Review of Economic Studies (2014) **81**, 919–943 doi:10.1093/restud/rdu011

Advance access publication 14 April 2014

Asymptotic Efficiency of Semiparametric Two-step GMM

DANIEL ACKERBERG
University of Michigan
XIAOHONG CHEN
Yale University
JINYONG HAHN
UCLA
and
ZHIPENG LIAO
UCLA

First version received December 2012*; final version accepted February* 2014 (*Eds.*)

Many structural economics models are semiparametric ones in which the unknown nuisance functions are identified via non-parametric conditional moment restrictions with possibly non-nested or overlapping conditioning sets, and the finite dimensional parameters of interest are over-identified via unconditional moment restrictions involving the nuisance functions. In this article we characterize the semiparametric efficiency bound for this class of models. We show that semiparametric two-step optimally weighted GMM estimators achieve the efficiency bound, where the nuisance functions could be estimated via any consistent non-parametric methods in the first step. Regardless of whether the efficiency bound has a closed form expression or not, we provide easy-to-compute sieve-based optimal weight matrices that lead to asymptotically efficient two-step GMM estimators.

Key words: Overlapping information sets, Semiparametric efficiency, Two-step GMM.

JEL Codes: C14, C31, C32

1. INTRODUCTION

We consider the semiparametric efficiency bound and efficient estimation of a finite dimensional parameter of interest θ_o that is (possibly over-) identified by the unconditional moment restrictions

$$E\left[g\left(Z;\theta_o,h_{1,o}(\cdot),...,h_{L,o}(\cdot)\right)\right]=0, \tag{1}$$

where the nuisance functions $h_o(\cdot)=\left(h_{1,o}(\cdot),...,h_{L,o}(\cdot)\right)$ are identified by the conditional moment restrictions

$$E\left[\rho_\ell\left(Z,h_{\ell,o}(X_\ell)\right)\middle|X_\ell\right]=0 \text{ almost surely } X_\ell, \quad \ell=1,...,L. \tag{2}$$

Here, $Z=\left(Y',X'\right)'$ are random vectors, X is the union of distinct elements of X_ℓ, $\ell=1,...,L$. The unknown functions $h_{\ell,o}(\cdot)$, $\ell=1,...,L$, are distinct from each other; while $h_{\ell,o}(\cdot)$ enters (2) through $h_{\ell,o}(X_\ell)$ only, it may enter (1) through its values at all support points of X_ℓ. This class of models is flexible enough to allow for models of semiparametric mean and quantile treatment effects, missing data, sample selection, default, entry, censoring, some models with semiparametric control function approach and many more.

Given a random sample $\{Z_i\}_{i=1}^n$ of Z, we can exploit the conditional moment restrictions (2), and estimate $h_{\ell,o}$ by any non-parametric estimator $\widehat{h}_\ell$ for $\ell=1,...,L$. We can then estimate θ_o in (1) by setting the sample analogue $n^{-1}\sum_{i=1}^n g\left(Z_i;\theta,\widehat{h}\right)$ of $E[g(Z;\theta,h_o)]$ as close to zero as possible. This intuitive strategy is called a semiparametric two-step GMM procedure.[1] Alternatively, one could compute an optimally weighted GMM estimator *jointly* using moment restrictions (1) and (a finite yet increasing number of unconditional moments implied by) (2).

The two-step procedure often has significant computational advantages over the joint estimation procedure, which explains its popularity among empirical researchers estimating complicated structural models. Examples are the recent literatures on estimating production functions (*e.g.* Olley and Pakes, 1996) and dynamic models (*e.g.* Hotz and Miller, 1993). In both cases, the joint approach would require a large-dimensional non-linear search over h and θ simultaneously. In contrast, the two-step approach can be computed with two sequential estimation procedures, the first-step estimating h, and the second-step estimating θ. Generally speaking, the latter is easier computationally, both in terms of computational time and in terms of reliability. Moreover, in many cases h can be conveniently specified such that the first step of the two-step approach is either analytically computable (*e.g.* least squares) or the solution to a globally concave optimization problem (*e.g.* a logit model). This can further decrease computational time and increase reliability (*e.g.* with a globally concave optimization problem, the global maximum is the only local maximum).

However, the two-step procedure may also have disadvantages relative to the joint procedure. Any inference based on a semiparametric two-step GMM estimator $\widehat{\theta}_n$ is a "limited information" inference in the sense that the information contained in moment conditions (1) and (2) are not simultaneously considered. Intuitively, the joint approach might be more efficient than a semiparametric two-step GMM estimator, but to the best of our knowledge, formal semiparametric efficiency results are so far only established for the cases where the non-parametric first stage (2) takes the form of sequential moment restrictions.[2] We pose a natural question whether the "limited information" estimation strategy in fact exhausts all the information in model (1) and (2). Such a question was posed earlier in a finite dimensional GMM context by Crepon *et al.* (1997), who noted that the limited information strategy in fact achieves full efficiency as long as the first step estimator is exactly identified. Newey and Powell (1999) considered optimality of the second step estimator conditional on a given first-step non-parametric estimator, and noted in some examples that the efficient second step estimator is fully efficient when the first step non-parametric estimator

1. The root-n asymptotic normality of a semiparametric two-step GMM estimator $\widehat{\theta}_n$ (of θ_o) and the consistent estimation of the asymptotic variance of $\widehat{\theta}_n$ have been studied in the existing literature. See, *e.g.* Andrews (1994), Newey (1994), Pakes and Olley (1995), Chen *et al.* (2003), Ackerberg *et al.* (2012) and the references therein.

2. See Chamberlain (1992) and Ai and Chen (2012), *e.g.* for the semiparametric efficiency bound and efficient estimation of such sequential moment restriction models. Even in the sequential moment restriction case, our approach and estimator has an advantage over Ai and Chen's (2012) two-step efficient estimator. Their efficient estimator takes a fairly complicated form involving a filtering/orthogonalization procedure in which additional non-parametric components (conditional covariances) are estimated. Our estimator does not require estimating these additional components and is very similar to two-step approaches typically used in the parametric literature. On the other hand, unlike Ai and Chen (2012), our efficiency results require that the non-parametric objects are exactly identified.

is exactly identified. We build on these papers, and show that Newey and Powell's (1999) insight holds in general.

We derive the semiparametric efficiency bound for θ_o in the model (1) and (2). The efficiency bound is calculated by establishing the bound for θ_o in a transformed model

$$E\left[\widetilde{g}\left(Z;\theta_o,h_{1,o}(\cdot),...,h_{L,o}(\cdot)\right)\right]=0$$

where $\left(h_{1,o},...,h_{L,o}\right)$ are known, and relating the bound there to the asymptotic variance of the two step estimator. The transformed model is such that it is orthogonal to the non-parametric moment (2).[3] As noted above, we find that when θ_o is estimated in the second step by GMM using the unconditional moment (1) with an optimal weight matrix that reflects the noise in estimating the nuisance functions h_o, the resulting semiparametric two-step GMM estimators achieve the semiparametric efficiency bound for θ_o. The semiparametric efficiency bound for θ_o may not have a closed form expression in general, and hence it may be difficult to compute a feasible optimal weight matrix based on any non-parametric first step. However, when the nuisance functions are estimated via a simple sieve M procedure in the first step, we provide easy-to-compute optimal weight matrices that lead to asymptotically efficient two-step GMM estimators.

Our result leads to a convenient practical implication that the two-step GMM estimator may reduce computational burden without sacrificing efficiency. As long as practitioners use the weight matrix reflecting the noise of estimating h_o, the "limited information" inference exhausts all the information in the model. Besides the practical implication, we believe that our result is of interest from a theoretical perspective as well; we allow the conditioning variables X_ℓ, $\ell=1,...,L$, to be nested, overlapping, or non-nested (different from each other and have arbitrary overlaps). To the best of our knowledge, our paper is the first to compute the efficiency bound for θ_o when the sets of non-parametric conditional moment restrictions (2) could be non-nested or overlapping. We illustrate how such non-nested conditional moment models often arise in the literatures based on Olley and Pakes (1996) and Hotz and Miller (1993). In a brief Monte-Carlo study based on Olley and Pakes, we compare the small sample properties of an efficient joint estimator, an efficient two-step estimator based on our methodology, and a slightly simpler "naive" two-step estimator that is not necessarily efficient. We find that the small sample properties of the efficient two-step estimator are similar to the efficient joint estimator, and superior to a naive two-step estimator.

The rest of the article is organized as follows. Section 2 establishes the semiparametric efficiency bound for θ_o. Readers who would like to avoid technical details can jump directly to Section 3, where the main result of Section 2 is rephrased in a more intuitive way and some of its practical implications are discussed. Section 4 presents examples and Monte Carlo results, and Section 5 provides a short summary. Additional proofs and technical derivations are gathered in the Appendix.

2. SEMIPARAMETRIC EFFICIENCY BOUND

In this section, we derive the semiparametric efficiency bound for θ_o when the unknown parameters $\alpha_o=(\theta_o,h_o)\in\Theta\times\mathcal{H}$ are identified by the sets of moment restrictions (1) and (2). We first introduce some notation and definitions used in this article. $E(\cdot)$ and $Var(\cdot)$ are computed with respect to the true unknown distribution F_o of Z. Let Θ be a compact set in $\mathcal{R}^{d_\theta}$ that contains an open ball centering at $\theta_o\in int(\Theta)$. For $\ell=1,...,L$, we assume that the nuisance function space $\mathcal{H}_\ell$ is a linear subspace of the space of square integrable functions with respect

3. The exact nature of the transformed model will be discussed later in Section 2.1.

to X_ℓ. The moment functions $g(\cdot)$ and $\rho_\ell(\cdot)$ are respectively $d_g \times 1$ and $d_\ell \times 1$ vector valued, with $d_g \geq d_\theta$ and $d_\ell = \dim(h_\ell(x_\ell))$ for $\ell = 1, ..., L$. Let $\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}$ be the $d_g \times d_\theta$ matrix valued ordinary (partial) derivative of the function $E[g(Z,\theta,h)]$ with respect to θ evaluated at (θ_o, h_o). Let $\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial h_\ell}[v_\ell]$ be the $d_g \times 1$ vector valued pathwise derivative of $E[g(Z,\theta,h)]$ with respect to h_ℓ, evaluated at (θ_o, h_o), in the direction $v_\ell \in \mathcal{H}_\ell - \{h_{\ell,o}\}$

$$\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial h_\ell}[v_\ell] = \left.\frac{\partial E\left[g\left(Z,\theta_o,h_{\ell,o}+\tau v_\ell, h_{-\ell,o}\right)\right]}{\partial \tau}\right|_{\tau=0} \tag{3}$$

where $h_{-\ell,o} = (h_{1,o}, ..., h_{\ell-1,o}, h_{\ell+1,o}, ..., h_{L,o})$. Let $m_\ell(X_\ell, h_\ell) = E\left[\rho_\ell\left(Z, h_{\ell,o}(X_\ell)\right)\middle| X_\ell\right]$. In this article, because any $h_\ell \in \mathcal{H}_\ell$ and $v_\ell \in \mathcal{V}_\ell$ are restricted to be measurable functions of X_ℓ, and because the conditional moment function $m_\ell(X_\ell, h_\ell)$ depends on h_ℓ only through $h_\ell(X_\ell)$, the pathwise derivative $\frac{\partial m_\ell(X_\ell, h_{\ell,o})}{\partial h_\ell}[v_\ell]$ takes a simple form $\frac{\partial m_\ell(X_\ell, h_{\ell,o}(X_\ell)+\tau v_\ell(X_\ell))}{\partial \tau}|_{\tau=0}$. To stress this fact, we let $\frac{\partial m_\ell(x_\ell, h_{\ell,o}(x_\ell))}{\partial h'_\ell}$ be a $d_\ell \times d_\ell$ matrix-valued (ordinary derivative) function such that

$$\frac{\partial m_\ell(X_\ell, h_{\ell,o}(X_\ell))}{\partial h'_\ell} v_\ell(X_\ell) = \left.\frac{\partial m_\ell(X_\ell, h_{\ell,o}+\tau v_\ell)}{\partial \tau}\right|_{\tau=0} \quad \text{for all } v_\ell \in \mathcal{V}_\ell, \tag{4}$$

where $v_\ell(X_\ell)$ is a $d_\ell \times 1$ vector-valued function of X_ℓ. For any $v_\ell, \widetilde{v}_\ell \in \mathcal{H}_\ell - \{h_{\ell,o}\}$, we define the following inner product

$$\langle v_\ell, \widetilde{v}_\ell \rangle_\ell = E\left[v_\ell(X_\ell)' \left(\frac{\partial m_\ell(X_\ell, h_{\ell,o}(X_\ell))}{\partial h'_\ell} \right)' \frac{\partial m_\ell(X_\ell, h_{\ell,o}(X_\ell))}{\partial h'_\ell} \widetilde{v}_\ell(X_\ell) \right]. \tag{5}$$

Finally, we say that $\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial h_\ell}[\cdot]$ is a bounded (or regular) linear functional on $\mathcal{V}_\ell$ if $\frac{\partial E[g_j(Z,\theta_o,h_o)]}{\partial h_\ell}[\cdot]$ is a bounded linear functional on $\mathcal{V}_\ell$ for all $j = 1, ..., d_g$, *i.e.*

$$\max_{1 \leq j \leq d_g} \sup_{v \neq 0, v \in \mathcal{V}_\ell} \frac{\left| \frac{\partial E[g_j(Z,\theta_o,h_o)]}{\partial h_\ell}[v] \right|^2}{\langle v, v \rangle_\ell} < \infty.$$

We impose the following basic regularity condition:

Condition 1. *(i) the data $\{Z_i\}_{i=1}^n$ is a random sample drawn from the unknown $F_o(\cdot)$; (ii) (θ_o, h_o) satisfies model (1) - (2), $\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}$ has full (column) rank d_θ; (iii) $\frac{\partial m_\ell(X_\ell, h_{\ell,o}(X_\ell))}{\partial h'_\ell}$ is invertible almost surely - X_ℓ for $\ell = 1, ..., L$; (iv) $\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial h_\ell}[\cdot]$ is a bounded linear functional on $\mathcal{V}_\ell$ for $\ell = 1, ..., L$.*

Under Conditions 1(ii) and (iii), the unknown θ_o could be over identified by the unconditional moment restrictions (1) if h_o were known, but the unknown function h_o is "exactly" identified by the conditional moment restrictions (2).

