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

April 29, 2007

Abstract

Empirical implications of economic theory are often characterized by a solution to fixed point problem. The major practical obstacle of estimating structural models with fixed point constraint lies in the computational burden. This paper considers a computationally attractive sequential estimation procedure for a class of extremum estimators with fixed point constraint. A sequential algorithm based on the pseudo-likelihood function is proposed by Aguirregabiria and Mira (2002, 2007) to develop an alternative estimator to the two-step estimator of Hotz and Miller (1993) and they show that the limit of the sequential estimators is more efficient than the two-step estimator if the convergence is achieved. To date, however, its convergence property has not been well understood. We analyze the conditions under which the convergence is achieved in the nested pseudo-likelihood algorithm and show that its convergence rate crucially depends on the contraction property of the operator defining the pseudo-likelihood function. We also show that the similar results hold in the context of the sequential algorithm based on GMM.

Keywords: contraction, fixed point, nested pseudo likelihood, nested generalized method of moments, nested minimum distance.

JEL Classification Numbers: C13, C14, C63.

1 Introduction

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

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

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 $E[g(Z_i,P^0)] = 0$ evaluated at the true probability P^0. We also consider classical minimum distance estimator (CMD, hereafter) with $Q_n(P) = -[\hat{P}_0 - P]'\hat{W}[\hat{P}_0 - P]$ where $\hat{P}_0$ is the initial consistent estimator of P^0. Here, $\{Z_i\}_{i=1}^n$ is the sample data drawn from P^0.

The fixed point constraint $P = \Psi(P,\theta)$ in (1) summarizes the set of structural restrictions of the model that is parametrized with a finite 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. The examples of operator $\Psi(\cdot,\theta)$ include the Bellman's operator and policy iteration operator for dynamic programming models (e.g., Rust (1987), Hotz and Miller (1993)), an operator defined by best response functions for games (e.g., Pakes, Ostrovsky and Berry (2005), Pesendorfer and Schmidt-Dengler (2006)), and an operator to define the fixed point problem for recursive competitive equilibrium in dynamic macroeconomic models (e.g., Prescott and Mehra (1980), Aiyagari (1994)).

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 parameters in the context of single agent dynamic programming model. 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 (cf., Bajari, Benkard, and Levin, 2005; Pakes, Ostrovsky, and Berry, 2005; Pesendorfer and Schmidt-Dengler, 2006; 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. One of the important econometric issues in this literature is to develop an estimation method that is computational simple and has good finite sample properties.

This paper studies a sequential estimation procedure obtained by reformulating (1) in terms of a sequence of semiparametric extremum estimators when an initial consistent estimator $\hat{P}_0$ is available:

Step 1: Given $\hat{P}_{j-1}$, update θ by

$$\hat{\theta}_j = \arg\max_{\theta \in \Theta} Q_n(\Psi(\hat{P}_{j-1}, \theta)). \tag{2}$$

Step 2: Update $\hat{P}_{j-1}$ using the obtained estimate $\hat{\theta}_j$:

$$\hat{P}_j = \Psi(\hat{P}_{j-1}, \hat{\theta}_j). \tag{3}$$

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

This algorithm is first proposed by Aguirregabiria and Mira (2002, AM02 hereafter) as a recursive extension of the two-step method developed by Hotz and Miller (1993) in the context of single agent dynamic programming model. Using the (pseudo-)likelihood function in the objective function, their algorithm is called the nested pseudo likelihood (NPL) algorithm. Aguirregabiria and Mira (2007, AM07 hereafter) apply the similar idea in the context of dynamic discrete games. Their analysis shows that 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 AM07 have obtained convergence in their simulations and illustrate that the NPL estimator performs very well relative to the two-step estimator, they neither provide the conditions under which the NPL algorithm converges nor analyze how fast the convergence occurs. On the other hand, the simulation results of Pesendorfer and Schmidt-Dengler (2006) provide some evidence that the NPL algorithm may not necessarily converge. To date, we do not know under which circumstances the NPL algorithm is applicable to obtain the more efficient estimator than the two-step estimators.

This paper analyzes the conditions under which the sequential algorithm of iterating (2) and (3) achieves convergence and derives the convergence rate of a sequence of estimators generated by the sequential algorithm. In the contest of single agent dynamic programming model, Kasahara and Shimotsu (2006, KS06 hereafter) derive the rate at which the sequence of the estimators generated from the NPL algorithm approaches the MLE. We extend the results of KS06 to a general class of structural models that are formulated as fixed point problem, including a model of dynamic games.

There are, however, important differences between a single agent model and a model of dynamic games. As AM02 and AM07 show, the NPL algorithm achieves the MLE in a single agent dynamic model while the limit of the sequences generated by the NPL algorithm in a model of dynamic games is asymptotically *less* efficient than the MLE. This difference reflects the fact that the shape of the pseudo-likelihood function is very similar to that of the true likelihood function in a single agent model while it is less so in a model of dynamic games. Furthermore, as KS06 show, one of the key properties that assures the fast convergence rate of the NPL algorithm in a single agent dynamic model is the (asymptotic) orthogonality between the parameter θ and the nuisance parameter P in the pseudo-likelihood function but this orthogonality condition is often violated in a model of dynamic games.

The key to understanding the convergence properties of the NPL algorithm, or more generally the sequential estimators of iterating (2) and (3), is a *contraction* property of the operator Ψ defining the fixed point problem. Intuitively, the faster the operator achieves contraction, the closer the the value obtained after one iteration is to the fixed point, and therefore a curvature of the pesudo-likelihood function gets closer to that of the true likelihood function. Moreover, the higher contraction rate of the operator Ψ implies that the initial value of P has less influence on the value obtained after one contraction and, thus, the degree of orthogonality between θ and P in the pseudo-likelihood function is higher.

Our main result is that the convergence of the NPL algorithm is achieved if the largest eigenvalue of the Jacobian matrix $\partial\Psi(P,\theta)/\partial P$ evaluated at the fixed point P_θ is less than one in absolute value. This is because the local contraction property of the operator Ψ is determined by the eigenvalues of the derivative of Ψ with respect to P. The closer the largest eigenvalue of $\partial\Psi(P_\theta,\theta)/\partial P$ to zero, the faster the convergence rate of the NPL algorithm. When the operator Ψ has the "zero Jacobian property" that $\partial\Psi(P_\theta,\theta)/\partial P = 0$, it is possible to achieve a superlinear convergence rate as shown in KS06.

We show that the similar results hold in the context of the sequential algorithm based on GMM. The convergence of the sequential GMM algorithm also requires that all the eigenvalues of $\partial\Psi(P_\theta,\theta)/\partial P$ are less than one in absolute value. The limit of the sequential GMM estimators may be more efficient than two-step estimator and it can be asymptotically equivalent to the efficient GMM estimator that is based on a full solution of fixed point problem.

The reminder of the paper is organized as follows. Section 2 presents preliminary analysis on contraction properties of operator Ψ. Section 3 shows the results on the sequential NPL algorithm. Section 4 extends our analysis to the sequential GMM estimator. Section 5 discusses the sequential CMD estimator.

2 Preliminary

This section provides various analytical results on the (local) contraction properties of operator Ψ. These results are used to analyze the convergence properties of the sequential extremum estimators in later sections.

2.1 Finite-dimensional Case

Consider the case that Z_i takes a L possible values, $Z_i \in \mathcal{Z} = \{z_1, z_2, ..., z_L\}$, so that P is a finite L-dimensional vector. Expanding $\Psi(P, \theta)$ around the fixed point, P_θ, gives

$$\Psi(P,\theta) - P_\theta = \frac{\partial \Psi(P,\theta)}{\partial P'}|_{P=P_\theta}(P - P_\theta) + O(||P - P_\theta||^2).$$

The first-order convergence property of $\Psi(P,\theta)$ depends on the matrix $\frac{\partial \Psi(P,\theta)}{\partial P'}|_{P=P_\theta}$. When $\frac{\partial \Psi(P_\theta,\theta)}{\partial P'} = 0$, the operator achieves a quadratic contraction. When $\frac{\partial \Psi(P,\theta)}{\partial P'}|_{P=P_\theta} \neq 0$, the contraction property of $\Psi(\cdot,\theta)$ is determined by the largest eigenvalue of the matrix $\frac{\partial \Psi(P,\theta)}{\partial P'}|_{P=P_\theta}$. Denote the eigenvalues of $\frac{\partial \Psi(P_\theta,\theta)}{\partial P'}$ by $\lambda_j(\theta)$ for $j = 1, 2, ..., L$. Let $\Lambda(\theta)$ be a diagonal matrix with $diag\{\Lambda(\theta)\} = (\lambda_1(\theta), ..., \lambda_L(\theta))'$. Let $M(\theta)$ be the modal matrix of $\frac{\partial \Psi(P_\theta,\theta)}{\partial P'}$, of which j-th column is the eigenvector corresponds to the eigenvalue $\lambda_j(\theta)$. Then we have $\frac{\partial \Psi(P_\theta,\theta)}{\partial P'} = M(\theta)\Lambda(\theta)M(\theta)^{-1}$. If we define a q-stage operator of Ψ by

