

Supplementary Appendix to “Sequential Estimation of Structural Models with a Fixed Point Constraint”

Hiroyuki Kasahara
Department of Economics
University of British Columbia
and University of Western Ontario
hkasahar@uwo.ca

Katsumi Shimotsu
Department of Economics
Hitotsubashi University
and Queen’s University
shimotsu@econ.hit-u.ac.jp

November 3, 2009

This supplementary appendix contains the following details omitted from the main paper due to space constraints: (A) numerical implementation of the sequential algorithm based on the RPM, (B) approximated fixed point (AFXP) algorithm, (C) the sequential GMM estimator, (D) the convergence properties of the NPL algorithm for models with unobserved heterogeneity, and (E) additional Monte Carlo results.

A Numerical Implementation of the Sequential Algorithm based on the RPM in Section 4.2

Implementing the sequential algorithm based on the RPM in Section 4.2 requires evaluating $(I - \Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})\nabla_{P'}\Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})\Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}))^{-1}$ as well as computing an orthonormal basis $Z(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ from the eigenvectors of $\nabla_{P'}\Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ for $j = 1, \dots, k$. This is potentially costly when the analytical expression of $\nabla_{P'}\Psi(\theta, P)$ is not available.

In this section, we discuss how to reduce the computational cost of implementing the RPM algorithm by updating $(I - \Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})\nabla_{P'}\Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})\Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}))^{-1}$ and $Z(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ without explicitly computing $\nabla_{P'}\Psi(\theta, P)$ in each iteration.

First, we provide theoretical underpinning. The following Corollary shows that, if an alternate preliminary consistent estimator (θ^*, P^*) is used in forming $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$, it only affects the remainder terms in Proposition 7. Therefore, if we use a root- n consistent (θ^*, P^*) to evaluate $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ and keep these estimates unchanged throughout iterations, the resulting sequence of estimators is only $O_p(n^{-1})$ away from the corresponding estimators generated by the approximate RPM algorithm.

Corollary 1 *Suppose Assumption 3 holds. Let (θ^*, P^*) be a consistent estimator of (θ^0, P^0) , and suppose we obtain $(\tilde{\theta}_j, \tilde{P}_j)$ by the approximate RPM algorithm with $\Pi(\theta^*, P^*)$ and $\nabla_{P'}\Psi(\theta^*, P^*)$ in place of $\Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ and $\nabla_{P'}\Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$. Then, there exists $c > 0$ such that, if $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}(c)$, then $\tilde{\theta}_j - \hat{\theta}_{RPM} = O_p(\|\tilde{P}_{j-1} - \hat{P}_{RPM}\| + n^{-1/2}\|\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}\| + \|\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}\|^2 + r_{nj}^*)$ and $\tilde{P}_j - \hat{P}_{RPM} = M_{\Gamma_\theta}\Gamma_P(\tilde{P}_{j-1} - \hat{P}_{RPM}) + O_p(n^{-1/2}\|\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}\| + \|\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}\|^2 + n^{-1/2}\|\tilde{P}_{j-1} - \hat{P}_{RPM}\| + \|\tilde{P}_{j-1} - \hat{P}_{RPM}\|^2) + r_{nj}^*$ uniformly in $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}(c)$, where $r_{nj}^* = O_p(n^{-1/2}\|\theta^* - \hat{\theta}_{RPM}\| + \|\theta^* - \hat{\theta}_{RPM}\|^2 + n^{-1/2}\|P^* - \hat{P}_{RPM}\| + \|P^* - \hat{P}_{RPM}\|^2)$.*

Proof of Corollary 1 The proof closely follows the proof of Proposition 7. Define $\Gamma(\theta, P, \eta, Q) \equiv \Psi(\theta, P) + [(I - \Pi(\eta, Q))\nabla_{P'}\Psi(\eta, Q)\Pi(\eta, Q)]^{-1} - I]\Pi(\eta, Q)(\Psi(\theta, P) - P)$, so that the objective function in Step 1 is written as $\bar{\gamma}(\theta, \tilde{P}_{j-1}, \theta^*, P^*) = n^{-1} \sum_{i=1}^n \ln \Gamma(\theta, \tilde{P}_{j-1}, \theta^*, P^*)(a_i | x_i)$. For $\epsilon_1 > 0$, define a neighborhood $\mathcal{N}_4(\epsilon_1) = \{(\theta, P, \eta, Q) : \max\{\|\theta - \theta^0\|, \|P - P^0\|, \|\eta - \theta^0\|, \|Q - P^0\|\} < \epsilon_1\}$. Then, for any $\epsilon_1 > 0$, we have $(\tilde{\theta}_j, \tilde{P}_{j-1}, \theta^*, P^*) \in \mathcal{N}_4(\epsilon_1)$ wpa1 if c is chosen sufficiently small by the same argument as the proof of Proposition 7.

Assuming $(\tilde{\theta}_j, \tilde{P}_{j-1}, \theta^*, P^*) \in \mathcal{N}_4(\epsilon_1)$, the stated result follows from starting from the first order condition $\nabla_{\theta'}\bar{\gamma}(\tilde{\theta}_j, \tilde{P}_{j-1}, \theta^*, P^*) = 0$, expanding it around $(\hat{\theta}, \hat{P}_{j-1}, \theta^*, P^*)$, and following the proof of Proposition 7 using $\nabla_{Q'}\Gamma(\theta^0, P^0, \theta^0, P^0) = 0$. \square

Using Corollary 1, in the following we discuss how to reduce the computational cost of implementing the RPM algorithm by updating $(I - \Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}))\nabla_{P'}\Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})\Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})^{-1}$ and $Z(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ without explicitly computing $\nabla_{P'}\Psi(\theta, P)$ in each iteration. Denote $\tilde{\Pi}_{j-1} \equiv \Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, $\tilde{Z}_{j-1} = Z(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, and $\tilde{\Psi}_{P,j-1} = \nabla_{P'}\Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$.

First, using $\tilde{\Pi}_{j-1} = \tilde{Z}_{j-1}(\tilde{Z}_{j-1})'$ and $(\tilde{Z}_{j-1})'\tilde{Z}_{j-1} = I$, we may verify that

$$(I - \tilde{\Pi}_{j-1}\tilde{\Psi}_{P,j-1}\tilde{\Pi}_{j-1})^{-1}\tilde{\Pi}_{j-1} = \tilde{Z}_{j-1}(I - (\tilde{Z}_{j-1})'\tilde{\Psi}_{P,j-1}\tilde{Z}_{j-1})^{-1}(\tilde{Z}_{j-1})'.$$

Let $\tilde{Z}_{j-1} = [\tilde{z}_{j-1}^1, \dots, \tilde{z}_{j-1}^m]$ and $\xi > 0$. The i th column of $\tilde{\Psi}_{P,j-1}\tilde{Z}_{j-1}$ can be approximated by $\tilde{\Psi}_{P,j-1}\tilde{z}_{j-1}^i \approx (1/\xi)[\Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1} + \xi\tilde{z}_{j-1}^i) - \Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})]$, which requires $(m+1)$ function evaluations of $\Psi(\theta, P)$. Further, evaluating $(I - \tilde{\Pi}_{j-1}\tilde{\Psi}_{P,j-1}\tilde{\Pi}_{j-1})^{-1}$ only requires the inversion of the $m \times m$ matrix $I - (\tilde{Z}_{j-1})'\tilde{\Psi}_{P,j-1}\tilde{Z}_{j-1}$ instead of an inversion of an $L \times L$ matrix. Thus, when m is small, numerically evaluating $(I - \tilde{\Pi}_{j-1}\tilde{\Psi}_{P,j-1}\tilde{\Pi}_{j-1})^{-1}$ is not computationally difficult.

Second, it is possible to use $\tilde{\Psi}_{P,j}\tilde{Z}_{j-1}$ to update an estimate of the orthogonal basis Z . Namely, given a preliminary estimate \tilde{Z}_{j-1} , we may obtain \tilde{Z}_j by performing one step of an orthogonal power iteration (see Shroff and Keller, 1993, p. 1107 and Golub and Van Loan, 1996) by computing $\tilde{Z}_j = \text{orth}(\tilde{\Psi}_{P,j}\tilde{Z}_{j-1})$, where “ $\text{orth}(B)$ ” denotes an orthonormal basis for the columns of B computed by Gram-Schmidt orthogonalization.

Our numerical implementation of the RPM sequential algorithm is summarized as follows.

Step 0 (Initialization): (a) Find the eigenvalues of $\tilde{\Psi}_{P,0} \equiv \nabla_{P'}\Psi(\tilde{P}_0, \tilde{\theta}_0)$ for which the mod-

ulus is larger than δ . Let $\{\tilde{\lambda}_{0,1}, \dots, \tilde{\lambda}_{0,m}\}$ denote them.¹ (b) Find the eigenvectors of $\tilde{\Psi}_{P,0}$ associated with $\tilde{\lambda}_{0,1}, \dots, \tilde{\lambda}_{0,m}$. (c) Using Gram-Schmidt orthogonalization, compute an orthonormal basis of the space spanned by these eigenvectors. Let $\{\tilde{z}_0^1, \dots, \tilde{z}_0^m\}$ denote the basis. (d) Compute $\tilde{Z}_0(I - \tilde{Z}'_0 \tilde{\Psi}_{P,0} \tilde{Z}_0)^{-1} \tilde{Z}'_0$ and $\tilde{\Pi}_0 = \tilde{Z}_0 \tilde{Z}'_0$, where $\tilde{Z}_0 = [\tilde{z}_0^1, \dots, \tilde{z}_0^m]$.

Step 1 (Update θ): Given $\tilde{Z}_{j-1}(I - \tilde{Z}'_{j-1} \tilde{\Psi}_{P,j-1} \tilde{Z}_{j-1})^{-1} \tilde{Z}'_{j-1}$ and $\tilde{\Pi}_{j-1} = \tilde{Z}_{j-1}(\tilde{Z}_{j-1})'$, update θ by $\tilde{\theta}_j = \arg \max_{\theta \in \Theta_j} n^{-1} \sum_{i=1}^n \ln \Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}, \tilde{Z}_{j-1})(a_i | x_i)$, where $\Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}, \tilde{Z}_{j-1}) = \tilde{\Pi}_{j-1} \tilde{P}_{j-1} + \tilde{Z}_{j-1}(I - \tilde{Z}'_{j-1} \tilde{\Psi}_{P,j-1} \tilde{Z}_{j-1})^{-1} \tilde{Z}'_{j-1} (\Psi(\theta, \tilde{P}_{j-1}) - \tilde{P}_{j-1}) + (I - \tilde{\Pi}_{j-1}) \Psi(\theta, \tilde{P}_{j-1})$ with $\tilde{\Psi}_{P,j-1} \equiv \nabla_{P'} \Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$.

Step 2 (Update P): Given $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}, \tilde{Z}_{j-1})$, update P by $\tilde{P}_j = \Gamma(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}, \tilde{Z}_{j-1})$.

