Sequential Estimation of Structural Models with Fixed Point Constraint*

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

Katsumi Shimotsu
Department of Economics
Queen's University
shimotsu@econ.queensu.ca

September 23, 2008

Abstract

This paper considers the estimation problem of structural models of which empirical restrictions are characterized in terms of a fixed point constraint, such as a structural dynamic discrete choice model and a model of dynamic games. We analyze the conditions under which the nested pseudo-likelihood (NPL) algorithm achieves convergence and derive its convergence rate. We find that the NPL algorithm may not necessarily converge when the fixed point mapping does not have a local contraction property. To address the issue of non-convergence, we propose alternative sequential estimation procedures that can achieve convergence even when the NPL algorithm does not. Upon convergence, some of our proposed estimation algorithms produce more efficient estimators than the NPL estimator. We also show that a similar convergence result holds for models with permanent unobserved heterogeneity. Further, we develop a recursive extension of popular two-step moment methods called the nested generalized method of moments (GMM) algorithm and show that its convergence properties are similar to those of the NPL algorithm.

Keywords: approximate maximum likelihood, contraction, dynamic games, nested generalized method of moments, nested pseudo likelihood, unobserved heterogeneity.

JEL Classification Numbers: C13, C14, C63.

*We are grateful to Victor Aguirregabiria, Kenneth Judd, Whitney Newey, and seminar participants at Far Eastern Summer Meeting of the Econometric Society, New York Camp Econometrics, North American Summer Meeting of the Econometric Society, Vienna Macroeconomic Workshop, University of British Columbia, Hitotsubashi University, University of Tokyo, University of Western Ontario, Yokohama National University for helpful comments. The authors thank the SSHRC for financial support.

1 Introduction

Empirical implications of economic theory are often characterized by fixed point problems. Upon estimating such models, researchers typically consider a class of extremum estimators with a fixed point constraint:

$$\max_{\theta\in\Theta} \quad Q_n(P) \quad s.t. \quad P = \Psi(P,\theta), \tag{1}$$

where $Q_n(P) = n^{-1}\sum_{i=1}^n \ln P(Z_i)$ for maximum likelihood estimator (MLE, hereafter) while $Q_n(P) = -\left[n^{-1}\sum_{i=1}^n g(Z_i,P)\right]' \hat{W} \left[n^{-1}\sum_{i=1}^n g(Z_i,P)\right]$ for the generalized method of moments estimator (GMM, hereafter) with the moment condition $E_{P^0}[g(Z_i,P^0)] = 0$.

The fixed point constraint $P = \Psi(P,\theta)$ in (1) summarizes the set of structural restrictions of the model that is parametrized by a vector $\theta \in \Theta$. When the model is correctly specified, the probability distribution obtained as the fixed point of the operator Ψ evaluated at the true parameter θ^0 generates the sample data. Examples of the operator $\Psi(\cdot,\theta)$ include the policy iteration operator for a single agent dynamic programming model (e.g., Rust (1987), Hotz and Miller (1993)), and an operator defined by the best response function of a games (e.g., Bajari, Benkard, and Levin (2007), Pakes, Ostrovsky and Berry (2007), Pesendorfer and Schmidt-Dengler (2007)).

In principle, we may estimate the parameter θ in (1) by repeatedly solving the fixed point P_θ of $P = \Psi(P,\theta)$ at each parameter value to maximize the objective function with respect to θ. The major practical obstacle of applying such an estimation procedure lies in the computational burden because solving the fixed point problem for a given parameter can be very costly.

To reduce the computational burden, Hotz and Miller (1993) developed a simpler two-step estimator that does not require solving the fixed point problem for each trial value of the parameter. A number of recent papers in empirical industrial organization build on the idea of Hotz and Miller (1993) to develop two-step estimators for models with multiple agents (e.g., Bajari, Benkard, and Levin, 2007; Pakes, Ostrovsky, and Berry, 2007; Pesendorfer and Schmidt-Dengler, 2007; Bajari, Chernozhukov, and Hong, 2006). These two-step estimators may suffer from substantial finite sample bias, however, when the choice probabilities are poorly estimated in the first step. This drawback is especially severe in estimating models with unobserved heterogeneity because it is difficult to obtain consistent initial estimates of the choice probabilities.

To address the limitations of two-step estimators, Aguirregabiria and Mira (2002)(2007, AM07 hereafter) develop a recursive extension of the two-step method of Hotz and Miller (1993), called the *nested pseudo likelihood (NPL) algorithm*. Starting from an initial estimate $\tilde{P}_0$, their algorithm iterates

Step 1: Given $\tilde{P}_{j-1}$, update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \ln[\Psi(\tilde{P}_{j-1},\theta)](Z_i)$.

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

until $j = k$. AM07 show that their method can be applied to models with unobserved heterogeneity in the context of dynamic discrete games, and the NPL estimator—defined as the limit of the sequence generated by the NPL algorithm—is more efficient than the two-step estimators *if the convergence is achieved*.

While the NPL algorithm provides an attractive apparatus for empirical researchers, little is known about its convergence property. In their simulations, AM07 report that the NPL algorithm converges and the NPL estimator performs very well relative to the two-step estimator. On the other hand, Pesendorfer and Schmidt-Dengler (2007) provide simulation evidence that the NPL algorithm may not necessarily converge. Collard-Wexler (2006) uses the NPL algorithm to estimate a structural model of entry and exit for the ready-mix concrete industry and finds that $\tilde{P}_j$'s "cycle around several values without converging." In view of these mixed evidences and its practical importance, it is imperative that we understand the convergence property of the NPL algorithm.

This paper makes two main contributions. First, this paper derives the condition under which the NPL algorithm converges. We show that a key determinant of the convergence property of the NPL algorithm is the *contraction* property of the mapping Ψ. Intuitively, the faster the operator achieves contraction, the closer the value obtained after one iteration is to the fixed point, and, therefore, we expect that the NPL algorithm works well if the operator has a good contraction property. The dominant eigenvalue of the Jacobian matrix $\partial\Psi(P,\theta)/\partial P$ determines the contraction property of Ψ, and it is shown that the NPL algorithm converges if the dominant eigenvalue of $\partial\Psi(P,\theta)/\partial P$ evaluated at the fixed point P_θ is sufficiently smaller than one in absolute value.

The violation of the convergence condition of the NPL algorithm is a practical concern. Using the dynamic discrete game model of AM07, we find that, when the degree of strategic substitutability is high, the smallest eigenvalue of the Jacobian matrix of the policy iteration mapping is less than -1, leading to no convergence of the NPL algorithm. In such cases, various two step estimators can be used but they may suffer from the finite sample bias.

Second, we propose alternative sequential algorithms (estimators) that are implementable even when the original NPL algorithm does not converge. The first estimator replaces the fixed point mapping $\Psi(P,\theta)$ in the NPL algorithm with $\Lambda(P,\theta) = [\Psi(P,\theta)]^\alpha P^{1-\alpha}$ that shares the same fixed point as Ψ. With an appropriate choice of α and under some conditions on Ψ, the mapping Λ has better contraction property than Ψ.

The second algorithm requires more computation than the first algorithm but converges under very general conditions. It uses the idea of the Recursive Projection Method (RPM) of Shroff and Keller (1993). The divergence of the fixed point mapping Ψ is often caused by a small number of eigenvalues of $\partial\Psi(P_\theta,\theta)/\partial P$ being outside of the unit circle. The key idea behind the RPM is to find the eigenvectors corresponding to the unstable modes and to decompose the probability space into the unstable subspace and its orthogonal complement. Then, it modifies

the fixed point mapping Ψ by taking a Newton-Raphson step on the unstable subspace while using the fixed point iteration on the stable subspace. The modified mapping then becomes contractive.

The third algorithm we propose requires more computation but gives the maximum likelihood estimator upon convergence. The algorithm is based on directly approximating the fixed point of the mapping. Given an initial consistent estimator, taking one step of this sequential algorithm leads to an estimator that is asymptotically equivalent to the MLE. Taking additional steps produces a sequence of estimators that approaches the MLE in higher orders.

The fourth estimator uses a pseudo-likelihood objective function that is defined in terms of multiple iterations of the mapping as opposed to one iteration. Such a modification, however, leads to an increase in the computational cost because repeated evaluations of the mapping is required for solving the optimization problem in Step 1. We, therefore, introduce an approximation method that requires evaluating the mapping and its Jacobian with respect to the parameter θ only once outside of the optimization routine. This algorithm converges faster than the original NPL algorithm and, upon convergence, the proposed estimator is more efficient than the NPL estimator.

We also analyze the convergence properties of the NPL algorithm when it is applied to models with permanent unobserved heterogeneity. Furthermore, we develop a recursive extension of two-step moment estimators, called the *nested generalized method of moments (GMM) algorithm* and show that the convergence of the nested GMM algorithm also requires that all the eigenvalues of $\partial\Psi(P_\theta, \theta)/\partial P$ are sufficiently smaller than one in absolute value.

The reminder of the paper is organized as follows. Section 2 introduces a class of models with fixed point constraints. Section 3 establishes the convergence property of the NPL algorithm. In Section 4, we develop alternative sequential algorithms that can be used even when the NPL algorithm has convergence problems. Section 5 extends our analysis to the sequential GMM estimator while section 6 applies our proposed methods to models with unobserved heterogeneity. Section 7 reports some simulation results. Section 8 concludes the paper.

2 The models with fixed point constraint and maximum likelihood estimator

We consider a class of parametric models of which restrictions are characterized in terms of fixed point problems in probability space. Upon estimating such models, researchers may consider the (conditional) maximum likelihood estimator (MLE) with fixed point constraint:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \left\{ \max_{P \in \mathcal{M}_\theta} \; n^{-1} \sum_{i=1}^{n} \ln P(a_i|x_i) \right\}, \tag{2}$$

where

$$\mathcal{M}_\theta = \{P \in B_P : \ P = \Psi(P, \theta)\} \tag{3}$$

is a set of fixed points of $\Psi(\cdot, \theta)$ given the value of $\theta \in \Theta \subset \mathbb{R}^K$. Here, B_P represents the space of conditional probabilities while Θ is the set of possible parameter values. The model space—the set of probabilities that are consistent with the parametric fixed point restrictions—is then defined as a union of $\mathcal{M}_\theta$ over Θ: $\mathcal{M} = \cup_{\theta \in \Theta} \mathcal{M}_\theta = \{P \in B_P : \ P = \Psi(P, \theta), \ \theta \in \Theta\}$. We assume that the model is correctly specified so that the conditional probability in population, denoted by P^0, belongs to the model space $\mathcal{M}$, i.e., $P^0 \in \mathcal{M}$.

The fixed point constraint $P = \Psi(P, \theta)$ in (3) summarizes the set of structural restrictions of the model that is parametrized with a finite K dimensional vector θ. For each θ, an operator $\Psi(\cdot, \theta)$ maps the space of conditional choice probabilities into itself. When the model is correctly specified, the true probability distribution P^0 is the fixed point of the operator Ψ evaluated at the true parameter θ^0, from which the sample data is generated. The examples of operator $\Psi(\cdot, \theta)$ include the policy iteration operator for single agent dynamic programming models (e.g., Rust (1987), Hotz and Miller (1993), Aguirregabiria and Mira (2002)) and an operator defined by best response functions for dynamic games (e.g., Aguirregabiria and Mira (2007), Bajari, Benkard and Levin (2007), Pakes, Ostrovsky and Berry (2007), Pesendorfer and Schmidt-Dengler (2007)).

Example 1 (A dynamic discrete choice model and the policy iteration mapping) *An agent maximizes the expected discounted sum of utilities, $E[\sum_{j=0}^{\infty} \beta^j \{u(x_{t+j}, a_{t+j}; \theta) + \epsilon_{t+j}(a_{t+j})\} | a_t, x_t; \theta]$, where x_t is an observable state variable and $\epsilon_t(a_t)$ is a state variable that are known to the agent but not to the researcher. The Bellman equation for this dynamic optimization problem is*

$$V(x) = \int \max_{a \in A} \left\{ u(x, a; \theta) + \epsilon(a) + \beta \sum_{x' \in X} V(x') f(x'|x, a; \theta) \right\} g(d\epsilon|x), \tag{4}$$

where $\beta \in (0, 1)$ is the discount factor, $g(\epsilon|x)$ is the joint distribution of $\epsilon = \{\epsilon(j) : j \in A\}$ and $f(x'|x, a; \theta)$ is transition function. For each value of θ, we may compute the fixed point of the Bellman equation and the conditional choice probability is given by

$$P_\theta(a|x) = \int 1 \left\{ a = \arg\max_{j \in A} \left[u(x, j; \theta) + \epsilon(j) + \beta \sum_{x' \in X} V_\theta(x') f(x'|x, j; \theta) \right] \right\} g(d\epsilon|x), \tag{5}$$

where V_θ is the fixed point of (4). Using the Hotz and Miller (1993)'s invertibility proposition, we may derive the policy iteration mapping in the space of conditional choice probability, $\Psi(\cdot, \theta)$, of which fixed point is the same as the conditional choice probabilities in (5), i.e., $P_\theta = \Psi(P_\theta, \theta)$.

Example 2 (A dynamic discrete game) *Consider the model of dynamic discrete games studied by Aguirregabiria and Mira (2007). There are N global firms who are the potential entrants*

in M separate markets. At the beginning of each period, a firm makes an entry/exit choice in each market, i.e., $a_{it} \in A = \{0, 1\}$. The profit of firm i operating in period t depends on the vector of current firms' current decision $a_t = (a_{1t}, ..., a_{Nt})'$, the market demand condition S_t, its previous entry decision $a_{i,t-1}$, and the vector of firms's state that is private information to each firm $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_{Nt})'$. Let $\tilde{\Pi}_i(a_t, S_t, a_{t-1}, \epsilon_t; \theta)$ be firm i's profit in period t. Then, firm i maximizes the expected discounted sum of profits $E\left[\sum_{t=0}^{\infty} \beta^t \Pi_i(a_t, S_{mt}, a_{i,t-1}, \epsilon_{it}; \theta) | S_{mt}, a_{m,t-1}; \theta\right]$. We assume that S_t follows an exogenous first-order Markov process $f_S(S_{t+1}|S_t, a_{t-1}; \theta)$, which is common knowledge while ϵ_{it} is iid across markets and firms conditional on S_t and a_{t-1}.

Let $\sigma^*(\theta) = \{\sigma_i^*(S_t, a_{t-1}, \epsilon_{it}; \theta) : i = 1, \ldots, N\}$ denote a set of strategy functions in a stationary Markov perfect equilibrium (MPE) given θ. Then, the equilibrium conditional choice probabilities are given by

$$P_i^{\sigma^*(\theta)}(a_i|S_t, a_{t-1}) = \int 1\{a_i = \sigma_i^*(S_t, a_{t-1}, \epsilon; \theta)\} g(d\epsilon|S_t, a_{t-1}), \tag{6}$$

where $g(\epsilon|S_t, a_{t-1})$ is the conditional distribution function for $\epsilon = \{\epsilon(a) : a \in A\}$. Aguirregabiria and Mira (2007) provides a best response mapping in probability space of which fixed point is identical to the equilibrium conditional choice probabilities in (6) so that $P_\theta = \Psi(P_\theta, \theta)$ where $P_\theta = \{P_i^{\sigma^*(\theta)} : i = 1, ..., N\}$.

The computation of the maximum likelihood estimator (MLE) in (2) requires repeatedly solving all the fixed points of $P = \Psi(P, \theta)$ at each parameter value to maximize the objective function with respect to θ. When there are multiple fixed points, finding all the fixed points of $P = \Psi(P, \theta)$ may be computationally infeasible. Even if there is a unique fixed point for each θ, the MLE could be extremely computationally intensive when evaluating the mapping Ψ is costly. For example, the MLE is often impractical in estimating models of dynamic game in example 2 with the modest number of players since the state space increases at exponential rate as the number of players increases. One of the major econometric issues in estimating models with fixed point constraint is to develop an estimator that is computational simple and has good finite sample properties as an alternative to the MLE.

3 The nested pseudo likelihood algorithm

3.1 Asymptotic properties of the NPL estimator

This section reviews the properties of the two-step pseudo maximum likelihood (PML) estimator and the estimator generated by the nested pseudo likelihood (NPL) algorithm as discussed in Aguirregabiria and Mira (2002, 2007). They are feasible alternatives to the MLE.

The pseudo maximum likelihood (PML) estimator is $\hat{\theta}_{PML} = \arg\max_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} \ln \Psi(\hat{P}_0, \theta)(a_i|x_i)$, where $\hat{P}_0$ is an initial consistent estimator for P^0. We assume that the support of (a_i, x_i)

is finite, $A \times X = \{a^1, a^2, ..., a^{|A|}\} \times \{x^1, x^2, \ldots, x^{|X|}\}$. Accordingly, P is represented with a $L \times 1$ vector while, given θ, the Jacobian $\nabla_{P'}\Psi(P,\theta)$ is a $L \times L$ matrix, where $L = |A||X|$. Let $\mathcal{N}$ denote a neighborhood of (P^0, θ^0), and let $\mathcal{N}_{\theta^0}$ denote a neighborhood of (θ^0).

Assumption 1 *(a) Θ is compact and, for any $\theta \in \Theta$, $\mathcal{M}_\theta$ is compact. (b) $\Psi(P,\theta)$ is three times continuously differentiable. (c) $\Psi(P,\theta)(a|x) > 0$ for any (a,x) and any $\{P,\theta\} \in B_P \times \Theta$. (d) (a_i, x_i) for $i = 1, 2, \ldots, N$, are independently and identically distributed, and $dF(x) > 0$ for any x in the support of x_i, where $F(x)$ is the distribution function of x_i. (e) There is a unique $\theta^0 \in int(\Theta)$ and a unique $P_{\theta^0} \in \mathcal{M}_{\theta^0}$ such that, for any $(a,x) \in A \times X$, $P_{\theta^0}(a|x) = P^0(a|x)$. For any $\theta \neq \theta^0$, $\Pr_{P^0}(\{(a,x) : \Psi(P^0,\theta)(a|x) \neq P^0(a|x)\}) > 0$. (g) $E_{\theta^0} \sup_{(P,\theta)\in\mathcal{N}} ||D^s\Psi(P,\theta)(a|x)||^2 < \infty$ for $s = 1, \ldots, 3$.*

As shown in Proposition 1 of AM07, under Assumption 1, the two-step PML estimator is consistent and, when a root-n consistent estimator of P^0 is available, it is asymptotically normal.

Proposition 1 *Assume Assumption 1 holds and $\hat{P}_0 \to_p P^0$. Then $\hat{\theta}_{PML} \to_p \theta^0$.*

Proposition 2 *Assume Assumption 1 holds and $\sqrt{n}(\hat{P}_0 - P^0) \to_d N(0, \Sigma)$. Then, $\sqrt{n}(\hat{\theta}_{PML} - \theta^0) \to N(0, V_{PML})$, where $V_{PML} = (\Omega_{\theta\theta})^{-1} + (\Omega_{\theta\theta})^{-1}\Omega_{\theta P}\Sigma(\Omega_{\theta P})'(\Omega_{\theta\theta})^{-1}$ with*

$$\begin{aligned}
\Omega_{\theta\theta} &\equiv E[(\partial/\partial\theta)\ln\Psi(P^0,\theta^0)(a|x)(\partial/\partial\theta')\ln\Psi(P^0,\theta^0)(a|x)] = -E[(\partial^2/\partial\theta\partial\theta')\ln\Psi(P^0,\theta^0)(a|x)], \\
\Omega_{\theta P} &\equiv E[(\partial/\partial\theta)\ln\Psi(P^0,\theta^0)(a|x)(\partial/\partial P')\ln\Psi(P^0,\theta^0)(a|x)] = -E[(\partial^2/\partial\theta\partial P')\ln\Psi(P^0,\theta^0)(a|x)].
\end{aligned}$$

The second term of the variance expression, $(\Omega_{\theta\theta})^{-1}\Omega_{\theta P}\Sigma(\Omega_{\theta P})'(\Omega_{\theta\theta})^{-1}$, captures the effect of the first step estimator $\hat{P}_0$ on $\hat{\theta}_{PML}$. When the estimator $\hat{P}_0$ is imprecise as is often the case in practice, the two-step PML estimator may perform poorly. The eigenvalues of the Jacobian matrix $\Psi_P \equiv (\partial/\partial P')\Psi(P^0,\theta^0)$ is another important determinant of the variance V_{PML}. If all the eigenvalues of Ψ_P are equal to zero, then $\Omega_{\theta P} = 0$ and there is no effect of $\hat{P}_0$ on $\hat{\theta}_{PL}$ in the first order asymptotic. In this case, the limiting distribution of the two-step estimator is the same as that of the MLE (cf., Aguirregabiria and Mira (2002)), which is true even under the weaker assumption that $\hat{P}_0 - P^0 = O_p(n^{-b})$ with $b > 1/4$ (see Kasahara and Shimotsu (2008a)).