Our main efficiency bound result is contained in the following theorem. We need to define an object $\mathbf{v}^*_\ell(X_\ell)$ for this purpose. By Condition 1(iv) and the Riesz representation theorem, we have: for each $j = 1, ..., d_g$, there is a unique $u^*_{\ell,j} \in \mathcal{V}_\ell$ such that

$$\frac{\partial E\left[g_j(Z,\theta_o,h_o)\right]}{\partial h_\ell}[v_\ell] = \left\langle u^*_{\ell,j}, v_\ell \right\rangle_\ell = E\left[\left(\frac{\partial m_\ell(X_\ell, h_{\ell,o})}{\partial h_\ell}[u^*_{\ell,j}] \right)' \left(\frac{\partial m_\ell(X_\ell, h_{\ell,o})}{\partial h_\ell}[v_\ell] \right) \right] \tag{6}$$

for all $v_\ell \in \mathcal{V}_\ell$. (See Ai and Chen (2003, 2007) for use of Riesz representation in a related context.)

Let

$$\mathbf{v}_\ell^*(X_\ell) \equiv \begin{bmatrix} v_{\ell,1}^*(X_\ell)' \\ \vdots \\ v_{\ell,d_g}^*(X_\ell)' \end{bmatrix} = \begin{bmatrix} \left(\frac{\partial m_\ell(X_\ell, h_{\ell,o})}{\partial h_\ell}[u_{\ell,1}^*]\right)' \\ \vdots \\ \left(\frac{\partial m_\ell(X_\ell, h_{\ell,o})}{\partial h_\ell}[u_{\ell,d_g}^*]\right)' \end{bmatrix}, \tag{7}$$

which is a $d_g \times d_\ell$ matrix valued function. Having defined $\mathbf{v}_\ell^*(X_\ell)$, we are able to state our main theorem.

Theorem 1. *Let Condition 1 hold. Define*

$$\widetilde{g}(Z,\theta,h) = g(Z,\theta,h) - \sum_{\ell=1}^{L} \mathbf{v}_\ell^*(X_\ell)\rho_\ell(Z, h_\ell(X_\ell)) \tag{8}$$

with $\mathbf{v}_\ell^*(\cdot)$ ($\ell = 1,...,L$) defined in equation (7). If $Var(\widetilde{g}(Z,\theta_o,h_o))$ is non-singular, then the semiparametric information bound for θ_o is

$$\left(\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}\right)' [Var(\widetilde{g}(Z,\theta_o,h_o))]^{-1} \left(\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}\right). \tag{9}$$

Proof Proof, along with discussion, is presented in Subsection 2.1. ‖

This semiparametric efficiency bound result is very general. In addition to allow for non-overlapping or arbitrarily overlapped conditional moment restrictions, to allow for over identified GMM restrictions, it also allows for moment functions $g(Z,\theta,h)$ and $\rho_\ell(Z,h_\ell(X_\ell)), \ell = 1,...,L$ to be pointwise non-smooth with respect to parameters. This efficiency bound is derived using a new technique based on an orthogonality argument. The orthogonalization has an interesting relationship to adjustment of the influence function for estimation of the unknown $h_o()$, which are discussed in Subsection 2.1.

2.1. *Proof of Theorem 1*

We first develop a semiparametric information bound under an extra zero derivative restriction (10).

Lemma 1. *Let Condition 1 hold and $Var(g(Z,\theta_o,h_o))$ be non-singular. If for all $\ell = 1,...,L$, the restriction*

$$\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial h_\ell}[v_\ell] = 0 \text{ for all } v_\ell \in \mathcal{V}_\ell \tag{10}$$

is satisfied, then the semiparametric information bound for θ_o is

$$\left(\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}\right)' [Var(g(Z,\theta_o,h_o))]^{-1} \left(\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}\right). \tag{11}$$

Proof Proof in Appendix. ‖

Lemma 1 shows that when the effects of estimating unknown h_o on the moment conditions $E[g(Z,\theta_o,h_o)] = 0$ are ruled out, the semiparametric efficiency bound of θ_o only relies on $E[g(Z,\theta_o,h_o)] = 0$ with assuming h_o to be known.

We now argue that the implication of Lemma 1 is not limited to the case where the zero derivative condition (10) is satisfied. This is because we can always transform the model such that the moment condition $E[g(Z,\theta_o,h_o)]=0$ is equivalent to $E[\widetilde{g}(Z,\theta_o,h_o)]=0$ under (2) and moreover

$$\frac{\partial E[\widetilde{g}(Z,\theta_o,h_o)]}{\partial h_\ell}[v_\ell]=0 \text{ for all } v_\ell\in\mathcal{V}_\ell,\ \ell=1,...,L, \tag{12}$$

where the pathwise derivative $\frac{\partial E[\widetilde{g}(Z,\theta_o,h_o)]}{\partial h_\ell}[v_\ell]$ of $\widetilde{g}(Z,\theta,h)$ is defined similarly to that in equation (3).

To prove Theorem 1, we present a systematic method of transforming the model (1) such that the zero derivative restriction (12) is always satisfied by the transformed moment $\widetilde{g}(Z,\theta,h)$ defined in equation (8). Equations (6)–(7) imply that $\mathbf{v}^*_\ell(\cdot)\,(\ell=1,...,L)$ can be equivalently defined as solution to

$$\frac{\partial E\left[g_j(Z,\theta_o,h_o)\right]}{\partial h_\ell}[v_\ell]=E\left[v^*_{\ell,j}(X_\ell)'\left(\frac{\partial m_\ell(X_\ell,h_{\ell,o})}{\partial h_\ell}[v_\ell]\right)\right] \quad \text{for all } v_\ell\in\mathcal{V}_\ell \tag{13}$$

for each $j=1,...,d_g$. We also have for each $j=1,...,d_g$,

$$\begin{aligned}\frac{\partial E\left[v^*_{\ell,j}(X_\ell)'\rho_\ell\left(Z,h_{\ell,o}(X_\ell)\right)\right]}{\partial h_\ell}[v_\ell]&=\left.\frac{\partial E\left[v^*_{\ell,j}(X_\ell)'\rho_\ell(Z,h_{\ell,o}(X_\ell)+\tau v_\ell(X_\ell))\right]}{\partial\tau}\right|_{\tau=0}\\&=\left.\frac{\partial E\left[v^*_{\ell,j}(X_\ell)'m_\ell(X_\ell,h_{\ell,o}(X_\ell)+\tau v_\ell(X_\ell))\right]}{\partial\tau}\right|_{\tau=0}\\&=E\left[v^*_{\ell,j}(X_\ell)'\left(\frac{\partial m_\ell(X_\ell,h_{\ell,o})}{\partial h_\ell}[v_\ell]\right)\right],\end{aligned}$$

where the last equal sign holds under the assumption allowing for interchanging the expectation and differentiation.

By definition of $\widetilde{g}(Z,\theta,h)=g(Z,\theta,h)-\sum_{\ell=1}^L\mathbf{v}^*_\ell(X_\ell)\rho_\ell(Z,h_\ell(X_\ell))$ in equation (8), we have for all $j=1,...,d_g$,

$$\begin{aligned}\frac{\partial E\left[\widetilde{g}_j(Z,\theta_o,h_o)\right]}{\partial h_\ell}[v_\ell]&=\frac{\partial E\left[g_j(Z,\theta_o,h_o)\right]}{\partial h_\ell}[v_\ell]-E\left[v^*_{\ell,j}(X_\ell)'\left(\frac{\partial m_\ell(X_\ell,h_{\ell,o})}{\partial h_\ell}[v_\ell]\right)\right]\\&=0 \quad \text{for all } v_\ell\in\mathcal{V}_\ell \text{ by equation (13)},\end{aligned}$$

which implies that

$$\frac{\partial E[\widetilde{g}(Z,\theta_o,h_o)]}{\partial h_\ell}[v_\ell]=0 \quad \text{for all } v_\ell\in\mathcal{V}_\ell,\ \ell=1,...,L. \tag{14}$$

We also have:

$$\frac{\partial E[\widetilde{g}(Z,\theta_o,h_o)]}{\partial\theta'}=\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial\theta'}. \tag{15}$$

Under the conditional moment restrictions (2), the original unconditional moment condition $E[g(Z,\theta_o,h_o)]=0$ and the transformed moment condition $E[\widetilde{g}(Z,\theta_o,h_o)]=0$ are equivalent;

$$E[\widetilde{g}(Z,\theta_o,h_o)]=0 \quad \text{if and only if} \quad E[g(Z,\theta_o,h_o)]=0. \tag{16}$$

From equations (14), (15), and (16), Lemma 1 is applicable with the transformed moment $E[\widetilde{g}(Z,\theta_o,h_o)]=0$ and hence Theorem 1 holds.

2.2. *Special cases*

The semiparametric efficiency bound stated in Theorem 1 depends on the functions $\mathbf{v}_\ell^*(\cdot)$ $(\ell=1,...,L)$, which are characterized by equation (13) but may not have simple closed form expressions in general.

We now consider a special case where the functions $\mathbf{v}_\ell^*(\cdot)$ $(\ell=1,...,L)$ and hence the efficiency bound could be solved more explicitly. In the following we let $\frac{\partial E[g(Z,\theta_o,h_o)|X_\ell]}{\partial h_\ell}[v_\ell]$ be the pathwise derivative of the function $E[g(Z,\theta_o,h_o)|X_\ell]$ with respective to h_ℓ in the direction $v_\ell \in \mathcal{V}_\ell$

$$\frac{\partial E[g(Z,\theta_o,h_o)|X_\ell]}{\partial h_\ell}[v_\ell] = \left.\frac{\partial E[g\big(Z,\theta_o,h_{\ell,o}+\tau v_\ell,h_{-\ell,o}\big)|X_\ell]}{\partial \tau}\right|_{\tau=0}.$$

Lemma 2. *Let all the conditions of Theorem 1 hold. If for all $\ell=1,...,L$ there is a $d_g \times d_\ell$ matrix valued square integrable function $D_\ell(X_\ell,\theta_o,h_o)$ of X_ℓ such that for all $v_\ell \in \mathcal{V}_\ell$,*

$$D_\ell(X_\ell,\theta_o,h_o)v_\ell(X_\ell) = \frac{\partial E[g(Z,\theta_o,h_o)|X_\ell]}{\partial h_\ell}[v_\ell]. \tag{17}$$

Then the conclusion of Theorem 1 holds with

$$\widetilde{g}(Z,\theta,h) = g(Z,\theta,h) - \sum_{\ell=1}^{L} D_\ell(X_\ell,\theta_o,h_o)\left(\frac{\partial m_\ell(X_\ell,h_{\ell,o}(X_\ell))}{\partial h_\ell'}\right)^{-1}\rho_\ell(Z,h_\ell(X_\ell)). \tag{18}$$