Step 3 (Update Z): (a) Update the orthonormal basis Z by $\tilde{Z}_j = \text{orth}(\tilde{\Psi}_{P,j} \tilde{Z}_{j-1})$, where the i -th column of $\tilde{\Psi}_{P,j} \tilde{Z}_{j-1}$ is computed by $\tilde{\Psi}_{P,j} \tilde{z}_{j-1}^i \approx (1/\xi)[\Psi(\tilde{\theta}_j, \tilde{P}_j + \xi \tilde{z}_{j-1}^i) - \Psi(\tilde{\theta}_j, \tilde{P}_j)]$ for small $\xi > 0$ with $\tilde{Z}_{j-1} = [\tilde{z}_{j-1}^1, \dots, \tilde{z}_{j-1}^m]$. (b) Compute $\tilde{\Pi}_j = \tilde{Z}_j(\tilde{Z}_j)'$ and $\tilde{Z}_j(I - \tilde{Z}'_j \tilde{\Psi}_{P,j} \tilde{Z}_j)^{-1} \tilde{Z}'_j$, where the i -th row of $\tilde{\Psi}_{P,j} \tilde{Z}_j$ is given by $\tilde{\Psi}_{P,j} \tilde{z}_j^i \approx (1/\xi)[\Psi(\tilde{\theta}_j, \tilde{P}_j + \xi \tilde{z}_j^i) - \Psi(\tilde{\theta}_j, \tilde{P}_j)]$. (c) Every J iterations, update the orthonormal basis Z using the algorithm of Step 0, where $(\tilde{\theta}_0, \tilde{P}_0)$ is replaced with $(\tilde{\theta}_j, \tilde{P}_j)$.

Step 4: Iterate Steps 1-3 k times.

When an initial estimate is not precise, the dominant eigenspace of $\tilde{\Psi}_{P,j}$ will change as iterations proceed. In Step 3(a), the orthonormal basis is updated to maintain the accuracy of the basis without changing the size of the orthonormal basis. If an initial estimate of the size of the orthonormal basis is smaller than the true size, however, the estimated subspace $\tilde{\mathbb{P}} = \tilde{\Pi} \mathbb{R}^L$ may not contain all the bases for which eigenvalues are outside the unit circle. In such a case, the algorithm may not converge. To safeguard against such a possibility, the basis size is updated every J iterations in Step 3(c). In our Monte Carlo experiments, we chose $J = 10$. Corollary 1 implies that this modified algorithm will converge.

B Approximate fixed point algorithm

We apply the idea of the approximate q -NPL algorithm to approximate the MLE. Define the fixed point of $\Psi(\theta, P)$ that maximizes the likelihood function among all the fixed points of $\Psi(\theta, P)$ as $\hat{P}(\theta) \equiv \arg \max_{P \in \mathcal{M}_\theta} n^{-1} \sum_{i=1}^n \ln P(a_i | x_i)$, where $\mathcal{M}_\theta \equiv \{P \in B_P : P = \Psi(\theta, P)\}$ as defined in the main text. Define $P(\theta)$ as the population counterpart of $\hat{P}(\theta)$, i.e., $P(\theta) \equiv \arg \max_{P \in \mathcal{M}_\theta} E \ln P(a_i | x_i)$. Using $\hat{P}(\theta)$ and $P(\theta)$, we may write the objective function of the MLE and its limit as $Q_n(\theta) = n^{-1} \sum_{i=1}^n \ln \hat{P}(\theta)(a_i | x_i)$ and $Q(\theta) = E \ln P(\theta)(a_i | x_i)$. If $P(\theta)$ is

¹Computing the m dominant eigenvalues of $\tilde{\Psi}_{P,0}$ is potentially costly. We follow the numerical procedure based on the power iteration method as discussed in section 4.1 of SK.

unique and continuously differentiable in a neighborhood of θ^0 , we can approximate $P(\theta)$ as $P(\theta) = P^0 + (I - \nabla_{P'}\Psi(\theta^0, P^0))^{-1}\nabla_{\theta'}\Psi(\theta^0, P^0)(\theta - \theta^0) + O(\|\theta - \theta^0\|^2)$ in a neighborhood of θ^0 , using the relations $\nabla_{\theta'}P(\theta) = (I - \nabla_{P'}\Psi(\theta, P(\theta)))^{-1}\nabla_{\theta'}\Psi(\theta, P(\theta))$ and $P(\theta^0) = P^0$. Therefore, if we have a consistent estimate of θ^0 and P^0 , we may approximate $P(\theta)$ by a linear function of θ with the mappings $\nabla_{P'}\Psi(\theta, P)$ and $\nabla_{\theta'}\Psi(\theta, P)$.

We consider an estimation algorithm, called the *Approximate Fixed Point (AFXP) algorithm*, based on the following objective function: $Q_n(\theta, P, \eta) \equiv n^{-1} \sum_{i=1}^n \ln \Phi(\theta, P, \eta)(a_i|x_i)$, where

$$\Phi(\theta, P, \eta) \equiv P + (I - \nabla_{P'}\Psi(\eta, P))^{-1}\nabla_{\theta'}\Psi(\eta, P)(\theta - \eta).$$

Let $\tilde{\theta}_0$ be an initial estimator of θ^0 , such as the PML estimator. The AFXP algorithm iterates the following steps until $j = k$:

Step 1: Given $\tilde{\theta}_{j-1}$, update P by solving the fixed point: $\tilde{P}_j = \hat{P}(\tilde{\theta}_{j-1})$.

Step 2: Given $(\tilde{P}_j, \tilde{\theta}_{j-1})$, update θ by $\tilde{\theta}_j = \arg \max_{\theta \in \Theta_j} Q_n(\theta, \tilde{P}_j, \tilde{\theta}_{j-1})$, where $\Theta_j \equiv \{\theta \in \Theta : \Phi(\theta, \tilde{P}_j, \tilde{\theta}_{j-1})(a|x) \in [\xi, 1 - \xi] \text{ for all } (a, x) \in A \times X\}$ for an arbitrary small $\xi > 0$.

First, we establish the asymptotic normality of the MLE.

Assumption 5 (a) *The observations $\{a_i, x_i : i = 1, \dots, n\}$ are independent and identically distributed, and $dF(x) > 0$ for any $x \in X$, where $F(x)$ is the distribution function of x_i .* (b) *$\Psi(\theta, P)(a|x) > 0$ for any $(a, x) \in A \times X$ and any $(\theta, P) \in \Theta \times B_P$.* (c) *$\Psi(\theta, P)$ is four times continuously differentiable.* (d) *Θ is compact and, for any $\theta \in \Theta$, \mathcal{M}_θ is compact.* (e) *There is a unique $\theta^0 \in \text{int}(\Theta)$ such that $P(\theta^0) = P^0$.* (f) *For any $\theta \neq \theta^0$, $P(\theta) \neq P^0$.* (g) *$P(\theta)$ is unique for any θ and continuous in θ .* (h) *$I - \Psi_P$ is nonsingular and $\text{rank}(\Psi_\theta) = K$.*

The uniqueness of $P(\theta)$ in Assumption 5(g) is potentially a strong assumption. This assumption holds if no equilibrium choice probabilities are observationally equivalent to $P(\theta)$. As discussed in the proof of Proposition 10, Assumption 5(h) implies that $P(\theta)$ is unique for any θ in a neighborhood of θ^0 . Therefore, if $\tilde{\theta}_{j-1}$ is consistent, $\hat{P}(\tilde{\theta}_{j-1})$ is also uniquely determined for a sufficiently large sample size.

The following proposition establishes the asymptotic normality of the MLE.

Proposition 10 *Suppose that Assumption 5 holds. Then $n^{1/2}(\hat{\theta}_{MLE} - \theta^0) \rightarrow_d N(0, (\mathcal{I}^0)^{-1})$, where $\mathcal{I}^0 \equiv E[\nabla_\theta \ln P(\theta^0)(a_i|x_i)\nabla_{\theta'} \ln P(\theta^0)(a_i|x_i)]$.*

Having established the asymptotic normality of the MLE, we derive the convergence properties of the estimators generated by the AFXP algorithm. For $\epsilon > 0$, define a neighborhood of θ^0 by $\mathcal{N}_\theta(\epsilon) = \{\theta : \|\theta - \theta^0\| < \epsilon\}$.

Proposition 11 *Suppose that Assumption 5 holds and we obtain $(\tilde{\theta}_j, \tilde{P}_j)$ by the AFXP algorithm. Then, there exists $c > 0$ such that $\tilde{P}_j - \hat{P}_{MLE} = O_p(\|\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}\|)$ and $\tilde{\theta}_j - \hat{\theta}_{MLE} = O_p(n^{-1/2}\|\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}\|) + O_p(\|\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}\|^2)$ uniformly in $\tilde{\theta}_{j-1} \in \mathcal{N}_\theta(c)$.*

Thus, the estimator generated by the AFXP algorithm is first-order equivalent to the MLE for all $k \geq 1$ if started from a root- n consistent estimate. This algorithm can be used to obtain the MLE because, upon convergence, its limit is identical to the MLE.

Implementing Step 1 of the AFXP algorithm may be impractical when it is computationally infeasible to find all the fixed points. In such cases, we may replace the solution to the fixed point in Step 1 with its consistent estimator. Define the q -AFXP algorithm by the same sequential algorithm as the AFXP algorithm except that, starting from an initial consistent estimate $(\tilde{\theta}_0, \tilde{P}_0)$, Step 1 updates P by $\tilde{P}_j = \Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ or $\tilde{P}_j = \Gamma^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$.

The following proposition establishes the convergence properties of the estimators generated by the q -AFXP algorithm. Define a $K \times L$ matrix \mathcal{J} as $\mathcal{J} \equiv E[\nabla_{\theta} \ln P(\theta^0)(a_i|x_i)I(a_i|x_i)/P^0(a_i|x_i)]$, where $I(a_i|x_i)$ is the row of an $L \times L$ identity matrix that corresponds to $(a_i|x_i)$. For $\epsilon > 0$, define a neighborhood of (θ^0, P^0) by $\mathcal{N}(\epsilon) = \{P : \max\{\|\theta - \theta^0\|, \|P - P^0\|\} < \epsilon\}$.

Proposition 12 *Suppose that Assumption 5 holds. Suppose we obtain $\tilde{\theta}_j$ by the q -AFXP algorithm with $\tilde{P}_j = \Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$. Then, there exists $c > 0$ such that*

$$\tilde{P}_j - \hat{P}_{MLE} = \Lambda_P^q(\tilde{P}_{j-1} - \hat{P}_{MLE}) + \Lambda_{\theta}^q(\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) + r_{nj}, \quad (22)$$

$$\tilde{\theta}_j - \hat{\theta}_{MLE} = (\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) - (\mathcal{I}^0)^{-1} \mathcal{J}(\tilde{P}_j - \hat{P}_{MLE}) + r_{nj}, \quad (23)$$

where r_{nj} denotes a remainder term satisfying $r_{nj} = O_p(n^{-1/2} \|\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}\| + \|\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}\|^2 + n^{-1/2} \|\tilde{P}_{j-1} - \hat{P}_{MLE}\| + \|\tilde{P}_{j-1} - \hat{P}_{MLE}\|^2)$ uniformly in $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}(c)$.