Aguirregabiria and Mira (2002, 2007) consider a recursive extension of the two-step PML estimator based on the nested pseudo likelihood (NPL) algorithm as follows. Assume that an initial consistent estimator $\tilde{P}_0$ is available.

Step 1: Given $\tilde{P}_{j-1}$, update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \ln\Psi(\tilde{P}_{j-1},\theta)(a_i|x_i)$.

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

Iterate Steps 1-2 until $j = k$.

This procedure generates a sequence of estimators $\{\tilde{P}_j, \tilde{\theta}_j\}_{j=1}^k$. If this sequence converges, its limit $(\hat{P}_{NPL}, \hat{\theta}_{NPL})$ is called the *NPL estimator*, satisfying the following two conditions:

$$\hat{\theta}_{NPL} = \arg\max_{\theta\in\Theta} n^{-1} \sum_{i=1}^{n} \ln \Psi(\hat{P}_{NPL}, \theta)(a_i|x_i) \quad \text{and} \quad \hat{\theta}_{NPL} = \Psi(\hat{P}_{NPL}, \hat{\theta}_{NPL}). \tag{7}$$

The following proposition is from AM07 and states that $\hat{\theta}_{NPL}$ is root-n consistent and more efficient than a two-step estimator if all the eigenvalues of Ψ_P are between 0 and 1.

Proposition 3 *Assume Assumption 1 holds. Then,* $\sqrt{n}(\hat{\theta}_{NPL} - \theta^0) \to N(0, V_{NPL})$, *where* $V_{NPL} = [\Omega_{\theta\theta}+\Omega_{\theta P}(I-\Psi_P)^{-1}\Psi_\theta]^{-1}\Omega_{\theta\theta}\{[\Omega_{\theta\theta}+\Omega_{\theta P}(I-\Psi_P)^{-1}\Psi_\theta]^{-1}\}'$ *with* $\Psi_\theta \equiv (\partial/\partial\theta')\Psi(P^0, \theta^0)$. *Furthermore, if all the eigenvalues of* Ψ_P *are less than one in absolute value, then* $V_{PML} - V_{NPL}$ *is positive definite.*

The estimator $\hat{\theta}_{NPL}$ can be obtained as a limit of iterating steps 1 and 2 *if the iterations converge.* Although AM07 have obtained convergence in their simulations and illustrate that the estimator $\hat{\theta}_{NPL}$ performs very well relative to the PML estimator, they neither provide the conditions under which the NPL algorithm converges nor analyze how fast the convergence occurs. On the other hand, some other studies find potential problems on the convergence of the NPL algorithm. The simulation results of Pesendorfer and Schmidt-Dengler (2007) provide some evidence that the NPL algorithm may not necessarily converge. Collard-Wexler (2006) uses the NPL method to estimate a structural model of entry and exit for the ready-mix concrete industry and finds that the NPL algorithm generates a sequence of $\hat{P}_j$'s that is oscillating without converging. To date, little is known about the convergence properties of the NPL algorithm.

3.2 Convergence properties of the NPL algorithm

We now analyze the conditions under which the NPL algorithm achieves convergence and derives its convergence rates. We show that its convergence property crucially depends on the eigenvalues of Ψ_P. In particular, if all the eigenvalues of Ψ_P are sufficiently smaller than 1 in absolute value, then the NPL algorithm converges.

First, we state the regularity conditions.

Assumption 2 *Assumption 1 holds, and* $\sup_{(P,\theta)\in\mathcal{N}} ||D^s\Psi(P,\theta)|| < \infty$ *for* $s = 1, 2$.

Define $f_x(x_l) = \Pr(x = x^l)$ and let f_x be a $L \times 1$ vector of $\Pr(x = x^l)$ whose elements are arranged conformably with $P_{\theta^0}(a^j|x^l)$. Let $\Delta_P = diag(P^0)^{-1}diag(f_x)$. With these notations, we may write $\Omega_{\theta\theta} = \Psi_\theta'\Delta_P\Psi_\theta$ and $\Omega_{\theta P} = \Psi_\theta'\Delta_P\Psi_P$.

The following lemma is one of the main results of this paper. It states the local convergence rate of the NPL algorithm.

Lemma 1 *Suppose Assumption 2 holds. Then, for* $j = 1, \ldots, k$,

$$\begin{aligned}\tilde{\theta}_j - \hat{\theta}_{NPL} &= O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||), \\ \tilde{P}_j - \hat{P}_{NPL} &= M_{\Psi_\theta}\Psi_P(\tilde{P}_{j-1} - \hat{P}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2),\end{aligned}$$

where $M_{\Psi_\theta} \equiv I - \Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P$.

It follows from induction that

$$\tilde{P}_k - \hat{P}_{NPL} = (M_{\Psi_\theta}\Psi_P)^k(\tilde{P}_0 - \hat{P}_{NPL}) + O((M_{\Psi_\theta}\Psi_P)^{k-1})[O_p(n^{-1/2}||\tilde{P}_0 - \hat{P}_{NPL}||) + O_p(||\tilde{P}_0 - \hat{P}_{NPL}||^2)].$$

If all the eigenvalues of $M_{\Psi_\theta}\Psi_P$ are less than 1 in absolute value, an iteration moves $\tilde{P}_j$ toward $\hat{P}_{NPL}$. As we discuss in the next section, since the eigenvalues of M_{Ψ_θ} are either zero or one, the convergence property of $(M_{\Psi_\theta}\Psi_P)^k$ as $k \to \infty$ is primarily determined by the dominant eigenvalues of Ψ_P. If all the eigenvalues of Ψ_P is sufficiently smaller than 1 in absolute value, then the dominant eigenvalue of $M_{\Psi_\theta}\Psi_P$ is smaller than 1 in absolute value, and $\tilde{P}_k, \tilde{\theta}_k$ converges to $\hat{P}_{NPL}, \hat{\theta}_{NPL}$ as $k \to \infty$. In contrast, if some eigenvalues of $M_{\Psi_\theta}\Psi_P$ are outside of the unit circle, then an iteration moves some elements of $\tilde{P}_j$ further away from $\hat{P}_{NPL}$. In this case, it is not clear whether the iterations eventually converge even when the initial estimate $\tilde{P}_0$ is in the neighborhood of $\hat{P}_{NPL}$.

Remark 1 $\Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P$ *is a generalized least squares projection matrix from a regression of an element of* B_P *onto the space spanned by* Ψ_θ *while* M_{Ψ_θ} *is the orthogonal projection matrix that generates the "residuals".*

Remark 2 *Even if the initial estimate,* $\tilde{P}_0$*, is not root-n consistent, iterations reduce the effect of the initial estimate on* $\tilde{\theta}_j$*, provided all the eigenvalues of* $M_{\Psi_\theta}\Psi_P$ *are smaller than 1 in absolute value.*

Remark 3 *If all the eigenvalues of* $M_{\Psi_\theta}\Psi_P$ *are smaller than 1 in absolute value and we choose* $k \to \infty$ *so that* $\log n = o(k)$*, then* $\tilde{P}_k - \hat{P}_{NPL} = o_p(n^{-1/2})$ *and the effect of* $\tilde{P}_0$ *on* $\hat{P}_{NPL}$ *vanishes in the limit. This is useful when some elements of* x *are continuously distributed and root-n consistent* $\tilde{P}_0$ *is not available.*

Remark 4 *When* $\Psi_P = 0$*, the convergence rate is faster than linear:* $\tilde{P}_j - \hat{P}_{NPL} = O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2)$.

Remark 5 *If at least one element of* x_i *is continuously distributed, one can prove the higher-order improvement by bootstrap as in Kasahara and Shimotsu (2008a).*

3.3 The convergence property of $(M_{\Psi_\theta}\Psi_P)^k$ as $k \to \infty$

For a given matrix A, the convergence property of A^k as $k \to \infty$ is determined by its spectral radius. The spectral radius of A is defined as (cf. Horn and Johnson, 1985): $\rho(A) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}$. Then $A^k \to 0$ as $k \to \infty$ if and only if $\rho(A) < 1$ (Horn and Johnson, 1985, Theorem 5.6.12). Hence, the NPL algorithm converges if and only if $\rho(M_{\Psi_\theta}\Psi_P) < 1$. Because Ψ_P is closely related to the property of the economic model, we want to find a bound of $\rho(M_{\Psi_\theta}\Psi_P)$ in terms of $\rho(\Psi_P)$. The spectral radius is, however, not submultiplicative; i.e., $\rho(AB) > \rho(A)\rho(B)$ is possible. Hence, we cannot simply bound $\rho(AB)$ by $\rho(A)\rho(B)$. In the following, we give two discussions on the relation between $\rho(M_{\Psi_\theta}\Psi_P)$ and $\rho(\Psi_P)$. The first one uses the fact that M_{Ψ_θ} is a projection matrix, and the second one relies more on matrix algebra.

3.3.1 Projection by M_{Ψ_θ} and eigenvalues of $M_{\Psi_\theta}\Psi_P$

Let $P_{\Psi_\theta} = \Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P$. P_{Ψ_θ} is an GLS projection matrix, or a projection matrix corresponding to a weighted least squares in which the weights are given by the elements of $\Delta_P^{1/2}$. Since P_{Ψ_θ} is a projection matrix, we may decompose any $L-$vector x into two: $x = x_1 + x_2$, where $x_1 = P_{\Psi_\theta}x \in \mathcal{S}(\Psi_\theta)$ (the column space of Ψ_θ) and $x_2 = (I - P_{\Psi_\theta})x = M_{\Psi_\theta}x \in \mathcal{S}^\perp(\Delta_P\Psi_\theta)$ (the orthogonal compliment of $\Delta_P\Psi_\theta$).

Suppose y is an eigenvector of Ψ_P with non-zero eigenvalue ν so that $\Psi_P y = \nu y$ and $\nu \neq 0$. Apply the above decomposition to y and write $y = y_1 + y_2$, where $y_1 = P_{\Psi_\theta}y$ and $y_2 = M_{\Psi_\theta}y$. It follows that

$$M_{\Psi_\theta}\Psi_P y = \nu M_{\Psi_\theta}y = \nu M_{\Psi_\theta}(y_1 + y_2) = \nu y_2 = \nu y - \nu y_1.$$

Consider two extreme cases. First, suppose $y_1 = 0$, namely, the GLS regression of y on regressing on Ψ_θ gives no fit. In this case, $M_{\Psi_\theta}\Psi_P y = \nu y$, and $M_{\Psi_\theta}\Psi_P$ and Ψ_P share the same eigenvector y with eigenvalue ν. Second, suppose $y_2 = 0$, namely, the GLS regression of y on regressing on Ψ_θ gives a perfect fit. In this case, $M_{\Psi_\theta}\Psi_P y = 0$, and y is an eigenvector of $M_{\Psi_\theta}\Psi_P$ with eigenvalue 0. Therefore, when $y_1/||y|| \simeq 0$, it is likely that $M_{\Psi_\theta}\Psi_P y \simeq \nu y$, and there exists an eigenvector y^* of $M_{\Psi_\theta}\Psi_P$ with eigenvalue ν^* such that $y \simeq y^*$ and $\nu \simeq \nu^*$.[1] On the other hand, when $y_2/||y|| \simeq 0$, it is likely that there exist an eigenvector y^* of $M_{\Psi_\theta}\Psi_P$ with eigenvalue $\nu^* \simeq 0$ such that $y \simeq y^*$.

Now we place the above discussion in the context of our model. Recall that Ψ_θ is a $K \times L$ matrix, and typically $L \gg K$ because the dimension of the state variable is much larger than the number of parameters. Then, for many $L-$vectors y, regressing y on K regressors gives a poor fit, and it is likely that $y_2/||y|| \simeq 0$ so that the eigenvalues of Ψ_P and $M_{\Psi_\theta}\Psi_P$ are likely to be close. For a few y, we may have a good fit and $y_1/||y|| \simeq 0$, then the eigenvalue of $M_{\Psi_\theta}\Psi_P$ associated with such y is close to zero and is not likely to be the dominant eigenvalue.

[1] Strictly speaking, this requires that the mapping $f(z) = \{y, \lambda : (M_{\Psi_\theta}\Psi_P - \lambda I)y = z\}$ is continuous in a neighborhood of (y^*, ν^*).

Hence, we expect that the dominant eigenvalues of Ψ_P and $M_{\Psi_\theta}\Psi_P$ are close to each other. In our simulation exercise of section 7 with $L = 72$ and $K = 2$, we find either $y_2/||y|| \simeq 0$ or $y_1/||y|| \simeq 0$ hold for most of the eigenvectors, and the spectral radius of $M_{\Psi_\theta}\Psi_P$ is very similar to the spectral radius of Ψ_P (see Table 1).

3.3.2 The case when Ψ_P is diagonalizable

We can obtain a bound of $\rho(M_{\Psi_\theta}\Psi_P)$ if we assume Ψ_P is diagonalizable, i.e., $\Psi_P = SDS^{-1}$ for a diagonal matrix D. A matrix A is diagonalizable if all the eigenvectors are linearly independent (Horn and Johnson, 1985, Theorem 1.3.7). A sufficient condition for the diagonalizability of A is that the eigenvalues of A are distinct (Horn and Johnson, 1985, Theorem 1.3.9). Although economic models do not give implications for the diagonalizability of Ψ_P, we expect that A is diagonalizable in some, possibly many, cases.

For a matrix A, let $||A||_s$ denote its spectral norm: $||A||_s = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A'A\}$, which satisfies $||AB||_s \leq ||A||_s||B||_s$. Since $\rho(S^{-1}AS) = \rho(A)$, $\rho(A) \leq ||A||$ for any matrix norm $||\cdot||$, and $||D||_s = \rho(D)$ if D is diagonal, we have, if Ψ_P is diagonalizable, $\rho(M_{\Psi_\theta}\Psi_P) = \rho(M_{\Psi_\theta}SDS^{-1}) = \rho(S^{-1}M_{\Psi_\theta}SD) \leq ||S^{-1}M_{\Psi_\theta}S||_s||D||_s = ||S^{-1}M_{\Psi_\theta}S||_s\rho(\Psi_P)$. Consequently, $(M_{\Psi_\theta}\Psi_P)^k$ converges to 0 if Ψ_P is diagonalizable and $\rho(\Psi_P)$ is sufficiently smaller than 1.

This bound might not be so useful. But suppose we replace $\Psi(P,\theta)$ with its q-iterated version, $\Psi(P,\theta)^q$, based on which we develop an alternative algorithm in section 4.4. An argument analogous to above then gives $\rho(M_{\Psi_\theta^q}\Psi_P^q) \leq ||S^{-1}M_{\Psi_\theta^q}S||_s\rho(\Psi_P^q)$, where $\Psi_P^q \equiv \partial\Psi^q(P_\theta,\theta)/\partial P' = (\Psi_P)^q = SD^qS^{-1}$ and $\Psi_\theta^q \equiv \partial\Psi^q(P_\theta,\theta)/\partial\theta' = (I-\Psi_P)^{-1}(I-\Psi_P^q)\Psi_\theta$. Therefore, if $\rho(\Psi_P) < 1$, we have $\rho(M_{\Psi_\theta^q}\Psi_P^q) < 1$ for a sufficiently large q because $\max_{q\geq 1}||\Psi_\theta^q||_s$ is bounded.

4 Alternative sequential likelihood-based estimators

When a mapping $\Psi(P,\theta)$ is not a contraction in the neighborhood of (P^0,θ^0), the NPL algorithm has a convergence problem and therefore may not be used in practice. While the PML or other two-step estimators can be used in such cases, the finite sample bias is often a serious concern in these estimators. This section discusses alternative sequential algorithms that are implementable even when the NPL algorithm encounters a convergence problem.

4.1 Locally contractive mapping with geometric average

In this section, we propose implementing the NPL algorithm by modifying the mapping $\Psi(P,\theta)$ so that its transformed mapping has better contraction property. We consider a class of mappings that are obtained as a log-linear combination of $\Psi(P,\theta)$ and P:

$$[\Lambda(P,\theta)](a|x) \equiv \{[\Psi(P,\theta)](a|x)\}^\alpha P(a|x)^{1-\alpha} \tag{8}$$

for all $(a,x) \in A \times X$, where $\alpha \in [0,1]$. Given θ, $\Lambda(P,\theta)$ is a mapping from B_P into itself. Since P is a fixed point of $\Psi(P,\theta)$ if and only if it is a fixed point of $\Lambda(P,\theta)$, we may obtain the fixed point of $\Psi(P,\theta)$ by solving the fixed point of $\Lambda(P,\theta)$.

The following proposition states that, under certain conditions, we may choose the value of α so that the mapping $\Lambda(P,\theta)$ may become locally contractive with its dominant eigenvalue less than 1 even when the mapping $\Psi(P,\theta)$ is not locally contractive. Denote the largest and the smallest eigenvalues of Ψ_P by $\lambda_{\max}$ and λ_{min} while let α^* denote the value of α that minimizes the spectral radius of $\Lambda_P \equiv \nabla_{P'}\{\Psi(P,\theta)^\alpha P^{1-\alpha}\}|_{(P,\theta)=(P^0,\theta^0)}$, i.e., $\alpha^* = \arg\min_{\alpha\in[0,1]} \rho\left(\Lambda_P\right)$.

Proposition 4 *If $\lambda_{\max} > 1 > \lambda_{\min}$, then there is no value of α such that the spectral radius of Λ_P is less than one. If $1 > \lambda_{\max} > \lambda_{\min}$, then $\alpha^* = \frac{2}{2-\lambda_{\max}-\lambda_{\min}}$ and $\rho(\Psi_P) > \rho(\Lambda_P) = \frac{\lambda_{\max}-\lambda_{\min}}{2-\lambda_{\max}-\lambda_{\min}}$.*

We may consider the NPL algorithm using $\Lambda(P,\theta)$ in place of $\Psi(P,\theta)$. When the condition that $1 > \lambda_{\max} > \lambda_{\min}$ is satisfied, the sequence of estimators generated by the NPL algorithm with $\Lambda(P,\theta)$ may converge even if the NPL algorithm with $\Psi(P,\theta)$ does not converge. Furthermore, the limit of a sequence of estimators generated by the NPL algorithm with $\Lambda(P,\theta)$ satisfies the same first order conditions as that of (7) and it is identical to the original NPL estimator with $\Psi(P,\theta)$ upon convergence (see the Appendix B).

The advantage of this method is its simplicity. Once an appropriate value of α is determined, it achieves better convergence property than the original NPL algorithm without adding computational burden. The condition $1 > \lambda_{\max} > \lambda_{\min}$ may be restrictive in some cases but, in our Monte Carlo experiments using the model of Example 2, we find that $\lambda_{\max}$ is less than 1 while $\lambda_{\min}$ may become less than -1 when the degree of strategic substitutability is high.

4.2 Recursive Projection Method

This section constructs a mapping based on Ψ so that the transformed mapping shares the same fixed point as Ψ and yet has a better local contraction property by using the Recursive Projection Method (RPM) of Shroff and Keller (1993) [SK, hereafter].

Suppose that a small number, m, of the eigenvalues of Ψ_P are larger than δ in absolute value:

$$|\lambda_1| \geq \cdots \geq |\lambda_m| > \delta \geq |\lambda_{m+1}| \geq \cdots \geq |\lambda_L|.$$

for some $\delta \in (0,1)$. Define a subspace $\mathbb{P}$ as the maximal invariant subspace of Ψ_P belonging to $\{\lambda^k\}_{k=1}^m$ while let $\mathbb{Q} = \mathbb{R}^L - \mathbb{P}$ be the orthogonal complement of $\mathbb{P}$ in $\mathbb{R}^L$. Let $Z \in \mathbb{R}^{L\times m}$ be an orthonormal basis for $\mathbb{P}$. Denote the orthogonal projector of $\mathbb{R}^L$ onto the subspace $\mathbb{P}$ by $\Pi = ZZ'$. Then, for each $P \in \mathbb{R}^L$, we have the unique decomposition $P = u + v$, where $u = \Pi P \in \mathbb{P}$ and $v = (I - \Pi)P \in \mathbb{Q}$.

Now apply Π and $I-\Pi$ to $P=\Psi(P,\theta)$ and decompose it as follows:

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

Since $g(u,v,\theta)$ is locally contracting in v (see Lemma 2.10 of SK), we can update $v_{j-1}=(I-\Pi)P_{j-1}$ by the recursion $v_j=g(u,v_{j-1},\theta)$. On the other hand, the recursion $u_j=f(u_{j-1},v,\theta)$ cannot be used to update $v_{j-1}=(I-\Pi)P_{j-1}$ since $f(u,v,\theta)$ is not locally contracting in u. Instead, the RPM takes a Newton-Raphson step on the system $u=f(u,v,\theta)$, which leads to the following updating procedure:

$$\begin{aligned} u_j &= u_{j-1}+(I-\Pi\nabla_{P'}\Psi(P_{j-1},\theta)\Pi)^{-1}(f(u_{j-1},v_{j-1},\theta)-u_{j-1})\equiv h(u_{j-1},v_{j-1},\theta), \\ v_j &= g(u_{j-1},v_{j-1},\theta). \end{aligned} \tag{9}$$

Lemma 3.11 of SK shows that the Jacobian of the stabilized iteration (9) has the dominant eigenvalue of which modulus is less than δ, and thus the iteration is locally converging.