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

then we have $\frac{\partial \Psi^q(P_\theta,\theta)}{\partial P'} = \left(\frac{\partial \Psi(P_\theta,\theta)}{\partial P'}\right)^q = M(\theta)\Lambda^q(\theta)M(\theta)^{-1}$ and it follows that

$$\Psi^q(P,\theta) - P_\theta \quad = \quad M(\theta)\Lambda^q(\theta)M(\theta)^{-1}(P - P_\theta) + O(||P - P_\theta||^2).$$

When P is in neighborhood of P_θ, the higher order terms are negligible and the local convergence property of a q-stage operator is given by:

$$||M(\theta)^{-1}(\Psi^q(P,\theta) - P_\theta)|| \leq (\lambda_{\max}(\theta))^q ||M(\theta)^{-1}(P - P_\theta)||,$$

where $\lambda_{\max}(\theta) = \max\{|\lambda_1(\theta)|, ..., |\lambda_L(\theta)|\}$ is the largest eigenvalue of $\Lambda(\theta)$ in absolute value. A contraction requires the value of $\lambda_{\max}(\theta)$ to be strictly less than one.

2.2 Infinite Dimensional Case

When Z_i is continuously distributed, P is infinite dimensional. Let the space of the probability distributions P belongs to be B_P. Given the value of θ, the operator $\Psi(\cdot,\theta)$ is a mapping from B_P to itself. The derivative of Ψ need to be defined as Fréchet (F-) derivatives. For a map $g : X \to Y$, where X and Y are Banach spaces, g is F-differentiable iff there exists a linear and continuous map T such that $g(x + h) - g(x) = Th + o(||h||)$ as $h \to 0$ for all h in neighborhood of zero. If it exists, T is called F-derivative of g at x. The operator norm is defined as $||g|| = \sup_{||x|| \leq 1} ||g(x)||$. Consequently, $||g(x)|| \leq ||g|| ||x||$ for all $x \in X$. See Zeidler (1986) for further details.

Proposition 1 *Suppose that an operator Ψ is a contraction with modulus λ such that, for any $\theta \in \Theta$ and $P \in B_P$: $||P_\theta - \Psi(P, \theta)|| \leq \lambda ||P_\theta - P||$. Then,*

$$\begin{aligned} ||D_P \Psi(P_\theta, \theta) h|| &\leq \lambda, \\ ||D_\theta P_\theta k - D_\theta \Psi(P, \theta)|_{P=P_\theta} k|| &\leq \lambda ||D_\theta P_\theta k|| \end{aligned}$$

for all $h \in B_P$ and $k \in \Theta$.

One important implication is that

$$\begin{aligned} ||D_P \Psi^q(P_\theta, \theta) h|| &\leq \lambda^q, \\ ||D_\theta P_\theta k - D_\theta \Psi^q(P, \theta)|_{P=P_\theta} k|| &\leq \lambda^q ||D_\theta P_\theta k||, \end{aligned}$$

for all $h \in B_P$ and $k \in \Theta$, since $\Psi^q(\cdot, \theta)$ is a contraction with modulus λ^q if $\Psi(\cdot, \theta)$ is a contraction with modulus λ.

2.3 Superlinear Contraction

In some cases, the operator Ψ has a superlinear contraction property. The following proposition states that a super linear contraction property implies "zero Jacobian property" of pseudo-likelihood function.

Proposition 2 *Suppose that there exists some $\epsilon > 0$ such that*

$$P_\theta - \Psi(P, \theta) = O(||P_\theta - P||^{1+\epsilon}),$$

for $\theta \in \Theta$ and P in neighborhood of P_θ. Then,

$$\begin{aligned} D_\theta \Psi(P, \theta)|_{P=P_\theta} &= D_\theta P_\theta, \\ D_P \Psi(P_\theta, \theta) &= 0. \end{aligned}$$

Furthermore,

$$D_{P\theta} \Psi^2(P_\theta, \theta) = 0.$$

As AM02 discuss, the condition that $D_P \Psi(P_\theta, \theta) = 0$ implies that θ and P are asymptotically orthogonal in the pseudo likelihood function. The condition that $D_{P\theta} \Psi^2(P_\theta, \theta) = 0$ implies that θ and P are orthogonal in any sample size. As shown in KS06, the orthogonality between θ and P is the key to understanding the convergence property of the NPL algorithm.

3 Maximum Likelihood Estimator

The maximum likelihood estimator solves the following constrained maximization problem:

$$\max_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\ln P(a_i|x_i) \qquad s.t. \quad P=\Psi(P,\theta).$$

Denote the fixed point $P_\theta=\Psi(P_\theta,\theta)$. Let the maximum likelihood estimator be

$$\hat{\theta}=\arg\max_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\ln P_\theta(a_i|x_i),$$

while let the two-step maximum pseudo-likelihood estimator be

$$\tilde{\theta}_1^q=\arg\max_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\ln\Psi^q(\hat{P}_0,\theta)(a_i|x_i).$$

Let's collect notations first.

$$\begin{aligned}
\psi^q(P,\theta)&=\ln\Psi^q(P,\theta), & \overline{\psi}^q(P,\theta)&=\frac{1}{n}\sum_{i=1}^{n}\ln\Psi^q(P,\theta)(a_i|x_i),\\
\Psi_\theta&=(\partial/\partial\theta')\Psi(P^0,\theta^0), & \Psi_P&=(\partial/\partial P')\Psi(P^0,\theta^0),\\
\overline{\psi}_\theta^q(P,\theta)&=\frac{1}{n}\sum_{i=1}^{n}(\partial/\partial\theta)\ln\Psi^q(P,\theta)(a_i|x_i), & \overline{\psi}_P^q(P,\theta)&=\frac{1}{n}\sum_{i=1}^{n}(\partial/\partial P)\ln\Psi^q(P,\theta)(a_i|x_i),\\
\overline{\psi}_{\theta P}^q(P,\theta)&=\frac{1}{n}\sum_{i=1}^{n}(\partial^2/\partial\theta\partial P')\ln\Psi^q(P,\theta)(a_i|x_i), & \overline{\psi}_{\theta\theta}^q(P,\theta)&=\frac{1}{n}\sum_{i=1}^{n}(\partial^2/\partial\theta\partial\theta')\ln\Psi^q(P,\theta)(a_i|x_i),
\end{aligned}$$

and

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

Note that the information matrix equality holds for $\ln\Psi^q(P,\theta)$, too, because $\Psi^q(P^0,\theta^0)(a|x)$ is also the true density of the data. In the case of $q=1$, we simply denote $\psi(P,\theta)=\psi^1(P,\theta)$, $\overline{\psi}(P,\theta)=\overline{\psi}^1(P,\theta)$, etc.. We focus on the case where the support of (a_i,x_i) is finite, $A\times X=\{a^1,a^2,...,a^{|A|}\}\times\{x^1,x^2,\ldots,x^{|X|}\}$.

In AM07, $\nabla_P\Psi$ denotes $(\partial/\partial P')\Psi(P^0,\theta^0)$, which corresponds to our Ψ_P. Similarly, $\nabla_\theta\Psi$ in AM07 corresponds to our Ψ_θ. Define $f_x(x_l)=\Pr(x=x^l)$ and let f_x be a $|A||X|\times 1$

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 $\Omega^q_{\theta\theta} = \Psi^{q'}_\theta \Delta_P \Psi^q_\theta$ and $\Omega^q_{\theta P} = \Psi^{q'}_\theta \Delta_P \Psi^q_P$.

3.1 Two-step estimator

Let P^0 be the true set of probabilities. Consider the following regularity conditions.

Assumption 1. (a) Θ is compact. (b) $\Psi^q(P,\theta)$ is three times continuously F-differentiable. (c) $\Psi^q(P,\theta)(a|x) > 0$ 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(Θ) such that, for any $(a,x) \in A \times X$, $P_{\theta^0}(a|x) = P^0(a|x)$. For any $\theta \neq \theta^0$, $\Pr_{\theta^0}(\{(a,x) : \Psi^q(P^0,\theta^0)(a|x) \neq P^0(a|x)\}) > 0$. (g) $E_{\theta^0} \sup_{(P,\theta)} ||D^s \Psi^q(P,\theta)(a|x)||^2 < \infty$ for $s = 1, \ldots, 4$.

Under Assumption 1, the two-step maximum pseudo-likelihood estimator is consistent and, when a root-n consistent estimator of P^0 is available, it is asymptotically normal.