Proof By equations (13) and (17), we have: for each $j=1,...,d_g$,

$$E\left\{\left[D_{\ell,j}(X_\ell,\theta_o,h_o) - v_{\ell,j}^*(X_\ell)'\left(\frac{\partial m_\ell(X_\ell,h_{\ell,o}(X_\ell))}{\partial h_\ell'}\right)\right]v_\ell(X_\ell)\right\} = 0$$

for all $v_\ell \in \mathcal{V}_\ell$. In particular, it holds for

$$v_\ell = D_{\ell,j}(X_\ell,\theta_o,h_o) - v_{\ell,j}^*(X_\ell)'\left(\frac{\partial m_\ell(X_\ell,h_{\ell,o}(X_\ell))}{\partial h_\ell'}\right),$$

which means that

$$D_{\ell,j}(X_\ell,\theta_o,h_o) = v_{\ell,j}^*(X_\ell)'\left(\frac{\partial m_\ell(X_\ell,h_{\ell,o}(X_\ell))}{\partial h_\ell'}\right) \text{ almost surely } X_\ell.$$

By Condition 1(iii), we obtain

$$\mathbf{v}_\ell^*(X_\ell) = D_\ell(X_\ell,\theta_o,h_o)\left(\frac{\partial m_\ell(X_\ell,h_{\ell,o}(X_\ell))}{\partial h_\ell'}\right)^{-1} \text{ almost surely } X_\ell. \tag{19}$$

The conclusion now follows immediately from Theorem 1 under equations (8) and (19). ‖

If the unconditional moment restrictions (1) take the special form

$$E[g\big(Z,\theta_o,h_{1,o}(X_1),...h_{L,o}(X_L)\big)] = 0, \tag{20}$$

i.e. if the moment function g depends on $h_o(\cdot)$ only through $\left(h_{1,o}(X_1),...h_{L,o}(X_L)\right)$, then equation (17) is trivially satisfied with

$$D_\ell(X_\ell,\theta_o,h_o)=\frac{\partial E[g\left(Z,\theta_o,h_{\ell,o}(X_\ell),h_{-\ell,o}\left(X_{-\ell}\right)\right)|X_\ell]}{\partial h'_\ell},\ \ell=1,...,L,$$

which could be viewed as an ordinary partial derivative defined similarly as that in equation (4). We next give two examples when the unconditional moment restrictions (1) is of the special form (20).

Example 1 (Non-parametric Regression) For $\ell=1,...,L$, the unknown function $h_{\ell,o}$ is identified by the conditional mean restriction: $E\left[\left.Y_\ell-h_{\ell,o}(X_\ell)\right|X_\ell\right]=0$. Then: $\frac{\partial m_\ell(X_\ell,h_{\ell,o}(X_\ell))}{\partial h'_\ell}=-1$ and

$$\widetilde{g}(Z,\theta,h)=g(Z,\theta,h)+\sum_{\ell=1}^{L}\frac{\partial E[g(Z,\theta_o,h_o)|X_\ell]}{\partial h'_\ell}(Y_\ell-h_\ell(X_\ell)).$$

Suppose further that (i) $L=2$, (ii) $Z=\left(Z_1,Z_2,Z'_3,Z'_4,Z'_5,Z'_6\right)$ such that Z_1 and Z_2 are scalars, and (iii) $g=\left(g'_1,g'_2\right)$ with

$$g_1(Z,\theta,h)=Z_5\times(Z_1-q_1(Z_3,h_1(X_1);\theta))$$
$$g_2(Z,\theta,h)=Z_6\times(Z_2-q_2(Z_4,h_2(X_2);\theta))$$

for some parametrically specified scalar-valued functions q_1 and q_2. (The models discussed in Section 4 follows a similar structure.) We then have $\widetilde{g}=\left(\widetilde{g}'_1,\widetilde{g}'_2\right)$ with

$$\widetilde{g}_1(Z,\theta,h)=Z_5\times(Z_1-q_1(Z_3,h_1(X_1);\theta))-Z_5\times\frac{\partial E[q_1\left(Z_3,h_{1,o}(X_1);\theta_o\right)|X_1]}{\partial h_1}(Y_1-h_1(X_1))$$

$$\widetilde{g}_2(Z,\theta,h)=Z_6\times(Z_2-q_2(Z_4,h_2(X_2);\theta))-Z_6\times\frac{\partial E[q_2\left(Z_4,h_{2,o}(X_2);\theta_o\right)|X_2]}{\partial h_2}(Y_2-h_2(X_2)).$$

Example 2 (Non-parametric Quantile Regression) For $\ell=1,...,L$, the unknown function $h_{\ell,o}$ is identified by the conditional quantile restriction: $E\left[\left.\tau-I\{Y_\ell\le h_{\ell,o}(X_\ell)\}\right|X_\ell\right]=0$. Denote $U_\ell=Y_\ell-h_{\ell,o}(X_\ell)$. Let $f_{U_\ell}(\cdot|X_\ell)$ be the conditional density of U_ℓ given X_ℓ. Then: $\frac{\partial m_\ell(X_\ell,h_{\ell,o}(X_\ell))}{\partial h'_\ell}=-f_{U_\ell}(0|X_\ell)$ and

$$\widetilde{g}(Z,\theta,h)=g(Z,\theta,h)+\sum_{\ell=1}^{L}\frac{\partial E[g(Z,\theta_o,h_o)|X_\ell]}{\partial h'_\ell}\frac{(\tau-I\{Y_\ell\le h_\ell(X_\ell)\})}{f_{U_\ell}(0|X_\ell)}.$$

3. DISCUSSION

3.1. *Intuition*

In order to gain an intuition underlying our result, consider a simple model

$$E[g(Z;\theta_o,\beta_o)]=0, \tag{21}$$

where the finite dimensional parameter β_o can be identified by

$$E[\rho(Z;\beta_o)]=0 \tag{22}$$

We assume that the nuisance parameter β_o is exactly identified[4] by (22) in the sense that $\dim(\rho) = \dim(\beta)$. We also assume that even if β_o were known, the finite-dimensional parameter of interest θ_o is possibly overidentified by (21) in the sense $\dim(g) \geq \dim(\theta)$. Note that if the distribution of X has known, finite support, then the semiparametric moment conditions in (1) and (2) can be written in (21) and (22).

The information bound for θ_o can be obtained by the inverse of the upper-left block of

$$\left(\frac{\partial E[\varphi(Z;\theta_o,\beta_o)]}{\partial(\theta',\beta')}\right)' E\left[\varphi(Z;\theta_o,\beta_o)\varphi(Z;\theta_o,\beta_o)'\right]^{-1}\left(\frac{\partial E[\varphi(Z;\theta_o,\beta_o)]}{\partial(\theta',\beta')}\right),$$

where we define φ by stacking ρ and g vertically, *i.e.* $\varphi' = (\rho', g')$.

Assume further that

$$\frac{\partial E[g(Z;\theta_o,\beta_o)]}{\partial\beta'} = 0. \tag{23}$$

Under the regularity condition that $\partial E[\rho(Z;\beta_o)]/\partial\beta'$ is non-singular, it is straightforward to show that the asymptotic variance bound of θ is equal to the inverse of

$$\left(\frac{\partial E[g(Z;\theta_o,\beta_o)]}{\partial\theta'}\right)' E\left[g(Z;\theta_o,\beta_o)g(Z;\theta_o,\beta_o)'\right]^{-1}\left(\frac{\partial E[g(Z;\theta_o,\beta_o)]}{\partial\theta'}\right). \tag{24}$$

Now, if the assumption (23) is violated, we can consider the following transformation:

$$\widetilde{g}(Z;\theta,\beta) = g(Z;\theta,\beta) - \mathbf{v}^*\rho(Z;\beta_o) \tag{25}$$

such that

$$\frac{\partial E[\widetilde{g}(Z;\theta_o,\beta_o)]}{\partial\beta'} = 0 \tag{26}$$

i.e.

$$\mathbf{v}^* = \left(\frac{\partial E[g(Z;\theta_o,\beta_o)]}{\partial\beta'}\right)\left(\frac{\partial E[\rho(Z;\beta_o)]}{\partial\beta'}\right)^{-1}$$

The asymptotic variance of the optimal GMM estimator for the moments

$$E[\widetilde{g}(Z;\theta_o,\beta_o)] = 0$$
$$E[\rho(Z;\beta_o)] = 0$$

which is obtained by a non-singular transformation of the model (21) and (22), is identical to the original model, but satisfies the zero derivative restriction (26). Therefore, we can conclude that the asymptotic variance bound of θ is in general equal to the inverse of

$$\left(\frac{\partial E[\widetilde{g}(Z;\theta_o,\beta_o)]}{\partial\theta'}\right)' E\left[\widetilde{g}(Z;\theta_o,\beta_o)\widetilde{g}(Z;\theta_o,\beta_o)'\right]^{-1}\left(\frac{\partial E[\widetilde{g}(Z;\theta_o,\beta_o)]}{\partial\theta'}\right)$$
$$= \left(\frac{\partial E[g(Z;\theta_o,\beta_o)]}{\partial\theta}\right)' E\left[\widetilde{g}(Z;\theta_o,\beta_o)\widetilde{g}(Z;\theta_o,\beta_o)'\right]^{-1}\left(\frac{\partial E[g(Z;\theta_o,\beta_o)]}{\partial\theta}\right),$$

whether the zero derivative restriction (23) is satisfied or not.

4. The discussion in this subsection reflects an anonymous referee's insight. It also reflects Whitney Newey's insight that he kindly shared with us in a private communication.

The $\mathbf{v}^*$ in (25) can also be given the following interpretation. Let $\widehat{\beta}$ denote a method of moments estimator solving the sample counterpart of the exactly identified model (22). Standard arguments can be used to show that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g\left(Z_i;\theta_o,\widehat{\beta}\right)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(Z_i;\theta_o,\beta_o)-\left(\frac{\partial E[g(Z;\theta_o,\beta_o)]}{\partial \beta'}\right)\left(\frac{\partial E[\rho(Z;\beta_o)]}{\partial \beta'}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho(Z_i;\beta_o)+o_p(1)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(g(Z_i;\theta_o,\beta_o)-\mathbf{v}^*\rho(Z_i;\beta_o)\right)+o_p(1)
\end{aligned}$$

In other words, the $\mathbf{v}^*$ can be understood to be a part of the adjustment of the influence function of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g\left(Z_i;\theta_o,\widehat{\beta}\right)$ to reflect the noise of estimating $\widehat{\beta}$.

The preceding discussion allows us to provide an alternative interpretation to $E\left[\widetilde{g}(Z;\theta_o,\beta_o)\widetilde{g}(Z;\theta_o,\beta_o)'\right]$, which is used later in Section 3.3. Suppose that θ_o is known and that we define a "parameter" ξ_o by the moment equation

$$E[g(Z,\theta_o,\beta_o)-\xi_o]=0.$$

A natural estimator of ξ_o is $\widehat{\xi}$ that sets the sample moment condition $\frac{1}{n}\sum_{i=1}^{n}\left(g\left(Z_i,\theta_o,\widehat{\beta}\right)-\widehat{\xi}\right)$ equal to zero. Note that the asymptotic variance of $\widehat{\xi}$ is equal to that of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g\left(Z_i;\theta_o,\widehat{\beta}\right)$, which in turn is equal to $E\left[\widetilde{g}(Z;\theta_o,\beta_o)\widetilde{g}(Z;\theta_o,\beta_o)'\right]$ according to the discussion above.

3.2. *Practical implication of Theorem 1*

Suppose that h_o were known, then we would estimate θ_o in (1) by Hansen's (1982) optimally weighted GMM

$$\min_{\theta\in\Theta}\left[n^{-1/2}\sum_{i=1}^{n} g(Z_i,\theta,h_o)\right]' W_n^{-1}\left[n^{-1/2}\sum_{i=1}^{n} g(Z_i,\theta,h_o)\right]$$

with $W_n = Var[g(Z;\theta_o,h_o)]+o_p(1)$. Because $Var[g(Z;\theta_o,h_o)]=Avar\left(n^{-1/2}\sum_{i=1}^{n} g(Z_i,\theta_o,h_o)\right)$, the asymptotic variance of such an infeasible GMM estimator would be equal to the inverse of

$$\left(\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}\right)'\left(Avar\left(n^{-1/2}\sum_{i=1}^{n} g(Z_i,\theta_o,h_o)\right)\right)^{-1}\left(\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}\right).$$

Now h_o is in fact unknown, we may consider a feasible version of the preceding GMM estimator by replacing h_o by any consistent non-parametric estimator $\widehat{h}$ and using a weight matrix such that its probability limit is the inverse of $Avar\left(n^{-1/2}\sum_{i=1}^{n} g(Z_i,\theta_o,\widehat{h})\right)$; the asymptotic variance of such a feasible GMM estimator would be the inverse of

$$\left(\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}\right)'\left(Avar\left(n^{-1/2}\sum_{i=1}^{n} g\left(Z_i,\theta_o,\widehat{h}\right)\right)\right)^{-1}\left(\frac{\partial E[g(Z,\theta_o,h_o)]}{\partial \theta'}\right). \tag{27}$$

This feasible GMM estimator was discussed by Newey (1994), Ackerberg *et al*. (2012), among others. Recall our motivation of the paper that it is not obvious (to us at least) whether the feasible

GMM estimator exploits all the information in model (1) and (2), because it does not seem to use, *e.g.* the (conditional) covariance of the moments between (1) and (2).

