Inference in dynamic discrete choice problems under local misspecification*

Federico A. Bugni
Department of Economics
Duke University
federico.bugni@duke.edu

Takuya Ura
Department of Economics
University of California, Davis
takura@ucdavis.edu

December 7, 2016

Abstract

Dynamic discrete choice models are typically estimated using heavily parametrized econometric frameworks, making them susceptible to model misspecification. This paper investigates how misspecification can affect the results of inference in these models. For tractability reasons, we consider a local misspecification framework in which specification errors are assumed to vanish with the sample size. However, we impose no restrictions on the rate at which these errors vanish.

We consider a general class of two-stage estimators based on the K-step sequential policy function iteration algorithm, where K denotes the number of iterations employed in the estimation. This class includes Rust (1987)'s nested fixed point estimator, Hotz and Miller (1993)'s conditional choice probability estimator, Aguirregabiria and Mira (2002)'s pseudo-likelihood estimator, and Pesendorfer and Schmidt-Dengler (2008)'s asymptotic least squares estimator.

We show that local misspecification can affect asymptotic bias, asymptotic variance, and even the rate of convergence of these estimators. However, our main finding is that the effect of the local misspecification is invariant to the number of iterations K. In practice, this means that the choice of the number of iterations K should not be guided by concerns of model misspecification.

Keywords: Dynamic discrete choice problems, estimation, inference, misspecification, local misspecification.

JEL Classification Codes: C13, C61, C73

*Thanks to Peter Arcidiacono, Joe Hotz, Shakeeb Khan, Matt Masten, Arnaud Maurel, Jia Li, and the participants at the Duke Microeconometrics Reading Group for useful comments and suggestions. Of course, any and all errors are our own.

1 Introduction

This paper investigates the effect of model misspecification on inference in dynamic discrete choice models. Our study is motivated by two observations regarding this literature.[1] First, typical econometric frameworks used in empirical studies are heavily parametrized and are therefore subject to misspecification. Second, there are several methods that can be used to estimate these models, including Rust (1987, 1988)'s nested fixed point estimator, Hotz and Miller (1993)'s conditional choice probability estimator, Aguirregabiria and Mira (2002)'s pseudo-likelihood estimator, and Pesendorfer and Schmidt-Dengler (2008)'s asymptotic least squares estimator. While the literature has studied the behavior of these estimators under correct specification, their properties under misspecification have not been explored. In fact, to the best of our knowledge, our paper is the first attempt to investigate the effect of misspecification on inference in these types of models.

In this paper, we propose a local approach to misspecification in dynamic discrete choice models. By local misspecification, we mean that the mistakes in the specification are assumed to vanish with the sample size. In other words, the econometric model is misspecified, but the amount of misspecification vanishes as the sample size increases. This is an asymptotic device that can provide concrete conclusions in the presence of misspecification while keeping the analysis tractable.[2] As with any other asymptotic device, local misspecification is just an approximation and should not be taken literally.

Our local approach to misspecification in dynamic discrete choice models yields relevant conclusions in several dimensions. First, local misspecification can constitute a reasonable approximation to the asymptotic behavior when there are small mistakes in the specification of the model. Second, there are multiple available estimation methods and their performance under misspecification is not well understood. Their relative performance under local misspecification should be a relevant comparison criterion. Finally, while our approach to local misspecification is admittedly local in that the misspecification is assumed to vanish, we allow the rate at which this occurs to be completely arbitrary; in particular, it can be faster, equal, or slower than the regular parametric convergence, i.e., $\sqrt{n}$.

We consider a class of two-stage estimators based on the K-step sequential policy function iteration algorithm along the lines of Aguirregabiria and Mira (2002), where K denotes the number of iterations employed in the estimation. By appropriate choice of the criterion function, this class captures the K-step maximum likelihood estimators (K-ML) and the K-step minimum distance estimators (K-MD). This is a general class that includes the previously mentioned estimators as special cases.

In this context, we show that local misspecification can affect asymptotic bias, asymptotic variance, and even the rate of convergence of these estimators. We obtain three general main results. First, we show that K-ML estimators are asymptotically equivalent regardless of the choice of K. While this result was shown under correct specification by Aguirregabiria and Mira (2002), its validity under an arbitrary amount of local misspecification is completely new. Second, we show that an analogous result holds for K-MD estimators, i.e., given the choice of weight matrix, K-MD estimators are asymptotically equivalent regardless of the choice of K. The combination of these two results implies that the choice of the number of iterations K in dynamic discrete choice models should not be guidad by concerns of model misspecification. To the best of our knowledge, this conclusion is completely new to the literature. Our third and final result is to compare K-MD and K-ML estimators in terms of asymptotic mean squared error. Under local misspecification, the

[1] See survey papers by Aguirregabiria and Mira (2010) and Arcidiacono and Ellickson (2011), and references therein.

[2] We are certainly not the first paper to propose a local approach to study the robustness of econometric inference to misspecification. See references below.

optimally weighted K-MD estimator depends on the unknown asymptotic bias and is thus unfeasible. In turn, feasible K-MD estimators could have an asymptotic mean squared error that could be higher or lower than that of K-ML estimators.

In practice, researchers often specify econometric models that may contain non-vanishing specification errors, i.e., global misspecification. Relative to the global misspecification analysis, our local misspecification approach has two related advantages. First, global misspecification analysis can be intractable, making it impossible to produce general findings. In contrast, our local misspecification yields concrete and general conclusions. Second, as we have already mentioned, there are multiple available estimation methods for the structural parameter of interest. Under correct specification, all of these estimators are (typically) consistent and they only differ in asymptotic distribution. Under global misspecification, the different estimation methods typically converge to different pseudo-true parameters, making results hard to interpret and compare. In contrast, under local misspecification, all of these estimators will be shown to be consistent to the same parameter value and so local misspecification can produce meaningful comparisons in terms of their asymptotic distributions.

This paper relates to an important literature on inference under model misspecification. For example, White (1982, 1996) and Newey (1985a,b) investigate the power properties of the specification tests based on the point identified models under model misspecification. We are also related to a literature that consider a local approach to investigate the robustness of inference to the assumptions of the econometric model. For example, Chesher (1991) considers a vanishing amount of measurement error in the covariates of a quantile regression model. Schorfheide (2005) considers a local misspecified vector autoregression process and proposes an information criterion for the lag length in the autoregression model. Bugni et al. (2012) compare inference methods in partially identified moment (in)equality models that are locally misspecified. Another example is Kitamura et al. (2013), who consider a class of estimators that are robust to local misspecification in the context of point identified moment condition models. In other related work, Kasahara and Shimotsu (2008) study the second order behavior of K-ML estimators under correct specification. Their analysis advocates using a large value of K to improve the convergence rate to the MLE (K-ML with $K = \infty$). The second order properties of our analysis under local misspecification might be an interesting extension to our project that we consider out of the scope of the present paper.

The remainder of the paper is structured as follows. Section 2 describes the econometric framework. Section 2.1 introduces the dynamic discrete choice model and Section 2.2 introduces the possibility of local misspecification into the econometric model. Section 3 studies the problem of inference in the locally misspecified model. This section contains the main result of the paper, which characterizes the asymptotic distribution of general two-stage estimators under local misspecification. Section 4 applies this result to the class of maximum likelihood estimators (Section 4.1) and the class of minimum distance estimators (Section 4.2). Section 5 presents results of Monte Carlo simulation and Section 6 concludes. The appendix of the paper collects all the proofs and intermediate results.