Proposition 3 *Assume Assumption 1 holds and $\hat{P}_0 \rightarrow_p P^0$. Then $\tilde{\theta}^q_1 \rightarrow_p \theta^0$.*

Proposition 4 *Assume Assumption 1 holds and $\sqrt{n}(\hat{P}_0 - P^0) \rightarrow_d N(0,\Sigma)$. Then, $\sqrt{n}(\tilde{\theta}^q_1 - \theta^0) \rightarrow N(0, V^q)$, where $V^q = (\Omega^q_{\theta\theta})^{-1} + (\Omega^q_{\theta\theta})^{-1}\Omega^q_{\theta P}\Sigma(\Omega^q_{\theta P})'(\Omega^q_{\theta\theta})^{-1}$.*

Remark 1 *When $\Psi^q_P = 0$, the limiting distribution of the two-step estimator is the same as that of the MLE even under the weaker assumption that $\hat{P}_0 - P^0 = O_p(n^{-b})$ with $b > 1/4$. See Proposition 4 of AM02 and Proposition 2 of KS06.*

3.2 Sequential Pseudo Maximum Likelihood Estimator

In this section, we analyze the asymptotic properties of the sequential pseudo maximum likelihood estimator that is defined as follows: assuming that an initial consistent estimator $\tilde{P}_0$ is available,

Step 1: Given $\tilde{P}_{j-1}$, update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta} \overline{\psi}(\tilde{P}_{j-1}, \theta)$.

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 $j = k$.

AM02 and AM07 propose the nested pseudo likelihood (NPL) estimator $\tilde{\theta}$ that is defined by the following properties:

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

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

Proposition 5 *Assume Assumption 1 holds. Then,* $\sqrt{n}(\tilde{\theta}-\theta^0) \to N(0, V_{NPL})$*, where* $V_{NPL} = [\Omega_{\theta\theta}+\Omega_{\theta P}(I-\Psi_P)^{-1}\Psi_\theta]^{-1}\Omega_{\theta\theta}\{[\Omega_{\theta\theta}+\Omega_{\theta P}(I-\Psi_P)^{-1}\Psi_\theta]^{-1}\}'$*. Furthermore, if all the eigenvalues of* Ψ_P *are less than one in absolute value, then* $V^1 - V_{NPL}$ *is positive definite.*

As noted by AM07, the NPL estimator can be obtained as a limit of iterating steps 1 and 2 if the iterations converge. In their simulation study, AM07 report that the iterations always converged. However, the convergence property of this algorithm has not been fully understood.

We show that its convergence property crucially depends on the eigenvalues of Ψ_P. When all the eigenvalues of Ψ_P are smaller than 1 in absolute value, iterating steps 1 and 2 converges.

First, we state the regularity conditions.

Assumption B Assumption 1 holds, and in addition

$$\begin{aligned}
\overline{\psi}_{\theta P}(P^0,\theta^0) &= -\Omega_{\theta P} + O_p(n^{-1/2}), \\
\overline{\psi}_{\theta\theta}(P^0,\theta^0) &= -\Omega_{\theta\theta} + O_p(n^{-1/2}), \\
\overline{\psi}_{\theta}(P^0,\theta^0) &= O_p(n^{-1/2}), \\
E\sup_{\theta,P}||D_{\theta P}\ln\Psi(P,\theta)|| &< \infty, \quad E\sup_{\theta,P}||D^3\ln\Psi(P,\theta)|| < \infty, \\
\sup_{\theta,P}||D^2\Psi(P,\theta)|| &= O(1),
\end{aligned}$$

$\overline{\psi}_{\theta\theta}(P,\theta)$ is invertible for all (P,θ).

All the assumptions but the last two are fairly weak. $\overline{\psi}_{\theta\theta}(P,\theta)$ should be invertible in many cases because $\overline{\psi}_{\theta\theta}(P,\theta)$ is an average of n matrices. If we assume $\tilde{P}_0$ is consistent, then the last assumption can be replaced by the invertibility of $\Omega_{\theta\theta}$. The following lemma shows the bound of $\tilde{\theta}_j - \tilde{\theta}$ and $\tilde{P}_j - \tilde{P}$.

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

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

Remark 2 $\Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P$ *is a generalized least squares projection matrix from a regression of an element of* B_P *onto the space spanned by* Ψ_θ*, where the "error variance matrix" is* Δ_P^{-1}.

Remark 3 *The eigenvalues of* $I - \Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P$ *are either zero or one. Hence, the convergence property of* $\tilde{P}_j$ *is determined by the eigenvalues of* Ψ_P*. If all the eigenvalues of* Ψ_P

are smaller than 1 in absolute value, an iteration moves $\tilde{P}_j$ toward $\tilde{P}$. It follows from induction that

$$\tilde{P}_k - \tilde{P} = \left\{ \left[I - \Psi_\theta (\Psi_\theta' \Delta_P \Psi_\theta)^{-1} \Psi_\theta' \Delta_P \right] \Psi_P \right\}^k (\tilde{P}_0 - \tilde{P}) + O(\Psi_P^{k-1})(O_p(n^{-1/2}||\tilde{P}_0 - \tilde{P}||) + O_p(||\tilde{P}_0 - \tilde{P}||^2)),$$

and $\tilde{P}_k, \tilde{\theta}_k$ converges to $\tilde{P}, \tilde{\theta}$ as $k \to \infty$. On the other hand, if some eigenvalues of Ψ_P are larger than 1, then an iteration moves some elements of $\tilde{P}_j$ further away from $\tilde{P}$. In this case, it is not clear whether the iterations eventually converge.

Remark 4 *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 Ψ_P are smaller than 1 in absolute value.*

Remark 5 *If all the eigenvalues of Ψ_P are smaller than 1 in absolute value and we choose $k \to \infty$ so that $\log n = o(k)$, then $\tilde{P}_k - \tilde{P} = o_p(n^{-1/2})$ and the effect of $\tilde{P}_0$ on $\tilde{P}$ 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 6 *When the operator Ψ has a superlinear contraction property, we have $\Psi_P = 0$ (c.f., Proposition 2). In such a case, the convergence rate is faster than linear:*

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

Remark 7 *If at least one element of x_i is continuously distributed, one can prove the higher-order improvement by bootstrap as in KS06.*

We can also construct a q–version of the sequential estimator.

Step 1: Given $\tilde{P}^q_{j-1}$, update θ by $\tilde{\theta}^q_j = \arg\max_{\theta \in \Theta} \overline{\psi}^q(\tilde{P}^q_{j-1}, \theta)$.

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

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

Note that the derivative of the q–stage mapping $\Psi^q(P_\theta, \theta)$ take the following form:

$$\begin{aligned} (\partial/\partial\theta')\Psi^q(P_\theta, \theta) &= (I + \Psi_P + \cdots \Psi_P^{q-1})\Psi_\theta = (I - \Psi_P^q)(1 - \Psi_P)^{-1}\Psi_\theta, \\ (\partial/\partial P')\Psi^q(P_\theta, \theta) &= \Psi_P^q. \end{aligned}$$

If the eigenvalues of Ψ_P are positive, iterating this mapping increases the curvature of the likelihood function. If they are all between 0 and 1, then $(\partial/\partial\theta')\Psi^q(P, \theta)$ approaches to $(\partial/\partial\theta')P_\theta$.

Corollary 1 *Suppose Assumption B holds with $\Psi^q(P,\theta)$ replacing $\Psi(P,\theta)$. Then, for $j = 1,\ldots,k$,*

$$
\begin{aligned}
\tilde{\theta}^q_j - \tilde{\theta}^q &= O_p(||\tilde{P}^q_{j-1} - \tilde{P}^q||), \\
\tilde{P}^q_j - \tilde{P}^q &= \left[I - \Psi^q_\theta((\Psi^q_\theta)'\Delta_P\Psi^q_\theta)^{-1}(\Psi^q_\theta)'\Delta_P\right]\Psi^q_P(\tilde{P}^q_{j-1} - \tilde{P}^q) \\
&\quad + O_p(n^{-1/2}||\tilde{P}^q_{j-1} - \tilde{P}^q||) + O_p(||\tilde{P}^q_{j-1} - \tilde{P}^q||^2).
\end{aligned}
$$

Note that using Ψ^q in place of Ψ accerelates the convergence of the sequential estimator when all the eigenvalues of Ψ_P are smaller than one in absolute value.

