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

Hiroyuki Kasahara
Department of Economics
University of British Columbia
hkasahar@mail.ubc.ca

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

July 24, 2011

This supplementary appendix contains the following details omitted from the main paper due to space constraints: (A) proof of the results in the paper, (B) auxiliary results and their proof, (C) additional alternative sequential algorithms, (D) the convergence properties of the NPL algorithm for models with unobserved heterogeneity, and (E) additional Monte Carlo results.

1 Appendix

Throughout the appendix, let a.s. abbreviate "almost surely," and let i.o. abbreviate "infinitely often." C denotes a generic positive and finite constant which may take different values in different places. For matrix and nonnegative scalar sequences of random variables $\{X_M, M \geq 1\}$ and $\{Y_M, M \geq 1\}$, respectively, we write $X_M = O(Y_M)$ (or $o(Y_M)$) a.s. if $||X_M|| \leq AY_M$ for some (or all) $A > 0$ a.s.. When Y_M belongs to a family of random variables indexed by $\tau \in T$, we say $X_M = (Y_M(\tau))$ (or $o(Y_M(\tau))$) a.s. *uniformly* in τ if the constant $A > 0$ can be chosen the same for every $\tau \in T$. For instance, in Proposition 7 below, we take $\tau = \tilde{P}_{j-1}$ and $Y_M(\tau) = ||\tilde{P}_{j-1} - \hat{P}_{NPL}||$.

A Proof of the results in the main text

Throughout the proof, the $O()$ terms are uniform, but we suppress the reference to their uniformity for brevity.

A.1 Proof of Proposition 1

We suppress the subscript NPL from $\hat{P}_{NPL}$. Let $b > 0$ be a constant such that $\rho(M_{\Psi_\theta}\Psi_P) + 2b < 1$. From Lemma 5.6.10 of Horn and Johnson (1985), there is a matrix norm $|| \cdot ||_\alpha$ such that

$||M_{\Psi_\theta}\Psi_P||_\alpha \leq \rho(M_{\Psi_\theta}\Psi_P) + b$. Define a vector norm $||\cdot||_\beta$ for $x \in \mathbb{R}^L$ as $||x||_\beta = ||[x\ 0 \ldots 0]||_\alpha$, then a direct calculation gives $||Ax||_\beta = ||A[x\ 0 \ldots 0]||_\alpha \leq ||A||_\alpha ||x||_\beta$ for any matrix A. From the equivalence of vector norms in $\mathbb{R}^L$ (see, for example, Corollary 5.4.5 of Horn and Johnson (1985)), we can restate Proposition 7 in terms of $||\cdot||_\beta$ as follows: there exists $c > 0$ such that $\tilde{P}_j - \hat{P} = M_{\Psi_\theta}\Psi_P(\tilde{P}_{j-1} - \hat{P}) + O(M^{-1/2}||\tilde{P}_{j-1} - \hat{P}||_\beta + ||\tilde{P}_{j-1} - \hat{P}||_\beta^2)$ a.s. holds uniformly in $\tilde{P}_{j-1} \in \{P : ||P - P^0||_\beta < c\}$. We rewrite this statement further so that it is amenable to recursive substitution. First, note that $||M_{\Psi_\theta}\Psi_P(\tilde{P}_{j-1} - \hat{P})||_\beta \leq ||M_{\Psi_\theta}\Psi_P||_\alpha ||\tilde{P}_{j-1} - \hat{P}||_\beta \leq (\rho(M_{\Psi_\theta}\Psi_P) + b)||\tilde{P}_{j-1} - \hat{P}||_\beta$. Second, rewrite the remainder term as $O(M^{-1/2} + ||\tilde{P}_{j-1} - \hat{P}||_\beta)||\tilde{P}_{j-1} - \hat{P}||_\beta$. Set $c < b$, then this term is smaller than $b||\tilde{P}_{j-1} - \hat{P}||_\beta$ a.s. Third, since $\hat{P}$ is consistent, $\{P : ||P - \hat{P}||_\beta < c/2\} \subset \{P : ||P - P^0||_\beta < c\}$ a.s. Consequently, $||\tilde{P}_j - \hat{P}||_\beta \leq (\rho(M_{\Psi_\theta}\Psi_P) + 2b)||\tilde{P}_{j-1} - \hat{P}||_\beta$ holds a.s. for all $\tilde{P}_{j-1}$ in $\{P : ||P - \hat{P}||_\beta < c/2\}$. Because each NPL updating of (θ, P) uses the same pseudo-likelihood function, we may recursively substitute for the $\tilde{P}_j$'s, and hence $\lim_{k\to\infty} \tilde{P}_k = \hat{P}$ a.s. if $||\tilde{P}_0 - \hat{P}||_\beta < c/2$. The stated result follows from applying the equivalence of vector norms in $\mathbb{R}^L$ to $||\tilde{P}_0 - \hat{P}||_\beta$ and $||\tilde{P}_0 - \hat{P}||$ and using the consistency of $\hat{P}$. □

A.2 Proof of Proposition 2

We prove that the stated result holds if $\tilde{P}_{j-1}$ is in a neighborhood $\mathcal{N}_c^{NPL}$ of $\hat{P}_{NPL}$. The stated result then follows from the strong consistency of $\hat{P}_{NPL}$.

Because Proposition 7 holds under the current assumption, we have

$$\tilde{P}_j - \hat{P}_{NPL} = M_{\Psi_\theta}\Psi_P^0(\tilde{P}_{j-1} - \hat{P}_{NPL}) + f(\tilde{P}_{j-1} - \hat{P}_{NPL}), \tag{5}$$

where $|f(x)| \leq C(x^2 + M^{-1/2}|x|)$ a.s. Let $\lambda_1, \lambda_2, \ldots, \lambda_L$ be the eigenvalues of $M_{\Psi_\theta}\Psi_P^0$ such that

$$|\lambda_1| \geq \ldots \geq |\lambda_r| > 1 \geq |\lambda_{r+1}| \geq \ldots \geq |\lambda_L|. \tag{6}$$

For any $\epsilon \neq 0$, we may apply a Jordan decomposition to $M_{\Psi_\theta}\Psi_P^0$ to obtain $H^{-1}M_{\Psi_\theta}\Psi_P^0 H = D + \epsilon J$, where $D = \text{diag}(\lambda_1, \ldots, \lambda_L)$, and J is a matrix with zeros and ones on immediately above the main diagonal (on the superdiagonal) and zeros everywhere else.

Define $y_j = H^{-1}(\tilde{P}_j - \hat{P}_{NPL})$ and $g(y) = H^{-1}f(Hy)$; then multiplying (5) by H^{-1} gives $y_j = (D + \epsilon J)y_{j-1} + g(y_{j-1})$ with $|g(y)| \leq C(|y|^2 + M^{-1/2}|y|)$ a.s. Let y_j^1 denote the first r elements of y, and rewrite this equation as

$$\begin{pmatrix} y_j^1 \\ y_j^2 \end{pmatrix} = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \begin{pmatrix} y_{j-1}^1 \\ y_{j-1}^2 \end{pmatrix} + \epsilon \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix} \begin{pmatrix} y_{j-1}^1 \\ y_{j-1}^2 \end{pmatrix} + \begin{pmatrix} g_1(y_{j-1}) \\ g_2(y_{j-1}) \end{pmatrix}, \tag{7}$$

where y_{j-1}^1 and $g_1(y_{j-1})$ are $r \times 1$ and D_1 and J_1 are $r \times r$ with $D_1 = \text{diag}(\lambda_1, \ldots, \lambda_r)$.

We first show that $H_1(\tilde{P}_j - \hat{P}_{NPL}) > H_1(\tilde{P}_{j-1} - \hat{P}_{NPL})$ by proving that $||y_j^1|| > ||y_{j-1}^1||$ under the stated assumptions. Applying the triangle inequality to the first equation of (7) gives

$$||y_j^1|| \geq ||D_1 y_{j-1}^1|| - ||\epsilon J_1 y_{j-1}^1|| - ||g_1(y_{j-1})||. \tag{8}$$

For the first two terms on the right hand side of (8), we have $||D_1 y_{j-1}^1|| = (\sum_{k=1}^r |\lambda_k|^2 (y_{j-1,k})^2)^{1/2} \geq (1+3\delta)||y_{j-1}^1||$ for some $\delta > 0$ from (6) and $||\epsilon J_1 y_{j-1}^1|| \leq \delta ||y_{j-1}^1||$ by choosing ϵ sufficiently small. For the last term of (8), observe that $\tilde{P}_{j-1} \in V(c)$ if and only if $||y_{j-1}^1|| \leq c||y_{j-1}^2||$, and hence

$$\tilde{P}_{j-1} \notin V(c) \Rightarrow ||y_{j-1}||^2 = ||y_{j-1}^1||^2 + ||y_{j-1}^2||^2 < (1 + c^{-2})||y_{j-1}^1||^2.$$

Therefore, $||g_1(y_{j-1})|| \leq \delta ||y_{j-1}^1||$ holds when $\mathcal{N}_c^{NPL}$ is sufficiently small and M is sufficiently large. It then follows from (8) that $||y_j^1|| \geq (1+\delta)||y_{j-1}^1|| > ||y_{j-1}^1||$.

It remains to show that $\tilde{P}_j \notin V(c)$. Applying the triangle inequality to the second equation of (7) gives $||y_j^2|| \leq ||D_2 y_{j-1}^2|| + ||\epsilon J_2 y_{j-1}^2 + g_2(y_{j-1})||$. For the first term on the right hand side, $||D_2 y_{j-1}^2|| = (\sum_{k=r+1}^L |\lambda_k|^2 (y_{j-1,k})^2)^{1/2} \leq ||y_{j-1}^2||$ from (6). For the second term, similar to the updating of y_j^1, by choosing ϵ and $\mathcal{N}_c^{NPL}$ sufficiently small, we have $||\epsilon J_2 y_{j-1}^2 + g_2(y_{j-1})|| \leq c^{-1}\delta ||y_{j-1}^1||$ if $\tilde{P}_{j-1} \in \mathcal{N}_c^{NPL} \setminus V(c)$ and M is sufficiently large. Therefore, $||y_j^2|| \leq ||y_{j-1}^2|| + c^{-1}\delta||y_{j-1}^1|| < c^{-1}(1+\delta)||y_{j-1}^1|| < c^{-1}||y_j^1||$ a.s., where the last two inequalities use $||y_{j-1}^2|| < c^{-1}||y_{j-1}^1||$ and $||y_{j-1}^1|| < ||y_j^1||$. This proves $\tilde{P}_{j-1} \notin V(c)$. □

A.3 Proof of Proposition 3

First, note that $\tilde{P}_j$ for $j \geq 1$ satisfies restriction (2) because it is generated by $\Psi(\theta, P)$. The restrictions (2)–(3) do not affect the validity of Propositions 1 and 2 because (i) the fixed point constraint in terms of $\Psi(\theta, P)$ and of $\Psi^+(\theta, P^+)$ are equivalent, and (ii) the restrictions (2)–(3) do not affect the order of magnitude of the derivatives of $\Psi(\theta, P)$.

For the equivalence of the eigenvalues, taking the derivative of (3) gives

$$\nabla_{P'}\Psi(\theta, P) = \begin{pmatrix} \nabla_{P^{+\prime}}\Psi^+(\theta, P^+) & 0 \\ -\mathcal{E}\nabla_{P^{+\prime}}\Psi^+(\theta, P^+) & 0 \end{pmatrix} = \begin{pmatrix} U\nabla_{P^{+\prime}}\Psi^+(\theta, P^+) & 0 \end{pmatrix}, \tag{9}$$

and $\nabla_{\theta'}\Psi(\theta, P) = U\nabla_{\theta'}\Psi^+(\theta, P^+)$. Substituting this into $M_{\Psi_\theta}\Psi_P^0$, using $\Psi_\theta^0 = U\Psi_\theta^+$, and rearranging terms give $M_{\Psi_\theta}\Psi_P^0 = [UM_{\Psi_\theta}^+\Psi_{P^+}^+ \vdots 0]$. Therefore, the updating formula of P^+ and P^- is given by $\tilde{P}_j^+ - \hat{P}_{NPL}^+ = M_{\Psi_\theta}^+\Psi_{P^+}^+(\tilde{P}_{j-1}^+ - \hat{P}_{NPL}^+) + O(M^{-1/2}||\tilde{P}_{j-1}^+ - \hat{P}_{NPL}^+|| + ||\tilde{P}_{j-1}^+ - \hat{P}_{NPL}^+||^2)$ a.s. and $\tilde{P}_j^- - \hat{P}_{NPL}^- = -\mathcal{E}(\tilde{P}_j^+ - \hat{P}_{NPL}^+)$, respectively. Finally, the equivalence of the eigenvalues follows from $\det(M_{\Psi_\theta}\Psi_P^0 - \lambda I_{\dim(P)}) = \det(M_{\Psi_\theta}^+\Psi_{P^+}^+ - \lambda I_{\dim(P^+)})\det(-\lambda I_{\dim(P^-)})$ and $\det(\Psi_P^0 - \lambda I_{\dim(P)}) = \det(\Psi_{P^+}^+ - \lambda I_{\dim(P^+)})\det(-\lambda I_{\dim(P^-)})$. □

A.4 Proof of Equation (4)

The notation follows p. 10 of Aguirregabiria and Mira (2007, henceforth AM07). Let $\pi_i^{P_{-i}}(a_i, x; \theta) = \sum_{a_{-i} \in A} P_{-i}(a_{-i}) \Pi_i(a_i, a_{-i}, x; \theta)$ and $e_i^{P_i}(a_i, x; \theta) = E[\epsilon_i(a_i)|x, P_i]$, where $E[\epsilon_i(a_i)|x, P] = E[\epsilon_i(a_i)|x, P_i]$ holds as discussed on pages 9-10 of AM07. Let $V_i(x)$ denote the solution of firm i's integrated Bellman equation:

$$V_i(x) = \int \max_{a_i \in A} \left\{ \pi_i^{P_{-i}}(a_i, x; \theta) + \beta \sum_{x' \in X} V_i(x') f_i^{P_{-i}}(x'|x, a_i) + \epsilon_i(a_i) \right\} g(d\epsilon_i; \theta), \tag{10}$$

where $f_i^{P_{-i}}(x'|x, a_i) = \sum_{a_{-i} \in A} P_{-i}(a_{-i}) f(x'|x, a_i, a_{-i})$. Let $\Pi_i(a_i, a_{-i}; \theta)$, $\pi_i^{P_{-i}}(a_i; \theta)$, $e_i^{P_i}(a_i; \theta)$, $P_i(a_i)$, and V_i denote the vectors of dimension $|X|$ that stack the corresponding state-specific elements of $\Pi_i(a_i, a_{-i}, x; \theta)$, $\pi_i^{P_{-i}}(a_i, x; \theta)$, $e_i^{P_i}(a_i, x; \theta)$, $P_i(a_i|x)$ and $V_i(x)$, respectively. Define the valuation operator as

$$\Gamma_i(\theta, P) = (I - \beta F^P)^{-1} \sum_{a_i \in A} P_i(a_i) * [\pi_i^{P_{-i}}(a_i; \theta) + e_i^{P_i}(a_i; \theta)],$$

where F^P is a matrix with transition probabilities $f^P(x'|x)$, and $*$ denotes the Hadamard product. $\Gamma_i(\theta, P)$ gives the solution of firm i's integrated Bellman equation given θ and P.

