Discrete Dynamic Models with Continuous State Variables

Patrick Bajari and Victor Chernozhukov and Han Hong

1 Introduction

We develop semiparametric estimators for dynamic discrete choice models and dynamic discrete games allowing for the presence of both continuous and discrete state variables. This paper builds on the two step estimators of Berry, Pakes, and Ostrovsky (2003), Aguirregabiria and Mira (2002), Bajari, Benkard, and Levin (2004), Pesendorfer and Schmidt-Dengler (2003) and Hotz and Miller (1993). Our contribution lies in the extension to allow for both discrete and continuous state variables. In the presence of continuous state variables, flexible nonparametric estimators are used to recover the conditional choice probabilities given the state variables. The estimator is semiparametric in the sense that the parametric components in the period utility functions can still be consistently estimated at the $\sqrt{T}$ and has an asymptotic normal distribution, despite the fact that the first stage nonparametric estimation can have a much slower than $\sqrt{T}$ rate of convergence.

This contribution is a technical one but nevertheles an important one. It is a conventional exercise in the literature of the estimation of dynamic model to discretize the state space. There are two implications to the discretization procedure. The first is a computational one and the second is a statistical one. First of all, the value functions are typically computed by discretization of the state space. The issue here is a computational one. The precision of the discretization procedure is only constrained by the computing power and not by the available of data. With sufficient computing power, it is always possible to use a sufficiently fine grid to discretize the value function so that it does not affect the statistical uncertainty of the parameter estimates.

The second issue, however, is a statistical one. Discretization is also frequently used in many recent two step procedures to estimate the conditional choice probabilities given the state variables in the first step. Increasing the number of grids in estimating the choice probabilities has two offseting effects on the estimation errors for the parameters. It reduces the bias in the first stage estimation but on the other hand increases the variance. It becomes necessary to balance the delicate tradeoff between bias and variance. In fact, when the dimension of continuous state variables is larger than four, it is impossible to obtain $\sqrt{T}$ consistent and asymptotic normal estimators for the period utility function parameters through discretization. Intuitively, the variance of the discretization procedure is of the magnitude of $1/\sqrt{T`h^d}$ where T is the sample size, d is the dimension of continuous state variables, and h is the window size used in the discretization. The bias, on the hand, is of the magnitude of $\sqrt{T}h^2$. $\sqrt{T}$ consistency of the parameter estimator requires than both the variance and the bias decrease to zero as the sample size n increases to ∞. This is impossible, however, when d is larger than 4.

The use of higher order kernel function and flexible parametric series expansions in obtaining the conditional choice probabilities allow us to reduce the bias in the first stage estimation, and permit $\sqrt{T}$ consistent estimation of the period utility function parameters in the second stage. For this purpose it is important to understand the assumptions needed on the underlying model and on the use of series approximations in order to achieve asymptotic normality in the second stage estimation procedure.

The assumptions of continuous or discrete state variables are arguable both approximations to the empirical data. For example, wage data can be considered either discrete or continuous. Which approximation offers a better one depends on the number of observations and the number of possible value the state variable can take. This is ultimatley an empirical question.

The paper is organized as follows. Section 2 describes the model and the estimation procedure. Section 3 provides conditions for the existence of markov perfect equilibrium in the presence of continous state variables. Section 4 develops the statistical properties of the semiparametric two step estimator. Section 5 contains an empirical application. Section 6 concludes.

2 The model and the estimator

To simplify notation we confine ourselves to stationary models. The primitive components of a dynamic discrete choice model or a dynamic discrete discrete game consists of the deterministic mean period utility function and random utility shock, the transition process of observed states. The mean per period utility function for player i when he takes action i and when other players take action a_{-i}, given that the current state is s, is given by a function known up to a parameter vector θ. The vector of random utility shocks for each player is assumed to be i.i.d. across players and time, and are independent of the sequence of state variables.

$$u_i\left(a, \theta, \epsilon_i, s\right) = \Pi_i\left(a_i, a_{-i}, \theta, s\right) + \epsilon_i\left(a_i\right).$$

Both the set of state spaces and action spaces are homogenuous across agents $i = 1, \ldots, n$. Therefore for all i, $a_i \in \{0, 1, \ldots, K\}$ and $s \in S$.

The goal of the semiparametric estimation procedure is to recover the parameters θ in the per period utility function $\Pi\left(a_i, a_{-i}, \theta\right)$, taking as given the estimated values of the conditional choice probabilities $p\left(a|s\right)$ and the transition process $p\left(s_{t+1}|a, s\right)$. To illustrate, define the following notations:

$$\pi\left(a_i, s|\theta, p\left(a_{-i}|s\right)\right) = \sum_{a_{-i}} \Pi\left(a_i, a_{-i}, s, \theta\right) p\left(a_{-i}|s\right)$$

as the expected period utility for player i given the conditional choice probabilities of other players. Next we define the ex ante value function and the choice specific value functions through a pair of recursive relations. Let the ex ante value function for player i when the state is s be denoted by $V_i\left(s\right)$, and let the choice specific value functions for player i when the state is s and when the action choice is a_i be denoted by $V_i\left(s, a_i\right)$. Then

$$V_i\left(a_i, s\right) = \pi\left(a_i, s\right) + \beta \int V_i\left(s'\right) dp\left(s'|s, a_i\right), \tag{1}$$

where $dp\left(s'|s, a_i\right) = dp\left(s'|s, a_{-i}, a_i\right) dp\left(a_{-i}|s\right)$, and

$$V_i\left(s\right) = \sum_{a_i} V_i\left(a_i, s\right) p\left(a_i|s\right) + \sum_{a_i} E\left(\epsilon_i\left(a_i\right)|a_i \text{ is chosen}\right) p\left(a_i|s\right). \tag{2}$$

The last summation, which is the expected unobserved utility component given that it is chosen optimally, can also be written as

$$\sum_{a_i=0}^{K} \int \epsilon_i\left(a_i\right) \Pi_{k \neq a_i} 1\left(\epsilon_i\left(a_i\right) - \epsilon_i\left(k\right) > V_i\left(k, s\right) - V_i\left(a_i, s\right)\right) dF\left(\epsilon_i\right),$$

which clearly depends only on the difference of choice specific utilities

$$V_i\left(k,s\right)-V_i\left(0,s\right),\forall k=1,\dots,K.$$

The relations (1) and (2) also define a contraction mapping for the ex ante value function

$$V_i\left(s\right)=\int\Pi_i\left(a,s,\theta\right)dp\left(a|s\right)+\beta\int V_i\left(s'\right)dp\left(s'|s\right)+\sum_{a_i}E\left(\epsilon_i\left(a_i\right)|a_i\text{ is chosen}\right)p\left(a_i|s\right).$$

Using these notations, the dynamic discrete model is observationally equivalent to a static discrete choice model where the per period utility $\pi_i\left(a_i,s\right)+\epsilon_i\left(a_i\right)$ is replaced by that with the choice specific value function:

$$V_i\left(a_i,s\right)+\epsilon_i\left(a_i\right).$$

It is well known from Hotz and Miller (1993) (recently cited in Pesendorfer and Schmidt-Dengler (2003)) that there exists a one-to-one mapping between the observed conditional choice probabilities and the differences in the choice specific value functions. In other words, it is possible to invert the system that for $a=1,\dots,K$,

$$p\left(a|s\right)=P\left(\epsilon_i\left(a\right)+V_i\left(a,s\right)-V_i\left(0,s\right)>\epsilon_i\left(k\right)+V_i\left(k,s\right)-V_i\left(0,s\right),\forall k=0,\dots,K,k\neq a\right)$$

to obtain the differences in the choice specific value functions as a function of the observed conditional choice probabilities given the state variables. Denote this mapping by

$$\left(p\left(a|s\right),a=1,\dots,K\right)=F\left(V_i\left(k,s\right)-V_i\left(0,s\right),k=1,\dots,K\right).$$

and the inverse mapping by

$$\left(V_i\left(k,s\right)-V_i\left(0,s\right),k=1,\dots,K\right)=\Phi\left(p\left(k|s\right),k=1,\dots,K\right). \tag{3}$$

In the following, let $p\left(k|s\right),k=1,\dots,K$ denote a nonparametric estimator for the conditional choice probabilities. This typically involves frequency summation for the discrete components of s and kernel or series approximations for the continous components of the state s. Consider approximation using sieve series expansions (Ai and Chen (2003)). To be more precise, suppose we have access to a data set with observations $t=1,\dots,T$. Let

$s = \left(s^d, s^c\right)$ denote the discrete and continuous components of s. Let $\{q_l\left(s^c\right), l = 1, 2, \ldots\}$ denote a sequence of known basis functions that can approximate any square-measurable function of s^c arbitrarily well. Also let

$$q^{\kappa(T)}\left(s^c\right) = \left(q_1\left(s^c\right), \ldots, q_{\kappa(T)}\left(s^c\right)\right),$$

be the approximating functions at value x^c and let

$$Q_T\left(s^d\right) = \left(q^{\kappa(T)}\left(s_1^c\right) 1\left(s_1^d = s^d\right), \ldots, q^{\kappa(T)}\left(s_T^c\right) 1\left(s_T^d = s^d\right)\right),$$

be the data matrix of approximating functions when for value of discrete component s^d, where for some integer $\kappa(T) \to \infty$, $\kappa(T)/T \to 0$ at appropriate rate to be specified below when $T \to \infty$. A linear probability model based nonparametric first step estimator for $P_i\left(k|s\right), k = 1, \ldots, K$ can then be given by

$$\hat{p}\left(k|s\right) = \sum_{t=1}^{T} 1\left(a_{it} = k, s_t^d = s^d\right) q^{\kappa(T)}\left(s_t^c\right) \left(Q_T\left(s^d\right)' Q_T\left(s^d\right)\right)^{-1} q^{\kappa(T)}\left(s^c\right).$$

In the presence of continuous components of the state variables s^c, $\hat{p}\left(k|s\right)$ typically converges at a nonparametric rate to the true $p\left(k|s\right)$, which is lower than $\sqrt{T}$. However, section 4 will give precise conditions on $\kappa(T)$ such that the resulting semiparametric estimator is $\sqrt{T}$ consistent and asymptotically normally distributed. Other link function can also be used. For example, a sieve logit model can be used in the first step to replace the sieve linear probability model.