3.3 Non-optimality of the NPL estimator

Define $\psi(P,\theta) = \log\Psi(P,\theta)$. The model we consider implies the following conditional moment restriction

$$
E\left[\nabla_\theta\psi(P^0,\theta^0)(a|x)\middle| x\right] = 0. \tag{4}
$$

Let $A(x)$ be a $k \times k$ nonsingular matrix of functions of x, and consider the following GMM estimator

$$
\tilde{\theta} = \arg\min_\theta \left\| \frac{1}{n}\sum_{i=1}^n A(x_i)\nabla_\theta\psi(\tilde{P},\theta)(a_i|x_i) \right\|^2 \quad s.t. \quad \tilde{P} = \Psi(\tilde{P},\tilde{\theta}).
$$

This estimator resembles the NPL estimator in that $\tilde{P}$ is set to satisfy the NPL fixed point condition. The efficient GMM estimator is obtained by choosing $A(x)$ to minimize the variance of $\tilde{\theta}$. As shown in the following proposition, the limiting variance of the efficient GMM estimator is different from that of the NPL estimator, and the NPL estimator may not be optimal within a class of the estimators that are based on the conditional moment restriction (4) and impose the NPL fixed point condition.

Proposition 6 *The optimal instrument is*

$$
\begin{aligned}
\bar{A}(x) &= E\left[\nabla_{\theta\theta}\psi(P^0,\theta^0)(a|x) + \nabla_{\theta P}\psi(P^0,\theta^0)(a|x)(I-\Psi_P)^{-1}\Psi_\theta|x\right]' \\
&\quad \times E\left[\nabla_\theta\psi(P^0,\theta^0)(a|x)\nabla_\theta\psi(P^0,\theta^0)(a|x)'|x\right]^{-1}
\end{aligned}
$$

and the limiting variance of the efficient GMM estimator is

$$
\bar{V} = \left\{E\left[\bar{A}(x)E\left[\nabla_\theta\psi(P^0,\theta^0)(a|x)\nabla_\theta\psi(P^0,\theta^0)(a|x)'|x\right]^{-1}\bar{A}(x)'\right]\right\}^{-1},
$$

which is different from the limiting variance of the NPL estimator

$$
V_{NPL} = \left\{E[\bar{A}(x)]E[\nabla_\theta\psi(P^0,\theta^0)(a|x)\nabla_\theta\psi(P^0,\theta^0)(a|x)']E[\bar{A}(x)]'\right\}^{-1}
$$

if $\Psi_P \neq 0$. When $\Psi_P = 0$, both $\bar{V}$ and V_{NPL} are equal to the asymptotic variance of the maximum likelihood estimator.

4 Generalized Method of Moments Estimator

The generalized method of moments estimator is defined by

$$\hat{\theta}_{GMM} = \arg\max_{\theta\in\Theta} \ -\bar{g}(P)'\hat{W}\bar{g}(P) \qquad s.t. \quad P = \Psi(P,\theta),$$

where $\hat{W} \to_p W$ positive semi-definite and

$$\bar{g}(P) = n^{-1}\sum_{i=1}^{n} g(a_i, x_i; P),$$

with $g(\cdot;P) = (g_1(\cdot;P), g_2(\cdot;P), ..., g_L(\cdot;P))'$ is a moment vector function representing L moment conditions. Specifically, we consider

$$g_l(a_i, x_i; P) = \rho_l(x_i)\left\{h_l(a_i) - \sum_{a\in A} h_l(a)P(a|x_i)\right\},$$

which satisfies $E[g_l(a_i, x_i; P^0)] = E[\rho_l(x_i)E[h_l(a_i) - E(h_l(a)|x_i)|x_i]] = 0$ for $l = 1, 2, ..., L$.

Let's collect notations first.

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

Define f_x as before so that its elements are arranged comformably 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 $\gamma_l(a,x) = \rho(x)h(a)$ and γ_l represent a vector of $|A||X|$ length. Finally, let $\Gamma = (\gamma_1', \gamma_2', ..., \gamma_L')'$ be a L by $|A||X|$ matrix.

With those notations, we may write $\bar{G}_\theta(\Psi^q(P,\theta)) = -\Gamma\hat{\Delta}_x(\partial/\partial\theta')\Psi^q(P,\theta)$, $\bar{G}_P(\Psi^q(P,\theta)) = -\Gamma\hat{\Delta}_x(\partial/\partial P')\Psi^q(P,\theta)$, $G_\theta^q = -\Gamma\Delta_x(I-\Psi_P^q)(I-\Psi_P)^{-1}\Psi_\theta$ and $G_P^q = -\Gamma\Delta_x\Psi_P^q$. In the case of $q = 1$, we write $G_\theta = G_\theta^1$, $G_P = G_P^1$, etc..

4.1 Two-step Estimator

Given an initial consistent estimator $\hat{P}_0$, a two-step GMM estimator based on the operator Ψ^q is defined by

$$\tilde{\theta}^q \quad = \quad \arg\max_{\theta\in\Theta} \ -\bar{g}(\Psi^q(\hat{P}_0,\theta))'\hat{W}\bar{g}(\Psi^q(\hat{P}_0,\theta)).$$

Let $r(a_i, x_i)$ be a vector of length $|A||X|$ whose elements are arranged comformably 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 $\hat{P}_0 = n^{-1} \sum_{i=1}^{n} r(a_i, x_i)$.

The asymptotic distribution of $\tilde{\theta}$ is given by

$$\sqrt{n}(\tilde{\theta} - \theta^0) \to_d N(0, V),$$

where

$$V = (G_\theta' W G_\theta)^{-1} G_\theta' W S W' G_\theta (G_\theta' W' G_\theta)^{-1}$$

with $S = E[(g(a_i, x_i; P^0) + G_P(r(a_i, x_i) - P^0))(g(a_i, x_i; P^0) + G_P(r(a_i, x_i) - P^0))']$. Using an optimal weighting matrix $W^* = S^{-1}$, the limiting variance is given by $V = (G_\theta' S^{-1} G_\theta)^{-1}$.

If we use $\Psi_q(P, \theta)$ in place of $\Psi(P, \theta)$, we have $\sqrt{n}(\tilde{\theta}^q - \theta^0) \to_d N(0, V^q)$ where

$$V^q = ((G_\theta^q)' W G_\theta^q)^{-1} (G_\theta^q)' W S^q W' G_\theta^q ((G_\theta^q)' W' G_\theta^q)^{-1}$$

with $S^q = E[(g(a_i, x_i; P^0) + G_P^q(r(a_i, x_i) - P^0))(g(a_i, x_i; P^0) + G_P^q(r(a_i, x_i) - P^0))']$.

When all the eigenvalues of Ψ_P are between 0 and 1, the GMM estimator $\hat{\theta}_{GMM}$ is obtained as the limit of $\tilde{\theta}^q$ as $q \to \infty$ and, therefore, the limiting variance of $\hat{\theta}_{GMM}$ is given by $V^\infty = ((G_\theta^\infty)' W G_\theta^\infty)^{-1} (G_\theta^\infty)' W \Omega W' G_\theta^\infty ((G_\theta^\infty)' W' G_\theta^\infty)^{-1}$. Using a weighting matrix $W = \Omega^{-1}$ leads to the efficient GMM estimator with the asymptotic variance $((G_\theta^\infty)' \Omega^{-1} G_\theta^\infty)^{-1}$.

4.2 Sequential GMM Estimator

Given an initial estimator $\tilde{P}_0$,

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

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

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

We first consider an nested generalized method of moments estimator (NGMM) which satisfies

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

Assumption C Assumption 1 holds, and in addition

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

Note that $\sup_{\theta,P} ||D^2\Psi(P,\theta)|| < \infty$ and $||\Gamma|| < \infty$ 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. The following lemma provides the limiting distribution of the NGMM estimator.

Lemma 2 *Suppose Assumption C holds. Then*

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

$\Omega = E[g(a_i, x_i; P^0)g(a_i, x_i; P^0)']$. *If* $W^* = (G_\theta^{-1})'(G_\theta^\infty)'\Omega^{-1}$ *is positive semi-definite, the minimized asymptotic variance is* $((G_\theta^\infty)'\Omega^{-1}G_\theta^\infty)^{-1}$ *by setting* $W = W^*$.

Remark 8 *When all the eigenvalues of* Ψ_P *are less than one in absolute value, the asymptotic variance of the efficient GMM estimator is also* $((G_\theta^\infty)'\Omega^{-1}G_\theta^\infty)^{-1}$. *Thus, the NGMM estimator is more efficient than two-step estimator and can be asymptotically equivalent to the efficient GMM estimator.*

Remark 9 *When* $\Psi_P = 0$, *the two-step GMM estimator with an optimal weighting matrix is also asymptotically equivalent to the efficient GMM estimator.*

The NGMM estimator can be obtained as the limit of the sequential GMM estimators upon convergence. The convergence property of sequential GMM estimators is given by the following lemma.