When \tilde{P}_j is obtained by $\tilde{P}_j = \Gamma^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ and $I - \Pi(\theta^0, P^0)\Psi_P\Pi(\theta^0, P^0)$ is nonsingular, the uniform bounds (22)-(23) hold in which Λ_P^q and Λ_{θ}^q are replaced with Γ_P^q and Γ_{θ}^q .

Ignoring r_{nj} , arranging the two updating relations into a system of equations, solving for $\tilde{P}_j - \hat{P}_{MLE}$ and $\tilde{\theta}_j - \hat{\theta}_{MLE}$, and using $\Lambda_P^q = (\Lambda_P)^q$, $\Lambda_{\theta}^q = (I + \Lambda_P + \dots + (\Lambda_P)^{q-1})\Lambda_{\theta} = (I - (\Lambda_P)^q)(I - \Lambda_P)^{-1}\Lambda_{\theta} = (I - (\Lambda_P)^q)\nabla_{\theta'} P(\theta^0)$, and $\mathcal{J}\nabla_{\theta'} P(\theta^0) = \mathcal{I}^0$, we obtain

$$\begin{pmatrix} \tilde{P}_j - \hat{P}_{MLE} \\ \tilde{\theta}_j - \hat{\theta}_{MLE} \end{pmatrix} = Q \begin{pmatrix} \tilde{P}_{j-1} - \hat{P}_{MLE} \\ \tilde{\theta}_{j-1} - \hat{\theta}_{MLE} \end{pmatrix}, \text{ where } Q = \begin{pmatrix} (\Lambda_P)^q & \Lambda_{\theta}^q \\ -(\mathcal{I}^0)^{-1} \mathcal{J}(\Lambda_P)^q & (\mathcal{I}^0)^{-1} \mathcal{J}(\Lambda_P)^q \nabla_{\theta'} P(\theta^0) \end{pmatrix}.$$

Suppose $\rho(\Lambda_P) < 1$. Then, as q increases, $(\Lambda_P)^q$ approaches zero, and all the eigenvalues of Q approach zero. Therefore, all of the eigenvalues of Q are inside the unit circle for sufficiently large q , and iterating the q -AFXP algorithm converges to the MLE.

We note that, while consistency allows us to focus on a neighborhood of θ^0 , the Taylor approximation might be a poor approximation in finite samples and could affect the convergence of the AFXP algorithm.²

²We further examine the finite sample issue in our simulation by conducting the Monte Carlo experiment with a small sample of $n = 100$, and we find that the estimator generated by the q -AFXP algorithm after $k = 50$ iterations still performs better than the NPL estimator using Λ across different parameters. Thus, the small

B.1 Proof of propositions in Section B

Proof of Proposition 10 The consistency of the MLE follows from Theorem 2.1 of Newey and McFadden (1994), because (i) θ^0 uniquely maximizes $Q(\theta)$ because $\Pr_{P^0}(\{(a, x) : P(\theta)(a|x) \neq P^0(a|x)\}) > 0$ for any $\theta \neq \theta^0$; (ii) Θ is compact; (iii) $Q(\theta) = E \ln P(\theta)(a_i|x_i)$ is continuous because $P(\theta)$ is continuous; (iv) $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| = \sup_{\theta \in \Theta} |\sup_{P \in \mathcal{M}_\theta} n^{-1} \sum_{i=1}^n \ln P(a_i|x_i) - \sup_{P \in \mathcal{M}_\theta} E \ln P(a_i|x_i)| \rightarrow_p 0$.

For the asymptotic normality of the MLE, consider the objective function $Q_n^*(\theta) = n^{-1} \sum_{i=1}^n \ln P(\theta)(a_i|x_i)$, in which $\hat{P}(\theta)(a_i|x_i)$ in $Q_n(\theta)$ is replaced with $P(\theta)(a_i|x_i)$, and an estimator $\theta^* = \arg \max_{\theta \in \Theta} Q_n^*(\theta)$. First, we derive the asymptotic normality of θ^* using Theorem 3.3 of Newey and McFadden (1994), and then we establish the asymptotic normality of the MLE by showing $\Pr(\theta^* = \hat{\theta}_{MLE}) \rightarrow 1$. The asymptotic normality of θ^* follows from confirming that $P(\theta)(a|x)$ satisfies conditions (i)-(v) of Theorem 3.3 of Newey and McFadden (1994). Condition (i) holds from Assumption 5(e). For condition (ii), observe that a direct calculation gives $\nabla_{\theta'} P(\theta) = (I - \nabla_{P'} \Psi(\theta, P(\theta)))^{-1} \nabla_{\theta'} \Psi(\theta, P(\theta))$. Differentiating both sides by θ , we find that $P(\theta)$ is twice continuously differentiable in \mathcal{N}_{θ^0} since $I - \Psi_P$ is nonsingular. Condition (ii) then holds because $P(\theta) > 0$ from Assumption 5(c). Conditions (iii) and (v) hold because both $\sup_{\theta \in \mathcal{N}_{\theta^0}} \|\nabla_{\theta'} P(\theta)\|$ and $\sup_{\theta \in \mathcal{N}_{\theta^0}} \|\nabla_{\theta\theta'} P(\theta)\|$ are finite. For condition (iv), observe that $E[\nabla_{\theta} \ln P(\theta^0)(a_i|x_i) \nabla_{\theta'} \ln P(\theta^0)(a_i|x_i)] = \Psi'_\theta (I - \Psi'_P)^{-1} \Delta_P (I - \Psi_P)^{-1} \Psi_\theta$, where Δ_P is defined on page 6 of the main text and is nonsingular. Condition (iv) then holds because $\Psi'_\theta (I - \Psi'_P)^{-1} \Delta_P (I - \Psi_P)^{-1} \Psi_\theta$ is nonsingular from Assumption 5(h).

It remains to show $\Pr(\theta^* = \hat{\theta}_{MLE}) \rightarrow 1$. First, note that the fixed point of $\Psi(\theta, P)$ is unique in a neighborhood of (θ^0, P^0) from the implicit function theorem because the mapping $\nu(\theta, P) = P - \Psi(\theta, P)$ has a nonsingular Jacobian matrix with respect to P at (θ^0, P^0) from Assumption 5(h). Second, using $\sup_{(\theta, P) \in (\Theta \times B_P)} |n^{-1} \sum_{i=1}^n \ln \Psi(\theta, P)(a_i|x_i) - E \ln \Psi(\theta, P)(a_i|x_i)| \rightarrow_p 0$ twice gives $\sup_{\theta \in \Theta} |E \ln \Psi(\theta, \hat{P}(\theta))(a_i|x_i) - E \ln \Psi(\theta, P(\theta))(a_i|x_i)| \rightarrow_p 0$. Therefore, $\sup_{\theta \in \Theta} |E \ln \hat{P}(\theta)(a_i|x_i) - E \ln P(\theta)(a_i|x_i)| \rightarrow_p 0$. Since the fixed point of $\Psi(\theta, P)$ is unique for $(\theta, P) \in \mathcal{N}$, we have $\Pr(\hat{P}(\theta) = P(\theta)) \rightarrow 1$ for $\theta \in \mathcal{N}_\theta$, and $\Pr(\theta^* = \hat{\theta}_{MLE}) \rightarrow 1$ follow because θ^* and $\hat{\theta}_{MLE}$ are consistent and $\Pr(\theta^*, \hat{\theta}_{MLE} \in \mathcal{N}_\theta) \rightarrow 1$. \square

Proof of Proposition 11 We suppress the subscript MLE from $\hat{\theta}_{MLE}$ and \hat{P}_{MLE} . Define $Q_n^*(\theta, \eta) = n^{-1} \sum_{i=1}^n \ln[P(\eta) + \nabla_{\theta'} P(\eta)(\theta - \eta)](a_i|x_i)$, where $\nabla_{\theta'} P(\eta)$ denotes the derivative of $P(\theta)$ evaluated at η , and define $Q^*(\theta, \eta) = E \ln[P(\eta) + \nabla_{\theta'} P(\eta)(\theta - \eta)](a_i|x_i)$. For $\epsilon > 0$, define a neighborhood $\mathcal{N}_{\theta\eta}(\epsilon) = \{(\theta, \eta) : \max\{\|\theta - \theta^0\|, \|\eta - \theta^0\|\} < \epsilon\}$. Then, there exists $\epsilon_1 > 0$ such that (i) $\Pr(\hat{P}(\theta) = P(\theta)) \rightarrow 1$ for $(\theta, \eta) \in \mathcal{N}_{\theta\eta}(\epsilon_1)$ from the proof of Proposition 10, (ii) $\sup_{(\theta, \eta) \in \mathcal{N}_{\theta\eta}(\epsilon_1)} \|\nabla_{\theta\theta'} Q^*(\theta, \eta)^{-1}\| < \infty$ and $\sup_{(\theta, \eta) \in \mathcal{N}_{\theta\eta}(\epsilon_1)} \|\nabla^3 Q^*(\theta, \eta)\| < \infty$ because $\nabla_{\theta\theta'} Q^*(\theta^0, \theta^0) = -\mathcal{I}^0$, $\nabla_{\theta'} P_\theta = (I - \nabla_{P'} \Psi(\theta, P_\theta))^{-1} \nabla_{\theta'} \Psi(\theta, P_\theta)$ for any fixed point P_θ , $I - \Psi_P$

sample issue regarding the Taylor approximation does not appear to be a major issue, at least, in our simulation setup of section 4.

is nonsingular, and $\Psi(\theta, P)$ is four times continuously differentiable.