The updating procedure (9) can be summarized into the following mapping:

$$\Gamma(P,\theta)=h(u,v,\theta)+g(u,v,\theta)=\Pi P+(I-\Pi\nabla_{P'}\Psi(P,\theta)\Pi)^{-1}(\Pi\Psi(P,\theta)-\Pi P)+(I-\Pi)\Psi(P,\theta).$$

An NPL-type algorithm that uses $\Gamma(P,\theta)$ in place of $\Psi(P,\theta)$ would look promising because of the following reasons. First, the mapping $\Gamma(P,\theta)$ has the same fixed point as $\Psi(P,\theta)$; $P^0=\Gamma(P^0,\theta^0)$. Second, $\Gamma(P,\theta)$ is locally contracting. Third, because $\Gamma(P^0,\theta^0)$ gives the true conditional choice probability, an information matrix equality holds with respect $\Gamma(P,\theta)$, so that $E\nabla_{\alpha\beta'}\ln\Gamma(P^0,\theta^0)(a|x)=-E\nabla_{\alpha}\ln\Gamma(P^0,\theta^0)(a|x)\nabla_{\beta'}\ln\Gamma(P^0,\theta^0)(a|x)$ for $\alpha,\beta\in\{\theta,P\}$.

However, using $\Gamma(P,\theta)$ in the objective function is computationally demanding, because it requires evaluating Π and $\nabla_{P'}\Psi(P,\theta)$ for all the trial values of θ. To reduce the computational burden, we evaluate Π and $\nabla_{P'}\Psi(P,\theta)$ at a preliminary estimate of θ and P outside of the optimization routine and use these estimates to approximately evaluate the objective function $\Gamma(P,\theta)$ across different values of θ.

Let η be a preliminary estimate of θ. Let $Z(P,\eta)$ be an orthonormal basis for the space spanned by the eigenvectors associated with the eigenvalues of $\nabla_{P'}\Psi(P,\eta)$ that are larger than δ in absolute value. By replacing Π and $\nabla_{P'}\Psi(P,\theta)$ in $\Gamma(P,\theta)$ with $\Pi(P,\eta)\equiv Z(P,\eta)Z(P,\eta)'$ and $\nabla_{P'}\Psi(P,\eta)$, respectively, we obtain

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

Let $(\tilde{P}_0,\tilde{\theta}_0)$ be an initial consistent estimator of (P^0,θ^0). For instance, $\tilde{\theta}_0$ can be the PML estimator. The sequential procedure based on RPM iterates

Step 1: Given $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta \in \bar{\Theta}_j} n^{-1} \sum_{i=1}^n \ln \Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_i|x_i)$, where $\bar{\Theta}_j = \{\theta \in \Theta : \Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_i|x_i) \in [\epsilon, 1-\epsilon] \text{ for all } (a,x) \in A \times X\}$ for an arbitrary small $\epsilon > 0$. We impose this restriction in order to avoid computing $\ln(0)$.[2]

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

until $j = k$. If this sequence converges, its limit $(\tilde{P}, \tilde{\theta})$ satisfies the following two conditions:

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \ln \Gamma(\theta, \tilde{P}, \tilde{\theta})(a_i|x_i), \quad \text{and} \quad \tilde{P} = \Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta}).$$

Somewhat surprisingly, using $\Gamma(\theta, P, \eta)$ in place of $\Gamma(P, \theta)$ has only a second-order effect on the convergence property of the algorithm and the asymptotic distribution of its limit. To see why, with a slight abuse of notation, take a partial derivative of $\Gamma(\theta, P, \eta)$ with respect to the (i,j)th element of $\Pi(P, \eta)$. Because $P^0 = \Psi(P^0, \theta^0)$, evaluating it at $(\theta^0, P^0, \theta^0)$ gives $\partial \Gamma(\theta^0, P^0, \theta^0)/\partial \Pi_{ij}(P, \eta) = 0$ for all (i,j). Similarly, taking a partial derivative of $\Gamma(\theta, P, \eta)$ with respect to the (i,j)th element of $\nabla_{P'} \Psi(P, \eta)$ at $(\theta^0, P^0, \theta^0)$ gives $\partial \Gamma(\theta^0, P^0, \theta^0)/\partial \nabla_{P'} \Psi_{ij}(P, \eta) = 0$ for all (i,j). It follows that the derivatives of $\Gamma(\theta, P, \eta)$ satisfy

$$\nabla_{P'} \Gamma(\theta^0, P^0, \theta^0) = \nabla_{P'} \Gamma(P^0, \theta^0), \quad \nabla_{\eta'} \Gamma(\theta^0, P^0, \theta^0) = 0. \tag{10}$$

Further, (10) implies that we can construct the information matrix with respect to $\ln \Gamma(P, \theta)(a|x)$ from $\ln \Gamma(\theta, P, \eta)(a|x)$. Define the following information matrices: $\Omega^{\Gamma}_{\theta P} = E \nabla_\theta \ln \Gamma(P^0, \theta^0)(a|x) \times \nabla_{P'} \ln \Gamma(P^0, \theta^0)(a|x)$ and $\Omega^{\Gamma}_{\theta\theta} = E \nabla_\theta \ln \Gamma(P^0, \theta^0)(a|x) \nabla_{\theta'} \ln \Gamma(P^0, \theta^0)(a|x)$. Since $\nabla_\theta \Gamma(P^0, \theta^0) = \nabla_\theta \Gamma(\theta^0, P^0, \theta^0) + \nabla_\eta \Gamma(\theta^0, P^0, \theta^0) = \nabla_\theta \Gamma(\theta^0, P^0, \theta^0)$ and $\nabla_P \Gamma(P^0, \theta^0) = \nabla_P \Gamma(\theta^0, P^0, \theta^0)$, we can write $\Omega^{\Gamma}_{\cdot\cdot}$ in terms of $\Gamma(\theta, P, \eta)$ as

$$\Omega^{\Gamma}_{\alpha\beta} = E \nabla_\alpha \ln \Gamma(\theta^0, P^0, \theta^0)(a|x) \nabla_{\beta'} \ln \Gamma(\theta^0, P^0, \theta^0)(a|x) \text{ for } \alpha, \beta \in \{\theta, P\}. \tag{11}$$

By the virtue of these properties, this sequential algorithm has the following convergence property that is superior to the NPL algorithm. Define Γ_P and Γ_θ analogously to Ψ_P and Ψ_θ.

Proposition 5 *Suppose Assumption 2 holds. Then, for $j = 1, \ldots, k$,*

$$\begin{aligned}
\tilde{\theta}_j - \tilde{\theta} &= O_p(||\tilde{P}_{j-1} - \tilde{P}||) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \tilde{\theta}||) + O_p(||\tilde{\theta}_{j-1} - \tilde{\theta}||^2), \\
\tilde{P}_j - \tilde{P} &= M_{\Gamma_\theta} \Gamma_P (\tilde{P}_{j-1} - \tilde{P}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2) \\
&\quad + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \tilde{\theta}||) + O_p(||\tilde{\theta}_{j-1} - \tilde{\theta}||^2),
\end{aligned}$$

[2] In practice, we may consider a penalized pseudo likelihood objective function by truncating $\Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ so that its value takes between ϵ and $1-\epsilon$, and adding a penalty term that is increasing in the distance between $\Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ and the set $[\epsilon, 1-\epsilon]$.

where $M_{\Gamma_\theta} = I - \Gamma_\theta(\Gamma_\theta'\Delta_P\Gamma_\theta)^{-1}\Gamma_\theta'\Delta_P$.

Since the dominant eigenvalue of Γ_P is less than δ, the sequential algorithm with Γ in place of Ψ is converging even when the original NPL algorithm with Ψ is not.

Implementing the proposed sequential algorithm based on RPM requires evaluating $(I - \Pi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\Pi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}))^{-1}$ as well as computing an orthonormal basis $Z(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ from the eigenvectors of $\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ for $j = 1, ..., k$. This is potentially costly when the analytical expression of $\nabla_{P'}\Psi(P, \theta)$ is not available. In the following, we discuss how to further reduce the computational burden of implementing the sequential algorithm based on RPM.

First, we may verify that $(I - \Pi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\Pi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}))^{-1} = \tilde{Z}_{j-1}(I - (\tilde{Z}_{j-1})'\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\tilde{Z}_{j-1})^{-1}(\tilde{Z}_{j-1})'$. With $\tilde{Z}_{j-1} = [\tilde{z}^1_{j-1}, ..., \tilde{z}^m_{j-1}]$ and $\epsilon > 0$, the i-th row of $\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\tilde{Z}_{j-1}$ can be approximated by $\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\tilde{z}^i_{j-1} \approx (1/\epsilon)[\Psi(\tilde{P}_{j-1} + \epsilon\tilde{z}^i_{j-1}, \tilde{\theta}_{j-1}) - \Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})]$. As a result, evaluating $(I - \Pi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\Pi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}))^{-1}$ only requires the $(m+1)$ function evaluations of $\Psi(P, \theta)$ even when the analytical expression of $\nabla_{P'}\Psi(P, \theta)$ is not available.

Second, it is also possible to use $\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\tilde{Z}_{j-1}$ to update an estimate of orthogonal basis Z. Namely, given a preliminary estimate $\tilde{Z}_{j-1}$, we may perform one step of an orthogonal power iteration (cf., Golub and Van Loan, 1996) by computing $\tilde{Z}_j = \text{orth}(\nabla_{P'}\Psi(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})\tilde{Z}_{j-1})$, where "orth($B$)" denotes computing an orthonormal basis for the columns of B using Gram-Schmidt orthogonalization.

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

Step 0 (initialization): (a) Using Gram-Schmidt orthogonalization, compute an orthonormal basis, denoted by $\{\tilde{z}^1_0, ..., \tilde{z}^m_0\}$, of the space spanned by the eigenvectors of $\tilde{\Psi}_{P,0} \equiv \nabla_{P'}\Psi(\tilde{P}_0, \tilde{\theta}_0)$ associated with the eigenvalues $\{\tilde{\lambda}_{0,1}, ..., \tilde{\lambda}_{0,m}\}$ of which absolute values are larger than δ.[3] (c) 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 $\tilde{Z}_0 = [\tilde{z}^1_0, ..., \tilde{z}^m_0]$.

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

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

Step 3 (Update Z): (a) Update the orthonormal basis Z by $\tilde{Z}_j = \text{orth}(\tilde{\Psi}_{P,j}\tilde{Z}_{j-1})$, where the i-th row of $\tilde{\Psi}_{P,j}\tilde{Z}_{j-1}$ is computed by $\tilde{\Psi}_{P,j}\tilde{z}^i_{j-1} \approx (1/\epsilon)[\Psi(\tilde{P}_j + \epsilon\tilde{z}^i_{j-1}, \tilde{\theta}_j) - \Psi(\tilde{P}_j, \tilde{\theta}_j)]$ for small $\epsilon > 0$ with $\tilde{Z}_{j-1} = [\tilde{z}^1_{j-1}, ..., \tilde{z}^m_{j-1}]$. (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 $\tilde{\Psi}_{P,j}\tilde{z}^i_j \approx (1/\epsilon)[\Psi(\tilde{P}_j + \epsilon\tilde{z}^i_j, \tilde{\theta}_j) - \Psi(\tilde{P}_j, \tilde{\theta}_j)]$. (c) For

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

every J iterations, update the orthonormal basis Z using the algorithm of Step 0, where $(\tilde{P}_0, \tilde{\theta}_0)$ is replaced with $(\tilde{P}_j, \tilde{\theta}_j)$.

Step 4: Iterate Steps 1-3 until $j = k$.

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 orthonormal basis. If an initial estimate of the size of 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 basis of which eigenvalues are outside of the unit disk. 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).

4.3 The q-NPL algorithm

In this section, we consider a possible extension of the NPL algorithm by defining a q-stage operator of Λ by

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

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

Given an initial consistent estimator $\tilde{P}_0$, the q-NPL algorithm iterates

Step 1: Update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \ln \Lambda^q(\tilde{P}_{j-1},\theta)(a_i|x_i)$ and

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

until $j = k$. The limit of this sequence of estimators, denoted by $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$, satisfies

$$\hat{\theta}_{qNPL} = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \ln \Lambda^q(\hat{P}_{qNPL},\theta)(a_i|x_i) \quad \text{and} \quad \hat{\theta}_{qNPL} = \Lambda^q(\hat{P}_{qNPL},\hat{\theta}_{qNPL}), \tag{12}$$

if iterations converge. The estimator $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ is called the *q-NPL estimator*. Since the result of Lemma 1 also applies here by replacing Ψ with Λ^q, the dominant eigenvalue of $\Lambda_P^q \equiv \nabla_{P'}\Lambda^q(P^0,\theta^0)$ is the main determinant of the convergence rate of the q-NPL algorithm. When the dominant eigenvalue of Λ_P, denoted by λ^*, is less than 1 in absolute value, the q-NPL algorithm converges faster than the NPL algorithm because the absolute value of dominant eigenvalue of Λ_P^q is equal to $|\lambda^*|^q$. Furthermore, the variance of the q-NPL estimator approaches to that of the MLE at the exponential rate of $|\lambda^*|^{2q}$ as $q \to \infty$. See the Appendix B.

A simple application of the q-NPL algorithm is computationally intensive because computing Step 1 of the q-NPL algorithm requires repeatedly evaluating the mapping Λ at many different values of the vector of probabilities P. In contrast, an iteration of the NPL algorithm often requires evaluating the mapping Λ only once as discussed in Aguirregabiria and Mira (2002, 2007). For this reason, we consider the following *approximate q-NPL algorithm*.

Suppose that a consistent estimate $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ is available. Expanding $\Lambda^q(\tilde{P}_{j-1}, \theta)$ in Step 1 of the q-NPL algorithm gives

$$\Lambda^q(\tilde{P}_{j-1}, \theta) = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta'}\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\theta - \tilde{\theta}_{j-1}) + O(||\theta - \tilde{\theta}_{j-1}||^2). \tag{13}$$

Thus, $\Lambda^q(\tilde{P}_{j-1}, \theta)$ can be approximated by $\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta'}\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\theta - \tilde{\theta}_{j-1})$, and this approximation becomes exact as $\theta \to \tilde{\theta}_{j-1}$.

We propose to estimate θ using this approximation of $\Lambda^q(P, \theta)$. Given an initial consistent estimator $(\tilde{P}_0, \tilde{\theta}_0)$, the approximate q-NPL algorithm iterates

Step 1: Given $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta \in \Theta_j^q} n^{-1} \sum_{i=1}^n \ln \tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_i|x_i)$, where $\tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \equiv \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta'}\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\theta - \tilde{\theta}_{j-1})$ and $\Theta_j^q = \{\theta \in \Theta : \tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a|x) \in [\epsilon, 1-\epsilon]$ for all $(a, x) \in A \times X\}$ for an arbitrary small $\epsilon > 0$.

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

until $j = k$. 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, it involves much fewer number of evaluations of $\Lambda(P, \theta)$ across different values of θ and P than the original q-NPL algorithm.[4]

To establish the consistency of the sequence of estimators generated by the approximate q-NPL algorithm, we need the following assumption in addition to Assumption 1.

Assumption 3 *(a) For any $\eta \in \mathbb{R}^K$ such that $\eta \neq 0$, $\nabla_{\theta'}\Lambda^q(P^0, \theta^0)(a|x)\eta \neq 0$ with positive probability. (b) $E \sup_{\theta \in \Theta, (P^*, \theta^*) \in \mathcal{N}} ||\nabla_{\theta'}\tilde{\Lambda}^q(\theta, P^*, \theta^*)(a|x)|| < \infty$, and $E \sup_{\theta \in \Theta, (P^*, \theta^*) \in \mathcal{N}} ||\nabla_{P^{*\prime}}\tilde{\Lambda}^q(\theta, P^*, \theta^*)(a|x)|| < \infty$.*

Assumption 3(a) is an identification condition for the probability limit of our objective function and is required because we use an approximation of $\Lambda^q(P, \theta)(a|x)$ in the objective function. If this assumption is violated, then there exists a direction of θ such that $\nabla_{\theta'}\Lambda^q(P^0, \theta^0)(a|x)(\theta - \theta^0) = 0$ even when $\theta \neq \theta^0$. Then, it is not possible to identify θ^0. Assumption 3(a) is satisfied if the

[4] $\Lambda^q(\tilde{P}_0, \tilde{\theta}_0)$ can be computed by just iterating $\Lambda(\tilde{P}_0, \tilde{\theta}_0)$ q times while $\nabla_{\theta'}\Lambda^q(\tilde{P}_0, \tilde{\theta}_0)$ can be computed by taking a numerical derivative of $\Lambda^q(\tilde{P}_0, \tilde{\theta}_0)$ with respect to the parameter vector θ. Using one-sided numerical derivatives, Step 1 requires the $(K+1)q$ function evaluations of $\Psi(P, \theta)$ across different parameter values.

following $|X| \times K$ matrix has full column rank:

$$\begin{bmatrix} \nabla_{\theta'} \Lambda^q(P^0, \theta^0)(a|x = X_1) \\ \vdots \\ \nabla_{\theta'} \Lambda^q(P^0, \theta^0)(a|x = X_{|X|}) \end{bmatrix}.$$

Since $|X| \gg K$ in general, this condition is likely to be satisfied in most cases. Assumption 3(b) is required for the uniform convergence of the objective function.

Under these assumptions, we may establish consistency:

Proposition 6 *Suppose that Assumptions 1 and 3 hold and $(\tilde{P}_0, \tilde{\theta}_0)$ is consistent. Suppose we obtain $\tilde{\theta}_k$ by the approximate q-NPL algorithm. Then $\tilde{\theta}_k - \theta^0 = o_p(1)$ for $k = 1, 2, \ldots$*

Let $\nabla^{(3)} \tilde{\Lambda}^q(\theta, P^*, \theta^*)$ denote the third derivatives of $\tilde{\Lambda}^q(\theta, P^*, \theta^*)$ with respect to (θ, P^*, θ^*). Under the follo

Assumption 4 $E \sup_{\theta \in \mathcal{N}_{\theta^0}, (P^*, \theta^*) \in \mathcal{N}} ||\nabla^{(3)} \tilde{\Lambda}^q(\theta, P^*, \theta^*)(a|x)|| < \infty$.

Proposition 7 *Suppose Assumptions 1-4 hold and $(\tilde{P}_0, \tilde{\theta}_0)$ is consistent. Suppose we obtain $\{\tilde{P}_j, \tilde{\theta}_j\}_{j=1}^k$ by the approximated q-NPL algorithm. Then, for $j = 1, \ldots, k$,*

$$\begin{aligned} \tilde{\theta}_j - \hat{\theta}_{qNPL} &= O_p(||\tilde{P}_{j-1} - \hat{P}_{qNPL}||), \\ \tilde{P}_j - \hat{P}_{qNPL} &= M_{\Lambda_\theta^q} \Lambda_P^q (\tilde{P}_{j-1} - \hat{P}_{qNPL}) + O_p(n^{-1/2} ||\tilde{P}_{j-1} - \hat{P}_{qNPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{qNPL}||^2), \end{aligned}$$

where $M_{\Lambda_\theta^q} \equiv I - \Lambda_\theta^q ((\Lambda_\theta^q)' \Delta_P \Lambda_\theta^q)^{-1} (\Lambda_\theta^q)' \Delta_P$ with $\Lambda_\theta^q = \nabla_{\theta'} \Lambda^q(P^0, \theta^0)$.

Thus, the approximate q-NPL algorithm achieves the same convergence rate as the original q-NPL algorithm, improving the convergence property of the NPL algorithm if the dominant eigenvalue of Λ_P is less than 1 in absolute value. Upon convergence, this algorithm generates the q-NPL estimator defined by (12), which is more efficient than the NPL estimator $\hat{\theta}_{NPL}$.