Given the estimates of $\hat{p}\left(k|s\right)$, the differences in the choice specific value functions can then be estimated by inverting the one-to-one mapping between these functions and conditional choice probabilities, for $i = 1, \ldots, n$:

$$\left(\hat{V}_i\left(k, s\right) - \hat{V}_i\left(0, s\right), k = 1, \ldots, K\right) = \Phi\left(\hat{p}_i\left(k|s\right), k = 1, \ldots, K\right).$$

Either $\hat{p}\left(k|s\right)$ or $\left(\hat{V}_i\left(k, s\right) - \hat{V}_i\left(0, s\right), k = 1, \ldots, K\right)$ contain the desired information in the data regarding the utility paramaters. In the second step, note that with knowledge of the parameters θ and the estimated choice probabilities $\hat{p}\left(k|s\right)$ and the transition probability process $\hat{p}\left(s'|s, k\right)$, the ex ante value function and the choice specific value functions can be recovered from $\Pi(a_i, a_{-i}, \theta)$ through the recursive relations (1) and (2) that define them. We will denote this abstract mapping by

$$\left(V_i(k,s), \forall k\right) = \Psi_i\left(k, s|\theta, \hat{p}(k|s), \forall k, \hat{p}(s'|s)\right).$$

If this mapping is either known or can be analytically computed, a semiparametric least square estimator for θ can be defined by

$$\min_{\theta} \sum_{t=1}^{T} \left|\left| \Psi_i\left(k-0, s|\theta, \hat{p}(k|s), \forall k, \hat{p}(s'|s)\right) - \left(\hat{V}_i(k,s) - \hat{V}_i(0,s)\right), \forall i, \forall k \right|\right|_t, \tag{4}$$

where the norm $||\cdot||$ can be data dependent on the heterogeneity of the market and applies to the vector of the differences across players and actions. In the above we have used the simplied notation

$$\begin{aligned}&\Psi_i\left(k-0, s|\theta, \hat{p}(k|s), \forall k, \hat{p}(s'|s)\right)\\ =&\Psi_i\left(k, s|\theta, \hat{p}(k|s), \forall k, \hat{p}(s'|s)\right) - \Psi_i\left(0, s|\theta, \hat{p}(k|s), \forall k, \hat{p}(s'|s)\right).\end{aligned}$$

For example, Bajari, Benkard, and Levin (2004) considered the important case where the per period payoff function is linear in the parameters:

$$\Pi_i(a_i, a_{-i}, s, \theta) = \Phi_i(a_i, a_{-i}, s)' \theta.$$

In this case the contraction mapping for the ex ante value function $V_i(s)$ takes the form of

$$V_i(s) = \Lambda_i(s) + \Phi_i(s)' \theta + \beta \int V_i(s')\, dp(s'|s)$$

where

$$\Lambda_i(s) = \sum_{a_i} E\left(\epsilon_i(a_i) | a_i \text{ is chosen}\right) p(a_i|s) \tag{5}$$

and

$$\Phi_i(s) = \int \Phi_i(a, s)\, dp(a|s). \tag{6}$$

Since the inverse of a linear operator has to be linear, for a given θ, the value function $V_i(s)$ can then be solved as

$$V_i(s) = \bar{\Lambda}_i(s) + \bar{\Phi}_i(s)' \theta,$$

where $\bar{\Lambda}(s)$ and $\bar{\Phi}_i(s)$ solve

$$\bar{\Lambda}_i(s) = \Lambda_i(s) + \beta \int \bar{\Lambda}_i(s')\, dp(s'|s), \tag{7}$$

and

$$\bar{\Phi}_i(s) = \Phi_i(s) + \beta \int \bar{\Phi}_i(s')\, dp(s'|s). \tag{8}$$

The choice specific value functions are then also linear functions of θ. In particular,

$$V_i(a_i, s) = \Phi_i(a_i, s)' \theta + \beta \tilde{\Lambda}_i(s; a_i) + \beta \tilde{\Phi}_i(s; a_i)' \theta,$$

where

$$\Phi_i(a_i, s) = E_{a_{-i}|s} \Phi_i(a_i, a_{-i}, s), \tag{9}$$

$$\tilde{\Lambda}_i(s; a_i) = \int \bar{\Lambda}_i(s')\, dp(s'|s, a_i) \tag{10}$$

and

$$\tilde{\Phi}_i(s, a_i) = \int \bar{\Phi}_i(s')\, dp(s'|s, a_i) = E\left[\bar{\Phi}_i(s')\,|s, a_i\right]. \tag{11}$$

Therefore we can write $\Psi_i(k - 0, s|\theta, \hat{p}(k|s), \forall k, \hat{p}(s'|s))$ as

$$\beta\left(\tilde{\Lambda}_i(s; k) - \tilde{\Lambda}_i(s, 0)\right) + \left(\Phi_i(k, s) - \Phi_i(0, s) + \beta\left(\tilde{\Phi}_i(s; k) - \tilde{\Phi}_i(s; 0)\right)\right)' \theta.$$

In this case the estimator defined in equation (4) amounts to nothing other than running a linear ordinary least square regression where the dependent variables are the collection of $\left(\hat{V}_i(k, s) - \hat{V}_i(0, s)\right) - \beta\left(\Lambda_i(s; k) - \Lambda_i(s, 0)\right)$ and the regressors are the collection of

$$\left(\Phi_i(k, s) - \Phi_i(0, s) + \beta\left(\tilde{\Phi}_i(s; k) - \tilde{\Phi}_i(s; 0)\right)\right).$$

This is computationally trivial and does not require any numerical optimization over the parameter space θ. Even if β is unknown, it can also be estimated together with θ using an ordinary least square regression, by regressing $\left(\hat{V}_i(k, s) - \hat{V}_i(0, s)\right)$ on $\left(\Lambda_i(s; k) - \Lambda_i(s, 0)\right)$, $\Phi_i(k, s) - \Phi_i(0, s)$ and $\left(\tilde{\Phi}_i(s; k) - \tilde{\Phi}_i(s; 0)\right)$. This also offers testable implications because the third set of coefficients should be the product of the first two. All the value function iterations only need to be done once before the second step linear regression to recover θ. The estimator is point consistent for the true parameters regardless of whether the regressors are discrete or continuous. In the case of a single player static model with discrete regressors, this reduces to the minimum chi square estimator discussed in Amemiya (1985).

3 Practical Inference

The theory of sieve approximation provides rigorous conditions for controlling the bias term of the approximation, so that which parametric models can be viewed as good approximations of the true transition probability function and choice probability functions. Given this theory, calculating the statistical properties of the first stage can be performed as if the first stage is estimated parametrically. In this section we describe the practical computation of standard errors for the two step estimator for the dynamic discrete choice model. It requires computing a fix pointed iteration only twice and involves no numerical optimization. The first time in calculating the estimator and the second time in calculating the standard errors. It is interesting to note that all the statistical noise in the estimation procedure derives from the first step parametric or nonparametric estimation of the transition probability functions and the choice probability functions. The second stage linear regression does not introduce any additional statistical uncertainties. Intuitively, this is because of the lack of an additional error term in the second stage of the estimation procedure. If the transition probabilities and the choice probabilities are exactly known, then the second stage estimation procedure is an identity that holds for all values of the state variables and all choice actions, at the true utility parameter.

The second stage estimator minimizes some suitably chosen norm of the entire infinite dimension vector of the difference across all players, all values of the state variables and all choice actions of the form:

$$\left|\Psi_i\left(k-0, s|\theta, \hat{p}\left(k|s\right), \forall k, \hat{p}\left(s'|s\right)\right) - \left(\hat{V}_i\left(k,s\right) - \hat{V}_i\left(0,s\right)\right)\right|_{s,i,k}.$$

For example, a Cramer-Von Mises integrated quadratic norm can be used:

$$\int \left[\Psi_i\left(k-0,s|\theta,\hat{p}\left(k|s\right),\forall k,\hat{p}\left(s'|s\right)\right)-\left(\hat{V}_i\left(k,s\right)-\hat{V}_i\left(0,s\right)\right)\right]^2 d\hat{W}\left(s,i,k\right),$$

where the weight functions $\hat{W}\left(s,i,k\right)$ can be random and can depend on data. For example, if the empirical distribution of s, i, and k are used for the weight functions, under the linearity assumption, the norm that is to be minimized takes the form of

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=1}^{K}\Bigg[\left(\hat{V}_i\left(k,s_t\right)-\hat{V}_i\left(0,s_t\right)\right)-\beta\left(\tilde{\Lambda}_i\left(s_t;k\right)-\tilde{\Lambda}_i\left(s_t;0\right)\right)$$
$$-\left(\Phi_i\left(k,s_t\right)-\Phi_i\left(0,s_t\right)+\beta\left(\tilde{\Phi}_i\left(s_t;k\right)-\tilde{\Phi}_i\left(s_t;0\right)\right)\right)'\theta\Bigg]^2.$$

We focus on this norm in this section to illustrate how practical inference can be performed. The dependent variables and the regressors in the above regression depend on preliminary estimates of the first stage chocie probabilities and transition probabilities. Denote the parameters estimated in the first stage by $\hat{\alpha}$. The parametric estimates $\hat{\theta}$ have the explicit solution of, where we use $\sum$ to denote $\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=1}^{K}$,

$$\hat{\theta}=\left(\sum X\left(s_t,i,k,\hat{\alpha}\right)X\left(s_t,i,k,\hat{\alpha}\right)'\right)^{-1}\left(\sum X\left(s_t,i,k,\hat{\alpha}\right)Y\left(s_t,i,k,\hat{\alpha}\right)'\right),$$

where the dependence on the first stage is made explicit by $\hat{\alpha}$, and

$$X\left(s_t,i,k\right)=\left(\Phi_i\left(k,s_t\right)-\Phi_i\left(0,s_t\right)+\beta\left(\tilde{\Phi}_i\left(s_t;k\right)-\tilde{\Phi}_i\left(s_t;0\right)\right)\right)$$
$$Y\left(s_t,i,k\right)=\left(\hat{V}_i\left(k,s_t\right)-\hat{V}_i\left(0,s_t\right)\right)-\beta\left(\tilde{\Lambda}_i\left(s_t;k\right)-\tilde{\Lambda}_i\left(s_t;0\right)\right).$$