Define firm i's best response mapping given V_i and P_{-i} as (cf. equation (15) of AM07)

$$[\Upsilon_i(\theta, V_i, P_{-i})](a_i|x) = \int I \left(a_i = \operatorname*{argmax}_{a \in A} \left\{ \pi_i^{P_{-i}}(a, x; \theta) + \epsilon_i(a) + \beta \sum_{x' \in X} V_i(x') f_i^{P_{-i}}(x'|x, a) \right\} \right) g(d\epsilon_i; \theta), \tag{11}$$

where $f_i^{P_{-i}}(x'|x, a_i) = \sum_{a_{-i} \in A} P_{-i}(a_{-i}) f(x'|x, a_i, a_{-i})$. Then, the mapping Ψ and its Jacobian matrix evaluated at (θ^0, P^0) are given by

$$\Psi(\theta, P) = \begin{pmatrix} \Psi_1(\theta, P) \\ \Psi_2(\theta, P) \end{pmatrix} = \begin{pmatrix} \Upsilon_1(\theta, \Gamma_1(\theta, P), P_2) \\ \Upsilon_2(\theta, \Gamma_2(\theta, P), P_1) \end{pmatrix} \text{ and } \Psi_P^0 = \begin{pmatrix} 0 & \nabla_{P_2'} \Psi_1(\theta^0, P^0) \\ \nabla_{P_1'} \Psi_2(\theta^0, P^0) & 0 \end{pmatrix},$$

where $\nabla_{P_i'} \Psi_i(\theta^0, P^0) = 0$ follows from $\nabla_{P_i'} \Gamma_i(\theta, P_i, P_{-i}) = 0$ (Aguirregabiria and Mira, 2002, Proposition 2). $\square$

A.5 Proof of Proposition 4

For part (a), let V_i^* denote the solution of the Bellman equation (10) given P_{-i}, and let P_i^* be the conditional choice probabilities associated with V_i^*. Since $x_t = (S_t, a_{1,t-1}, a_{2,t-1})$ holds and S_t follows an exogenous process, we may verify under Assumption 3(c) that (i) $V_i^*(S_t, a_{i,t-1}, a_{-i,t-1}^\dagger) = V_i^*(S_t, a_{i,t-1}, a_{-i,t-1}^\ddagger)$ for $a_{-i,t-1}^\dagger \neq a_{-i,t-1}^\ddagger$, (ii) V_i^* does not depend on P_{-i}, (iii) $\Gamma_i(\theta^*, P_i, P_{-i}^\dagger) = \Gamma_i(\theta^*, P_i, P_{-i}^\ddagger) = V_i^*$, and (iv) $\Upsilon_i(\theta^*, V_i^*, P_{-i}^\dagger) = \Upsilon_i(\theta^*, V_i^*, P_{-i}^\ddagger) = P_i^*$

for any $P_{-i}^{\dagger}$ and $P_{-i}^{\ddagger}$ in the space of P_{-i}'s. It follows from (i)-(iv) that the model becomes a single-agent model for each of player and that there exists a unique Markov perfect equilibrium characterized by a unique fixed point $P_i^* = \Psi_i(\theta^*, P_i^*, P_{-i}) = \Upsilon_i(\theta^*, \Gamma_i(\theta^*, P_i^*, P_{-i}), P_{-i})$ for $i = 1, 2$, where the fixed point P_i^* does not depend on the value of P_{-i}.

Define $F(\theta, P) = P - \Psi(\theta, P)$. Since $\nabla_{P'} F(\theta^*, P^*) = I - \nabla_{P'}\Psi(\theta^*, P^*) = I$, we may apply the implicit function theorem to $F(\theta, P) = P - \Psi(\theta, P)$ at $(\theta, P) = (\theta^*, P^*)$ under Assumption 2(b), and there exists an open set $\mathcal{N}_{\theta^*}$ containing θ^*, an open set $\mathcal{N}_{P^*}$ containing P^*, and a unique continuously differentiable function $P(\theta) : \mathcal{N}_{\theta^*} \to \mathcal{N}_{P^*}$ such that $P(\theta) = \Psi(\theta, P(\theta))$ for any $\theta \in \mathcal{N}_{\theta^*}$. Therefore, a Markov perfect equilibrium exists in $\mathcal{N}_{P^*}$ when the true parameter θ^0 is in $\mathcal{N}_{\theta^*}$.

The mapping $\rho(M_{\Psi_\theta}\nabla_{P'}\Psi(\theta, P(\theta)))$ is a continuous function of $\theta \in \mathcal{N}_{\theta^*}$ because $P(\theta)$ is continuous in $\theta \in \mathcal{N}_{\theta^*}$, $\Psi(\theta, P)$ is continuously differentiable by Assumption 2(b), $||M_{\Psi_\theta}|| < \infty$ by Assumption 2(b), and the spectral radius of a matrix is a continuous function of the elements of the matrix. The stated result then follows from $\nabla_{P'}\Psi(\theta^*, P(\theta^*)) = 0$ (Remark 1) and the continuity of $\rho(M_{\Psi_\theta}\nabla_{P'}\Psi(\theta, P(\theta)))$.

For part (b), under Assumption 3(d), $\theta = \theta^\diamond$, and $\beta = 0$, the model becomes a single-agent model for each player. Therefore, repeating the argument for part (a) gives the stated result. □

A.6 Proof of Proposition 5

Let $\lambda = \mathrm{Re}(\lambda) + i\mathrm{Im}(\lambda) = r\cos\theta + ir\sin\theta$ be an eigenvalue of Ψ_P^0. Then, the corresponding eigenvalue of Λ_P is $\lambda(\alpha) = \alpha r\cos\theta + i\alpha r\sin\theta + (1-\alpha)$. Let $f(\alpha) = |\lambda(\alpha)|^2$, then the stated result holds because $f(0) = 1$ and $\nabla_\alpha f(0) = 2(r\cos\theta - 1) < 0$ if $r\cos\theta < 1$ and $\nabla_\alpha f(0) > 0$ if $r\cos\theta > 1$. □

B Auxiliary results and their proof

Proposition 6 strengthens the weak consistency result of Proposition 2 of AM07 to strong consistency. Proposition 7 describes how an NPL step updates θ and P.

Proposition 6 *Suppose that Assumption 1 holds. Then,* $(\hat{\theta}_{NPL}, \hat{P}_{NPL}) \to (\theta^0, P^0)$ *a.s.*

Proof Proposition 2 of AM07 showed weak consistency of $(\hat{\theta}_{NPL}, \hat{P}_{NPL})$. Therefore, strong consistency $(\hat{\theta}_{NPL}, \hat{P}_{NPL})$ follows from strengthening "in probability" and "with probability approaching 1" statements in Steps 2-5 of the proof of Proposition 2 of AM07 to "almost surely."

First, observe that AM07 (pp. 44-45) showed that $Q_M(\theta, P)$ converges to $Q_0(\theta, P)$ a.s. and uniformly in (θ, P). Thus, the events A_M's defined in Steps 2-3 and 5 of AM07 satisfy $\Pr(A_M^c \ i.o.) = 0$. In Step 2, we can strengthen $\Pr((\theta_M^*, P_M^*) \in \Im) \to 1$ of AM07 to $(\theta_M^*, P_M^*) \in \Im$

a.s. because AM07 (pp. 46-47) showed $A_M \Rightarrow \{(\theta_M^*, P_M^*) \in \Im\}$ and we have $\Pr(A_M^c \; i.o.) = 0$. In Step 3, an analogous argument strengthens $\Pr(\sup_{P \in N(P^0)} ||\tilde{\theta}_M(P) - \tilde{\theta}_0(P)|| < \varepsilon) \to 1$ in AM07 to $\sup_{P \in N(P^0)} ||\tilde{\theta}_M(P) - \tilde{\theta}_0(P)|| < \varepsilon$ a.s. Similarly, we can strengthen "with probability approaching 1" in Steps 4 and 5 to "almost surely," and strong consistency of the NPL estimator follows. □

Proposition 7 *Suppose that Assumption 2 holds. Then, there exists a neighborhood $\mathcal{N}_1$ of P^0 such that $\tilde{\theta}_j - \hat{\theta}_{NPL} = O(||\tilde{P}_{j-1} - \hat{P}_{NPL}||)$ a.s. and $\tilde{P}_j - \hat{P}_{NPL} = M_{\Psi_\theta} \Psi_P^0 (\tilde{P}_{j-1} - \hat{P}_{NPL}) + O(M^{-1/2} ||\tilde{P}_{j-1} - \hat{P}_{NPL}|| + ||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2)$ a.s. uniformly in $\tilde{P}_{j-1} \in \mathcal{N}_1$.*

Proof We suppress the subscript NPL from $\hat{P}_{NPL}$ and $\hat{\theta}_{NPL}$. For $\epsilon > 0$, define a neighborhood $\mathcal{N}(\epsilon) = \{(\theta, P) : ||\theta - \theta^0|| + ||P - P^0|| < \epsilon\}$. Then, there exists $\epsilon_1 > 0$ such that $\mathcal{N}(\epsilon_1) \subset \mathcal{N}$ and $\sup_{(\theta,P) \in \mathcal{N}(\epsilon_1)} ||\nabla_{\theta\theta'} Q_0(\theta, P)^{-1}|| < \infty$ because $\nabla_{\theta\theta'} Q_0(\theta, P)$ is continuous and $\nabla_{\theta\theta'} Q_0(\theta^0, P^0)$ is nonsingular.

First, we assume $(\tilde{\theta}_j, \tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ and derive the stated representation of $\tilde{\theta}_j - \hat{\theta}$ and $\tilde{P}_j - \hat{P}$. We later show $(\tilde{\theta}_j, \tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ a.s. if $\mathcal{N}_1$ is sufficiently small. The first order condition for $\tilde{\theta}_j$ is $\nabla_\theta Q_M(\tilde{\theta}_j, \tilde{P}_{j-1}) = 0$. Expanding it around $(\hat{\theta}, \hat{P})$ and using $\nabla_\theta Q_M(\hat{\theta}, \hat{P}) = 0$ gives

$$0 = \nabla_{\theta\theta'} Q_M(\bar{\theta}, \bar{P})(\tilde{\theta}_j - \hat{\theta}) + \nabla_{\theta P'} Q_M(\bar{\theta}, \bar{P})(\tilde{P}_{j-1} - \hat{P}), \tag{12}$$

where $(\bar{\theta}, \bar{P})$ lie between $(\tilde{\theta}_j, \tilde{P}_{j-1})$ and $(\hat{\theta}, \hat{P})$. Write (12) as $\tilde{\theta}_j - \hat{\theta} = -\nabla_{\theta\theta'} Q_M(\bar{\theta}, \bar{P})^{-1} \nabla_{\theta P'} Q_M(\bar{\theta}, \bar{P})(\tilde{P}_{j-1} - \hat{P})$, then the stated uniform bound of $\tilde{\theta}_j - \hat{\theta}$ follows because (i) $(\bar{\theta}, \bar{P}) \in \mathcal{N}(\epsilon_1)$ a.s. since $(\tilde{\theta}_j, \tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ and $(\hat{\theta}, \hat{P})$ is strongly consistent from Proposition 6, and (ii) $\sup_{(\theta,P) \in \mathcal{N}(\epsilon_1)} ||\nabla_{\theta\theta'} Q_M(\theta, P)^{-1} \nabla_{\theta P'} Q_M(\theta, P)|| = O(1)$ a.s. since $\sup_{(\theta,P) \in \mathcal{N}(\epsilon_1)} ||\nabla_{\theta\theta'} Q_0(\theta, P)^{-1}|| < \infty$ and $\sup_{(\theta,P) \in \mathcal{N}} ||\nabla^2 Q_M(\theta, P) - \nabla^2 Q_0(\theta, P)|| = o(1)$ a.s., where the latter follows from Kolomogorov's strong law of large numbers and Theorem 2 and Lemma 1 of Andrews (1992).