4.4 Approximate fixed point algorithm

The approximation method similar to the approximate q-NPL algorithm can be directly applied to the fixed point, $P_\theta = \Psi(P_\theta, \theta)$, resulting in the approximation of the MLE. From Taylor expansion and using $\nabla_{\theta'} P_\theta = (I - \nabla_{P'} \Psi(P_\theta, \theta))^{-1} \nabla_{\theta'} \Psi(P_\theta, \theta)$, we can approximate P_θ as

$$P_\theta = P_{\theta^0} + (I - \nabla_{P'} \Psi(P_{\theta^0}, \theta^0))^{-1} \nabla_{\theta'} \Psi(P_{\theta^0}, \theta^0)(\theta - \theta^0) + O(||\theta - \theta^0||^2), \tag{14}$$

where $\nabla_{\theta'} P_{\theta^0}$ denotes the derivative of P_θ evaluated at $\theta = \theta^0$. Therefore, if we have a consistent estimate of θ^0 and P^0, we may approximate P_θ with the mappings $\nabla_{P'} \Psi(P, \theta)$ and $\nabla_{\theta'} \Psi(P, \theta)$. This approximation method is particularly useful when it is possible to derive an analytical expression for $\nabla_{P'} \Psi(P, \theta)$ and $\nabla_{\theta'} \Psi(P, \theta)$.

Consider the following objective function based on (14):

$$Q_n(\theta, P^*, \theta^*) = n^{-1} \sum_{i=1}^{n} \ln \Phi(\theta, P^*, \theta^*)(a_i|x_i),$$

where

$$\Phi(\theta, P^*, \theta^*) = P^* + (I - \nabla_{P'} \Psi(P^*, \theta^*))^{-1} \nabla_{\theta'} \Psi(P^*, \theta^*)(\theta - \theta^*). \tag{15}$$

We call the estimation algorithm using $Q_n(\theta, P^*, \theta^*)$ the *Approximate Fixed Point Algorithm (AFXP)* because it is based on the approximation of the fixed point P_θ.

Let $\tilde{\theta}_0$ be an initial estimator of θ^0, such as the PML estimator. For $j \geq 1$, consider the following sequential procedure.

Step 1: Given $\tilde{\theta}_{j-1}$, update P by solving the fixed point $P_{\tilde{\theta}_{j-1}} = \Psi(P_{\tilde{\theta}_{j-1}}, \tilde{\theta}_{j-1})$. If there are multiple fixed points, choose the one that maximizes the likelihood: $\tilde{P}_j = \arg\max_{P \in \mathcal{M}_{\tilde{\theta}_{j-1}}} \ln P(a_i|x_i)$, where $\mathcal{M}_\theta$ is defined in (3).

Step 2: Given $(\tilde{P}_j, \tilde{\theta}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta \in \Theta_j} Q_n(\theta, \tilde{P}_j, \tilde{\theta}_{j-1})$, where

$$\Theta_j = \{\theta \in \Theta : \Phi(\theta, \tilde{\theta}_{j-1}, \tilde{P}_j)(a|x) \in [\epsilon, 1-\epsilon] \text{ for all } (a,x) \in A \times X\} \tag{16}$$

for an arbitrary small $\epsilon > 0$.

Iterate Steps 1-2 until $j = k$.

To establish the consistency of sequential estimators generated by the AFXP algorithm, consider the following assumptions. The first set of the assumptions is regularity conditions for the consistency of the MLE. The second set of the assumptions is concerned with the NPL algorithm.

Assumption 5 *(a) Θ is compact and, for any $\theta \in \Theta$, $\mathcal{M}_\theta$ is compact. (b) (a_i, x_i) for $i = 1, \ldots, M$, are independently and identically distributed, and $\Pr(x_i = x) > 0$ for any $x \in X$. (c) There is a unique $\theta^0 \in int(\Theta)$ and a unique $P_{\theta^0} \in \mathcal{M}_{\theta^0}$ such that, for any $(a,x) \in A \times X$, $P_{\theta^0}(a|x) = P^0(a|x)$. (d) For any $P_\theta \in \mathcal{M}_\theta$ given any $\theta \neq \theta^0$, $\Pr_{P^0}(\{(a,x) : P_\theta(a|x) \neq P^0(a|x)\}) > 0$. (e) $E \sup_{\theta \in \Theta} |P_\theta(a|x)| < \infty$.*

Assumption 6 *(a) $\Psi(P,\theta)(a|x) > 0$ for any $(a,x) \in A \times X$ and any $\{P, \theta\} \in B_P \times \Theta$. (b) $\Psi(P,\theta)$ is continuously differentiable in $(P,\theta) \in \mathcal{N}$, and $\sup_{(P,\theta) \in \mathcal{N}} ||\nabla_{P'} \Psi(P,\theta)|| < \infty$ and $\sup_{(P,\theta) \in \mathcal{N}} ||\nabla_{\theta'} \Psi(P,\theta)|| < \infty$.*

The consistency of AFXP estimator requires the following additional assumptions:

Assumption 7 *(a) For any $\eta \in \mathbb{R}^K$ such that $\eta \neq 0$, $\nabla_{\theta'} P_{\theta^0}(a|x)\eta \neq 0$ with positive probability. (b) $E \sup_{\theta \in \Theta, (P^*,\theta^*) \in \mathcal{N}} ||\nabla_{\theta^{*\prime}} \Phi(\theta, P^*, \theta^*)(a|x)|| < \infty$, and $E \sup_{\theta \in \Theta, (P^*,\theta^*) \in \mathcal{N}} ||\nabla_{P^{*\prime}} \Phi(\theta, P^*, \theta^*)(a|x)|| < \infty$. (c) $E||\nabla_{\theta'} P_{\theta^0}(a|x)|| < \infty$.*

Assumption 7(a) is similar to Assumption 3 and is an identification condition for the probability limit of our objective function. Assumption 7(b)-(c) are regularity conditions required for the uniform convergence of the objective function. Assumption 7(b) is stated in terms of the conditions on the derivatives of Φ to simplify the presentation, but it is possible to state it in terms of the conditions on the derivatives of $\Psi(P, \theta)$.

Under these assumptions, the sequential estimators generated by the AFXP algorithm is consistent:

Proposition 8 *Suppose that Assumptions 5-7 hold and $\tilde{\theta}_0$ is consistent. Suppose we obtain $\tilde{\theta}_k$ by the AFXP algorithm. Then $\tilde{\theta}_k - \theta^0 = o_p(1)$ for $k = 1, 2, \ldots$*

If a sequence of estimators generated by the AFXP algorithm converges, it converges to the MLE. We now analyze the convergence property of the AFXP algorithm. We introduce the following additional regularity conditions. Let $\nabla^{(3)}\Phi(\theta, P^*, \theta^*)$ denote the third derivatives of $\Phi(\theta, P^*, \theta^*)$ with respect to (θ, P^*, θ^*). Assumption 8(a) is required for the asymptotic normality of the NFXP estimator.

Assumption 8 *(a) $E \sup_{\theta \in \mathcal{N}_{\theta^0}} ||\nabla_{\theta'} P_\theta(a|x)||^2 < \infty$, and $E \sup_{\theta \in \mathcal{N}_{\theta^0}} ||\nabla_{\theta\theta'} P_\theta(a|x)|| < \infty$. (b) $\Psi(P, \theta)$ is twice continuously differentiable in $(P, \theta) \in \mathcal{N}$ with a bounded second derivative. (c) $E \sup_{\theta \in \mathcal{N}_{\theta^0}, (P^*,\theta^*) \in \mathcal{N}} ||\nabla^{(3)} \Phi(\theta, P^*, \theta^*)(a|x)|| < \infty$.*

The following proposition establishes the convergence rate of the AFXP algorithm. Let θ_{MLE} be the MLE of θ and define $\hat{P}_{MLE} = P_{\hat{\theta}_{MLE}}$, the MLE of P.

Proposition 9 *Suppose that Assumptions 5-8 hold and $\tilde{\theta}_0$ is consistent. Suppose we obtain $\{\tilde{P}_j, \tilde{\theta}_j\}_{j=1}^k$ by the AFXP algorithm. Then, for $j = 1, 2, \ldots, k$,*

$$\begin{aligned} \tilde{\theta}_j - \hat{\theta}_{MLE} &= O_p(||\tilde{P}_j - \hat{P}_{MLE}||), \\ \tilde{P}_j - \hat{P}_{MLE} &= O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{MLE}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{MLE}||^2). \end{aligned}$$

Thus, the estimator generated by the AFXP algorithm is first-order equivalent to the MLE for all $k \geq 1$. This convergence rate is also the same as that of the NPL algorithm for a single-agent model with $\nabla_{P'} \Psi(P^0, \theta^0) = 0$ (Kasahara and Shimotsu, 2008a). This algorithm can be used to obtain the MLE because, upon convergence, its limit is identical to the MLE.

Implementing Step 1 of the AFXP algorithm may be impractical when finding all the fixed points is computationally infeasible. In such cases, we may replace the solution to the fixed point in Step 1 with its consistent estimator as follows. Let $(\tilde{P}_0, \tilde{\theta}_0)$ be an initial estimator of (P^0, θ^0). For $j \geq 1$, consider the *q-AFXP* algorithm which iterates

Step 1: Given $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$, update P by $\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$, and

Step 2: Given $(\tilde{P}_j, \tilde{\theta}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta \in \Theta_j} Q_n(\theta, \tilde{P}_j, \tilde{\theta}_{j-1})$, where Θ_j is given by (16),

until $j = k$. We may define the q-AFXP algorithm using Γ^q in place of Λ^q and the same argument applies to Γ^q. The sequential estimators generated by the q-AFXP algorithm is consistent.

Proposition 10 *Suppose that Assumptions 5-7 hold and $(\tilde{P}_0, \tilde{\theta}_0)$ is consistent. Suppose we obtain $\tilde{\theta}_k$ by the q-AFXP algorithm. Then $\tilde{\theta}_k - \theta^0 = o_p(1)$ for $k = 1, 2, \ldots$*

We now derive the convergence property of the q-AFXP algorithm. First, we introduce some notations. Define the information matrix for the MLE as $\mathcal{I}^0 = E[\nabla_\theta \ln P_{\theta^0}(a_i|x_i) \nabla_{\theta'} \ln P_{\theta^0}(a_i|x_i)]$. Under the standard regularity conditions, the MLE satisfies $\sqrt{n}(\hat{\theta}_{MLE} - \theta^0) \to_d N(0, (\mathcal{I}^0)^{-1})$. Define a $K \times L$ matrix $\mathcal{J}$ as (we state it in terms of $\mathcal{J}'$ for notational convenience)

$$\mathcal{J}' = E\left[\nabla_P \left\{ \frac{[(I - \nabla_{P'}\Psi(P^0, \theta^0))^{-1} \nabla_{\theta'} \Psi(P^0, \theta^0)](a|x)}{P^0(a|x)} \right\}\right].$$

The following proposition establishes the convergence rate of the q-AFXP algorithm.

Proposition 11 *Suppose that Assumptions 5-8 hold and $(\tilde{P}_0, \tilde{\theta}_0)$ is consistent. Suppose we obtain $\tilde{\theta}_k$ by the q-AFXP algorithm. Then, for $k = 1, 2, \ldots$,*

$$\begin{aligned} \tilde{P}_j - \hat{P}_{MLE} &= \Lambda^q_P(\tilde{P}_{j-1} - \hat{P}_{MLE}) + \Lambda^q_\theta(\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) + R_{n,j}, \\ (\mathcal{I}^0 + o_p(1))(\tilde{\theta}_j - \hat{\theta}_{MLE}) &= -\mathcal{J}\nabla_{\theta'} P_{\theta^0}(\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) + \mathcal{J}(\tilde{P}_j - \hat{P}_{MLE}) + R_{n,j}, \end{aligned}$$

where $R_{n,j}$ denotes a generic reminder term satisfying $R_{n,j} = O_p(||\tilde{P}_{j-1} - \hat{P}_{MLE}||^2) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{MLE}||) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||)$.

Ignoring $R_{n,j}$ and $o_p(1)$ term and arranging the two updating relations into a system of equations, we obtain

$$\begin{pmatrix} I_L & 0 \\ -\mathcal{J} & \mathcal{I}^0 \end{pmatrix} \begin{pmatrix} \tilde{P}_j - \hat{P}_{MLE} \\ \tilde{\theta}_j - \hat{\theta}_{MLE} \end{pmatrix} = \begin{pmatrix} \Lambda^q_P & \Lambda^q_\theta \\ 0 & -\mathcal{J}\nabla_{\theta'} P_{\theta^0} \end{pmatrix} \begin{pmatrix} \tilde{P}_{j-1} - \hat{P}_{MLE} \\ \tilde{\theta}_{j-1} - \hat{\theta}_{MLE} \end{pmatrix}.$$

It follows that

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

Therefore, the convergence property of the q-AFXP algorithm, or the eigenvalues of Q, depends on three factors: (i) the magnitude of Λ^q_P and Λ^q_θ, (ii) the magnitude of $\mathcal{I}^0$ and $\mathcal{J}$, and

(iii) difference between Λ_θ^q and $\nabla_{\theta'} P_{\theta^0}$. From the properties of Λ, we obtain $\Lambda_P^q = (\Lambda_P)^q$ and $\Lambda_\theta^q - \nabla_{\theta'} P_{\theta^0} = -\nabla_{\theta'} P_{\theta^0} (\Lambda_P)^{q-1}$. Since the trace of a matrix is the sum of its eigenvalues, the sum of the eigenvalues of Q is the sum of the trace of $(\Lambda_P)^q$ and the trace of $-(\mathcal{I}^0)^{-1} \mathcal{J} \nabla_{\theta'} P_{\theta^0} (\Lambda_P)^{q-1}$. Therefore, when all the eigenvalues of Λ_P are smaller than 1 in absolute value, then, for sufficiently large q, all the eigenvalues of Q are smaller than 1, and iterating the q-AFXP algorithm converges to the MLE.

5 Sequential GMM estimators

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

Given the conditional probabilities P^0 in population, for any function $h : A \to \mathbb{R}$, the following conditional moment condition always holds: $E\left[h(a) - \sum_{a' \in A} h(a') P^0(a'|x) \mid x\right] = 0$. For example, we may choose $h(a) = a$ or $h(a) = a^2$. The conditional moment condition imply unconditional moment conditions of the form $E\left[g_l(x, a; P^0)\right] = 0$, where

$$g_l(x, a; P^0) = \rho_l(x) \left(h_l(a) - \sum_{a' \in A} h_l(a') P^0(a'|x) \right) \tag{17}$$

for any function $\rho_l : X \to \mathbb{R}$ and $h_l : A \to \mathbb{R}$.

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

The generalized method of moments (GMM) estimator with fixed point constraint is defined as: $\hat{\theta}_{GMM} = \arg\min_{\theta \in \Theta} \left\{ \min_{P \in \mathcal{M}_\theta} \ \bar{g}(P)' \hat{W} \bar{g}(P) \right\}$, where $\mathcal{M}_\theta$ is given in (3), $\hat{W} \to_p W$ positive semi-definite, and $\bar{g}(P) = n^{-1} \sum_{i=1}^n g(a_i, x_i; P)$. Here, $g(\cdot; P) = (g_1(\cdot; P), g_2(\cdot; P), ..., g_L(\cdot; P))'$ is a moment vector function representing L moment conditions with $g_l(\cdot)$ function given by (17) for $l = 1, .., L$.

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

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

In the following, we use the following notations.

$$\bar{G}_\theta(\Psi(P,\theta)) = (\partial/\partial\theta')\bar{g}(\Psi(P,\theta)), \qquad \bar{G}_P(\Psi(P,\theta)) = (\partial/\partial P')\bar{g}(\Psi(P,\theta)),$$
$$G_\theta = E[(\partial/\partial\theta')g(a_i,x_i;\Psi(P^0,\theta^0))], \qquad G_P = E[(\partial/\partial P')g(a_i,x_i;\Psi(P^0,\theta^0))].$$

Define f_x as before so that its elements are arranged conformably with $P^0(j|x^l)$ while let $\hat{f}_x$ be a frequency estimator of f_x. Denote $\Delta_x = diag(f_x)$ and $\hat{\Delta}_x = diag(\hat{f}_x)$. Let $\gamma_l(a,x) = \rho_l(x)h_l(a)$ and γ_l represent a vector of $|A||X|$ length. Finally, let $H = (\gamma_1', \gamma_2', ..., \gamma_L')'$ be a L by $|A||X|$ matrix. With those notations, we may write $\bar{G}_\theta(\Psi(P,\theta)) = -H\hat{\Delta}_x(\partial/\partial\theta')\Psi(P,\theta)$, $\bar{G}_P(\Psi(P,\theta)) = -H\hat{\Delta}_x(\partial/\partial P')\Psi(P,\theta)$, $G_\theta = -H\Delta_x\Psi_\theta$ and $G_P = -H\Delta_x\Psi_P$. Let $r(a_i,x_i)$ be a vector of length $|A||X|$ whose elements are arranged conformably with $P^0(a|x)$ and be equal to zero except for the element of $(a,x) = (a_i,x_i)$ which takes a value of one. With this notation, we can write $\hat{P}_0 = n^{-1}\sum_{i=1}^n r(a_i,x_i)$.

Assumption 9 *(a) For any $\theta \neq \theta^0$, $WE[g(a,x;\Psi(P^0,\theta))] \neq 0$; (b) $G_\theta'WG_\theta$ is nonsingular; (c) $E[||g(a,x;P^0)||^2] < \infty$; (d) $E[\sup_{\theta\in\Theta}||g(a,x;\Psi(P^0,\theta))||] < \infty$; (e) $E[\sup_{\theta\in\Theta}||\nabla_{\theta'}g(a,x;\Psi(P^0,\theta))||] < \infty$.*

Under Assumptions 1 and 9, $\hat{\theta}_{2GMM}$ is consistent and asymptotic normal: $\sqrt{n}(\hat{\theta}_{2GMM} - \theta^0) \to_d N(0, V_{2GMM})$, where $V_{2GMM} = (G_\theta'WG_\theta)^{-1}G_\theta'WSWG_\theta(G_\theta'WG_\theta)^{-1}$ with $S = E[(g(a_i,x_i;P^0) + G_P(r(a_i,x_i) - P^0))(g(a_i,x_i;P^0) + G_P(r(a_i,x_i) - P^0))']$. Using an optimal weighting matrix $W = S^{-1}$, the limiting variance is given by $V_{2GMM} = (G_\theta'S^{-1}G_\theta)^{-1}$.

We now consider a recursive extension of the two-step GMM estimator called the *nested GMM algorithm* which iterates

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

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

until $j = k$. Under regularity conditions similar to the ones in Assumption 1, the sequence of estimators generated by this algorithm are consistent. If the sequence converges, the limit is called the nested GMM (NGMM) estimator. The NGMM estimator $(\hat{P}_{NGMM}, \hat{\theta}_{NGMM})$ satisfies

$$\begin{aligned}\hat{\theta}_{NGMM} &= \arg\min_{\theta\in\Theta} \quad \bar{g}(\Psi(\hat{P}_{NGMM},\theta))'\hat{W}\bar{g}(\Psi(\hat{P}_{NGMM},\theta)),\\ \hat{P}_{NGMM} &= \Psi(\hat{P}_{NGMM},\hat{\theta}_{NGMM}).\end{aligned}$$

Under the following additional assumptions, we derive the limiting distribution of the NGMM estimator.

Assumption 10

$$\bar{g}(P^0) = O_p(n^{-1/2}), \quad \sup_{\theta,P}||D^2\Psi(P,\theta)|| < \infty, \quad ||H|| < \infty, \quad rank((\partial/\partial\theta')\Psi(P,\theta)) = k \text{ for all } P$$

Note that $\sup_{\theta,P}||D^2\Psi(P,\theta)|| < \infty$ and $||H|| < \infty$ imply that $\sup_{\theta,P}||D\bar{G}_\theta(\Psi(P,\theta))|| < \infty$. The rank condition on $(\partial/\partial\theta')\Psi(P,\theta)$ guarantees that $(\bar{G}_\theta(P))'\hat{W}\bar{G}_\theta(P)$ is invertible.

Proposition 12 *Suppose Assumptions 1, 9 and 10 hold. Then*

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

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

Remark 6 *When* $\Psi_P = 0$, *the two-step GMM estimator with an optimal weighting matrix is asymptotically equivalent to the NGMM estimator with* $W = \Omega^{-1}$.

The NGMM estimator can be obtained as the limit of the sequence of estimators generated by the NGMM algorithm upon convergence. The convergence property of the NGMM estimator is given by the following proposition.

Proposition 13 *Suppose Assumptions 1 and 10 hold. Then, for* $j = 1, \ldots, k$,

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

Remark 7 *Observe that* $-\Psi_\theta(G_\theta'\hat{W}G_\theta)^{-1}G_\theta'\hat{W}H\Delta_x = \Psi_\theta(\Psi_\theta'\Delta_x'H'\hat{W}H\Delta_x\Psi_\theta)^{-1}\Psi_\theta'\Delta_x'H'\hat{W}H\Delta_x$ *is a projection matrix, and the sequence of estimators here also has the convergence property similar to the estimators generated by the NPL algorithm. Again, the convergence rate is primarily determined by the eigenvalues of* Ψ_P.