This can also be written as

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right)=\left(\frac{1}{T}\sum X\left(s_t,i,k,\hat{\alpha}\right)X\left(s_t,i,k,\hat{\alpha}\right)'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum X\left(s_t,i,k,\hat{\alpha}\right)\epsilon\left(s_t,i,k,\hat{\alpha}\right)\right)$$

where

$$\epsilon\left(s_t,i,k,\hat{\alpha}\right)=\left(\hat{V}_i\left(k,s_t\right)-\hat{V}_i\left(0,s_t\right)\right)-\beta\left(\tilde{\Lambda}_i\left(s_t;k\right)-\tilde{\Lambda}_i\left(s_t;0\right)\right)$$
$$-\left(\Phi_i\left(k,s_t\right)-\Phi_i\left(0,s_t\right)+\beta\left(\tilde{\Phi}_i\left(s_t;k\right)-\tilde{\Phi}_i\left(s_t;0\right)\right)\right)'\theta_0.$$

In general, we can also use an instrumental variable interpretation for the identity relation $\epsilon\left(s_t, i, k, \alpha_0, \theta_0\right) \equiv 0$ for all s_t and i and k. The analog identity relation we are imposing in the sample is that for all s_t, i and k,

$$\epsilon\left(s_t, i, k, \hat{\alpha}, \theta_0\right) \equiv 0.$$

In order to use this to solve for an estimate of $\hat{\theta}$, we can construct any set of (potentially random) functions $A\left(s_t, i, k\right)$ with the same dimension as θ, and solve $\hat{\theta}$ from the sampling implications:

$$\sum_t A\left(s_t, i, k\right) \epsilon\left(s_t, i, k, \hat{\alpha}, \hat{\theta}\right) = 0.$$

Therefore, in general using a set of instruments $A\left(s_t, i, k\right)$,

$$\hat{\theta} = \left(\sum A\left(s_t, i, k\right) X\left(s_t, i, k, \hat{\alpha}\right)'\right)^{-1} \left(\sum A\left(s_t, i, k\right) Y\left(s_t, i, k, \hat{\alpha}\right)'\right),$$

In the following, we focus on the case where $X\left(s_t, i, k, \hat{\alpha}\right)$ is used as the instruments. We will see that the estimation of $\hat{\alpha}$ in the random instrument will have no impact on the variance.

First of all, under suitable regularity conditions, it is easy to show that

$$\frac{1}{T} \sum X\left(s_t, i, k, \hat{\alpha}\right) X\left(s_t, i, k, \hat{\alpha}\right)' \xrightarrow{p} G \equiv \sum_{i=1}^{n} \sum_{k=1}^{K} E_s X\left(s_t, i, k\right) X\left(s_t, i, k\right)'$$

where E_s is taken with respect to the stationary distribution of the state variables, and the random quantities in $X\left(s_t, i, k\right)$ are evaluated at the population value.

The next step is to evaluate the asymptotic distribution of

$$\frac{1}{\sqrt{T}} \sum X\left(s_t, i, k, \hat{\alpha}\right) \epsilon\left(s_t, i, k, \hat{\alpha}\right) \xrightarrow{d} N\left(0, \Omega\right).$$

and find the analytic expression for Ω. For this purpose, note that if α is known for certain, then

$$X\left(s_t, i, k, \alpha_0\right) \epsilon\left(s_t, i, k, \alpha_0\right) \equiv 0$$

and

$$\frac{\partial X\left(s_t, i, k, \alpha_0\right)}{\partial \alpha} \epsilon\left(s_t, i, k, \alpha_0\right) \equiv 0.$$

Therefore all the first order statistical uncertainty in $\hat{\theta}$ is induced by the estimation of α in $\epsilon\left(s_t, i, k, \alpha\right)$. We can then write

$$\frac{1}{\sqrt{T}} \sum X\left(s_t, i, k, \hat{\alpha}\right) \epsilon\left(s_t, i, k, \hat{\alpha}\right)=\sum_{i, k} E_s X\left(s_t, i, k, \hat{\alpha}\right) \frac{\partial \epsilon\left(s_t, i, k, \hat{\alpha}\right)}{\partial \alpha} \sqrt{T}(\hat{\alpha}-\alpha)+o_p(1) .$$

Hence, $\Omega=H \Sigma H^{\prime}$, where $\sqrt{T}\left(\hat{\alpha}-\alpha_0\right) \xrightarrow{d} N(0, \Sigma)$ and

$$H=\sum_{i, k} E_s X\left(s_t, i, k, \hat{\alpha}\right) \frac{\partial \epsilon\left(s_t, i, k, \alpha\right)}{\partial \alpha} .$$

Obviously $\hat{\Sigma}$ can be read off the variance covariance of α. To construct the asymptotic variance one only needs to be able to estimate H. Apparently, H can be estimated by

$$\hat{H}=\frac{1}{T} \sum X\left(s_t, i, k, \hat{\alpha}\right) \frac{\partial \hat{\epsilon}\left(s_t, i, k, \alpha\right)}{\partial \hat{\alpha}} .$$

The only remaining question is how to compute $\frac{\partial \hat{\epsilon}\left(s_t, i, k, \alpha\right)}{\partial \hat{\alpha}}$ efficiently. Recall the definition of $\epsilon\left(s_t, i, k, \alpha\right)$, we note that this derivative depends on the derivatives of the various components of the value function with respect to α:

$$\begin{aligned} \frac{\partial \hat{\epsilon}\left(s_t, i, k, \hat{\alpha}\right)}{\partial \alpha}= & \left(\frac{\partial \hat{V}_i\left(k, s_t, \hat{\alpha}\right)}{\partial \alpha}-\frac{\partial \hat{V}_i\left(0, s_t, \hat{\alpha}\right)}{\partial \alpha}\right)-\beta\left(\frac{\partial \tilde{\Lambda}_i\left(s_t ; k, \hat{\alpha}\right)}{\partial \alpha}-\frac{\partial \tilde{\Lambda}_i\left(s_t ; 0, \hat{\alpha}\right)}{\partial \alpha}\right) \\ & -\left(\frac{\partial \Phi_i\left(k, s_t, \hat{\alpha}\right)}{\partial \alpha}-\frac{\partial \Phi_i\left(0, s_t, \hat{\alpha}\right)}{\partial \alpha}+\beta\left(\frac{\partial \tilde{\Phi}_i\left(s_t ; k, \hat{\alpha}\right)}{\partial \alpha}-\frac{\partial \tilde{\Phi}_i\left(s_t ; 0, \hat{\alpha}\right)}{\partial \alpha}\right)\right)^{\prime} \hat{\theta} . \end{aligned}$$

We now describe how each of the above components can be calculated analytically.

First of all, it is clear that

$$\frac{\partial\left[V_i(k, s, \hat{\alpha})-V_i(0, s, \hat{\alpha})\right]}{\partial \alpha}=\left[\frac{\partial F\left(V_i(k, s, \hat{\alpha})-V_i(0, s, \hat{\alpha})\right)}{\partial\left[V_i(k, s, \hat{\alpha})-V_i(0, s, \hat{\alpha}), k=1, \ldots, K\right]}\right]^{-1} \frac{\partial[P(a \mid s), a=1, \ldots K]}{\partial \alpha} .$$

The term inside the inverse clearly involves the density of the joint density of the error terms ϵ_i and can be easily evaluated analytically. The second term can be read off from the first step estimation of the choice probabilities.

If the dimension of α increases as the sample size increases, the parametric estimate of $\hat{P}(a|s)$ converges to the true nonparametric $P(a|s)$. Using the arguments in Newey (1994a), (see section 4, in particular, proposition 4 in page 1361), one can show that in the limit, regardless of the nonparametric method used to estimate $p(a|s)$, the first term

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X(s_t, i, k)\left(\hat{V}_i(k, s_t) - \hat{V}_i(0, s_t)\right)$$

will be asymptotically equivalent to

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X(s_t, i, k)\left[\frac{\partial F(V_i(k, s_t, \hat{\alpha}) - V_i(0, s_t, \hat{\alpha}))}{\partial [V_i(k, s_t, \hat{\alpha}) - V_i(0, s_t, \hat{\alpha})]}\right]^{-1}(a_t - p(a_t|s_t)).$$

This is also the limit when the dimension of α increases to infinity of

$$\sum_{i,k} E_s X(s, i, k)\frac{\partial [V_i(k, s, \hat{\alpha}) - V_i(0, s, \hat{\alpha})]}{\partial \alpha}\sqrt{T}(\hat{\alpha} - \alpha),$$

where the relevant components of α solves the score equation

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{a_t - p(a|s_t)}{p(a|s_t)(1 - p(a|s_t))}p_\alpha(a_t|s_t) = 0.$$

Next note that

$$\Lambda_i(s) = \sum_{a_i} E(\epsilon_i(a_i)|a_i \text{ is chosen}) P(a_i|s)$$

is also a function of $p_i(a|s)$, hence $V_i(k, s) - V_i(0, s)$. It derivative $\frac{\partial \Lambda_i(s,\alpha)}{\partial \alpha}$ can also be calculated analytically in principle.

Next, consider

$$\frac{\partial \tilde{\Lambda}_i(s; k, \hat{\alpha})}{\partial \alpha} = \int \frac{\partial \bar{\Lambda}(s'; \hat{\alpha})}{\partial \alpha} dp(s'|s, \hat{\alpha}) + \int \bar{\Lambda}(s'; \hat{\alpha})\frac{\partial \log p(s'|s, \hat{\alpha})}{\partial \alpha} dp(s'|s, \hat{\alpha}).$$

To calculate the first term, one needs to calculate the entire function of $\frac{\partial \bar{\Lambda}(s;\hat{\alpha})}{\partial \alpha}$. For this purpose, note that

$$\frac{\partial \bar{\Lambda}_i(s)}{\partial \alpha} = \frac{\partial \Lambda(s)}{\partial \alpha} + \beta \int \bar{\Lambda}_i(s')\frac{\partial \log p(s'|s, \alpha)}{\partial \alpha} dp(s'|s, \alpha) + \beta \int \frac{\partial \bar{\Lambda}_i(s')}{\partial \alpha} dp(s'|s, \hat{\alpha}).$$

Clearly, this involves another fixed point evaluation for $\frac{\partial \bar{\Lambda}_i(s)}{\partial \alpha}$, given the values of

$$\frac{\partial \Lambda(s)}{\partial \alpha} + \beta \int \bar{\Lambda}_i\left(s'\right) \frac{\partial \log p\left(s'|s,\alpha\right)}{\partial \alpha} dp\left(s'|s,\alpha\right).$$

The other terms follow a similar calculation.