A practical implication of our Theorem 1 is that (27) is indeed the semiparametric information bound for model (1) and (2), and therefore, the feasible GMM estimator discussed above is actually semiparametrically efficient. In order to understand this implication, we need to relate $Var(\widetilde{g}(Z,\theta_o,h_o))$ in the middle of (9) in Theorem 1 to the $Avar\left(n^{-1/2}\sum_{i=1}^n g\left(Z_i,\theta_o,\widehat{h}\right)\right)$ in the middle of (27):

Proposition 1. *For the model (1)–(2), suppose that Condition 1 (i) and (iv) are satisfied. We then have*

$$Var(\widetilde{g}(Z,\theta_o,h_o)) = Avar\left(n^{-1/2}\sum_{i=1}^n g\left(Z_i,\theta_o,\widehat{h}\right)\right), \tag{28}$$

where $\widetilde{g}$ is defined in (8) and $\widehat{h}$ is any consistent non-parametric estimator of h_o satisfying (2).

Proof We derive the adjustment to the influence function following Newey (1994, pp. 1360–1361). For simplicity, we will assume that $L=1$ and that h is a scalar, noting that generalization can be done in an additive way as discussed in Newey (1994, p. 1357). For a path $\{F_\tau(z)\}$ of the distribution of random variable Z, let h_τ to be the function indexed by τ such that $E_\tau[\rho(Z,h_\tau(X))|X]=0$, where $E_\tau[\cdot|X]$ denotes the conditional expectation taken under $F_\tau(z)$ with the corresponding score $S(Z)$. It follows that

$$E_\tau[w(X)\rho(Z,h_\tau(X))]=0 \tag{29}$$

for any square integrable $w(X)$. Differentiating (29) with respect to τ, we obtain

$$\frac{\partial}{\partial\tau}E_\tau[w(X)\rho(Z,h_o(X))]+\frac{\partial}{\partial\tau}E[w(X)m(X,h_\tau)]=0. \tag{30}$$

We can recall the definition of $\mathbf{v}^*(X)$, and write

$$\frac{\partial}{\partial\tau}E[g(Z,\theta_o,h_\tau)]=E\left[\mathbf{v}^*(X)\left(\frac{\partial}{\partial\tau}m(X,h_\tau)\right)\right]=\frac{\partial}{\partial\tau}E\left[\mathbf{v}^*(X)m(X,h_\tau)\right], \tag{31}$$

which together with (30) implies that

$$\frac{\partial}{\partial\tau}E[g(Z,\theta_o,h_\tau)]=-\frac{\partial}{\partial\tau}E_\tau\left[\mathbf{v}^*(X)\rho(Z,h_o(X))\right]=E_\tau\left[-\mathbf{v}^*(X)\rho(Z,h_o(X))S(Z)\right]. \tag{32}$$

It follows that the adjustment term (α in Newey's notation) is equal to $-\mathbf{v}^*(X)\rho(Z,h_o(X))$, and the influence function of $n^{-1/2}\sum_{i=1}^n g\left(Z_i,\theta_o,\widehat{h}\right)$ is equal to

$$g(Z,\theta_o,h_o)-\mathbf{v}^*(X)\rho(Z,h_o(X))=\widetilde{g}(Z,\theta_o,h_o).$$

Next, we note that the $Avar\left(n^{-1/2}\sum_{i=1}^n g\left(Z_i,\theta_o,\widehat{h}\right)\right)$ is invariant to the choice of any consistent nonparametric estimator $\widehat{h}$ of h_o, which follows from Newey's (1994, Proposition 1) observation that the asymptotic variance of a semiparametric $\sqrt{n}$-consistent estimator is independent of the types of first step consistent non-parametric estimators. ‖

Remark 1. *Note that $n^{-1}\sum_{i=1}^n g\left(Z_i,\widehat{\theta},\widehat{h}\right)g\left(Z_i,\widehat{\theta},\widehat{h}\right)'$ usually converges in probability to $Var(g(Z,\theta_o,h_o))$, which is often different from $Avar\left(n^{-1/2}\sum_{i=1}^n g\left(Z_i,\theta_o,\widehat{h}\right)\right)$.*

3.3. *Implementation*

The general expression of the information bound of θ_o in (27) indicates that under suitable regularity conditions, the second step GMM estimator $\widehat{\theta}_n$ that solves

$$\min_{\theta\in\Theta}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}g\left(Z_i,\theta,\widehat{h}\right)\right]'W_n^{-1}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}g\left(Z_i,\theta,\widehat{h}\right)\right], \tag{33}$$

is semiparametric efficient as long as the weighting matrix W_n satisfies

$$W_n = Avar\left(n^{-1/2}\sum_{i=1}^{n}g\left(Z_i,\theta_o,\widehat{h}\right)\right)+o_p(1) \tag{34}$$

for any consistent non-parametric estimator $\widehat{h}$ of h_o.

For efficient semiparametric estimation of θ_o, the crucial step is therefore to consistently estimate the asymptotic variance of $n^{-1/2}\sum_{i=1}^{n}g\left(Z_i,\theta_o,\widehat{h}\right)$. We describe how this objective can be achieved using a simple algorithm described in Ackerberg *et al*. (2012).[5] For simplicity of illustration, we assume that for $\ell=1,...,L$, the unknown function $h_{\ell,o}$ is identified by the conditional mean restriction: $E\left[Y_\ell-h_{\ell,o}(X_\ell)\middle|X_\ell\right]=0$, *i.e*. we use Example 1.

We imagine a researcher, who "pretends" that $h_\ell(X_\ell)=p_{\ell,1}(X_\ell)\beta_{(\ell),1}+\cdots+p_{\ell,K_\ell}(X_\ell)\beta_{(\ell),K_\ell}=p_\ell^{K_\ell}(x_{\ell,i})'\beta_{(\ell)}=h_\ell\left(X_\ell,\beta_{(\ell)}\right)$.[6] Our researcher equates $\widehat{h}$ with $\widehat{\beta}$, and perceives the latter to be a simple M-estimator solving the moment equation $E[\rho(Z;\beta_o)]=0$, where

$$\rho(Z;\beta)=\begin{bmatrix}p_1^{K_1}(X_1)\left(Y_1-h_1\left(X_1,\beta_{(1)}\right)\right)\\ \vdots\\ p_L^{K_L}(X_L)\left(Y_L-h_L\left(X_L,\beta_{(L)}\right)\right)\end{bmatrix}. \tag{35}$$

The researcher then perceives the problem to be a parametric problem characterized by (21) and (22) with $\rho(Z;\beta)$ defined above in (35).

Using Ackerberg *et al*. (2012), it can be seen that the following algorithm produces a feasible estimator of $Avar\left(n^{-1/2}\sum_{i=1}^{n}g\left(Z_i,\theta_o,\widehat{h}\right)\right)$:

1. Estimate $\widehat{\beta}$ by an M-estimator (e.g. OLS) solving the moment equation $E[\rho(Z;\beta_o)]=0$, which is equivalent to solving

$$\min_{\beta_{(\ell)}}\sum_{i=1}^{n}\left(Y_{\ell,i}-h_\ell\left(X_{\ell,i},\beta_{(\ell)}\right)\right)^2 \qquad \ell=1,...,L.$$

2. Using an arbitrary weight matrix, minimize the sample moment $\frac{1}{n}\sum_{i=1}^{n}g\left(Z_i,\theta,\widehat{\beta}\right)$ over θ to obtain a preliminary estimator $\overline{\theta}$ of θ_o. Note that our researcher uses a parametric specification of $\widehat{h}$, so without loss of generality, we can write $g(Z,\theta,h)=g(Z,\theta,\beta)$.

5. Ackerberg *et al*. (2012) obtain a convenient estimator of standard errors regardless of efficiency issues, but they do not discuss efficient estimation of θ_o.

6. The functions $p_{\ell,1}(X_\ell),p_{\ell,2}(X_\ell),\ldots$ are such that $h_\ell(X_\ell)$ can be well approximated by their linear combination, and $K_\ell=K_{\ell,n}$ is a function of n to be theoretically correct, although it is perceived to be fixed for our fictitious researcher.

3. Let the "parameter" ξ_o be defined by the moment

$$E[g(Z,\theta_o,\beta_o)-\xi_o]=0.$$

Pretend that $\overline{\theta}=\theta_o$ and "estimate" ξ_o with the $\widehat{\xi}$ that sets the sample moment condition $\frac{1}{n}\sum_{i=1}^{n}\left(g\left(Z_i,\overline{\theta},\widehat{\beta}\right)-\widehat{\xi}\right)$ equal to zero. Note that this estimation problem is exactly identified. In fact, it is just the mean of the moment conditions evaluated at $(\overline{\theta},\widehat{\beta})$.

4. Again consider $\overline{\theta}$ to be fixed. Note that $(\widehat{\beta},\widehat{\xi})$ from Steps 1 and 3 can be thought of as an exactly identified "parametric" estimator of (β_o,ξ_o) using the moments

$$E[\rho(Z;\beta_o)]=0,$$

$$E\left[g\left(Z,\overline{\theta},\beta_o\right)-\xi_o\right]=0.$$

Use the standard parametric GMM asymptotic variance formula[7] to estimate the variance of $(\widehat{\beta},\widehat{\xi})$. Denote by $\widehat{W}_n$ the portion of this variance matrix corresponding to $\widehat{\xi}$. $\widehat{W}_n$ is a consistent estimator of $Avar\left(n^{-1/2}\sum_{i=1}^{n}g\left(Z_i,\theta_o,\widehat{h}\right)\right)$.

5. Our second-step efficient estimator for θ_o is simply the solution to

$$\min_{\theta}\left(\frac{1}{n}\sum_{i=1}^{n}g\left(Z_i,\theta,\widehat{\beta}\right)\right)'\widehat{W}_n^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}g\left(Z_i,\theta,\widehat{\beta}\right)\right).$$

The components of the above procedure are all very familiar from the parametric GMM literature, and the procedure is not much harder than a "naive" two-step approach that computes $\widehat{W}_n$ by assessing the variance of the second-step moments assuming that $(\overline{\theta},\widehat{\beta})$ are fixed. The "artificial" parameter ξ_o is just a tool to obtain an estimate of the variance of the second-step moments that includes the variance contribution of $\widehat{\beta}$ $(\widehat{h})$. In our Monte-Carlo example, we compare this efficient two-step procedure to both a naive two-step estimator that is not necessarily efficient and the efficient joint estimator.

Remark 2. *Step 4 does require using the standard GMM asymptotic variance formula, which as usual requires computing the derivative of the moments. If analytic derivatives are not possible or feasible and one prefers not to do numeric differentiation,*[8] *an alternative way to compute* $Avar\left(n^{-1/2}\sum_{i=1}^{n}g\left(Z_i,\theta_o,\widehat{h}\right)\right)$ *in Step 4 is the bootstrap.*[9] *We have to be careful in using the bootstrap, though. Weak convergence does not imply that the bootstrap variance converges to the asymptotic variance. See Ghosh et al. (1984) or Wu (1986). Hence, if the bootstrap is to be used, a practitioner may want to avoid using the standard bootstrap variance. Alternative methods that have been shown to produce a consistent estimator of the variance include versions of truncation as suggested in Shao (1992) or Gonçalves and White (2005), and a percentile method as in Machado and Parente (2005).*

3.4. *Comparison with Chamberlain (1992) and Ai and Chen (2012)*

Chamberlain (1992) derived the efficiency bound of θ_o for the sequential moment restrictions

$$E\left[\rho_t(Y,X;\theta_o)|X^{(t)}\right]=0 \quad \text{with } \{1\}\subseteq X^{(0)}\subset X^{(1)}\subset\cdots\subset X^{(L)}$$

7. See Newey (1984), Murphy and Topel (1985), Section II of Ackerberg *et al*. (2012), or standard textbooks such as Wooldridge (2002, Chapter 12.4). Note that even though $(\widehat{\beta},\widehat{\xi})$ have been estimated in two steps, the model is exactly identified so it is equivalent to joint estimation and thus the standard GMM asymptotic variance formula is used.