For the bound of $\tilde{P}_j - \hat{P}$, first we collect the following results, which follow from the Taylor expansion around (θ^0, P^0), root-M consistency of $(\hat{\theta}, \hat{P})$, and the information matrix equality.

$$\begin{aligned} &\nabla_{\theta\theta'} Q_M(\hat{\theta}, \hat{P}) = -\Omega_{\theta\theta} + O(M^{-1/2}) \; a.s., && \nabla_{\theta P'} Q_M(\hat{\theta}, \hat{P}) = -\Omega_{\theta P} + O(M^{-1/2}) \; a.s., \\ &\nabla_{\theta'} \Psi(\hat{\theta}, \hat{P}) = \Psi_\theta^0 + O(M^{-1/2}) \; a.s., && \nabla_{P'} \Psi(\hat{\theta}, \hat{P}) = \Psi_P^0 + O(M^{-1/2}) \; a.s. \end{aligned} \tag{13}$$

Expand the right hand side of $\tilde{P}_j = \Psi(\tilde{\theta}_j, \tilde{P}_{j-1})$ twice around $(\hat{\theta}, \hat{P})$ and use $\Psi(\hat{\theta}, \hat{P}) = \hat{P}$ and $\tilde{\theta}_j - \hat{\theta} = O(||\tilde{P}_{j-1} - \hat{P}||)$ a.s., then we obtain $\tilde{P}_j - \hat{P} = \nabla_{\theta'} \Psi(\hat{\theta}, \hat{P})(\tilde{\theta}_j - \hat{\theta}) + \nabla_{P'} \Psi(\hat{\theta}, \hat{P})(\tilde{P}_{j-1} - \hat{P}) + O(||\tilde{P}_{j-1} - \hat{P}||^2)$ a.s. since $\sup_{(\theta,P) \in \mathcal{N}(\epsilon_1)} \nabla^3 \Psi(\theta, P) < \infty$. Applying (13) and $\tilde{\theta}_j - \hat{\theta} = O(||\tilde{P}_{j-1} - \hat{P}||)$ a.s. to the right hand side gives

$$\tilde{P}_j - \hat{P} = \Psi_\theta(\tilde{\theta}_j - \hat{\theta}) + \Psi_P^0(\tilde{P}_{j-1} - \hat{P}) + O(||\tilde{P}_{j-1} - \hat{P}||^2) + O(M^{-1/2} ||\tilde{P}_{j-1} - \hat{P}||) \; a.s. \tag{14}$$

We proceed to refine (12) to write $\tilde{\theta}_j - \hat{\theta}$ in terms of $\tilde{P}_{j-1} - \hat{P}$ and substitute it into (14).

Expanding $\nabla_{\theta\theta'} Q_M(\bar{\theta},\bar{P})$ in (12) around $(\hat{\theta},\hat{P})$, noting that $||\bar{\theta}-\hat{\theta}|| \le ||\tilde{\theta}_j-\hat{\theta}||$ and $||\bar{P}-\hat{P}|| \le ||\tilde{P}_{j-1}-\hat{P}||$, and using $\tilde{\theta}_j-\hat{\theta}=O(||\tilde{P}_{j-1}-\hat{P}||)$ a.s., we obtain $\nabla_{\theta\theta'} Q_M(\bar{\theta},\bar{P}) = \nabla_{\theta\theta'} Q_M(\hat{\theta},\hat{P}) + O(||\tilde{P}_{j-1}-\hat{P}||)$ a.s. Further, applying (13) gives $\nabla_{\theta\theta'} Q_M(\bar{\theta},\bar{P}) = -\Omega_{\theta\theta} + O(M^{-1/2}) + O(||\tilde{P}_{j-1}-\hat{P}||)$ a.s. Similarly, we obtain $\nabla_{\theta P'} Q_M(\bar{\theta},\bar{P}) = -\Omega_{\theta P} + O(M^{-1/2}) + O(||\tilde{P}_{j-1}-\hat{P}||)$ a.s. Using these results, refine (12) as $\tilde{\theta}_j-\hat{\theta} = -\Omega_{\theta\theta}^{-1}\Omega_{\theta P}(\tilde{P}_{j-1}-\hat{P}) + O(M^{-1/2}||\tilde{P}_{j-1}-\hat{P}|| + ||\tilde{P}_{j-1}-\hat{P}||^2)$ a.s. Substituting this into (14) in conjunction with $\Omega_{\theta\theta}^{-1}\Omega_{\theta P} = (\Psi_\theta^{0\prime}\Delta_P\Psi_\theta^0)^{-1}\Psi_\theta^{0\prime}\Delta_P\Psi_P^0$ gives the stated result.

It remains to show $(\tilde{\theta}_j,\tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ a.s. if $\mathcal{N}_1$ is sufficiently small. Let $\mathcal{N}_\theta \equiv \{\theta : ||\theta-\theta^0|| < \epsilon_1/2\}$ and define $\Delta = Q_0(\theta^0,P^0) - \sup_{\theta\in\mathcal{N}_\theta^c\cap\Theta} Q_0(\theta,P^0) > 0$, where the last inequality follows from information inequality, compactness of $\mathcal{N}_\theta^c\cap\Theta$, and continuity of $Q_0(\theta,P)$. It follows that $\{\tilde{\theta}_j \notin \mathcal{N}_\theta\} \Rightarrow \{Q_0(\theta^0,P^0) - Q_0(\tilde{\theta}_j,P^0) \ge \Delta\}$. Further, observe that $Q_0(\theta^0,P^0) - Q_0(\tilde{\theta}_j,P^0) \le Q_M(\theta^0,\tilde{P}_{j-1}) - Q_M(\tilde{\theta}_j,\tilde{P}_{j-1}) + 2\sup_{\theta\in\Theta}|Q_0(\theta,P^0)-Q_0(\theta,\tilde{P}_{j-1})| + 2\sup_{(\theta,P)\in\Theta\times B_P}|Q_M(\theta,P)-Q_0(\theta,P)| \le 2\sup_{\theta\in\Theta}|Q_0(\theta,P^0)-Q_0(\theta,\tilde{P}_{j-1})| + 2\sup_{(\theta,P)\in\Theta\times B_P}|Q_M(\theta,P)-Q_0(\theta,P)|$, where the second inequality follows from the definition of $\tilde{\theta}_j$. From continuity of $Q_0(\theta,P)$, there exists $\epsilon_\Delta > 0$ such that the first term on the right is smaller than $\Delta/2$ if $||P^0-\tilde{P}_{j-1}|| \le \epsilon_\Delta$. The second term on the right is $o(1)$ a.s. from Kolomogorov's strong law of large numbers and Theorem 2 and Lemma 1 of Andrews (1992). Hence, $\Pr(\tilde{\theta}_j \notin \mathcal{N}_\theta\ i.o.) = 0$ if $||P^0-\tilde{P}_{j-1}|| \le \epsilon_\Delta$, and setting $\mathcal{N}_1 = \{P : ||P-P^0|| \le \min\{\epsilon_1/2,\epsilon_\Delta\}\}$ gives $(\tilde{\theta}_j,\tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ a.s. $\square$

C Additional alternative sequential algorithms

C.1 Recursive Projection Method

In this subsection, we construct a mapping that has a better local contraction property than Ψ, building upon the Recursive Projection Method (RPM) of Shroff and Keller (1993) (henceforth SK).

First, fix θ. Let P_θ denote an element of $M_\theta = \{P \in B_P : P = \Psi(\theta,P)\}$ so that P_θ is one of the fixed points of $\Psi(\theta,P)$ when there are multiple fixed points. Consider finding P_θ by iterating $P_j = \Psi(P_{j-1},\theta)$ starting from a neighborhood of P_θ. If some eigenvalues of $\nabla_{P'}\Psi(\theta,P_\theta)$ are outside the unit circle, this iteration does not converge to P_θ in general. Suppose that, counting multiplicity, there are r eigenvalues of $\nabla_{P'}\Psi(\theta,P_\theta)$ that are larger than $\delta \in (0,1)$ in modulus:

$$|\lambda_1| \ge \cdots \ge |\lambda_r| > \delta \ge |\lambda_{r+1}| \ge \cdots \ge |\lambda_L|. \qquad (15)$$

Define $\mathbb{P} \subseteq \mathbb{R}^L$ as the maximum invariant subspace of $\nabla_{P'}\Psi(\theta,P_\theta)$ belonging to $\{\lambda_k\}_{k=1}^r$, and let $\mathbb{Q} \equiv \mathbb{R}^L - \mathbb{P}$ be the orthogonal complement of $\mathbb{P}$. Let Π_θ denote the orthogonal projector from $\mathbb{R}^L$ on $\mathbb{P}$. We may write $\Pi_\theta = Z_\theta Z_\theta'$, where $Z_\theta \in \mathbb{R}^{L\times r}$ is an orthonormal basis of $\mathbb{P}$. Then, for each $P \in \mathbb{R}^L$, we have the unique decomposition $P = u + v$, where $u \equiv \Pi_\theta P \in \mathbb{P}$ and $v \equiv (I-\Pi_\theta)P \in \mathbb{Q}$.

Now apply Π_θ and $I - \Pi_\theta$ to $P = \Psi(\theta, P)$, and decompose the system as follows:

$$\begin{aligned} u &= f(u, v, \theta) \equiv \Pi_\theta \Psi(\theta, u + v), \\ v &= g(u, v, \theta) \equiv (I - \Pi_\theta)\Psi(\theta, u + v). \end{aligned}$$

For a given P_{j-1}, decompose it into $u_{j-1} = \Pi_\theta P_{j-1}$ and $v_{j-1} = (I - \Pi_\theta)P_{j-1}$. Since $g(u, v, \theta)$ is contractive in v (see Lemma 2.10 of SK), we can update v_{j-1} by the recursion $v_j = g(u, v_{j-1}, \theta)$. On the other hand, when the dominant eigenvalue of Ψ_P^0 is outside the unit circle, the recursion $u_j = f(u_{j-1}, v, \theta)$ cannot be used to update u_{j-1} because $f(u, v, \theta)$ is not a contraction in u. Instead, the RPM performs a single Newton step on the system $u = f(u, v, \theta)$, leading to the following updating procedure:

$$\begin{aligned} u_j &= u_{j-1} + (I - \Pi_\theta \nabla_{P'}\Psi(\theta, P_{j-1})\Pi_\theta)^{-1}(f(u_{j-1}, v_{j-1}, \theta) - u_{j-1}) \equiv h(u_{j-1}, v_{j-1}, \theta), \\ v_j &= g(u_{j-1}, v_{j-1}, \theta). \end{aligned} \tag{16}$$

Lemma 3.11 of SK shows that the spectral radius of the Jacobian of the stabilized iteration (16) is no larger than δ, and thus the iteration $P_j = h(\Pi_\theta P_{j-1}, (I - \Pi_\theta)P_{j-1}, \theta) + g(\Pi_\theta P_{j-1}, (I - \Pi_\theta)P_{j-1}, \theta)$ converges locally. In the following, we develop a sequential algorithm building upon the updating procedure (16).

Let $\Pi(\theta, P)$ be the orthogonal projector from $\mathbb{R}^L$ onto the maximum invariant subspace of $\nabla_{P'}\Psi(\theta, P)$ belonging to its r largest (in modulus) eigenvalues, counting multiplicity. Define u^*, v^*, $h^*(u^*, v^*, \theta)$, and $g^*(u^*, v^*, \theta)$ by replacing Π_θ in u, v, $h(u, v, \theta)$, and $g(u, v, \theta)$ with $\Pi(\theta, P)$, and define

$$\begin{aligned} \Gamma(\theta, P) &\equiv h^*(u^*, v^*, \theta) + g^*(u^*, v^*, \theta) \\ &= \Psi(\theta, P) + [(I - \Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P))^{-1} - I]\Pi(\theta, P)(\Psi(\theta, P) - P). \end{aligned} \tag{17}$$

P^0 is a fixed point of $\Gamma(\theta^0, P)$, because all the fixed points of $\Psi(\theta, P)$ are also fixed points of $\Gamma(\theta, P)$. The following proposition shows two important properties of $\Gamma(\theta, P)$: local contraction and the equivalence of fixed points of $\Gamma(\theta, P)$ and $\Psi(\theta, P)$.

Proposition 8 *(a) Suppose that $I - \Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P)$ is nonsingular and hence $\Gamma(\theta, P)$ is well-defined. Then $\Gamma(\theta, P)$ and $\Psi(\theta, P)$ have the same fixed points; i.e., $\Gamma(\theta, P) = P$ if and only if $\Psi(\theta, P) = P$. (b) $\rho(\nabla_{P'}\Gamma(\theta^0, P^0)) \leq \delta^0$, where δ^0 is defined by (15) in terms of the eigenvalues of $\nabla_{P'}\Psi(\theta^0, P^0)$. Hence, $\Gamma(\theta, P)$ is locally contractive.*

Define $Q_M^\Gamma(\theta, P) \equiv M^{-1}\sum_{m=1}^M \sum_{t=1}^T \ln \Gamma(\theta, P)(a_{mt}|x_{mt})$. Define an *RPM fixed point* as a pair $(\check{\theta}, \check{P})$ that satisfies $\check{\theta} = \arg\max_{\theta \in \Theta} Q_M^\Gamma(\theta, \check{P})$ and $\check{P} = \Gamma(\check{\theta}, \check{P})$. The *RPM estimator*, denoted by $(\hat{\theta}_{RPM}, \hat{P}_{RPM})$, is defined as the RPM fixed point with the highest value of the pseudo likelihood among all the RPM fixed points. Define the *RPM algorithm* by the same sequential algorithm as the NPL algorithm except that it uses $\Gamma(\theta, P)$ in place of $\Psi(\theta, P)$.

Proposition 9 shows the asymptotic properties of the RPM estimator and the convergence properties of the RPM algorithm. Define the RPM counterparts of $\tilde{\theta}_0(P)$, $\phi_0(P)$, $\Omega_{\theta\theta}$, and $\Omega_{\theta P}$ as $\tilde{\theta}_0^\Gamma(P) \equiv \arg\max_{\theta\in\Theta} EQ_M^\Gamma(\theta, P)$, $\phi_0^\Gamma(P) = \Gamma(\tilde{\theta}_0^\Gamma(P), P)$, $\Omega_{\theta\theta}^\Gamma \equiv E(\nabla_\theta s_{mt}^\Gamma \nabla_{\theta'} s_{mt}^\Gamma)$, and $\Omega_{\theta P}^\Gamma \equiv E(\nabla_\theta s_{mt}^\Gamma \nabla_{P'} s_{mt}^\Gamma)$, where $s_{mt}^\Gamma = \sum_{t=1}^T \ln \Gamma(\theta^0, P^0)(a_{mt}|x_{mt})$. Define $\Gamma_P^0 \equiv \nabla_{P'}\Gamma(\theta^0, P^0)$ and $\Gamma_\theta^0 \equiv \nabla_{\theta'}\Gamma(\theta^0, P^0)$. We outline the assumptions first.