4 Semiparametric Variance

The general framework of Newey (1994a) provides the powerful result that the influence of the first stage nonparametric estimation on the second stage parametric variance depends only on the nonparametric functional to be estimated, and not on the particular nonparametric estimation method for estimating it. This suggests that we can give the form of the limit semiparametric variance for the dynamic discrete models regardless of the nonparametric methods used to estimate the choice probabilities and the transition process of state variables.

In the framework of Newey (1994a), one is interested in finding the influence function representation for

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[m\left(z_t, \theta_0, \hat{h}\right) - m\left(z_t, \theta_0, h_0\right) \right],$$

where $\hat{h}$ is a vector of nonparametric estimate of conditional expectation functions $h(x)$ of y given x, a subset of z. Suppose there exists a function $D(z,h)$ that is linear in h such that for any parametric sub path $h(\theta)$ through the space of nonparametric functions h,

$$\frac{\partial}{\partial \theta} Em\left(z, h(\theta)\right) = \frac{\partial}{\partial \theta} ED\left(z, h(\theta)\right).$$

If one can find a set of function $\delta(x)$ such that for all g,

$$ED(z,g) = E\left(\delta(x) g(x)\right),$$

then under suitable regularity conditions

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[m\left(z_t, \theta_0, \hat{h}\right) - m\left(z_t, \theta_0, h_0\right) \right] = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \delta\left(x_t\right)\left(y_t - h_0\left(x_t\right)\right) + o_p(1).$$

We now cast the dynamic model we consider into this general framework. We are concerned with the asymptotic distribution of

$$\frac{1}{\sqrt{T}}\sum X\left(s_t,i,k\right)\epsilon\left(s_t,i,k,\hat{h}\right),$$

where $h=\left(h_1,\ldots,h_5\right)$ such that h_1 corresponds to the $p_i\left(k|s\right)$ in the estimation of $\Phi\left(\hat{p}_i\left(k|s\right)\right)$ for $\hat{V}_i\left(k,s\right)-\hat{V}_i\left(0,s\right)$ in equation (3). h_2 corresponds to the estimation of $p\left(a_i|s\right)$ in the estimation of $\Lambda_i\left(s\right)$ in equation (5) which is subsequently transmitted to $\tilde{\Lambda}_i\left(s;a_i\right)$ through equations (7) and (10). h_3 corresponds to the estimation of $p\left(s'|s\right)$ in the calculation of $\bar{\Lambda}_i\left(s\right)$ in equation (7) which is then transmitted to $\tilde{\Lambda}_i\left(s;a_i\right)$ through equation (10). h_4 corresponds to the estimation of $p\left(s'|s,a_i\right)$ in calculating $\tilde{\Lambda}_i\left(s;a_i\right)$ in equation (10). h_5 corresponds to the estimation of $p\left(a_{-i}|s\right)$ in equation (9) for $\Phi_i\left(a_i,s\right)$. h_6 corresponds to the estimation of $p\left(a|s\right)$ in equation (6) which is then transmitted to $\tilde{\Phi}_i\left(s,a_i\right)$ through equations (8) and (11). h_7 corresponds to the estimation of $p\left(s'|s\right)$ in the equation (8) which is then transmitted through equation (11) to $\tilde{\Phi}_i\left(s,a_i\right)$. h_8 corresponds to the estimation of $p\left(s'|s,a_i\right)$ in equation (11) for $\tilde{\Phi}_i\left(s,a_i\right)$. Therefore for $J=8$ we are looking for a set of J linear functions in $h^j, j=1,\ldots,J$, $\xi\left(s_t,i,k,h^j\right)$, such that for any parametric path h_θ of the unknown conditional expectation function

$$\frac{\partial}{\partial\theta}E_{s_t}X\left(s_t,i,k\right)\epsilon\left(s_t,i,k,h_\theta\right)=\sum_{j=1}^{J}\frac{\partial}{\partial\theta}E_{s_t}\xi\left(s_t,i,k,h_\theta^j\right).$$

If such linear functions can be found, the results of Newey (1994a) show that the asymptotic linear representation is given by, where $h^j\left(s_t\right)=E\left(y_t^j|s_t\right)$,

$$\frac{1}{\sqrt{T}}\sum X\left(s_t,i,k\right)\epsilon\left(s_t,i,k,\hat{h}\right)=\frac{1}{\sqrt{T}}\sum\sum_{j=1}^{J}\frac{\xi\left(s_t,i,k,h^j\right)}{h^j\left(s_t\right)}\left(y_t^j-h^j\left(s_t\right)\right)+o_p\left(1\right).$$

We describe each of these J functions that are linear in h^j in the following.

$\xi\left(\mathbf{s_t},\mathbf{i},\mathbf{k},\mathbf{h^1}\right)$ can be seen as a special case of example 2 in page 1362 of Newey (1994a), and has be verified previously to be

$$X\left(s_t,i,k\right)\left[\frac{\partial F\left(V_i\left(k,s_t,\hat{\alpha}\right)-V_i\left(0,s_t,\hat{\alpha}\right)\right)}{\partial\left[V_i\left(k,s_t,\hat{\alpha}\right)-V_i\left(0,s_t,\hat{\alpha}\right)\right]}\right]^{-1}p\left(a_t|s_t\right)$$

Therefore its asymptotic linear representation is

$$\eta\left(s_t, i, k, h^1\right) = X\left(s_t, i, k\right)\left[\frac{\partial F\left(V_i\left(k, s_t, \hat{\alpha}\right) - V_i\left(0, s_t, \hat{\alpha}\right)\right)}{\partial\left[V_i\left(k, s_t, \hat{\alpha}\right) - V_i\left(0, s_t, \hat{\alpha}\right)\right]}\right]^{-1}\left(a_t - p\left(a_t|s_t\right)\right)$$

$\xi\left(\mathbf{s_t}, \mathbf{i}, \mathbf{k}, \mathbf{h^2}\right)$**.** Note that equations (8) and (10) defines a linear functional from $\Lambda_i\left(s\right)$ to $\tilde{\Lambda}_i\left(s; a_i\right)$ for each a_i, which we denote by

$$\tilde{\Lambda}_i^{s,a_i}\left(\Lambda_i\left(\cdot\right)\right)\left(s\right).$$

This in turns defines a linear mapping from $p\left(a_i|s\right)$ to $\tilde{\Lambda}_i\left(s; a_i\right)$ by

$$\tilde{\Lambda}_i^{s,a_i}\left(\frac{\partial\Lambda_i\left(\cdot\right)}{\partial p\left(a_i|\cdot\right)}p\left(a_i|\cdot\right)\right)\left(s\right).$$

Therefore

$$\xi\left(s_t, i, k, h^2\right) = -\beta X\left(s_t, i, k\right)\tilde{\Lambda}_i^{s,a_i}\left(\frac{\partial\Lambda_i\left(\cdot\right)}{\partial p\left(a_i|\cdot\right)}p\left(a_i|\cdot\right)\right)\left(s_t\right)$$

and the corresponding asymptotic linear representation is

$$\eta\left(s_t, i, k, h^2\right) = -\beta X\left(s_t, i, k\right)\tilde{\Lambda}_i^{s,a_i}\left(\frac{\partial\Lambda_i\left(\cdot\right)}{\partial p\left(a_i|\cdot\right)}p\left(a_i|\cdot\right)\right)\left(s_t\right)\frac{a_{it} - s_t}{p\left(a_i|s_t\right)}.$$

$\xi\left(\mathbf{s_t}, \mathbf{i}, \mathbf{k}, \mathbf{h^3}\right)$**.** The mapping from $p\left(s'|s\right)$ to $\bar{\Lambda}_i\left(s\right)$ in equation (7) is nonlinear. It can be linearized around the true values $\bar{\Lambda}_i^0\left(s\right)$ and $p^0\left(s'|s\right)$ of $\bar{\Lambda}_i\left(s\right)$ and $p\left(s'|s\right)$ as

$$\begin{aligned}\bar{\Lambda}_i\left(s\right) =& \Lambda_i\left(s\right) + \beta\int\bar{\Lambda}_i\left(s'\right)dp^0\left(s'|s\right) + \beta\int\bar{\Lambda}_i^0\left(s'\right)dp\left(s'|s\right)\\ =& \Lambda_i\left(s\right) + \beta\int\bar{\Lambda}_i\left(s'\right)dp^0\left(s'|s\right) + \beta E\left[\bar{\Lambda}_i^0\left(s'\right)|s\right].\end{aligned} \tag{12}$$

Equations (12) and (10) therefore define a linear mapping from $E\left[\bar{\Lambda}_i^0\left(s'\right)|s\right]$ to $\tilde{\Lambda}_i\left(s, a_i\right)$ for all a_i, which we will denote by

$$\tilde{\Lambda}_i^{s,a_i}\left(E\left[\bar{\Lambda}_i^0\left(s'\right)|\cdot\right]\right)\left[s\right].$$

Therefore

$$\xi\left(s_t, i, k, h^3\right) = -\beta X\left(s_t, i, k\right)\tilde{\Lambda}_i^{s,a_i}\left(E\left[\bar{\Lambda}_i^0\left(s'\right)|\cdot\right]\right)\left(s_t\right),$$

and the corresponding asymptotic linear representation is given by

$$\eta\left(s_t, i, k, h^3\right) = -\beta X\left(s_t, i, k\right)\tilde{\Lambda}_i^{s,a_i}\left(E\left[\bar{\Lambda}_i^0\left(s'\right)|\cdot\right]\right)\left(s_t\right)\frac{\bar{\Lambda}_i^0\left(s_t'\right) - E\left[\bar{\Lambda}_i^0\left(s'\right)|s_t\right]}{E\left[\bar{\Lambda}_i^0\left(s'\right)|s_t\right]}.$$