First, we assume $(\theta, \eta) \in \mathcal{N}_{\theta\eta}(\epsilon_1)$ and derive the stated representation of $\tilde{\theta}_j - \hat{\theta}$ and $\tilde{P}_j - \hat{P}$. $\tilde{P}_j - \hat{P} = O_p(\|\tilde{\theta}_{j-1} - \hat{\theta}\|)$ follows easily from a Taylor expansion. To show the bound of $\tilde{\theta}_j - \hat{\theta}$, note that $Q_n^*(\theta, \eta) = Q_n(\theta, \hat{P}(\eta), \eta)$ with probability approaching one (wpa1 henceforth) from (i) above. Therefore, $\tilde{\theta}_j = \arg \max_{\theta \in \Theta_j} Q_n^*(\theta, \tilde{\theta}_{j-1})$ wpa1. Expanding the first order condition $\nabla_{\theta} Q_n^*(\tilde{\theta}_j, \tilde{\theta}_{j-1}) = 0$ around $(\hat{\theta}, \tilde{\theta}_{j-1})$ gives

$$0 = \nabla_{\theta} Q_n^*(\hat{\theta}, \tilde{\theta}_{j-1}) + \nabla_{\theta\theta'} Q_n^*(\bar{\theta}, \tilde{\theta}_{j-1})(\tilde{\theta}_j - \hat{\theta}), \quad \bar{\theta} \in [\tilde{\theta}_j, \hat{\theta}]. \quad (24)$$

writing $\bar{\theta} = \bar{\theta}(\tilde{\theta}_j)$ and proceeding as in the proof of Proposition 7 in conjunction with (ii) above, we obtain $\sup_{(\tilde{\theta}_j, \tilde{\theta}_{j-1}) \in \mathcal{N}_{\theta\eta}(\epsilon_1)} \|\nabla_{\theta\theta'} Q_n^*(\bar{\theta}(\tilde{\theta}_j), \tilde{\theta}_{j-1})^{-1}\| = O_p(1)$. For $\nabla_{\theta} Q_n^*(\hat{\theta}, \tilde{\theta}_{j-1})$, since the MLE satisfies $\nabla_{\theta} Q_n^*(\hat{\theta}, \hat{\theta}) = 0$ wpa1, expanding $\nabla_{\theta} Q_n^*(\hat{\theta}, \tilde{\theta}_{j-1})$ around $(\hat{\theta}, \hat{\theta})$ gives $\nabla_{\theta} Q_n^*(\hat{\theta}, \tilde{\theta}_{j-1}) = \nabla_{\theta\eta'} Q_n^*(\hat{\theta}, \hat{\theta})(\tilde{\theta}_{j-1} - \hat{\theta}) + O_p(\|\tilde{\theta}_{j-1} - \hat{\theta}\|^2)$. Now, $\nabla_{\theta\eta'} Q_n^*(\hat{\theta}, \hat{\theta}) = n^{-1} \sum_{i=1}^n \nabla_{\theta\theta'} P(\hat{\theta})(a_i|x_i)/P(\hat{\theta})(a_i|x_i) = O_p(n^{-1/2})$, where the last equality follows from the root- n consistency of $\hat{\theta}$ because the information matrix equality implies $E[\nabla_{\theta\theta'} P(\theta^0)(a_i|x_i)/P(\theta^0)(a_i|x_i)] = 0$. Therefore, $\nabla_{\theta} Q_n^*(\hat{\theta}, \tilde{\theta}_{j-1}) = O_p(n^{-1/2}\|\tilde{\theta}_{j-1} - \hat{\theta}\| + \|\tilde{\theta}_{j-1} - \hat{\theta}\|^2)$, and the stated uniform bound of $\tilde{\theta}_j - \hat{\theta}$ follows from (24).

It remains to show $(\theta, \eta) \in \mathcal{N}_{\theta\eta}(\epsilon_1)$ wpa1 if $c > 0$ is taken sufficiently small. The proof is essentially the same as the proof of $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ wpa1 in the proof of Proposition 9(b). The argument of the proof of Proposition 9(b) carries through if we replace $\Lambda^q(\theta, P, \eta)$ with $\Phi(\theta, P, \eta)$, because (i) $\ln \Phi(\theta, P, \eta)$ is continuous in $(\theta, P, \eta) \in \Theta_j \times \mathcal{N}$ from Assumptions 5(b)(c)(h), (ii) $E \sup_{(\theta, P, \eta) \in \Theta_j \times \mathcal{N}} |\ln \Phi(\theta, P, \eta)(a_i|x_i)| < \infty$ from the compactness of Θ_j and the continuity of $\ln \Phi(\theta, P, \eta)$, and (iii) $\nabla_{\theta'} P(\theta^0)\nu = (I - \Psi_P)^{-1}\Psi_{\theta}\nu \neq 0$ for any K -vector $\nu \neq 0$ from Assumption 5(h). \square

Proof of Proposition 12 We suppress the subscript MLE from $\hat{\theta}_{MLE}$ and \hat{P}_{MLE} . The updating formula for \tilde{P}_j follows from expanding the right hand side of $\tilde{P}_j = \Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ and $\tilde{P}_j = \Gamma^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ twice around $(\hat{\theta}, \hat{P})$ and using $\hat{P} = \Lambda^q(\hat{\theta}, \hat{P}) = \Gamma^q(\hat{\theta}, \hat{P})$. Both $\Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ and $\Gamma^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ are continuously differentiable from Assumption 5(c) and the nonsingularity of $I - \Pi(\theta^0, P^0)\Psi_P\Pi(\theta^0, P^0)$.

We proceed to derive the bound of $\tilde{\theta}_j - \hat{\theta}$. We focus on the case when $\tilde{P}_j = \Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ is used. The same argument carries through for $\tilde{P}_j = \Gamma^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$. The proof closely follows the proof of Proposition 7. Define

$Q_n(\theta, P, \eta) = n^{-1} \sum_{i=1}^n \ln[\Lambda^q(\eta, P) + (I - \nabla_{P'} \Psi(\eta, \Lambda^q(\eta, P)))^{-1} \nabla_{\theta'} \Psi(\eta, \Lambda^q(\eta, P))(\theta - \eta)](a_i|x_i)$, so that $Q_n(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ is the objective function for $\tilde{\theta}_j$, and define $Q(\theta, P, \eta) = E \ln[\Lambda^q(\eta, P) + (I - \nabla_{P'} \Psi(\eta, \Lambda^q(\eta, P)))^{-1} \nabla_{\theta'} \Psi(\eta, \Lambda^q(\eta, P))(\theta - \eta)](a_i|x_i)$. For $\epsilon > 0$, define a neighborhood $\mathcal{N}_3(\epsilon) = \{(\theta, P, \eta) : \max\{\|\theta - \theta^0\|, \|P - P^0\|, \|\eta - \theta^0\|\} < \epsilon\}$. Then, there exists $\epsilon_1 > 0$ such that $\sup_{(\theta, P, \eta) \in \mathcal{N}_3(\epsilon_1)} \|\nabla_{\theta\theta'} Q(\theta, P, \eta)^{-1}\| < \infty$ and $\sup_{(\theta, P, \eta) \in \mathcal{N}_3(\epsilon_1)} \|\nabla^3 Q(\theta, P, \eta)\| < \infty$ because $\nabla_{\theta\theta'} Q(\theta^0, P^0, \theta^0) = -\mathcal{I}^0$, $I - \Psi_P$ is nonsingular, and $\Psi(\theta, P)$ is four times continuously

differentiable.

We first assume $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ and derive the stated representation of $\tilde{\theta}_j - \hat{\theta}$. We later show $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ wpa1 if $c > 0$ is taken sufficiently small. Expanding the first order condition $\nabla_{\theta} Q_n(\tilde{\theta}_j, \tilde{P}_j, \tilde{\theta}_{j-1}) = 0$ around $(\hat{\theta}, \hat{P}_j, \hat{\theta}_{j-1})$ gives

$$0 = \nabla_{\theta} Q_n(\hat{\theta}, \hat{P}_j, \hat{\theta}_{j-1}) + \nabla_{\theta\theta'} Q_n(\bar{\theta}, \tilde{P}_j, \tilde{\theta}_{j-1})(\tilde{\theta}_j - \hat{\theta}), \quad \bar{\theta} \in [\tilde{\theta}_j, \hat{\theta}]. \quad (25)$$

Consider $\nabla_{\theta} Q_n(\hat{\theta}, \tilde{P}_j, \tilde{\theta}_{j-1})$ on the right of (25). Note that $\nabla_{\theta} Q_n(\hat{\theta}, \hat{P}, \hat{\theta}) = 0$ and $E[\nabla^j \Phi(\theta, P, \eta)(a_i|x_i)/\Phi(\theta, P, \eta)(a_i|x_i)] = 0$ for $j \geq 1$ when evaluated at $(\theta, P, \eta) = (\theta^0, P^0, \theta^0)$. Then expanding $\nabla_{\theta} Q_n(\hat{\theta}, \tilde{P}_j, \tilde{\theta}_{j-1})$ twice around $(\hat{\theta}, \hat{P}, \hat{\theta})$ and using the root- n consistency of $(\hat{\theta}, \hat{P})$ and the information matrix equality gives

$$\nabla_{\theta} Q_n(\hat{\theta}, \tilde{P}_j, \tilde{\theta}_{j-1}) = -E[\nabla_{\theta} \ln P(\theta^0)(a_i|x_i)I(a_i|x_i)/P^0(a_i|x_i)](\tilde{P}_j - \hat{P}) + \mathcal{I}^0(\tilde{\theta}_{j-1} - \hat{\theta}) + r_{nj}, \quad (26)$$

where $I(a_i|x_i)$ is the row of an $L \times L$ identity matrix corresponding to $(a_i|x_i)$, and r_{nj} is a remainder term of $O_p(n^{-1/2} \|\tilde{\theta}_{j-1} - \hat{\theta}\| + \|\tilde{\theta}_{j-1} - \hat{\theta}\|^2 + n^{-1/2} \|\tilde{P}_j - \hat{P}\| + \|\tilde{P}_j - \hat{P}\|^2)$. For $\nabla_{\theta\theta'} Q_n(\bar{\theta}, \tilde{P}_j, \tilde{\theta}_{j-1})$ on the right of (25), writing $\bar{\theta} = \bar{\theta}(\tilde{\theta}_j)$ and proceeding as in the proof of Proposition 7 in conjunction with $\sup_{(\theta, P, \eta) \in \mathcal{N}_3(\epsilon_1)} \|\nabla_{\theta\theta'} Q(\theta, P, \eta)^{-1}\| < \infty$, we obtain

$\sup_{(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)} \|\nabla_{\theta\theta'} Q_n(\bar{\theta}(\tilde{\theta}_j), \tilde{P}_j, \tilde{\theta}_{j-1})^{-1}\| = O_p(1)$. Therefore, we have $\tilde{\theta}_j - \hat{\theta} = O_p(\|\tilde{\theta}_{j-1} - \hat{\theta}\| + \|\tilde{P}_j - \hat{P}\|)$ from (25) and (26). Having established $\tilde{\theta}_j - \hat{\theta} = O_p(\|\tilde{\theta}_{j-1} - \hat{\theta}\| + \|\tilde{P}_j - \hat{P}\|)$, refine (25) as $0 = \nabla_{\theta} Q_n(\hat{\theta}, \tilde{P}_j, \tilde{\theta}_{j-1}) - \mathcal{I}^0(\tilde{\theta}_j - \hat{\theta}) + r_n$ by expanding $\nabla_{\theta\theta'} Q_n(\bar{\theta}, \tilde{P}_j, \tilde{\theta}_{j-1})$ around $(\hat{\theta}, \hat{P}, \hat{\theta})$. Then, the stated updating formula for $\tilde{\theta}_j$ follows from substituting it into (26).