The following notation is used throughout the paper. For any $s \in \mathbb{N}$, $\mathbf{0}_s$ and $\mathbf{1}_s$ denote a column vector of size $s \times 1$ composed of zeros and ones, respectively, and $\mathbf{I}_s$ denotes the identity matrix of size $s \times s$. We use $||\cdot||$ to denote the Euclidean norm. For sets of finite indices $S_1 = \{1, \ldots, |S_1|\}$ and $S_2 = \{1, \ldots, |S_2|\}$, $\{M(s_1, s_2)\}_{(s_1,s_2)\in S_1\times S_2}$ denotes the column vector equal to the vectorization of $\{\{M(s_1, s_2)\}_{s_1=1}^{|S_1|}\}_{s_2=1}^{|S_2|}$. Finally, "w.p.a.1" abbreviates "with probability approaching one".

2 Setup

The researcher is interested in modeling the behavior of an agent solving a dynamic discrete choice problem. He assumes an econometric model as described in Section 2.1. Unfortunately, the specified model is incorrect and the nature of the misspecification is characterized in Section 2.2.

2.1 The econometric model

The researcher assumes that the agent behaves according to the discrete Markov decision framework in Aguirregabiria and Mira (2002). In each period $t = 1, \ldots, T \equiv \infty$, the agent is assumed to observe a vector of state variables s_t and to choose an action $a_t \in A \equiv \{1, \ldots, |A|\}$ with the objective of maximizing the expected discounted utility. The vector of state variables $s_t = (x_t, \epsilon_t)$ is composed by two subvectors. The subvector $x_t \in X \equiv \{1, \ldots, |X|\}$ represents a scalar state variables observed by the agent and the researcher, whereas the subvector $\epsilon_t \in \mathbb{R}^{|A|}$ represents an action-specific state vector only observed by the agent.

The uncertainty about the agent's future state variables $(x_{t+1}, \epsilon_{t+1})$ are modeled by a Markov transition probability density $d\Pr(x_{t+1}, \epsilon_{t+1}|x_t, \epsilon_t, a_t)$ that factors in the following manner:

$$d\Pr(x_{t+1}, \epsilon_{t+1}|x_t, \epsilon_t, a_t) \;=\; g_{\theta_g}(\epsilon_{t+1}|x_{t+1}) f_{\theta_f}(x_{t+1}|x_t, a_t),$$

where $g_{\theta_g}(\cdot)$ is the (conditional) distribution of the unobserved state variable and $f_{\theta_f}(\cdot)$ is the transition probability of the observed state variable, with parameters θ_g and θ_f, respectively.

The utility is assumed to be time separable and the agent discounts future utility by a known discount factor $\beta \in (0, 1)$.[3] The current utility function of choosing action a_t under state variables (x_t, ϵ_t) is given by:

$$u_{\theta_u}(x_t, a_t) \;+\; \epsilon_t(a_t),$$

where $u_{\theta_u}(\cdot)$ is non-stochastic component of the current utility with parameter θ_u, and $\epsilon_t(a_t)$ denotes the $a_t th$ coordinate of ϵ_t.

The researcher's goal is to estimate the unknown parameters in the model, $\theta \equiv (\theta_g, \theta_u, \theta_f) \in \Theta$, where Θ is the compact parameter space. For the sake of notation, we use $\theta = (\alpha, \theta_f) \in \Theta \equiv \Theta_\alpha \times \Theta_f$ with $\alpha \equiv (\theta_u, \theta_g) \in \Theta_\alpha$ and $\theta_f \in \Theta_f$.

Following Aguirregabiria and Mira (2002), we impose the following regularity conditions.

Assumption 1. (Regularity conditions) For every $\theta \in \Theta$, assume that:

(a) For every $x \in X$, $g_{\theta_g}(\epsilon|x)$ has finite first moments and is twice differentiable in ϵ.

(b) $\epsilon = \{\epsilon(a)\}_{a \in A}$ has full support.

(c) $g_{\theta_g}(\epsilon|x)$, $f_{\theta_f}(x'|x, a)$, and $u_{\theta_u}(x, a)$ are twice continuously differentiable with respect to θ.

By Blackwell (1965)'s theorem and its generalization by Rust (1988), the model implies that the agent has a stationary and Markovian optimal decision rule. Therefore, the researcher can drop the time subscript from the model and use prime to denote future periods. Furthermore, the agent's optimal value function V_θ is the unique solution of the following Bellman equation.

[3] This follows Aguirregabiria and Mira (2002, Footnote 12) and the identification analysis in Magnac and Thesmar (2002).

$$V_\theta(x,\epsilon) = \max_{a\in A}\{u_{\theta_u}(x,a) + \epsilon(a) + \beta \int_{(x',\epsilon')} V_\theta(x',\epsilon') g_{\theta_g}(\epsilon'|x') f_{\theta_f}(x'|x,a) d(x',\epsilon')\}, \tag{2.1}$$

By integrating out the unobserved error, we obtain the smoothed value function:

$$V_\theta(x) \equiv \int_\epsilon V_\theta(x,\epsilon) g_{\theta_g}(\epsilon|x) d\epsilon,$$

which is the unique solution of the smoothed Bellman equation, given by:

$$V_\theta(x) = \int_\epsilon \max_{a\in A}\{u_{\theta_u}(x,a) + \epsilon(a) + \beta \sum_{x'\in X} V_\theta(x') f_{\theta_f}(x'|x,a)\} g_{\theta_g}(\epsilon|x) d\epsilon. \tag{2.2}$$

We now turn to the description of the conditional choice probability (CCP), denoted by $P_\theta(a|x)$, which is the model implied probability that an agent chooses action a when the observed state is x. Since the agent chooses among the actions in A, $P_\theta(|A||x) = 1 - \sum_{a\in\tilde{A}} P_\theta(a|x)$ for all $x \in X$. As a consequence, the vector of model implied conditional choice probabilities (CCPs) is completely characterized by $\{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X}$ with $\tilde{A} \equiv \{1,\ldots,|A|-1\}$. For the remainder of the paper, we use $\Theta_P \subset [0,1]^{|\tilde{A}X|}$ to denote the parameter space for the CCPs.