8. See, *e.g*. Hong *et al*. (2010) for suggestions and caveats regarding numeric differentiation

9. See Armstrong *et al*. (2012), who established the weak convergence of the bootstrap in this case.

for $t=0,1,...,L$. Ai and Chen (2012) extended the result such that ρ_t function may depend on a nuisance function $h_o(\cdot)$. In order to derive the efficiency bound, they proceed by "orthogonalizing" the moments by working with forward filtering as in Hayashi and Sims (1983):

$$\varepsilon_L(Z,\alpha)=\rho_L(Z,\alpha),$$

$$\varepsilon_s(Z,\alpha)=\rho_s(Z,\alpha)-\sum_{t=s+1}^{L}\Gamma_{s,t}\left(X^{(t)}\right)\varepsilon_t(Z,\alpha),$$

where $\alpha=(\theta,h)$ and

$$\Gamma_{s,t}\left(X^{(t)}\right)=E\left[\rho_s(Z,\alpha_o)\varepsilon_t(Z,\alpha_o)'\middle|X^{(t)}\right]\left(\Sigma_t\left(X^{(t)}\right)\right)^{-1},$$

$$\Sigma_t\left(X^{(t)}\right)=E\left[\varepsilon_t(Z,\alpha_o)\varepsilon_t(Z,\alpha_o)'\middle|X^{(t)}\right].$$

Their efficient estimator is based on forming counterparts of $\varepsilon_s(Z,\alpha)$, which requires nonparametric estimation of the Γ's and Σ's.

We can see that our orthogonalization is quite different than Ai and Chen (2012). Theirs is with respect to covariances of the moments, whereas our orthogonalization is with respect to derivatives of the moments. Perhaps more important from an applied perspective, our orthogonalization is just a proof technique that can be bypassed in practice by exploiting the algorithm in Section 3.3. In contrast, the procedure of Ai and Chen (2012) requires nonparametric implementation of orthogonalization for estimation. As noted in the introduction, our orthogonalization is also available when the first-step conditioning variables are non-nested. On the other hand, our results are limited to the situation where the h is exactly identified, whereas the results in Ai and Chen (2012) do not have such limitation.

4. EXAMPLES AND MONTE CARLO RESULTS

We now illustrate the usefulness of our results and estimators by showing how they can be applied to two recent methodological literatures that are based on two-step semiparametric techniques. In both examples, our results imply that two-step methods do not need to sacrifice efficiency relative to joint estimation. We then do a brief Monte-Carlo study.

4.1. *Example 1: Two-Step Estimation of Dynamic Models*

Hotz and Miller (1993) initiated a large literature that uses two-step semiparametric estimators to estimate single agent dynamic programming problems and dynamic games. The main benefit of the semiparametric approach is to avoid the computational burden associated with solving dynamic programming problems. With these approaches, one can estimate structural parameters without ever having to explicitly solve agents' dynamic programming problems.

In this literature, two-step estimators are typically preferred to estimators that jointly use all the moment conditions because of a different computational issue. With a two-step approach, the non-parametric parts of the problem can often be estimated in the first step using analytic estimators (*e.g.* least squares) or estimators with a globally concave objective function (*e.g.* logit). This can not only save time but also alleviates concern regarding the reliability of a non-linear search over a large (*i.e.* asymptotically increasing) set of parameters representing the non-parametric parts of the problem.

One might worry that such two-step approaches have an efficiency cost, but our results show that this may not be the case. We illustrate this with a simple finite horizon, single agent, dynamic binary choice model. This might be appropriate, *e.g.* for a model of retirement or fertility decisions. One could also apply our results to problems with multiple agents, infinite horizons, and multinomial choice. However, since our first step relies on estimating conditional expectations, the results do not directly apply to problems with continuous choices (*e.g.* Bajari *et al.*, 2007).

Suppose that the per-period utility function for agent i making choice $y_{it} \in \{0, 1\}$ in period $t = 1, ..., T$ is given by

$$U = \begin{cases} U_0(x_{it}; \theta) + \epsilon_{i0t} & \text{if } y_{it} = 0 \\ U_1(x_{it}; \theta) + \epsilon_{i1t} & \text{if } y_{it} = 1 \end{cases}.$$

The vector x_{it} includes the state variables of the problem (*e.g.* work experience, number of children, wealth) that are observed by the econometrician. $\epsilon_{it} = \{\epsilon_{i0t}, \epsilon_{i1t}\}$ represent state variables that are not observed by the econometrician. U_0 and U_1 are known up to the finite dimensional structural parameters θ. The majority of the empirical literature thus far has assumed that ϵ_{it} are independent of x_{it} and i.i.d. over time[10], *e.g.* Type 1 Extreme Value random variables.

Assuming the evolution of the state variable x_{it} is first-order Markov, the optimal policy function in this problem is

$$y_{it} = y_t(x_{it}, \epsilon_{it}).$$

The function is indexed by t because of the finite horizon. One can also consider a "conditional choice probability"

$$E[y_{it}|x_{it}] = h_t(x_{it}) = \int y_t(x_{it}, \epsilon_{it}) p(\epsilon_{it}) d\epsilon_{it}$$

which is the probability of making choice 1 in time t given state x_{it} (prior to the agent's realization of ϵ_{it}).

The key result of Hotz and Miller (1993) is that under certain conditions, one can rewrite the dynamic programming Bellman equation in terms of conditional choice probabilities, *i.e.*

$$h_t(x_{it}) = g_t\left(x_{it}, h_{t+1}(\cdot); \theta\right) \tag{36}$$

where the g_t's are known, (relatively) easily computable, functions.[11] This representation is possible because there is a one-to-one mapping between value functions and conditional choice probabilities.

Estimation can then proceed with the following two-step procedure. In the first step, one non-parametrically estimates the T conditional choice probabilities using

$$E[y_{i1} - h_1(x_{i1}) \mid x_{i1}] = 0, \tag{37}$$

$$\vdots$$

$$E[y_{iT} - h_T(x_{iT}) \mid x_{iT}] = 0.$$

10. see, *e.g.* Pakes *et al.* (2007), Pesendorfer and Schmidt-Dengler (2008), Ryan (2012), Collard-Wexler (2012), Fang and Wang (2012). Only recently has the literature considered allowing correlation in unobservables over time, *e.g.* Aguirregabiria and Mira (2007), Kasahara and Shimotsu (2009), Hu *et al.* (2010), and Arcidiacono and Miller (2011), and it is challenging.

11. For the particularly simple form in the text, one needs one of the choices to lead to a terminal state. Hotz *et al.* (1994) consider finite horizon models without this condition—in that case, all future h's enter the right hand side. In an infinite horizon problem, the equation also has a very simple form, since the h function does not depend on time (see, *e.g.* Aguirregabiria and Mira, 2002)

In the simplest case, the h_t functions might be represented by a linear sieve, in which case the first step can be performed using simple least squares.[12] Note that this set of first stage moments falls directly into our framework of non-nested first-step conditioning sets.

In the second step, one can estimate the structural parameters θ using the following moment conditions implied by (36)

$$E\left[\left(y_{it}-g_t\left(x_{it},\widehat{h}_{t+1}(\cdot);\theta\right)\right)\otimes r(x_{it})\ \right]=0 \tag{38}$$

where the unknown functions h_{t+1} have been replaced with their estimates from the first step, $\widehat{h}_{t+1}$. This does not require explicitly solving the agents' dynamic programming problems. While the g_t function needs to be computed, this is relatively simple.[13]

In this context, our results imply that as long as one uses an appropriate weight matrix for the second step (*e.g.* the procedure described in Section 3.3), this two-step procedure does not sacrifice asymptotic efficiency relative to a joint procedure that considers both sets of moments (37) and (38) simultaneously. Again, this is important because the joint procedure requires a non-linear search over the entire parameter space $(\theta, h_1, ..., h_T)$, which is likely both more computationally demanding and less reliable than the two-step approach (which requires a non-linear search over just θ in the case where linear sieves are used in the first stage[14]).

4.2. *Example 2: two-step estimation of production functions*

We next show how our results can be applied to a version of the Olley and Pakes (1996) two-step methodology for estimating production functions. Consider a panel of firms indexed by i producing output y_{it} using inputs x_{it} across time t. The model can be described with three equations. First is the production function

$$\text{Production Function:}\quad y_{it}=F(x_{it};\theta_1)+\omega_{it}+\epsilon_{it}. \tag{39}$$

The production function contains two scalar econometric unobservables, ω_{it} and ϵ_{it}. ω_{it} is firm i's "productivity" shock in period t, and will be permitted to be correlated with input choices x_{it}. In contrast, ϵ_{it} is noise in output (*e.g.* measurement error) that is assumed to be mean independent of the firm's information set at t, I_{it}.

The second equation describes how productivity ω_{it} evolves over time. Specifically, ω_{it} is assumed to follow a first-order Markov process from the firm's perspective, *i.e.*

$$\text{Productivity Evolution:}\quad E\left[\omega_{it}|I_{it-1}\right]=\mu(\omega_{it-1};\theta_2). \tag{40}$$

The last equation describes how some other variable i_{it} is chosen by the firm at t, *i.e.*

$$\text{Proxy Choice:}\quad i_{it}=i(x_{it},\omega_{it}). \tag{41}$$

This precise definition of this "proxy" variable i_{it} differs across different formulations of these estimators. For example, in Olley and Pakes (1996), i_{it} is the firm's current investment towards

12. Alternatively, one could use a sieve logit or probit.

13. See Hotz and Miller (1993) for details. Note that our efficiency result is conditional on a given set of moments (38), *i.e.* we do not consider the optimal choice of instrument function $r(x_{it})$—for this see Pesendorfer and Schmidt-Dengler (2008) in a finite dimensional parameter context.

14. In the case where sieve logits are used in the first stage, the first step would require solving T globally concave optimization problems, and the second step would be a non-linear search over θ. Again, this is generally quicker and more reliable than a non-linear search over the full $(\theta, h_1, ..., h_T)$ space. Moreover, even if the first step does not have a global concave objective function, it will generally be easier computationally to estimate the parameters in two steps.

future physical capital. In Levinsohn and Petrin (2003), i_{it} is the firm's choice of an intermediate input, *e.g.* electricity or material input.

In the current formulation, we treat the functions F and μ parametrically, *i.e.* known up to the finite dimensional parameters θ_1 and θ_2. In contrast, the optimal proxy choice function i is treated non-parametrically. This seems somewhat natural since both F and μ can be considered economic primitives of the model, while the i function is not an economic primitive (*e.g.* in Olley and Pakes it is the solution to a complicated dynamic investment problem). That said, we should note that this differs slightly from most of the existing empirical literature, which treat both i and μ non-parametrically, and only F parametrically.[15]

The two key assumptions regarding the Proxy Choice equation are that (i) $i(x_{it}, \omega_{it})$ is strictly monotonic in ω_{it}, and (ii) ω_{it} is the only econometric unobservable in $i(x_{it}, \omega_{it})$. This implies that two firms with the same x_{it} and i_{it} have the same ω_{it}. To make these assumptions more plausible, most empirical researchers using this methodology have allowed the proxy choice equation to vary across t.

$$i_{it} = i_t(x_{it}, \omega_{it}). \tag{42}$$

This allows for the general economic environment (*e.g.* input prices, costs of investment, industry level demand, industry structure) firms are operating in to change over time. It also means that two firms with the same x_{it} and i_{it} do not necessarily have the same ω_{it} (if they are operating in different time periods).[16]

We also make the assumption that $x_{it} \in I_{it-1}$ - this is a "timing" assumption that the inputs used in production at time t were decided upon (*i.e.* committed to) at time $t-1$.[17] This is often partially relaxed in the literature, *e.g.* in OP and LP, the labour input is not decided until t. In some case our results can apply to this more general model, but we do not elaborate here to keep things simple.[18]

To derive the first-step estimating equation, substitute the inverted (42) into (39), obtaining:

$$\begin{aligned} y_{it} &= F(x_{it}; \theta_1) + i_t^{-1}(x_{it}, i_{it}) + \epsilon_{it} \\ &= h_t(x_{it}, i_{it}) + \epsilon_{it}. \end{aligned}$$

Note that since x_{it} enters this equation both parametrically (through F), and non-parametrically (through i_t^{-1}), θ_1 and i_t^{-1} cannot be separately identified at this stage. Hence, the first step involves non-parametrically estimating the "composite" functions h_t. Common practice in the

15. We need μ to be parametric to fit the model into a two-step procedure in which the second step only requires estimating a finite dimensional set of parameters. In practice, as long as the parametric μ is specified flexibly, the difference should be minor. But strictly speaking, our results only apply when μ is assumed parametric.

16. One can also allow the production function F to depend on t - our results would generalize to this model as well.

17. This assumption helps provide identification because although x_{it} is correlated with ω_{it}, it implies that x_{it} is not correlated with the innovation in ω_{it}, *i.e.* $\omega_{it} - E[\omega_{it}|I_{it-1}]$.