It remains to show $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ wpa1 if $c > 0$ is taken sufficiently small. The proof is essentially the same as the proof of $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ wpa1 in the proof of Proposition 9(b). The argument of the proof of Proposition 9(b) carries through if we replace $\Lambda^q(\theta, P, \eta)$ with $\Phi(\theta, P, \eta)$, because (i) $\ln \Phi(\theta, P, \eta)$ is continuous in $(\theta, P, \eta) \in \Theta_j \times \mathcal{N}$ from Assumptions 5(b)(c)(h), (ii) $E \sup_{(\theta, P, \eta) \in \Theta_j \times \mathcal{N}} |\ln \Phi(\theta, P, \eta)(a_i|x_i)| < \infty$ from the compactness of Θ_j and the continuity of $\ln \Phi(\theta, P, \eta)$, and (iii) $\nabla_{\theta'} P(\theta^0)\nu = (I - \Psi_P)^{-1} \Psi_{\theta\nu} \neq 0$ for any K -vector $\nu \neq 0$ from Assumption 5(h). \square

C Sequential GMM estimators

Recently, many researchers extend the Hotz-Miller CCP estimator and develop various two-step moment estimators for dynamic games (see Bajari, Benkard and Levin, 2007; Pakes, Ostrovsky and Berry, 2007; Pesendorfer and Schmidt-Dengler, 2008). These estimators often suffer from finite sample bias, especially when the initial estimator of P^0 is imprecise. This section develops a recursive extension of two-step moment estimators called the *nested GMM estimator* using an idea similar to that in the NPL algorithm.

Let $P^0 = \{P^0(a|x)\}_{(a,x) \in A \times X}$ denote the equilibrium conditional choice probabilities. Then,

for any function $h(a)$, the following conditional moment condition holds:

$$E \left[h(a) - \sum_{a' \in A} h(a') P^0(a'|x) \middle| x \right] = E[h(a)|x] - E \left[\sum_{a' \in A} h(a') P^0(a'|x) \right] = 0.$$

Here, $E[h(a)|x]$ represents the model-free conditional expectation of $h(a)$, whereas $E[\sum_{a' \in A} h(a') P^0(a'|x)]$ represents the conditional expectation of $h(a)$ implied by the model P^0 . For example, we may choose $h(a) = a$ or $h(a) = a^2$. The conditional moment condition implies that the following unconditional moment condition holds for any function $\rho_m(x)$ and $h_m(a)$, where $m = 1, \dots, M$:

$$E[g_m(a, x; P^0)] = 0, \quad g_m(a, x; P^0) = \rho_m(x) \left(h_m(a) - \sum_{a' \in A} h_m(a') P^0(a'|x) \right). \quad (27)$$

We consider a generalized method of moments (GMM) estimator based on these moment conditions when the equilibrium conditional choice probabilities belong to a parametric class with a fixed point constraint: $\mathcal{M} = \cup_{\theta \in \Theta} \mathcal{M}_\theta$, where $\mathcal{M}_\theta = \{P \in B_P : P = \Psi(\theta, P)\}$. Define the GMM estimator as:

$$\hat{\theta}_{GMM} = \arg \min_{\theta \in \Theta} \left\{ \min_{P \in \mathcal{M}_\theta} \bar{g}(P)' \hat{W} \bar{g}(P) \right\},$$

where $\bar{g}(P) = n^{-1} \sum_{i=1}^n g(a_i, x_i; P)$, and $\hat{W} \rightarrow_p W$, which is positive definite. Here, $g(a, x; P) = (g_1(a, x; P), \dots, g_M(a, x; P))'$ is an M -vector of moment conditions, where the $g_m(a, x; P)$'s are defined in (27).

To compute the GMM estimator, we need to repeatedly solve the fixed point $P = \Psi(\theta, P)$ for each candidate parameter value θ until one finds the parameter that minimizes the GMM objective function. When solving the fixed point is costly, this estimator is impractical.

The two-step GMM estimator is defined as $\hat{\theta}_{2GMM} = \arg \min_{\theta \in \Theta} \bar{g}(\Psi(\theta, \hat{P}_0))' \hat{W} \bar{g}(\Psi(\theta, \hat{P}_0))$, where \hat{P}_0 is an initial consistent estimator for P^0 .

We introduce the following notation:

$$\begin{aligned} \bar{G}_\theta(\Psi(\theta, P)) &= \nabla_{\theta'} \bar{g}(\Psi(\theta, P)), & \bar{G}_P(\Psi(\theta, P)) &= \nabla_{P'} \bar{g}(\Psi(\theta, P)), \\ G_\theta &= E[\nabla_{\theta'} g(a_i, x_i; \Psi(\theta^0, P^0))], & G_P &= E[\nabla_{P'} g(a_i, x_i; \Psi(\theta^0, P^0))]. \end{aligned}$$

Define $L = |A||X|$. Let f_x be an $L \times 1$ vector of $\Pr(x = x^s)$, $s = 1, \dots, |X|$, whose elements are arranged conformably with $P^0(a^j|x^s)$, and let \hat{f}_x be the frequency estimator of f_x . Denote $\Delta_x = \text{diag}(f_x)$ and $\hat{\Delta}_x = \text{diag}(\hat{f}_x)$. Let γ_m be an $L \times 1$ vector of $\rho_m(x^s) h_m(a^j)$ whose elements are ordered conformably with $P^0(a^j|x^s)$, and let $H = (\gamma_1, \dots, \gamma_M)'$, which is an M by L matrix. With this notation, we may write $\bar{G}_\theta(\Psi(\theta, P)) = -H \hat{\Delta}_x \nabla_{\theta'} \Psi(\theta, P)$, $\bar{G}_P(\Psi(\theta, P)) = -H \hat{\Delta}_x \nabla_{P'} \Psi(\theta, P)$, $G_\theta = -H \Delta_x \Psi_\theta$ and $G_P = -H \Delta_x \Psi_P$. Let $r(a, x)$ be an $L \times 1$ vector of indicator functions whose elements are ordered conformably with $P^0(a^j|x^s)$, so

that $\hat{P}_0 - P^0 = n^{-1} \sum_{i=1}^n r(a_i, x_i) + o_p(n^{-1/2})$. The explicit form of $r(a, x)$ can be found by expanding $\hat{P}_0 - P^0$.

Assumption 6 (a) For any $\theta \neq \theta^0$, $E[g(a, x; \Psi(\theta, P^0))] \neq 0$; (b) $G'_\theta W G_\theta$ is nonsingular; (c) $E \sup_{\theta \in \Theta} \|g(a, x; \Psi(\theta, P^0))\| < \infty$; (d) $E \sup_{\theta \in \Theta} \|\nabla_{\theta'} g(a, x; \Psi(\theta, P^0))\| < \infty$, $E \sup_{\theta \in \Theta} \|\nabla_{P'} g(a, x; \Psi(\theta, P^0))\| < \infty$; (e) $E \|g(a, x; P^0)\|^2 < \infty$.

Under Assumptions 1 and 6, $\hat{\theta}_{2GMM}$ is consistent and asymptotically normal: $\sqrt{n}(\hat{\theta}_{2GMM} - \theta^0) \rightarrow_d N(0, V_{2GMM})$, where $V_{2GMM} = (G'_\theta W G_\theta)^{-1} G'_\theta W S W G_\theta (G'_\theta W G_\theta)^{-1}$ with $S = E[(g(a_i, x_i; P^0) - G_P(r(a_i, x_i) - P^0))(g(a_i, x_i; P^0) - G_P(r(a_i, x_i) - P^0))']$. Using the optimal weighting matrix $W = S^{-1}$, the limiting variance is given by $V_{2GMM} = (G'_\theta S^{-1} G_\theta)^{-1}$.

We now consider a recursive extension of the two-step GMM estimator called the *nested GMM algorithm* which iterates the following steps until $j = k$:

Step 1: Given \tilde{P}_{j-1} , update θ by $\tilde{\theta}_j = \arg \min_{\theta} \bar{g}(\Psi(\theta, \tilde{P}_{j-1}))' \hat{W} \bar{g}(\Psi(\theta, \tilde{P}_{j-1}))$.

Step 2: Update P using the obtained estimate $\tilde{\theta}_j$: $\tilde{P}_j = \Psi(\tilde{\theta}_j, \tilde{P}_{j-1})$.

If the iterations converge, the limit satisfies $\check{\theta} = \arg \min_{\theta \in \Theta} \bar{g}(\Psi(\theta, \check{P}))' \check{W} \bar{g}(\Psi(\theta, \check{P}))$ and $\check{P} = \Psi(\check{\theta}, \check{P})$. Among the pairs $(\check{\theta}, \check{P})$ that satisfy these two conditions, the one that minimizes the value of the criterion function $\bar{g}(\Psi(\theta, P))' \hat{W} \bar{g}(\Psi(\theta, P))$ is called the *nested GMM (NGMM) estimator*, which we denote by $(\hat{\theta}_{NGMM}, \hat{P}_{NGMM})$.

Under regularity conditions similar to the ones in Assumption 1, the sequence of estimators generated by this algorithm is consistent. The following proposition establishes the limiting distribution of the NGMM estimator.

Proposition 13 *Suppose Assumptions 1 and 6 hold. Then*

$$\sqrt{n}(\hat{\theta}_{NGMM} - \theta^0) \rightarrow_d N(0, (G'_\theta W G_\theta^\infty)^{-1} G'_\theta W \Omega W' G_\theta ((G_\theta^\infty)' W' G_\theta)^{-1}),$$

where $\Omega = E[g(a_i, x_i; P^0)g(a_i, x_i; P^0)']$ and $G_\theta^\infty = -H \Delta_x (I - \Psi_P)^{-1} \Psi_\theta$. If we choose $W = \Omega^{-1}$, the asymptotic variance is given by $(G'_\theta \Omega^{-1} G_\theta^\infty)^{-1} G'_\theta \Omega^{-1} G_\theta ((G_\theta^\infty)' \Omega^{-1} G_\theta)^{-1}$.

Remark 2 *When $\Psi_P = 0$, the two-step GMM estimator with the optimal weighting matrix is asymptotically equivalent to the NGMM estimator with $W = \Omega^{-1}$.*

The NGMM estimator can be obtained as the limit of the sequence of estimators generated by the NGMM algorithm if iterations converge. The convergence properties of the NGMM estimator is given by the following proposition.