Remark 8 *Analogous remarks to Remarks 1-5 apply here.*

6 Unobserved Heterogeneity

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

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

Consider a panel data set $\{\{a_{it}, x_{it}, x_{i,t+1}\}_{t=1}^T\}_{i=1}^n$ such that $w_i = \{a_{it}, x_{it}, x_{i,t+1}\}_{t=1}^T$ is randomly drawn across i's from the population. The conditional probability distribution of a_{it} given x_{it} for type m agent is given by a fixed point of $P_{\theta^m} = \Psi(P_{\theta^m}, \theta^m)$. On the other hand, to simplify our analysis, we assume that the transition probability function of x_{it} is independent of types and given by $f_x(x_{i,t+1}|a_{it}, x_{it})$ and is known to researchers.[5]

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

Stack P^m's as $\mathbf{P} = (P^{1\prime}, \ldots, P^{M\prime})'$, and let $\mathbf{P}^0$ denote its true value. Define $\mathbf{\Psi}(\mathbf{P}, \theta) = (\Psi(P^1, \theta^1)', \ldots, \Psi(P^M, \theta^M)')'$. Then, the set of possible probabilities consistent with the fixed point constraints given the value of θ is given by $\mathcal{M}_\theta^* = \{\mathbf{P} \in B_P^M : \mathbf{P} = \mathbf{\Psi}(\mathbf{P}, \theta)\}$.

The maximum likelihood estimator for a model with unobserved heterogeneity is:

$$\hat{\zeta}_{MLE} = \arg\max_{\zeta \in \Theta_\zeta} \left\{ \max_{\mathbf{P} \in \mathcal{M}_\theta^*} \ln \left([L(\mathbf{P}, \pi)](w_i) \right) \right\}, \tag{18}$$

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

Assumption 11 *(a) Θ_ζ is compact and, for any $\theta \in \Theta^M$, $\mathcal{M}_\theta^*$ is compact. (b) $[L(\mathbf{P}, \pi)](w) > 0$ for any w and for any $(\mathbf{P}, \pi) \in \cup_{\theta \in \Theta^M} \mathcal{M}_\theta^* \times \Theta_\pi$. (c) $w_i = \{(a_{it}, x_{it}, x_{i,t+1}) : t = 1, \ldots, T\}$ for $i = 1, \ldots, n$, are independently and identically distributed. (d) For any $P \in B_P$, there exists a unique fixed point for $T(\cdot, P)$. (e) There is a unique $\zeta^0 \in int(\Theta_\zeta)$ and a unique $\mathbf{P}^0 \in \mathcal{M}_{\theta^0}^*$ such that, for any $w = \{(a_t, x_t, x_{t+1}) : t = 1, \ldots, T\}$, $[L(\mathbf{P}^0, \pi^0)](w) = L^0(w)$, where L^0 is the true probability for w. Given any $(\theta, \pi) \neq (\theta^0, \pi^0)$, for any $\mathbf{P} \in \mathcal{M}_\theta^*$, $\Pr_{L^0}(\{w : [L(P, \pi)](w) \neq L^0(w)\}) > 0$. (f) $\tilde{\mathbf{P}}_0 - \mathbf{P}^0 = o_p(1)$, and the MLE denoted by $\hat{\zeta}_{MLE}$ satisfies $\sqrt{n}(\hat{\zeta}_{MLE} - \zeta^0) \to_d N(0, \Omega_\zeta)$.*

Assumption 11(f) requires an initial consistent estimators for the type-specific conditional probabilities. Kasahara and Shimotsu (2006, 2008b) 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 for $\mathbf{P}$. On the other hand, as AM07 argue, if the NPL algorithm

[5] 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 (2008a) analyze the case in which the transition probability function is also type-dependent in the context of a single agent dynamic programming model with unobserved heterogeneity.

converges, then the limit may provide a consistent estimate for the parameter ζ even when $\tilde{\mathbf{P}}_0$ is not consistent.

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

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

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

until $j = k$. If the iterations converge, its limit $(\hat{\mathbf{P}}_{NPL}, \hat{\zeta}_{NPL})$ is the NPL estimator for models with unobserved heterogeneity and satisfies the conditions analogous to (7).

We now establish the convergence property of the NPL algorithm for models with unobserved heterogeneity. Let $[l(\mathbf{P}, \zeta)](w) \equiv \ln \Big([L(\mathbf{\Psi}(\tilde{\mathbf{P}}_{j-1}, \theta), \pi)](w) \Big)$. Then, $\Omega_{\zeta\zeta} = E[(\partial/\partial\zeta)[l(\mathbf{P}^0, \zeta^0)](w) \; (\partial/\partial\zeta')[l(\mathbf{P}^0, \zeta^0)](w)]$ and $\Omega_{\zeta P} = E\left[(\partial/\partial\zeta)[l(\mathbf{P}^0, \zeta^0)](w)(\partial/\partial\mathbf{P}')[l(\mathbf{P}^0, \zeta^0)](w)\right]$ are written as

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

where $\mathbf{\Psi}_P \equiv (\partial/\partial\mathbf{P}')\mathbf{\Psi}(\mathbf{P}^0, \theta^0)$, $\mathbf{\Psi}_\theta \equiv (\partial/\partial\theta')\mathbf{\Psi}(\mathbf{P}^0, \theta^0)$, $\Delta_L = diag((L^0)^{-1}) = diag((L(\mathbf{P}^0, \pi^0))^{-1})$, $L_P = \nabla_{P'} L(\mathbf{P}^0, \pi^0)$, and $L_\pi = \nabla_{\pi'} L(\mathbf{P}^0, \pi^0)$.

Assumption 12

$$\bar{l}_\zeta(\mathbf{P}^0, \zeta^0) = O_p(n^{-1/2}), \qquad \bar{l}_{\zeta\zeta}(\mathbf{P}^0, \zeta^0) = -\Omega_{\zeta\zeta} + O_p(n^{-1/2}), \quad \bar{l}_{\zeta P}(\mathbf{P}^0, \zeta^0) = -\Omega_{\zeta P} + O_p(n^{-1/2}),$$
$$E \sup_{\zeta, \mathbf{P}} ||D_{\zeta\mathbf{P}}[l(\mathbf{P}, \zeta)](w)|| < \infty, \qquad E \sup_{\zeta, \mathbf{P}} ||D^3[l(\mathbf{P}, \zeta)](w)|| < \infty,$$
$$\sup_{\theta, P} ||D^2\Psi(P, \theta)|| = O(1), \qquad \bar{l}_{\zeta\zeta}(P, \theta) \text{ is invertible for all } (P, \theta).$$

where $\bar{l}_\zeta(\mathbf{P}, \zeta) = n^{-1} \sum_{i=1}^n (\partial/\partial\zeta)[l(\mathbf{P}, \zeta)](w_i)$, $\bar{l}_{\zeta\zeta}(\mathbf{P}, \zeta) = n^{-1} \sum_{i=1}^n (\partial^2/\partial\zeta\partial\zeta')[l(\mathbf{P}, \zeta)](w_i)$, *and* $\bar{l}_{\zeta\mathbf{P}}(\mathbf{P}, \zeta) = n^{-1} \sum_{i=1}^n (\partial^2/\partial\zeta\partial\mathbf{P}')[l(\mathbf{P}, \zeta)](w_i)$.

The following result states the convergence properties of the NPL algorithm for models with unobserved heterogeneity.

Lemma 2 *Suppose Assumptions 11-12 hold. Then, for* $j = 1, \ldots, k$,

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

where $D = (\mathbf{\Psi}_\theta' L_P' \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P \mathbf{\Psi}_\theta)^{-1}$ *and* $M_{L_\pi} = I - \Delta_L^{1/2} L_\pi (L_\pi' \Delta_L L_\pi)^{-1} L_\pi \Delta_L^{1/2}$.

Since $I - \mathbf{\Psi}_\theta D \mathbf{\Psi}'_\theta L'_P \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P$ is an idempotent matrix, its eigenvalues are either zero or one. Consequently, the convergence rate of the NPL algorithm for models with unobserved heterogeneity is primarily determined by the dominant eigenvalue of $\mathbf{\Psi}_P$. When the NPL algorithm encounters a convergence problem, replacing $\Psi(P,\theta)$ with $\Lambda(P,\theta)$ or $\Gamma(P,\theta)$ improves convergence property.

Remark 9 *It is possible to relax the stationarity assumption on the initial states by estimating the type-specific initial distributions of x, denoted by $\{p^{*m}\}_{m=1}^M$, without imposing stationarity restriction in Step 1 of the NPL algorithm. In this case, the NPL algorithm has the convergence rates similar to those of Lemma 2.*

7 Monte Carlo experiments

We consider the model of Example 2. The profit function of firm i operating in market m in period t is specified as $\tilde{\Pi}_i(a_{mt}, S_{mt}, a_{m,t-1}, \epsilon_{imt}; \theta) = \Pi_i(a_{imt}, a_{-i,mt}, S_{mt}, a_{m,t-1}; \theta) + \epsilon_{imt}$ with

$$\begin{aligned}&\Pi_i(a_{imt}=1, a_{-i,mt}, S_{mt}, a_{m,t-1}; \theta) + \epsilon_{imt}(1) = \\ &\quad \theta_{RS} \ln S_{mt} - \theta_{RN} \ln(1 + \sum_{j \neq i} a_{jmt}) - \theta_{FC,i} - \theta_{EC}(1 - a_{im,t-1}) + \epsilon_{imt}(1)\end{aligned}$$

while, if the firm is not operating in market m, its profit is $\Pi_i(a_{imt}=0, a_{-i,mt}, S_{mt}, a_{m,t-1}; \theta) + \epsilon_{imt}(0) = \epsilon_{imt}(0)$. We assume that $\{\epsilon_{imt}\}$ are i.i.d. extreme value type I with zero mean and unit variance and S_{mt} follows an exogenous first-order Markov process $f_S(S_{m,t+1}|S_{mt})$. The explicit expression for the fixed mapping Ψ in this model is provided in the Appendix B.

We set the number of firms $N = 3$. The state space for the market size S_{mt} is $\{2, 6, 10\}$.[6] The discount factor is set to $\beta = 0.96$. Fixed operating costs are $\theta_{FC,1} = 1.0$, $\theta_{FC,2} = 0.9$, and $\theta_{FC,3} = 0.8$ while we set both θ_{RS} and θ_{EC} equal to 1.

The value of parameter θ_{RN} determines the degree of strategic substitutability among firms and is the main determinant of the dominant eigenvalues of Ψ_P. We therefore vary the values of θ_{RN} across experiments and examine the performance of different estimators across different parameter values of θ_{RN}. In particular, we consider $\theta_{RN} = 1, 2, 4$, and 6. All the eigenvalues of Ψ_P are less than 1 in absolute value for $\theta_{RN} = 1$ and 2 while the smallest eigenvalues are less than -1 for $\theta_{RN} = 4$ and 6. The fourth and the fifth column of Table 1 respectively show the largest and the smallest eigenvalues of Ψ_P. We estimate θ_{RS} and θ_{RN} while the other parameters are not estimated but fixed at the true values.

[6] The transition probability matrix of S_{mt} is given by

$$\begin{bmatrix} 0.8 & 0.2 & 0.0 \\ 0.2 & 0.6 & 0.2 \\ 0.0 & 0.2 & 0.8 \end{bmatrix}.$$

Given the equilibrium choice probabilities obtained as the fixed point of Ψ and the transition probabilities for market size S, we obtain the steady state distribution. To generate an observation in market m, we first randomly draw $x_m = \{S_{m1}, a_{1m0}, a_{2m0}, a_{3m0}\}$ from the steady-state distribution and then, conditioning on the realized value of x_m, the choice a_{im1} for $i = 1, 2, 3$ is randomly drawn from the equilibrium choice probabilities. Repeating the procedure for $m = 1, 2, ..., n$, we obtain a data set with a sample size n: $\{S_{m1}, a_{im0}, a_{im1} : i = 1, 2, 3; m = 1, 2, ..., n\}$. In our experiment, we produce 500 simulated samples, each of which contains $n = 500$, 2000, and 8000 observations.

To generate the data for each experiment, we need to compute a fixed point of $\Psi(P, \theta)$. For $\theta_{RN} = 1$ and 2, the fixed point is obtained by iterating the mapping $\Psi(P, \theta)$ starting from an initial vector of choice probabilities, P_0, with all probabilities equal to 0.5. For $\theta_{RN} = 4$ and 6, the smallest eigenvalues of Ψ_P evaluated at the fixed point are smaller than negative one at -1.18 and -1.48, respectively, and hence the sequence $\{P_k\}$ obtained by iterating the mapping $\Psi(P, \theta)$ does not converge. To obtain a fixed point of $\Psi(P, \theta)$ for $\theta_{RN} = 4$ and 6, we consider an alternative mapping $[\Lambda(P, \theta)](a = 1|x) \equiv \{[\Psi(P, \theta)](a = 1|x)\}^{\alpha^*}\{P(a = 1|x)\}^{1-\alpha^*}$ with the optimal value of α^*, which has better convergence property than Ψ.

To estimate the value of α^* for the mapping Λ, we simulate a sequence $P^j = \Lambda(P^{j-1}, \theta)$ for $j = 1, ..., k$ across different values of α, say $\alpha = 0.01, 0.02, ..., 0.99, 1.00$, where k is the number of iterations at convergence. Then, the value of α that leads to the smallest value of the mean of $||P^{j+1} - P^k||/||P^j - P^k||$'s is our estimate for α^*. The estimates of α^* using this procedure are reported in the second column of Table 1 and they are very close to the true value of α^* reported in the third column. As reported in the last two columns of Table 1, the absolute value of the dominant eigenvalue of $M_{\Psi_\theta}\Psi_P$ and $M_{\Lambda_\theta}\Lambda_P$ are similar to the corresponding eigenvalues of Ψ_P and Λ_P that are reported in the fourth to seventh columns. Thus, in view of Lemma 1, the convergence rate of the NPL algorithm is primarily determined by the dominant eigenvalue of Ψ_P or Λ_P.

Table 2 compares the bias and the mean squared errors (MSE) across different estimators for $\theta_{RN} = 2$ or 4 with the sample size of $n = 500$, 2000, and 8000. The maximum number of iterations for sequential estimators is set to $k = 50$. For $\theta_{RN} = 2$, the NPL estimator with Ψ substantially improves the performance of the two-step PML estimator and it converges to the same estimate as the NPL estimator with Λ.

For $\theta_{RN} = 4$, however, reflecting its non-convergence property, the estimator generated by 50 iterations of the NPL algorithm with Ψ performs substantially worse than the NPL estimator with Λ. In particular, with the sample size of $n = 500$, the square root of the MSE for the estimates of $\hat{P}$ generated by the NPL algorithm with Ψ is more than thirty times larger than those with Λ. Furthermore, as the sample size increases from $n = 500$ to $n = 2000$, and then to $n = 8000$, the MSE for the NPL estimator with Λ decreases to one-half, and then one-fourth, approximately at the rate of square root-n, but the MSE for the NPL estimator with Ψ decrease

at much slower rate than square root-n. With the sample size of $n = 8000$, the MSE of the NPL estimator with Ψ is even larger than that of the two-step PML estimator when $\theta_{RN} = 4$.

The fourth and the fifth rows of each panel of Table 2 report the performance of the estimator generated by the RPM for $\delta = 0.5$ and 0.8, respectively, where the orthonormal basis Z is updated fully every $J = 10$ iterations. Both estimators perform better than the NPL estimator with Ψ, especially when $\theta_{RN} = 4$, and their performance is comparable to that of the NPL estimator with Λ. We also note that the NPL estimator based on the RPM with $\delta = 0.5$ performs better than that with $\delta = 0.8$ because the former achieves faster contraction.

The sixth and the seventh rows of each panel of Table 2 report the performance of the q-NPL estimator and the q-AFXP estimator with Λ^q, respectively, where q is set to 4. The estimator generated by the q-NPL algorithm as well as the q-AFXP algorithm perform better than the estimators generated by the NPL algorithm with Ψ or Λ for both $\theta_{RN} = 2$ and $\theta_{RN} = 4$, suggesting their efficiency gains over the NPL estimator.

8 Conclusion

In this paper, we analyze the convergence property of the recursive method pioneered by Aguirregabiria and Mira (2007) to estimate a class of structural models characterized by the fixed point constraint, such as a dynamic game model. We show that, when the fixed point mapping has a local contraction property, then the convergence of the NPL algorithm is achieved in the neighborhood of the true value. Even in the presence of (a finite number of) multiple equilibria, the initial consistent estimator $\tilde{P}_0$ is asymptotically in the neighborhood of the true equilibrium probabilities when the data is generated from a single equilibrium; therefore, the local contraction property of the NPL algorithm guarantees the consistency of the NPL estimator.

In practice, the violation of the convergence condition is a concern. We develop alternative sequential estimators that can be used even when the original fixed point mapping is not locally contractive. As our Monte Carlo experiments show, these alternative sequential estimators work well even when the NPL algorithm has a convergence problem and their performance can be substantially better than that of the two-step estimator.

9 Appendix A: Proofs

In the following, C denotes a generic positive and finite constant, and it may take different values in different places.

9.1 Proof of Proposition 1

Assumption 1 (a), (b), and (d) with $\hat{P}_0 \to_p P^0$ imply that $\overline{\psi}(\hat{P}_0, \theta)$ converges uniformly in probability in θ to $E(\ln \Psi(P^0, \theta))$ (cf., Lemma 24.1 of Gourieroux and Monfort, 1995). Then,

the rest of the proof follows the proof of Theorem 2.1 of Newey and McFadden (1994). □

9.2 Proof of Propositions 2 and 3

See pp.49-52 of Aguirregabiria and Mira (2007). □

9.3 Proof of Lemma 1

Define $\overline{\psi}_\theta(P,\theta) \equiv n^{-1}\sum_{i=1}^n (\partial/\partial\theta)\ln\Psi(P,\theta)(a_i|x_i)$, $\overline{\psi}_{\theta P}(P,\theta) \equiv n^{-1}\sum_{i=1}^n (\partial^2/\partial\theta\partial P')\ln\Psi(P,\theta)(a_i|x_i)$, and $\overline{\psi}_{\theta\theta}(P,\theta) \equiv n^{-1}\sum_{i=1}^n (\partial^2/\partial\theta\partial\theta')\ln\Psi(P,\theta)(a_i|x_i)$.

Recall that $\tilde{\theta}_j$ satisfies the first order condition

$$\overline{\psi}_\theta(\tilde{P}_{j-1},\tilde{\theta}_j) = 0. \tag{19}$$

Expanding this around $(\hat{P}_{NPL},\hat{\theta}_{NPL})$ and using $\overline{\psi}_\theta(\hat{P}_{NPL},\hat{\theta}_{NPL}) = 0$ gives $0 = \overline{\psi}_{\theta P}(\bar{P},\bar{\theta})(\tilde{P}_{j-1} - \hat{P}_{NPL}) + \overline{\psi}_{\theta\theta}(\bar{P},\bar{\theta})(\tilde{\theta}_j - \hat{\theta}_{NPL})$, where $(\bar{P},\bar{\theta})$ lie between $(\tilde{P}_{j-1},\tilde{\theta}_j)$ and $(\hat{P}_{NPL},\hat{\theta}_{NPL})$. Inverting $\overline{\psi}_{\theta\theta}(\bar{P},\bar{\theta})$, we obtain

$$\tilde{\theta}_j - \tilde{\theta} = -\overline{\psi}_{\theta\theta}(\bar{P},\bar{\theta})^{-1}\overline{\psi}_{\theta P}(\bar{P},\bar{\theta})(\tilde{P}_{j-1} - \hat{P}_{NPL}) = O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||), \tag{20}$$

where the last equality follows from the last two assumptions of Assumption 2.