18. Briefly, whether our efficiency result holds depends on whether the structural parameters related to the "variable" inputs can be identified using only the first-step moment condition. If they are, as in the first-step moment of Olley and Pakes (1996), Levinsohn and Petrin (2003), and Wooldridge (2009), our efficiency results does not hold. If they are not, as in the first-step moment of Ackerberg *et al*. (2006), then our efficiency result does hold.

applied literature is to use the moment conditions

$$E[\epsilon_{i1}|x_{i1}, i_{i1}] = E[y_{i1} - h_1(x_{i1}, i_{i1})|x_{i1}, i_{i1}] = 0 \tag{43}$$

$$\vdots$$

$$E[\epsilon_{iT}|x_{iT}, i_{iT}] = E[y_{iT} - h_T(x_{iT}, i_{iT})|x_{iT}, i_{iT}] = 0$$

and simple kernel or polynomial series regressions of y_{it} on $(x_{it},\ i_{it})$ to estimate each of the h_t's separately. Thus, this again falls into our framework of non-nested conditioning sets.

For the second step estimating equation, take the conditional expectation of (39) given I_{it-1}, substitute in (40), and then substitute in the inverted (42), *i.e.*

$$\begin{aligned} E\left[y_{it}|I_{it-1}\right] &= E\left[F(x_{it};\theta_1) + \omega_{it} + \epsilon_{it}|I_{it-1}\right] \\ &= F(x_{it};\theta_1) + E\left[\omega_{it}|I_{it-1}\right] + 0 \\ &= F(x_{it};\theta_1) + \mu(\omega_{it-1};\theta_2) \\ &= F(x_{it};\theta_1) + \mu(i_{t-1}^{-1}(x_{it-1}, i_{it-1});\theta_2) \\ &= F(x_{it};\theta_1) + \mu(h_{t-1}(x_{it-1}, i_{it-1}) - F(x_{it-1};\theta_1);\theta_2). \end{aligned}$$

The finite dimensional parameters θ_1 and θ_2 are then estimated using the moment condition:

$$E\left[\left(y_{it} - F(x_{it};\theta_1) - \mu(\widehat{h}_{t-1}(x_{it-1}, i_{it-1}) - F(x_{it-1};\theta_1);\theta_2)\right) \otimes r(I_{it-1})\right] = 0 \tag{44}$$

where the unknown functions h_t have been replaced with their estimates from the first step, $\widehat{h}_t$.

Note that since we started by assuming $E[\epsilon_{it}|I_{it}] = 0$ (and I_{it} includes past i's and x's), the first step moments (43) likely do not exhaust all the information in the model. But our results show that if, as typically done in practice, one only uses this limited set of first-step moments[19], the two-step procedure (with appropriate second-step weight matrix) does not sacrifice asymptotic efficiency relative to a joint procedure. Again, this is important because the joint procedure would requires non-linear optimization over both θ and the h_t's simultaneously, which is considerably more computationally burdensome (and prone to error) than the two-step approach, which in most cases only requires non-linear optimization over θ.

4.3. *Small Monte Carlo experiment*

We perform a brief Monte-Carlo experiment in the context of the above production function example to examine the performance of the various estimators in a small sample context. We consider the following Cobb–Douglas production function in logs

$$y_{it} = \theta_0 + \theta_1 k_{it} + \omega_{it} + \epsilon_{it}$$

where $\theta_0 = 0$ and $\theta_1 = 1$. Firms accumulate capital according to (note that uppercase variables are not logged)

$$K_{it} = \delta K_{it-1} + \kappa_{it} I_{it-1}$$

19. Presumably applied researchers do this because of the ease of running simple kernel or series regressions. It would be more complicated to enforce all the moment conditions (*i.e.* w.r.t. the full I_{it}) to estimate the h_t functions.

TABLE 1
Monte Carlo results

		Naive		Efficient two-step		Joint	
	Truth	Mean	S.D.	Mean	S.D.	Mean	S.D.
		Exact polynomial approximation (1000 reps)					
θ_0	0	−0.0010	0.0522	−0.0009	0.0484	−0.0006	0.0484
θ_1	1	1.0002	0.0202	1.0004	0.0186	1.0002	0.0187
θ_2	0.7	0.6972	0.0361	0.6974	0.0314	0.6999	0.0316
		Non-exact polynomial approximation (1000 reps)					
θ_0	0	0.0143	0.0659	−0.0216	0.0565	−0.0146	0.0563
θ_1	1	0.9930	0.0259	1.0089	0.0222	1.0058	0.0220
θ_2	0.7	0.7245	0.0433	0.7098	0.0344	0.7243	0.0351

where $\delta = 0.9$ and κ_{it} is a lognormal shock to the capital accumulation process. Firms investment decisions are assumed to follow

$$i_{it} = \gamma_0 + \gamma_1 k_{it} + \gamma_2 \omega_{it} \tag{45}$$

where $\gamma_0 = 0$, $\gamma_1 = -0.1$, and $\gamma_2 = 1$. This investment process is admittedly ad hoc but very convenient since it (i) allows us to do the Monte-Carlos without having to solve firms' dynamic programming problems, and (ii) allows us to run a specification where the non-parametric approximation is exact. We consider 1000 firms, and assume we observe two periods of full data for each firm (plus a lag - period 0). We do a 1000 period run-in prior to the observed data, so the data can be thought of as coming from the steady-state distribution given the specified investment process.

The productivity shock ω_{it} is assumed to follow a normal AR(1) process with depreciation parameter $\theta_2 = 0.7$. The variance of the innovation term in the AR(1) is set such that $\sigma_\omega = 0.1$. The measurement error in output, ϵ_{it} is normal and i.i.d. over i and t. We vary σ_ϵ across the three relevant periods in the data - $\sigma_{\epsilon_0} = 0.2$, $\sigma_{\epsilon_1} = 0.05$ and $\sigma_{\epsilon_2} = 0.1$. This is important because in our simple model, this heterogeneity in σ_ϵ generates an efficiency advantage of joint estimation relative to naive two-step estimation. Intuitively, the heterogeneity in σ_ϵ means that the different h's are estimated with different precision, which is accounted for in joint estimation (and our procedure), but not in the "naive" two-step approach. The lognormal capital accumulation shock κ_{it} is assumed to be i.i.d. over i and t and where the variance of the underlying normal is 1.[20]

Following the discussion in the prior subsection, we use i_{it} as the "proxy" variable. This leads to the first-step moment conditions

$$\begin{aligned} E[\epsilon_{i0} | k_{i0}, i_{i0}] &= E[y_{i0} - h_0(k_{i0}, i_{i0}) | k_{i0}, i_{i0}] = 0, \\ E[\epsilon_{i1} | k_{i1}, i_{i1}] &= E[y_{i1} - h_1(k_{i1}, i_{i1}) | k_{i1}, i_{i1}] = 0. \end{aligned} \tag{46}$$

We model h_0 and h_1 as second-order polynomials in the two arguments. Because (45) is linear, the non-parametric approximation is exact in this case (and while we still estimate the second-order terms, they are irrelevant). We also consider a case where we replace k_{it} and i_{it} with K_{it} and I_{it}

20. The relatively low variance of ω and ϵ and relatively high variance of κ (which generates more variation in observed k_{it}) helps lower the variance of all the estimators. Relatedly, it also makes the objective function more concave, which helps the reliability of the numeric optimization algorithm. This is particularly important to have confidence in the results of joint estimation procedure because that requires a non-linear search over 15 parameters.

(non-logged variables) in (46) and assume the h's are linear. Since ω_{it} is not linear in K_{it} and I_{it}, in this case the polynomial approximation is not exact. The second-step moment conditions are

$$E\left[\left(y_{i1}-\theta_0-\theta_1 k_{i1}-\theta_2(\widehat{h}_0(k_{i0},i_{i0})-\theta_0-\theta_1 k_{i0})\right)\otimes[1,k_{i0},k_{i1},i_{i0}]\right]=0, \tag{47}$$

$$E\left[\left(y_{i2}-\theta_0-\theta_1 k_{i2}-\theta_2(\widehat{h}_1(k_{i1},i_{i1})-\theta_0-\theta_1 k_{i1})\right)\otimes[1,k_{i1},k_{i2},i_{i1}]\right]=0.$$

Results are in Table 1. We estimate the model three ways—efficient joint estimation of both sets of moments ("Joint"), our proposed two-step efficient estimator from Section 3.3 ("Efficient Two-Step"), and a naive two-step estimator that does not consider the effect of $\widehat{h}$ in constructing the second-step weight matrix ("Naive").[21] Confirming our theoretical results, the efficient two-step estimator performs almost identically to the joint estimator, and both have a smaller small sample variance than the naive two-step estimator. This is true regardless of whether the polynomial approximation subsumes the true specification.

5. SUMMARY

This article studies the efficiency issue of a general two-step GMM estimation procedure, where the "exactly identified" unknown nuisance functions are estimated first and the finite dimensional parameters of interest are estimated by GMM with the first-step non-parametric estimators. We calculate the semiparametric efficiency bound for these models, and show that semiparametric two-step optimally weighted GMM estimators achieve the efficiency bound, where the nuisance functions could be estimated via any consistent nonparametric methods in the first step. Regardless of whether the efficiency bound has a closed form expression or not, we provide easy-to-compute sieve based optimal weight matrices that lead to asymptotically efficient two-step GMM estimators.

It is not yet clear whether or how the results would generalize to the case where the first-step non-parametric estimator takes the form of a non-parametric instrumental variables (NPIV) estimator. This is an important challenge that we leave as a future research agenda.

APPENDIX

A. PROOF OF THE MAIN RESULTS IN SECTION 2

Let $F_o(\cdot)$ be the unknown true probability distribution of Z. For $\ell=1,...,L$ with a fixed finite L, let $F_{\ell,o}(\cdot|x_\ell)$ be the unknown true conditional probability distribution of $Z_{-\ell}$ given $X_\ell=x_\ell$, where $Z_{-\ell}$ denotes the components of Z not in the conditioning variable X_ℓ, and hence $Z=(Z'_{-\ell},X'_\ell)'$. The model (1)–(2) can be rewritten

$$\int g\left(z,\theta_o,h_{1,o}(\cdot),...,h_{L,o}(\cdot)\right)dF_o(z)=0, \tag{A.1}$$

$$\int \rho_\ell\left(z_{-\ell},x_\ell,h_{\ell,o}(x_\ell)\right)dF_{\ell,o}(z_{-\ell}|x_\ell)=0 \text{ for almost all } x_\ell,\ \ell=1,...,L. \tag{A.2}$$

We note that although the unknown functions $h_{\ell,o}(\cdot)$, $\ell=1,...,L$ enter the conditional moment restrictions (2) (*i.e.* (A.2)) through $h_{\ell,o}(X_\ell)$ only, they could enter the unconditional moment restrictions (1) (*i.e.* (A.1)) in a very flexible way. We assume that the infinite dimensional nuisance functions $h_o(\cdot)=(h_{1,o}(\cdot),...,h_{L,o}(\cdot))\in\mathcal{H}=\mathcal{H}_1\times\cdots\times\mathcal{H}_L$ are identified by the conditional moment restrictions (A.2), and that if $h_o(\cdot)$ were known, the finite dimensional parameter $\theta_o\in\Theta$ is (possibly) over identified by the unconditional moment restrictions (A.1).

21. In all three cases, we need an initial consistent estimate of (θ,h) to form weight matrices. For all three cases, we use a $\overline{h}$ from the first-step polynomial OLS regression, and a $\overline{\theta}$ obtained by minimizing the second-step moments with a weight matrix given by $Var([\xi_{0i},\xi_{0i}k_{i0},\xi_{0i}k_{i1},\xi_{0i}i_{i0},\xi_{1i},\xi_{1i}k_{i1},\xi_{1i}k_{i2},\xi_{1i}i_{i1}])^{-1}$ where ξ_{0i} and ξ_{1i} are i.i.d. standard normals.

Proof of Lemma 1. For the ease of notation and without loss of generality, we assume in this proof that $L=2$. We assume that the regularity condition as in Newey (1990), Definition A.1) is satisfied.