Proposition 14 *Suppose Assumptions 1 and 6 hold, and $\tilde{P}_0 - P^0 = o_p(1)$. Then, for $j = 1, \dots, k$,*

$$\begin{aligned}\tilde{\theta}_j - \hat{\theta}_{NGMM} &= O_p(\|\tilde{P}_{j-1} - \hat{P}_{NGMM}\|), \\ \tilde{P}_j - \hat{P}_{NGMM} &= [I + \Psi_\theta(G'_\theta \hat{W} G_\theta)^{-1} G'_\theta \hat{W} H \Delta_x] \Psi_P(\tilde{P}_{j-1} - \hat{P}_{NGMM}) \\ &\quad + O_p(n^{-1/2} \|\tilde{P}_{j-1} - \hat{P}_{NGMM}\|) + O_p(\|\tilde{P}_{j-1} - \hat{P}_{NGMM}\|^2).\end{aligned}$$

Remark 3 *Because $-\Psi_\theta(G'_\theta \hat{W} G_\theta)^{-1} G'_\theta \hat{W} H \Delta_x = \Psi_\theta(\Psi'_\theta \Delta'_x H' \hat{W} H \Delta_x \Psi_\theta)^{-1} \Psi'_\theta \Delta'_x H' \hat{W} H \Delta_x$ is a projection matrix, the convergence properties of the NGMM algorithm is analogous to that of the NPL algorithm. Again, the convergence rate is primarily determined by the eigenvalues of Ψ_P .*

C.1 Proof of propositions in Section C

Proof of Proposition 13 Suppress the subscript NGMM from $\hat{\theta}_{NGMM}$ and \hat{P}_{NGMM} . The consistency of $(\hat{\theta}, \hat{P})$ follows from applying the proof of Proposition 2 of Aguirregabiria and Mira (2007).

For the asymptotic distribution of $(\hat{\theta}, \hat{P})$, observe that $(\hat{\theta}, \hat{P})$ satisfies $\bar{G}_\theta(\Psi(\hat{\theta}, \hat{P}))' \hat{W} \bar{g}(\Psi(\hat{\theta}, \hat{P})) = 0$ and $\hat{P} - \Psi(\hat{\theta}, \hat{P}) = 0$. Expanding $\bar{g}(\Psi(\hat{\theta}, \hat{P}))$ around (θ^0, P^0) and using the consistency of $(\hat{\theta}, \hat{P})$ gives

$$\begin{aligned}G'_\theta W \bar{g}(\Psi(\theta^0, P^0)) + G'_\theta W G_\theta(\hat{\theta} - \theta^0) + G'_\theta W G_P(\hat{P} - P^0) &= o_p(n^{-1/2}), \\ (I - \Psi_P)(\hat{P} - P^0) - \Psi_\theta(\hat{\theta} - \theta^0) &= o_p(n^{-1/2}).\end{aligned}$$

Eliminating $(\hat{P} - P^0)$ from these equations and using $G'_\theta W G_\theta + G'_\theta W G_P(I - \Psi_P)^{-1} \Psi_\theta = G'_\theta W G_\theta^\infty$, where $G_\theta^\infty = \nabla_{\theta'} \bar{g}(P(\theta^0)) = -H \Delta_x (I - \Psi_P)^{-1} \Psi_\theta$, we have $\sqrt{n}(\hat{\theta} - \theta^0) \rightarrow_d N(0, (G'_\theta W G_\theta^\infty)^{-1} G'_\theta W \Omega W' G_\theta ((G_\theta^\infty)' W' G_\theta)^{-1})$, where $\Omega = E[g(a_i, x_i; P^0) g(a_i, x_i; P^0)']$. \square

Proof of Proposition 14 Suppress the subscript NGMM from $\hat{\theta}_{NGMM}$ and \hat{P}_{NGMM} . We use induction. Assume \tilde{P}_{j-1} is consistent. Then, $\tilde{\theta}_j$ is consistent because $\bar{g}(\Psi(\theta, \tilde{P}_{j-1}))' \hat{W} \bar{g}(\Psi(\theta, \tilde{P}_{j-1}))$ converges uniformly to $Eg(a_i, x_i; \Psi(\theta, P^0))' W E g(a_i, x_i; \Psi(\theta, P^0))$.

For the bound of $\tilde{\theta}_j$, recall that $\tilde{\theta}_j$ satisfies the first order condition

$$\bar{G}'_\theta(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1})) \hat{W} \bar{g}(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1})) = 0. \quad (28)$$

Expanding $\bar{g}(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1}))$ around $(\hat{\theta}, \hat{P})$ in (28) and using $\bar{G}'_\theta(\Psi(\hat{\theta}, \hat{P})) \hat{W} \bar{g}(\Psi(\hat{\theta}, \hat{P})) = 0$ gives

$$\begin{aligned}\tilde{\theta}_j - \hat{\theta} &= [\bar{G}'_\theta(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1})) \hat{W} \bar{G}_\theta(\Psi(\bar{P}, \bar{\theta})) + o_p(1)]^{-1} [\bar{G}'_\theta(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1})) \hat{W} \bar{G}_P(\Psi(\bar{P}, \bar{\theta})) + o_p(1)] (\tilde{P}_{j-1} - \bar{P}) \\ &= O_p(\|\tilde{P}_{j-1} - \bar{P}\|),\end{aligned} \quad (29)$$

where $(\bar{\theta}, \bar{P})$ lies between $(\tilde{\theta}_j, \tilde{P}_{j-1})$ and $(\hat{\theta}, \hat{P})$.

For the second result, we begin by using (29) to obtain

$$\tilde{P}_j - \hat{P} = \Psi_\theta(\tilde{\theta}_j - \hat{\theta}) + \Psi_P(\tilde{P}_{j-1} - \hat{P}) + O_p(n^{-1/2} \|\tilde{P}_{j-1} - \hat{P}\|) + O_p(\|\tilde{P}_{j-1} - \hat{P}\|^2) \quad (30)$$

Expanding $\bar{g}(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1}))$ in (28) twice around $(\hat{\theta}, \hat{P})$ and using $\bar{G}'_\theta(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1}))\hat{W}\bar{g}(\Psi(\hat{\theta}, \hat{P})) = O_p(n^{-1/2} \|\tilde{\theta}_j - \hat{\theta}\|) + O_p(n^{-1/2} \|\tilde{P}_{j-1} - \hat{P}\|)$,

$$\bar{G}_P(\Psi(\hat{\theta}, \hat{P})) = G_P + O_p(n^{-1/2}), \quad \bar{G}_\theta(\Psi(\hat{\theta}, \hat{P})) = G_\theta + O_p(n^{-1/2}), \quad (31)$$

and (29) gives

$$\begin{aligned} 0 &= \bar{G}'_\theta(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1}))\hat{W}G_P(\tilde{P}_{j-1} - \hat{P}) + \bar{G}'_\theta(\Psi(\tilde{\theta}_j, \tilde{P}_{j-1}))\hat{W}G_\theta(\tilde{\theta}_j - \hat{\theta}) \\ &\quad + O_p(n^{-1/2} \|\tilde{P}_{j-1} - \hat{P}\|) + O_p(\|\tilde{P}_{j-1} - \hat{P}\|^2). \end{aligned} \quad (32)$$

Expanding $\Psi(\tilde{\theta}_j, \tilde{P}_{j-1})$ around $(\hat{\theta}, \hat{P})$ and using (29) and (31) in (32), we have

$$\tilde{\theta}_j - \hat{\theta} = -(G'_\theta \hat{W} G_\theta)^{-1} G'_\theta \hat{W} G_P(\tilde{P}_{j-1} - \hat{P}) + O_p(n^{-1/2} \|\tilde{P}_{j-1} - \hat{P}\|) + O_p(\|\tilde{P}_{j-1} - \hat{P}\|^2).$$

Substituting this into (30) and noting that $G_\theta = -H\Delta_x\Psi_\theta$ and $G_P = -H\Delta_x\Psi_P$, we obtain

$$\tilde{P}_j - \hat{P} = [I + \Psi_\theta(G'_\theta \hat{W} G_\theta)^{-1} G'_\theta \hat{W} H\Delta_x] \Psi_P(\tilde{P}_{j-1} - \hat{P}) + O_p(n^{-1/2} \|\tilde{P}_{j-1} - \hat{P}\|) + O_p(\|\tilde{P}_{j-1} - \hat{P}\|^2),$$

and the second result follows. \square

D Unobserved Heterogeneity

This section extends our analysis to models with unobserved heterogeneity. The NPL algorithm has an important advantage over two step methods in estimating models with unobserved heterogeneity because it is difficult to obtain a reliable initial estimate of P in this context.

Suppose that there are M types of agents, where type m is characterized by a type-specific parameter θ^m , and the probability of being type m is π^m with $\sum_{m=1}^M \pi^m = 1$. These types capture time-invariant state variables that are unobserved by researchers. With a slight abuse of notation, denote $\theta = (\theta^1, \dots, \theta^M)' \in \Theta^M$ and $\pi = (\pi^1, \dots, \pi^M)' \in \Theta_\pi$. Then, $\zeta = (\theta', \pi')'$ is the parameter to be estimated, and let $\Theta_\zeta = \Theta^M \times \Theta_\pi$ denote the set of possible values of ζ . The true parameter is denoted by ζ^0 .

Consider a panel data set $\{\{a_{it}, x_{it}, x_{i,t+1}\}_{t=1}^T\}_{i=1}^n$ such that $w_i = \{a_{it}, x_{it}, x_{i,t+1}\}_{t=1}^T$ is randomly drawn across i 's from the population. The conditional probability distribution of a_{it} given x_{it} for a type m agent is given by a fixed point of $P_{\theta^m} = \Psi(\theta^m, P_{\theta^m})$. To simplify our analysis, we assume that the transition probability function of x_{it} is independent of types and given by

$f_x(x_{i,t+1}|a_{it}, x_{it})$ and is known to researchers.³

In this framework, the initial state x_{i1} is correlated with the unobserved type (i.e., the initial conditions problem of Heckman (1981)). We assume that x_{i1} for type m is randomly drawn from the type m stationary distribution characterized by a fixed point of the following equation: $p^*(x) = \sum_{x' \in X} p^*(x') (\sum_{a' \in A} P_{\theta^m}(a'|x') f_x(x|a', x')) \equiv [T(p^*, P_{\theta^m})](x)$. Since solving the fixed point of $T(\cdot, P)$ for given P is often less computationally intensive than computing the fixed point of $\Psi(\cdot, \theta)$, we assume the full solution of the fixed point of $T(\cdot, P)$ is available given P .

Let P^m denote type m 's conditional choice probabilities, stack the P^m 's as $\mathbf{P} = (P^1, \dots, P^M)'$, and let \mathbf{P}^0 denote its true value. Define $\Psi(\theta, \mathbf{P}) = (\Psi(\theta^1, P^1)', \dots, \Psi(\theta^M, P^M)')'$. Then, for a value of θ , the set of possible conditional choice probabilities consistent with the fixed point constraints is given by $\mathcal{M}_\theta^* = \{\mathbf{P} \in B_P^M : \mathbf{P} = \Psi(\theta, \mathbf{P})\}$. The maximum likelihood estimator for a model with unobserved heterogeneity is:

$$\hat{\zeta}_{MLE} = \arg \max_{\zeta \in \Theta_\zeta} \left\{ \max_{\mathbf{P} \in \mathcal{M}_\theta^*} n^{-1} \sum_{i=1}^n \ln ([L(\pi, \mathbf{P})](w_i)) \right\}, \quad (33)$$

where $[L(\pi, \mathbf{P})](w_i) = \sum_{m=1}^M \pi^m p_{P^m}^*(x_{i1}) \prod_{t=1}^T P^m(a_{it}|x_{it}) f_x(x_{i,t+1}|a_{it}, x_{it})$, and $p_{P^m}^* = T(p_{P^m}^*, P^m)$ is the type m stationary distribution of x when the conditional choice probability is P^m . If \mathbf{P}^0 is the true conditional choice probability distribution and π^0 is the true mixing distribution, then $L^0 = L(\pi^0, \mathbf{P}^0)$ represents the true probability distribution of w .