The vector of CCPs is a central equilibrium object in the model. Lemma 2.1 shows that the CCPs are the unique fixed point of the policy function mapping. As a result of utility maximization, the vector of CCPs is determined by the following equation:

$$P_\theta(a|x) \equiv \int_\epsilon 1\left[a = \arg\max_{\tilde{a}\in A}[u_{\theta_u}(x,\tilde{a}) + \beta \sum_{x'\in X} V_\theta(x') f_{\theta_f}(x'|x,\tilde{a}) + \epsilon_{\tilde{a}}]\right] dg_{\theta_g}(\epsilon|x),$$

which can be succinctly represented as follows:

$$\{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X} \;=\; \Lambda_\theta(\{V_\theta(x)\}_{x\in X}). \tag{2.3}$$

Also, notice that Eq. (2.2) can be re-written as:

$$V_\theta(x) = \sum_{a\in A} P_\theta(a|x)\left\{u_{\theta_u}(x,a) + E_\theta[\epsilon(a)|x,a] + \beta \sum_{x'\in X} V_\theta(x') f_{\theta_f}(x'|x,a)\right\}, \tag{2.4}$$

where $E_\theta[\epsilon(a)|x,a]$ denotes the expectation of the unobservable $\epsilon(a)$ conditional on the state being x and on the optimal action being a. Under our assumptions, Hotz and Miller (1993) show that there is a one-to-one mapping relating the CCPs and the (normalized) smoothed value function. The inverse of this mapping allows us to re-express $\{E_\theta[\epsilon(a)|x,a]\}_{(a,x)\in AX}$ as a function of the vector of CCPs. By combining this re-expression and Eq. (2.4), we can express the vector $\{V_\theta(x)\}_{x\in X}$ as a function of the vector $\{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X}$. An explicit formula for such function is provided in Aguirregabiria and Mira (2002, Equation (8)), which we succinctly express as follows:

$$\{V_\theta(x)\}_{x\in X} \;=\; \varphi_\theta(\{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X}). \tag{2.5}$$

By combining Eqs. (2.3) and (2.5), we obtain the following fixed point representation of $\{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X}$:

$$\{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X} = \Psi_\theta(\{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X}), \tag{2.6}$$

where $\Psi_\theta \equiv \Lambda_\theta \circ \varphi_\theta$ is the policy function mapping. As explained by Aguirregabiria and Mira (2002), this operator can be evaluated at any vector of conditional choice probabilities, optimal or not. For any arbitrary $P \equiv \{P(a|x)\}_{(a,x)\in\tilde{A}X}$, $\Psi_\theta(P)$ provides the current optimal choice probabilities of an agent whose future behavior is distributed according to P.

Under the current assumptions, the policy function mapping Ψ_θ in Eq. (2.6) has several properties that are central to the results of this paper.

Lemma 2.1. *Assume Assumption 1. Then, Ψ_θ satisfies the following properties:*

(a) Ψ_θ has a unique fixed point $P_\theta \equiv \{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X}$,

(b) The sequence $P^K = \Psi_\theta(P^{K-1})$ for $K \geq 1$, converges to P_θ for any initial P^0,

(c) $\Psi_\theta(P)$ is twice continuously differentiable in θ and P,

(d) The Jacobian matrix of Ψ_θ with respect to P is zero at P_θ,

(e) $\Psi_\theta(P)(a|x) > 0$ for any $(a,x) \in AX$ and any θ and P.[4]

Following the literature on estimation of dynamic discrete choice models, estimates $\theta = (\alpha, \theta_f)$ using a two-stage procedure Ψ_θ. In a first step, he uses f_{θ_f} to estimate θ_f. In a second step, he uses $\Psi_{(\alpha,\theta_f)}(P)$ and the first-step results to estimate α. To this end, the following assumption is imposed.

Assumption 2. (Identification) The parameter $\theta = (\alpha, \theta_f) \in \Theta$ is identified as follows:

(a) θ_f is identified by f_{θ_f}, i.e., $f_{\theta_{f,a}} = f_{\theta_{f,b}}$ implies $\theta_{f,a} = \theta_{f,b}$.

(b) α is identified by the fixed point condition $\Psi_{(\alpha,\theta_f)}(P) = P$ for any $(\theta_f, P) \in \Theta_f \times \Theta_P$, i.e., $\Psi_{(\alpha_a,\theta_f)}(P) = P$ and $\Psi_{(\alpha_b,\theta_f)}(P) = P$ implies $\alpha_a = \alpha_b$.

Magnac and Thesmar (2002) provide sufficient conditions for Assumption 2. Also, Assumption 2 implies the higher level condition used by Aguirregabiria and Mira (2002, conditions (e)-(f) in Proposition 4). Under these conditions, we can deduce certain important properties for the model implied CCPs.

Lemma 2.2. *Assume Assumptions 1-2. Then,*

(a) $P_\theta(a|x) > 0$ for any $(a,x) \in \tilde{A}X$ and $\sum_{a\in\tilde{A}} P_\theta(a|x) = 1$ for any $x \in X$,

(b) P_θ is continuously differentiable,

(c) $\partial P_\theta/\partial\theta = \partial\Psi_\theta(P_\theta)/\partial\theta$,

(d) α is identified by $P_{(\alpha,\theta_f)}$ for any $\theta_f \in \Theta_f$, i.e., $\forall\theta_f \in \Theta_f$, $P_{(\alpha_a,\theta_f)} = P_{(\alpha_b,\theta_f)}$ implies $\alpha_a = \alpha_b$.

Thus far, we have described how the model specifies two conditional distributions: the CCPs and the transition probabilities. The characterization of the model is completed by the marginal distribution of the observed state variables, which the researcher leaves completely unspecified.

[4] This expression is an abuse of the notation for $a = |A|$ in the sense that $\Psi_\theta(P)(a|x)$ is only defined for $(a,x) \in \tilde{A}X$, i.e., it is not defined when $a = |A|$. To complete the definition, we use $\Psi_\theta(P)(|A||x) \equiv 1 - \sum_{a\in\tilde{A}} \Psi_\theta(P)(a|x)$ for any $x \in X$.

2.2 Local misspecification

We now describe the true underlying data generating process (DGP), denoted by $\Pi_n^*(a, x, x')$, and characterize its relationship to the econometric model. Here and for the rest of the paper, the superscript with asterisk denotes true value. It is important that to acknowledge that the DGP, i.e., a population object, is indexed by the sample size n. As we explain in this section, this indexing serves an important purpose in the locally misspecification framework.

By definition, the DGP is the product of the transition probability, the CCPs, and the marginal distribution of the state variable, i.e.,

$$\Pi_n^*(a, x, x') \;=\; f_n^*(x'|a, x) \;\times\; P_n^*(a|x) \;\times\; m_n^*(x), \tag{2.7}$$

where:

$$\begin{aligned} f_n^*(x'|a, x) \quad &\equiv \quad \frac{\Pi_n^*(a, x, x')}{\sum_{\tilde{x}' \in X} \Pi_n^*(a, x, \tilde{x}')} \quad \forall (a, x, x') \in AX^2, \\ P_n^*(a|x) \quad &\equiv \quad \frac{\sum_{\tilde{x}' \in X} \Pi_n^*(a, x, \tilde{x}')}{\sum_{(\check{a}, \check{x}') \in AX} \Pi_n^*(\check{a}, x, \check{x}')} \quad \forall (a, x) \in AX, \\ m_n^*(x) \quad &\equiv \quad \sum_{(a, x') \in AX} \Pi_n^*(a, x, x') \quad \forall x \in X. \end{aligned} \tag{2.8}$$

The econometric model in Section 2.1 specifies $P_\theta(a|x)$ as the model for $P_n^*(a|x)$, $f_{\theta_f}(x'|a, x)$ as the model for $f_n^*(x'|a, x)$, and imposes no restrictions on $m_n^*(x)$. This paper allows the econometric model to be misspecified, i.e.,

$$\inf_{(\alpha, \theta_f) \in \Theta_\alpha \times \Theta_f} \| \; (P_{(\alpha, \theta_f)} - P_n^*) \; , \; (f_{\theta_f} - f_n^*) \; \| \geq 0, \tag{2.9}$$

but requires the misspecification to vanishes asymptotically according to the following assumption.