Let $f_o(z)$ to be the true density of Z with respect to a sigma finite dominating measure $\mu(z)$, and $f_o(z_{-\ell}|x_\ell)$ be the true conditional density of $Z_{-\ell}$ given $X_\ell = x_\ell$ ($\ell = 1,2$). Here $\mathcal{F}$ denotes a class of candidate density function of Z with $f_o \in \mathcal{F}$. Define a class of density functions $\mathcal{F}_\alpha$ that satisfy the conditional and unconditional moment conditions:

$$\mathcal{F}_\alpha = \Big\{ f \in \mathcal{F} : \int \rho_1(z_{-1}, h_1(x_1)) f(z_{-1}|x_1) d\mu(z_{-1}) = 0, \int \rho_2(z_{-2}, h_2(x_2)) f(z_{-2}|x_2) d\mu(z_{-2}) = 0, \int g(z, \theta, h_1, h_2) f(z) d\mu(z) = 0 \Big\}. \tag{A.3}$$

We will consider a class of densities of Z indexed by (θ, h_1, h_2, η), where η denotes the parameter that determines the features of the distribution of Z other than the restriction above. More precisely, let $\mathcal{G}$ denote a class of real-valued measurable function of Z such that

$$\mathcal{F}_\alpha = \{ f(z|\theta, h_1, h_2, \eta) : \eta \in \mathcal{G} \} \tag{A.4}$$

for any $\alpha = (\theta, h_1, h_2) \in \Theta \times \mathcal{H}_1 \times \mathcal{H}_2$. Let $\mathcal{V}_\theta \times \mathcal{V}_1 \times \mathcal{V}_2 \times \mathcal{V}_\eta$ denote the completion (with respect to the L_2-norm) of $\Theta \times \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{G} - \{(\theta_o, h_{1,o}, h_{2,o}, \eta_o)\}$ where η_o satisfies

$$f\left(z|\theta_o, h_{1,o}, h_{2,o}, \eta_o\right) = f_o(z).$$

We will consider the parametric family $f\left(z|\theta_o + \theta, h_{1,o} + \tau_1 v_1, h_{2,o} + \tau_2 v_2, \eta_o + \tau_\eta v_\eta\right)$ in $\mathcal{F}_\alpha$. The scores in the direction of θ, τ_1, τ_2, τ_η of this family are such that

$$\begin{aligned}
s_\theta(Z) &= c_{\theta,1}(Z_{-1}|X_1) + d_{\theta,1}(X_1) \\
&= c_{\theta,2}(Z_{-2}|X_2) + d_{\theta,2}(X_2), \\
s_{h_1}(Z)[v_1] &= c_{h_1,1}(Z_{-1}|X_1)[v_1] + d_{h_1,1}(X_1)[v_1] \\
&= c_{h_1,2}(Z_{-2}|X_2)[v_1] + d_{h_1,2}(X_2)[v_1], \\
s_{h_2}(Z)[v_2] &= c_{h_2,1}(Z_{-1}|X_1)[v_2] + d_{h_2,1}(X_1)[v_2] \\
&= c_{h_2,2}(Z_{-2}|X_2)[v_2] + d_{h_2,2}(X_2)[v_2], \\
s_\eta(Z)[v_\eta] &= c_{\eta,1}(Z_{-1}|X_1)[v_\eta] + d_{\eta,1}(X_1)[v_\eta] \\
&= c_{\eta,2}(Z_{-2}|X_2)[v_\eta] + d_{\eta,2}(X_2)[v_\eta],
\end{aligned}$$

with

$$E[c_{\theta,1}(Z_{-1}, X_1)|X_1] = 0, \tag{A.5}$$
$$E[d_{\theta,1}(X_1)] = 0, \tag{A.6}$$
$$E[c_{\theta,2}(Z_{-2}, X_2)|X_2] = 0, \tag{A.7}$$
$$E[d_{\theta,2}(X_2)] = 0, \tag{A.8}$$

$$E[c_{h_1,1}(Z_{-1}|X_1)[v_1]|X_1] = 0, \tag{A.9}$$
$$E[d_{h_1,1}(X_1)[v_1]] = 0, \tag{A.10}$$
$$E[c_{h_2,1}(Z_{-1}|X_1)[v_1]|X_1] = 0, \tag{A.11}$$
$$E[d_{h_2,1}(X_1)[v_1]] = 0, \tag{A.12}$$

$$E[c_{h_2,2}(Z_{-2}|X_2)[v_2]|X_2] = 0, \tag{A.13}$$
$$E[d_{h_2,2}(X_2)[v_2]] = 0, \tag{A.14}$$
$$E[c_{h_1,2}(Z_{-2}|X_2)[v_2]|X_2] = 0, \tag{A.15}$$
$$E[d_{h_1,2}(X_2)[v_2]] = 0, \tag{A.16}$$

and

$$E\left[\, c_{\eta,1}(Z_{-1},X_1)\left[v_\eta\right]\middle|X_1\right]=0, \tag{A.17}$$

$$E\left[d_{\eta,1}(X_1)\left[v_\eta\right]\right]=0, \tag{A.18}$$

$$E\left[\, c_{\eta,2}(Z_{-2},X_2)\left[v_\eta\right]\middle|X_2\right]=0, \tag{A.19}$$

$$E\left[d_{\eta,2}(X_2)\left[v_\eta\right]\right]=0. \tag{A.20}$$

Here, $c_{h_1}(Z_{-1}|X_1)[v_1]$ and $d_{h_1}(X_1)[v_1]$ denote the conditional score of Z_{-1} given X_1 and the marginal score of X_1, obtained by differentiating the log-likelihood with respect to τ_1, for example. Note that $s_\theta(Z)$ is a $d_\theta\times 1$ vector of functions. Below, we will write $c_{h_1}(Z)[v_1]\equiv c_{h_1}(Z_{-1}|X_1)[v_1]$, *e.g.* for simplicity of notations.

Differentiating the moment restrictions in (A.3), we obtain the non-parametric tangent space $\mathcal{T}$ as the completion of the set consisting of $s_{h_1}(Z)[v_1]+s_{h_2}(Z)[v_2]+s_\eta(Z)\left[v_\eta\right]$, where s's satisfy (A.5)–(A.20) as well as

$$E\left[\,\rho_1\left(Z,h_{1,o}\right)c_{\theta,1}(Z)'\middle|X_1\right]=0, \tag{A.21}$$

$$\frac{\partial m_1(X_1,h_{1,o}(X_1))}{\partial h_1'}v_1(X_1)+E\left[\,\rho_1\left(Z,h_{1,o}\right)c_{h_1,1}(Z)[v_1]\middle|X_1\right]=0, \tag{A.22}$$

$$E\left[\,\rho_1\left(Z,h_{1,o}\right)c_{h_2,1}(Z)[v_2]\middle|X_1\right]=0, \tag{A.23}$$

$$E\left[\,\rho_1(Z,h_{1,o})c_{\eta,1}(Z)\left[v_\eta\right]\middle|X_1\right]=0, \tag{A.24}$$

$$E\left[\,\rho_2(Z,h_{2,o})c_{\theta,2}(Z)'\middle|X_2\right]=0, \tag{A.25}$$

$$E\left[\,\rho_2(Z,h_{2,o})c_{h_1,2}(Z)[v_1]\middle|X_2\right]=0, \tag{A.26}$$

$$\frac{\partial m_2(X_2,h_{2,o}(X_2))}{\partial h_2'}v_2(X_2)+E\left[\,\rho_2(Z,h_{2,o})c_{h_2,2}(Z)[v_2]\middle|X_2\right]=0, \tag{A.27}$$

$$E\left[\,\rho_2(Z,h_{2,o})c_{\eta,2}(Z)\left[v_\eta\right]\middle|X_2\right]=0, \tag{A.28}$$

and

$$\frac{\partial E\left[g\left(Z,\theta_o,h_{1,o},h_{2,o}\right)\right]}{\partial\theta'}+E\left[g\left(Z,\theta_o,h_{1,o},h_{2,o}\right)s_\theta(Z)'\right]=0, \tag{A.29}$$

$$E[g\left(Z,\theta_o,h_{1,o},h_{2,o}\right)s_{h_1}(Z)[v_1]]=0, \tag{A.30}$$

$$E\left[g\left(Z,\theta_o,h_{1,o},h_{2,o}\right)s_{h_2}(Z)[v_2]\right]=0, \tag{A.31}$$

$$E\left[g\left(Z,\theta_o,h_{1,o},h_{2,o}\right)s_\eta(Z)\left[v_\eta\right]\right]=0, \tag{A.32}$$

for any $(v_1,v_2,v_\eta)\in\mathcal{V}_1\times\mathcal{V}_2\times\mathcal{V}_\eta$, where $\frac{\partial m_\ell(X_\ell,h_{\ell,o}(X_\ell))}{\partial h_\ell'}$ and $v_\ell(X_\ell)$ are $d_\ell\times d_\ell$ matrix of functions and $d_\ell\times 1$ vector of functions respectively. Note that (14) is used in (A.30) and (A.31).

The residual of the projection of s_θ on $\mathcal{T}$, $s_\theta(Z)-proj[\,s_\theta(Z)|\mathcal{T}]$ will give the semiparametric score $S_\theta^*(Z)$ and the semiparametric information bound of θ_o will be $E[S_\theta^*(Z)S_\theta^*(Z)']$. See Bickel *et al*. (1993) and Newey (1990). We show that the residual of the projection of s_θ on $\mathcal{T}$ is equal to

$$S_\theta^*(Z)=-\left(\frac{\partial E[g(Z)]}{\partial\theta'}\right)'\left\{E\left[g(Z)g(Z)'\right]\right\}^{-1}g(Z) \tag{A.33}$$

where $g(Z)=g\left(Z,\theta_o,h_{1,o},h_{2,o}\right)$.

We now define $\Lambda_1^*(X_1)$ and $\Lambda_2^*(X_2)$ as solutions to

$$0=E\left[\,\rho_1(Z,h_{1,o})\left\{c_{\theta,1}(Z)'-S_\theta^*(Z)'-c_{h_1,1}(Z)\left[\Lambda_1^*\right]-c_{h_2,1}(Z)\left[\Lambda_2^*\right]\right\}\middle|X_1\right] \tag{A.34}$$

and

$$0=E\left[\,\rho_2(Z,h_{2,o})\left\{c_{\theta,2}(Z)'-S_\theta^*(Z)'-c_{h_1,2}(Z)\left[\Lambda_1^*\right]-c_{h_2,2}(Z)\left[\Lambda_2^*\right]\right\}\middle|X_2\right]. \tag{A.35}$$

Note that for $\ell=1,2$, $\Lambda_\ell^*(X_\ell)$ is a $d_\ell\times d_\theta$ matrix of functions and

$$c_{h_\ell,\ell}(Z)\left[\Lambda_\ell^*\right]=\left[c_{h_\ell,\ell}(Z)\left[\Lambda_{\ell,1}^*\right],\ldots,c_{h_\ell,\ell}(Z)\left[\Lambda_{\ell,d_\theta}^*\right]\right]$$

is a $1\times d_\theta$ vector of functions, where $\Lambda_{\ell,j}^*(X_\ell)$ denotes the j-th row of $\Lambda_\ell^*(X_\ell)$ for $j=1,\ldots,d_\theta$.

We argue that such $\Lambda_1^*(X_1)$ and $\Lambda_2^*(X_2)$ exist as unique objects almost surely for the following reason. Letting $v_1 = \Lambda_{1,j}^*(X_1)$ in (A.22) and $v_2 = \Lambda_{2,j}^*(X_2)$ in (A.23) for $j = 1, \ldots, d_\theta$, we get

$$\frac{\partial m_1(X_1, h_{1,o}(X_1))}{\partial h_1'} \Lambda_1^*(X_1) + E\left[\rho_1\left(Z, h_{1,o}\right) c_{h_1,1}(Z)\left[\Lambda_1^*\right] \middle| X_1\right] = 0 \tag{A.36}$$

and

$$E\left[\rho_1\left(Z, h_{1,o}\right) c_{h_2,1}(Z)\left[\Lambda_2^*\right] \middle| X_1\right] = 0. \tag{A.37}$$

Using (A.21), (A.33), (A.36), and (A.37), we note that

$$\begin{aligned}
&E\left[\rho_1(Z, h_{1,o}) c_{\theta,1}(Z)' \middle| X_1\right] - E\left[\rho_1(Z, h_{1,o}) S_\theta^*(Z)' \middle| X_1\right] \\
&- E\left[\rho_1(Z, h_{1,o}) c_{h_1,1}(Z)\left[\Lambda_1^*\right] \middle| X_1\right] - E\left[\rho_1(Z, h_{1,o}) c_{h_2,1}(Z)\left[\Lambda_2^*\right] \middle| X_1\right] \\
=& 0 + \left(\frac{\partial E[g(Z)]}{\partial \theta'}\right)' \left\{E\left[g(Z) g(Z)'\right]\right\}^{-1} E\left[g(Z) \rho_1(Z, h_{1,o})' \middle| X_1\right] \\
&+ \left(\frac{\partial m_1(X_1, h_{1,o}(X_1))}{\partial h_1'} \Lambda_1^*(X_1)\right)' + 0,
\end{aligned}$$

so we rewrite (A.34) as

$$0 = E\left[\rho_1\left(Z, h_{1,o}\right) g(Z)' \middle| X_1\right] \left\{E\left[g(Z) g(Z)'\right]\right\}^{-1} \frac{\partial E[g(Z)]}{\partial \theta'} + \frac{\partial m_1(X_1, h_{1,o}(X_1))}{\partial h_1'} \Lambda_1^*(X_1),$$

which can be solved for $\Lambda_1^*(X_1)$ as long as $\partial m_1\left(X_1, h_{1,o}(X_1)\right) / \partial h_1'$ is invertible almost surely. Similarly, we can solve for $\Lambda_2^*(X_2)$ as long as $\partial m_2\left(X_2, h_{2,o}(X_2)\right) / \partial h_2'$ is invertible almost surely.