Lemma 3 *Suppose Assumption C holds. Then, for* $j = 1, \ldots, k$,

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

Remark 10 *Observe that* $-\Psi_\theta(G_\theta'\hat{W}G_\theta)^{-1}G_\theta'\hat{W}\Gamma\Delta_x = \Psi_\theta(\Psi_\theta'\Delta_x'\Gamma'\hat{W}\Gamma\Delta_x\Psi_\theta)^{-1}\Psi_\theta'\Delta_x'\Gamma'\hat{W}\Gamma\Delta_x$ *is a projection matrix, and the sequential GMM estimator has the same convergence property as the sequential ML estimator. Again, the eigenvalues of* Ψ_P *determine the convergence.*

Remark 11 *Analogous remarks to Remarks 3-6 apply here.*

A q-stage version of the sequential GMM estimator is described as follows.

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

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

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

Let $(\tilde{P}^q, \tilde{\theta}^q) = \lim_{k\to\infty}(\tilde{P}^q_k, \tilde{\theta}^q_k)$. We may show that

$$\sqrt{n}(\tilde{\theta}^q - \theta^0) \to_d N(0, ((G^q_\theta)'WG^\infty_\theta)^{-1}(G^q_\theta)'W\Omega W'G^q_\theta((G^\infty_\theta)'W'G^q_\theta)^{-1}),$$

where $\Omega = E[g(a_i, x_i; P^0)g(a_i, x_i; P^0)']$. Choosing $W^* = ((G^q_\theta)^{-1})'(G^\infty_\theta)'\Omega^{-1}$ as a weighting matrix (if it is positive semi-definite) leads to the asymptotic variance $((G^\infty_\theta)'\Omega^{-1}G^\infty_\theta)^{-1}$.

Corollary 2 *Suppose Assumption C holds with $\Psi^q(P,\theta)$ replacing $\Psi(P,\theta)$. Then, for $j = 1, \ldots, k$,*

$$\begin{aligned}
\tilde{\theta}^q_j - \tilde{\theta}^q &= O_p(||\tilde{P}^q_{j-1} - \tilde{P}^q||),\\
\tilde{P}^q_j - \tilde{P}^q &= [I + \Psi^q_\theta((G^q_\theta)'\hat{W}G^q_\theta)^{-1}(G^q_\theta)'\hat{W}\Gamma\Delta_x]\Psi^q_P(\tilde{P}^q_{j-1} - \tilde{P}^q)\\
&\quad + O_p(n^{-1/2}||\tilde{P}^q_{j-1} - \tilde{P}^q||) + O_p(||\tilde{P}^q_{j-1} - \tilde{P}^q||^2).
\end{aligned}$$

5 Classical Minimum Distance Estimator (or Asymptotic Least Squares Estimator)

Let $\hat{P}_0$ be a frequency estimator of P^0 satisfying $n^{1/2}(\hat{P}_0 - P^0) \to_d N(0, \Sigma)$. For parameters P and θ, define the difference between the empirical frequency and the value after q iterations of the operator as

$$\bar{g}^q(\theta, P) = \hat{P}_0 - \Psi^q(P, \theta).$$

Substitute $\hat{P}_0$ into $\bar{g}^q(\theta, P)$, and consider estimating θ by

$$\tilde{\theta}^q = \arg\min_\theta \bar{g}^q(\theta, \hat{P}_0)'\hat{W}\bar{g}^q(\theta, \hat{P}_0),$$

where $\hat{W} \to_p W$ is a positive definite weighting matrix. We call $\tilde{\theta}^q$ the q–classical minimum distance (q–CMD) estimator. The asymptotic least square estimator of Pesendorfer and Schmidt-Dengler (2006) corresponds to the case $q = 1$. Pesendorfer and Schmidt-Dengler (2006, Propositions 4 and 5) show that the efficient least squares estimator, $\tilde{\theta}_{LS}$, that uses $W = (I - \Psi'_P)^{-1}\Sigma^{-1}(I - \Psi_P)^{-1}$ satisfies

$$\sqrt{n}(\tilde{\theta}_{LS} - \theta^0) \to_d N(0, V_{LS}), \quad V_{LS} = (\Psi'_\theta(I - \Psi'_P)^{-1}\Sigma^{-1}(I - \Psi_P)^{-1}\Psi_\theta)^{-1}.$$

We use the following assumptions:

Assumption D Assumption 1 holds, and in addition

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

Note that $\sup_{\theta,P}||D^2\Psi^q(P,\theta)|| < \infty$ implies that $\sup_{\theta,P}||D_{\theta\theta}\bar{g}^q(\theta,P)|| < \infty$. Let $\Psi_\theta^{(q)}$ and $\Psi_P^{(q)}$ denote $D_\theta\Psi^q(P^0,\theta^0)$ and $D_P\Psi^q(P^0,\theta^0)$, respectively.

The following lemma shows the asymptotic distribution of the q–CMD estimator. It turns out, when the optimal weighting matrix is used, $\tilde{\theta}^q$ has the same efficiency as $\tilde{\theta}_{LS}$ and there is no efficiency improvement from using Ψ^q.

Lemma 4 *Suppose Assumption D holds. Then*

$$\sqrt{n}(\tilde{\theta}^q-\theta^0) \to_d N(0,(\Psi_\theta^{(q)\prime}W\Psi_\theta^{(q)})^{-1}(\Psi_\theta^{(q)\prime}W(I-\Psi_P^{(q)})\Sigma(I-\Psi_P^{(q)\prime})W'\Psi_\theta^{(q)})(\Psi_\theta^{(q)\prime}(I-\Psi_P^{(q)\prime})^{-1}W'\Psi_\theta^{(q)})^{-1}).$$

The asymptotic variance of $\tilde{\theta}^q$ is minimized by choosing $W = (I-\Psi_P^{(q)\prime})^{-1}\Sigma^{-1}(I-\Psi_P^{(q)})^{-1}$. With this choice of W,

$$\sqrt{n}(\tilde{\theta}^q-\theta^0) \to N(0,(\Psi_\theta'(I-\Psi_P')^{-1}\Sigma^{-1}(I-\Psi_P)^{-1}\Psi_\theta)^{-1})),$$

and $\tilde{\theta}^q$ has the same efficiency as $\tilde{\theta}_{LS}$.

An efficiency gain is achieved if one uses Ψ^q with $q \geq 2$ while imposing an NPL fixed point restriction on P. Consider the following nested minimum distance (NMD) estimator.

$$\tilde{\theta}_{NMD}^q = \arg\min_\theta \bar{g}^q(\theta,\tilde{P})'\hat{W}\bar{g}^q(\theta,\tilde{P}) \quad s.t. \quad \tilde{P} = \Psi(\tilde{P},\tilde{\theta}).$$

The following lemma shows the asymptotic distribution of the NMD estimator.

Lemma 5 *Suppose Assumption D holds. Then*

$$\sqrt{n}(\tilde{\theta}_{NMD}^q-\theta^0) \to_d N(0,(\Psi_\theta^{(q)\prime}W(I-\Psi_P)^{-1}\Psi_\theta^{(q)})^{-1}(\Psi_\theta^{(q)\prime}W\Sigma W'\Psi_\theta^{(q)})(\Psi_\theta^{(q)\prime}(I-\Psi_P')^{-1}W'\Psi_\theta^{(q)})^{-1}).$$

The limiting variance of $\tilde{\theta}_{NMD}^q$ is minimized by choosing $W^* \equiv (I-\Psi_P')^{-1}\Sigma^{-1}$ as the weighting matrix, and then

$$\sqrt{n}(\tilde{\theta}_{NMD}^q-\theta^0) \to_d N(0,V_{NMD}), \quad V_{NMD} = (\Psi_\theta^{(q)\prime}(I-\Psi_P')^{-1}\Sigma^{-1}(I-\Psi_P)^{-1}\Psi_\theta^{(q)})^{-1}.$$

When $q = 1$, this limiting variance is identical to that of $\tilde{\theta}_{LS}$, and there is no efficiency gain from imposing an NPL fixed point constraint. When $q \geq 2$, the NMD estimator is more efficient than $\tilde{\theta}_{LS}$ if all the eigenvalues of Ψ_P are between 0 and 1. Recall $\Psi_\theta^{(q)} = (I+\Psi_P+\cdots+\Psi_P^{q-1})\Psi_\theta$. Let $\Omega = (I-\Psi_P')^{-1}\Sigma^{-1}(I-\Psi_P)^{-1}$, and write V_{NMD}^{-1} and V_{LS}^{-1} as

$$V_{LS}^{-1} = \Psi_\theta'\Omega\Psi_\theta, \quad V_{NMD}^{-1} = \Psi_\theta'(I+\Psi_P+\cdots+\Psi_P^{q-1})'\Omega(I+\Psi_P+\cdots+\Psi_P^{q-1})\Psi_\theta$$

Thus

$$\begin{aligned} V_{NMD}^{-1} - V_{LS}^{-1} &= \Psi_\theta'(\Psi_P + \cdots + \Psi_P^{q-1})'\Omega(\Psi_P + \cdots + \Psi_P^{q-1})\Psi_\theta \\ &\quad + \Psi_\theta'[(\Psi_P + \cdots + \Psi_P^{q-1})'\Omega + \Omega'(\Psi_P + \cdots + \Psi_P^{q-1})]\Psi_\theta, \end{aligned}$$

which is positive definite if all the eigenvalues of Ψ_P are between 0 and 1. Note, however, that the positive definiteness of W^* requires the positive definiteness of $(I - \Psi_P')^{-1}$. If this is not the case, choosing $W = W^*$ is not possible, and imposing an NPL fixed point constraint has a detrimental effect on the efficiency of the NMD estimator.