$\xi\left(\mathbf{s_t},\mathbf{i},\mathbf{k},\mathbf{h^4}\right)$**.** The mapping from $p\left(s'|s,a_i\right)$ to $\tilde{\Lambda}_i\left(s;a_i\right)$ in equation (10) is already a linear conditional expectation operation:

$$\tilde{\Lambda}_i\left(s;a_i\right)=E\left[\Lambda_i\left(s'\right)|s,a_i\right].$$

Therefore

$$\xi\left(s_t,i,k,h^4\right)=-\beta X\left(s,i,k\right)E\left[\Lambda_i\left(s'\right)|s,a_i\right],$$

and the corresponding asymptotic linear representation is given by

$$\eta\left(s_t,i,k,h^4\right)=-\beta X\left(s_t,i,k\right)\left(\Lambda_i\left(s_t'\right)-E\left[\Lambda_i\left(s'\right)|s_t,a_i\right]\right).$$

$\xi\left(\mathbf{s_t},\mathbf{i},\mathbf{k},\mathbf{h^5}\right)$**.** Since the mapping from $p\left(a_{-i}|s\right)$ to $\Phi_i\left(a_i,s\right)$ in equation (9) is a linear conditional expectation,

$$\xi\left(s_t,i,k,h^5\right)=-X\left(s_t,i,k\right)\Phi_i\left(a_i,s\right)=-X\left(s_t,i,k\right)E_{a_{-i}|s}\Phi_i\left(a_i,a_{-i},s\right),$$

and the corresponding asymptotic linear representation is

$$\eta\left(s_t,i,k,h^5\right)=-X\left(s_t,i,k\right)\left[\Phi_i\left(a_{it},a_{-it},s_t\right)-\Phi_i\left(a_{it},s_t\right)\right].$$

$\xi\left(\mathbf{s_t},\mathbf{i},\mathbf{k},\mathbf{h^6}\right)$**.** Note that the mapping from $E\left(\Phi_i\left(a,s\right)|s\right)$ in equation (6) to $\tilde{\Phi}_i\left(s,a_i\right)$ through equations (6), (8) and (11) is linear in every step, and therefore the mapping is a linear one. Denote this linear mapping by

$$\tilde{\Phi}_i^{s,a_i}\left(E\left[\Phi_i\left(a,s\right)|\cdot\right]\right)\left(s_t\right).$$

Therefore

$$\xi\left(s_t,i,k,h^6\right)=-\beta X\left(s_t,i,k\right)\tilde{\Phi}_i^{s,a_i}\left(E\left[\Phi_i\left(a,s\right)|\cdot\right]\right)\left(s_t\right)\theta_0$$

and its corresponding asymptotic representation is

$$\eta\left(s_t,i,k,h^6\right)=-\beta X\left(s_t,i,k\right)\tilde{\Phi}_i^{s,a_i}\left(E\left[\Phi_i\left(a,s\right)|\cdot\right]\right)\left(s_t\right)\frac{\Phi_i\left(a,s_t\right)-E\left[\Phi_i\left(a,s_t\right)|s_t\right]}{E\left[\Phi_i\left(a,s_t\right)|s_t\right]}\theta_0.$$

$\xi\left(\mathbf{s_t},\mathbf{i},\mathbf{k},\mathbf{h^7}\right)$**.** The mapping from $p\left(s'|s\right)$ to $\bar{\Phi}_i\left(s\right)$ in equation (8) is nonlinear. It can be linearized around the true values $\bar{\Phi}_i^0\left(s\right)$ and $p^0\left(s'|s\right)$ of $\bar{\Phi}_i\left(s\right)$ and $p\left(s'|s\right)$ as

$$\bar{\Phi}_i\left(s\right)=\Phi_i\left(s\right)+\beta\int\bar{\Phi}_i\left(s'\right)dp^0\left(s'|s\right)+\beta E\left[\bar{\Phi}_i^0\left(s'\right)|s\right]. \tag{13}$$

Equations (13) and (11) define a linear mapping from $E\left[\bar{\Phi}_i^0\left(s'\right)|s\right]$ to $\tilde{\Phi}_i\left(s,a_i\right)$. We denote this linear mapping by

$$\tilde{\Phi}_i^{s,a_i}\left(E\left[\bar{\Phi}_i^0\left(s'\right)|\cdot\right]\right)\left(s_t\right).$$

Therefore

$$\xi\left(s_t,i,k,h^7\right)=-\beta X\left(s_t,i,k\right)\tilde{\Phi}_i^{s,a_i}\left(E\left[\bar{\Phi}_i^0\left(s'\right)|\cdot\right]\right)\left(s_t\right)\theta_0.$$

and the asymptotic linear representation is

$$\eta\left(s_t,i,k,h^7\right)=-\beta X\left(s_t,i,k\right)\tilde{\Phi}_i^{s,a_i}\left(E\left[\bar{\Phi}_i^0\left(s'\right)|\cdot\right]\right)\left(s_t\right)\frac{\bar{\Phi}_i^0\left(s_t'\right)-E\left[\bar{\Phi}_i^0\left(s'\right)|s_t\right]}{E\left[\bar{\Phi}_i^0\left(s'\right)|s_t\right]}\theta_0.$$

$\xi\left(\mathbf{s_t},\mathbf{i},\mathbf{k},\mathbf{h^8}\right)$**.** The mapping from $p\left(s'|s,a_i\right)$ to $\tilde{\Phi}_i\left(s,a_i\right)$ in equation (11) is a linear conditional expectation operator. Therefore,

$$\xi\left(s_t,i,k,h^8\right)=-\beta X\left(s_t,i,k\right)E\left(\bar{\Phi}_i\left(s'\right)|s_t,a_i\right)\theta_0$$

and the corresponding asymptotic linear representation is given by

$$\eta\left(s_t,i,k,h^8\right)=-\beta X\left(s_t,i,k\right)\left[\bar{\Phi}_i\left(s_t'\right)-E\left(\bar{\Phi}_i\left(s'\right)|s_t,a_{it}\right)\right]\theta_0.$$

To summarize, we can write

$$\frac{1}{\sqrt{T}}\sum_t X\left(s_t,i,k\right)\epsilon\left(s_t,i,k,\hat{h}\right)=\frac{1}{\sqrt{T}}\sum_t\sum_{j=1}^J\eta\left(s_t,i,k,h^j\right)+o_p\left(1\right),$$

which implies that

$$\frac{1}{\sqrt{T}}\sum_t X\left(s_t,i,k\right)\epsilon\left(s_t,i,k,\hat{h}\right)\xrightarrow{d}N\left(0,\Omega\right)$$

where

$$\Omega = Var\left(\sum_{j=1}^{J} \eta\left(s_t, i, k, h^j\right)\right).$$

The asymptotic distribution of the two step semiparametric estimator follows as

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, G\Omega G'\right).$$

Each of the linear representations $\eta\left(s_t, i, k, h^j\right)$ can be consistently estimated by replacing population quantities with sample analogs, denote this by $\hat{\eta}\left(s_t, i, k, \hat{h}^j\right)$. Ω can then be consistently estimated by

$$\hat{\Omega} = \frac{1}{T}\sum_t \left(\sum_{j=1}^{J} \eta\left(s_t, i, k, h^j\right)\right)\left(\sum_{j=1}^{J} \eta\left(s_t, i, k, h^j\right)\right)'.$$

G can also be consistently estimated by

$$\hat{G} = \frac{1}{T}\sum X\left(s_t, i, k, \hat{\alpha}\right) X\left(s_t, i, k, \hat{\alpha}\right)'.$$

5 Existence of Equilibrium

6 Regularity Conditions

Newey (1994a) gave a general framework for providing regularity conditions for the asymptotic normality of a sample moment condition that depends on nonparametric estimate of nuisance parameters in the first stage:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} m\left(z_t, \hat{h}\right),$$

where $\hat{h}$ is a nonparametric estimate of $h_0 = E\left(y|z\right)$ such that $Em\left(z_t, h_0\right) = 0$. Several high level conditions are required:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(m\left(z_t, \hat{h}\right) - m\left(z_t, h\right)\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} D\left(z_t, \hat{h} - h\right) + o_p\left(1\right) \tag{14}$$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} D\left(z_t, \hat{h}-h\right) = \sqrt{T} E D\left(z_t, \hat{h}-h\right) + o_p(1) \tag{15}$$

where the expectation is understood to be taken with respect to z_t holding $\hat{h}$ fixed.

There exists square integrable mean zero function $\alpha(y,z)$, and a random probability measure $\hat{F}(y,z)$, such that

$$\sqrt{T} E D\left(z_t, \hat{h}-h\right) = \sqrt{T}\int \alpha(y,z)\, d\hat{F}(y,z) + o_p(1). \tag{16}$$

Finally, we require that

$$\sqrt{T}\int \alpha(y,z)\, d\hat{F}(y,z) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \alpha(y_t, z_t) + o_p(1). \tag{17}$$

When the moment condition $m(z_t, h)$ is a smooth functional of the nonparametric function h, condition (14) is typically satisfied by the following two conditions. First, for all h close enough to h_0, there is an integrable function $b(z)$ such that

$$\left|\left| m(z,h) - m(z,h_0) - D(z, h-h_0) \right|\right| \leq b(z) \left|\left| h - h_0 \right|\right|^2 \tag{18}$$

where $||\cdot||$ denotes the Euclidean norm of a scalar or a vector, and

$$\sqrt{T}\left|\left|\hat{h} - h\right|\right|^2 \xrightarrow{p} 0. \tag{19}$$

In the terminologies of Newey (1994a), because our second stage is linear in the parametric parameters β, only his assumptions 5.1, 5.2 and 5.3 are relevant. His assumptions 5.4, 5.5 and 5.6 are not relevant for us because they are only relevant for second stage estimators that are nonlinear in the parametric parameters β.

Newey (1994a) also provides sufficient assumptions for sieve estimation such that these high level conditions are satisfied when $\hat{h}$ is estimated using a series estimator. In the following, we explain the primitive conditions in Newey (1994a) which imply the above conditions (14), (15), (16) and (17).

Condition (14) is implied by (18) and (19). Condition (18) is a property of the model itself and needs to be verified using primitive assumptions regarding the latent distributions in the dynamic game model. We leave verification for a later section. Condition (19) is implied by the following primitive assumptions.

