V}^{-1}\left(\widetilde{g}_1 \mid X^{(1)}\right)\ \widetilde{g}_1 \\ &\quad -\partial_\theta \mathbb{E}\left(g_2' \mid X^{(2)}\right)\ \mathbb{V}^{-1}\left(g_2 \mid X^{(2)}\right)\ g_2. \end{aligned}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

In the particular case where $X^{(1)} = X^{(2)} = X$,

$$\begin{aligned}\overline{S}_{\theta_0} &= \widetilde{a}_1^*(X)\cdot\widetilde{g}_1 + \widetilde{a}_2^*(X)\cdot g_2\\ &= -\partial_\theta \mathbb{E}\left(\widetilde{g}_1' \mid X\right)\ \mathbb{V}^{-1}\left(\widetilde{g}_1 \mid X^{(1)}\right)\ \widetilde{g}_1 - \partial_\theta \mathbb{E}\left(g_2' \mid X\right)\ \mathbb{V}^{-1}\left(g_2 \mid X\right)\ g_2\\ &= \begin{pmatrix} -\partial_\theta \mathbb{E}\left(\widetilde{g}_1' \mid X\right)\\ -\partial_\theta \mathbb{E}\left(g_2' \mid X\right)\end{pmatrix}' \mathbb{V}^{-1}\left(\begin{pmatrix}\widetilde{g}_1\\ g_2\end{pmatrix} \mid X\right)\begin{pmatrix}\widetilde{g}_1\\ g_2\end{pmatrix}\\ &= -\partial_\theta \mathbb{E}\left(g'\ C'(X) \mid X\right)\mathbb{V}^{-1}\left(C(X)g \mid X\right)\ C(X)g\\ &= -\partial_\theta \mathbb{E}\left(g' \mid X\right)\ \mathbb{V}^{-1}\left(g \mid X\right)g,\end{aligned}$$

where $g' = \left(g_1'\ g_2'\right)$ and

$$C(X) = \begin{pmatrix} I & -\mathbb{E}\left(g_1\ g_2' \mid X\right)\mathbb{V}^{-1}\left(g_2 \mid X\right)\\ 0 & I\end{pmatrix}$$

is a nonsingular random matrix. This expression of the efficient score directly yields the efficiency bound derived in Chamberlain (1987).

Another important particular case of formulae (10) is provided by models defined by sequential conditional moments; see Chamberlain (1992b) and Ai and Chen (2012). Taking $X^{(1)} = X_1$ and $X^{(2)} = \left(X_1', X_2'\right)'$, one obtains:

$$\begin{aligned}\overline{S}_{\theta_0} &= \widetilde{a}_1^*(X)\cdot\widetilde{g}_1 + \widetilde{a}_2^*(X)\cdot g_2\\ &= -\partial_\theta \mathbb{E}\left(\widetilde{g}_1' \mid X_1\right)\ \mathbb{V}^{-1}\left(\widetilde{g}_1 \mid X_1\right)\ \widetilde{g}_1 - \partial_\theta \mathbb{E}\left(g_2' \mid X_1, X_2\right)\ \mathbb{V}^{-1}\left(g_2 \mid X_1, X_2\right)\ g_2.\end{aligned}$$

Let us point out that Chamberlain (1992b) only proves this result for discrete distributions and Ai and Chen (2012) obtain the result in a more general framework (allowing for unknown infinite dimensional parameters in the equations defining the model) but under slightly more restrictive assumptions than in our setting.[2]

4.2. Regression-Like Models with Missing Data

Consider now a regression-like model defined by the equations:

$$\mathbb{E}\left[\rho\left(Y, X^*, \alpha\right) \mid X^*\right] = 0, \tag{11}$$

where $\rho(\cdot,\cdot,\cdot)$ is some measurable vector-valued function, α is a (finite-dimensional) vector of parameters, and the vector $\left(Y', X^{*\prime}\right) = \left(Y', X', V'\right)$ is not always completely observed. We also assume that a non-missing indicator δ and some other variable V^0 are always observed. In the following examples we consider two random missingness mechanisms considered respectively, by Tan (2011) and Robins et al. (1994).

Example 1

(i) The vector Y is observed iff $\delta = 1$;

(ii) The vector $W = \begin{pmatrix} X^* \\ V^0\end{pmatrix}$ is always observed and we have:

$$P\left(\delta = 1 \mid Y, W\right) = P\left(\delta = 1 \mid W\right) = \pi\left(W\right). \tag{12}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

Example 2

(i) Let $X^* = \begin{pmatrix} X \\ V \end{pmatrix}$ where X is observed iff $\delta = 1$;

(ii) The vector $W = \begin{pmatrix} Y \\ V \\ V^0 \end{pmatrix}$ is always observed and we have:

$$P(\delta = 1 \mid X, W) = P(\delta = 1 \mid W) = \pi(W). \tag{13}$$

Let α_0 be the true value of the parameter identified by the model (11). The equation (11) and each of (12) or (13) imply:

$$\mathbb{E}\left[\frac{\delta}{\pi(W)}\,\rho\left(Y, X^*, \alpha_0\right) \mid X^*\right] = 0. \tag{14}$$

We can consider this equation at the observational level even for missing X^*, since for missing values of X^*, we have $\delta = 0$ which renders the equation noninformative. Note also that (12) and (13) can be written under the unified form:

$$P\left(\delta = 1 \mid Y, X^*, W\right) = \pi(W).$$

Therefore, at the observational level, with any of the two examples we obtain a model such as:

$$\begin{cases} \mathbb{E}\left[\dfrac{\delta}{\pi(W)}\,\rho\left(Y, X^*, \alpha_0\right) \mid X^*\right] = 0 \\ \mathbb{E}\left[\dfrac{\delta}{\pi(W)} - 1 \mid W\right] = 0. \end{cases} \tag{15}$$

Moreover, as in Graham (2011, fn. 8, p. 442), it can be shown that, at the observational level, a model given by equation (11) and any of the missing data mechanisms described in Example 1 or Example 2, is equivalent to the model defined by (15).

With our notation, Z is the vector built as the union of all the variables contained in Y, X^*, W and δ, $\theta = \alpha$, $g_1(Z, \theta) = \{\delta/\pi(W)\}\rho(Y, X^*, \alpha)$, $g_2(Z, \theta) = \{\delta/\pi(W)\} - 1$, $X^{(1)} = X^*$, and $X^{(2)} = W$. Let ρ be short for $\rho(Y, X^*, \alpha_0)$. Then the functions a_1^* and a_2^* defining the efficient score are given by the following equations obtained (see also equations (9)) from equations (7):

$$\begin{aligned} a_1^*\left(X^*\right) &= a_1^*\left(X^{(1)}\right) \\ &= -\partial_\alpha \mathbb{E}\left(\rho' \mid X^*\right)\, \mathbb{E}^{-1}\left(\frac{1}{\pi(W)}\,\rho\,\rho' \mid X^*\right) \\ &\quad + \mathbb{E}\left\{\mathbb{E}\left[a_1^*\left(X^*\right)\,\rho \mid W\right]\,\frac{1-\pi(W)}{\pi(W)}\,\rho' \mid X^*\right\}\,\mathbb{E}^{-1}\left(\frac{1}{\pi(W)}\,\rho\,\rho' \mid X^*\right); \end{aligned}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

$$
\begin{aligned}
a_2^*(W) &= a_2^*\left(X^{(2)}\right) \\
&= \mathbb{E}\left[a_1^*\left(X^*\right)\rho\,\frac{\delta}{\pi(W)}\left(\frac{\delta}{\pi(W)}-1\right) \mid W\right]\mathbb{E}^{-1}\left[\left(\frac{\delta}{\pi(W)}-1\right)^2 \mid W\right] \\
&= -\mathbb{E}\left[a_1^*\left(X^*\right)\ \rho \mid W\right].
\end{aligned}
$$

In the particular case, where $\rho = \rho(Y, X^*, \alpha_0) = Y - g(X^*, \alpha_0)$ and the selection probability $\pi(W)$ is known, these are exactly the equations obtained in Robins et al. (1994). They showed that for the regression case, the equation for a_1^* corresponds to a contraction (see the proof of their Proposition 4.2). In Section A.4 in the Appendix, we show that such a contraction property holds for a more general ρ. Hence, we could include further interesting examples in our framework, *e.g.* quantile regressions. The contraction property allows us to solve the equations in $a_1^*(X^*)$ and $a_2^*(W)$ by successive approximations.

Let us consider the extended framework where the selection probability is known up to an unknown finite-dimensional parameter γ_0, that is:

$$P(\delta = 1 \mid W) = \pi(W, \gamma_0),$$

(see also Robins et al., 1994, eqn. 18). In Section A.5 in the Appendix, we show that the efficient score for α_0 has the same expression regardless of the selection probability function π is given or depends on the unknown parameter γ_0. Thus, we extend a result of Robins et al. (1994), see also Tan (2011), obtained in the particular case of mean regressions.