Assumption 4 *(a) Assumption 1 holds. (b) $\Psi(\theta, P)$ is four times continuously differentiable in $\mathcal{N}$. (c) $I - \Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P)$ is nonsingular. (d) $\Gamma(\theta, P) > 0$ for any $(a, x) \in A \times X$ and $(\theta, P) \in \Theta \times B_P$. (e) The operator $\phi_0^\Gamma(P) - P$ has a nonsingular Jacobian matrix at P^0.*

Assumption 4(c) is required for $\Gamma(\theta, P)$ to be well-defined. It would be possible to drop Assumption 4(d) by considering a trimmed version of $\Gamma(\theta, P)$, but for brevity we do not pursue it.

Proposition 9 *Suppose that Assumption 4 holds. Then (a) $\hat{P}_{RPM} - P^0 = O(M^{-1/2})$ a.s. and $M^{-1/2}(\hat{\theta}_{RPM} - \theta^0) \to_d N(0, V_{RPM})$, where $V_{RPM} = [\Omega_{\theta\theta}^\Gamma + \Omega_{\theta P}^\Gamma(I - \Gamma_P^0)^{-1}\Gamma_\theta^0]^{-1}\Omega_{\theta\theta}^\Gamma\{[\Omega_{\theta\theta}^\Gamma + \Omega_{\theta P}^\Gamma(I - \Gamma_P^0)^{-1}\Gamma_\theta^0]^{-1}\}'$. (b) Suppose we obtain $(\tilde{\theta}_j, \tilde{P}_j)$ from $\tilde{P}_{j-1}$ by the RPM algorithm. Then, there exists a neighborhood $\mathcal{N}_1$ of P^0 such that $\tilde{\theta}_j - \hat{\theta}_{RPM} = O(||\tilde{P}_{j-1} - \hat{P}_{RPM}||)$ and $\tilde{P}_j - \hat{P}_{RPM} = M_{\Gamma_\theta}\Gamma_P^0(\tilde{P}_{j-1} - \hat{P}_{RPM}) + O(M^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{RPM}|| + ||\tilde{P}_{j-1} - \hat{P}_{RPM}||^2)$ a.s. uniformly in $\tilde{P}_{j-1} \in \mathcal{N}_1$, where $M_{\Gamma_\theta} \equiv I - \Gamma_\theta^0(\Gamma_\theta^{0\prime}\Delta_P\Gamma_\theta^0)^{-1}\Gamma_\theta^{0\prime}\Delta_P$.*

C.2 Approximate RPM algorithm

Implementing the RPM algorithm is costly because it requires evaluating $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ for all the trial values of θ. We reduce the computational burden by evaluating $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ outside the optimization routine by using a preliminary estimate of θ. This modification has only a second-order effect on the convergence of the algorithm because the derivatives of $\Gamma(\theta, P)$ with respect to $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ are zero when evaluated at $P = \Psi(\theta, P)$; see the second term in (17). Let η be a preliminary estimate of θ. Replacing θ in $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ with η, we define the following mapping:

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

Once $\Pi(\eta, P)$ and $\nabla_{P'}\Psi(\eta, P)$ are computed, the computational cost of evaluating $\Gamma(\theta, P, \eta)$ across different values of θ would be similar to that of evaluating $\Psi(\theta, P)$.

Let $(\tilde{\theta}_0, \tilde{P}_0)$ be an initial estimator of (θ^0, P^0). For instance, $\tilde{\theta}_0$ can be the PML estimator. The *approximate RPM algorithm* iterates the following steps until $j = k$:

Step 1: Given $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\bar{\Theta}_j} M^{-1}\sum_{m=1}^M \sum_{t=1}^T \ln \Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_{mt}|x_{mt})$, where $\bar{\Theta}_j \equiv \{\theta \in \Theta : \Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a|x) \in [\xi, 1-\xi] \text{ for all } (a,x) \in A \times X\}$ for an arbitrary small $\xi > 0$. We impose this restriction in order to avoid computing $\ln(0)$.[1]

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

The following proposition shows that the approximate RPM algorithm achieves the same convergence rate as the original RPM algorithm in the first order.

Proposition 10 *Suppose that Assumption 4 holds and we obtain* $(\tilde{\theta}_j, \tilde{P}_j)$ *from* $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ *by the approximate RPM algorithm. Then, there exists a neighborhood* $\mathcal{N}_2$ *of* (θ^0, P^0) *such that* $\tilde{\theta}_j - \hat{\theta}_{RPM} = O(||\tilde{P}_{j-1} - \hat{P}_{RPM}|| + M^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}||^2)$ *a.s. and* $\tilde{P}_j - \hat{P}_{RPM} = M_{\Gamma_\theta}\Gamma_P^0(\tilde{P}_{j-1} - \hat{P}_{RPM}) + O(M^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}||^2 + M^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{RPM}|| + ||\tilde{P}_{j-1} - \hat{P}_{RPM}||^2)$ *a.s. uniformly in* $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}_2$.

By choosing δ sufficiently small, the dominant eigenvalue of $M_{\Gamma_\theta}\Gamma_P$ lies inside the unit circle, and the approximate RPM algorithm can converge to a consistent estimator even when the NPL algorithm diverges away from the true value. The following proposition states the local convergence of the approximate RPM algorithm when $\rho(M_{\Gamma_\theta}\Gamma_P) < 1$.

Proposition 11 *Suppose that Assumption 4 holds,* $\rho(M_{\Gamma_\theta}\Gamma_P^0) < 1$*, and* $\{\tilde{\theta}_k, \tilde{P}_k\}$ *is generated by the approximate RPM algorithm starting from* $(\tilde{\theta}_0, \tilde{P}_0)$*. Then, there exists a neighborhood* $\mathcal{N}_3$ *of* (θ^0, P^0) *such that, for any initial value* $(\tilde{\theta}_0, \tilde{P}_0) \in \mathcal{N}_3$*, we have* $\lim_{k\to\infty}(\tilde{\theta}_k, \tilde{P}_k) = (\hat{\theta}_{RPM}, \hat{P}_{RPM})$ *a.s.*

C.3 Numerical implementation of the approximate RPM algorithm

Implementing the approximate RPM algorithm 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 $j = 1, \ldots, k$. 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 approximate 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 alternative 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 10. Therefore, if we use a root-M 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 $O(M^{-1})$ away a.s. from the corresponding estimators generated by the approximate RPM algorithm.

[1]In practice, we may consider a penalized objective function by truncating $\Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ so that it takes a value between ξ and $1-\xi$, and adding a penalty term that penalizes θ such that $\Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \notin [\xi, 1-\xi]$.

Corollary 1 *Suppose that Assumption 4 holds. Let* (θ^*, P^*) *be a strongly 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 a neighborhood* $\mathcal{N}_4$ *of* (θ^0, P^0) *such that* $\tilde{\theta}_j - \hat{\theta}_{RPM} = O(||\tilde{P}_{j-1} - \hat{P}_{RPM}|| + r_{Mj})$ *a.s. and* $\tilde{P}_j - \hat{P}_{RPM} = M_{\Gamma_\theta}\Gamma_P(\tilde{P}_{j-1} - \hat{P}_{RPM}) + O(M^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{RPM}|| + ||\tilde{P}_{j-1} - \hat{P}_{RPM}||^2 + r_{Mj})$ *a.s. uniformly in* $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}_4$*, where* $r_{Mj} = M^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}||^2 + M^{-1/2}||\theta^* - \hat{\theta}_{RPM}|| + ||\theta^* - \hat{\theta}_{RPM}||^2 + M^{-1/2}||P^* - \hat{P}_{RPM}|| + ||P^* - \hat{P}_{RPM}||^2$*.*

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 $\tilde{\Pi}_{j-1} = \Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, $\tilde{Z}_{j-1} = Z(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, 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 $\tilde{Z}_{j-1} = [\tilde{z}_{j-1}^1, \ldots, \tilde{z}_{j-1}^r]$ and $\xi > 0$. The ith 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 $(r+1)$ 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 $r \times r$ 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 r 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 "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 modulus is larger than δ. Let $\{\tilde{\lambda}_{0,1}, \ldots, \tilde{\lambda}_{0,r}\}$ denote them.[2] (b) Find the eigenvectors of $\tilde{\Psi}_{P,0}$ associated with $\tilde{\lambda}_{0,1}, \ldots, \tilde{\lambda}_{0,r}$. (c) Using Gram-Schmidt orthogonalization, compute an orthonormal basis of the space spanned by these eigenvectors. Let $\{\tilde{z}_0^1, \ldots, \tilde{z}_0^r\}$ 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 $\tilde{Z}_0 = [\tilde{z}_0^1, \ldots, \tilde{z}_0^r]$.

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 $\tilde{\theta}_j = \arg\max_{\theta \in \Theta_j} M^{-1}\sum_{m=1}^M \sum_{t=1}^T \ln \Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}, \tilde{Z}_{j-1})(a_{mt}|x_{mt})$, 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 $\tilde{\Psi}_{P,j-1} \equiv \nabla_{P'}\Psi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$.

[2] Computing the r 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.

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 $\tilde{Z}_j$ =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 $\tilde{\Psi}_{P,j}\tilde{z}^i_{j-1} \approx (1/\xi)[\Psi(\tilde{\theta}_j, \tilde{P}_j + \xi\tilde{z}^i_{j-1}) - \Psi(\tilde{\theta}_j, \tilde{P}_j)]$ for small $\xi > 0$ with $\tilde{Z}_{j-1} = [\tilde{z}^1_{j-1}, \ldots, \tilde{z}^r_{j-1}]$. (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 $\tilde{\Psi}_{P,j}\tilde{z}^i_j \approx (1/\xi)[\Psi(\tilde{\theta}_j, \tilde{P}_j + \xi\tilde{z}^i_j) - \Psi(\tilde{\theta}_j, \tilde{P}_j)]$. (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.

C.4 Applying RPM to the example of Pesendorfer and Schmidt-Dengler (2010)

This subsection illustrates how the RPM algorithm can be applied to the example of Pesendorfer and Schmidt-Dengler (2010). We first derive the relation between $(\Gamma^+_\theta, \Gamma^+_{P^+})$ and $(\Psi^+_\theta, \Psi^+_{P^+})$. Define $\Pi^+(\theta, P^+)$ as the orthogonal projector from $\mathbb{R}^{\dim(P^+)}$ onto the maximum invariant subspace of $\nabla_{P^{+\prime}}\Psi(\theta, P^+)$ belonging to its r largest (in modulus) eigenvalues, and let $Z(\theta, P^+)$ be an orthonormal basis of the column space of $\Pi^+(\theta, P^+)$ so that $\Pi^+(\theta, P^+) = Z(\theta, P^+)Z(\theta, P^+)'$. From the proof of Proposition 6, we have $\Gamma^+(\theta, P^+) - P^+ = A(\theta, P^+)(\Psi^+(\theta, P^+) - P^+)$, where $A(\theta, P^+) = Z(\theta, P^+)[I - Z(\theta, P^+)'\nabla_{P^{+\prime}}\Psi(\theta, P^+)Z(\theta, P^+)]^{-1}Z(\theta, P^+)' + I - \Pi(\theta, P^+)$. Consequently, $\Gamma^+_\theta = A(\theta^0, P^{0+})\Psi^+_\theta$ and $\Gamma^+_{P^+} = A(\theta^0, P^{0+})(\Psi^+_{P^+} - I) + I$.

We proceed to derive $M^+_{\Gamma_\theta}\Gamma^+_{P^+}$. Recall

$$\Psi^+_\theta = p^0\mathbf{1}_2, \quad \Psi^+_{P^+} = \begin{pmatrix} 0 & \theta^0 \\ \theta^0 & 0 \end{pmatrix}, \quad M^+_{\Psi_\theta} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

The eigenvectors and eigenvalues of $\Psi^+_{P^+}$ are given by

$$z_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_1 = \theta^0, \quad z_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \lambda_2 = -\theta^0.$$

Because the eigenvector z_1 is annihilated by $M^+_{\Psi_\theta}$, we may take $Z(\theta^0, P^{0+}) = z_2$. Suppress (θ^0, P^{0+}) from $Z(\theta^0, P^{0+})$, $\Pi(\theta^0, P^{0+})$, and $A(\theta^0, P^{0+})$. Since Z is the eigenvector of Ψ^+_{P+} with eigenvalue $-\theta^0$, we have $Z'\Psi^+_{P+}Z = -\theta^0 Z'Z = -\theta^0$ and hence

$$A = Z(1 - Z'\Psi^+_{P+}Z)^{-1}Z' + (I - \Pi) = (1+\theta^0)^{-1}\Pi + (I - \Pi) = I - \theta^0(1+\theta^0)^{-1}\Pi.$$

Because Ψ^+_{P+} is symmetric, we may apply the eigenvalue decomposition to it and write $\Psi^+_{P+} = \theta^0 z_1 z_1' - \theta^0 z_2 z_2' = \theta^0 z_1 z_1' - \theta^0 \Pi$. In view of $Az_1 = z_1$ and $A\Pi = (1+\theta^0)^{-1}\Pi$, we have $\Gamma^+_{P+} = A(\Psi^+_{P+} - I) + I = \theta^0 z_1 z_1' - \theta^0(1+\theta^0)^{-1}\Pi - A + I = \theta^0 z_1 z_1'$. Further, from $\Psi^+_\theta = p^0\sqrt{2}z_1$, we have $\Gamma^+_\theta = A\Psi^+_\theta = \Psi^+_\theta$ and hence $M^+_{\Gamma_\theta} = I - z_1(z_1 z_1')^{-1}z_1'$. It follows that $M^+_{\Gamma_\theta}\Gamma^+_{P+} = 0$, and the local convergence condition holds.

C.5 q-NPL algorithm and approximate q-NPL algorithm

When the spectral radius of Λ^0_P or Ψ^0_P is smaller than but close to 1, the convergence of the NPL algorithm could be slow and the generated sequence could behave erratically. Furthermore, in such a case, the efficiency loss from using the NPL estimator compared to the MLE is substantial. To overcome these problems, consider a q-fold operator of Λ as

$$\Lambda^q(\theta, P) \equiv \underbrace{\Lambda(\theta, (\Lambda(\theta, \ldots \Lambda(\theta, \Lambda}_{q \text{ times}}(\theta, P)) \ldots))).$$

We may define $\Gamma^q(\theta, P)$ and $\Psi^q(\theta, P)$ analogously. Define the *q-NPL (q-RPM) algorithm* by using a q-fold operator Λ^q, Γ^q, and Ψ^q in place of Λ, Γ, or Ψ in the original NPL (RPM) algorithm. In the following, we focus on Λ^q but the same argument applies to Γ^q and Ψ^q.