For the second result, expand the second-step updating equation $\tilde{P}_j = \Psi(\tilde{P}_{j-1},\tilde{\theta}_j)$ twice around $(\hat{P}_{NPL},\hat{\theta}_{NPL})$ and use $\Psi(\hat{P}_{NPL},\hat{\theta}_{NPL}) = \hat{P}_{NPL}$, root-$n$ consistency of $(\hat{P}_{NPL},\hat{\theta}_{NPL})$, and (20), then it follows that

$$\tilde{P}_j - \hat{P}_{NPL} = \Psi_P(\tilde{P}_{j-1} - \hat{P}_{NPL}) + \Psi_\theta(\tilde{\theta}_j - \hat{\theta}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2). \tag{21}$$

Rewriting (20) by using $\overline{\psi}_{\theta P}(\hat{P}_{NPL},\hat{\theta}_{NPL}) = -\Omega_{\theta P} + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(n^{-1/2})$ and $\overline{\psi}_{\theta\theta}(\hat{P}_{NPL},\hat{\theta}_{NPL}) = -\Omega_{\theta\theta} + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(n^{-1/2})$, we obtain

$$\tilde{\theta}_j - \tilde{\theta} = -\Omega_{\theta\theta}^{-1}\Omega_{\theta P}(\tilde{P}_{j-1} - \hat{P}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2).$$

Substituting this into (21) in conjunction with $\Omega_{\theta\theta}^{-1}\Omega_{\theta P} = (\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P\Psi_P$ gives

$$\tilde{P}_j - \hat{P}_{NPL} = \left[I - \Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P\right]\Psi_P(\tilde{P}_{j-1} - \hat{P}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2),$$

giving the stated result. □

9.4 Proof of Proposition 4

For any eigenvalue λ of Ψ_P, the corresponding eigenvalue of Λ_P is $\alpha\lambda + (1-\alpha) = \alpha(\lambda-1)+1$. Suppose $\lambda_{\max} > 1 > \lambda_{\min}$. If $\alpha \geq 0$, then $\alpha(\lambda_{max}-1)+1 \geq 1$. If $\alpha < 0$, then $\alpha(\lambda_{min}-1)+1 > 1$.

Therefore, there is no value of α such that $\alpha(\lambda-1)+1<1$ for both $\lambda=\lambda_{max}$ and λ_{min}, giving the first result.

Now, assume that $1>\lambda_{\max}>\lambda_{\min}$. We derive the value of α that minimizes the absolute value of the dominant eigenvalue of Λ_P. Suppose that $\alpha(\lambda_{\min}-1)+1<0$. Then, the absolute value of the dominant eigenvalue of Λ is $\max\{\alpha(\lambda_{\max}-1)+1, -\alpha(\lambda_{\min}-1)-1\}$. If $\alpha(\lambda_{\max}-1)+1 > -\alpha(\lambda_{\min}-1)-1$, then it is possible to reduce the absolute value of the dominant eigenvalue by slightly increasing the value of α, and such a choice of α is not optimal. Similarly, if $\alpha(\lambda_{\max}-1)+1 < -\alpha(\lambda_{\min}-1)-1$, then such a choice of α is not optimal. Therefore, the optimal value of α satisfies $\alpha(\lambda_{\max}-1)+1=-\alpha(\lambda_{\min}-1)-1$ and $\alpha^*=\frac{2}{2-\lambda_{\max}-\lambda_{\min}}$. The largest and smallest eigenvalues of Λ_P with α^* is given by $\frac{\lambda_{\max}-\lambda_{\min}}{2-\lambda_{\max}-\lambda_{\min}}$ and $-\frac{\lambda_{\max}-\lambda_{\min}}{2-\lambda_{\max}-\lambda_{\min}}$, both of which are between -1 and 1. If $\lambda_{\max}+\lambda_{\min}>0$, then $\lambda_{\max}$ is the dominant eigenvalue of Ψ_P and $\lambda_{\max}>\frac{\lambda_{\max}-\lambda_{\min}}{2-\lambda_{\max}-\lambda_{\min}}$ holds. If $\lambda_{\max}+\lambda_{\min}<0$, then $\lambda_{\min}<0$ is the dominant eigenvalue of Ψ_P and $-\frac{\lambda_{\max}-\lambda_{\min}}{2-\lambda_{\max}-\lambda_{\min}}>\lambda_{\min}$ holds. It follows that the absolute value of the dominant eigenvalue of Λ_P with α^* is less than that of Ψ_P.

Suppose that $\alpha(\lambda_{\min}-1)+1\geq 0$. Then, the value of $\alpha(\lambda-1)+1\geq 0$ for any eigenvalue λ of Ψ and $\alpha=0$ must be the optimal choice, but this is not optimal because the value of the dominant eigenvalue of Λ_P with $\alpha=0$ is equal to 1. □

9.5 Proof of Proposition 5

We use induction. Write the objective function as $\bar{\gamma}(\theta,P,\eta)=n^{-1}\sum_{i=1}^n \ln\Gamma(\theta,P,\eta)(a_i|x_i)$. Let $\bar{\gamma}_\theta(\theta,P,\eta)$ denote $\nabla_{\theta'}\bar{\gamma}(\theta,P,\eta)$, and similarly for $\bar{\gamma}_P(\theta,P,\eta)$, $\bar{\gamma}_{\theta\theta}(\theta,P,\eta)$, etc.

First, we show that $\tilde{\theta}_j$ is consistent if $(\tilde{P}_{j-1},\tilde{\theta}_{j-1})$ is consistent. First, $\sup_\theta|\bar{\gamma}(\theta,\tilde{P}_{j-1},\tilde{\theta}_{j-1})-E\ln\Gamma(\theta,P^0,\theta^0)|=o_p(1)$ because $\sup_{(\theta,P,\eta)}|\bar{\gamma}(\theta,P,\eta)-E\ln\Gamma(\theta,P,\eta)|=o_p(1)$ and $(\tilde{P}_{j-1},\tilde{\theta}_{j-1})$ is consistent. Second, θ^0 uniquely maximizes $E\ln\Gamma(\theta,P^0,\theta^0)$ because θ^0 is the only parameter such that $P^0=\Gamma(\theta,P^0,\theta^0)$. Third, $\Gamma(\theta,P,\theta)$ is continuous and $\bar{\Theta}_j$, B_P and Θ are compact by assumption. Therefore, $\tilde{\theta}_j-\theta^0=o_p(1)$ follows from Theorem 2.1 of Newey and McFadden (1994).

We proceed to analyze $\tilde{\theta}_j-\tilde{\theta}$ and $\tilde{P}_j-\tilde{P}$. First, note that the first order condition for $\tilde{\theta}_j$ and $\tilde{\theta}$ implies $\bar{\gamma}_\theta(\tilde{\theta}_j,\tilde{P}_{j-1},\tilde{\theta}_{j-1})=0$ and $\bar{\gamma}_\theta(\tilde{\theta},\tilde{P},\tilde{\theta})=0$. Expanding $\bar{\gamma}_\theta(\tilde{\theta}_j,\tilde{P}_{j-1},\tilde{\theta}_{j-1})$ twice around $(\tilde{\theta},\tilde{P},\tilde{\theta})$ and using $2|ab|\leq a^2+b^2$, we obtain

$$\begin{aligned}0 &= \bar{\gamma}_{\theta P}(\tilde{\theta},\tilde{P},\tilde{\theta})(\tilde{P}_{j-1}-\tilde{P})+\bar{\gamma}_{\theta\theta}(\tilde{\theta},\tilde{P},\tilde{\theta})(\tilde{\theta}_j-\tilde{\theta})+\bar{\gamma}_{\theta\eta}(\tilde{\theta},\tilde{P},\tilde{\theta})(\tilde{\theta}_{j-1}-\tilde{\theta})\\ &\quad +O_p(||\tilde{P}_{j-1}-\tilde{P}||^2)+O_p(||\tilde{\theta}_j-\tilde{\theta}||^2)+O_p(||\tilde{\theta}_{j-1}-\tilde{\theta}||^2). \qquad (22)\end{aligned}$$

For the first term on the right of (22), we have $\bar{\gamma}_{\theta P}(\tilde{\theta},\tilde{P},\tilde{\theta})=\bar{\gamma}_{\theta P}(\theta^0,P^0,\theta^0)+O_p(n^{-1/2})=-\Omega^\Gamma_{\theta P}+O_p(n^{-1/2})$ from the root-n consistency of $(\tilde{P},\tilde{\theta})$, equation (11), and Assumption 2. For the second and third terms on the right of (22), a similar analysis gives $\bar{\gamma}_{\theta\theta}(\tilde{\theta},\tilde{P},\tilde{\theta})=-\Omega^\Gamma_{\theta\theta}+O_p(n^{-1/2})$ and $\bar{\gamma}_{\theta\eta}(\tilde{\theta},\tilde{P},\tilde{\theta})=\bar{\gamma}_{\theta\eta}(\theta^0,P^0,\theta^0)+O_p(n^{-1/2})=O_p(n^{-1/2})$, where the last

equality follows because (10) implies $E\nabla_\theta \ln \Gamma(\theta^0, P^0, \theta^0)(a|x)\nabla_{\eta'} \ln \Gamma(\theta^0, P^0, \theta^0)(a|x) = 0$.

Consequently, in view of the consistency of $\tilde{\theta}_j$, we have

$$\begin{aligned}[\Omega^\Gamma_{\theta\theta} + o_p(1)](\tilde{\theta}_j - \tilde{\theta}) &= -\Omega^\Gamma_{\theta P}(\tilde{P}_{j-1} - \tilde{P}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2) \\ &\quad + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \tilde{\theta}||) + O_p(||\tilde{\theta}_{j-1} - \tilde{\theta}||^2). \qquad (23)\end{aligned}$$

Since $\Omega^\Gamma_{\theta\theta}$ is positive definite, it follows that

$$\tilde{\theta}_j - \tilde{\theta} = O_p(||\tilde{P}_{j-1} - \tilde{P}||) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \tilde{\theta}||) + O_p(||\tilde{\theta}_{j-1} - \tilde{\theta}||^2), \qquad (24)$$

giving the first result of the proposition.

For the bound of $\tilde{P}_j - \tilde{P}$, recall that $\tilde{P}_j = \Gamma(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ and $\tilde{P} = \Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta})$. Expanding $\Gamma(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ twice around $(\tilde{\theta}, \tilde{P}, \tilde{\theta})$ gives

$$\begin{aligned}\tilde{P}_j - \tilde{P} &= \Gamma(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) - \Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta}) \\ &= \nabla_{P'}\Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta})(\tilde{P}_{j-1} - \tilde{P}) + \nabla_{\theta'}\Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta})(\tilde{\theta}_j - \tilde{\theta}) + \nabla_{\eta'}\Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta})(\tilde{\theta}_{j-1} - \tilde{\theta}) \\ &\quad + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2) + O_p(||\tilde{\theta}_{j-1} - \tilde{\theta}||^2),\end{aligned}$$

where the order of the $O_p()$ terms follows from (24). Since $\nabla_{P'}\Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta}) = \nabla_{P'}\Gamma(\theta^0, P^0, \theta^0) + O_P(n^{-1/2}) = \nabla_{P'}\Gamma(P^0, \theta^0) + O_P(n^{-1/2})$, $\nabla_{\theta'}\Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta}) = \nabla_{\theta'}\Gamma(\theta^0, P^0, \theta^0) + O_P(n^{-1/2}) = \nabla_{\theta'}\Gamma(P^0, \theta^0) + O_P(n^{-1/2})$, and $\nabla_{\eta'}\Gamma(\tilde{\theta}, \tilde{P}, \tilde{\theta}) = \nabla_{\eta'}\Gamma(\theta^0, P^0, \theta^0) + O_P(n^{-1/2}) = O_P(n^{-1/2})$, it follows that

$$\tilde{P}_j - \tilde{P} = \nabla_{P'}\Gamma(P^0, \theta^0)(\tilde{P}_{j-1} - \tilde{P}) + \nabla_{\theta'}\Gamma(P^0, \theta^0)(\tilde{\theta}_j - \tilde{\theta}) + R_{nj},$$

where R_n denotes a generic reminder term satisfying $R_{nj} = O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \tilde{\theta}||) + O_p(||\tilde{\theta}_{j-1} - \tilde{\theta}||^2)$. The second result of the proposition follows from substituting (23) into the above and noting $\Omega^\Gamma_{\theta\theta} = \Gamma'_\theta \Delta_P \Gamma_\theta$ and $\Omega^\Gamma_{\theta P} = \Gamma'_\theta \Delta_P \Gamma_P$. □

9.6 Proof of Proposition 6

We show, for $j \geq 1$, $(\tilde{P}_j, \tilde{\theta}_j) \to_p (P^0, \theta^0)$ if $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \to_p (P^0, \theta^0)$. The stated result then follows from induction and $(\tilde{P}_0, \tilde{\theta}_0) \to_p (P^0, \theta^0)$.

Assume $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \to_p (P^0, \theta^0)$. First, $\tilde{P}_j \to_p P^0$ follows from $\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \to_p \Lambda^q(P^0, \theta^0) = P^0$.

We proceed to show $\tilde{\theta}_j \to_p \theta^0$. Define $Q^q_n(\theta, P^*, \theta^*) = n^{-1}\sum_{i=1}^n \ln \tilde{\Lambda}^q(\theta, P^*, \theta^*)(a_i|x_i)$ and $Q^q(\theta) = E \ln \tilde{\Lambda}^q(\theta, P^0, \theta^0)(a_i|x_i)$. From Theorem 2.1 of Newey and McFadden (1994) and the

compactness of Θ_j^q, the consistency of $\tilde{\theta}_j$ follows if we show

$$Q_n^q(\theta, \tilde{P}_j, \tilde{\theta}_{j-1}) - Q_n^q(\theta, P^0, \theta^0) = o_p(1) \quad \text{uniformly in } \theta \in \Theta_j^q, \tag{25}$$

$$Q_n^q(\theta, P^0, \theta^0) - Q^q(\theta) = o_p(1) \quad \text{uniformly in } \theta \in \Theta_j^q, \tag{26}$$

$$Q^q(\theta) \text{ is continuous in } \theta \text{ and uniquely maximized at } \theta^0. \tag{27}$$

We show (25) first. It follows from the mean value theorem that

$$Q_n^q(\theta, \tilde{P}_j, \tilde{\theta}_{j-1}) - Q_n^q(\theta, P^0, \theta^0) = D_{P,n}^q(\theta, \bar{P}, \bar{\theta})(\tilde{P}_j - P^0) + D_{\theta,n}^q(\theta, \bar{P}, \bar{\theta})(\tilde{\theta}_{j-1} - \theta^0), \tag{28}$$

where $\bar{P} \in [\tilde{P}_j, P^0]$, $\bar{\theta} \in [\tilde{\theta}_{j-1}, \theta^0]$, $D_{P,n}^q(\theta, \bar{P}, \bar{\theta}) = n^{-1} \sum_{i=1}^n \frac{\nabla_{P^{*\prime}} \tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)}{\tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)}$, and $D_{\theta,n}^q(\theta, \bar{P}, \bar{\theta}) = n^{-1} \sum_{i=1}^n \frac{\nabla_{\theta^{*\prime}} \tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)}{\tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)}$. Because $(\tilde{P}_j, \tilde{\theta}_{j-1}) \to_p (P^0, \theta^0)$ and $\tilde{\Lambda}^q(\theta, P^*, \theta^*)$ is continuous in (P^*, θ^*), the definition of Θ_j^q implies that, for all $(a, x) \in A \times X$,

$$\tilde{\Lambda}^q(\theta, \tilde{P}, \tilde{\theta})(a|x) \in [\epsilon/2, 1 - \epsilon/2] \text{ uniformly in } \tilde{P} \in [\tilde{P}_j, P^0], \tilde{\theta} \in [\tilde{\theta}_{j-1}, \theta^0] \text{ and } \theta \in \Theta_j^q. \tag{29}$$

Consequently, $||D_{P,n}^q(\theta, \bar{P}, \bar{\theta})|| < C||n^{-1} \sum_{i=1}^n \nabla_{P^{*\prime}} \tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)|| = O_p(1)$ uniformly in $\theta \in \Theta_j^q$, where the last equality follows from Assumption 3(b) and the consistency of $(\tilde{P}_j, \tilde{\theta}_{j-1})$. Similarly, $||D_{\theta,n}^q(\theta, \bar{P}, \bar{\theta})|| = O_p(1)$ uniformly in $\theta \in \Theta_j^q$. Then, (25) follows from (28) and $(\tilde{P}_j, \tilde{\theta}_{j-1}) \to_p (P^0, \theta^0)$.

We proceed to show (26). Note that, since $\Lambda^q(P^0, \theta^0) = P^0$,

$$Q_n^q(\theta, P^0, \theta^0) = n^{-1} \sum_{i=1}^n \ln \tilde{\Lambda}^q(\theta, P^0, \theta^0)(a_i|x_i) = n^{-1} \sum_{i=1}^n \ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)(a_i|x_i).$$

Since Θ_j^q is compact and $\ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)$ is continuous in $\theta \in \Theta_j^q$, (26) follows from Lemma 2.4 of Newey and McFadden (1994) if we show $E \sup_{\theta \in \Theta_j^q} |\ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)(a_i|x_i)| < \infty$. Recall $\ln(a) < a$ for all $a > 0$. Applying this inequality with $a = |(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)(a_i|x_i)|$ in conjunction with the definition of Θ_j^q, we have

$$\ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)(a_i|x_i) < (\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)(a_i|x_i) < 1,$$

for any $\theta \in \Theta_j^q$. Then, since Assumption 3(a) implies $E \sup_{\theta \in \Theta_j^q} |(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0))(a_i|x_i)| > 0$, we have $E \sup_{\theta \in \Theta_j} |\ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)(a_i|x_i)| < 1 \leq CE \sup_{\theta \in \Theta_j^q} |\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0)(a_i|x_i)| \leq CE||\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(a_i|x_i)|| \sup_{\theta \in \Theta_j^q} ||\theta - \theta^0|| < \infty$, and (26) follows.

It remains to show (27). $Q^q(\theta)$ is continuous in θ from Lemma 2.4 of Newey and McFadden

(1994) and the proof of (26). Note that

$$\begin{aligned} Q^q(\theta) - Q^q(\theta^0) &= E\ln(\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(\theta-\theta^0) + P^0)(a_i|x_i) - E\ln P^0(a_i|x_i) \\ &= E\ln\left(\frac{\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0)}{P^0(a_i|x_i)} + 1\right). \end{aligned}$$

We show

$$E\ln\left(\frac{\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0)}{P^0(a_i|x_i)} + 1\right) < E\left[\frac{\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0)}{P^0(a_i|x_i)}\right] \quad \text{for all } \theta \neq \theta^0, \tag{30}$$

then $Q^q(\theta) - Q^q(\theta^0) < 0$ for all $\theta \neq \theta^0$ because $E[\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)/P^0(a_i|x_i)] = 0$.

Recall $\ln(1+x) \leq x$ for all $x > -1$ where the inequality is strict if $x \neq 0$. Thus, (30) holds if, for all $\theta \neq \theta^0$, we have $\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0)/P^0(a_i|x_i) \neq 0$ with positive probability. Since $P^0(a_i|x_i)$ is bounded away from both 0 and ∞, this is implied by Assumption 3. Hence, (30) holds, and (27) is shown. Therefore, $\tilde{\theta}_j \to_p \theta^0$. □

9.7 Proof of Proposition 7

To analyze $\tilde{\theta}_j$, let us introduce a simplified notation for the objective function in the jth iteration:

$$Q_n^{q(j)}(\theta) \equiv Q_n^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = n^{-1}\sum_{i=1}^n \ln\tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_i|x_i),$$

where $\tilde{\Lambda}^q(\theta, P^*, \theta^*) = \nabla_{\theta'}\Lambda^q(P^*,\theta^*)(\theta-\theta^*) + \Lambda^q(P^*,\theta^*)$.