We close this example with the following remark. Robins et al. (1994) considered the case where missingness arises only in covariables $X^\star$ (that is also the case considered in our Example 2) and derived the efficient score equations. Tan (2011) formally obtained the same equations with missing regressors *or* missing responses (the case corresponding to our Example 1) using the corresponding definition of W. However, our approach, based on conditional moments, points to an important difference between Examples 1 and 2. *I.e.*, in the possibly missing responses case, we have $\sigma(X^*) \subset \sigma(W)$, so that Example 1 falls in the sequential conditional moments framework where the solutions for a_1^* and a_2^* are explicit. On the other hand, such explicit solutions are *no longer* available in the framework considered by Robins et al. (1994) and in our Example 2.

4.3. Simultaneous Equations Models

Consider the linear simultaneous equations:

$$
\begin{cases}
Q = a\,P + b'\,X + c'\,T + \varepsilon \\
P = \alpha\,Q + \beta'\,X + \gamma'\,U + \eta,
\end{cases}
\tag{16}
$$

that could represent, for instance, the equations of a supply (P) and demand (Q) model. Usually, it is admitted that the errors ε and η are correlated with the

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

endogenous variables P and Q but are uncorrelated with the exogenous regressors, which means:

$$\mathbb{E}(\varepsilon \mid X, T, U) = \mathbb{E}(\eta \mid X, T, U) = 0. \tag{17}$$

Under this strong assumption, the model can be written as:

$$\mathbb{E}\left[\rho_1(Z, \theta) \mid X^{(1)}\right] = 0, \tag{18}$$

with a two components function ρ_1, the random vectors $Z = (Q, P, X, T, U)'$, $X^{(1)} = (X, T, U)'$ and parameter $\theta = (a, b', c', \alpha, \beta', \gamma')'$. The model (18) allows for standard inference methods (GMM, instrumental variables, 2SLS) for the efficient estimation of θ. With our approach, we are able to relax these assumptions by letting the errors in each equation be uncorrelated with the corresponding exogenous variables of the same equation. *I.e.*, we only impose:

$$\mathbb{E}(\varepsilon \mid X, T) = 0 \qquad \text{and} \qquad \mathbb{E}(\eta \mid X, U) = 0, \tag{19}$$

that represent more appealing conditions in the econometric literature. See for instance Wooldridge (2010, eqn. 14.33–14.36).

We then obtain the following system of equations of the form given in (1):

$$\begin{cases} \mathbb{E}\left[\rho_1(Z, \theta) \mid X^{(1)}\right] = 0 \\ \mathbb{E}\left[\rho_2(Z, \theta) \mid X^{(2)}\right] = 0, \end{cases} \tag{20}$$

where Z and θ are the same as above, but $X^{(1)} = (X, T)'$, $X^{(2)} = (X, U)'$, $\rho_1(Z, \theta) = Q - a\,P - b'\,X - c'\,T$, and $\rho_2(Z, \theta) = P - \alpha\,Q - \beta'\,X - \gamma'\,U$.

Let us point out that, on the one hand, one could also take into account additional exogenous instruments for each of the two equations in (16). Let I_1, I_2 denote such instruments. Then one could add the moment equations $\mathbb{E}[\rho_1(Z, \theta)\, I_1] = 0$, $\mathbb{E}[\rho_2(Z, \theta)\, I_2] = 0$ to the system (20). On the other hand, one could also take into account additional observed variables and moment conditions not containing unknown parameters as in Qian and Schmidt (1999, Sect. 4). However, in order to provide a readable illustration of our methodology, for the remainder of this example, we stay with the system (16).

If $1 - a\alpha \neq 0$, the system (16) can be written under the restricted reduced form:

$$\begin{cases} Q = \delta(b' + a\,\beta')\,X + \delta\,c'\,T + \delta\,a\,\gamma'\,U + \tilde{\varepsilon} \\ P = \delta(\alpha\,b' + \beta')\,X + \delta\,\alpha\,c'\,T + \delta\,\gamma'\,U + \tilde{\eta}, \end{cases} \tag{21}$$

where $\delta = (1 - a\alpha)^{-1}$, $\tilde{\varepsilon} = \delta(\varepsilon + a\,\eta)$, and $\tilde{\eta} = \delta(\alpha\,\varepsilon + \eta)$. Under the conditions (17), we also have $\mathbb{E}(\tilde{\varepsilon} \mid X, T, U) = \mathbb{E}(\tilde{\eta} \mid X, T, U) = 0$. If the matrix $\mathbb{E}(WW')$ has full rank, where $W' = (X', T', U')$, then the coefficients

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

$s_1 = \delta(b + a\beta), \ldots, s_6 = \delta\gamma$ are identified and can be consistently estimated by OLS, and efficiently estimated in a second stage, using nonparametric estimates of the conditional variances $\mathbb{V}(\tilde{\varepsilon} \mid X, T, U)$ and $\mathbb{V}(\tilde{\eta} \mid X, T, U)$. This two-stage procedure leads to efficient estimators for the vector coefficient θ in the original system (16).[3]

Suppose now, that we only impose the conditions from equations (19). Then $\mathbb{E}(\tilde{\varepsilon} \mid X) = \mathbb{E}(\tilde{\eta} \mid X) = 0$ and one could consistently estimate θ, but the two-stage estimation procedure no longer yields an efficient estimator. An approximatively efficient estimator can be obtained by the approaches proposed in this paper, *i.e.*, either by increasing the number of instruments for each equation (depending on X and T for the first equation in the system (20) and on X and U for the second one), or by the iterative procedure described in Section 3 above. To illustrate the iterative approach, let us assume that the conditional variances $\sigma_\varepsilon^2(X, T) = \mathbb{V}(\varepsilon \mid X, T)$ and $\sigma_\eta^2(X, U) = \mathbb{V}(\eta \mid X, U)$ are strictly positive. Noting that the functions a_1^* and a_2^* appearing in equations (7) can be arbitrarily defined on the sets $g_1 = 0$ and $g_2 = 0$, respectively, we obtain the following system for a_1^* and a_2^*:

$$\begin{cases} a_1^*(X, T)\, \sigma_\varepsilon^2(X, T) = (\mathbb{E}(P \mid X, T),\, X',\, T',\, 0,\, 0,\, 0)' - \mathbb{E}\left[a_2^*(X, U)\, g_2\, g_1 \mid X, T\right] \\ a_2^*(X, U)\, \sigma_\eta^2(X, U) = (0,\, 0,\, 0,\, \mathbb{E}(Q \mid X, U),\, X',\, U')' - \mathbb{E}\left[a_1^*(X, T)\, g_1\, g_2 \mid X, U\right]. \end{cases}$$

4.4. Generalization of Bivariate Probit Models

Consider the following system involving two binary outcomes Y_1 and Y_2:

$$\begin{cases} Y_1 = \mathbf{1}\left\{X_1^T \beta_1 - \varepsilon_1 > 0\right\} \\ Y_2 = \mathbf{1}\left\{X_2^T \beta_2 - \varepsilon_2 > 0\right\}. \end{cases} \tag{23}$$

If $\varepsilon = (\varepsilon_1, \varepsilon_2)$ is bivariate normal and independent of (X_1, X_2) this is the classical bivariate probit model (see, for example, Wooldridge, 2010, p. 595).

Assuming only that ε_i is independent of X_i and the distribution function F_{ε_i} of ε_i is known, for $i \in \{1, 2\}$, then the model becomes:

$$\begin{cases} \mathbb{E}\left[Y_1 - F_{\varepsilon_1}(X_1^T \beta_1) \mid X_1\right] = 0 \\ \mathbb{E}\left[Y_2 - F_{\varepsilon_2}(X_2^T \beta_2) \mid X_2\right] = 0. \end{cases} \tag{24}$$

Since the joint law of $(\varepsilon_1, \varepsilon_2)$ is not necessarily known, as in the classical bivariate probit model, a maximum likelihood approach to estimate the parameters in (23) is no longer available, but one can (approximately) efficiently estimate the parameters with our approach. Note that in this form (24) no specific assumption is needed on the joint law of $(\varepsilon_1, \varepsilon_2)$. In particular we do not need to consider

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

the nuisance parameter $\rho = cov(\varepsilon_1, \varepsilon_2)$ that is usually integrated in the Gaussian likelihood in order to estimate β_1 and β_2. This approach still works if ε_i and X_i are not independent, assuming instead that the conditional distribution function $F_{\varepsilon_i}(\cdot, x_i)$ of ε_i, given that $X_i = x_i$, is known, for $i \in \{1, 2\}$.

NOTES

1. Here we provide a counterexample showing that the set of bounded functions is not dense in any subspace of $L^2(P_0)$. Indeed, if $suppP_0 = suppZ = [0,1]^q \subset \mathbb{R}^q$, let $\phi_0(u) = ||u||^{-1/4}\mathbf{1}\{0 < ||u|| < 1\}$ and take $\mathcal{S}$ the linear span of $\{\phi_0(||Z|| - k), k \in \mathbb{Z}\}$. Then $\mathcal{S}$ in an infinite dimensional subspace of $L^2(P_0)$ and the subspace $\mathcal{M} \subset L^2(P_0)$ of bounded functions is not dense in $\mathcal{S}$ since $\mathcal{M} \cap \mathcal{S} = \{0\}$, so that $\overline{\mathcal{M} \cap \mathcal{S}} = \{0\}$.