If the sequence of estimators generated by the q-NPL algorithm converges, its limit satisfies $\check{\theta} = \arg\max_{\theta\in\Theta} M^{-1}\sum_{m=1}^{M}\sum_{t=1}^{T} \ln \Lambda^q(\theta, \check{P})(a_{mt}|x_{mt})$ and $\check{\theta} = \Lambda^q(\check{\theta}, \check{P})$. Among the pairs $(\hat{\theta}, \hat{P})$ that satisfy these two conditions, the one that maximizes the value of the pseudo likelihood is called the *q-NPL estimator* and denoted by $(\hat{\theta}_{qNPL}, \hat{P}_{qNPL})$.

Since the result of Proposition 7 also applies here by replacing Ψ with Λ^q, the local convergence property of the q-NPL algorithm is primarily determined by the spectral radius of $\Lambda^q_P \equiv \nabla_{P'}\Lambda^q(\theta^0, P^0)$. When $\rho(\Lambda^0_P)$ is less than 1, the q-NPL algorithm converges faster than the NPL algorithm because $\rho(\Lambda^q_P) = (\rho(\Lambda^0_P))^q$. Moreover, the variance of the q-NPL estimator approaches that of the MLE as $q \to \infty$.

Applying the q-NPL algorithm, as defined above, is computationally intensive because the q-NPL Step 1 requires evaluating Λ^q at many different values of θ. We reduce the computational burden by introducing a linear approximation of $\Lambda^q(\theta, P)$ around (η, P), where η is a preliminary estimate of θ: $\Lambda^q(\theta, P, \eta) \equiv \Lambda^q(\eta, P) + \nabla_{\theta'}\Lambda^q(\eta, P)(\theta - \eta)$.

Given an initial estimator $(\tilde{\theta}_0, \tilde{P}_0)$, the *approximate q-NPL algorithm* iterates the following steps until $j = k$:

Step 1: Given $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta \in \Theta_j^q} M^{-1} \sum_{m=1}^{M} \sum_{t=1}^{T} \ln \Lambda^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_{mt}|x_{mt})$, where $\Theta_j^q \equiv \{\theta \in \Theta : \tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a|x) \in [\xi, 1-\xi]$ for all $(a, x) \in A \times X\}$ for an arbitrary small $\xi > 0$.

Step 2: Given $(\tilde{\theta}_j, \tilde{P}_{j-1})$, update P using the obtained estimate $\tilde{\theta}_j$ by $\tilde{P}_j = \Lambda^q(\tilde{\theta}_j, \tilde{P}_{j-1})$.

Implementing Step 1 requires evaluating $\Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ and $\nabla_{\theta'} \Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ only once outside of the optimization routine for θ and thus involves much fewer evaluations of $\Lambda(\theta, P)$ across different values of P and θ, compared to the original q-NPL algorithm.[3]

Define the q-NPL counterparts of $\tilde{\theta}_0(P)$, $\phi_0(P)$, and $\Omega_{\theta\theta}$ as $\tilde{\theta}_0^q(P) \equiv \arg\max_{\theta \in \Theta} E[\sum_{t=1}^{T} \ln \Lambda^q(\theta, P)(a_{mt}|x_{mt})]$, $\phi_0^q(P) = \Lambda^q(\tilde{\theta}_0^q(P), P)$, and $\Omega_{\theta\theta}^q \equiv E(\nabla_\theta s_{mt}^\Lambda \nabla_{\theta'} s_{mt}^\Lambda)$ with $s_{mt}^\Lambda = \sum_{t=1}^{T} \ln \Lambda^q(\theta^0, P^0)(a_{mt}|x_{mt})$, respectively. Define $\Omega_{\theta P}^q$ analogously.

Assumption 5 *(a) Assumption 1 holds. (b) $\Psi(\theta, P)$ is four times continuously differentiable in $\mathcal{N}$. (c) There is a unique θ^0 such that $\Lambda^q(\theta^0, P^0) = P^0$. (d) $I - (\alpha \Psi_P^0 + (1-\alpha) I)^q$ and $I - \Psi_P^0$ are nonsingular. (e) The operator $\phi_0^q(P) - P$ has a nonsingular Jacobian matrix at P^0.*

Assumption 5(c) is necessary for identifying θ^0 when the conditional probability is given by $\Lambda^q(\theta, P)$. This assumption rules out $\theta^1 \neq \theta^0$ that satisfies $\Lambda^q(\theta^1, P^0) = P^0$ even if $\Lambda(\theta^1, P^0) \neq P^0$. This occurs, for example, if $\Lambda(\theta^1, P^0) = P^1$ and $\Lambda(\theta^1, P^1) = P^0$ hold for $\theta^1 \neq \theta^0$ and $P^1 \neq P^0$. Assumption 5(d) is necessary for $\Omega_{\theta\theta}^q$ to be nonsingular. Since $\Lambda_P^q = (\alpha \Psi_P^0 + (1-\alpha) I)^q$, the first condition holds if $\rho(\Lambda_P^q) < 1$ from 19.15 of Seber (2007).

The following proposition establishes the asymptotics of the q-NPL estimator and the convergence property of the approximate q-NPL algorithm. Proposition 12(c) implies that, when q is sufficiently large, the q-NPL estimator is more efficient than the NPL estimator, provided that additional conditions in Assumption 5 hold. Proposition 12(d) corresponds to Proposition 1.

Proposition 12 *Suppose that Assumption 5 holds. Then (a) $\hat{P}_{qNPL} - P^0 = O(M^{-1/2})$ a.s. and $M^{-1/2}(\hat{\theta}_{qNPL} - \theta^0) \to_d N(0, V_{qNPL})$, where $V_{qNPL} = [\Omega_{\theta\theta}^q + \Omega_{\theta P}^q (I - \Lambda_P^0)^{-1} \Lambda_\theta^q]^{-1} \Omega_{\theta\theta}^q \{[\Omega_{\theta\theta}^q + \Omega_{\theta P}^q (I - \Lambda_P^0)^{-1} \Lambda_\theta^q]^{-1}\}'$. (b) Suppose we obtain $(\tilde{\theta}_j, \tilde{P}_j)$ from $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ by the approximate q-NPL algorithm. Then, there exists a neighborhood $\mathcal{N}_6$ of (θ^0, P^0) such that $\tilde{\theta}_j - \hat{\theta}_{qNPL} = O(||\tilde{P}_{j-1} - \hat{P}_{qNPL}|| + M^{-1/2} ||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2)$ a.s. and $\tilde{P}_j - \hat{P}_{qNPL} = M_{\Lambda_\theta^q} \Lambda_P^q (\tilde{P}_{j-1} - \hat{P}_{qNPL}) + O(M^{-1/2} ||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2 + M^{-1/2} ||\tilde{P}_{j-1} - \hat{P}_{qNPL}|| + ||\tilde{P}_{j-1} - \hat{P}_{qNPL}||^2)$ a.s. uniformly in $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}_6$, where $M_{\Lambda_\theta^q} \equiv I - \Lambda_\theta^q (\Lambda_\theta^{q\prime} \Delta_P \Lambda_\theta^q)^{-1} \Lambda_\theta^{q\prime} \Delta_P$ with $\Lambda_\theta^q \equiv \nabla_{\theta'} \Lambda^q(\theta^0, P^0)$. (c) If $\rho(\Lambda_P^0) < 1$, then $V_{qNPL} \to V_{MLE}$ as $q \to \infty$. (d) Suppose*

[3] Using one-sided numerical derivatives, evaluating $\nabla_{\theta'} \Lambda^q(\tilde{\theta}_j, \tilde{P}_j)$ requires $(K+1)q$ function evaluations of $\Psi(\theta, P)$.

$\{\tilde{\theta}_k, \tilde{P}_k\}$ *is generated by the approximate q-NPL algorithm starting from* $(\tilde{\theta}_0, \tilde{P}_0)$ *and* $\rho(M_{\Lambda_\theta^q}\Lambda_P^q) < 1$*. Then, there exists a neighborhood* $\mathcal{N}_7$ *of* (θ^0, P^0) *such that, for any starting value* $(\tilde{\theta}_0, \tilde{P}_0) \in \mathcal{N}_7$*, we have* $\lim_{k\to\infty}(\tilde{\theta}_k, \tilde{P}_k) = (\hat{\theta}_{qNPL}, \hat{P}_{qNPL})$ *a.s.*

C.6 Proof of Propositions in Section C

Proof of Proposition 8 For part (a), write $\Gamma(\theta, P) - P$ as $\Gamma(\theta, P) - P = A(\theta, P)(\Psi(\theta, P) - P)$, where $A(\theta, P) \equiv (I - \Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P))^{-1}\Pi(\theta, P) + (I - \Pi(\theta, P))$. Let $Z(\theta, P)$ denote an orthonormal basis of the column space of $\Pi(\theta, P)$, so that $Z(\theta, P)Z(\theta, P)' = \Pi(\theta, P)$ and $Z(\theta, P)'Z(\theta, P) = I_r$. Suppress (θ, P) from $\Pi(\theta, P)$, $Z(\theta, P)$, and $\nabla_{P'}\Psi(\theta, P)$. A direct calculation gives $(I - \Pi\nabla_{P'}\Psi\Pi)^{-1}\Pi = Z(I - Z'\nabla_{P'}\Psi Z)^{-1}Z'$, so we can write $A(\theta, P)$ as $A(\theta, P) = Z(I - Z'\nabla_{P'}\Psi Z)^{-1}Z' + (I - \Pi)$. The stated result follows since $A(\theta, P)$ is nonsingular because $\text{rank}[Z(I - Z'\nabla_{P'}\Psi Z)^{-1}Z'] = r$, $\text{rank}(I - \Pi) = L - r$, and $Z(I - Z'\nabla_{P'}\Psi Z)^{-1}Z'$ and $I - \Pi$ are orthogonal to each other.

For part (b), define $\Gamma_P^0 \equiv \nabla_{P'}\Gamma(\theta^0, P^0)$ and $\Pi^0 \equiv \Pi(\theta^0, P^0)$. Define $\mathbb{P}$ with respect to $\Psi_P^0 \equiv \nabla_{P'}\Psi(\theta^0, P^0)$. Computing $\nabla_{P'}\Gamma(\theta, P)$ and noting that $\Psi(\theta^0, P^0) = P^0$, we find $\Gamma_P^0 = \Pi^0 + (I - \Pi^0\Psi_P^0\Pi^0)^{-1}\Pi^0(\Psi_P^0 - I) + (I - \Pi^0)\Psi_P^0$. Observe that $\Gamma_P^0\Pi^0 = (I - \Pi^0)\Psi_P^0\Pi^0 = 0$, where the last equality follows because $\Psi_P^0\Pi^0 P \in \mathbb{P}$ for any $P \in \mathbb{R}^L$ by the definition of Π^0. Hence, $\Gamma_P^0 = \Gamma_P^0(I - \Pi^0)$. We also have $(I - \Pi^0)\Gamma_P^0 = (I - \Pi^0)\Psi_P^0$ because a direct calculation gives $(I - \Pi^0\Psi_P^0\Pi^0)^{-1}\Pi^0 = Z^0(I - (Z^0)'\Psi_P^0 Z^0)^{-1}(Z^0)'$ where $Z^0 = Z(\theta^0, P^0)$, and hence $(I - \Pi^0)(I - \Pi^0\Psi_P^0\Pi^0)^{-1}\Pi^0 = 0$. Then, in conjunction with $\Gamma_P^0 = \Gamma_P^0(I - \Pi^0)$, we obtain $(I - \Pi^0)\Gamma_P^0 = (I - \Pi^0)\Psi_P^0(I - \Pi^0)$. Since $\Gamma_P^0(I - \Pi^0)$ has the same eigenvalues as $(I - \Pi^0)\Gamma_P^0$ (see Theorem 1.3.20 of Horn and Johnson, 1985), we have $\rho(\Gamma_P^0) = \rho(\Gamma_P^0(I - \Pi^0)) = \rho((I - \Pi^0)\Gamma_P^0) = \rho[(I - \Pi^0)\Psi_P^0(I - \Pi^0)] \leq \delta^0$, where the last inequality follows from Lemma 2.10 of SK: P, Q, and F_u^* in SK correspond to our Π^0, $I - \Pi^0$, and Ψ_P^0. $\square$

Proof of Proposition 9 The stated results follow from Proposition 2 of AM07 and our Proposition 7 if Assumptions 1(b)-(c) and 1(e)-(h) and Assumptions 2(b)-(c) hold when $\Psi(\theta, P)$ is replaced with $\Gamma(\theta, P)$.

We check Assumptions 2(b)-(c) first because they are used in showing the other conditions. First, note that Chu (1990, Section 4.2, in particular line 17 on page 1377) proved the following: if a matrix $A(t)$ is ℓ times continuously differentiable with respect to t, and if $X(t)$ spans the invariant subspace corresponding to a subset of eigenvalues of $A(t)$, then $X(t)$ is also ℓ times continuously differentiable with respect to t. Consequently, $\Pi(\theta, P)$ is three times continuously differentiable in $\mathcal{N}$ (we suppress "in $\mathcal{N}$" henceforth) since $\nabla_{P'}\Psi(\theta, P)$ is three times continuously differentiable from Assumption 4(b). Further, $I - \Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P)$ is nonsingular and three times continuously differentiable from Assumptions 4(b)-(c), and hence Assumption 2(b) holds for $\Gamma(\theta, P)$. For Assumption 2(c), a direct calculation gives $\Omega_{\theta\theta}^\Gamma = \Psi_\theta^{0\prime}A(\theta^0, P^0)'\Delta_P A(\theta^0, P^0)\Psi_\theta^0$, where $A(\theta, P)$ is defined in the proof of Proposition 2 and shown to be nonsingular. Since $\text{rank}(\Psi_\theta^0) = K$ from nonsingularity of $\Omega_{\theta\theta} = \Psi_\theta^{0\prime} \Delta_P \Psi_\theta^0$, positive definiteness of $\Omega_{\theta\theta}^\Gamma$ follows.