Assumption (19)

1. For some constant C, for all K there is π such that $|h(z) - p^K(z)' \pi|_0 \leq CK^{-\alpha}$. Here the norm of a scalar function $|f(z)|_0$ is defined as

   $$|f(z)|_0 = \max_{z \in \mathcal{Z}} |f(z)|$$

   In other words, if we define $\Sigma_K = Ep^K(z) p^K(z)$, $\pi_K = \Sigma_K^{-1} Ep^K(z) h(z)$, and

   $$h_K(z) = p^K(z)' \pi_K,$$

   as the population projection of $h(Z)$ into the linear space spanned by $p^K(z)$, then we can write

   $$|h(z) - h_K(z)|_0 \leq CK^{-\alpha}.$$

2. $\sqrt{T}\zeta_0(K)^2 \left[\left(\frac{K}{T}\right) + K^{-2\alpha}\right] \longrightarrow 0$, where

   $$\zeta_0(K) = \sup_{z \in \mathcal{Z}} ||p^K(z)||,$$

   and $p^K(z)$ is the vector of sieve functions for the first K terms.

Assumption (19) implies condition (19) because Newey (1994b) showed that

$$||\hat{h} - h|| \leq |\hat{h} - h|_0 \equiv \max_{z \in \mathcal{Z}} |\hat{h}(z) - h(z)| = O_p\left(\zeta_0(K)\left[\sqrt{\frac{K}{T}} + K^{-\alpha}\right]\right).$$

Condition (15), which is a stochastic equicontinuity condition, is implied by the following assumption (essentially assumption 6.5 of pp1371 of Newey (1994a)):

Assumption (15)

1. For a square integrable function $b(z)$, $||D(z,h) - D(z,h_0)|| \leq b(z)|h - h_0|_0$.

2. $K^{-\alpha} \to 0$, which holds as long as $K \to \infty$ and $\alpha > 0$.
3. $\left(\sum_{k=1}^{K} |p_k(z)|_0^2\right)^{1/2} \left[(K/T)^{1/2} + K^{-\alpha}\right] \longrightarrow 0.$

Proof that assumption (15) implies condition (15): The goal is to prove equation (15). We break down equation (15) into two parts:

$$
\begin{aligned}
&\frac{1}{\sqrt{T}} \sum_{t=1}^{T} D\left(z_t, \hat{h} - h\right) - \sqrt{T} ED\left(z_t, \hat{h} - h\right) \\
=& \underbrace{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} D\left(z_t, \hat{h} - h_K\right) - \sqrt{T} ED\left(z_t, \hat{h} - h_K\right)}_{(a)} + \underbrace{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} D\left(z_t, h_K - h\right) - \sqrt{T} ED\left(z_t, h_K - h\right)}_{(b)}.
\end{aligned}
$$

Essentially (a) controls for the variance while (b) controls for the bias. Next the stochastic orders of parts (a) and (b) can be calculated, using the fact that both have zero means.

$$
\begin{aligned}
E||(b)||^2 =& E||D\left(z_t, h_K - h\right) - ED\left(z_t, h_K - h\right)||^2 \\
\leq& E||D\left(z_t, h_K - h\right)||^2 \leq Eb\left(z\right)^2 |h_K - h|_0^2 \leq CK^{-2\alpha} \to 0,
\end{aligned}
$$

where the last inequality follows from part 1 of assumption (19) and the last convergence follows from part 2 of assumption (15). Therefore $(b) = o_p(1)$ by the usual Markov inequality.

Regarding part (a), note that $\hat{h}(z) - h_K(z) = p^K(z)' (\hat{\pi}_K - \pi_K)$, where

$$
\hat{\pi}_K = \hat{\Sigma}_K^{-1} \frac{1}{T} \sum_{t=1}^{T} p^K(z_t)' y_t, \qquad \hat{\Sigma}_K = \frac{1}{T} \sum_{t=1}^{T} p^K(z_t) p^K(z)'.
$$

Because of the linearity of $D(z, h)$ in h,

$$
(a) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(D\left(z_t, p^K(z)\right) - ED\left(z_t, p^K(z)\right)\right)' (\hat{\pi}_K - \pi_K).
$$

Therefore because of its zero mean, $(1) = O_p\left(\sqrt{E(1)^2}\right)$ and

$$
E(a)^2 \leq E\left[D\left(z_t, p^K(z_t)\right)' (\hat{\pi}_K - \pi_K)\right]^2.
$$

Then using Cauchy-Schwartz inequality,

$$\begin{aligned} E\left(a\right)^2 \leq & Eb\left(z\right)^2 ||p^K\left(z_t\right)||^2 ||\left(\hat{\pi}_K - \pi_K\right)||^2 \\ \leq & C\sum_{k=1}^{K} |p_k\left(z\right)|_0^2 ||\left(\hat{\pi}_K - \pi_K\right)||^2. \end{aligned}$$

Since Newey (1994b) showed that $||\left(\hat{\pi}_K - \pi_K\right)|| = O_p\left(\left(K/T\right)^{1/2} + K^{-\alpha}\right)$, that $(a) = o_p(1)$ follows from part 3 of assumption (15).

Therefore equation (15) is verified. We are left to explain how to verify conditions (16) and (17). Consider (16) first. It turns out in our case, (16) holds as an exact identity without the $o_p(1)$ term, and no additional assumption is needed. In other words, equation (16) is a tautology and can be combined by equation (17). Just to illustrate (16) for $\alpha\left(y_t, z_t\right) = \left(y_t - h\left(z_t\right)\right)\delta\left(z_t\right)$, write

$$\begin{aligned} ED\left(z, \hat{h} - h\right) = & E\delta\left(z\right)\left(\hat{h}\left(z\right) - h\left(z\right)\right) \\ = & Ep^K\left(z\right)' \hat{\Sigma}_K^{-1} \frac{1}{T}\sum_{t=1}^{T} p^K\left(z_t\right)\delta\left(z\right)\left(y_t - h\left(z\right)\right) \equiv \int \alpha\left(z, y\right) d\hat{F}\left(y, z\right). \end{aligned}$$

The second equality follows from the definition of $\hat{h}(z)$ as a least square prediction and the fact that as long as the first term in $p^K(z)$ is the constant term, $p_1(z) \equiv 1$, then

$$p^K\left(z\right)' \hat{\Sigma}_K^{-1} \frac{1}{T} p^K\left(z_t\right) = 1. \tag{20}$$

This is because the left hand is the predicted value at z of regression the vector of 1's on the sieve functions $p^K(z_t)$. The third equality is the definition of the measure $\hat{F}(y,z)$ and the integral against $\hat{F}(y,z)$. Here the sequence of z_t on the left hand side is understood to be a fixed sequence of constants that do not enter the definition of the integration. Integration over y is the summation over y_t, while integration over z is given by the expectation E operator.

Finally we are left to explain assumptions that imply condition (17). Assumption 6.6 of Newey (1994a) (page 1371) serves this purpose, which we restate as:

Assumption (17)

1. Define $\xi_K = \Sigma_K^{-1} E p^K(z)\,\delta(z)$ and $\delta_K(z) = \xi_K' p^K(z)$ to be the population projection of $\delta(z)$ into the linear space spanned by $p^K(z)$. Then K is chosen such that

$$TE\left[|\delta(z) - \delta_K(z)|^2\right] E\left[|h(z) - h_K(z)|^2\right] \to 0.$$

2. $\zeta_0(K)^4 K/T \to 0$.

3. $\zeta_0(K)^2 E(h(z) - h_K(z))^2 \to 0$ and $E(\delta(z) - \delta_K(z))^2 \to 0$

The left hand side of (17), which is equal to the left hand side of (16), can be broken down into

$$\begin{aligned} ED\left(z, \hat{h} - h\right) =& ED\left(z, \hat{h} - h_K\right) + ED(z, h_K - h) \\ =& \underbrace{E\delta(z)\left(\hat{h}(z) - h_K(z)\right)}_{(c)} + \underbrace{E\delta(z)(h_K(z) - h(z))}_{(d)}. \end{aligned}$$

Part (c) controls variance term while part (d) controls the bias term. As usual, the bias term needs to be relatively small. First note that because of the orthogonality definition of linear projection,

$$E\delta_K(z)(h(z) - h_K(z)) = 0,$$

therefore $(d) = E(\delta(z) - \delta_K(z))(h_K(z) - h(z))$, and

$$\begin{aligned} |\sqrt{T}(d)|^2 =& TE|(\delta(z) - \delta_K(z))(h_K(z) - h(z))|^2 \\ \leq& TE\left[|\delta(z) - \delta_K(z)|^2\right] E\left[|h(z) - h_K(z)|^2\right] \to 0, \end{aligned}$$

according to part 1 of assumption (17).

Finally consider part (c). Define $\Psi_K = E\delta(z)\,p^K(z)$ and note that

$$\begin{aligned} E\delta(z)\,h_K(z) = \Psi_K'\pi_K =& \Psi_K'\hat{\Sigma}_K^{-1}\frac{1}{T}\sum_{t=1}^{T} p^K(z_t)\underbrace{p^K(z_t)'\pi_k}_{h_K(z_t)} \\ =& \Psi_K'\hat{\Sigma}_K^{-1}\frac{1}{T}\sum_{t=1}^{T} p^K(z_t)\,h_K(z_t). \end{aligned}$$

Then we can write

$$(c) = \Psi_K' \hat{\Sigma}_K^{-1} \frac{1}{T} \sum_{t=1}^{T} p^K (z_t) (y_t - h_K (z_t)) .$$

There are two remaining steps involved here. In the first one, one shows that $\hat{\Sigma}_K$ can be replaced by Σ_K:

$$\Psi_K' \left(\hat{\Sigma}_K^{-1} - \Sigma_K^{-1} \right) \frac{1}{T} \sum_{t=1}^{T} p^K (z_t) (y_t - h_K (z_t)) .$$