The estimate $\tilde{\theta}_j$ satisfies the first order condition: $\nabla_{\theta'}Q_n^{q(j)}(\tilde{\theta}_j) = 0$. Applying a second-order Taylor expansion to each element of $\nabla_{\theta'}Q_n^{q(j)}(\tilde{\theta}_j)$ around $\tilde{\theta}_{j-1}$, we obtain

$$\begin{aligned} 0 &= \nabla_{\theta'}Q_n^{q(j)}(\tilde{\theta}_j) = \nabla_{\theta'}Q_n^{q(j)}(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j - \tilde{\theta}_{j-1})'\nabla_{\theta\theta'}Q_n^{q(j)}(\tilde{\theta}_{j-1}) \\ &\quad + [(\tilde{\theta}_j - \tilde{\theta}_{j-1})'B_1(\tilde{\theta}_j - \tilde{\theta}_{j-1}), \cdots, (\tilde{\theta}_j - \tilde{\theta}_{j-1})'B_K(\tilde{\theta}_j - \tilde{\theta}_{j-1})], \end{aligned} \tag{31}$$

where B_k, $k = 1, \ldots, K$, is the second derivative of the kth element of $\nabla_{\theta'}Q_n^{q(j)}(\theta)$ evaluated at $\bar{\theta} \in [\tilde{\theta}_j, \tilde{\theta}_{j-1}]$. We find an alternate expression for the last term on the right. Note that

$$\begin{aligned} (\tilde{\theta}_j - \tilde{\theta}_{j-1})'B_k(\tilde{\theta}_j - \tilde{\theta}_{j-1}) &= (\tilde{\theta}_j - \hat{\theta}_{qNPL} + \hat{\theta}_{qNPL} - \tilde{\theta}_{j-1})'B_k(\tilde{\theta}_j - \hat{\theta}_{qNPL} + \hat{\theta}_{qNPL} - \tilde{\theta}_{j-1}) \\ &= (\tilde{\theta}_j - \hat{\theta}_{qNPL})'C_k + (\hat{\theta}_{qNPL} - \tilde{\theta}_{j-1})'B_k(\hat{\theta}_{qNPL} - \tilde{\theta}_{j-1}), \end{aligned} \tag{32}$$

where $C_k = B_k[(\tilde{\theta}_j - \hat{\theta}_{qNPL}) + 2(\hat{\theta}_{qNPL} - \tilde{\theta}_{j-1})] = o_p(1)$ for all k because $\hat{\theta}_{qNPL}$, $\tilde{\theta}_j$, and $\tilde{\theta}_{j-1}$ are consistent, and Assumption 4 implies $B_k = O_p(1)$. Substituting this to the last term on the

right of (31), we get

$$0 = \nabla_{\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j - \tilde{\theta}_{j-1})' \nabla_{\theta\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j - \hat{\theta}_{qNPL})' o_p(1) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2). \quad (33)$$

For the first term on the right of (33), define $L_n^q(P,\theta) = n^{-1}\sum_{i=1}^n \nabla_{\theta'} \ln \Lambda^q(P,\theta)(a_i|x_i)$, and then we can write $\nabla_{\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) = L_n^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$. Expanding this around $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ and using $L_n^q(\hat{P}_{qNPL}, \hat{\theta}_{qNPL}) = n^{-1}\sum_{i=1}^n \nabla_{\theta'} \ln \Lambda^q(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})(a_i|x_i) = 0$ gives

$$\begin{aligned}
\nabla_{\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) &= (\tilde{P}_{j-1} - \hat{P}_{qNPL})' \nabla_P L_n^q(\hat{\theta}_{qNPL}, \hat{P}_{qNPL}) + (\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL})' \nabla_\theta L_n^q(\hat{\theta}_{qNPL}, \hat{P}_{qNPL}) \\
&\quad + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2) + O_p(||\tilde{P}_{j-1} - \hat{P}_{qNPL}||^2) \\
&= (\tilde{P}_{j-1} - \hat{P}_{qNPL})' E \nabla_{P\theta'} \ln \Lambda^q(P^0, \theta^0) + (\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL})' E \nabla_{\theta\theta'} \ln \Lambda^q(P^0, \theta^0) + r_{n,j}, \quad (34)
\end{aligned}$$

where $r_{n,j}$ denotes a generic reminder term of the form

$$r_{n,j} = O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{qNPL}||^2) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{qNPL}||),$$

and the last equality of (34) follows from expanding $\nabla_P L_n(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ and $\nabla_\theta L_n(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ around (P^0, θ^0) and using the root-n consistency of $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$.

For the second term on the right of (33), define a $1 \times K$ vector $g_i^q = \nabla_{\theta'} \Lambda^q(\tilde{P}_j, \tilde{\theta}_{j-1})(a_i|x_i)$, then $\nabla_{\theta\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) = -n^{-1}\sum_{i=1}^n \frac{g_i^{q'} g_i^q}{(\Lambda^q(\tilde{P}_j, \tilde{\theta}_{j-1})(a_i|x_i))^2}$. Therefore, in view of the root-n consistency of $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$, we obtain

$$\begin{aligned}
\nabla_{\theta\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) &= -E[\nabla_\theta \ln \Lambda^q(P^0, \theta^0)(a_i|x_i) \nabla_{\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] \\
&\quad + O_p(n^{-1/2}) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||) + O_p(||\tilde{P}_j - \hat{P}_{qNPL}||). \quad (35)
\end{aligned}$$

Substituting (34) and (35) into (33) and using $E[\nabla_\theta \ln \Lambda^q(P^0, \theta^0)(a_i|x_i) \nabla_{\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)]$ $+ E[\nabla_{\theta\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] = 0$ gives

$$\begin{aligned}
&\{E[\nabla_\theta \ln \Lambda^q(P^0, \theta^0)(a_i|x_i) \nabla_{\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] + O_p(n^{-1/2})\}(\tilde{\theta}_j - \hat{\theta}_{qNPL}) \\
= \; & E[\nabla_{\theta P'} \Lambda^q(P^0, \theta^0)(a_i|x_i)](\tilde{P}_{j-1} - \hat{P}_{qNPL}) + r_{n,j}. \quad (36)
\end{aligned}$$

It follows that $\tilde{\theta}_j - \hat{\theta}_{qNPL} = O_p(||\tilde{P}_{j-1} - \hat{P}_{qNPL}||)$.

To obtain the updating formula of $\tilde{P}_j$, expand $\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_j)$ around $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ and use the root-n consistency of $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ to get

$$\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_j) = \hat{P}_{qNPL} + \Lambda_P^q(\tilde{P}_{j-1} - \hat{P}_{qNPL}) + \Lambda_\theta^q(\tilde{\theta}_j - \hat{\theta}_{qNPL}) + r_{n,j}, \quad (37)$$

where $\Lambda_P^q \equiv \nabla_{P'} \Lambda^q(P^0, \theta^0)$ and $\Lambda_\theta^q \equiv \nabla_{\theta'} \Lambda^q(P^0, \theta^0)$.

Using matrix notations of $E[\nabla_\theta \ln \Lambda^q(P^0, \theta^0)(a_i|x_i) \nabla_{\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] = (\Lambda_\theta^q)' \Delta_P \Lambda_\theta^q$

and $E[\nabla_{\theta P'}\Lambda^q(P^0,\theta^0)(a_i|x_i)] = -(\Lambda^q_\theta)'\Delta_P\Lambda^q_P$, (36) is written as $\tilde{\theta}_j - \hat{\theta}_{qNPL} = -\{(\Lambda^q_\theta)'\Delta_P\Lambda^q_\theta + O_p(n^{-1/2})\}^{-1}(\Lambda^q_\theta)'\Delta_P\Lambda^q_P(\tilde{P}_{j-1} - \hat{P}_{qNPL}) + r_{n,j}$. Substituting this expression for $\tilde{\theta}_j - \hat{\theta}_{qNPL}$ into (37) gives the stated convergence rate of $\tilde{P}_j$. □

9.8 Proof of Proposition 8

The proof is similar to the proof of Proposition 6 and omitted. □

9.9 Proof of Proposition 9

The proof is similar to the proof of Proposition 7 and omitted. Note that $\Lambda^q(P,\theta)$, $\nabla_{\theta'}\Lambda^q(P,\theta)$, and $\nabla_{P'}\Lambda^q(P,\theta)$ are replaced with P_θ, $\nabla_{\theta'}P_\theta$ and $\nabla_{P'}P_\theta = 0$, respectively, and $\nabla_{P'}P_{\theta^0} = 0$ in this proposition corresponds to $\Lambda^q_P = 0$ in Proposition 7.

9.10 Proof of Proposition 10

The proof is similar to that of Proposition 6 and omitted. □

9.11 Proof of Proposition 11

The updating formula of $\tilde{P}_j$ follows simply from expanding $\Lambda^q(\tilde{P}_{j-1},\tilde{\theta}_{j-1})$ around $(\hat{P}_{MLE},\hat{\theta}_{MLE})$:

$$\begin{aligned}\tilde{P}_j &= \Lambda^q(\tilde{P}_{j-1},\tilde{\theta}_{j-1}) \\ &= \hat{P}_{MLE} + \nabla_{P'}\Lambda^q(P^0,\theta^0)(\tilde{P}_{j-1} - \hat{P}_{MLE}) + \nabla_{\theta'}\Lambda^q(P^0,\theta^0)(\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) + O_p(||\tilde{P}_{j-1} - \hat{P}_{MLE}||^2) \\ &\quad + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{MLE}||) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||),\end{aligned}$$

where the order of $O_p(\cdot)$ terms in the second equality follows from Assumption 8(b) and the root-n consistency of $(\hat{P}_{MLE},\hat{\theta}_{MLE})$.

Define the objective function in the jth iteration by

$$Q^{(j)}_n(\theta) \equiv Q_n(\theta,\tilde{P}_j,\tilde{\theta}_{j-1}) = n^{-1}\sum_{i=1}^n \ln\Phi(\theta,\tilde{P}_j,\tilde{\theta}_{j-1})(a_i|x_i),$$

where $\Phi(\theta,P^*,\theta^*) = (I - \nabla_{P'}\Psi(P^*,\theta^*))^{-1}\nabla_{\theta'}\Psi(P^*,\theta^*)(\theta-\theta^*) + P^*$ as defined in (15). Expanding the first order condition $0 = \nabla_{\theta'}Q^{(j)}_n(\tilde{\theta}_j)$ around $\tilde{\theta}_{j-1}$, we obtain

$$\begin{aligned}0 &= \nabla_{\theta'}Q^{(j)}_n(\tilde{\theta}_j) = \nabla_{\theta'}Q^{(j)}_n(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j - \tilde{\theta}_{j-1})'\nabla_{\theta\theta'}Q^{(j)}_n(\tilde{\theta}_{j-1}) \\ &\quad + [(\tilde{\theta}_j - \tilde{\theta}_{j-1})'B_1(\tilde{\theta}_j - \tilde{\theta}_{j-1}),\cdots,(\tilde{\theta}_j - \tilde{\theta}_{j-1})'B_K(\tilde{\theta}_j - \tilde{\theta}_{j-1})], \qquad (38)\end{aligned}$$

where B_k, $k = 1,\ldots,K$, is the second derivative of the kth element of $\nabla_{\theta'}Q^{(j)}_n(\theta)$ evaluated at

$\bar{\theta} \in [\tilde{\theta}_j, \tilde{\theta}_{j-1}]$. Analogous to (32), each column of the last term on the right of (38) is written as

$$(\tilde{\theta}_j - \tilde{\theta}_{j-1})' B_k (\tilde{\theta}_j - \tilde{\theta}_{j-1}) = (\tilde{\theta}_j - \hat{\theta}_{MLE})' C_k + (\hat{\theta}_{MLE} - \tilde{\theta}_{j-1})' B_k (\hat{\theta}_{MLE} - \tilde{\theta}_{j-1}), \tag{39}$$

where $C_k = B_k[(\tilde{\theta}_j - \hat{\theta}_{MLE}) + 2(\hat{\theta}_{MLE} - \tilde{\theta}_{j-1})] = o_p(1)$ for all k because $\tilde{\theta}_j$, $\tilde{\theta}_{j-1}$, and $\hat{\theta}$, are consistent and Assumption 8(c) implies $B_k = O_p(1)$ for all k. Substituting this to the last term on the right of (38), we can rewrite the first order condition (38) as

$$0 = \nabla_{\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j - \tilde{\theta}_{j-1})' \nabla_{\theta\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j - \hat{\theta}_{MLE})' o_p(1) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2). \tag{40}$$

For the first term on the right of (40), define

$$L_n(P, \theta) = n^{-1} \sum_{i=1}^{n} \frac{[(I - \nabla_{P'} \Psi(P, \theta))^{-1} \nabla_{\theta'} \Psi(P, \theta)](a_i|x_i)}{P(a_i|x_i)},$$

then we have $\nabla_{\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) = L_n(\tilde{P}_j, \tilde{\theta}_{j-1})$. Since the MLE $(\hat{P}_{MLE}, \hat{\theta}_{MLE})$ satisfies $L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE}) = n^{-1} \sum_{i=1}^{n} \nabla_{\theta'} \ln P_{\hat{\theta}_{MLE}}(a_i|x_i) = 0$, expanding $L_n(\tilde{P}_j, \tilde{\theta}_{j-1})$ around $(\hat{P}_{MLE}, \hat{\theta}_{MLE})$ gives

$$\begin{aligned} \nabla_{\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) &= (\tilde{P}_j - \hat{P}_{MLE})' \nabla_P L_n(\hat{\theta}_{MLE}, \hat{P}_{MLE}) + (\tilde{\theta}_{j-1} - \hat{\theta}_{MLE})' \nabla_\theta L_n(\hat{\theta}_{MLE}, \hat{P}_{MLE}) \\ &\quad + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) + O_p(||\tilde{P}_j - \hat{P}_{MLE}||^2). \end{aligned} \tag{41}$$

where the order of the $O_p(\cdot)$ term follows from Assumption 8(c).

We proceed to obtain approximations of $\nabla_P L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE})$ and $\nabla_\theta L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE})$ in (41). First, $\nabla_P L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE}) = \mathcal{J}' + O_p(n^{-1/2})$ from expanding it around (P^0, θ^0) and using the root-n consistency of $(\hat{P}_{MLE}, \hat{\theta}_{MLE})$. For $\nabla_{\theta'} L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE})$, note that $\nabla_{\theta'} \ln P_\theta(a|x) = \frac{[(I - \nabla_{P'} \Psi(P_\theta, \theta))^{-1} \nabla_{\theta'} \Psi(P_\theta, \theta)](a|x)}{P_\theta(a|x)}$ and

$$\begin{aligned} \nabla_\theta [\nabla_{\theta'} \ln P_\theta(a|x)] &= \nabla_\theta \left\{ \frac{[(I - \nabla_{P'} \Psi(P, \theta))^{-1} \nabla_{\theta'} \Psi(P, \theta)](a|x)}{P(a|x)} \right\} \Bigg|_{P = P_\theta} \\ &\quad + \nabla_\theta (P_\theta)' \nabla_P \left\{ \frac{[(I - \nabla_{P'} \Psi(P, \theta))^{-1} \nabla_{\theta'} \Psi(P, \theta)](a|x)}{P(a|x)} \right\} \Bigg|_{P = P_\theta}. \end{aligned}$$

Consequently, in light of $\hat{P}_{MLE} = P_{\hat{\theta}_{MLE}}$, we have $\nabla_\theta L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE}) = n^{-1} \sum_{i=1}^{n} \nabla_{\theta\theta'} \ln P_{\hat{\theta}_{MLE}}(a_i|x_i) - \nabla_\theta (P_{\hat{\theta}_{MLE}})' \nabla_P L_n(P_{\hat{\theta}_{MLE}}, \hat{\theta}_{MLE}) = E \nabla_{\theta\theta'} \ln P_{\theta^0}(a_i|x_i) - \nabla_\theta (P_{\theta^0})' \mathcal{J}' + O_p(n^{-1/2})$. Substituting these into the right hand side of (41) gives

$$\nabla_{\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) = (\tilde{P}_j - \hat{P}_{MLE})' \mathcal{J}' + (\tilde{\theta}_{j-1} - \hat{\theta}_{MLE})' (E \nabla_{\theta\theta'} \ln P_{\theta^0}(a_i|x_i) - \nabla_\theta (P_{\theta^0})' \mathcal{J}') + r_n, \tag{42}$$

where $r_n = O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) + O_p(||\tilde{P}_j - \hat{P}_{MLE}||^2) + O_p(n^{-1/2} ||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||) + O_p(n^{-1/2} ||\tilde{P}_j - \hat{P}_{MLE}||)$.

For the second term on the right of (40), define a $1 \times K$ vector $g_i = [(I - \nabla_{P'}\Psi(\tilde{P}_j, \tilde{\theta}_{j-1}))^{-1} \nabla_{\theta'}\Psi(\tilde{P}_j, \tilde{\theta}_{j-1})](a_i|x_i)$, then $\nabla_{\theta\theta'}Q_n^{(j)}(\tilde{\theta}_{j-1}) = -n^{-1}\sum_{i=1}^n \frac{g_i'g_i}{(\tilde{P}_j(a_i|x_i))^2}$. Therefore, in view of Assumption 8(c), $\hat{P}_{MLE} = P_{\hat{\theta}_{MLE}}$, and the root-$n$ consistency of $\hat{\theta}_{MLE}$, we obtain

$$\begin{aligned}\nabla_{\theta\theta'}Q_n^{(j)}(\tilde{\theta}_{j-1}) &= -E[\nabla_\theta \ln P_{\theta^0}(a_i|x_i)\nabla_{\theta'} \ln P_{\theta^0}(a_i|x_i)] \\ &\quad +O_p(n^{-1/2}) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||) + O_p(||\tilde{P}_j - \hat{P}_{MLE}||). \qquad (43)\end{aligned}$$

Substituting (42) and (43) into (40) and using $E[\nabla_\theta \ln P_{\theta^0}(a_i|x_i)\nabla_{\theta'} \ln P_{\theta^0}(a_i|x_i)]$ $+ E[\nabla_{\theta\theta'} \ln P_{\theta^0}(a_i|x_i)] = 0$ gives

$$\begin{aligned}&\{E[\nabla_\theta \ln P_{\theta^0}(a_i|x_i)\nabla_{\theta'} \ln P_{\theta^0}(a_i|x_i)] + o_p(1)\}(\tilde{\theta}_j - \hat{\theta}_{MLE}) \\ = \; &-\mathcal{J}\nabla_{\theta'}P_{\theta^0}(\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) + \mathcal{J}(\tilde{P}_j - \hat{P}_{MLE}) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) + O_p(||\tilde{P}_j - \hat{P}_{MLE}||^2) \\ &+O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||) + O_p(n^{-1/2}||\tilde{P}_j - \hat{P}_{MLE}||),\end{aligned}$$

giving the stated result. □

9.12 Proof of Proposition 12

The marginal conditions are given by

$$\begin{aligned}\bar{G}_\theta(\Psi(\tilde{P}, \tilde{\theta}))'\hat{W}\bar{g}(\Psi(\tilde{P}, \tilde{\theta})) &= 0, \\ \tilde{P} - \Psi(\tilde{P}, \tilde{\theta}) &= 0.\end{aligned}$$

Expanding $\bar{g}(\Psi(\tilde{P}, \tilde{\theta}))$ around (P^0, θ^0) and using $||\hat{f}_x - f_x|| = O_p(n^{-1/2})$ give

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

Eliminating $(\tilde{P} - P^0)$ from these equations and using $G_\theta'WG_\theta + G_\theta'WG_P(I - \Psi_P)^{-1}\Psi_\theta = G_\theta'WG_\theta^\infty$, where $G_\theta^\infty = (\partial/\partial\theta')\bar{g}(P_{\theta^0}) = -H\Delta_x(I - \Psi_P)^{-1}\Psi_\theta$, we have

$$\sqrt{n}(\tilde{\theta} - \theta^0) \to_d N(0, (G_\theta'WG_\theta^\infty)^{-1}G_\theta'W\Omega W'G_\theta((G_\theta^\infty)'W'G_\theta)^{-1}),$$

where $\Omega = E[g(a_i, x_i; P^0)g(a_i, x_i; P^0)']$. □

9.13 Proof of Proposition 13

Recall that $\tilde{\theta}_j$ satisfies the first order condition

$$\bar{G}_\theta(\Psi(\tilde{P}_{j-1}, \tilde{\theta}_j))\hat{W}\bar{g}(\Psi(\tilde{P}_{j-1}, \tilde{\theta}_j)) = 0. \qquad (44)$$

Expanding $\bar{g}(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))$ around $(\tilde{P},\tilde{\theta})$ in (44) and using $\bar{G}'_\theta(\Psi(\tilde{P},\tilde{\theta}))\hat{W}\bar{g}(\Psi(\tilde{P},\tilde{\theta}))=0$ gives

$$\begin{aligned}\tilde{\theta}_j-\tilde{\theta} &= [\bar{G}'_\theta(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{G}_\theta(\Psi(\bar{P},\bar{\theta}))+o_p(1)]^{-1}[\bar{G}'_\theta(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{G}_P(\Psi(\bar{P},\bar{\theta}))+o_p(1)](\tilde{P}_{j-1}-\tilde{P})\\ &= O_p(||\tilde{P}_{j-1}-\tilde{P}||), \qquad (45)\end{aligned}$$

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

For the second result, first, using (45), we obtain the same approximation as (21):

$$\tilde{P}_j-\tilde{P}=\Psi_P(\tilde{P}_{j-1}-\tilde{P})+\Psi_\theta(\tilde{\theta}_j-\tilde{\theta})+O_p(n^{-1/2}||\tilde{P}_{j-1}-\tilde{P}||)+O_p(||\tilde{P}_{j-1}-\tilde{P}||^2) \qquad (46)$$

Expanding $\bar{g}(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))$ in (44) twice around $(\tilde{P},\tilde{\theta})$ and using $\bar{G}'_\theta(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{g}(\Psi(\tilde{P},\tilde{\theta}))=O_p(n^{-1/2}||\tilde{\theta}_j-\tilde{\theta}||)+O_p(n^{-1/2}||\tilde{P}_{j-1}-\tilde{P}||)$,

$$\bar{G}_P(\Psi(\tilde{P},\tilde{\theta}))=G_P+O_p(n^{-1/2}), \qquad \bar{G}_\theta(\Psi(\tilde{P},\tilde{\theta}))=G_\theta+O_p(n^{-1/2}) \qquad (47)$$

and (45) gives

$$\begin{aligned}0 &= \bar{G}'_\theta(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}G_P(\tilde{P}_{j-1}-\tilde{P})+\bar{G}'_\theta(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}G_\theta(\tilde{\theta}_j-\tilde{\theta})\\ &\quad +O_p(n^{-1/2}||\tilde{P}_{j-1}-\tilde{P}||)+O_p(||\tilde{P}_{j-1}-\tilde{P}||^2). \qquad (48)\end{aligned}$$