2. In Ai and Chen (2012), it is implicitly required that the class $\mathcal{G}$ appearing in their Assumption A in the Mathematical Appendix, is the same for each value of their model parameter α. This variation independent parametrization assumption represents an additional restriction that is unnecessary in our approach. See also van der Laan and Robins (2003, p. 18), for some lucid comments on the existence of a variation independent parametrization.

3. One could also consider the weaker conditions

$$\mathbb{E}(\varepsilon \mid X) = \mathbb{E}(\eta \mid X) = 0, \tag{22}$$

in which case $\mathbb{E}(\tilde{\varepsilon} \mid X) = \mathbb{E}(\tilde{\eta} \mid X) = 0$. A two-stage efficient procedure based on the nonparametric estimation of the conditional variances with respect to X, is still possible. However, weaker conditions on the errors such as these are less intuitive and justified for modeling purposes. The point we make with this example is that assumptions (19) may appear as the most natural compromise between the 'extreme' conditions (22), the less restrictive, and (19), the most restrictive.

4. As in Lavergne (2008), let $\Lambda = \mathbb{E}^{-1}(A'A)\mathbb{E}(A'B)$. Then

$$\mathbb{E}[(B - A\Lambda)'(B - A\Lambda) = \mathbb{E}(B'B) - \mathbb{E}(B'A)\mathbb{E}^{-1}(A'A)\mathbb{E}(A'B)$$

is clearly positive semidefinite, and is zero iff $B = A\Lambda$.

REFERENCES

Ahn, S.C. & P. Schmidt (1999) Estimation of linear panel data models using GMM. In L. Mátyás (ed.), *Generalized Method of Moments Estimation*. Cambridge University Press.

Ai, C. & X. Chen (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71, 1795–1843.

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

Bickel, P.J., C.A.J. Klaassen, Y. Ritov, & J.A. Wellner (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. The John Hopkins University Press.

Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *Econometrica* 34, 305–334.

Chamberlain, G. (1992a) Efficiency bounds for semiparametric regression. *Econometrica* 60, 567–596.

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

Chen, X. & D. Pouzo (2009) Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics* 152, 46–60.

Gallant, A.R. (1975) Seemingly unrelated nonlinear regressions. *Journal of Econometrics* 3, 35–50.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

Graham, B. (2011) Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* 79, 437–452.
Hahn, J. (1997) Efficient estimation of panel data models with sequential moment restrictions. *Journal of Econometrics* 79, 1–21.
Han, C. & P.C.B. Phillips (2006) GMM with many moment conditions. *Econometrica* 74, 147–192.
Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–1054.
Hansen, L.P. (2008) Generalized method of moments. In S.N. Durlauf & L.E. Blume (eds.), *The New Palgrave Dictionary of Economics*, 2nd ed. Palgrave Macmillan.
Hansen, L.P. & T.J. Sargent (1991) *Rational Expectations Econometrics*. Westview Press.
Ibragimov, I.A. & R.Z. Has'minskii (1981) *Statistical Estimation. Asymptotic Theory*. Springer-Verlag.
Jun, S. & J. Pinske (2009) Efficient semiparametric seemingly unrelated quantile regression estimation. *Econometric Theory* 25, 1392–1414.
Lavergne, P. (2008) A Cauchy-Schwarz inequality for expectation of matrices. http://econpapers.repec.org/RePEc:sfu:sfudps:dp08-07.
Lavergne, P. & V. Patilea (2013) Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory. *Journal of Econometrics* 177, 47–59.
Müller, U.U. (2009) Estimating linear functionals in nonlinear regression with responses missing at random. *Annals of Statistics* 37, 2245–2277.
Müller, U.U. & I. Van Keilegom (2012) Efficient parameter estimation in regression with missing responses. *Electronic Journal of Statistics* 6, 1200–1219.
Newey, W.K. (1990) Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5, 99–135.
Newey, W.K. (1993) Efficient estimation of models with conditional moment restrictions. *Handbook of Statistics*, vol. 11, pp. 419–454.
Newey, W.K. (2001) Conditional moment restrictions in censored and truncated regression models. *Econometric Theory* 17, 863–888.
Newey, W.K. (2004) Efficient semiparametric estimation via moment restrictions. *Econometrica* 72, 1877–1897.
Qian, H. & P. Schmidt (1999) Improved instrumental variables and generalized method of moments estimators. *Journal of Econometrics* 91, 145–169.
Robins, J.M., A. Rotnitzky, & L.P. Zhao (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846–866.
Tan, Z. (2011) Efficient restricted estimators for conditional mean models with missing data. *Biometrika* 98, 663–684.
Tsiatis, A.A. (2006) *Semiparametric Theory and Missing Data*. Springer.
van der Laan, M.J. & J.M. Robins (2003) *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag.
van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge University Press.
Wei, Y., Y. Ma, & R.J. Carroll (2012) Multiple imputation in quantile regression. *Biometrika* 99, 423–438.
Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press.

APPENDIX

For a model $\mathcal{P}$ (resp. $\mathcal{P}_j$) and a probability measure P in the model, let $\dot{\mathcal{P}}_P$ (resp. $\dot{\mathcal{P}}_{j,P}$) denote the tangent cone of the model $\mathcal{P}$ (resp. $\mathcal{P}_j$) at P. When there is no possible confusion, we simply write $\dot{\mathcal{P}}$ (resp. $\dot{\mathcal{P}}_j$). Let $\mathcal{T}(\mathcal{P}, P)$ denote the tangent space of a model $\mathcal{P}$ at some probability measure $P \in \mathcal{P}$, that means the closure of the linear span of the tangent set $\dot{\mathcal{P}}_P$. By definition, both the tangent cone and the tangent space are subsets of $L^2(P)$.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

A.1. A General Lemma

The following result is a generalization of Newey (2004, Thm. 1) where only the case of models defined by conditional moments, with the same conditioning variables, was considered.

LEMMA A.1. *Let $P_0 \in \mathcal{P} \subset \mathcal{P}_1$ be the true law of the vector $Z \in \mathcal{Z}$ and $\theta_0 = \psi(P_0)$ for a map $\psi : \mathcal{P}_1 \to \mathbb{R}^d$ differentiable at P_0 relative to the tangent cone $\dot{\mathcal{P}}_{1,P_0}$. Let $\{\mathcal{P}_k\}_{k\in\mathbb{N}^*}$ be a decreasing family of statistical models such that:*

$$\mathcal{P}_1 \supset \mathcal{P}_2 \supset \cdots \supset \mathcal{P}_j \supset \mathcal{P}_{k+1} \supset \cdots \supset \bigcap_{k=1}^{\infty} \mathcal{P}_k \supset \mathcal{P} \ni P_0 \tag{A.1}$$

and

$$\bigcap_{k=1}^{\infty} \mathcal{T}_k = \mathcal{T}, \tag{A.2}$$

where $\mathcal{T} = \mathcal{T}(\mathcal{P}, P_0)$ and $\mathcal{T}_k = \mathcal{T}(\mathcal{P}_k, P_0)$, $k \in \mathbb{N}^*$. Then

$$I_{\theta_0}(\mathcal{P}) = \lim_{k\to\infty} I_{\theta_0}(\mathcal{P}_k),$$

where $I_{\theta_0}(\mathcal{P})$ stands for the Fisher information on $\theta_0 = \psi(P_0)$ in the model $\mathcal{P}$.

For the definition of the Fisher information $I_{\theta_0}(\mathcal{P})$ on $\theta_0 = \psi(P_0)$ in the model $\mathcal{P}$, we refer to Bickel et al. (1993) or van der Vaart (1998); see also Newey (1990). When the models $\mathcal{P}_k$, $k \in \mathbb{N}^*$, are defined by an increasing number of moment conditions with the same conditioning vectors, condition (A.2) is exactly the so-called spanning condition of Newey (2004).