I believe this follows from the facts that (1) $||\hat{\Sigma} - \Sigma|| = O_p \left(\zeta_0 (K)^2 / \sqrt{T} \right) = o_p (1)$, see equation (A.1) of page 1376 of Newey (1994a); (2) $||\Psi_K' \hat{\Sigma}_K^{-1} - \Psi_K' \Sigma_K^{-1}|| = o_p (1)$ and $||\Psi' \hat{\Sigma}^{-1}|| = O_p (1)$, see the top of page 1378 of Newey (1994a); and (3)

$$\Sigma^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} p^K (z_t) (y_t - h_K (z_t)) = O_p \left(\sqrt{K + \zeta_0 (K)^2 E (h (z) - h_K (z))^2} \right) ,$$

as in equation (A.8) of page 1377 of Newey (1994a). I am not 100% sure how (1), (2) and (3) are put together. But taking it for granted, the replacement of $\hat{\Sigma}_K^{-1}$ by Σ_K^{-1} follows from parts 2 and 3 of assumption (17) since

$$\begin{aligned} & \Psi_K' \left(\hat{\Sigma}_K^{-1} - \Sigma_K^{-1} \right) \frac{1}{T} \sum_{t=1}^{T} p^K (z_t) (y_t - h_K (z_t)) \\ =& O_p \left(\frac{\zeta_0^4 (K)}{T} \left(K + \zeta_0 (K)^2 E (h (z) - h_K (z))^2 \right) \right) . \end{aligned}$$

Finally, note that

$$\Psi_K' \Sigma_K^{-1} \frac{1}{T} \sum_{t=1}^{T} p^K (z_t) (y_t - h_K (z_t)) = \frac{1}{T} \sum_{t=1}^{T} \delta_K (z_t) (y_t - h_K (z_t)) ,$$

the last step is to show that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \delta_K (z_t) (y_t - h_K (z_t)) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \delta (z_t) (y_t - h (z_t)) + o_p (1) .$$

The difference between the two sides has two terms, both with zero means:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\delta_K\left(z_t\right)-\delta\left(z_t\right)\right)\left(y_t-h\left(z_t\right)\right)+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\delta_K\left(z_t\right)\left(h\left(z_t\right)-h_K\left(z_t\right)\right).$$

The first term is $o_p\left(1\right)$ because the $E\left[\left(y_t-h\left(z_t\right)\right)^2|z_t\right]$ is uniformly bounded, and therefore the variance of the first term is bounded by

$$CE|\delta_K\left(z_t\right)-\delta\left(z_t\right)|^2\to 0.$$

The second term is $o_p\left(1\right)$ because

$$\begin{aligned}E\delta_K\left(z_t\right)\left(h\left(z_t\right)-h_K\left(z_t\right)\right)^2 =&E\left(\Psi_K'\Sigma_K^{-1}p^K\left(z\right)\left(h\left(z\right)-h_K\left(z\right)\right)\right)^2\\ &\leq\left(\Psi_K'\Sigma_K^{-1}\Psi_K\right)^2Ep^K\left(z\right)'\Sigma_K^{-1}p^K\left(z\right)\left(h\left(z\right)-h_K\left(z\right)\right)^2.\end{aligned}$$

The first part $\left(\Psi_K'\Sigma_K^{-1}\Psi_K\right)=E\delta_K\left(z\right)'\delta_K\left(z\right)$ is bounded, $p^K\left(z\right)'\Sigma_K^{-1}p^K\left(z\right)\leq\xi_0\left(K\right)^2$ being a quadratic norm where the largest eigen value of Σ_K^{-1} is bounded. Therefore the whole term is bounded by

$$\xi_0\left(K\right)^2E\left(h\left(z\right)-h_K\left(z\right)\right)^2\to 0$$

by part 3 of assumption (17).

A Linear probability model and two stage least square

In the standard linear instrument variable model with heteroscedastic error terms,

$$y_t=x_t'\beta+Y_t'\gamma+\epsilon_t$$

where y_t plays against the realized value of Y_t. Let Z_t be a set of instruments that include all the variables in x_t. In other words, x_t is a strict subset of z_t. The reduced form first stage for Y_t is given by

$$Y_t=Z_t'\Pi+v_t.$$

Stata implements two stage least square in the following way. In the first stage, $\hat{\Pi} = (Z'Z)^{-1} Z'Y$, so that $\hat{\Pi} - \Pi = (Z'Z)^{-1} Z'v$. In the second stage, $\hat{\beta}, \hat{\gamma}$ are estimated by equating

$$\sum_{t=1}^{T} \begin{pmatrix} x_t \\ Z_t'\hat{\Pi} \end{pmatrix} \left(y_t - x_t'\hat{\beta} - \left(Z_t'\hat{\Pi} \right) \hat{\gamma} \right) = 0. \tag{21}$$

It is easy to show that

$$\begin{pmatrix} \hat{\beta} - \beta \\ \hat{\gamma} - \gamma \end{pmatrix} \overset{A}{\sim} N\left(0, \hat{A}^{-1}\Omega\hat{A}^{-1}\right)$$

where

$$\hat{A} = \sum_{t=1}^{T} \begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix} \left(x_t, \Pi' Z_t \right) \tag{22}$$

and Ω is the approximate variance covariance matrix of

$$\begin{aligned} & \sum_{t=1}^{T} \begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix} \left(y_t - x_t'\beta - \left(Z_t'\hat{\Pi} \right) \gamma \right) \\ = & \sum_{t=1}^{T} \begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix} \left(\epsilon_t + v_t'\gamma - Z_t' \left(\hat{\Pi} - \Pi \right) \gamma \right) = \sum_{t=1}^{T} \begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix} \left(\epsilon_t + v_t'\gamma - Z_t' \left(Z'Z \right)^{-1} Z'v\gamma \right) \\ = & \sum_{t=1}^{T} \begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix} \left(\epsilon_t + v_t'\gamma - v_t'\gamma \right) = \sum_{t=1}^{T} \begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix} \epsilon_t. \end{aligned}$$

Since ϵ_t can be consistently estimated by $y_t - x_t'\hat{\beta} - Y_t'\hat{\gamma}$, stata implements a heteroscedasticity robust estimate of Ω by

$$\sum_{t=1}^{T} \begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix} \left(x_t, \Pi' Z_t \right) \left(y_t - x_t'\hat{\beta} - Y_t'\hat{\gamma} \right)^2.$$

Now let's see why the same inference procedure is valid for a binary interaction linear probability model where y_t depends on the expected value of Y_t given the information set of y_t in Z_t instead of on the realized value of Y_t. Let Z_t be the information set of a player in game t who makes a choice of y_t. For simplicity let y_t be binary. The incomplete information static game model we consider can in general be written as

$$P\left(y_t = 1|Z_t\right) = E\left(y_t|Z_t\right) = \Lambda\left(x_t'\beta + E\left(Y_t|Z_t\right)'\gamma\right),$$

where Y_t can be the number of competitors, in which case $E\left(Y_t|Z_t\right)$ is the expected number of competitors. If there is only one competitor, $E\left(Y_t|Z_t\right)$ is the probability that the competitor enters the market given the information set Z_t available to the player. $\Lambda\left(\cdot\right)$ can be the logit or probit function, or even the identity function that gives rise to the linear probability model, as we consider here.

Further, let's assume for now that $E\left(Y_t|Z_t\right)$ is correctly specified as a linear function of Z_t (or additional functions of Z_t). Later we will show that this linearity assumption, together with the linear probability model assumption, are both innocuous when we consider a nonparametric version of this model using the theory of sieve approximation.

In additional defense of the linear probability model, even in standard binary choice models, researchers rarely perform model selection tests. It is not even clear that either logit or probit are more correctly specified than the linear probability model. In addition, in the exogenuous case, the linear probability model has the interpretation of best linear predictor even without any assumption of correct model specification. Such interpretations are not available for logit or probit models.

Therefore let $E\left(Y_t|Z_t\right) = Z_t'\Pi$, in other words, $Y_t = Z_t'\Pi + v_t$ such that $E\left(v_t|Z_t\right) = 0$ but v_t can be heteroscedastic conditional on Z_t, which is necessarily the case when Y_t is either binary or discrete. Furthermore, if we assume the linear probability model $\Lambda\left(x\right) = x$, then

$$\begin{aligned} y_t =& x_t'\beta + E\left(Y_t|Z_t\right)'\gamma + \epsilon_t \\ =& x_t'\beta + \left(Z_t'\Pi\right)'\gamma + \epsilon_t \end{aligned}$$

where $E\left(\epsilon_t|Z_t\right) = 0$ but ϵ_t is necessarily heteroscedastic when y_t is binary or discrete.

An estimate for $\hat{\beta}, \hat{\gamma}$ can be obtained using exactly the same stata two stage least square procedure. The next question is whether the stata heteroscedasticity robust standard errors are also correct for this incomplete information game model. The answer turns out to be yes. The normal equation that is used to solve for $\hat{\beta}, \hat{\gamma}$ is the same as (21), and $\hat{A}$ is the same as before in (22). Ω again is the approximate variance covariance matrix of

$$\sum_{t=1}^{T}\begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix}\left(y_t - x_t'\beta - \left(Z_t'\hat{\Pi}\right)\gamma\right)$$

which can be now written as

$$\sum_{t=1}^{T}\begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix}\left(\epsilon_t - Z_t'\left(\hat{\Pi} - \Pi\right)\gamma\right) = \sum_{t=1}^{T}\begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix}\left(\epsilon_t - Z_t'\left(Z'Z\right)^{-1}Z'v\gamma\right)$$

$$=\sum_{t=1}^{T}\begin{pmatrix} x_t \\ Z_t'\Pi \end{pmatrix}\left(\epsilon_t - v_t'\gamma\right).$$

Now note that the $y_t - x_t'\hat{\beta} - Y_t'\hat{\gamma}$ will consistently estimate $\epsilon_t - v_t'\gamma$ instead of ϵ_t, because of the incomplete information version assumption. Therefore the same calculation performed in stata for heteroscedasticity robust two stage least square gives rise to the correct influence in the incomplete information version of the model, and the stata robust standard errors turn out to be the correct one!

The discussion above depends on several crucial assumptions. The first one is the linear probability model, so that ϵ_t appears linearly in the model. The second one is that $E\left(Y_t|Z_t\right)$ is correctly specified as a linear function of Z_t. The third one is that utility index is linear in both x_t and $E\left(Y_t|Z_t\right)$.