Taking $q \to \infty$ in $\tilde\theta^q$ gives the full-solution minimum distance estimator that minimizes the distance between $\hat P_0$ and P_θ. Interestingly, even using the full fixed point solution does not achieve an efficiency gain.

5.1 Sequential CMD estimator

We consider the sequential CMD estimator recursively as follows. Set $\tilde P_0 = \hat P_0$, and

Step 1: Given $\tilde P_{j-1}$, estimate θ by $\tilde\theta_j = \arg\min_\theta \bar g(\theta, \tilde P_{j-1})'\hat W \bar g(\theta, \tilde P_{j-1})$.

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 $j = k$.

Lemma 6 *Suppose Assumption D holds. Then, for* $j = 1, \ldots, k$,

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

We may also consider a q-stage version of the sequential CMD estimator as follows.

Step 1: Given $\tilde P_{j-1}$, estimate θ by $\tilde\theta_j^q = \arg\min_\theta \bar g^q(\theta, \tilde P_{j-1})'\hat W \bar g^q(\theta, \tilde P_{j-1})$.

Step 2: Update P using the obtained estimate $\tilde\theta_j^q$ by $\tilde P_j = \Psi^q(\tilde P_{j-1}, \tilde\theta_j)$.

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

Corollary 3 *Suppose Assumption D holds. Then, for* $j = 1, \ldots, k$,

$$\begin{aligned} \tilde\theta_j^q - \tilde\theta^q &= O_p(||\tilde P_{j-1} - \tilde P||), \\ \tilde P_j - \tilde P &= [I - \Psi_\theta^q((\Psi_\theta^q)'\hat W\Psi_\theta^q)^{-1}(\Psi_\theta^q)'\hat W]\Psi_P^q(\tilde P_{j-1} - \tilde P) + O_p(n^{-1/2}||\tilde P_{j-1} - \tilde P||) + O_p(||\tilde P_{j-1} - \tilde P||^2). \end{aligned}$$

6 Appendix

6.1 Proof of Proposition 1

For the first result, the definition of the F-derivative implies that, for any $h \in B_P$, $\Psi(P_\theta+h,\theta)-\Psi(P_\theta,\theta) = D_P\Psi(P_\theta,\theta)h + o(||h||)$ as $h \to 0$. Hence

$$||D_P\Psi(P_\theta,\theta)h|| \leq \lambda||h|| + o(||h||).$$

Dividing both sides by $||h||$ gives $||D_P\Psi(P_\theta,\theta)(h/||h||)|| \leq \lambda + o(||h||)/||h||$, and letting $h \to 0$ and using the definition of the operator norm gives the required result.[1] For the second result, observe that, for any $k \in \Theta$,

$$\begin{array}{rl} & ||P_{\theta+k} - P_\theta - \Psi(P_\theta,\theta+k) + \Psi(P_\theta,\theta)|| \\ \leq & ||\Psi(P_{\theta+k},\theta+k) - \Psi(P_\theta,\theta+k)|| \leq \lambda||P_{\theta+k} - P_\theta||. \end{array}$$

Therefore, $||D_\theta P_\theta k - D_\theta\Psi(P_\theta,\theta)k|| \leq \lambda||D_\theta P_\theta k|| + o(||k||)$ as $k \to 0$, and dividing both sides by $||k||$ and letting $k \to 0$ gives the stated result. □

6.2 Proof of Proposition 2

For any $h \in B_P$, $\Psi(P_\theta + h,\theta) - \Psi(P_\theta,\theta) = O(||h||^{1+\epsilon}) = o(||h||)$. Thus, $D_P\Psi(P_\theta,\theta) = 0$. For any $k \in \Theta$, $\Psi(P_\theta,\theta+k) - \Psi(P_\theta,\theta) = [P_{\theta+k} - O(||P_{\theta+k} - P_\theta||^{1+\epsilon})] - P_\theta = D_\theta P_\theta k + o(||k||)$. Thus, $D_\theta\Psi(P,\theta)|_{P=P_\theta} = D_\theta P_\theta$. For any $h \in B_P$

$$\begin{array}{rcl} D_{P\theta}\Psi(\Psi(P,\theta),\theta)h & = & D_{PP}\Psi(\Psi(P,\theta),\theta)D_P\Psi(P,\theta)h \cdot D_\theta\Psi(P,\theta) \\ & & +D_P\Psi(\Psi(P,\theta),\theta)D_{P\theta}\Psi(P,\theta)h + D_{P\theta}\Psi(\Psi(P,\theta),\theta)D_P\Psi(P,\theta)h. \end{array}$$

Therefore, if evaluated at the fixed point $P = P_\theta$, we have $D_{P\theta}\Psi^2(P_\theta,\theta) = 0$. □

6.3 Proof of Proposition 3

Assumption 1 (a), (b), and (d) with $\hat{P}_0 \to_p P^0$ imply that $\overline{\psi}^q(\hat{P}_0,\theta)$ converges uniformly in probability in θ to $E(\ln\Psi^q(P^0,\theta))$ (c.f., Lemma 24.1 of Gourieroux and Monfort, 1989). Then, the rest of the proof follows the proof of Theorem 2.1 of Newey and McFadden (1994). □

6.4 Proof of Propositions 4 and 5

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

[1] If $f(x) \leq a + o(1)$ as $x \to 0$, then $f(x) \leq a$. The proof is by contradiction (suppose $f(x) > a$, then ...).

6.5 Proof of Lemma 1 and Corollary 1

We prove Lemma 1 first. Recall that $\tilde{\theta}_j$ satisfies the first order condition

$$\overline{\psi}_{\theta}(\tilde{P}_{j-1}, \tilde{\theta}_j) = 0. \tag{5}$$

Expanding this around $(\tilde{P}, \tilde{\theta})$ and using $\overline{\psi}_{\theta}(\tilde{P}, \tilde{\theta}) = 0$ gives

$$0 = \overline{\psi}_{\theta P}(\bar{P}, \bar{\theta})(\tilde{P}_{j-1} - \tilde{P}) + \overline{\psi}_{\theta\theta}(\bar{P}, \bar{\theta})(\tilde{\theta}_j - \tilde{\theta}),$$

where $(\bar{P}, \bar{\theta})$ lie between $(\tilde{P}_{j-1}, \tilde{\theta}_j)$ and $(\tilde{P}, \tilde{\theta})$. Inverting $\overline{\psi}_{\theta\theta}(\bar{P}, \bar{\theta})$, we obtain

$$\tilde{\theta}_j - \tilde{\theta} = -\overline{\psi}_{\theta\theta}(\bar{P}, \bar{\theta})^{-1}\overline{\psi}_{\theta P}(\bar{P}, \bar{\theta})(\tilde{P}_{j-1} - \tilde{P}) = O_p(||\tilde{P}_{j-1} - \tilde{P}||), \tag{6}$$

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

For the second result, expand the second-step updating equation $\tilde{P}_j = \Psi(\tilde{P}_{j-1}, \tilde{\theta}_j)$ twice around $(\tilde{P}, \tilde{\theta})$ and use $\Psi(\tilde{P}, \tilde{\theta}) = \tilde{P}$, root-$n$ consistency of $(\tilde{P}, \tilde{\theta})$, and (6), then it follows that

$$\tilde{P}_j - \tilde{P} = \Psi_P(\tilde{P}_{j-1} - \tilde{P}) + \Psi_\theta(\tilde{\theta}_j - \tilde{\theta}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2). \tag{7}$$

Rewriting (6) by using $\overline{\psi}_{\theta P}(\tilde{P}, \tilde{\theta}) = -\Omega_{\theta P} + O_p(||\tilde{P}_{j-1} - \tilde{P}||) + O_p(n^{-1/2})$ and $\overline{\psi}_{\theta\theta}(\tilde{P}, \tilde{\theta}) = -\Omega_{\theta\theta} + O_p(||\tilde{P}_{j-1} - \tilde{P}||) + O_p(n^{-1/2})$, we obtain

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

Substituting this into (7) in conjunction with $\Omega_{\theta\theta}^{-1}\Omega_{\theta P} = (\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P\Psi_P$ gives

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

giving the stated result. For the corollary, first note that $(\partial/\partial\theta)\Psi^q(P^0, \theta^0)\Delta_P(\partial/\partial\theta')\Psi^q(P^0, \theta^0) = \Omega_{\theta\theta}^q$. Therefore, replacing $\Psi(P, \theta)$ with $\Psi^q(P, \theta)$ and repeating the argument proves the corollary. □

6.6 Proof of Proposition 6

The consistency of the efficient GMM estimator follows from analogous argument to the proof of the consistency of the NPL estimator. The condition associated with $\tilde{\theta}$ is

$$\frac{1}{n}\sum_{i=1}^{n} A(x_i)\nabla_\theta\psi(\tilde{P}, \tilde{\theta})(a_i|x_i) = 0.$$

[2] If we assume $\tilde{P}_{j-1}$ is consistent, the second equality follows from consistency of $\tilde{P}$ and $\tilde{P}_{j-1}$.