Proof of Lemma A.1. By definition (van der Vaart, 1998, p. 363), there exists a continuous linear map $\dot{\psi} : L^2(P_0) \to \mathbb{R}^d$ such that for any $g \in \dot{\mathcal{P}}_{P_0} \subset L^2(P_0)$ and a submodel $(-\varepsilon, \varepsilon) \ni t \mapsto P_t$ with score function g,

$$\frac{\psi(P_t) - \psi(P_0)}{t} \underset{t\to 0}{\longrightarrow} \dot{\psi}(g).$$

By the Riesz representation theorem, there exists a unique d-dimensional vector-valued function having the components in $L^2(P_0)$ such that $\dot{\psi}(h) = \mathbb{E}_{P_0}(\overline{\psi} h)$ for every $h \in L^2(P_0)$. In particular,

$$\dot{\psi}(g) = \mathbb{E}_{P_0}(\overline{\psi} g) = \int \overline{\psi} g \, dP_0, \qquad \forall g \in \dot{\mathcal{P}}_{P_0} \subset L^2(P_0).$$

Let $\widetilde{\psi}$ and $\widetilde{\psi}_k$ denote the elements of $\left[L^2(P_0)\right]^d$ obtained by componentwise projections of $\overline{\psi}$ on the tangent spaces $\mathcal{T} \subset L^2(P_0)$ and $\mathcal{T}_k \subset L^2(P_0)$, respectively. The Fisher information matrices on $\theta_0 = \psi(P_0)$ in the models $\mathcal{P}$, $\mathcal{P}_k$ at P_0, are then defined by:

$$I_{\theta_0}^{-1}(\mathcal{P}) = \mathbb{V}_{P_0}(\widetilde{\psi}) = \mathbb{E}_{P_0}(\widetilde{\psi}\widetilde{\psi}'), \qquad I_{\theta_0}^{-1}(\mathcal{P}_k) = \mathbb{V}_{P_0}(\widetilde{\psi}_k), \quad k \in \mathbb{N}^*.$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

From

$$\mathcal{P}_1 \supset \mathcal{P}_2 \supset \cdots \supset \mathcal{P}_k \supset \mathcal{P}_{k+1} \supset \cdots \supset \bigcap_{k=1}^{\infty} \mathcal{P}_k \supset \mathcal{P}$$

we deduce that

$$\dot{\mathcal{P}}_1 \supset \dot{\mathcal{P}}_2 \supset \cdots \supset \dot{\mathcal{P}}_k \supset \dot{\mathcal{P}}_{k+1} \supset \cdots \supset \bigcap_{k=1}^{\infty} \dot{\mathcal{P}}_k \supset \dot{\mathcal{P}},$$

and

$$\mathcal{T}_1 \supset \mathcal{T}_2 \supset \cdots \supset \mathcal{T}_k \supset \mathcal{T}_{k+1} \supset \cdots \supset \bigcap_{k=1}^{\infty} \mathcal{T}_k = \mathcal{T},$$

where the last equality is due to (4). By Hansen and Sargent (1991, Lemma 4.5),

$$\lim_{k\to\infty} I_{\theta_0}^{-1}(\mathcal{P}_k) = \lim_{k\to\infty} \mathbb{V}_{P_0}\left(\Pi\left(\overline{\psi}|\mathcal{T}_k\right)\right) = \mathbb{V}_{P_0}\left(\Pi\left(\overline{\psi}|\mathcal{T}\right)\right) = \mathbb{V}_{P_0}\left(\widetilde{\psi}\right) = I_{\theta_0}^{-1}(\mathcal{P}). \quad \blacksquare$$

Remark A.1. Even if $\bigcap_{k=1}^{\infty} \mathcal{P}_k = \mathcal{P}$, condition (A.2) is not necessarily fulfilled. To see this, consider a symmetric density f_0 on the real line and let s_1, s_2 be two odd functions such that $|s_1|, |s_2| \leq 1$ $\left(\text{e.g. } s_l(x) = x^{2l-1}\mathbf{1}\{|x| \leq 1\}, l = 1, 2\right)$. For any $k \in \mathbb{N}^*$ and $t \in [-1, 1]$, define:

$$f_t(x) = f_0(x)[1 + t\, s_2(x)], \qquad f_{t;k}(x) = k f_0(kx)[1 + t\, s_1(x)]$$

and consider the following models defined by their densities with respect to $\lambda_{\mathbb{R}}$ the Lebesgue measure on the real line: $\mathcal{Q}_k = \left\{f_{t;k} \cdot \lambda_{\mathbb{R}} : t \in [-1, 1]\right\}$, $k \in \mathbb{N}^*$, and

$$\mathcal{P} = \{f_t \cdot \lambda_{\mathbb{R}} : t \in [-1, 1]\}, \qquad \mathcal{P}_k = \mathcal{P} \cup \bigcup_{m=k}^{\infty} \mathcal{Q}_m, \quad k \in \mathbb{N}^*.$$

We then have:

$$\mathcal{P}_1 \supset \mathcal{P}_2 \supset \cdots \supset \mathcal{P}_k \supset \mathcal{P}_{k+1} \supset \cdots \supset \bigcap_{k=1}^{\infty} \mathcal{P}_k = \mathcal{P}.$$

To describe the corresponding tangent spaces, notice that:

$$\forall k \geq 1, \ \partial_t \log f_{t;k}(x)\big|_{t=0} = s_1(x) \quad \text{and} \quad \partial_t \log f_t(x)|_{t=0} = s_2(x),$$

and thus $\dot{\mathcal{P}} = \{a\, s_2(x) : a \in \mathbb{R}\}$,

$$\dot{\mathcal{P}}_k = \{a\, s_2(x) : a \in \mathbb{R}\} \cup \{b\, s_1(x) : b \in \mathbb{R}\}, \quad k \in \mathbb{N}^*.$$

Then $\mathcal{T} = \{a\, s_2(x) : a \in \mathbb{R}\}$,

$$\mathcal{T}_k = \{a\, s_2(x) + b\, s_1(x) : a, b \in \mathbb{R}\}, \quad k \in \mathbb{N}^*.$$

This shows that:

$$\bigcap_{k=1}^{\infty} \dot{\mathcal{P}}_k \supsetneq \dot{\mathcal{P}} \qquad \text{and} \qquad \bigcap_{k=1}^{\infty} \mathcal{T}_k \supsetneq \mathcal{T},$$

even if the decreasing sequence of models $\{\mathcal{P}_k\}_{k\in\mathbb{N}^*}$ is such that $\bigcap_{k=1}^{\infty} \mathcal{P}_k = \mathcal{P}$.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

A.2. Comments on the Assumptions T and SP

Let $\varphi_1(Z), \ldots, \varphi_J(Z)$ be bounded vector-valued functions of dimensions $d_1, \ldots, d_J$, respectively. For each $j, k \in \{1, \ldots, J\}$, we are looking for a $p_j \times d_k-$matrix $\alpha_k^{(j)}(\underline{X})$ with bounded entries such that:

$$b_j(Z) = \alpha_1^{(j)}(\underline{X})\varphi_1(Z) + \cdots + \alpha_J^{(j)}(\underline{X})\varphi_J(Z) \in \mathbb{R}^{p_j}, \qquad j \in \{1, \ldots, J\},$$

satisfy a strong form of Assumption T.1, *i.e.*:

$$\mathbb{E}\left(b_j\, g_i' \,|\, \underline{X}\right) = 0, \qquad \forall i, j \in \{1, \ldots, J\},\ i \neq j.$$

Recall that for each $i \in \{1, \ldots, J\}$, g_i takes values in $\mathbb{R}^{p_i}$. Then one has the following system for $\alpha_1^{(j)}, \ldots, \alpha_J^{(j)}$:

$$\sum_{k=1}^{J} \alpha_k^{(j)}\, \mathbb{E}\left(\varphi_k\, g_i' \,|\, \underline{X}\right) = 0, \qquad \forall i, j \in \{1, \ldots, J\},\ i \neq j. \tag{A.3}$$

For a given $j \in \{1, \ldots, J\}$, there are $p_j \times \sum_{i=1, i\neq j}^{J} p_j$ equations with $p_j \times \sum_{k=1}^{J} d_k$ unknown functions (the components of $\alpha_k^{(j)}$'s). If $d_k = p_k$ and the components of $\alpha_k^{(k)}$ are taken as constant (for example $\alpha_k^{(k)} \equiv 1$), $1 \leq k \leq J$, then one obtains a unique solution for $\alpha_1^{(j)}, \ldots, \alpha_J^{(j)}$ from equations (A.3), given a suitable choice for the φ_k's. The functions $g_1, \ldots, g_J$ defining the model, would be natural candidates for the φ_k's if $d_k = p_k$. However, the functions $g_1, \ldots, g_J$ are not necessarily bounded and some adjustments are required. For a subset $A \subset \text{supp}Z$, we use the following notations: $g_{i,A} = g_i(Z, \theta_0)\mathbf{1}\{Z \in A\}$, $i = 1, \ldots, J$. If we take the subset A of the support of Z such that $g_{1,A}, \ldots, g_{J,A}$ are bounded, we can choose $\varphi_k = g_{k,A}$, $k = 1, \ldots, J$. Noting that

$$\mathbb{E}\left(g_{j,A}\, g_i' \,|\, \underline{X}\right) = \mathbb{E}\left(g_{j,A}\, g_{i,A}' \,|\, \underline{X}\right), \qquad \forall i, j \in \{1, \ldots, J\},$$

with a suitable choice of A, one can take:

$$b_j = \alpha_1^{(j)}\, g_{1,A} + \cdots + g_{j,A} + \cdots + \alpha_J^{(j)}\, g_{J,A} \in \mathbb{R}^{p_j}, \qquad j \in \{1, \ldots, J\}, \tag{A.4}$$

in Assumption T. Here, $\alpha_k^{(j)} = \alpha_k^{(j)}(\underline{X})$ are the solutions of the system:

$$\sum_{k=1}^{J} \alpha_k^{(j)}\, \mathbb{E}\left(g_{k,A}\, g_{i,A}' \,|\, \underline{X}\right) = 0, \qquad \forall i, j \in \{1, \ldots, J\},\ i \neq j, \tag{A.5}$$

where $\alpha_j^{(j)} = (1, \ldots, 1)' \in \mathbb{R}^{p_j}$ for $j \in \{1, \ldots, J\}$. With this particular choice of functions b_j, Assumption T can be replaced by the following assumption.