Expanding $\Psi(\tilde{P}_{j-1},\tilde{\theta}_j)$ around $(\tilde{P},\tilde{\theta})$ and using (45) and (47) in (48), we have

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

Substituting this into (46) and noting that $G_\theta=-H\Delta_x\Psi_\theta$ and $G_P=-H\Delta_x\Psi_P$, we obtain

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

and the second result follows. □

9.14 Proof of Lemma 2

The proof follows the proof of Lemma 1. Expanding the first order condition $\bar{l}_\zeta(\tilde{\mathbf{P}}_{j-1},\tilde{\zeta}_j)=\bar{l}_\zeta(\hat{\mathbf{P}}_{NPL},\hat{\zeta}_{NPL})=0$ gives

$$\tilde{\zeta}_j-\hat{\zeta}_{NPL}=-\bar{l}_{\zeta\zeta}(\bar{\mathbf{P}},\bar{\zeta})^{-1}\bar{l}_{\zeta P}(\bar{\mathbf{P}},\bar{\zeta})(\tilde{\mathbf{P}}_{j-1}-\hat{\mathbf{P}}_{NPL})=O_p(||\tilde{\mathbf{P}}_{j-1}-\hat{\mathbf{P}}_{NPL}||). \qquad (49)$$

where $(\bar{\mathbf{P}},\bar{\zeta})$ is between $(\tilde{\mathbf{P}}_{j-1},\tilde{\zeta}_j)$ and $(\hat{\mathbf{P}}_{NPL},\hat{\zeta}_{NPL})$. This gives the bound for $\tilde{\zeta}_j-\hat{\zeta}_{NPL}$. Rewriting this further using the first three assumptions of Assumption 12 gives

$$\tilde{\zeta}_j-\hat{\zeta}_{NPL}=-\Omega_{\zeta\zeta}^{-1}\Omega_{\zeta P}(\tilde{\mathbf{P}}_{j-1}-\hat{\mathbf{P}}_{NPL})+O_p(n^{-1/2}||\tilde{\mathbf{P}}_{j-1}-\hat{\mathbf{P}}_{NPL}||)+O_p(||\tilde{\mathbf{P}}_{j-1}-\hat{\mathbf{P}}_{NPL}||^2). \qquad (50)$$

On the other hand, expanding the second step equation $\tilde{\mathbf{P}}_j = \mathbf{\Psi}(\tilde{\mathbf{P}}_{j-1}, \tilde{\zeta}_j)$ twice around $(\hat{\mathbf{P}}_{NPL}, \hat{\zeta}_{NPL})$, using root-n consistency of $(\hat{\mathbf{P}}_{NPL}, \hat{\zeta}_{NPL})$ and (49) give

$$\tilde{\mathbf{P}}_j - \hat{\mathbf{P}}_{NPL} = \mathbf{\Psi}_P(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) + \mathbf{\Psi}_\zeta(\tilde{\zeta}_j - \hat{\zeta}_{NPL}) + O_p(n^{-1/2}||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||) + O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||^2), \tag{51}$$

where $\mathbf{\Psi}_\zeta \equiv (\partial/\partial\zeta')\mathbf{\Psi}(\mathbf{P}^0, \theta^0) = [\mathbf{\Psi}_\theta, \mathbf{0}]$. Substituting (50) into (51) gives

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

Note that

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

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

10 Appendix B: Additional Results

10.1 Relative efficiency of NPL, q-NPL, and MLE

The variance of the NPL estimator is given by

$$\begin{aligned} V_{NPL} &= [\Omega_{\theta\theta} + \Omega_{\theta P}(I - \Psi_P)^{-1}\Psi_\theta]^{-1} \Omega_{\theta\theta} [\Omega_{\theta\theta} + \Psi_\theta(I - \Psi_P')^{-1}\Omega_{\theta P}']^{-1} \\ &= \Psi_\theta'(I - \Psi_P)^{-1} \Delta_P \Psi_\theta (\Psi_\theta' \Delta_P \Psi_\theta)^{-1} \Psi_\theta' \Delta_P (I - \Psi_P')^{-1} \Psi_\theta \end{aligned}$$

while the variance of the MLE is

$$V_{MLE} = \left(E\left[\frac{\Psi_\theta'(I - \Psi_P)^{-1}(a|x)}{P_\theta(a|x)} \frac{(I - \Psi_P')^{-1}\Psi_\theta(a|x)}{P_\theta(a|x)} \right] \right)^{-1} = \left(\Psi_\theta'(I - \Psi_P)^{-1} \Delta_P (I - \Psi_P')^{-1} \Psi_\theta \right)^{-1}.$$

Define $B = \Delta_P^{1/2}\Psi_\theta$ and $D = \Delta_P^{1/2}(I - \Psi_P)^{-1}\Psi_\theta$. Then $V_{NPL}^{-1} = D'B(B'B)^{-1}B'D$, $V_{MLE}^{-1} = D'D = D'D(D'D)^{-1}D'D$, and $V_{MLE}^{-1} - V_{NPL}^{-1} = D'[I - B(B'B)^{-1}B']D = UU'$, where $U = D'[I - B(B'B)^{-1}B']$. Therefore, $V_{MLE}^{-1} - V_{NPL}^{-1}$ is positive semi-definite.

Next, consider the variance of q-NPL estimator, denoted by V_{qNPL}. First, evaluating the derivatives at $P = P_\theta$, we have $\Psi_\theta^q \equiv \nabla_{\theta'}\Psi^q(P_\theta, \theta) = (I - \Psi_P)^{-1}(I - \Psi_P^q)\Psi_\theta$ and $\Psi_P^q \equiv \nabla_{P'}\Psi^q(P_\theta, \theta) = (\Psi_P)^q$. Taking a derivative of $P_\theta = \Psi^q(P_\theta, \theta) = \Psi(P_\theta, \theta)$ with respect to θ gives $(\Psi_\theta^q)'(I - \Psi_P^q)^{-1} = \Psi_\theta'(I - \Psi_P)^{-1}$. Using this and defining $B_q \equiv \Delta_P^{1/2}\Psi_\theta^q = \Delta_P^{1/2}(I - \Psi_P)^{-1}(I - \Psi_P^q)\Psi_\theta$, we have $V_{qNPL}^{-1} = D'B_q(B_q'B_q)^{-1}B_q'D$. It follows that $V_{MLE}^{-1} - V_{qNPL}^{-1} = U_qU_q'$ with $U_q = D'[I - B_q(B_q'B_q)^{-1}B_q']$.

Note that $D - B_q = \Delta_P^{1/2}(I - \Psi_P)^{-1}\Psi_P^q\Psi_\theta = O(|\lambda^*|^q)$, where λ^* is the dominant eigenvalue

of Ψ_P. If all the eigenvalues of Ψ_P are less than one in absolute value, then $B_q \to D$ as $q \to \infty$ so that $V_{qNPL} \to V_{MLE}$ as $q \to \infty$. Expanding $D'B_q(B_q'B_q)^{-1}B_q'D$ around $B_q = D$ gives $V_{qNPL}^{-1} - V_{MLE}^{-1} = O(||B_q - D||^2) = O(|\lambda^*|^{2q})$.

10.2 The first order condition of (7) with Ψ and Λ

Without loss of generality, let $A = \{1, 2, ..., J\}$. Then, using that $[\Psi(P,\theta)](J|x) = 1 - \sum_{j=1}^{J-1}[\Psi(P,\theta)](j|x)$, the first order condition of the maximization problem in (7) is given by

$$n^{-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{J-1}\frac{1(a_i=j)[\nabla_{\theta'}\Psi(P,\theta)](j|x_i)}{[\Psi(P,\theta)](j|x_i)} - \frac{1(a_i=J)\sum_{s=1}^{J-1}[\nabla_{\theta'}\Psi(P,\theta)](s|x_i)}{1-\sum_{s=1}^{J-1}[\Psi(P,\theta)](s|x)}\right) = 0.$$

When the mapping Ψ is replaced with $\Lambda(P,\theta) = \{\Psi(P,\theta)\}^{\alpha}P^{1-\alpha}$, the corresponding first order condition becomes $n^{-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{J-1}\frac{1(a_i=j)[\nabla_{\theta'}\Lambda(P,\theta)](j|x_i)}{[\Lambda(P,\theta)](j|x_i)} - \frac{1(a_i=J)\sum_{s=1}^{J-1}[\nabla_{\theta'}\Lambda(P,\theta)](s|x_i)}{1-\sum_{s=1}^{J-1}[\Lambda(P,\theta)](s|x)}\right) = 0$, where $\nabla_{\theta'}\Lambda(P,\theta) = \alpha\{\Psi(P,\theta)\}^{\alpha-1}P^{1-\alpha}\nabla_{\theta'}\Psi(P,\theta)$. Evaluated at the fixed point $\hat{P}_{NPL} = \Psi(\hat{P}_{NPL},\hat{\theta}_{NPL}) = \Lambda(\hat{P}_{NPL},\hat{\theta}_{NPL})$, we have $\nabla_{\theta'}\Lambda(\hat{P}_{NPL},\hat{\theta}_{NPL}) = \alpha\nabla_{\theta'}\Psi(\hat{P}_{NPL},\hat{\theta}_{NPL})$ and these two first order conditions becomes identical.

10.3 Fixed point mapping Ψ for Monte Carlo Experiments

Denote equilibrium best response probabilities by $P^* = \{P_i^*(a_i|x), i = 1, ..., N\}$ and firm's value functions associated with this equilibrium by $V_1^{P^*}, ..., V_N^{P^*}$. Then,

$$V_i^{P^*}(x_t) = \sum_{a_{it}\in A} P_i^*(a_{it}|x_t)[\pi_i^{P^*}(a_{it},x_t;\theta) + e_i^{P^*}(a_{it},x_t)] + \beta\sum_{x_{t+1}\in X} V_i^{P^*}(x_{t+1})f^{P^*}(x_{t+1}|x_t)$$

where $e_i^{P^*}(a_{it},x_t) = \text{Euler's constant} - \ln(P_i^*(a_{it},x_t))$, $\pi_i^{P^*}(a_{it},x_t;\theta) = \sum_{a_{-i}\in A^{N-1}}\left(\prod_{j\neq i}P_j^*(a_j|x_t)\right)$ $\Pi(a_{it},a_{-i},x_t;\theta)$, and $f^{P^*}(x_{t+1}|x_t) = \left(\prod_{j=1}^{N}P_j^*(a_{jt}|x_t)\right)f_S(S_{t+1}|S_t)$.

We now derive the fixed point mapping Ψ for this model. In terms of matrix notation, denote $F_S = \{f_S(S'|S)\}$, $P_i = \{P_i(a|x)\}$, $P_{-i} = \{\prod_{j\neq i}P_j(a_j|x)\}$, $P = \{\prod_{i=1}^{N}P_i(a_i|x)\}$, and $\iota_k = (1, ..., 1)'$ be a $k \times 1$ vector. Both $e_i^P = \text{Euler's constant} - \ln(P_i)$ and $\pi_i^P(\theta)$ are $|A^N||S| \times |A|$ matrices, where the (i,j)-th element represents the value of $e_i^{P^*}(a_i,x)$ and $\pi_i^{P^*}(a_i,x;\theta)$ corresponding to a pair of the i-th state variable x and the j-th choice a.

Using these notations, we may write $\sum_{a_{it}\in A}P_i^*(a_{it}|x_t)[\pi_i^{P^*}(a_{it},x_t;\theta)+e_i^{P^*}(a_{it},x_t)]$ as $[\pi_i^P(\theta)+e_i^P]P_i'$ while $F^P = (\iota_{|A^N|}\iota_{|A^N|}'\otimes F_S)*(P\otimes\iota_{|S|}')$ represents the transition matrix for $x_t = (a_{t-1},S_t)$, where $*$ represents an element-by-element multiplication. The vector of values V_i^P can be computed as $V_i^P = (I - \beta F^P)^{-1}[\pi_i^P(\theta) + e_i^P]P_i' \equiv T_i(P,\theta)$.

Then, for $i = 1, 2, ..., N$, a fixed point mapping is given by

$$[\Psi_i(P,\theta)](a_i = j|x) = \frac{\exp(\pi_i^{P^*}(j,x;\theta) + \beta \sum_{x' \in X} [T_i(P,\theta)](x') f_i^{P^*}(x'|x,j))}{\sum_{a \in A} \exp(\pi_i^{P^*}(a,x;\theta) + \beta \sum_{x' \in X} [T_i(P,\theta)](x') f_i^{P^*}(x'|x,a))},$$

where $f_i^{P^*}(x_{t+1}|x_t, a_{it}) = \left(\prod_{j \neq i} P_j^*(a_{jt}|x_t)\right) f_S(S_{t+1}|S_t)$.

References

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.

Arcidiacono, P. and R. A. Miller (2008). CCP estimation of dynamic discrete choice models with unobserved heterogeneity. Mimeographed, Duke university.

Bajari, P., Benkard, C.L., and Levin, J. (2007). "Estimating dynamic models of imperfect competition." *Econometrica* 75(5): 1331-1370.

Bajari, P., V. Chernozhukov, and H. Hong (2006). "Semiparametric estimation of a dynamic game of incomplete information." NBER Technical Working Paper 320.

Collard-Wexler, A. (2006) Demand fluctuations and plant turnover in the Ready-Mix concrete industry. Mimeographed, NYU.

Golub, G. and C. Van Loan (1996) *Matrix Computations*, 3rd ed., Baltimore: Johns University Press.

Gourieroux, C., A. Monfort (1995) *Statistics and Econometric Models: Volume Two*, Cambridge: Cambridge 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.

Hotz, J. and R. A. Miller (1993). "Conditional choice probabilities and the estimation of dynamic models." *Review of Economic Studies* 60: 497-529.

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 (2008a) "Nested pseudo-likelihood estimation and bootstrap-based inference for structural discrete Markov decision models," *Journal of Econometrics*, forthcoming.

Kasahara, H. and K. Shimotsu (2008b) "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, forthcoming.

Lütkepohl, H (1996) *Handbook of Matrices*. Wiley.

Newey, W. K. and D. McFadden (1994). "Large Sample Estimation and Hypothesis Testing," in R. F. Engle and D. L. McFadden (eds.) Handbook of Econometrics, Vol. 4, Elsevier.

Pakes, A., M. Ostrovsky, and S. Berry (2007). "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)." *RAND Journal of Economics* 38(2): 373-399.

Pesendorfer, M. and P. Schmidt-Dengler (2007). Asymptotic least squares estimators for dynamic games. Mimeographed, LSE.

Rust, J. (1987). "Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher." *Econometrica* 55(5): 999-1033.

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.

Zeidler, E. (1986) *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems*. New York, Springer-Verlag.

Table 1: The Largest and Smallest Eigenvalues of Ψ_P and Λ_P

θ_{RN}	α^*	$\hat{\alpha}^*$	Eig(Ψ_P)		Eig(Λ_P)		Eig($M_{\Psi_\theta}\Psi_P$)	Eig($M_{\Lambda_\theta}\Lambda_P$)
			λ_{max}	λ_{min}	λ_{max}	λ_{min}		
1	0.9407	0.92	0.2104	-0.3365	0.2572	-0.2572	0.2922	0.2555
2	0.8830	0.83	0.4275	-0.6925	0.4945	-0.4945	0.5996	0.4937
4	0.8250	0.80	0.7596	-1.1839	0.8017	-0.8017	1.1788	0.8056
6	0.7730	0.71	0.8914	-1.4788	0.9161	-0.9161	1.4775	0.9150

A pair $(\lambda_{max}, \lambda_{min})$ represents the largest and the smallest eigenvalues of Ψ_P or Λ_P, while Λ is defined under the value of $\alpha = \alpha^*$ reported in the first column. The last two columns report the absolute value of the dominant eigenvalue of $M_{\Psi_\theta}\Psi_P$ and $M_{\Lambda_\theta}\Lambda_P$, where $M_{\Psi_\theta} = I - \Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P$ and $M_{\Lambda_\theta} = I - \Lambda_\theta(\Lambda_\theta'\Delta_P\Lambda_\theta)^{-1}\Lambda_\theta'\Delta_P$.

Table 2: Bias and MSE

	Estimator	$\theta_{RN} = 2$						$\theta_{RN} = 4$					
		$n = 500$		$n = 2000$		$n = 8000$		$n = 500$		$n = 2000$		$n = 8000$	
		Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$
$\hat{\theta}_{RS}$	PML with Ψ	-0.2277	0.2703	-0.0752	0.1125	-0.0258	0.0502	-0.1162	0.1438	-0.0323	0.0508	-0.0065	0.0196
	NPL with Ψ	-0.0147	0.1415	-0.0038	0.0646	-0.0037	0.0335	-0.0098	0.0685	-0.0056	0.0472	-0.0019	0.0403
	NPL with Λ	-0.0147	0.1415	-0.0038	0.0646	-0.0037	0.0335	0.0036	0.0593	-0.0015	0.0296	0.0011	0.0144
	RPM ($\delta = 0.5$)	-0.0162	0.1399	-0.0063	0.0636	-0.0041	0.0325	0.0033	0.0586	-0.0019	0.0280	0.0008	0.0140
	RPM ($\delta = 0.8$)	-0.0150	0.1410	-0.0038	0.0645	-0.0038	0.0334	0.0016	0.0617	-0.0027	0.0299	0.0010	0.0143
	q-NPL with Λ^q	-0.0135	0.1296	-0.0046	0.0595	-0.0023	0.0301	0.0024	0.0569	-0.0016	0.0278	0.0009	0.0139
	q-AFXP with Λ^q	-0.0131	0.1299	-0.0045	0.0596	-0.0023	0.0302	0.0021	0.0561	-0.0018	0.0276	0.0007	0.0137
$\hat{\theta}_{RN}$	PML with Ψ	-0.8116	0.9555	-0.2681	0.3988	-0.0935	0.1789	-0.7167	0.8270	-0.1798	0.2447	-0.0403	0.0871
	NPL with Ψ	-0.0450	0.4840	-0.0131	0.2285	-0.0144	0.1180	-0.1569	0.2753	-0.1168	0.1956	-0.0982	0.1624
	NPL with Λ	-0.0450	0.4840	-0.0131	0.2285	-0.0144	0.1180	0.0187	0.1346	0.0055	0.0678	0.0043	0.0350
	RPM ($\delta = 0.5$)	-0.0502	0.4798	-0.0223	0.2242	-0.0161	0.1144	0.0196	0.1462	0.0042	0.0688	0.0038	0.0350
	RPM ($\delta = 0.8$)	-0.0451	0.4843	-0.0132	0.2285	-0.0144	0.1181	-0.0099	0.1657	-0.0008	0.0727	0.0043	0.0357
	q-NPL with Λ^q	-0.0413	0.4411	-0.0165	0.2090	-0.0094	0.1052	0.0196	0.1267	0.0049	0.0651	0.0038	0.0330
	q-AFXP with Λ^q	-0.0403	0.4418	-0.0164	0.2094	-0.0092	0.1052	0.0184	0.1221	0.0046	0.0643	0.0034	0.0326
$\hat{P}$	PML with Ψ	-0.0007	0.0215	-0.0001	0.0056	0.0002	0.0019	-0.0010	0.0570	-0.0008	0.0194	-0.0002	0.0047
	NPL with Ψ	0.0002	0.0016	0.0002	0.0004	0.0002	0.0004	-0.0055	0.0346	-0.0020	0.0301	-0.0001	0.0289
	NPL with Λ	0.0002	0.0016	0.0002	0.0004	0.0002	0.0004	0.0000	0.0011	-0.0005	0.0005	0.0000	0.0004
	RPM ($\delta = 0.5$)	0.0002	0.0016	0.0002	0.0005	0.0002	0.0004	0.0000	0.0018	-0.0005	0.0006	0.0000	0.0003
	RPM ($\delta = 0.8$)	0.0002	0.0016	0.0002	0.0004	0.0001	0.0004	-0.0024	0.0102	-0.0009	0.0032	0.0000	0.0004
	q-NPL with Λ^q	0.0002	0.0014	0.0002	0.0004	0.0001	0.0002	-0.0002	0.0010	-0.0005	0.0005	0.0000	0.0003
	q-AFXP with Λ^q	0.0002	0.0014	0.0002	0.0004	0.0001	0.0002	-0.0002	0.0010	-0.0005	0.0006	0.0000	0.0003

The result is based on 500 simulated samples. The maximum number of iterations is set to 50. For RPM, the projection matrix is updated every $J = 10$ iterations. For q-NPL and q-AFXP, we set $q = 4$.