Assumption 3. (Local misspecification)

(a) The sequence of DGPs $\{\Pi_n^*\}_{n \geq 1}$ converges to a limiting DGP Π^* at n^δ-rate for some $\delta > 0$. In particular,

$$n^\delta (\Pi_n^* - \Pi^*) \to B_\Pi \in \mathbb{R}^{AX^2}.$$

(b) According to the limiting DGP Π^*, the econometric model is correctly specified, i.e.,

$$\inf_{(\alpha, \theta_f) \in \Theta_\alpha \times \Theta_f} \| \; (P_{(\alpha, \theta_f)} - P^*) \; , \; (f_{\theta_f} - f^*) \; \| = 0,$$

where the limiting CCPs P^* and limiting transition probabilities f^* are defined from the limiting DGP Π^* according to:

$$\begin{aligned} P^*(a|x) \quad &\equiv \quad \frac{\sum_{\tilde{x}' \in X} \Pi^*(a, x, \tilde{x}')}{\sum_{(\check{a}, \tilde{x}') \in AX} \Pi^*(\check{a}, x, \tilde{x}')} \quad \forall (a, x) \in AX, \\ f^*(x'|a, x) \quad &\equiv \quad \frac{\Pi^*(a, x, x')}{\sum_{\tilde{x}' \in X} \Pi^*(a, x, \tilde{x}')} \quad \forall (a, x, x') \in AX^2. \end{aligned}$$

Assumption 3 defines the local misspecification framework used in this paper. While Eq. (2.9) allows the

econometric model to be misspecified, Assumption 3 implies that the limiting distribution of the data can be correctly represented by the model. Note that Assumption 3(a) allows for misspecification to vanish at an arbitrary rate denoted by δ.

Under these assumptions, Theorem 2.1 implies that there is a unique true limiting parameter value (α^*, θ_f^*), regardless of the rate of local misspecification δ.

Theorem 2.1. *Under Assumptions 2-3(b), there is a unique* $(\alpha^*, \theta_f^*) \in \Theta$ *such that* $P_{(\alpha^*, \theta_f^*)} = P^*$ *and* $f_{\theta_f^*} = f^*$.

3 Inference in the locally misspecified model

Our paper focuses on two-stage estimators based on the K-step sequential policy function iteration (PI) algorithm developed by Aguirregabiria and Mira (2002).[5] Let $\hat{\theta}_{f_n}$ denote the first-step estimator. Then, the K-step PI estimator of α is given by:

$$\hat{\alpha}_n^K \equiv \arg\max_{\alpha \in \Theta_\alpha} Q_n(\alpha, \hat{\theta}_{f_n}, \hat{P}_n^{K-1}), \tag{3.1}$$

where $Q_n : \Theta_\alpha \times \Theta_f \times \Theta_P \to \mathbb{R}$ is a sample objective function chosen by the researcher, and $\hat{P}_n^K$ denotes the K-step estimator of the CCPs which is defined as follows. If we let $\hat{P}_n^0$ denote the zero-step or preliminary CCP estimator, the K-step CCP estimator $\hat{P}_n^K$ for some $K \in \mathbb{N}$ is iteratively defined as follows:

$$\hat{P}_n^K \equiv \Psi_{(\hat{\alpha}_n^K, \hat{\theta}_{f_n})}(\hat{P}_n^{K-1}).$$

Section 4 considers two possible choices for the sample objective function Q_n, leading to different estimation procedures. In particular, Section 4.1 considers maximum likelihood type (ML) estimation, while Section 4.2 considers minimum distance (MD) estimation.

We impose two assumptions that involve the sample objective function Q_n in Eq. (3.1) and the first-step estimator in θ_f^*. While these conditions are admittedly high-level, they are standard in extremum estimator problems and they will be verified for ML and MD estimators in Sections 4.1 and 4.2, respectively.

Assumption 4. (Interiority) α^* belongs to the interior of Θ_α.

Assumption 5. (Regularity for extremum estimators) Let $\mathcal{N}$ denote an arbitrary small neighborhood of (α, θ_f, P) around $(\alpha^*, \theta_f^*, P^*)$. Then,

(a) $\sup_{\alpha \in \Theta_\alpha} |Q_n(\alpha, \hat{\theta}_{f_n}, \tilde{P}_n) - Q_\infty(\alpha, \theta_f^*, P^*)| = o_{P_n}(1)$, provided that $\tilde{P}_n = P^* + o_{P_n}(1)$,

(b) $Q_\infty(\alpha, \theta_f^*, P^*)$ is uniquely maximized at α^*.

(c) $\partial^2 Q_n(\alpha, \theta_f, P)/\partial\alpha\partial\lambda$ is continuous on $\mathcal{N}$ for $\lambda \in \{\alpha, \theta_f, P\}$ w.p.a.1.

(d) $\sup_{(\alpha, \theta_f, P) \in \mathcal{N}} \|\partial^2 Q_n(\alpha, \theta_f, P)/\partial\alpha\partial\lambda - \partial^2 Q_\infty(\alpha, \theta_f, P)/\partial\alpha\partial\lambda\| = o_{P_n}(1)$ for $\lambda \in \{\alpha, \theta_f, P\}$.

(e) $\partial^2 Q_\infty(\alpha, \theta_f, P)/\partial\alpha\partial\alpha'$ is continuous and non-singular at $(\alpha^*, \theta_f^*, P^*)$.

[5] Results for single stage estimators procedures are relatively easy to deduce from our analysis by considering the special case in which the entire parameter vector is estimated on the second stage.

(f) $\partial^2 Q_\infty(\alpha^*, \theta_f^*, P^*)/\partial\lambda\partial P = \mathbf{0}_{d_\lambda \times |AX|}$ for $\lambda \in \{\alpha, \theta_f\}$.

Assumption 6. (Baseline convergence for extremum estimators)

(a) $n^{\min\{1/2,\delta\}}[\partial Q_n(\alpha^*, \theta_f^*, P^*)/\partial\alpha', (\hat{\theta}_{f,n} - \theta_f^*)] \xrightarrow{d} [\zeta_1, \zeta_2]$.

(b) $n^{\min\{1/2,\delta\}}(\hat{P}_n^0 - P^*) = O_{P_n}(1)$.

Assumption 4 is a standard assumption in extremum estimators and is also required by Aguirregabiria and Mira (2002). Assumptions 5-6 are high-level conditions regarding the asymptotic behavior of K-step PI estimators. These assumptions will be verified using low-level conditions for both ML and MD estimators in Section 4. In particular, Assumption 5(f) is shown to be a consequence of the zero Jacobian property derived in Lemma 2.1(d). Assumption 6 reflects the fact that the econometric model is locally misspecified at a rate of n^δ. It indicates that certain random variables converge in distribution of a limiting value at a rate of $n^{\min\{1/2,\delta\}}$, i.e., the slowest rate between the local misspecification and the regular rate for parametric estimation.

Under these assumptions, Theorem 3.1 establishes the asymptotic distributions of general two-stage K-step policy function iteration estimator.