The first assumption, the linear probability model, can be considered innocuous if one is interested in a nonparametric model where $E\left(y_t|Z_t\right)$ is a nonparametric function of x_t but is a linear index of $E\left(Y_t|Z_t\right)$:

$$P\left(y_t = 1|Z_t\right) = g\left(x_t\right) + E\left(Y_t|Z_t\right)'\gamma,$$

and one is interested in estimating $g\left(\cdot\right)$ nonparametrically and the parameter γ. This is because if we are willing to let the number of terms in x_t to increasing as a function of the sample size (without loss of generality we may interchangably use x_t to denote x_t itself or functions of x_t), under suitable conditions $g\left(x_t\right)$ can be approximated arbitrary well as a linear index of x_t and basis functions created from x_t. Under these regularity conditions which ensure that the bias term is small enough compared to the variances, the conventional two stage least square will not only give consistent estimates of $g\left(\cdot\right)$ and γ but will also give the correct standard errors of both $\hat{g}(\cdot)$ and $\hat{\gamma}$. The regularity conditions needed for this validity are implied by the regularity conditions we gave in the main text of this paper. Interestingly, as we can easily show, while $g(x_t)$ can only be estimated with a nonparametric rate, γ can be estimated at the parametric $\sqrt{T}$ rate, this makes the above semilinear model semiparametric.

The second implied assumption, that $E(Y_t|Z_t)$ is a linear index of Z_t, is also innocuous, because any sufficiently regular nonparametric function $E(Y_t|Z_t)$ can be approximated arbitrarily well with the correct standard errors with basis functions of Z_t under the regularity conditions we stated in the main text. In principle, $E(Y_t|Z_t)$ is defined in equilibrium as a functional of all the $g(x_t)$'s and the distributions of ϵ_t for all the players in the model. However, as pointed out in the vast recent literature of estimating static and dynamic incomplete information games, the variable $E(Y_t|Z_t)$ can be nonparametrically identified from the data as long as only one equilibrium (out of possibly many) is being played in the data set. Existence and uniqeness of the equilibrium expectations $E(Y_t|Z_t)$ are also easy to demonstrate because of the linearity assumption (only needed linearity in $E(Y_t|Z_t)$ and the additive separability of $g(x_t)$ from $E(Y_t|Z_t)$). Of course the precise conditions for existence and uniqueness depend on how Y_t is related to other players' y_t's. In the special that Y_t is a $n-1$ vector of other players actions y_t, where n is the total number of players in each market or time period, the necessary and sufficient condition for existence and uniqueness of equilibrium can be stated as the square matrix $I_n - \Gamma_n$ being nonsingular. Here Γ_n is a $n \times n$ matrix where each row corresponds to the response equation for each player and each column corresponds to the individual y_t for each player. The diagonal elements of Γ_n are all zeros and the i,jth component of Γ_n, for $i \neq j$, represents how the action of player i depends on the action of player j. Note that in the semi-linear specification the equilibrium condition does not depend on $g(x_t)$ at all.

The model is akin to the social interaction model of Manski (1993). In contrast to the nonidentification results of Manski (1993), we achieve identification by having some variables in Z_t that are excluded from x_t. All that is needed for identification is linear independence of $E(Y_t|Z_t)$ from x_t.

Precisely, identification requires that

$$Var\left(E\left(Y_t|Z_t\right)|x_t\right) > 0$$

for a set of x_t with positive probability. When Y_t is (in general) a vector, this notation is understood to require that the conditional variance matrix of $E\left(Y_t|Z_t\right)$ given x_t is strictly positive definite for a range of x_t. If the linear parametric model is correctly specified, then this identification will be reduced to the usual rank condition for simultaneous equation models. To remind one of this condition, partition Π into $\left[\Pi^*, \bar{\Pi}\right]$ where $\bar{\Pi}$ corresponds to the variables included in x_t and Π^* corresponds to the variables in Z_t that are excluded from x_t, which we will denote Z_t^*. The rank condition requires that the variance-covariance matrix of Z_t^* is strictly positive definite, and that Π^* has full column rank, i.e. has rank equal to the number of variables in Y_t.

The third assumption, that $E\left(y_t|Z_t\right)$ is a linear index in $E\left(Y_t|Z_t\right)$, contains two restrictions, the first being that there is no nonlinear terms of $E\left(Y_t|Z_t\right)$ that enter $E\left(y_t|Z_t\right)$ and the second being that there is no interactions between $g\left(x_t\right)$ and $E\left(Y_t|Z_t\right)$. Both of these restrictions can be relaxed for the purpose of obtaining a consistent estimator of $E\left(y_t|Z_t\right)$ as a function of $g\left(x_t\right)$ and $E\left(Y_t|Z_t\right)$, but one can no longer rely on stata 2SLS to produce the correct standard errors. Correct standard errors are still easy to compute but one needs to do some work beyond the stata 2SLS routine.

To illustrate the issue with consistency consider the more general model

$$P\left(y_t = 1|Z_t\right) = g\left(x_t, E\left(Y_t|Z_t\right)\right).$$

Compared to a pure nonparametric model of $p\left(y_t = 1|Z_t\right)$, this "structural" model implies a dimension reduction device, when there are more than one variables in Z_t that are excluded from x_t. This dimension reduction does not come for free. The cost of this device is of course that one needs to estimate $E\left(Y_t|Z_t\right)$ in the first stage, which will also introduce potentially large sampling errors. An additional importance difficulty of this general model, as compared to the model that is linear in $E\left(Y_t|Z_t\right)$, is that existence and uniqueness of the equilibrium solutions $E\left(Y_t|Z_t\right)$ are very difficult to establish in the general nonlinear setting.

Assuming that existence and uniqueness of equilibrium can be resolved, it is easy to estimate the responce function $g(\cdot,\cdot)$ in two steps. In the first step, a nonparametric regression using either kernel, local polynomial, or sieve of Y_t on Z_t can be used to obtain a consistent estimate $\hat{E}(Y_t|Z_t)$. In the second step, another nonparametric regression of y_t on x_t and $E(Y_t|Z_t)$ can be performed to provide a consistent estimate $\hat{g}\left(x_t, \hat{E}(Y_t|Z_t)\right)$.

There are two possibilities with standard errors in this approach. In the first case, if the nonparametric smoothing parameters (e.g. the number of terms in the sieve expansion, or the bandwidth in the kernel estimation, or the degree and neighborhood size in local polynomial estimation) are chosen such that the first stage estimation of $\hat{E}(Y_t|Z_t)$ is undersmoothed relative to the second stage estimation of $\hat{g}\left(x, \hat{E}(Y|Z)\right)$, and assuming the bias terms are smaller than the sample errors in both stages, the standard heteroscedasticity robust standard errors from the second stage ordinary least square will be asymptotically valid. In the second case, if the sample errors from the first stage estimation dominate or are equal in magnitude to the sample errors from the second stage estimation, then one needs to write down the approximate sample errors manually to compute it, or to rely on some resampling scheme for both stages.

To further illustrate the issues with standard errors in a model that is nonlinear in $E(Y_t|Z_t)$, consider a simple quadratic specification. Consider two models, the first being a complete information version and the second being an incomplete information version, in both cases instruments Z_t are available:

$$y_t = x_t'\beta + Y_t'\gamma_1 + Y_t^2\gamma_2 + \epsilon_t \tag{23}$$

$$y_t = x_t'\beta + E(Y_t|Z_t)'\gamma_1 + E(Y_t|Z_t)^2\gamma_2 + \epsilon_t \tag{24}$$

In both cases, assume that $E(Y_t|Z_t) = Z_t'\Pi$. Consider two estimation procedures:
Procedure 1: First regress Y_t on Z_t and obtain the fitted values $\hat{Y}_t$, next regress y_t on x_t, $\hat{Y}_t$ and $\hat{Y}_t^2$.
Procedure 2: Regress Y_t on Z_t to obtain fitted values $\hat{Y}_t$, also regress Y_t^2 on Z_t to obtain fittevd values which we will call $\widehat{Y_t^2}$. In the second step, regress y_t on x_t, $\hat{Y}_t$ and $\widehat{Y_t^2}$.

In model (23), procedure 2 is consistent and will also give the correct standard errors as long as it is performed using the 2SLS routine in Stata. But procedure 1 is invalid (inconsistent). In fact, procedure 2 is consistent as long as $Cov\left(Z_t, \epsilon_t\right) = 0$, regardless of whether $E\left(Y_t|Z_t\right)$ or $E\left(Y_t^2|Z_t\right)$ are linear functions of Z_t at all. Note that this is another difference between the complete information (Y_t) version and the incomplete information $(E\left(Y_t|Z_t\right))$ version of the simultaneous equation model. In the complete information (Y_t) version, only the assumption of $E\left(Z_t, \epsilon_t\right) = 0$ is needed for the validity of 2SLS, for both consistency and correct standard errors. On the other hand, in the incomplete information $E\left(Y_t|Z_t\right)$ version, the linearity of $E\left(Y_t|Z_t\right)$ in Z_t is crucial even for the consistency of "2SLS" both in models that are linear in $E\left(Y_t|Z_t\right)$ (in which case 2SLS std errors will be correct) and in models that are nonlinear in $E\left(Y_t|Z_t\right)$ (in which case std errors need a bit more work).

In contrast, in model (24), procedure 2 is invalid (inconsistent). Procedure 1, instead, is valid (consistent), but one needs to correct the standard errors in the second stage by taking into account the sampling errors from the first stage. This is easy to do, but one can not just use OLS or 2SLS in stata to obtain the correct standard errors, unlike in the linear model where the quadratic term of $E\left(Y_t|Z_t\right)^2$ is absent.

References

Aguirregabiria, V., and P. Mira (2002): "Sequential simulation-based estimation of dynamic discrete games," Technical Report, Boston University.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.

Bajari, P., L. Benkard, and J. Levin (2004): "Estimating Dynamic Models of Imperfect Competition," working paper.

Berry, S., A. Pakes, and M. Ostrovsky (2003): "Simple estimators for the parameters of dynamic games (with entry/exit examples)," Technical Report, Harvard University.

Hotz, J., and R. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 60, 497–529.

Manski, C. (1993): “Identification of Endogenous Social Effects: The Reflection Problem,” *Review of Economic Studies*, 60, 531–542.

Newey, W. (1994a): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–82.

Newey, W. (1994b): “Convergence Rates for Series Estimators,” in *Statistical Methods of Economics and Quantitative Economics: Essays in Honore of C. R. Rao.*

Pesendorfer, M., and P. Schmidt-Dengler (2003): “Identification and Estimation of Dynamic Games,” NBER working paper No. w9726.