We proceed to confirm Assumptions 1(b)-(c) and 1(e)-(h) hold for $\Gamma(\theta, P)$. Assumption 1(b) for $\Gamma(\theta, P)$ follows from Assumption 4(d). Assumption 1(c) holds because we have already shown that $\Gamma(\theta, P)$ is three times continuously differentiable. Assumption 1(e) holds because $\Psi(\theta, P)$ and $\Gamma(\theta, P)$ have the same fixed points by Proposition 8. As discussed in page 21 of AM07, Assumption 1(f) is implied by Assumption 4(e). Assumption 1(g) for $\tilde{\theta}_0^\Gamma(P)$ follows from the positive definiteness of $\Omega_{\theta\theta}^\Gamma$ and by the implicit function theorem applied to the first order condition for θ. Assumption 1(h) follows from Assumption 4(e). □

Proof of Proposition 10 Write the objective function as $Q_M^\Gamma(\theta, P, \eta) \equiv M^{-1} \sum_{m=1}^M \sum_{t=1}^T \ln \Gamma(\theta, P, \eta)(a_{mt}|x_{mt})$, and define $Q_0^\Gamma(\theta, P, \eta) \equiv E Q_M^\Gamma(\theta, P, \eta)$. 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 (i) $\Psi(\theta, P)$ is four times continuously differentiable in (θ, P) if $(\theta, P, \eta) \in \mathcal{N}_3(\epsilon_1)$, (ii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} Q_0^\Gamma(\theta, P, \eta)^{-1}|| < \infty$, and (iii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla^3 Q_0^\Gamma(\theta, P, \eta)|| < \infty$ because $\Gamma(\theta^0, P^0, \theta^0)(a|x) = P^0(a|x) > 0$, $\Gamma(\theta, P, \eta)$ is three times continuously differentiable (see the proof of Proposition 9), and $\nabla_{\theta\theta'} Q_0^\Gamma(\theta^0, P^0, \theta^0) = \nabla_{\theta\theta'} Q_0^\Gamma(\theta^0, P^0)$ is nonsingular.

First, we assume $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ and derive the stated representation of $\tilde{\theta}_j - \hat{\theta}$ and $\tilde{P}_j - \hat{P}$. We later show $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ a.s. if $\mathcal{N}_2$ is taken sufficiently small. Henceforth, we suppress the subscript RPM from $\hat{\theta}_{RPM}$ and $\hat{P}_{RPM}$. Expanding the first order condition $\nabla_\theta Q_M^\Gamma(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = 0$ around $(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ gives

$$0 = \nabla_\theta Q_M^\Gamma(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta\theta'} Q_M^\Gamma(\bar{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\tilde{\theta}_j - \hat{\theta}), \tag{18}$$

where $\bar{\theta} \in [\tilde{\theta}_j, \hat{\theta}]$. Writing $\bar{\theta} = \bar{\theta}(\tilde{\theta}_j)$, we obtain $\sup_{(\tilde{\theta}_j,\tilde{P}_{j-1},\tilde{\theta}_{j-1})\in\mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} Q_M^\Gamma(\bar{\theta}(\tilde{\theta}_j), \tilde{P}_{j-1}, \tilde{\theta}_{j-1})^{-1}|| = O(1)$ a.s. because (i) $||\bar{\theta}(\tilde{\theta}_j) - \theta^0|| < \epsilon_1$ a.s. since $||\tilde{\theta}_j - \theta^0|| < \epsilon_1$ and $\hat{\theta}$ is strongly consistent, and (ii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} Q_M^\Gamma(\theta, P, \eta)^{-1}|| = O(1)$ a.s. since $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} Q_0^\Gamma(\theta, P, \eta)^{-1}|| < \infty$ and $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla^2 Q_M^\Gamma(\theta, P, \eta) - \nabla^2 Q_0^\Gamma(\theta, P, \eta)|| = o(1)$ a.s. Therefore, the stated representation of $\tilde{\theta}_j - \hat{\theta}$ follows if we show

$$\nabla_\theta Q_M^\Gamma(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = -\Omega_{\theta P}^\Gamma(\tilde{P}_{j-1} - \hat{P}) + r_{Mj}^*, \tag{19}$$

where r_{Mj}^* denotes a generic remainder term that is $O(M^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}||^2 + M^{-1/2}||\tilde{P}_{j-1} - \hat{P}|| + ||\tilde{P}_{j-1} - \hat{P}||^2)$ a.s. uniformly in $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}_2$.

We proceed to show (19). Expanding $\nabla_\theta Q_M^\Gamma(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ twice around $(\hat{\theta}, \hat{P}, \hat{\theta})$ gives $\nabla_\theta Q_M^\Gamma(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = \nabla_\theta Q_M^\Gamma(\hat{\theta}, \hat{P}, \hat{\theta}) + \nabla_{\theta P'} Q_M^\Gamma(\hat{\theta}, \hat{P}, \hat{\theta})(\tilde{P}_{j-1} - \hat{P}) + \nabla_{\theta\eta'} Q_M^\Gamma(\hat{\theta}, \hat{P}, \hat{\theta})(\tilde{\theta}_{j-1} - \hat{\theta}) + O(||\tilde{\theta}_{j-1} - \hat{\theta}||^2 + ||\tilde{P}_{j-1} - \hat{P}||^2)$ a.s. For the first term on the right, the RPM estimator satisfies $\nabla_\theta Q_M^\Gamma(\hat{\theta}, \hat{P}, \hat{\theta}) = 0$ a.s. because $\nabla_{\theta'} Q_M^\Gamma(\hat{\theta}, \hat{P}) = 0$ from the first order condition, and Proposition 8(a) implies $\Psi(\hat{\theta}, \hat{P}) = \hat{P}$ a.s. and hence $\nabla_{\theta'}\Gamma(\hat{\theta}, \hat{P}, \hat{\theta}) = \nabla_{\theta'}\Gamma(\hat{\theta}, \hat{P})$ a.s. For the second and third terms on the right, we have $E[\sum_{t=1}^{T} \nabla_{\theta P'} \ln \Gamma(\theta^0, P^0, \theta^0)(a_{mt}|x_{mt})] = -\Omega_{\theta P}^{\Gamma}$ and $E[\sum_{t=1}^{T} \nabla_{\theta\eta'} \ln \Gamma(\theta^0, P^0, \theta^0)(a_{mt}|x_{mt})] = 0$ by the information matrix equality because $\Gamma(\theta^0, P^0, \theta^0) = \Gamma(\theta^0, P^0)$, $\nabla_{\theta'}\Gamma(\theta^0, P^0, \theta^0) = \nabla_{\theta'}\Gamma(\theta^0, P^0)$, $\nabla_{P'}\Gamma(\theta^0, P^0, \theta^0) = \nabla_{P'}\Gamma(\theta^0, P^0)$, and $\nabla_{\eta'}\Gamma(\theta^0, P^0, \theta^0) = 0$ from $P^0 = \Psi(\theta^0, P^0)$. Therefore, (19) follows from the root-M consistency of $(\hat{\theta}, \hat{P})$.

For the representation of $\tilde{P}_j - \hat{P}$, first we have

$$\tilde{P}_j = \hat{P} + \Gamma_\theta^0(\tilde{\theta}_j - \hat{\theta}) + \Gamma_P^0(\tilde{P}_{j-1} - \hat{P}) + r_{Mj}^*, \tag{20}$$

by expanding $\tilde{P}_j = \Gamma(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_j)$ around $(\hat{\theta}, \hat{P}, \hat{\theta})$ and using $\Gamma(\hat{\theta}, \hat{P}, \hat{\theta}) = \hat{P}$. Next, refine (18) as $0 = \nabla_\theta Q_M^\Gamma(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) - \Omega_{\theta\theta}^\Gamma(\tilde{\theta}_j - \hat{\theta}) + r_{Mj}^*$ by expanding $\nabla_{\theta\theta'} Q_M^\Gamma(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ in (18) around $(\hat{\theta}, \hat{P}, \hat{\theta})$ to write it as $\nabla_{\theta\theta'} Q_M^\Gamma(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = -\Omega_{\theta\theta}^\Gamma + O(M^{-1/2}) + O(||\tilde{\theta}_{j-1} - \hat{\theta}||) + O(||\tilde{P}_{j-1} - \hat{P}||)$ a.s. and using the bound of $\tilde{\theta}_j - \hat{\theta}$ obtained above. Substituting this into (19) gives

$$\tilde{\theta}_j - \hat{\theta} = -(\Omega_{\theta\theta}^\Gamma)^{-1}\Omega_{\theta P}^\Gamma(\tilde{P}_{j-1} - \hat{P}) + r_{Mj}^*. \tag{21}$$

The stated result follows from substituting this into (20) in conjunction with $(\Omega_{\theta\theta}^\Gamma)^{-1}\Omega_{\theta P}^\Gamma = (\Gamma_\theta^{0\prime}\Delta_P\Gamma_\theta^0)^{-1}\Gamma_\theta^{0\prime}\Delta_P\Gamma_P^0$.

It remains to show $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ a.s. if $\mathcal{N}_2$ is taken sufficiently small. We first show that

$$\sup_{(\theta,\eta,P)\in\bar{\Theta}_j\times\mathcal{N}_2} |Q_M^\Gamma(\theta, P, \eta) - Q_0^\Gamma(\theta, P, \eta)| = o(1) \ a.s., \quad Q_0^\Gamma(\theta, P, \eta) \text{ is continuous in } (\theta, \eta, P) \in \bar{\Theta}_j \times \mathcal{N}_2. \tag{22}$$

Take $\mathcal{N}_2$ sufficiently small, then it follows from the strong consistency of $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ and the continuity of $\Gamma(\theta, P, \eta)$ that $\Gamma(\theta, P, \eta)(a|x) \in [\xi/2, 1 - \xi/2]$ for all $(a, x) \in A \times X$ and $(\theta, P, \eta) \in \bar{\Theta}_j \times \mathcal{N}$ a.s. Observe that (i) $\bar{\Theta}_j \times \mathcal{N}$ is compact because it is an intersection of the compact set Θ and $|A||X|$ closed sets, (ii) $\ln \Gamma(\theta, P, \eta)$ is continuous in $(\theta, P, \eta) \in \bar{\Theta}_j \times \mathcal{N}$, and (iii) $E \sup_{(\theta,P,\eta)\in\bar{\Theta}_j\times\mathcal{N}} |\ln \Gamma(\theta, P, \eta)(a_i|x_i)| \leq |\ln(\xi/2)| + |\ln(1 - \xi/2)| < \infty$ because of the way we choose $\mathcal{N}$. Therefore, (22) follows from Kolomogorov's strong law of large numbers and Theorem 2 and Lemma 1 of Andrews (1992).

Finally, we show $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ a.s. under (22) by applying the argument in the proof of Proposition 7. Define $\Delta = Q_0^\Gamma(\theta^0, P^0, \theta^0) - \sup_{\theta\in\mathcal{N}_\theta(\epsilon_1)^c\cap\Theta} Q_0^\Gamma(\theta, P^0, \theta^0) > 0$, where the last inequality follows from the information inequality because $Q_0^\Gamma(\theta, P^0, \theta^0)$ is uniquely maximized at θ^0 and $\mathcal{N}_\theta(\epsilon_1)^c \cap \Theta$ is compact. It follows that $\{\tilde{\theta}_j \notin \mathcal{N}_\theta(\epsilon_1)\} \Rightarrow \{Q_0^\Gamma(\theta^0, P^0, \theta^0) - Q_0^\Gamma(\tilde{\theta}_j, P^0, \theta^0) \geq \Delta\}$. Proceeding as in the proof of Proposition 7, we find that, if $\mathcal{N}_2$ is taken sufficiently small, then $Q_0^\Gamma(\theta^0, P^0, \theta^0) - Q_0^\Gamma(\tilde{\theta}_j, P^0, \theta^0) \leq \Delta/2 + o(1)$ a.s. and hence $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ a.s. $\square$

Proof of Proposition 11 The proof closely follows the proof of Proposition 1. We suppress the subscript RPM from $\hat{\theta}_{RPM}$ and $\hat{P}_{RPM}$. Define

$$D = \begin{pmatrix} 0 & -(\Omega^{\Gamma}_{\theta\theta})^{-1}\Omega^{\Gamma}_{\theta P} \\ 0 & M_{\Gamma_\theta}\Gamma^0_P \end{pmatrix}. \tag{23}$$

Note that $\rho(D) = \rho(M_{\Gamma_\theta}\Gamma_P)$ and there exists a matrix norm $||\cdot||_\alpha$ such that $\|D\|_\alpha \leq \rho(D) + b = \rho(M_{\Gamma_\theta}\Gamma_P) + b$. We define the vector norm for $x \in \mathbb{R}^{k+L}$ as $||x||_\beta = ||[x\ 0\ldots 0]||_\alpha$, then $||Ax||_\beta \leq ||A||_\alpha ||x||_\beta$ for any matrix A.