Theorem 3.1. *Assume Assumptions 1-6. For any $K \geq 1$,*

$$
\begin{aligned}
&n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^K - \alpha^*) \\
&= (-\frac{\partial^2 Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial\alpha\partial\alpha'})^{-1} n^{\min\{1/2,\delta\}}[\frac{\partial Q_n(\alpha^*, \theta_f^*, P^*)}{\partial\alpha'} + \frac{\partial^2 Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial\alpha'\partial\theta_f}(\hat{\theta}_{f,n} - \theta_f^*)] + o_{p_n}(1) \\
&\xrightarrow{d} (-\frac{\partial^2 Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial\alpha\partial\alpha'})^{-1}[\zeta_1 + \frac{\partial^2 Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial\alpha'\partial\theta_f}\zeta_2].
\end{aligned}
$$

The result reveals two important properties of the asymptotic distributions of two-stage K-step policy function iteration estimator. First, local misspecification at the rate of n^δ causes the estimator to converge at a rate of $n^{\min\{1/2,\delta\}}$. Second, the asymptotic distribution of the two-stage K-step policy function iteration estimator does not depend on the number of iterations K.[6] In fact, the invariance the number of iterations K holds regardless of the rate of local misspecification δ.

The invariance of the asymptotic distribution to the number of iteration steps K is one of the main findings of this paper. The intuition of this result is as follows. As Eq. (3.1) reveals, local misspecification can affect the asymptotic distribution of $(\hat{\alpha}_n^K - \alpha^*)$ by three channels. The first channel is the first-step estimator $\hat{\theta}_{f,n}$, the second channel is the sample criterion function Q_n, and the third channel is the $(K-1)$-step CCP estimator $\hat{P}_n^{K-1}$. Our asymptotic arguments reveal that this third channel is the only one that could potentially be affected by the number of iteration steps K. However, the the zero Jacobian property behind Assumption 5(f) implies that the influence of $\hat{P}_n^{K-1}$ is asymptotically reset with every new iteration. As a consequence, the effect of the local misspecification at each iteration step will remain constant regardless of the number of iterations.

[6] In fact, the first equality in Theorem 3.1 shows that all K-step PI estimators are asymptotically equivalent.

4 Applications of the general result

In this section, we apply the main result in Theorem 3.1 to classes of estimators used in practice. Section 4.1 considers maximum likelihood type (ML) estimation, while Section 4.2 considers minimum distance (MD) estimation.

Throughout this section, we presume that the researcher observes an i.i.d. sample distributed according to the true DGP.

Assumption 7. (I.i.d.) $\{(a_i, x_i, x_i')\}_{i \leq n}$ is an i.i.d. sample distributed according to $\Pi_n^*(a, x, x')$.

Under this assumption, it is natural to consider the sample analogue estimators of the true DGP, CCPs, and transition probabilities, given by:

$$\begin{aligned} \hat{\Pi}_n(a, x, x') &\equiv \sum_{i=1}^{n} 1[x_i = x, a_i = a, x_i' = x']/n \quad \forall (a, x, x') \in AX^2, \\ \hat{P}_n(a|x) &\equiv \frac{\sum_{\tilde{x}' \in X} \hat{\Pi}_n(a, x, \tilde{x}')}{\sum_{(\check{a}, \check{x}') \in AX} \hat{\Pi}_n(\check{a}, x, \check{x}')} \quad \forall (a, x) \in AX, \\ \hat{f}_n(x'|a, x) &\equiv \frac{\hat{\Pi}_n(a, x, x')}{\sum_{\tilde{x}' \in X} \hat{\Pi}_n(a, x, \tilde{x}')} \quad \forall (a, x, x') \in AX^2. \end{aligned} \tag{4.1}$$

We now propose a formal framework of the first-step estimator $\hat{\theta}_{f,n}$ and the zero*th*-step CCP estimator $\hat{P}_n^0$. Rather than assuming a specific estimator for these objects, we consider a general framework that covers several popular examples of estimators and that satisfy our high-level Assumption 6.

Assumption 8. (Preliminary estimators) The preliminary first-step and CCP estimators are defined according to the following equation:

$$(\hat{\theta}_{f,n}, \hat{P}_n^0) \;=\; G_n(\hat{\Pi}_n),$$

where $\{G_n : \mathbb{R}^{|AX^2|} \to \mathbb{R}^{d_f} \times \mathbb{R}^{|AX^2|}\}_{n \geq 1}$ is a sequence of functions that satisfies the following properties. For an arbitrary small neighborhood of Π^* denoted by $\mathcal{N}_{\Pi^*}$,

(a) $\sup_{\Pi \in \mathcal{N}_{\Pi^*}} ||G_n(\Pi) - G(\Pi)|| = o_{P_n}(n^{-min\{1/2, \delta\}})$.

(b) $G(\Pi)$ is continuously differentiable for any $\Pi \in \mathcal{N}_{\Pi^*}$,

(c) $(\theta_f^*, P^*) = G(\Pi^*)$.

Assumption 8 is very mild. Under Assumptions 2-3, Theorem 2.1 shows that the true limiting CCPs P^* and limiting parameter θ_f^* are identified. Under these conditions and Assumption 7, it is reasonable to presume that the researcher proposes preliminary estimators of θ_f^* and P^* that are smooth functions of the sample analogue estimator of $\hat{\Pi}_n$.

In fact, given that the state and action spaces are finite, this is automatically satisfied if we use a sample analogue estimator of the CCP, i.e., $\hat{P}_n^0 = \hat{P}_n$, and a non-parametric model for the first-step, i.e., $\theta_f \equiv \{f(x'|a, x)\}_{(a,x,x') \in AX^2}$ that is also estimated by sample analogue estimation, i.e., $\hat{\theta}_{f,n} = \hat{f}_n$. Lemma A.4 in the appendix formally shows that these sample analogue estimators satisfy Assumption 8.[7]

[7] One practical problem with the sample analogue estimators is that they would not be properly defined if $\sum_{x' \in X} \hat{\Pi}_n(a, x, x') = 0$ for any $(a, x) \in AX$. See Hotz et al. (1994) and Pesendorfer and Schmidt-Dengler (2008, Page 914). Of course, this is a small sample problem that disappears asymptotically.

4.1 ML estimation

We now specialize the general results for two-stage K-step PI estimator to the K-ML estimators considered by Aguirregabiria and Mira (2002). This is achieved by setting the sample objective function Q_n to the pseudo-likelihood function, i.e.,

$$Q_n^{ML}(\alpha, \theta_f, P) \equiv n^{-1} \sum_{i=1}^{n} \ln \Psi_{(\alpha,\theta_f)}(P)(a_i|x_i).$$

In this sense, Aguirregabiria and Mira (2002)'s pseudo-likelihood estimator is a special case of the K-step PI estimator. In addition, Aguirregabiria and Mira (2002) show that if the K-step ML estimator converges as $K \to \infty$, it will do so to Rust (1987)'s nested fixed point estimator.

Deriving the asymptotic distribution of the K-ML estimator requires the following regularity condition.

Assumption 9. (Full rank gradient) $\partial \Psi_\theta(P)/\partial \alpha$ is a full rank matrix at (θ^*, P_{θ^*}).