Assumption T'. There exist a subset $A \subset \text{supp}Z$ such that the functions $b_1, \ldots, b_J$ defined by (A.4) and (A.5), satisfy Assumptions T.2 and T.3.

Let us detail the construction of the functions $b_1, \ldots, b_J$ in the particular case $J = 2$. Define:

$$b_i = g_{i,A} - \mathbb{E}\left(g_{i,A}\, g_{j,A}' \mid X^{(1)}, X^{(2)}\right) \mathbb{E}^{-1}\left(g_{j,A}\, g_{j,A}' \mid X^{(1)}, X^{(2)}\right) g_{j,A}, \tag{A.6}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

for $(i,j) \in \{(1,2),(2,1)\}$ where $\mathbb{E}^{-1}\left(g_{j,A}\, g'_{j,A} \mid X^{(1)}, X^{(2)}\right)$ stands for the inverse of the matrix $\mathbb{E}\left(g_{j,A}\, g'_{j,A} \mid X^{(1)}, X^{(2)}\right)$ that is supposed to exist. Assumption T could then be replaced by the following assumption.

Assumption T" (The case $J = 2$). There exist a subset $A \subset \text{supp}Z$ such that the functions $g_{1,A},\ g_{2,A}$ are bounded and $b_1,\ b_2$ defined by (A.6) satisfy Assumptions T.2 and T.3 and, for $i \in \{1,2\}$, $\mathbb{E}\left(g_{i,A}\, g'_{i,A} \mid X^{(1)}, X^{(2)}\right)$ is invertible (a.s.) and

$$\left\| \mathbb{E}^{-1}\left(g_{i,A}\, g'_{i,A} \mid X^{(1)}, X^{(2)}\right) \right\|_{\infty} < \infty \quad (a.s.).$$

If the functions g_1 and g_2 are bounded, Assumption T".1 with $A = suppZ$, is slightly stronger than usual assumptions appearing in the literature (see, for example, Assumptions 1 and 2 of Newey (2004), Assumptions 2 and 3 of Ai and Chen (2012)). Meanwhile, our results are derived under Assumption T, which is also more general than Assumption T", being closely related to the usual regularity conditions appearing in the literature, but not exactly comparable to them. Let us also point out that Assumption T.1 is not required in the case $J = 1$ where one can simply take $b_1 = g_{1,A}$.

Note that Assumption SP.1 does not necessarily mean that the parameters θ and η are completely separated, as is sometimes implicitly assumed in the literature. In fact θ and η are connected since the functional parameter η can have θ among its arguments. Assumption SP only means that when considering the density of $P_{\theta,\eta}$ with respect to a dominating measure μ, we could write it in the form

$$f(\cdot,\theta,\eta(\upsilon(\cdot,\theta))),$$

with f and υ having a known form, where $f(\cdot,\theta_0,\eta(\upsilon(\cdot,\theta_0)))$ and $f(\cdot,\theta,\eta_0(\upsilon(\cdot,\theta)))$ belong to the model $\mathcal{P}$ for every $\theta \in \Theta$ and $\eta \in H$. For example, in the conditional mean setting with one conditioning vector

$$\mathbb{E}[Y - m(X,\theta) \mid X] = 0,$$

we can take H as the set of zero conditional mean densities of $Z = (Y', X')'$, *i.e.,*

$$H = \left\{ p(y,x)\cdot\gamma(x) :\ p \geq 0,\ \gamma \geq 0,\ \int p(y,x)\,dy = 1,\ \int y p(y,x)\,dy = 0,\ \forall x, \right.$$
$$\left. \int \gamma(x)\,dx = 1 \right\}$$

and $\upsilon(y,x,\theta) = (y - m(x,\theta), x)$, so that:

$$\eta(\upsilon(z,\theta)) = \eta(y - m(x,\theta), x) = p(y - m(x,\theta), x)\cdot\gamma(x)$$

and

$$f(z,\theta,\eta(\upsilon(z,\theta))) = \eta(\upsilon(z,\theta)).$$

In the proof of Theorem 1, we identify the density $f(\cdot,\theta,\eta(\upsilon(\cdot,\theta)))$ with the infinite dimensional nuisance parameter η, which is itself a density.

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

A.3. Proof of Theorem 1

For any $k \in \mathbb{N}^*$, let $\mathcal{P}_k$ be the model defined by equation (4) and $\mathcal{P}$ the model defined by equation (1). Then:

$$\mathcal{P}_1 \supset \mathcal{P}_2 \supset \cdots \supset \mathcal{P}_k \supset \mathcal{P}_{k+1} \supset \cdots \supset \bigcap_{k=1}^{\infty} \mathcal{P}_k = \mathcal{P}.$$

Hence the stated result is a direct consequence of Lemma A.1, provided that condition (A.2) holds for the tangent spaces of $\mathcal{P}$ and $\mathcal{P}_k$, $k \in \mathbb{N}^*$, at θ_0.

For each $j \in \{1, \ldots, J\}$, any $z \in \mathcal{Z} \subset \mathbb{R}^q$ could be partitioned into two subvectors $y^{(j)} \in \mathbb{R}^{q-q_j}$ and $x^{(j)} \in \mathbb{R}^{q_j}$ with $x^{(j)}$ in the support of $X^{(j)}$. Recall that $P_{X^{(j)}}$ denotes the law of $X^{(j)}$. Model $\mathcal{P}$ is then defined by the set of conditions:

$$\int g_j(z,\theta)\, f(z,\theta)\, dy^{(j)} = 0 \quad P_{X^{(j)}} - a.s., \quad j \in \{1, \ldots, J\}; \tag{A.7}$$

for a fixed k, the model $\mathcal{P}_k$ is defined by:

$$\int g_j(z,\theta)\, f(z,\theta)\, w_r^{(j)}\left(x^{(j)}\right) dz = 0, \quad j \in \{1, \ldots, J\}, \quad r \in \{1, \ldots, k\}. \tag{A.8}$$

Consider now a regular parametric family $\{f_t\}_{t \in (-\varepsilon, \varepsilon)}$ of densities satisfying (A.7), that means that there exist parameters $\theta_t \in \Theta$ such that, for any $t \in (-\varepsilon, \varepsilon)$ and $P_{X^{(j)}} - a.s.$,

$$\int g_j(z,\theta_t)\ f_t(z,\theta_t)\ dy^{(j)} = 0, \quad \forall j \in \{1, \ldots, J\}. \tag{A.9}$$

Let

$$\dot{\theta} = \left.\frac{\partial \theta_t}{\partial t}\right|_{t=0},$$

$$s = \partial_t \log f_t(Z,\theta_0)|_{t=0}, \quad S_{\theta_0} = \partial_\theta \log f(Z,\theta)|_{\theta=\theta_0},$$

$$s_1 = \partial_t \log f_t(Z,\theta_t)|_{t=0} = s + S'_{\theta_0}\, \dot{\theta}.$$

Here and in the following, the derivatives of the log-densities are to be understood in the mean square sense, see Ibragimov and Has'minskii (1981, p. 64). Differentiating with respect to t in (A.9), we obtain:

$$\partial_{\theta'} \mathbb{E}\left[g_j(Z,\theta_0)\,|X^{(j)}\right] \dot{\theta} + \mathbb{E}\left[g_j(Z,\theta_0)\ s_1(Z) \mid X^{(j)}\right] = 0, \tag{A.10}$$

$\forall j \in \{1, \ldots, J\}$. Since $\dot{\theta} \in \mathbb{R}^d$ could be arbitrary, we deduce that:

$$\partial_{\theta'} \mathbb{E}\left[g_j(Z,\theta_0) \mid X^{(j)}\right] + \mathbb{E}\left[g_j(Z,\theta_0)\, S'_{\theta_0}(Z) \mid X^{(j)}\right] = 0,$$

$$\mathbb{E}\left[g_j(Z,\theta_0)\, s(Z) \mid X^{(j)}\right] = 0,$$

for each $j \in \{1, \ldots, J\}$. The last equation and the expression of the score functions s_1, suggest a tangent space $\mathcal{T} = \mathcal{T}(\mathcal{P}, P_0)$ of the form:

$$\mathcal{T} = \overline{\text{lin}}\, S_{\theta_0} + \Big\{ s : \mathbb{E}\left(s^2\right) < \infty,\ \mathbb{E}(s) = 0, \\ \mathbb{E}\left[g_j(Z,\theta_0)\, s(Z) \mid X^{(j)}\right] = 0,\ 1 \le j \le J \Big\}. \tag{A.11}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