Now let

$$\delta' = s_\theta(Z)' - S_\theta^*(Z)' - s_{h_1}(Z)\left[\Lambda_1^*\right] - s_{h_2}(Z)\left[\Lambda_2^*\right].$$

We will show that δ satisfies the properties (A.17)–(A.20), (A.24), (A.28), and (A.32) of the $s_\eta(Z)\left[v_\eta\right]$ for some v_η.

- By construction, we have $E[\delta] = 0$. Taking

$$\begin{aligned}
\widetilde{d}_{\eta,1}(X_1)\left[v_\eta\right] &= E[\,\delta | X_1] \\
&= d_{\theta,1}(X_1) - \left(d_{h_1,1}(X_1)\left[\Lambda_1^*\right]\right)' - \left(d_{h_2,1}(X_1)\left[\Lambda_2^*\right]\right)' \\
&+ E\left[c_{\theta,1}(Z) - S_\theta^*(Z) - \left(c_{h_1,1}(Z)\left[\Lambda_1^*\right]\right)' - \left(c_{h_2,1}(Z)\left[\Lambda_2^*\right]\right)' \middle| X_1\right] \\
&= d_{\theta,1}(X_1) - \left(d_{h_1,1}(X_1)\left[\Lambda_1^*\right]\right)' - \left(d_{h_2,1}(X_1)\left[\Lambda_2^*\right]\right)' - E\left[S_\theta^*(Z) | X_1\right],
\end{aligned}$$

  and

$$\begin{aligned}
\widetilde{c}_{\eta,1}(Z)\left[v_\eta\right] &= \delta - \widetilde{d}_{\eta,1}(X_1)\left[v_\eta\right] \\
&= c_{\theta,1}(X_1) - \left(c_{h_1,1}(Z)\left[\Lambda_1^*\right]\right)' - \left(c_{h_2,1}(Z)\left[\Lambda_2^*\right]\right)' - S_\theta^*(Z) + E\left[S_\theta^*(Z) | X_1\right],
\end{aligned}$$

  we can see that properties (A.17) and (A.18) are satisfied for

$$\delta = \widetilde{c}_{\eta,1}(Z)\left[v_\eta\right] + \widetilde{d}_{\eta,1}(X_1)\left[v_\eta\right].$$

  With $\widetilde{c}_{\eta,2}(Z)\left[v_\eta\right]$ and $\widetilde{d}_{\eta,2}(X_2)\left[v_\eta\right]$ similarly defined, we can see that properties (A.19) and (A.20) are also satisfied.
- Equations (A.34) implies that

$$\begin{aligned}
&E\left[\rho_1(Z, h_{1,o}) \widetilde{c}_{\eta,1}(z)\left[v_\eta\right]' \middle| X_1\right] \\
&= E\left[\rho_1(Z, h_{1,o}) \left\{c_{\theta,1}(Z)' - S_\theta^*(Z)' - c_{h_1,1}(Z)\left[\Lambda_1^*\right] - c_{h_2,1}(Z)\left[\Lambda_2^*\right]\right\} \middle| X_1\right] \\
&+ E\left[\rho_1(Z, h_{1,o}) \middle| X_1\right] E\left[S_\theta^*(Z)' \middle| X_1\right] \\
&= 0.
\end{aligned}$$

  which implies that the property (A.24) is satisfied by δ. Likewise, (A.28) are satisfied by δ.

- Using (A.29)–(A.31), we obtain

$$\begin{aligned} E\left[\delta g(Z)'\right] &= E\left[s_\theta(Z)g(Z)'\right] - E\left[S_\theta^*(Z)g(Z)'\right] \\ &= -\left(\frac{\partial E[g(Z)]}{\partial \theta'}\right)' + \left(\frac{\partial E[g(Z)]}{\partial \theta'}\right)' \left\{E[g(Z)g(Z)']\right\}^{-1} \left\{E[g(Z)g(Z)']\right\} \\ &= 0, \end{aligned} \tag{A.38}$$

  which shows that the property (A.32) is satisfied.

These observations lead us to conclude that

$$s_{h_1}(Z)\left[\Lambda_1^*\right] + s_{h_2}(Z)\left[\Lambda_2^*\right] + \delta \in \mathcal{T}. \tag{A.39}$$

Because $S_\theta^*(Z)$ is proportional to $g(Z)$, we can deduce from (A.30) to (A.32) that $S_\theta^*(Z) \perp \mathcal{T}$. Along with (A.39), this implies that $S_\theta^*(Z)$ is the residual of the projection of s_θ on $\mathcal{T}$. Thus the semiparametric information bound of θ_o is

$$E[S_\theta^*(Z)S_\theta^*(Z)'] = \left(\frac{\partial E[g(Z)]}{\partial \theta'}\right)' \left\{E[g(Z)g(Z)']\right\}^{-1} \left(\frac{\partial E[g(Z)]}{\partial \theta'}\right). \tag{A.40}$$

‖

Acknowledgments. We acknowledge helpful comments from R. Blundell, T. Christensen, A. Gandhi, W. Newey, A. Pakes, D. Pouzo, J. Powell, E. Renault and participants in May 2012 Cemmap Masterclass on Semi-nonparametric Models and Methods in London, in August 2012 NBER-NSF CEME Conference on The Econometrics of Dynamic Games at NYU, in November 2012 Info-metrics & Nonparametric Inference at UC Riverside and econometrics workshops at Boston University, UC-Berkeley, UC-Davis, USC and University of Washington. Any errors are the responsibility of the authors.

Supplementary Data

Replication files for the simulations of Section 4.3 are included in Supplementary data, which are available at *Review of Economic Studies* online.

REFERENCES

ACKERBERG, D., CAVES, K. and FRAZER, G. (2006), "Structural Identification of Production Functions" (mimeo, UCLA).

ACKERBERG, D., CHEN, X. and HAHN, J. (2012), "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators", *Review of Economics and Statistics*, **94**, 481–498.

AGUIRREGABIRIA, V. and MIRA, P. (2002), "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models", *Econometrica*, **70**, 1519–1543.

AGUIRREGABIRIA, V., and MIRA, P. (2007), "Sequential Estimation of Dynamic Discrete Games", *Econometrica*, **75**, 1–53.

AI, C. and CHEN, X. (2003), "Efficient Estimation of Conditional Moment Restrictions Models Containing Unknown Functions", *Econometrica*, **71**, 1795–1843.

AI, C. and CHEN, X. (2007), "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables", *Journal of Econometrics*, **141**, 5–43.

AI, C. and CHEN, X. (2012), "The Semiparametric Efficiency Bound for Models of Sequential Moment Restrictions Containing Unknown Functions", *Journal of Econometrics*, **170**, 442–457.

ANDREWS, D. (1994), "Asymptotics for Semi-parametric Econometric Models via Stochastic Equicontinuity", *Econometrica*, **62**, 43–72.

ARCIDIACONO, P. and MILLER, R. (2011), "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity", *Econometrica*, **79**, 1823–1868.

ARMSTRONG, T. B., BERTANHA, M. and HONG, H. (2012), "A Fast Resample Method for Parametric and Semiparametric Models" (Unpublished working paper).

BAJARI, P., BENKARD, L. and LEVIN, J. (2007), "Estimating Dynamic Models of Imperfect Competition", *Econometrica*, **75**, 1331–1370.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993), *Efficient and Adaptive Inference in Semiparametric Models* (Baltimore: Johns Hopkins University Press).

CHAMBERLAIN, G. (1992), "Comment: Sequential Moment Restrictions in Panel Data", *Journal of Business and Economic Statistics*, **10**, 20–26.

CHEN, X., LINTON, O. and VAN KEILEGOM, I. (2003), "Estimation of Semiparametric Models when the Criterion Function is not Smooth", *Econometrica*, **71**, 1591–1608.

COLLARD-WEXLER, A. (2012), "Demand Fluctuations in the Ready-Mix Concrete Industry" *Econometrica* (forthcoming).
CREPON, B., KRAMARZ, F. and TROGNON, A. (1997), "Parameters of Interest, Nuisance Parameters, and Orthogonality Conditions: An Application to Autoregressive Error Component Models", *Journal of Econometrics*, **82**, 135–156.
FANG, H. and WANG, Y. (2012), "Estimating Dynamic Discrete Choice Models with Hyperbolic Discounting with an Application to Mammorgraphy Decisions" (Working paper, University of Pennsylvania).
GHOSH, M., PARR, W.C. and SINGH, K. (1984), "A Note on Bootstrapping the Sample Median", *Annals of Statistics*, **12**, 1130–1135.
GONÇALVES, S. and WHITE, H. (2005), "Bootstrap Standard Error Estimates for Linear Regression", *Journal of American Statistical Association*, **100**, 970–979.
HAYASHI, F. and SIMS, C. (1983), "Nearly Efficient Estimation of Time Series Models with Predetermined, but not Exogenous Instruments", *Econometrica*, **51**, 783–798.
HONG, H., MAHAJAN, A. and NEKIPELOV, D. (2010), "Extremum Estimation and Numerical Derivatives" (mimeo, Stanford University).
HOTZ, J. and MILLER, R. (1993), "Conditional Choice Probabilities and the Estimation of Dynamic Models", *Review of Economic Studies*, **60**, 497–529.
HOTZ, J., MILLER, R., SANDERS, S. and SMITH, J. (1994), "A Simulation Estimator for Dynamic Models of Discrete Choice", *Review of Economic Studies*, **61**, 265–289.
Hu, Y., Shum, M., and Tan, W. (2010), "A Simple Estimator for Dynamic Models with Serially Correlated Unobservables" (Unpublished working paper).
HANSEN, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, **50**, 1029–1054.
KASAHARA, H. and SHIMOTSU, K. (2009), "Nonparametric Identification and Estimation of Finite Mixture Models of Dynamic Discrete Choices", *Econometrica*, **77**, 135–175.
LEVINSOHN, J. and PETRIN, A. (2003), "Estimating Production Functions using Inputs to Control for Unobservables", *Review of Economic Studies*, **70**, 317–341.
MACHADO, J. A. F., and PARENTE, P. (2005), "Bootstrap Estimation of Covariance Matrices via the Percentile Method", *Econometrics Journal*, **8**, 70–78.
MURPHY, K. M. and TOPEL, R. H. (1985), "Estimation and Inference in Two-Step Econometric Models", *Journal of Business and Economic Statistics*, **3**, 370–379.
NEWEY, W. K. (1984), "A Method of Moments Interpretation of Sequential Estimators", *Economics Letters*, **14**, 201–206.
NEWEY, W. K. (1990), "Semiparametric Efficiency Bounds", *Journal of Applied Econometrics*, **5**, 99–135.
NEWEY, W. K. (1994), "The Asymptotic Variance of Semiparametric Estimators", *Econometrica*, **62**, 1349–1382.
NEWEY, W. K. and POWELL, J. L. (1999), "Two-step Estimation, Optimal Moment Conditions, and Sample Selection Models" (Working paper, MIT).
OLLEY, G. and PAKES, A. (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry", *Econometrica*, **64**, 1263–1297.
PAKES, A. and OLLEY, G. (1995), "A Limit Theorem for a Smooth Class of Semiparametric Estimators", *Journal of Econometrics*, **65**, 295–332.
PAKES, A., OSTROVSKY, M. and BERRY, S. (2007), "Simple Estimators for the Parameters of Discrete Dynamic Games, with Entry/Exit Examples", *RAND Journal of Economics*, **38**, 373–399.
PESENDORFER, M. and SCHMIDT-DENGLER, P. (2008), "Asymptotic Least Squares Estimators for Dynamic Games", *Review of Economic Studies*, **75**, 901–928.
RYAN, S. (2012), "The Costs of Environmental Regulation in a Concentrated Industry", *Econometrica*, **80**, 1019–1061.
SHAO, J. (1992), "Bootstrap Variance Estimators with Truncation", *Statistics & Probability Letters*, **15**, 95–101.
WOOLDRIDGE, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data* (Cambridge: MIT Press).
WOOLDRIDGE, J. M. (2009), "On Estimating Firm-Level Production Functions Using Proxy Variables to Control for.Unobservables", *Economics Letters*, **104**, 112–114.
WU, C. F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis", *Annals of Statistics*, **14**, 1261–1295.