Assumption 9 is the low-level condition connected to the non-singularity requirement in Assumption 5(e). This condition is critical for the consistency of the K-ML estimators and it is also assumed in Aguirregabiria and Mira (2002). Since the α has been assumed to be identified by $\Psi_\theta(P) = P$ (and, thus, locally identified by it), Assumption 9 is equivalent to the regularity conditions in Rothenberg (1971, Theorem 1).

Theorem 4.1 is a corollary of Theorem 3.1 and characterizes the asymptotic distribution of the K-ML estimator under local misspecification.

Theorem 4.1. *Assume Assumptions 1-4 and 7-9. Then, for any* $K, \tilde{K} \geq 1$,

$$\begin{aligned} n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^{K-ML} - \alpha^*) &= n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^{\tilde{K}-ML} - \alpha^*) + o_{P_n}(1) \\ &\xrightarrow{d} \Upsilon_{ML} \times N\left(\begin{pmatrix} B_J \\ B_{\theta_f} \end{pmatrix} \times 1[\delta \leq 1/2] \ , \ \begin{pmatrix} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f} \end{pmatrix} \times 1[\delta \geq 1/2] \right), \end{aligned}$$

where B_J, B_{θ_f}, $\Omega_{J,J}$, Ω_{J,θ_f}, *and* $\Omega_{\theta_f,\theta_f}$ *are as defined in Lemma A.1, and for* Σ *as in Eq. (A.1),* Υ_{ML} *is defined by:*

$$\Upsilon_{ML} \equiv \left(\frac{\partial P_{\theta^*}}{\partial \alpha} (\Sigma \Omega_{J,J} \Sigma')^{-1} \frac{\partial P_{\theta^*}}{\partial \alpha}' \right)^{-1} \frac{\partial P_{\theta^*}}{\partial \alpha} (\Sigma \Omega_{J,J} \Sigma')^{-1} \left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f}' \right].$$

The qualitative conclusions of Theorem 4.1 are as in Theorem 3.1: All K-ML estimators are asymptotically equivalent and thus have the same rate of convergence $n^{\min\{1/2,\delta\}}$ and asymptotic distribution. In quantitative terms, Theorem 4.1 specifies the asymptotic distribution to be normal with mean and variance given by:

$$\begin{aligned} AB_{ML} &= \Upsilon_{ML} \begin{pmatrix} B_J \\ B_{\theta_f} \end{pmatrix} \times 1[\delta \leq 1/2], \\ AV_{ML} &= \Upsilon_{ML} \begin{pmatrix} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f} \end{pmatrix} \Upsilon'_{ML} \times 1[\delta \geq 1/2]. \end{aligned}$$

The presence of (asymptotic) bias and variance depends on the rate of local misspecification δ. In the case of $\delta > 1/2$, the asymptotic distribution has zero bias and the variance coincides exactly with the one obtained

under correct specification. In this case, the local misspecification is irrelevant relative to sampling error and does not affect the asymptotic distribution. The opposite situation occurs when $\delta < 1/2$, as the asymptotic distribution has zero variance and the estimator converges in probability to the bias. In such a case, the local misspecification is overwhelming relative to sampling error and dominates the asymptotic distribution. Finally, we have the knife-edge case with $\delta = 1/2$, in which asymptotic variance and bias coexist. In such a case, an adequate characterization of the precision of the estimator is given by the asymptotic mean squared error, given by:

$$\Upsilon_{ML}\left[\begin{pmatrix} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f} \end{pmatrix} + \begin{pmatrix} B_J \\ B_{\theta_f} \end{pmatrix}\begin{pmatrix} B_J \\ B_{\theta_f} \end{pmatrix}'\right]\Upsilon'_{ML}. \tag{4.2}$$

We conclude the subsection with a comment regarding the asymptotic optimality of the K-ML estimator. The K-ML estimator considered in this section is a "partial" ML estimator in the sense that we are plugging in the first-step estimator into the second stage, effectively ignoring its sample variation. Because of this feature, the usual optimality results for maximum likelihood estimation need not apply. In fact, the next subsection will describe a K-MD estimator that is more efficient than the K-ML, both in asymptotic variance and mean squared error.

4.2 MD estimation

We now specialize the general results for two-stage K-step PI estimator to the K-MD estimators. This is achieved by setting the sample objective function Q_n to the following function:

$$Q_n^{MD}(\alpha, \theta_f, P) \equiv -[\hat{P}_n - \Psi_{(\alpha,\theta_f)}(P)]' \hat{W}_n [\hat{P}_n - \Psi_{(\alpha,\theta_f)}(P)]. \tag{4.3}$$

where $\hat{W}_n \in \mathbb{R}^{|AX|\times|AX|}$ is the weight matrix. In the special case with $K = 1$ and preliminary estimator $\hat{P}_n^0 = \hat{P}_n$, the two step K-MD estimator coincides with the asymptotic least-squares estimator considered in Hotz and Miller (1993) and Pesendorfer and Schmidt-Dengler (2008).[8]

We impose the following condition regarding the weight matrix.[9]

Assumption 10. $\hat{W}_n = W^* + o_{P_n}(1)$, where W^* is positive definite and symmetric.

Theorem 4.2 is a corollary of Theorem 3.1 and characterizes the asymptotic distribution of the K-MD estimator under local misspecification.

Theorem 4.2. *Assume Assumptions 1-4 and 7-10. Then, for any $K, \tilde{K} \geq 1$,*

$$\begin{aligned} n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^{K-MD} - \alpha^*) &= n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^{\tilde{K}-MD} - \alpha^*) + o_{P_n}(1) \\ &\xrightarrow{d} \Upsilon_{MD}(W^*) \times N\left(\begin{pmatrix} B_J \\ B_{\theta_f} \end{pmatrix} \times 1[\delta \leq 1/2] \ , \ \begin{pmatrix} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f} \end{pmatrix} \times 1[\delta \geq 1/2]\right), \end{aligned}$$

where B_J, B_{θ_f}, $\Omega_{J,J}$, Ω_{J,θ_f}, and $\Omega_{\theta_f,\theta_f}$ are as defined in Lemma A.1, and for Σ as in Eq. (A.1), $\Upsilon_{MD}(W^*)$

[8] In all fairness, Pesendorfer and Schmidt-Dengler (2008, Eqs. (18)-(19)) allow for $\hat{P}_n$ in Eq. (4.3) to differ from the sample frequency estimator. Nonetheless, this observation could also apply to our K-MD estimation framework.

[9] In principle, we could allow for the weight matrix to be functions of the parameters of the problem. In such cases, one could obtain the same results by imposing additional conditions in Assumption 10 and by using slightly longer theoretical arguments.

is defined by:

$$\Upsilon_{MD}(W^*) \equiv \left(\frac{\partial P_{\theta^*}}{\partial \alpha} W^* \frac{\partial P_{\theta^*}}{\partial \alpha}'\right)^{-1} \frac{\partial P_{\theta^*}}{\partial \alpha} W^* \times \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta^*}}{\partial \theta_f}' \end{array}\right].$$

Remark 4.1. The asymptotic distribution of the K-ML estimator is a special case of that of the K-MD estimator with $W^* = W_{ML} \equiv (\Sigma \Omega_{J,J} \Sigma')^{-1}$.