On the other hand, the tangent space $\mathcal{T}_k = \mathcal{T}(\mathcal{P}_k, P_0)$ corresponding to the model defined by the equations (A.8), is given by vectors satisfying the unconditional moment equations:

$$\partial_{\theta'} \mathbb{E}\left[g_j(Z,\theta_0)\ w_r^{(j)}\left(X^{(j)}\right)\right]\dot{\theta} + \mathbb{E}\left[g_j(Z,\theta_0)\, s_1(Z)\ w_r^{(j)}\left(X^{(j)}\right)\right] = 0, \tag{A.12}$$

$1 \le j \le J$, $1 \le r \le k$. This yields the tangent spaces:

$$\mathcal{T}_k = \overline{\text{lin}}\, S_{\theta_0} + \Big\{s : \mathbb{E}(s^2) < \infty,\ \mathbb{E}(s) = 0,\ \mathbb{E}\left[g_j(Z,\theta_0)\ s(Z)\ w_r^{(j)}\left(X^{(j)}\right)\right] = 0,$$
$$\forall\, 1 \le j \le J,\ \forall\, 1 \le r \le k\Big\};$$

see for instance Bickel et al. (1993, Example 3, Sect. 3.2). Since the functions $w_r^{(j)}\left(X^{(j)}\right)$, $r \in \mathbb{N}^*$, span $L^2\left(P_{X^{(j)}}\right)$, equations (A.10) are satisfied if and only if equations (A.12) are satisfied for any $k \in \mathbb{N}^*$. In other words, the equivalent of the spanning condition of Newey (2004), see our equation (A.2) above, is satisfied and we can apply Lemma A.1 to conclude that $I_{\theta_0} = \lim_{k\to\infty} I_{\theta_0}^{(k)}$.

The proof will be complete if we show that the tangent space $\mathcal{T} = \mathcal{T}(\mathcal{P}, P_0)$ is indeed the set described in equation (A.11). It is quite easy to see that equations (A.10) guarantee the inclusion "$\subset$" in display (A.11). To show the reverse inclusion, since $\overline{\text{lin}}\, S_{\theta_0} \subset \mathcal{T}$ by the definition of the tangent space, it suffices to prove that $\mathcal{T}' \subset \mathcal{T}$, where:

$$\mathcal{T}' = \mathcal{T}'(\mathcal{P}, P_0)$$
$$= \left\{s : \mathbb{E}(s^2) < \infty,\ \mathbb{E}(s) = 0,\ \mathbb{E}\left[g_j(Z,\theta_0)\, s(Z) \mid X^{(j)}\right] = 0,\ \ 1 \le j \le J\right\}.$$

Let f_0 denote the true density of the vector Z. Take $s \in \mathcal{T}'$ and suppose for the moment, that s is bounded. Then, for real numbers t with sufficiently small absolute values, the functions $f_t = (1 + t \cdot s)\, f_0$ are densities on $\mathcal{Z}$ and if $\mathbb{E}_{f_t}$ denotes expectation with respect to the law defined by f_t,

$$\mathbb{E}_{f_t}\left[g_j(Z,\theta_0)\ a\left(X^{(j)}\right)\right] = \mathbb{E}\left[g_j(Z,\theta_0)\ a\left(X^{(j)}\right)\right]$$
$$+ t\ \mathbb{E}\left[g_j(Z,\theta_0)\ s(Z)\ a\left(X^{(j)}\right)\right] = 0,$$

for any square integrable function $a\left(X^{(j)}\right)$, so that $\mathbb{E}_{f_t}\left[g_j(Z,\theta_0) | X^{(j)}\right] = 0$, $1 \le j \le J$. Moreover,

$$\partial_t \log f_t|_{t=0} = \partial_t \log(1 + t \cdot s)|_{t=0} = s,$$

which means that the family of densities $\{f_t\}_{|t|<\varepsilon}$ defines a submodel of model (1), for which the tangent vector at $t = 0$ is exactly s. Next, we have to extend the argument to unbounded functions s. If $\mathcal{M} \subset L^2(P_0)$ is the subspace of bounded functions of Z, it remains for us to show that $\mathcal{M} \cap \mathcal{T}'$ is dense in $\mathcal{T}'$. One may consider this step obvious since any unbounded square integrable function can be approximated by a sequence of bounded functions, see for instance Ai and Chen (2003, p. 1838). We argue that this well-known approximation result, cannot be directly applied to our context, as is also the case in other contexts considered in the efficiency bounds literature. Indeed, here we are in the following situation: we have J infinite-dimensional closed subspaces $\mathcal{T}_1', \ldots, \mathcal{T}_J'$ such that

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

$\mathcal{T}' = \mathcal{T}'_1 \cap \ldots \cap \mathcal{T}'_J$, $\overline{\mathcal{M} \cap \mathcal{T}'_j} = \mathcal{T}'_j$, $1 \le j \le J$, and we need $\overline{\mathcal{M} \cap \mathcal{T}'} = \mathcal{T}'$. To the best of our knowledge, there is no general mathematical result which would allow us to claim that $\mathcal{M} \cap \mathcal{T}'$ is dense in $\mathcal{T}'$ without any further argument. That is why we have to provide a proof adapted to the case we consider herein, for which our Assumption T is well suited.

Herein, the norm of a vector (or matrix) should be understand as the sum of componentwise norms. Since $\mathcal{M}$ is dense in $L^2(P_0)$, for a fixed $s \in \mathcal{T}'$ there exist a sequence $\{t_n\}_n \subset \mathcal{M}$ such that:

$$\|s - t_n\|_{L^2(P_0)} \underset{n\to\infty}{\longrightarrow} 0 \quad \text{and} \quad \max_{1\le k\le J} \left\| \mathbb{E}\left(t_n g'_k \mid X^{(k)}\right) \right\|_\infty < \infty \quad (a.s.). \tag{A.13}$$

We define

$$u_n = t_n - \sum_{k=1}^{J} \mathbb{E}\left(t_n g'_k \mid X^{(k)}\right) \mathbb{E}^{-1}\left(b_k g'_k \mid X^{(k)}\right) b_k.$$

Since $\{t_n\}_n \subset \mathcal{M}$ satisfies the condition (A.13), by Assumption T.2 and T.3 we have, for each $k \in \{1, \ldots, J\}$,

$$\left\| \mathbb{E}\left(t_n g_k \mid X^{(k)}\right) \mathbb{E}^{-1}\left(b_k g'_k \mid X^{(k)}\right) b_k \right\|_\infty < \infty \quad (a.s.),$$

and thus $u_n \in \mathcal{M}$. Then, for $j \in \{1, \ldots, J\}$, using Assumption T.1,

$$\begin{aligned}
&\mathbb{E}\left(g_j\, u'_n \,|\, X^{(j)}\right) \\
&\quad = \mathbb{E}\left(g_j\, t'_n \mid X^{(j)}\right) - \sum_{k=1}^{J} \mathbb{E}\Big[g_j\, b'_k\, \mathbb{E}^{-1}\left(g_k\, b'_k \,|\, X^{(k)}\right) \mathbb{E}\left(g_k\, t'_n \mid X^{(k)}\right) \,|\, X^{(j)} \Big] \\
&\quad = \mathbb{E}\left(g_j\, t'_n \mid X^{(j)}\right) \\
&\qquad - \sum_{k=1}^{J} \mathbb{E}\left[\underbrace{\mathbb{E}\left(g_j b'_k \,|\, X^{(j)}, X^{(k)}\right)}_{=0 \text{ if } j\neq k} \mathbb{E}^{-1}\left(g_k b'_k \,|\, X^{(k)}\right) \mathbb{E}\left(g_k t'_n \,|\, X^{(k)}\right) \,|\, X^{(j)} \right] \\
&\quad = \underline{\mathbb{E}\left(g_j t'_n \mid X^{(j)}\right)} - \underline{\mathbb{E}\left(g_j b'_j \mid X^{(j)}\right)} \underline{\mathbb{E}^{-1}\left(g_j b'_j \,|\, X^{(j)}\right)} \underline{\mathbb{E}\left(g_j t'_n \,|\, X^{(j)}\right)} = 0.
\end{aligned} \tag{A.14}$$

Moreover,

$$\begin{aligned}
s - u_n &= s - t_n + t_n - u_n \\
&= s - t_n + \sum_{k=1}^{J} \mathbb{E}\Big[(t_n - s)\, g'_k \,|\, X^{(k)} \Big] \mathbb{E}^{-1}\left(b_k g'_k \,|\, X^{(k)}\right) b_k,
\end{aligned}$$

which, by Assumption T.2, entails

$$\begin{aligned}
&\|s - u_n\|_{L^2(P_0)} \\
&\quad \le \|s - t_n\|_{L^2(P_0)} + C \cdot c_b \sum_{k=1}^{J} \|b_k\|_\infty \left\| \mathbb{E}\Big[(t_n - s)\, g'_k \,|\, X^{(k)} \Big] \cdot \mathbf{1}\{b_k(Z) \neq 0\} \right\|_{L^2(P_0)},
\end{aligned}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

for some constant C depending on $d_1, \ldots, d_J$. Let

$$\delta_k^2\left(X^{(k)}\right) = \mathbb{E}\left(\mathbf{1}\{b_k(Z) \neq 0\} \mid X^{(k)}\right).$$