Expanding this around (P^0, θ^0) gives, with $(\bar{P}, \bar{\theta})$ lying between (P^0, θ^0) and $(\tilde{P}, \tilde{\theta})$,

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} A(x_i)\nabla_\theta \psi(P^0, \theta^0)(a_i|x_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} A(x_i)\nabla_{\theta\theta}\psi(\bar{P}, \bar{\theta})(a_i|x_i)(\tilde{\theta} - \theta^0) + \frac{1}{n}\sum_{i=1}^{n} A(x_i)\nabla_{\theta P}\psi(\bar{P}, \bar{\theta})(a_i|x_i)(\tilde{P} - P^0). \end{aligned}$$

The NPL fixed point restriction implies

$$\tilde{P} - P^0 = \Psi(\tilde{P}, \tilde{\theta}) - \Psi(P^0, \theta^0) = \nabla_P \Psi(\bar{P}, \bar{\theta})(\tilde{P} - P^0) + \nabla_\theta \Psi(\bar{P}, \bar{\theta})(\tilde{\theta} - \theta^0),$$

thus

$$\tilde{P} - P^0 = (I - \nabla_P \Psi(\bar{P}, \bar{\theta}))^{-1} \nabla_\theta \Psi(\bar{P}, \bar{\theta})(\tilde{\theta} - \theta^0).$$

Consequently, the first order condition is, in conjunction with $(\bar{P}, \bar{\theta}) \rightarrow_p (P^0, \theta^0)$,

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} A(x_i)\nabla_\theta \psi(P^0, \theta^0)(a_i|x_i) + o_p(1)(\tilde{\theta} - \theta^0) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} A(x_i)\left[\nabla_{\theta\theta}\psi(P^0, \theta^0)(a_i|x_i) + \nabla_{\theta P}\psi(P^0, \theta^0)(a_i|x_i)(I - \Psi_P)^{-1}\Psi_\theta\right](\tilde{\theta} - \theta^0). \end{aligned}$$

It follows that

$$\sqrt{n}(\tilde{\theta} - \theta^0) \rightarrow_d N(0, D(\tau)^{-1} E[m(z, \tau)m(z, \tau)'](D(\tau)^{-1})'),$$

where $\tau = A(x)$, $z = (a, x)$, and

$$\begin{aligned} D(\tau) &= E\left[A(x)\left(\nabla_{\theta\theta}\psi(P^0, \theta^0)(a|x) + \nabla_{\theta P}\psi(P^0, \theta^0)(a|x)(I - \Psi_P)^{-1}\Psi_\theta\right)\right], \\ m(z, \tau) &= A(x)\nabla_\theta \psi(P^0, \theta^0)(a|x). \end{aligned}$$

From Theorem 5.3 of Newey and McFadden (1994), the optimal instrument $\bar{\tau}$ satisfies $D(\tau) = E[m(Z, \tau)m(Z, \bar{\tau})']$ for all τ. Using an argument analogous to Newey and McFadden (1994, pp. 2168-2170), the optimal instrument $\bar{\tau} = \bar{A}(x)$ is

$$\begin{aligned} \bar{A}(x) &= E\left[\nabla_{\theta\theta}\psi(P^0, \theta^0)(a|x) + \nabla_{\theta P}\psi(P^0, \theta^0)(a|x)(I - \Psi_P)^{-1}\Psi_\theta | x\right]' \\ &\quad \times E\left[\nabla_\theta \psi(P^0, \theta^0)(a|x)\nabla_\theta \psi(P^0, \theta^0)(a|x)'|x\right]^{-1} \end{aligned}$$

With this instrument,

$$D(\bar{\tau}) = E\left\{\bar{A}(x) E\left[\nabla_\theta \psi(P^0, \theta^0)(a|x)\nabla_\theta \psi(P^0, \theta^0)(a|x)'|x\right]^{-1} \bar{A}(x)'\right\},$$

and the asymptotic variance of the efficient GMM estimator is $D(\bar{\tau})^{-1}$, which is different from the asymptotic variance of the NPL estimator. □

6.7 Proof of Lemma 2

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

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

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

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

where $\Omega = E[g(a_i,x_i;P^0)g(a_i,x_i;P^0)']$. From Theorem 5.3 of Newey and McFadden (1994), the limiting variance is minimized by using a weighting matrix $W^* = (G_\theta^{-1})'(G_\theta^\infty)'\Omega^{-1}$ if it is positive semi-definite. □

6.8 Proof of Lemma 3 and Corollary 2

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

$$\bar{G}_\theta(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{g}(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j)) = 0. \tag{8}$$

Expanding $\bar{g}(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))$ around $(\tilde{P},\tilde{\theta})$ in (8) and using $\bar{G}_\theta'(\Psi(\tilde{P},\tilde{\theta}))\hat{W}\bar{g}(\Psi(\tilde{P},\tilde{\theta})) = 0$ gives

$$\begin{aligned} \tilde{\theta}_j - \tilde{\theta} &= [\bar{G}_\theta'(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{G}_\theta(\Psi(\bar{P},\bar{\theta})) + o_p(1)]^{-1}[\bar{G}_\theta'(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{G}_P(\Psi(\bar{P},\bar{\theta})) + o_p(1)](\tilde{P}_{j-1}-\tilde{P}) \\ &= O_p(||\tilde{P}_{j-1}-\tilde{P}||), \end{aligned} \tag{9}$$

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

For the second result, first, using (9), we obtain the same approximation as (7):

$$\tilde{P}_j - \tilde{P} = \Psi_P(\tilde{P}_{j-1}-\tilde{P}) + \Psi_\theta(\tilde{\theta}_j - \tilde{\theta}) + O_p(n^{-1/2}||\tilde{P}_{j-1}-\tilde{P}||) + O_p(||\tilde{P}_{j-1}-\tilde{P}||^2) \tag{10}$$

Expanding $\bar{g}(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))$ in (8) twice around $(\tilde{P},\tilde{\theta})$ and using $\bar{G}_\theta'(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{g}(\Psi(\tilde{P},\tilde{\theta})) =$

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

$$\bar{G}_P(\Psi(\tilde{P},\tilde{\theta})) = G_P + O_p(n^{-1/2}), \qquad \bar{G}_\theta(\Psi(\tilde{P},\tilde{\theta})) = G_\theta + O_p(n^{-1/2}) \tag{11}$$

and (9) gives

$$\begin{aligned} 0 \;=\;& \bar{G}_\theta'(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}G_P(\tilde{P}_{j-1} - \tilde{P}) + \bar{G}_\theta'(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}G_\theta(\tilde{\theta}_j - \tilde{\theta}) \\ & + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2). \end{aligned} \tag{12}$$

Expanding $\Psi(\tilde{P}_{j-1},\tilde{\theta}_j)$ around $(\tilde{P},\tilde{\theta})$ and using (9) and (11) in (12), 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 (10) and noting that $G_\theta = -\Gamma\Delta_x\Psi_\theta$ and $G_P = -\Gamma\Delta_x\Psi_P$, we obtain

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

and the second result follows. Repeating the similar argument proves the corollary. □

6.9 Proof of Lemma 4

The consistency of $\tilde{\theta}^q$ follows from a standard argument. The first order condition is

$$0 = (\partial/\partial\theta)\Psi^q(\hat{P}_0,\tilde{\theta}^q)'\hat{W}\bar{g}^q(\tilde{\theta}^q,\hat{P}_0) = (\Psi_\theta^{(q)} + o_p(1))'\hat{W}[\hat{P}_0 - \Psi^q(\hat{P}_0,\tilde{\theta}^q)].$$