We consider a version of the NPL algorithm for models with unobserved heterogeneity originally developed by AM07 as follows. Assume that an initial consistent estimator $\tilde{\mathbf{P}}_0 = (\tilde{P}_0^1, \dots, \tilde{P}_0^M)$ is available. For $j = 1, 2, \dots$, iterate

Step 1: Given $\tilde{\mathbf{P}}_{j-1}$, update $\zeta = (\theta', \pi')'$ by $\tilde{\zeta}_j = \arg \max_{\zeta \in \Theta_\zeta} n^{-1} \sum_{i=1}^n \ln ([L(\pi, \Psi(\theta, \tilde{\mathbf{P}}_{j-1}))](w_i))$,

Step 2: Update \mathbf{P} using the obtained estimate $\tilde{\theta}_j$ by $\tilde{\mathbf{P}}_j = \Psi(\tilde{\theta}_j, \tilde{\mathbf{P}}_{j-1})$,

until $j = k$. If iterations converge, the limit satisfies $\hat{\zeta} = \arg \max_{\zeta \in \Theta_\zeta} n^{-1} \sum_{i=1}^n \ln ([L(\pi, \Psi(\theta, \hat{\mathbf{P}}))](w_i))$ and $\hat{\mathbf{P}} = \Psi(\hat{\theta}, \hat{\mathbf{P}})$. Among the pairs that satisfy these two conditions, the one that maximizes the pseudo likelihood is called the *NPL estimator*, which we denote by $(\hat{\zeta}_{NPL}, \hat{\mathbf{P}}_{NPL})$.

Let us introduce the assumptions required for the consistency and asymptotic normality of the NPL estimator. They are analogous to the assumptions used in Aguirregabiria and Mira (2007). Define $\tilde{\zeta}_0(\mathbf{P})$ and $\phi_0(\mathbf{P})$ similar to $\tilde{\theta}_0(P)$ and $\phi_0(P)$ in the main paper.

Assumption 7 (a) $w_i = \{(a_{it}, x_{it}, x_{i,t+1}) : t = 1, \dots, T\}$ for $i = 1, \dots, n$, are independently and identically distributed, and $dF(x) > 0$ for any $x \in X$, where $F(x)$ is the distribution

³When the transition probability function is independent of types, it can be directly estimated from transition data without solving the fixed point problem. Kasahara and Shimotsu (2008) analyze the case in which the transition probability function is also type-dependent in the context of a single-agent dynamic programming model with unobserved heterogeneity.

function of x_i . (b) $[L(\pi, \mathbf{P})](w) > 0$ for any w and for any $(\pi, \mathbf{P}) \in \Theta_\pi \times B_P^M$. (c) $\Psi(\theta, P)$ is twice continuously differentiable. (d) Θ_ζ is compact and B_P^M is a compact and convex subset of $[0, 1]^{LM}$. (e) There is a unique $\zeta^0 \in \text{int}(\Theta_\zeta)$ such that $[L(\pi^0, \mathbf{P}^0)](w) = [L(\pi^0, \Psi(\theta^0, \mathbf{P}^0))](w)$. (f) (ζ^0, \mathbf{P}^0) is an isolated population NPL fixed point. (g) $\tilde{\zeta}_0(\mathbf{P})$ is a single-valued and continuous function of \mathbf{P} in a neighborhood of \mathbf{P}^0 . (h) the operator $\phi_0(\mathbf{P}) - \mathbf{P}$ has a nonsingular Jacobian matrix at \mathbf{P}^0 . (i) For any $P \in B_P$, there exists a unique fixed point for $T(\cdot, P)$.

Under Assumption 7, the consistency and asymptotic normality of the NPL estimator can be shown by following the proof of Proposition 2 of Aguirregabiria and Mira (2007).

We now establish the convergence properties of the NPL algorithm for models with unobserved heterogeneity. Let $l(\zeta, \mathbf{P})(w) \equiv \ln(L(\pi, \Psi(\theta, \mathbf{P}))(w))$, and $\Omega_{\zeta\zeta} = E[\nabla_\zeta l(\zeta^0, \mathbf{P}^0)(w_i) \nabla_\zeta l(\zeta^0, \mathbf{P}^0)(w_i)]$.

Assumption 8 (a) Assumption 7 holds. (b) $\Psi(\theta, P)$ is three times continuously differentiable. (c) $\Omega_{\zeta\zeta}$ is nonsingular.

Assumption 8 requires an initial consistent estimator of the type-specific conditional probabilities. Kasahara and Shimotsu (2006, 2009) derive sufficient conditions for nonparametric identification of a finite mixture model and suggest a sieve estimator which can be used to obtain an initial consistent estimate of \mathbf{P} . On the other hand, as Aguirregabiria and Mira (2007) argue, if the NPL algorithm converges, then the limit may provide a consistent estimate of the parameter ζ even when $\tilde{\mathbf{P}}_0$ is not consistent.

The following proposition states the convergence properties of the NPL algorithm for models with unobserved heterogeneity. For $\epsilon > 0$, define a neighborhood of \mathbf{P}^0 by $\mathcal{N}_{\mathbf{P}}(\epsilon) = \{P : \|\mathbf{P} - \mathbf{P}^0\| < \epsilon\}$.

Proposition 15 Suppose Assumptions 7-8 hold. Then, there exists $c > 0$ such that

$$\begin{aligned} \tilde{\zeta}_j - \hat{\zeta}_{NPL} &= O_p(\|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}\|), \\ \tilde{\mathbf{P}}_j - \hat{\mathbf{P}}_{NPL} &= [I - \Psi_\theta D \Psi_\theta' L_P \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P] \Psi_P (\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) \\ &\quad + O_p(n^{-1/2} \|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}\|) + O_p(\|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}\|^2), \end{aligned}$$

uniformly in $\tilde{\mathbf{P}}_{j-1} \in \mathcal{N}_{\mathbf{P}}(c)$, where $D = (\Psi_\theta' L_P \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P \Psi_\theta)^{-1}$, $M_{L_\pi} = I - \Delta_L^{1/2} L_\pi (L_\pi' \Delta_L L_\pi)^{-1} L_\pi \Delta_L^{1/2}$, and $\Psi_\theta \equiv \nabla_{\theta'} \Psi(\theta^0, \mathbf{P}^0)$, $\Psi_P \equiv \nabla_{\mathbf{P}'} \Psi(\theta^0, \mathbf{P}^0)$, $\Delta_L = \text{diag}((L^0)^{-1})$, $L_P = \nabla_{\mathbf{P}'} L(\pi^0, \mathbf{P}^0)$, and $L_\pi = \nabla_{\pi'} L(\pi^0, \mathbf{P}^0)$.

Note that $I - \Psi_\theta D \Psi_\theta' L_P \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P$ is a projection matrix. The convergence rate of the NPL algorithm for models with unobserved heterogeneity is primarily determined by the dominant eigenvalue of Ψ_P . When the NPL algorithm encounters a convergence problem, replacing $\Psi(\theta, P)$ with $\Lambda(\theta, P)$ or $\Gamma(\theta, P)$ improves the convergence.

Remark 4 *It is possible to relax the stationarity assumption on the initial states by estimating the type-specific initial distributions of x , denoted by $\{p^{*m}\}_{m=1}^M$, without imposing a stationarity restriction in Step 1 of the NPL algorithm. In this case, Proposition 15 holds with additional remainder terms.*

Proof of Proposition 15 We suppress the subscript NPL from $\hat{\zeta}_{NPL}$ and $\hat{\mathbf{P}}_{NPL}$. The proof closely follows the proof of Lemma 1. Define $\bar{l}_\zeta(\zeta, \mathbf{P}) = n^{-1} \sum_{i=1}^n \nabla_\zeta l(\zeta, \mathbf{P})(w_i)$, $\bar{l}_{\zeta\zeta}(\zeta, \mathbf{P}) = n^{-1} \sum_{i=1}^n \nabla_{\zeta\zeta} l(\zeta, \mathbf{P})(w_i)$, and $\bar{l}_{\zeta\mathbf{P}}(\zeta, \mathbf{P}) = n^{-1} \sum_{i=1}^n \nabla_{\zeta\mathbf{P}} l(\zeta, \mathbf{P})(w_i)$. Expanding the first order condition $\bar{l}_\zeta(\tilde{\zeta}_j, \tilde{\mathbf{P}}_{j-1}) = \bar{l}_\zeta(\hat{\zeta}, \hat{\mathbf{P}}) = 0$ gives

$$0 = \bar{l}_{\zeta\zeta}(\bar{\zeta}, \bar{\mathbf{P}})(\tilde{\zeta}_j - \hat{\zeta}) + \bar{l}_{\zeta\mathbf{P}}(\bar{\zeta}, \bar{\mathbf{P}})(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}), \quad (34)$$

where $(\bar{\zeta}, \bar{\mathbf{P}})$ is between $(\tilde{\zeta}_j, \tilde{\mathbf{P}}_{j-1})$ and $(\hat{\zeta}, \hat{\mathbf{P}})$. Then, proceeding as in the proof of Lemma 1 gives the bound of $\tilde{\zeta}_j - \hat{\zeta}$.