As in the previous section, the qualitative conclusions of Theorem 4.2 are as in Theorem 3.1: Given the asymptotic weight matrix, all K-MD estimators are asymptotically equivalent and thus have the same rate of convergence $n^{\min\{1/2,\delta\}}$ and asymptotic distribution. In quantitative terms, Theorem 4.2 shows that the asymptotic distribution of the K-MD estimator is normal with mean and variance given by:

$$AB_{MD}(W^*) = \Upsilon_{MD}(W^*) \left(\begin{array}{c} B_J \\ B_{\theta_f} \end{array}\right) \times 1[\delta \leq 1/2], \tag{4.4}$$

$$AV_{MD}(W^*) = \Upsilon_{MD}(W^*) \left(\begin{array}{cc} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f} \end{array}\right) \Upsilon_{MD}(W^*)' \times 1[\delta \geq 1/2]. \tag{4.5}$$

The presence of (asymptotic) bias and variance depends on the rate of local misspecification δ in the same way as with K-ML estimator. In the knife-edge case with $\delta = 1/2$, the asymptotic variance and bias coexist, and the asymptotic mean-squared error is given by:

$$\Upsilon_{MD}(W^*) \left[\left(\begin{array}{cc} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f} \end{array}\right) + \left(\begin{array}{c} B_J \\ B_{\theta_f} \end{array}\right)\left(\begin{array}{c} B_J \\ B_{\theta_f} \end{array}\right)'\right] \Upsilon_{MD}(W^*)'. \tag{4.6}$$

We can now briefly discuss the optimality in the choice of W^* in K-MD estimation. Since these estimators can have bias, variance, and both, we deem the asymptotic mean squared error to be a reasonable optimality criterion. First, consider the case in which local misspecification is asymptotically irrelevant, i.e., $\delta > 1/2$. In this case, the K-MD estimator presents no asymptotic bias and the asymptotic variance and mean squared error are both equal to Eq. (4.5). By standard argument in GMM estimation (e.g. McFadden and Newey (1994, Page 2165)), the minimum asymptotic variance and mean squared error (in matrix sense) among K-MD estimators are given by:

$$\left(\frac{\partial P_{\theta^*}}{\partial \alpha}\left[\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta^*}}{\partial \theta_f}' \end{array}\right]\left(\begin{array}{cc} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f} \end{array}\right)\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta^*}}{\partial \theta_f}' \end{array}\right]'\right]^{-1} \frac{\partial P_{\theta^*}}{\partial \alpha}'\right)^{-1}.$$

This minimum can be achieved by the following (feasible) choice of W^*:

$$W^*_{AV} \equiv \left[\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta^*}}{\partial \theta_f}' \end{array}\right]\left(\begin{array}{cc} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f} \end{array}\right)\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta^*}}{\partial \theta_f}' \end{array}\right]'\right]^{-1}. \tag{4.7}$$

We say that Eq. (4.7) is feasible because can be consistently estimated by the researcher. As pointed out in Remark 4.1, the K-ML estimator the same asymptotic distribution as the K-MD estimator with $W^*_{ML} =$

$(\Sigma\Omega_{J,J}\Sigma')^{-1}$. Unless there are special conditions on the econometric model (e.g. $\partial P_{\theta^*}/\partial\theta_f = \mathbf{0}_{|\tilde{A}X|\times d_f}$), the K-ML is not necessarily optimal among K-MD estimators in the sense of achieving a minimum variance.

Next, consider the knife-edge case in which local misspecification is of the same rate as sampling error, i.e., $\delta = 1/2$. Once again, standard arguments imply that the minimum asymptotic mean squared error (in matrix sense) among all K-MD estimators is given by:

$$\left(\frac{\partial P_{\theta^*}}{\partial\alpha}\left[\left[\begin{array}{cc}\Sigma & -\frac{\partial P_{\theta^*}}{\partial\theta_f}'\end{array}\right]\left[\left(\begin{array}{cc}\Omega_{J,J} & \Omega_{J,\theta_f}\\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f}\end{array}\right)+\left(\begin{array}{c}B_J\\ B_{\theta_f}\end{array}\right)\left(\begin{array}{c}B_J\\ B_{\theta_f}\end{array}\right)'\right]\left[\begin{array}{cc}\Sigma & -\frac{\partial P_{\theta^*}}{\partial\theta_f}'\end{array}\right]'\right]^{-1}\frac{\partial P_{\theta^*}}{\partial\alpha}'\right)^{-1}.$$

This minimum can be achieved by the following (infeasible) choice of W^*:

$$W^*_{AMSE} \equiv \left[\left[\begin{array}{cc}\Sigma & -\frac{\partial P_{\theta^*}}{\partial\theta_f}'\end{array}\right]\left[\left(\begin{array}{cc}\Omega_{J,J} & \Omega_{J,\theta_f}\\ \Omega'_{J,\theta_f} & \Omega_{\theta_f,\theta_f}\end{array}\right)+\left(\begin{array}{c}B_J\\ B_{\theta_f}\end{array}\right)\left(\begin{array}{c}B_J\\ B_{\theta_f}\end{array}\right)'\right]\left[\begin{array}{cc}\Sigma & -\frac{\partial P_{\theta^*}}{\partial\theta_f}'\end{array}\right]'\right]^{-1}. \quad (4.8)$$

We say that Eq. (4.8) is infeasible because it depends on the asymptotic bias which is a population feature that the researcher is is unaware of. However, it is relevant to point out that the presence of asymptotic bias might generate a situation in which a feasible choice of weight matrix that minimizes asymptotic variance could result in a large asymptotic bias and, consequently, asymptotic mean squared error.[10]

Finally, we could consider the case in which local misspecification is asymptotically overwhelming, i.e., $\delta < 1/2$. In this case, the K-MD estimator presents no asymptotic variance and it converges at a pure bias term equal to Eq. (4.4). By definition, the asymptotic squared bias coincides with the asymptotic mean squared error. As in the previous case, minimizing asymptotic mean squared error is infeasible in the sense that it depends on the unknown bias.

While the weight matrix that minimizes asymptotic mean squared error in Eq. (4.8) is infeasible, one might wonder whether there are alternative feasible weight matrices that can achieve this minimum. Lemma A.7 in the Appendix provides the necessary and sufficient condition under which W^* is a minimizer of asymptotic mean squared error (W^*_{AMSE} is just a special case). Unless local misspecification is irrelevant (i.e. $\delta > 1/2$), minimizing asymptotic mean squared error is always infeasible in the sense that it depends on the unknown bias.

5 Monte Carlo simulations

This section investigates the finite sample performance of the two-stage estimators considered in previous sections under local misspecification. We simulate data using the classical bus engine replacement problem studied by Rust (1987).

5.1 A misspecified econometric model

In each period $t = 1, \ldots, T \equiv \infty$, the bus owner has to decide whether to replace the bus engine or not to minimize the discounted present value of his costs. In any representative period, his choice is denoted by

[10] This is the case in some of our Monte Carlo simulations, in which using a weight matrix that consistently estimates W^*_{AV} produces an asymptotic mean squared error that is larger than that the one obtained by using $\hat{W}_n = \mathbf{I}$.

$a \in A = \{1, 2\}$, where $a = 2$ represents replacing the engine and $a = 1$ represents not replacing the engine, and the current engine mileage is denoted by $x \in X \equiv \{1, \ldots, 20\}$.