Noting that, for $k \in \{1, \ldots, J\}$,

$$\begin{aligned}
&\left\| \mathbb{E}\left[(t_n - s)\, g_k \mid X^{(k)}\right] \cdot \mathbf{1}\{b_k \neq 0\} \right\|^2_{L^2(P_0)} \\
&\qquad = \mathbb{E}\left\{ \mathbb{E}\left[(t_n - s)\, \delta_k\, g_k' \mid X^{(k)}\right] \cdot \mathbb{E}\left[(t_n - s)\, \delta_k\, g_k \mid X^{(k)}\right] \right\} \\
\text{(Cauchy}-\text{Schwarz)} \quad &\leq \mathbb{E}\left\{ \mathbb{E}\left[(t_n - s)^2 \mid X^{(k)}\right] \mathbb{E}\left(\delta_k^2\, g_k'\, g_k \mid X^{(k)}\right) \right\} \\
\text{(Assumption T.3)} \quad &\leq \|t_n - s\|^2_{L^2(P_0)} \left\| \mathbb{E}\left(g_k'\, g_k \mid X^{(k)}\right) \delta_k^2 \right\|_\infty^2,
\end{aligned}$$

we finally obtain $\|s - u_n\|_{L^2(P_0)} \longrightarrow 0$ as $n \to \infty$. In particular, deduce that $\mathbb{E}(u_n) \to 0$. Now, since all the previous equations and inequalities involving u_n also hold with u_n replaced by $u_n - \mathbb{E}(u_n)$, we deduce that $\{u_n - \mathbb{E}(u_n)\}_n \subset \mathcal{M} \cap \mathcal{T}'$, which implies that $s \in \overline{\mathcal{M} \cap \mathcal{T}'}$. The proof is now complete. ■

A.4. Contraction Property in Regression-Like Models with Missing Data

With the same notation of Section 4.2, we shall prove that the equation

$$\begin{aligned}
a_1^*\left(X^*\right) = \mathbb{E}\left\{ \mathbb{E}\left[a_1^*\left(X^*\right)\, \rho\left(Z, \theta_0\right) \mid W\right] \frac{1 - \pi\left(W\right)}{\pi\left(W\right)}\, \rho'\left(Z, \theta_0\right) \mid X^* \right\} \\
\times\, \mathbb{E}^{-1}\left[\frac{1}{\pi\left(W\right)}\, \rho\left(Z, \theta_0\right)\, \rho'\left(Z, \theta_0\right) \mid X^* \right]
\end{aligned} \tag{A.15}$$

has a unique solution which can be obtained by successive approximation, under the additional assumption:

$$\inf_w \pi\left(w\right) = 1 - \beta > 0, \tag{A.16}$$

the infimum being taken over all possible values of W. For simplicity, in the remainder of this subsection we drop the arguments of the functions. Let $\widetilde{\rho} = \pi^{-1/2}\rho$. Assuming that $\mathbb{E}\left(\widetilde{\rho}\, \widetilde{\rho}\,' \mid X^*\right)$ is invertible, equation (A.15) can be equivalently written in the form:

$$\begin{aligned}
a_1^*\, \widetilde{\rho} &= \mathbb{E}\left[\mathbb{E}\left(a_1^*\, \rho \mid W\right) \frac{1 - \pi}{\pi}\, \rho' \mid X^* \right] \mathbb{E}^{-1}\left(\frac{1}{\pi}\, \rho\, \rho' \mid X^* \right) \widetilde{\rho} \\
&= \mathbb{E}\left[\mathbb{E}\left(a_1^*\, \widetilde{\rho} \mid W\right) (1 - \pi)\, \widetilde{\rho}\,' \mid X^* \right] \mathbb{E}^{-1}\left(\widetilde{\rho}\, \widetilde{\rho}\,' \mid X^*\right) \widetilde{\rho} \\
&=: \widetilde{T}\left(a_1^*\, \widetilde{\rho}\right).
\end{aligned}$$

We will show that the map $\widetilde{T}$ is a contraction. Before that, let us state a Cauchy-Schwarz inequality for matrix valued random variables, a version of an inequality in Lavergne (2008): let $\mathbb{E}$ denote the conditional expectation given an arbitrary $\sigma-$field, let $A \in \mathbb{R}^n \times \mathbb{R}^p$ and $B \in \mathbb{R}^n \times \mathbb{R}^q$ be random matrices such that $\mathbb{E}(tr(A'A)), \mathbb{E}(tr(B'B)) < \infty$ and $\mathbb{E}(A'A)$ is

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

nonsingular. Then the matrix $\mathbb{E}(B'B) - \mathbb{E}(B'A)\mathbb{E}^{-1}(A'A)\mathbb{E}(A'B)$ is positive semidefinite, with equality if and only if $B = A\mathbb{E}^{-1}(A'A)\mathbb{E}(A'B)$.[4] We also use the following notation: for any symmetric matrices B_1, B_2, $B_1 \gg B_2$ means $B_1 - B_2$ is positive semidefinite. Let us write:

$$\begin{aligned}\mathbb{E}\left[\widetilde{T}\left(a_1^*\ \widetilde{\rho}\right)\ \widetilde{T}'\left(a_1^*\ \widetilde{\rho}\right)\right] &= \mathbb{E}\left\{\left[\mathbb{E}\left(a_1^*\ \widetilde{\rho} \mid W\right)\ (1-\pi)\ \widetilde{\rho}\,' \mid X^*\right]\mathbb{E}^{-1}\left(\widetilde{\rho}\ \widetilde{\rho}\,' \mid X^*\right)\ \widetilde{\rho}\right.\\ &\qquad \left.\times\ \widetilde{\rho}\,'\,\mathbb{E}^{-1}\left(\widetilde{\rho}\ \widetilde{\rho}\,' \mid X^*\right)\ \left\{\left[\mathbb{E}\left(a_1^*\ \widetilde{\rho} \mid W\right)\ (1-\pi)\ \widetilde{\rho}\,' \mid X^*\right]\right\}'\right\}\\ &= \mathbb{E}\left\{\left[\mathbb{E}\left(a_1^*\ \widetilde{\rho} \mid W\right)\ (1-\pi)\ \widetilde{\rho}\,' \mid X^*\right]\mathbb{E}^{-1}\left(\widetilde{\rho}\ \widetilde{\rho}\,' \mid X^*\right)\right.\\ &\qquad \left.\times\ \left\{\left[\mathbb{E}\left(a_1^*\ \widetilde{\rho} \mid W\right)\ (1-\pi)\ \widetilde{\rho}\,' \mid X^*\right]\right\}'\right\}\end{aligned}$$

$$\begin{aligned}\text{(Cauchy-Schwarz)}\quad &\ll \mathbb{E}\left\{\mathbb{E}\left[\mathbb{E}\left(a_1^*\ \widetilde{\rho} \mid W\right)\ (1-\pi)^2\ \mathbb{E}\left(\widetilde{\rho}\,'\ a_1^{*\prime} \mid W\right) \mid X^*\right]\right\}\\ &= \mathbb{E}\left[\mathbb{E}\left(a_1^*\ \widetilde{\rho} \mid W\right)\ (1-\pi)^2\ \mathbb{E}\left(\widetilde{\rho}\,'\ a_1^{*\prime} \mid W\right)\right]\\ \text{(Cauchy-Schwarz)}\quad &\ll \mathbb{E}\left[(1-\pi)^2\ \left(a_1^*\ \widetilde{\rho}\right)\ \left(a_1^*\ \widetilde{\rho}\right)'\right].\end{aligned}$$

This implies

$$\begin{aligned}\left\|\widetilde{T}\left(a_1^*\ \widetilde{\rho}\right)\right\|_{L^2}^2 &= \mathbb{E}\left\{tr\left[\widetilde{T}'\left(a_1^*\ \widetilde{\rho}\right)\ \widetilde{T}\left(a_1^*\ \widetilde{\rho}\right)\right]\right\} = tr\left\{\mathbb{E}\left[\widetilde{T}\left(a_1^*\ \widetilde{\rho}\right)\ \widetilde{T}'\left(a_1^*\ \widetilde{\rho}\right)\right]\right\}\\ &\le \sup_w\left[1-\pi\ (w)\right]\ \left\|a_1^*\ \widetilde{\rho}\right\|_{L^2}^2 \le \beta\ \left\|a_1^*\ \widetilde{\rho}\right\|_{L^2}^2,\end{aligned}$$

where $\beta = \sup_w[1-\pi\ (w)] = 1 - \inf_w \pi\ (w) < 1$ by assumption (A.16). Deduce that $\widetilde{T}$ is a contracting map.