For the bound of $\tilde{\mathbf{P}}_j - \hat{\mathbf{P}}$, expanding the second step equation $\tilde{\mathbf{P}}_j = \Psi(\tilde{\zeta}_j, \tilde{\mathbf{P}}_{j-1})$ twice around $(\hat{\zeta}, \hat{\mathbf{P}})$, using $\hat{\mathbf{P}} = \Psi(\hat{\zeta}, \hat{\mathbf{P}})$, and proceeding as in the proof of Lemma 1 gives

$$\tilde{\mathbf{P}}_j - \hat{\mathbf{P}} = \Psi_P(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}) + \Psi_\zeta(\tilde{\zeta}_j - \hat{\zeta}) + O_p(n^{-1/2} \|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}\|) + O_p(\|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}\|^2), \quad (35)$$

where $\Psi_\zeta \equiv \nabla_{\zeta'} \Psi(\theta^0, \mathbf{P}^0) = [\Psi_\theta, \mathbf{0}]$. As in the proof of Lemma 1, refine (34) further as $\tilde{\zeta}_j - \hat{\zeta} = -\Omega_{\zeta\zeta}^{-1} \Omega_{\zeta\mathbf{P}}(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}) + O_p(n^{-1/2} \|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}\|) + O_p(\|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}\|^2)$, where $\Omega_{\zeta\mathbf{P}} = E[\nabla_{\zeta} l(\zeta^0, \mathbf{P}^0)(w_i) \nabla_{\mathbf{P}} l(\zeta^0, \mathbf{P}^0)(w_i)]$. Substituting this into (35) gives

$$\tilde{\mathbf{P}}_j - \hat{\mathbf{P}} = [\Psi_P - \Psi_\zeta \Omega_{\zeta\zeta}^{-1} \Omega_{\zeta\mathbf{P}}](\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}) + O_p(n^{-1/2} \|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}\|) + O_p(\|\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}\|^2).$$

Note that $\Omega_{\zeta\zeta}$ and $\Omega_{\zeta\mathbf{P}}$ are written as

$$\Omega_{\zeta\zeta} = \begin{bmatrix} \Omega_{\theta\theta} & \Omega_{\theta\pi} \\ \Omega_{\pi\theta} & \Omega_{\pi\pi} \end{bmatrix} = \begin{bmatrix} \Psi'_\theta L'_P \Delta_L L_P \Psi_\theta & \Psi'_\theta L'_P \Delta_L L_\pi \\ L'_\pi \Delta_L L_P \Psi_\theta & L'_\pi \Delta_L L_\pi \end{bmatrix}, \quad \Omega_{\zeta\mathbf{P}} = \begin{bmatrix} \Omega_{\theta P} \\ \Omega_{\pi P} \end{bmatrix} = \begin{bmatrix} \Psi'_\theta L'_P \Delta_L L_P \Psi_P \\ L'_\pi \Delta_L L_P \Psi_P \end{bmatrix},$$

and

$$\Omega_{\zeta\zeta}^{-1} = \begin{bmatrix} D & -D\Omega_{\theta\pi}\Omega_{\pi\pi}^{-1} \\ -\Omega_{\pi\pi}^{-1}\Omega_{\pi\theta}D & \Omega_{\pi\pi}^{-1} + \Omega_{\pi\pi}^{-1}\Omega_{\pi\theta}D\Omega_{\theta\pi}\Omega_{\pi\pi}^{-1} \end{bmatrix},$$

where $D = (\Psi'_\theta L'_P \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P \Psi_\theta)^{-1}$ with $M_{L_\pi} = I - \Delta_L^{1/2} L_\pi (L'_\pi \Delta_L L_\pi)^{-1} L_\pi \Delta_L^{1/2}$. Then, using $\Psi_\zeta = [\Psi_\theta, \mathbf{0}]$ gives $\Psi_\zeta \Omega_{\zeta\zeta}^{-1} \Omega_{\zeta\mathbf{P}} = \Psi_\theta D \Psi'_\theta L'_P \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P \Psi_P$, and the stated result follows. \square

E Additional Monte Carlo results

Table 4 reports some additional results of our Monte Carlo experiments. In particular, Table 4 includes the performance of the estimator generated by the q -AFXP algorithm after $k = 50$ iterations and that of its PML version obtained by taking one iteration of the q -AFXP algorithm from the original PML estimator (i.e., the PML- Ψ). They are denoted by q -AFXP and PML-AFXP, respectively. The last panel of Table 4 reports the bias and the RMSE of P across different estimators, including those of the frequency estimator of P .

In most cases, the performance of the q -AFXP is comparable to that of the NPL- Λ^q and is better than that of the NPL- Λ or the NPL- Ψ . The RMSE's of the PML-AFXP are much smaller than those of the PML- Ψ , suggesting that taking one iteration of the approximate AFXP algorithm from the PML- Ψ substantially improves its finite sample performance.

Table 4: Bias and RMSE

	Estimator	$\theta_{RN} = 2$						$\theta_{RN} = 4$					
		$n = 500$		$n = 2000$		$n = 8000$		$n = 500$		$n = 2000$		$n = 8000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\hat{\theta}_{RS}$	NPL- Ψ	-0.0151	0.1347	-0.0002	0.0660	-0.0023	0.0323	-0.0095	0.0676	-0.0062	0.0490	-0.0005	0.0408
	NPL- Λ	-0.0151	0.1347	-0.0002	0.0660	-0.0023	0.0323	0.0028	0.0575	-0.0006	0.0294	-0.0003	0.0143
	RPM	-0.0174	0.1331	-0.0028	0.0642	-0.0027	0.0320	0.0029	0.0576	-0.0012	0.0284	0.0000	0.0136
	q -NPL- Λ^q	-0.0117	0.1240	0.0002	0.0606	-0.0018	0.0305	0.0015	0.0542	-0.0009	0.0277	0.0000	0.0136
	q -AFXP	-0.0115	0.1239	0.0002	0.0607	-0.0019	0.0305	0.0011	0.0534	-0.0007	0.0275	0.0000	0.0134
	PML- Ψ	-0.2215	0.2698	-0.0717	0.1112	-0.0229	0.0474	-0.1280	0.1557	-0.0341	0.0514	-0.0082	0.0207
	PML- Λ	-0.2215	0.2698	-0.0717	0.1112	-0.0229	0.0474	-0.1280	0.1557	-0.0341	0.0514	-0.0082	0.0207
	PML-RPM	0.1353	0.2380	0.0658	0.1072	0.0203	0.0403	0.1166	0.1823	0.0211	0.0457	0.0043	0.0176
	PML- Λ^q	-0.0133	0.1475	0.0016	0.0629	-0.0018	0.0307	0.0142	0.0783	-0.0035	0.0290	-0.0003	0.0141
PML-AFXP	-0.0112	0.1471	0.0018	0.0626	-0.0018	0.0306	-0.0089	0.0736	-0.0081	0.0304	-0.0014	0.0143	
$\hat{\theta}_{RN}$	NPL- Ψ	-0.0467	0.4705	-0.0009	0.2339	-0.0095	0.1130	-0.1417	0.2572	-0.1414	0.2314	-0.0918	0.1612
	NPL- Λ	-0.0467	0.4705	-0.0009	0.2339	-0.0095	0.1130	0.0241	0.1424	-0.0001	0.0739	0.0013	0.0352
	RPM	-0.0544	0.4642	-0.0102	0.2274	-0.0111	0.1116	0.0249	0.1604	-0.0003	0.0841	0.0014	0.0342
	q -NPL- Λ^q	-0.0358	0.4280	0.0002	0.2131	-0.0079	0.1052	0.0228	0.1351	0.0000	0.0690	0.0014	0.0328
	q -AFXP	-0.0352	0.4282	0.0003	0.2134	-0.0080	0.1052	0.0209	0.1319	0.0000	0.0673	0.0017	0.0324
	PML- Ψ	-0.7895	0.9604	-0.2565	0.3949	-0.0828	0.1687	-0.7713	0.9094	-0.1964	0.2599	-0.0462	0.0937
	PML- Λ	-0.7895	0.9604	-0.2565	0.3949	-0.0828	0.1687	-0.7713	0.9094	-0.1964	0.2599	-0.0462	0.0937
	PML-RPM	0.4523	0.8255	0.2232	0.3754	0.0687	0.1401	0.6101	0.7821	0.1282	0.1848	0.0335	0.0600
	PML- Λ^q	-0.0603	0.5177	0.0021	0.2215	-0.0083	0.1061	0.1619	0.2704	0.0044	0.0745	0.0035	0.0366
PML-AFXP	-0.0533	0.5158	0.0020	0.2196	-0.0088	0.1056	-0.0210	0.2011	-0.0423	0.0978	-0.0072	0.0430	
$100 \times \hat{P}$	Frequency	-0.0425	2.1609	0.0203	0.5128	0.0244	0.1550	-0.0880	5.8734	-0.0025	1.9222	0.0066	0.4413
	NPL- Ψ	0.0322	0.1561	0.0229	0.0436	0.0156	0.0256	-0.6258	3.4992	-0.1544	3.1243	0.0052	2.9592
	NPL- Λ	0.0321	0.1560	0.0229	0.0436	0.0156	0.0256	-0.0318	0.1393	-0.0094	0.0414	-0.0094	0.0113
	RPM	0.0243	0.1627	0.0228	0.0384	0.0160	0.0291	-0.0498	0.2053	-0.0163	0.0731	-0.0053	0.0085
	q -NPL- Λ^q	0.0249	0.1276	0.0207	0.0380	0.0146	0.0222	-0.0487	0.1278	-0.0136	0.0407	-0.0051	0.0081
	q -AFXP	0.0244	0.1274	0.0210	0.0383	0.0147	0.0224	-0.0484	0.1150	-0.0117	0.0375	-0.0057	0.0105
	PML- Ψ	0.5558	1.9337	0.2180	0.6582	0.0686	0.2039	1.0331	3.6736	0.3606	1.3925	0.0682	0.3655
	PML- Λ	-0.1169	1.4388	0.1300	0.5271	0.0494	0.1739	-2.3132	4.3659	-0.5331	1.4651	-0.0564	0.2695
	PML-RPM	-0.6515	1.5933	-0.1964	0.5612	-0.0352	0.1280	-0.7598	1.9386	-0.2679	0.7829	-0.0523	0.2549
	PML- Λ^q	0.3133	0.3525	0.0701	0.0741	0.0253	0.0335	0.8506	2.1484	0.0919	0.5831	0.0150	0.1161
	PML-AFXP	0.3147	0.3482	0.0741	0.0781	0.0277	0.0363	0.7317	2.8405	0.1217	0.8471	0.0239	0.1772

Based on 1000 simulated samples. The maximum number of iterations is set to 50.

References

- Aguirregabiria, V. and P. Mira (2007). "Sequential estimation of dynamic discrete games." *Econometrica* 75(1): 1-53.
- Bajari, P., Benkard, C.L., and Levin, J. (2007). "Estimating dynamic models of imperfect competition." *Econometrica* 75(5): 1331-1370.
- Golub, G. and C. Van Loan (1996). *Matrix Computations*, 3rd ed., Baltimore: Johns University Press.
- Heckman, J. (1981). "The incidental parameter problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden. Cambridge: MIT Press.
- Kasahara, H. and K. Shimotsu (2006). Nonparametric Identification and Estimation of Finite Mixture Models of Dynamic Discrete Choices. Mimeographed, Queen's University.
- Kasahara, H. and K. Shimotsu (2008). "Pseudo-likelihood Estimation and Bootstrap Inference for Structural Discrete Markov Decision Models." *Journal of Econometrics* 146: 92-106.
- Kasahara, H. and K. Shimotsu (2009). "Nonparametric identification of finite mixture models of dynamic discrete choices." *Econometrica* 77(1): 135-175.
- Pakes, A., M. Ostrovsky, and S. Berry (2007). "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)." *RAND Journal of Economics* 38(2): 373-399.
- Pesendorfer, M. and P. Schmidt-Dengler (2008). "Asymptotic least squares estimators for dynamic games." *Review of Economic Studies* 75: 901-928.
- Shroff, G. M. and H. B. Keller (1993). "Stabilization of unstable procedures: the recursive projection method." *SIAM Journal of Numerical Analysis* 30(4): 1099-1120.