From the representation of $\tilde{P}_j - \hat{P}$ and $\tilde{\theta}_j - \hat{\theta}$ in Proposition 10 and (21), there exists a neighborhood $\mathcal{N}_\zeta$ of ζ^0 such that $\tilde{\zeta}_j - \hat{\zeta} = D(\tilde{\zeta}_{j-1} - \hat{\zeta}) + O(M^{-1/2}||\tilde{\zeta}_{j-1} - \hat{\zeta}||_\beta + ||\tilde{\zeta}_{j-1} - \hat{\zeta}||^2_\beta)$ holds a.s. uniformly in $\tilde{\zeta}_{j-1} \in \mathcal{N}_\zeta$. The stated result then follows from repeating the proof of Proposition 1. $\square$

Proof of Corollary 1 The proof closely follows the proof of Proposition 10. 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 $Q^{\Gamma}_M(\theta, \tilde{P}_{j-1}, \theta^*, P^*) = M^{-1}\sum_{m=1}^M \sum_{t=1}^T \ln \Gamma(\theta, \tilde{P}_{j-1}, \theta^*, P^*)(a_{mt}|x_{mt})$. For $\epsilon_1 > 0$, define a neighborhood $\mathcal{N}_5(\epsilon_1) = \{(\theta, P, \eta, Q) : \max\{||\theta - \theta^0||, ||P - P^0||, ||\eta - \theta^0||, ||Q - P^0||\} < \epsilon_1\}$. Then, for any $\epsilon_1 > 0$, we have $(\tilde{\theta}_j, \tilde{P}_{j-1}, \theta^*, P^*) \in \mathcal{N}_5(\epsilon_1)$ a.s. if $\mathcal{N}_4$ is chosen sufficiently small by the same argument as the proof of Proposition 10.

Assuming $(\tilde{\theta}_j, \tilde{P}_{j-1}, \theta^*, P^*) \in \mathcal{N}_5(\epsilon_1)$, the stated result follows from starting from the first order condition $\nabla_{\theta'} Q^{\Gamma}_M(\tilde{\theta}_j, \tilde{P}_{j-1}, \theta^*, P^*) = 0$, expanding it around $(\hat{\theta}, \tilde{P}_{j-1}, \theta^*, P^*)$, and following the proof of Proposition 10 using $\nabla_{Q'}\Gamma(\theta^0, P^0, \theta^0, P^0) = 0$. $\square$

Proof of Proposition 12 Part (a) follows from Proposition 2 of AM07 if Assumptions 1(b)-(c) and 1(e)-(h) and Assumptions 2(b)-(c) hold when $\Psi(\theta, P)$ is replaced with $\Lambda^q(\theta, P)$. Similar to the proof of Proposition 10, we check Assumptions 2(b)-(c) first. Assumption 2(b) holds for $\Lambda^q(\theta, P)$ because $\Psi(\theta, P)$ is three times continuously differentiable in $\mathcal{N}$ from Assumption 5(b). For Assumption 2(c), a direct calculation gives $\Omega^q_{\theta\theta} = (\nabla_{\theta'}\Lambda^q(\theta^0, P^0))'\Delta_P \nabla_{\theta'}\Lambda^q(\theta^0, P^0) = \Lambda^{0\prime}_\theta (I - (\Lambda^0_P)^q)'(I - \Lambda^{0\prime}_P)^{-1}\Delta_P (I - \Lambda^0_P)^{-1}(I - (\Lambda^0_P)^q)\Lambda_\theta = \Psi^{0\prime}_\theta (I - (\alpha\Psi^0_P + (1-\alpha)I)^q)'(I - \Psi^{0\prime}_P)^{-1}\Delta_P (I - \Psi^0_P)^{-1}(I - (\alpha\Psi^0_P + (1-\alpha)I)^q)\Psi^0_\theta$, where the second equality follows from $\nabla_{\theta'}\Lambda^q(\theta^0, P^0) = (\sum_{j=0}^{q-1}(\Lambda^0_P)^j)\Lambda^0_\theta = (I - \Lambda^0_P)^{-1}(I - (\Lambda^0_P)^q)\Lambda^0_\theta$, and the third equality follows from $\Lambda^0_\theta = \alpha\Psi^0_\theta$ and $\Lambda^0_P = \alpha\Psi^0_P + (1-\alpha)I$. Since $\text{rank}(\Psi^0_\theta) = K$ from nonsingularity of $\Omega_{\theta\theta} = \Psi^{0\prime}_\theta \Delta_P \Psi^0_\theta$, positive definiteness of $\Omega^q_{\theta\theta}$ follows from Assumption 5(d).

The proof of part (a) is completed by confirming that Assumptions 1(b)-(c) and 1(e)-(h) hold for $\Lambda^q(\theta, P)$. Assumptions 1(b)-(c) hold for $\Lambda^q(\theta, P)$ because Assumptions 1(b)-(c) hold for $\Psi(\theta, P)$. Assumption 1(e) for $\Lambda^q(\theta, P)$ follows from Assumption 5(c). As discussed in page 21 of AM07, Assumption 1(f) for $\Lambda^q(\theta, P)$ is implied by Assumption 5(e). Assumption 1(g) for $\tilde{\theta}^q_0(P)$ follows from the positive definiteness of $\Omega^q_{\theta\theta}$ and applying the implicit function theorem to the first order condition for θ. Assumption 1(h) follows from Assumption 5(e). This completes the proof of part (a).

We proceed to prove part (b). Define the objective function and its limit as $Q_M^q(\theta, P, \eta) \equiv M^{-1}\sum_{m=1}^M \sum_{t=1}^T \ln \Lambda^q(\theta, P, \eta)(a_{mt}|x_{mt})$ and $Q_0^q(\theta, P, \eta) \equiv EQ_M^q(\theta, P, \eta)$. 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 (i) $\Psi(\theta, P)$ is four times continuously differentiable in (θ, P) if $(\theta, P, \eta) \in \mathcal{N}_3(\epsilon_1)$, (ii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} Q_0^q(\theta, P, \eta)^{-1}|| < \infty$, and (iii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla^3 Q_0^q(\theta, P, \eta)|| < \infty$ because $\Lambda^q(\theta^0, P^0, \theta^0)(a|x) = P^0(a|x) > 0$, $\Lambda^q(\theta, P, \eta)$ is three times continuously differentiable, and $\nabla_{\theta\theta'} Q_0^q(\theta^0, P^0, \theta^0) = \nabla_{\theta\theta'} Q_0^q(\theta^0, P^0)$ is nonsingular.

First, we assume $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ and derive the stated representation of $\tilde{\theta}_j - \hat{\theta}$ and $\tilde{P}_j - \hat{P}$. We later show $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ a.s. if $\mathcal{N}_6$ is taken sufficiently small. Henceforth, we suppress the subscript qNPL from $\hat{\theta}_{qNPL}$ and $\hat{P}_{qNPL}$. The proof is similar to the proof of the updating formula of Proposition 10. For the representation of $\tilde{\theta}_j - \hat{\theta}$, expanding the first order condition $0 = \nabla_\theta Q_M^q(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ around $(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ gives $0 = \nabla_\theta Q_M^q(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta\theta'} Q_M^q(\bar{\theta}(\tilde{\theta}_j), \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\tilde{\theta}_j - \hat{\theta})$, which corresponds to (18) in the proof of Proposition 10. Proceeding as in the proof of Proposition 10, we obtain $\sup_{(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})\in\mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} Q_M^q(\bar{\theta}(\tilde{\theta}_j), \tilde{P}_{j-1}, \tilde{\theta}_{j-1})^{-1}|| = O(1)$ a.s. Therefore, the stated representation of $\tilde{\theta}_j - \hat{\theta}$ follows if we show $\nabla_\theta Q_M^q(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = -\Omega_{\theta P}^q(\tilde{P}_{j-1} - \hat{P}) + r_{Mj}^*$, where r_{Mj}^* denotes a remainder term of $O(M^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}||^2 + M^{-1/2}||\tilde{P}_{j-1} - \hat{P}|| + ||\tilde{P}_{j-1} - \hat{P}||^2)$ a.s. uniformly in $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}_6$. This representation corresponds to (19) in the proof of Proposition 10 and follows from the same argument. Namely, expanding $\nabla_\theta Q_M^q(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ twice around $(\hat{\theta}, \hat{P}, \hat{\theta})$ and noting that (i) the q-NPL estimator satisfies $\nabla_\theta Q_M^q(\hat{\theta}, \hat{P}, \hat{\theta}) = 0$, (ii) $\Lambda^q(\theta^0, P^0, \theta^0) = \Lambda^q(\theta^0, P^0)$, $\nabla_{\theta'}\Lambda^q(\theta^0, P^0, \theta^0) = \nabla_{\theta'}\Lambda^q(\theta^0, P^0)$, $\nabla_{P'}\Lambda^q(\theta^0, P^0, \theta^0) = \nabla_{P'}\Lambda^q(\theta^0, P^0)$, and $\nabla_{\eta'}\Lambda^q(\theta^0, P^0, \theta^0) = 0$, and using the information matrix equality and the root-M consistency of $(\hat{\theta}, \hat{P})$ gives the required result.

The proof of the representation of $\tilde{P}_j - \hat{P}$ follows from the proof of Proposition 10, because (i) $\tilde{P}_j = \hat{P} + \Lambda_\theta^q(\tilde{\theta}_j - \hat{\theta}) + \Lambda_P^q(\tilde{P}_{j-1} - \hat{P}) + r_{Mj}^*$, which corresponds to (20) in the proof of Proposition 10, from expanding $\Lambda^q(\tilde{\theta}_j, \tilde{P}_{j-1})$ twice around $(\hat{\theta}, \hat{P})$ and using $\hat{P} = \Lambda^q(\hat{\theta}, \hat{P})$, (ii) $\nabla_{\theta\theta'} Q_M^q(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\tilde{\theta}_j - \hat{\theta}) = -\Omega_{\theta\theta}^q(\tilde{\theta}_j - \hat{\theta}) + r_{Mj}^*$ from expanding $\nabla_{\theta\theta'} Q_M^q(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ around $(\hat{\theta}, \hat{P}, \hat{\theta})$ and using the bound of $\tilde{\theta}_j - \hat{\theta}$ obtained above, and (iii) $(\Omega_{\theta\theta}^q)^{-1}\Omega_{\theta P}^q = (\Lambda_\theta^{q\prime}\Delta_P\Lambda_\theta^q)^{-1}\Lambda_\theta^{q\prime}\Delta_P\Lambda_P^q$.

The proof of part (b) is completed by showing $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ a.s. if $\mathcal{N}_6$ is taken sufficiently small. First, observe that (22) in the proof of Proposition 10 holds with $Q_M^\Gamma(\theta, P, \eta)$ and $Q_0^\Gamma(\theta, P, \eta)$ replacing $Q_M^q(\theta, P, \eta)$ and $Q_0^q(\theta, P, \eta)$ if we take $\mathcal{N}_6$ sufficiently small. Therefore, $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ a.s. follows from repeating the argument in the last paragraph of the proof of Proposition 10 if we show that θ^0 uniquely maximizes $Q_0^q(\theta, P^0, \theta^0)$. Note that

$$
\begin{aligned}
Q_0^q(\theta, P^0, \theta^0) - Q_0^q(\theta^0, P^0, \theta^0) &= TE\ln(\nabla_{\theta'}\Lambda^q(\theta^0, P^0)(\theta - \theta^0) + P^0)(a_{mt}|x_{mt}) - TE\ln P^0(a_{mt}|x_{mt}) \\
&= TE\ln\left(\frac{\nabla_{\theta'}\Lambda^q(\theta^0, P^0)(a_{mt}|x_{mt})(\theta - \theta^0)}{P^0(a_{mt}|x_{mt})} + 1\right). \qquad (24)
\end{aligned}
$$

Recall that $\ln(y+1) \leq y$ for all $y > -1$ where the inequality is strict if $y \neq 0$. Since $\text{rank}(\nabla_{\theta'}\Lambda^q(\theta^0, P^0)) = K$ from the positive definiteness of $\Omega^q_{\theta\theta}$, it follows that $\nabla_{\theta'}\Lambda^q(\theta^0, P^0)\nu \neq 0$ for any K-vector $\nu \neq 0$. Therefore, $\nabla_{\theta'}\Lambda^q(\theta^0, P^0)(a_{mt}|x_{mt})(\theta - \theta^0) \neq 0$ for at least one (a_{mt}, x_{mt}) for all $\theta \neq \theta^0$. Consequently, the right hand side of (24) is strictly smaller than $TE[\nabla_{\theta'}\Lambda^q(\theta^0, P^0)(a_{mt}|x_{mt})(\theta - \theta^0)/P^0(a_{mt}|x_{mt})]$ for all $\theta \neq \theta^0$.
Because $E[\nabla_{\theta'}\Lambda^q(\theta^0, P^0)(a_{mt}|x_{mt})/P^0(a_{mt}|x_{mt})] = 0$, we have $Q^q_0(\theta, P^0, \theta^0) - Q^q_0(\theta^0, P^0, \theta^0) < 0$ for all $\theta \neq \theta^0$. Therefore, θ^0 uniquely maximizes $Q^q_0(\theta, P^0, \theta^0)$, and we complete the proof of part (b).

We prove part (c). From the proof of part (a) in conjunction with the relation $\Lambda^0_P = \alpha\Psi^0_P + (1-\alpha)I$, we may write $\Omega^q_{\theta\theta}$ as $\Omega^q_{\theta\theta} = T\Psi^{0\prime}_\theta(I - (\Lambda^0_P)^q)'(I - \Psi^{0\prime}_P)^{-1}\Delta_P(I - \Psi^0_P)^{-1}(I - (\Lambda^0_P)^q)\Psi^0_\theta$. Similarly, using the relation $\nabla_{P'}\Lambda^q(\theta^0, P^0) = (\Lambda^0_P)^q$, we obtain $\Omega^q_{\theta P} = T\Lambda^{0\prime}_\theta(I - (\Lambda^0_P)^q)'(I - \Lambda^{0\prime}_P)^{-1}\Delta_P(\Lambda^0_P)^q$. Therefore, if $\rho(\Lambda^0_P) < 1$, then $\Omega^q_{\theta\theta} \to T\Psi^{0\prime}_\theta(I - \Psi^{0\prime}_P)^{-1}\Delta_P(I - \Psi^0_P)^{-1}\Psi^0_\theta$ and $\Omega^q_{\theta P} \to 0$ as $q \to \infty$, and it follows that $V_{qNPL} \to [T\Psi^{0\prime}_\theta(I - \Psi^{0\prime}_P)^{-1}\Delta_P(I - \Psi^0_P)^{-1}\Psi^0_\theta]^{-1}$ as $q \to \infty$. This limit is the same as $V_{MLE} = (TE[\nabla_\theta \ln P(\theta^0)(a_{mt}|x_{mt})\nabla_{\theta'} \ln P(\theta^0)(a_{mt}|x_{mt})])^{-1}$, where $P(\theta) \equiv \arg\max_{P \in \mathcal{M}_\theta} E \ln P(a_{mt}|x_{mt})$ with $\mathcal{M}_\theta \equiv \{P \in B_P : P = \Psi(\theta, P)\}$, because $\nabla_{\theta'}P(\theta) = (I - \nabla_{P'}\Psi(\theta, P(\theta)))^{-1}\nabla_{\theta'}\Psi(\theta, P(\theta))$ holds in a neighborhood of $\theta = \theta^0$.