A.5. Efficient Score with Parametric Selection Probability in Regression-Like Models with Missing Data

Let $X^{(1)} = X^*$, $X^{(2)} = W$ and the parameter vector $\theta = \left(\alpha', \gamma'\right)'$. Moreover, let:

$$g_1(Z,\theta) = \frac{\delta}{\pi\ (W,\gamma)}\ \rho\left(Y, X^*, \alpha\right), \qquad g_2(Z,\theta) = \frac{\delta}{\pi\ (W,\gamma)} - 1,$$

$$\overline{S}_\theta = \overline{a}_1\left(X^*\right)\ g_1(Z,\theta) + \overline{a}_2\left(W\right)\ g_2(Z,\theta) = \begin{pmatrix}\overline{S}_\alpha\\ \overline{S}_\gamma\end{pmatrix},$$

where

$$\begin{aligned}\overline{a}_1\left(X^*\right) &= \overline{a}_1\left(X^{(1)}\right)\\ &= \begin{pmatrix}-\partial_\alpha\mathbb{E}\left[\mathbb{E}\left(\pi^{-1}(W,\gamma_0)\delta \mid X^*, W\right)\rho' \mid X^*\right]\mathbb{E}^{-1}\left(\pi^{-1}(W,\gamma_0)\rho\ \rho' \mid X^*\right)\\ 0\end{pmatrix}\\ &\quad + \mathbb{E}\left\{\mathbb{E}\left[\overline{a}_1\left(X^*\right)\rho \mid W\right]\ \left(\pi^{-1}(W,\gamma_0) - 1\right)\rho' \mid X^*\right\}\mathbb{E}^{-1}\left(\pi^{-1}(W,\gamma_0)\rho\rho' \mid X^*\right).\end{aligned}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

If we partition $\overline{a}_1(X^*)$ in $\overline{a}_1(X^*) = \begin{pmatrix} \overline{a}_{1,\alpha}(X^*) \\ \overline{a}_{1,\gamma}(X^*) \end{pmatrix}$ and we use the same short notation as previously, the preceding equations can be written as:

$$\begin{aligned}
\overline{a}_{1,\alpha}(X^*) = &-\partial_\alpha \mathbb{E}\left(\mathbb{E}\left(\frac{\delta}{\pi} \mid X^*, W\right)\rho' \mid X^*\right) \mathbb{E}^{-1}\left(\frac{1}{\pi}\,\rho\,\rho' \mid X^*\right) \\
&+\mathbb{E}\left\{\mathbb{E}\left[\overline{a}_{1,\alpha}(X^*)\,\rho \mid W\right]\left(\frac{1}{\pi}-1\right)\rho' \mid X^*\right\} \\
&\times \mathbb{E}^{-1}\left(\frac{1}{\pi}\,\rho\,\rho' \mid X^*\right), \\
\overline{a}_{1,\gamma}(X^*) = &\,\mathbb{E}\left\{\mathbb{E}\left[\overline{a}_{1,\gamma}(X^*)\,\rho \mid W\right]\left(\frac{1}{\pi}-1\right)\rho' \mid X^*\right\} \\
&\times \mathbb{E}^{-1}\left(\frac{1}{\pi}\,\rho\,\rho' \mid X^*\right),
\end{aligned}$$

with the obvious solution $\overline{a}_{1,\gamma} \equiv 0$ for the subvector of $\overline{a}_1$ corresponding to γ (possibly not the unique solution, but any solution yields the same efficient score $\overline{S}_\theta$). Similar calculations can be done for $\overline{a}_2(W)$:

$$\begin{aligned}
\overline{a}_2(W) &= \overline{a}_2\left(X^{(2)}\right) \\
&= \begin{pmatrix} 0 \\ \\ \dfrac{1}{\pi}\,\partial_\gamma \pi \end{pmatrix} \frac{\pi}{1-\pi} - \mathbb{E}\left[\overline{a}_1(X^*)\,\rho \mid W\right],
\end{aligned}$$

which gives, for $\overline{a}_2(W) = \begin{pmatrix} \overline{a}_{2,\alpha}(W) \\ \overline{a}_{2,\gamma}(W) \end{pmatrix}$,

$$\overline{a}_{2,\alpha}(W) = -\mathbb{E}\left[\overline{a}_{1,\alpha}(X^*)\,\rho \mid W\right]$$

$$\overline{a}_{2,\gamma}(W) = \frac{1}{1-\pi}\,\partial_\gamma \pi - \mathbb{E}\left[\overline{a}_{1,\gamma}(X^*)\,\rho \mid W\right] = \frac{1}{1-\pi}\,\partial_\gamma \pi.$$

Therefore,

$$\overline{S}_\theta = \overline{a}_1(X^*)\,g_1 + \overline{a}_2(W)\,g_2 = \begin{pmatrix} \overline{S}_\alpha \\ \overline{S}_\gamma \end{pmatrix} = \begin{pmatrix} \overline{a}_{1,\alpha}(X^*)\,g_1 + \overline{a}_{2,\alpha}(W)\,g_2 \\ \\ \overline{a}_{2,\gamma}(W)\,g_2 \end{pmatrix},$$

where

$$\begin{aligned}
\overline{a}_{1,\alpha}(X^*) = &-\partial_\alpha \mathbb{E}\left(\rho' \mid X^*\right) \mathbb{E}^{-1}\left(\frac{1}{\pi}\,\rho\,\rho' \mid X^*\right) \\
&+\mathbb{E}\left\{\mathbb{E}\left[\overline{a}_{1,\alpha}(X^*)\,\rho \mid W\right]\frac{1-\pi}{\pi}\,\rho' \mid X^*\right\} \\
&\times \mathbb{E}^{-1}\left(\frac{1}{\pi}\,\rho\,\rho' \mid X^*\right),
\end{aligned}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press

$$\overline{a}_{2,\alpha}(W) = -\mathbb{E}\left[\overline{a}_{1,\alpha}(X^*)\ \rho \mid W\right],$$

$$\overline{a}_{2,\gamma}(W) = \frac{\pi}{1-\pi}\,\partial_\gamma \pi \qquad \left(= \frac{\pi(W,\gamma_0)}{1-\pi(W,\gamma_0)}\,\partial_\gamma \pi(W,\gamma_0)\right).$$

Now, for any $s = b(W)\cdot g_2 = b(W)\left(\frac{\delta}{\pi(W,\gamma_0)} - 1\right) \in \mathcal{T}_2^{\perp}$, we have

$$\begin{aligned}
\mathbb{E}\left(\overline{S}_\alpha\ s' \mid W\right) &= \mathbb{E}\left[\overline{S}_\alpha\ \left(\frac{\delta}{\pi(W,\gamma_0)} - 1\right) \mid W\right] b'(W) \\
&= \mathbb{E}\left\{\left[\overline{a}_{1,\alpha}(X^*)\ \frac{\delta}{\pi}\,\rho + \overline{a}_{2,\alpha}(W)\left(\frac{\delta}{\pi} - 1\right)\right]\left(\frac{\delta}{\pi} - 1\right) \mid W\right\} b'(W) \\
&= \left\{\mathbb{E}\left[\overline{a}_{1,\alpha}(X^*)\ \rho \mid W\right] + \overline{a}_{2,\alpha}(W)\right\}\left(\frac{1}{\pi} - 1\right) b'(W) \\
&= \left\{\mathbb{E}\left[\overline{a}_{1,\alpha}(X^*)\ \rho \mid W\right] - \mathbb{E}\left[\overline{a}_{1,\alpha}(X^*)\ \rho \mid W\right]\right\}\left(\frac{1}{\pi} - 1\right) b'(W) \\
&= 0,
\end{aligned}$$

so that, since $\overline{S}_\gamma = \overline{a}_{2,\gamma}(W)\cdot g_2$, we obtain:

$$\mathbb{E}\left(\overline{S}_\alpha\ \overline{S}'_\gamma\right) = \mathbb{E}\left[\mathbb{E}\left(\overline{S}_\alpha\ \overline{S}'_\gamma \mid W\right)\right] = 0.$$

This means that the efficient score S^*_α for α, equal to the residual of the (componentwise) projection of $\overline{S}_\alpha$ on $\overline{S}_\gamma$, coincides with $\overline{S}_\alpha$,

$$S^*_\alpha = \overline{S}_\alpha - \mathbb{E}\left(\overline{S}_\alpha\ \overline{S}'_\gamma\right)\mathbb{V}^{-1}\left(\overline{S}_\gamma\right)\ \overline{S}_\gamma = \overline{S}_\alpha,$$

and has the same expression, as already noted in Robins et al. (1994), as in the case where $\pi(W)$ is completely known:

$$\begin{aligned}
S^*_\alpha = \overline{S}_\alpha &= \overline{a}_{1,\alpha}(X^*)\ g_1 + \overline{a}_{2,\alpha}(W)\ g_2 \\
&= a^*_1(X^*)\ g_1 + a^*_2(W)\ g_2.
\end{aligned}$$

https://doi.org/10.1017/S0266466615000080 Published online by Cambridge University Press