Expanding $\Psi^q(\hat{P}_0,\tilde{\theta}^q)$ on the right hand side around (P^0,θ^0) gives

$$0 = (\Psi_\theta^{(q)} + o_p(1))'\hat{W}(I - D_P\Psi^q(\bar{P},\bar{\theta}))(\hat{P}_0 - P^0) + (\Psi_\theta^{(q)} + o_p(1))'\hat{W}D_\theta\Psi^q(\bar{P},\bar{\theta})(\tilde{\theta}^q - \theta^0),$$

where $(\bar{P},\bar{\theta})$ lies between $(\hat{P}_0,\tilde{\theta}^q)$ and (P^0,θ^0). Then the asymptotic distribution of $n^{1/2}(\tilde{\theta}^q - \theta^0)$ follows from $n^{1/2}(\tilde{P} - P^0) \to_d N(0,\Sigma)$ and Slutsky's theorem. The choice of the optimal W follows from Theorem 5.3 of Newey and McFadden (1994, p.2166), and the limiting variance of $\tilde{\theta}^q$ has the stated form because by $(I - \Psi_P^{(q)})^{-1}\Psi_\theta^{(q)} = (I - \Psi_P)^{-1}\Psi_\theta$ by $\Psi_\theta^{(q)} = (I - \Psi_P^q)(I - \Psi_P)^{-1}\Psi_\theta$ and $\Psi_P^{(q)} = \Psi_P^q$. □

6.10 Proof of Lemma 5

For notational simplicity, we use $\tilde{\theta}$ to denote $\tilde{\theta}_{NMD}^q$. First, we write $\tilde{P} - P^0$ in terms of $\tilde{\theta} - \theta^0$. Note that

$$\Psi(\tilde{P},\tilde{\theta}) = \Psi(P^0,\theta^0) + D_P\Psi(\bar{P},\bar{\theta})(\tilde{P} - P^0) + D_\theta\Psi(\bar{P},\bar{\theta})(\tilde{\theta} - \theta^0),$$

where $(\bar{P},\bar{\theta})$ denotes a generic point between (P^0,θ^0) and $(\tilde{P},\tilde{\theta})$. In conjunction with $\tilde{P} = \Psi(\tilde{P},\tilde{\theta})$ and $P^0 = \Psi(P^0,\theta^0)$, it follows that

$$\tilde{P} - P^0 = (I - D_P\Psi(\bar{P},\bar{\theta}))^{-1} D_\theta \Psi(\bar{P},\bar{\theta})(\tilde{\theta} - \theta^0). \tag{13}$$

We proceed to show the asymptotic distribution of $\tilde{\theta}$. The consistency of $\tilde{\theta}$ follows from AM. The first order conditions

$$\begin{aligned} 0 &= (D_\theta \Psi^q(\tilde{P},\tilde{\theta}))'\hat{W}\bar{g}^q(\tilde{\theta},\tilde{P}) \\ &= (D_\theta \Psi^q(\tilde{P},\tilde{\theta}))'\hat{W}[\hat{P}_0 - \tilde{P}] \\ &= (D_\theta \Psi^q(\tilde{P},\tilde{\theta}))'\hat{W}(\hat{P}_0 - P^0) - (D_\theta \Psi^q(\tilde{P},\tilde{\theta}))'\hat{W}(\tilde{P} - P^0). \end{aligned} \tag{14}$$

Since $D_\theta \Psi^q(\tilde{P},\tilde{\theta}) \to_p \Psi_\theta^{(q)}$, for the first term on the right of (14) we have

$$n^{1/2}(D_\theta^{(q)}\Psi(\tilde{P},\tilde{\theta}))'\hat{W}(\hat{P}_0 - P^0) \to_d N(0, \Psi_\theta^{(q)\prime} W \Sigma W' \Psi_\theta^{(q)}).$$

For the second term on the right of (14), using (13) gives

$$(D_\theta \Psi^q(\tilde{P},\tilde{\theta}))'\hat{W}(\tilde{P} - P^0) = (\Psi_\theta^{(q)} + o_p(1))'\hat{W}(I - \Psi_P + o_p(1))^{-1}(\Psi_\theta^{(q)} + o_p(1))(\tilde{\theta} - \theta^0),$$

and the asymptotic distribution of $n^{1/2}(\tilde{\theta} - \theta^0)$ follows from $n^{1/2}(\tilde{P} - P^0) \to_d N(0,\Sigma)$ and Slutsky's theorem. The choice of the optimal weighting matrix follows from Theorem 5.3 of Newey and McFadden (1994, p.2166).[3] □

6.11 Proof of Lemma 6 and Corollary 3

The proof closely follows the proof of the GMM estimator. First, note that $(\partial/\partial\theta')\bar{g}(\theta,P) = -D_\theta\Psi(P,\theta)$ and $(\partial/\partial\theta')\bar{g}(\theta,P) = -D_P\Psi(P,\theta)$.

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

$$[D_\theta \Psi(\tilde{P}_{j-1},\tilde{\theta}_j)]'\hat{W}\bar{g}(\tilde{\theta}_j,\tilde{P}_{j-1}) = 0. \tag{15}$$

Expanding $\bar{g}(\tilde{\theta}_j,\tilde{P}_{j-1})$ around $(\tilde{P},\tilde{\theta})$ gives, with $(\bar{P},\bar{\theta})$ between $(\tilde{P}_{j-1},\tilde{\theta}_j)$ and $(\tilde{P},\tilde{\theta})$,

$$\begin{aligned} 0 \;=\; & (D_\theta\Psi(\tilde{P},\tilde{\theta}))'\hat{W}\bar{g}(\tilde{\theta},\tilde{P}) + [D_\theta\Psi(\tilde{P}_{j-1},\tilde{\theta}_j) - D_\theta\Psi(\tilde{P},\tilde{\theta})]'\hat{W}\bar{g}(\tilde{\theta},\tilde{P}) \\ & + (D_\theta\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))'\hat{W}D_P\Psi(\bar{P},\bar{\theta})(\tilde{P}_{j-1} - \tilde{P}) + (D_\theta\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))'\hat{W}[D_\theta\Psi(\bar{P},\bar{\theta})](\tilde{\theta}_j - \tilde{\theta}). \end{aligned} \tag{16}$$

Since $\tilde{P}_{j-1} - \tilde{P} = o_p(1)$ and $[D_\theta\Psi(\tilde{P},\tilde{\theta})]'\hat{W}\bar{g}(\tilde{\theta},\tilde{P}) = 0$ (by the first order condition of a nested

[3] $D(\tau)$ and $m(Z,\tau)$ in Newey and McFadden (1994) corresponds to our $\Psi_\theta^{(q)\prime}W(I - \Psi_P)^{-1}\Psi_\theta^{(q)}$ and $\Psi_\theta^{(q)\prime}W(\hat{P}_0 - P^0)$, respectively.

CMD estimator), (16) gives

$$\tilde{\theta}_j - \tilde{\theta} = -[\Psi_\theta' W \Psi_\theta + o_p(1)]^{-1}[\Psi_\theta' W \Psi_P + o_p(1)](\tilde{P}_{j-1} - \tilde{P}) = O_p(||\tilde{P}_{j-1} - \tilde{P}||). \tag{17}$$

This proves the first result.

For the second result, first using the bound of $\tilde{\theta}_j - \tilde{\theta}$, we obtain

$$\tilde{P}_j - \tilde{P} = \Psi_P(\tilde{P}_{j-1} - \tilde{P}) + \Psi_\theta(\tilde{\theta}_j - \tilde{\theta}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2). \tag{18}$$

By expanding $D_\theta \Psi(\tilde{P}_{j-1}, \tilde{\theta}_j)$, $D_\theta \Psi(\bar{P}, \bar{\theta})$, and $D_P \Psi(\bar{P}, \bar{\theta})$ in (16) around $(\tilde{P}, \tilde{\theta})$, and using

$$D_\theta \Psi(\tilde{P}, \tilde{\theta}) = \Psi_\theta + O_p(n^{-1/2}), \quad D_P \Psi(\tilde{P}, \tilde{\theta}) = \Psi_P + O_p(n^{-1/2}),$$

we obtain a finer expression for (17):

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

Substituting this into (14), we obtain the second result:

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

Repeating the similar argument proves the corollary. □

References

Aiyagari, S. Rao (1994). “Uninsured idiosyncratic risk and aggregate saving.” *Quarterly Journal of Economics*, 109(3): 659-684.

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.

Bajari, P., Benkard, C.L., and Levin, J. (2005). “Estimating dynamic models of imperfect competition.” Mimeographed, Stanford University.

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

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) “Nested pseudo-likelihood estimation and bootstrap-based inference for structural discrete Markov decision models.” Queen’s University Working Paper.

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 (2005). Simple estimators for the parameters of discrete dynamic games (with entry/exit examples). Mimeographed, Harvard University.

Pesendorfer, M. and P. Schmidt-Dengler (2006). Asymptotic least squares estimators for dynamic games. Mimeographed, LSE.

Prescott, E.C. and R. Mehra (1980). “Recursive competitive equilibrium: the case of homogeneous households.” *Econometrica* 48(6): 1365-1379.

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

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