We omit the proof of part (d) because it is identical to the proof of Proposition 11 except that $\hat{\theta}_{RPM}$, $\hat{P}_{RPM}$, $(\Omega^\Gamma_{\theta\theta})^{-1}\Omega^\Gamma_{\theta P}$, and $M_{\Gamma_\theta}\Gamma_P$ are replaced with $\hat{\theta}_{qNPL}$, $\hat{P}_{qNPL}$, $(\Omega^q_{\theta\theta})^{-1}\Omega^q_{\theta P}$, and $M_{\Lambda^q_\theta}\Lambda^q_P$, respectively. □

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

Consider a panel data set $\{\{a_{mt}, x_{mt}, x_{m,t+1}\}^T_{t=1}\}^M_{m=1}$ such that $w_m = \{a_{mt}, x_{mt}, x_{m,t+1}\}^T_{t=1}$ is randomly drawn across m's from the population. The conditional probability distribution of a_{mt} given x_{mt} for a type k agent is given by a fixed point of $P_{\theta^k} = \Psi(\theta^k, P_{\theta^k})$. To simplify our analysis, we assume that the transition probability function of x_{mt} is independent of types and given by $f_x(x_{m,t+1}|a_{mt}, x_{mt})$ and is known to researchers.[4]

[4] When 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

In this framework, the initial state x_{m1} is correlated with the unobserved type (i.e., the initial conditions problem of Heckman (1981)). We assume that x_{m1} for type k is randomly drawn from the type k stationary distribution characterized by a fixed point of the following equation: $p^*(x) = \sum_{x' \in X} p^*(x') \left(\sum_{a' \in A} P_{\theta^k}(a'|x') f_x(x|a', x') \right) \equiv [T(p^*, P_{\theta^k})](x)$. 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^k denote type k's conditional choice probabilities, stack the P^k's as $\mathbf{P} = (P^{1\prime}, \ldots, P^{K\prime})'$, and let $\mathbf{P}^0$ denote its true value. Define $\mathbf{\Psi}(\theta, \mathbf{P}) = (\Psi(\theta^1, P^1)', \ldots, \Psi(\theta^K, P^K)')'$. Then, for a value of θ, the set of possible conditional choice probabilities consistent with the fixed point constraints is given by $\mathcal{M}_\theta^* = \{\mathbf{P} \in B_P^K : \ \mathbf{P} = \mathbf{\Psi}(\theta, \mathbf{P})\}$. The maximum likelihood estimator for a model with unobserved heterogeneity is:

$$\hat{\zeta}_{MLE} = \arg\max_{\zeta \in \Theta_\zeta} \left\{ \max_{\mathbf{P} \in \mathcal{M}_\theta^*} M^{-1} \sum_{m=1}^{M} \ln \left([L(\pi, \mathbf{P})](w_m) \right) \right\}, \tag{25}$$

where $[L(\pi, \mathbf{P})](w_m) = \sum_{k=1}^K \pi^k p^*_{P^k}(x_{m1}) \prod_{t=1}^T P^k(a_{mt}|x_{mt}) f_x(x_{m,t+1}|a_{mt}, x_{mt})$, and $p^*_{P^k} = T(p^*_{P^k}, P^k)$ is the type k stationary distribution of x when the conditional choice probability is P^k. 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 $\tilde{\mathbf{P}}_0 = (\tilde{P}_0^1, \ldots, \tilde{P}_0^K)$ is available. For $j = 1, 2, \ldots$, iterate

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

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

until $j = \ell$. If iterations converge, the limit satisfies $\hat{\zeta} = \arg\max_{\zeta \in \Theta_\zeta} \ M^{-1} \sum_{m=1}^M \ln([L(\pi, \mathbf{\Psi}(\theta, \hat{\mathbf{P}}))](w_m))$ and $\hat{\mathbf{P}} = \mathbf{\Psi}(\hat{\theta}, \hat{\mathbf{P}})$. 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 AM07. 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 6 *(a)* $w_m = \{(a_{mt}, x_{mt}, x_{m,t+1}) : m = 1, \ldots, M; t = 1, \ldots, T\}$ *are independent across* m *and stationary over* t*, and* $\Pr(x_{mt} = x) > 0$ *for any* $x \in X$*. (b)* $[L(\pi, \mathbf{P})](w) > 0$ *for any* w *and for any* $(\pi, \mathbf{P}) \in \Theta_\pi \times B_P^K$*. (c)* $\Psi(\theta, P)$ *is twice continuously differentiable. (d)* Θ_ζ *is*

transition probability function is also type-dependent in the context of a single-agent dynamic programming model with unobserved heterogeneity.

compact and B_P^K *is a compact and convex subset of* $[0,1]^{LK}$*. (e) There is a unique* $\zeta^0 \in int(\Theta_\zeta)$ *such that* $[L(\pi^0, \mathbf{P}^0)](w) = [L(\pi^0, \mathbf{\Psi}(\theta^0, \mathbf{P}^0))](w)$*. (f)* $(\zeta^0, \mathbf{P}^0)$ *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 6, the consistency and asymptotic normality of the NPL estimator can be shown by following the proof of Proposition 2 of AM07.

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

Assumption 7 *(a) Assumption 6 holds. (b)* $\Psi(\theta, P)$ *is three times continuously differentiable. (c)* $\Omega_{\zeta\zeta}$ *is nonsingular.*

Assumption 7 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.

Proposition 13 *Suppose that Assumptions 6-7 hold. Then, there exists a neighborhood* $\mathcal{N}_{\mathbf{P}}$ *of* $\mathbf{P}^0$ *such that*

$$\begin{aligned} \tilde{\zeta}_j - \hat{\zeta}_{NPL} &= O(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||), \\ \tilde{\mathbf{P}}_j - \hat{\mathbf{P}}_{NPL} &= [I - \mathbf{\Psi}_\theta^0 D \mathbf{\Psi}_\theta^{0\prime} L_P' \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P] \mathbf{\Psi}_P^0 (\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) \\ &\quad + O(M^{-1/2} ||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}|| + ||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||^2), \end{aligned}$$

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

Note that $I - \mathbf{\Psi}_\theta^0 D \mathbf{\Psi}_\theta^{0\prime} L_P' \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P$ is a projection matrix. The convergence rate of the NPL algorithm for models with unobserved heterogeneity is primarily determined by the dominant eigenvalue of $\mathbf{\Psi}_P^0$. When the NPL algorithm encounters a convergence problem, replacing $\Psi(\theta, P)$ with $\Lambda(\theta, P)$ or $\Gamma(\theta, P)$ improves the convergence.

Remark 2 *It is possible to relax the stationarity assumption on the initial states by estimating the type-specific initial distributions of* x*, denoted by* $\{p^{*k}\}_{k=1}^K$*, without imposing a stationarity* *restriction in Step 1 of the NPL algorithm. In this case, Proposition 13 holds with additional remainder terms.*

Proof of Proposition 13 We suppress the subscript NPL from $\hat{\zeta}_{NPL}$ and $\hat{\mathbf{P}}_{NPL}$. The proof closely follows the proof of Proposition 7. Define $\bar{l}_\zeta(\zeta,\mathbf{P}) = M^{-1}\sum_{m=1}^{M}\nabla_\zeta l(\zeta,\mathbf{P})(w_m)$, $\bar{l}_{\zeta\zeta}(\zeta,\mathbf{P}) = M^{-1}\sum_{m=1}^{M}\nabla_{\zeta\zeta'} l(\zeta,\mathbf{P})(w_m)$, and $\bar{l}_{\zeta\mathbf{P}}(\zeta,\mathbf{P}) = M^{-1}\sum_{m=1}^{M}\nabla_{\zeta\mathbf{P}'} l(\zeta,\mathbf{P})(w_m)$. 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

$$0 = \bar{l}_{\zeta\zeta}(\bar{\zeta},\bar{\mathbf{P}})(\tilde{\zeta}_j - \hat{\zeta}) + \bar{l}_{\zeta P}(\bar{\zeta},\bar{\mathbf{P}})(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}), \tag{26}$$

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 Proposition 7 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 $\tilde{\mathbf{P}}_j = \mathbf{\Psi}(\tilde{\zeta}_j,\tilde{\mathbf{P}}_{j-1})$ twice around $(\hat{\zeta},\hat{\mathbf{P}})$, using $\hat{\mathbf{P}} = \mathbf{\Psi}(\hat{\zeta},\hat{\mathbf{P}})$, and proceeding as in the proof of Proposition 7 gives

$$\tilde{\mathbf{P}}_j - \hat{\mathbf{P}} = \mathbf{\Psi}_P^0(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}) + \mathbf{\Psi}_\zeta^0(\tilde{\zeta}_j - \hat{\zeta}) + O(M^{-1/2}||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}||) + O(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}||^2), \tag{27}$$

a.s, where $\mathbf{\Psi}_\zeta^0 \equiv \nabla_{\zeta'}\mathbf{\Psi}(\theta^0,\mathbf{P}^0) = [\mathbf{\Psi}_\theta^0,\mathbf{0}]$. As in the proof of Proposition 7, refine (26) further as $\tilde{\zeta}_j - \hat{\zeta} = -\Omega_{\zeta\zeta}^{-1}\Omega_{\zeta P}(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}) + O(M^{-1/2}||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}||) + O(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}||^2)$ a.s., where $\Omega_{\zeta P} = E\left[\nabla_\zeta l(\zeta^0,\mathbf{P}^0)(w_m)\nabla_{\mathbf{P}'} l(\zeta^0,\mathbf{P}^0)(w_m)\right]$. Substituting this into (27) gives $\tilde{\mathbf{P}}_j - \hat{\mathbf{P}} = [\mathbf{\Psi}_P^0 - \mathbf{\Psi}_\zeta^0\Omega_{\zeta\zeta}^{-1}\Omega_{\zeta P}](\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}) + O(M^{-1/2}||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}||) + O(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}||^2)$ a.s. Note that $\Omega_{\zeta\zeta}$ and $\Omega_{\zeta P}$ are written as

$$\Omega_{\zeta\zeta} = \begin{bmatrix} \Omega_{\theta\theta} & \Omega_{\theta\pi} \\ \Omega_{\pi\theta} & \Omega_{\pi\pi} \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi}_\theta^{0\prime}L_P'\Delta_L L_P\mathbf{\Psi}_\theta^0 & \mathbf{\Psi}_\theta^{0\prime}L_P'\Delta_L L_\pi \\ L_\pi'\Delta_L L_P\mathbf{\Psi}_\theta^0 & L_\pi'\Delta_L L_\pi \end{bmatrix}, \quad \Omega_{\zeta P} = \begin{bmatrix} \Omega_{\theta P} \\ \Omega_{\pi P} \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi}_\theta^{0\prime}L_P'\Delta_L L_P\mathbf{\Psi}_P^0 \\ L_\pi'\Delta_L L_P\mathbf{\Psi}_P^0 \end{bmatrix},$$

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 $D = (\mathbf{\Psi}_\theta^{0\prime}L_P'\Delta_L^{1/2}M_{L_\pi}\Delta_L^{1/2}L_P\mathbf{\Psi}_\theta^0)^{-1}$ 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 $\mathbf{\Psi}_\zeta^0 = [\mathbf{\Psi}_\theta^0,\mathbf{0}]$ gives $\mathbf{\Psi}_\zeta^0\Omega_{\zeta\zeta}^{-1}\Omega_{\zeta P} = \mathbf{\Psi}_\theta^0 D\mathbf{\Psi}_\theta^{0\prime}L_P'\Delta_L^{1/2}M_{L_\pi}\Delta_L^{1/2}L_P\mathbf{\Psi}_P^0$, 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 two-step (PML) version of the four estimators (NPL, NPL-Λ, approximate RPM, approximate q-NPL) discussed in the paper and Appendix C. These estimators are included for reference; they do not need iteration but require a root-M consistent initial nonparametric estimate of P. They are denoted by "PML-Ψ," "PML-Λ," "PML-RPM," and "PML-Λ^q," respectively. We do not report PML-Λ estimate of θ because it is identical to PML-Ψ. The PML-RPM and the PML-Λ^q take one approximate RPM and approximate q-NPL step, respectively, from the original PML estimator with Ψ and, thus, they are three step estimators. Their asymptotic properties can be easily derived from Proposition 1 of AM07, apart from changes in regularity conditions. 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.

The PML-RPM and the PML-Λ^q perform substantially better than the PML-Ψ, suggesting that our proposed alternative sequential methods are useful even when the researcher wants to make just one NPL iteration rather than iterate the NPL algorithm until convergence.

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
	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-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
$\hat{\theta}_{RN}$	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
	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-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
$100\times \hat{P}$	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
	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

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

References

Andrews, D. (1992). "Generic Uniform Convergence." *Econometric Theory* 8: 241-257.

Aguirregabiria, V. and P. Mira (2002). "Swapping the nested fixed point algorithm: a class of estimators for discrete Markov decision models." *Econometrica* 70(4): 1519-1543.

Aguirregabiria, V. and P. Mira (2007). "Sequential estimation of dynamic discrete games." *Econometrica* 75(1): 1-53.

Chu, K-W E. (1990). "On multiple eigenvalues of matrices depending on several parameters," *SIAM Journal of Numerical Analysis* 27(5): 1368-1385.

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.

Horn R. A. and C. R. Johnson (1985) *Matrix Analysis*. Cambridge University 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 (2009). "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, 77(1): 135-175.

Pesendorfer, M. and P. Schmidt-Dengler (2010). "Sequential estimation of dynamic discrete games: a comment," *Econometrica* 78, 833-842.

Seber, G. A. F. (2007). *A Matrix Handbook for Statisticians*. Wiley.

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.