The researcher assumes the following specification for the deterministic part of the utility (profit) function:

$$u_{\theta_u}(x, a) \;=\; -\theta_{u,1} \cdot 1[a = 2] \; -\theta_{u,2} \cdot 1[a = 1]x, \tag{5.1}$$

where $\theta_u \equiv (\theta_{u,1}, \theta_{u,2}) \in \Theta_u \equiv [-B, B]^2$ with $B = 10$. In addition, the researcher also assumes that the unobserved errors are distributed according to an extreme value type I distribution, independent of x, i.e.,

$$g(\epsilon = e|x) = \prod_{a \in A} \exp(e(a)) \exp(-\exp(e(a))), \tag{5.2}$$

which does not have unknown parameters. Finally, the observed state is assumed to evolve according to the following Markov chain:

$$f_{\theta_f}(x'|x, a) \;=\; (1 - \theta_f) \cdot 1[a = 1, x' = \min\{x + 1, |X|\}] \;+\; \theta_f \cdot 1[a = 1, x' = x] \;+\; 1[a = 2, x' = 1], \tag{5.3}$$

where $\theta_f \in \Theta_f \equiv [0, 1]$. The researcher correctly assumes that $\beta = 0.9999$. His goal is to estimate $\theta = (\alpha, \theta_f) \in \Theta = \Theta_\alpha \times \Theta_f$ with $\alpha \equiv \theta_u \in \Theta_\alpha = \Theta_u$ and $\theta_f \in \Theta_f$.

The researcher has correctly specified the error distribution and state transition probabilities, which satisfy Eq. (5.3) with $\theta_f = 0.25$. Unfortunately, he has misspecified the utility function. Instead of the linear function in Eq. (5.1), the utility function is the following quadratic function:

$$u_{\theta_u,n}(x, a) \;=\; -\theta_{u,1} \cdot 1[a = 2] \; -\theta_{u,2} \cdot 1[a = 1]x \; -\theta_{u,3,n} \cdot 1[a = 1]x^2 \tag{5.4}$$

with $\theta_{u,1} = 1$, $\theta_{u,2} = 0.05$, and $\theta_{u,3,n} = 0.025n^{-\delta}$ with $\delta \in \{1/4, 1/3, 1/2, 1\}$. The fact that $\{\theta_{u,3,n} n^\delta\}_{n \geq 1}$ is a constant sequence implies that the econometric model is misspecified in the sense of Assumption 3. Our choices of δ include a case in which the local misspecification is asymptotically irrelevant (i.e. $\delta = 1$), one case in which local misspecification is the knife-edge case (i.e. $\delta = 1/2$), and one case in which the local misspecification is overwhelming (i.e. $\delta \in 1/3$). For the sake of completeness, we also consider a case in which the econometric model is correctly specified, i.e., $\theta_{u,3,n} = 0$.

By the arguments in Section 2, the true CCPs P_n^* is determined by the true error distribution (Eq. (5.2)), true state transition probabilities (Eq. (5.3) with $\theta_f = 0.25$), and the true utility function (Eq. (5.4) with $\theta_{u,n} = (1, 0.05, 0.025n^{-\delta})$). We generate marginal observations of the state variables according to the following distribution:

$$m_n^*(x) \;\propto\; 1 + \log(x).^{11}$$

In combination with previous elements, this determines the true joint DGP Π_n^* according to Eq. (2.7).

Our simulation results will be the average of $S = 20,000$ independent datasets of observations $\{(a_i, x_i, x_i')\}_{i \leq n}$ that are i.i.d. distributed according to Π_n^*. We present simulation results for sample sizes of $n \in \{200,\ 500,\ 1,000\}$.

[11] Recall from Section 2 that this aspect of the model is left unspecified by the researcher.

5.2 Estimation

Given any sample of observations $\{(a_i, x_i, x_i')\}_{i \leq n}$, the researcher estimates the parameters of interest $\theta = (\theta_{u,1}, \theta_{u,2}, \theta_f)$ using a two-stage K-step policy function iteration estimator described in Sections 3-4.

- In the first stage, the researcher estimates θ_f using the following (consistent) estimator:

$$\hat{\theta}_{f,n} \equiv \frac{\sum_{i=1}^{n} 1[a_i = 1, x_i' = x_i, x_i \neq |X|]}{\sum_{i=1}^{n} 1[a_i = 1, x_i \neq |X|]}. \tag{5.5}$$

- In the second stage, the researcher estimates $(\theta_{u,1}, \theta_{u,2})$ using the K-step policy function iteration estimator in Eq. (3.1). In particular, he computes the policy function mapping Ψ_θ as in Eq. (2.6), and solves the estimation problem in Eq. (3.1) with:
  - The zero*th*-step CCP estimator $\hat{P}_n^0$ is set to be the sample analogue estimator $\hat{P}_n$.
  - The number of steps used is $K \in \{1, 2, 3, 10\}$.[12]
  - The criterion function Q_n is equal to: (a) pseudo-likelihood function Q_n^{ML} in Section 4.1 and (b) the weighted minimum distance function Q_n^{MD} in Section 4.2 with three choices of limiting weight matrix W^*: identity matrix (i.e. $W^* = \mathbf{I}_{|AX|}$) and asymptotic variance minimizer (i.e. $W^* = W_{AV}^*$ as in Eq. (4.7)).[13]

5.3 Results

We used our simulations to investigate the finite sample behavior of the estimators for $\theta_u = (\theta_{u,1}, \theta_{u,2})$. For reasons of brevity, we focus on the main text on the coefficient $\theta_{u,2}$ as we expect it to be more affected by the misspecification of the utility function with respect to $x \in X$.[14]

Table 1 describes results under correct specification. When scaled by $\sqrt{n}$, all estimators appear to converge to a distribution with zero mean and finite variance. As expected, the number of iterations K does not appear to affect the bias or variance of the estimators under consideration. In addition, the optimal MD and ML estimators are similar and more efficient than the MD estimator with identity weight matrix.

Table 2 provides results under asymptotically irrelevant local misspecification, i.e., $\delta = 1$. According to our theoretical results, the asymptotic behavior of all estimators should identical to the correctly specified model. These predictions are confirmed in our simulations, as Tables 1 and 2 are virtually identical.

Table 3 provides results under local misspecification that vanishes at the knife-edge rate, i.e., $\delta = 1/2$. According to our theoretical results, this should result in an asymptotic distribution that has non-zero bias and is not affected by the number of iterations K. By and large, this result is confirmed in our simulations. Given the presence of asymptotic bias, we now use the mean squared error to evaluate the efficiency. In this simulation design, the MD estimator with identity weight matrix is now also slightly more efficient than the ML and minimum variance MD estimator. While the MD estimator with identity matrix has more variance than the ML estimator it appears to have less bias, resulting in less mean squared error.

[12] In accordance to our asymptotic theory, the simulation results with $K \in \{4, \ldots, 9\}$ are almost identical to those with $K \in \{3, 10\}$. These were eliminated from the paper for reasons of brevity but they are available from the authors upon request.

[13] These were approximated using Monte Carlo integration and numerical derivates with a sample size that is significantly larger than those used in the actual Monte Carlo simulations.

[14] The results for $\theta_{u,1}$ are qualitatively similar and are thus omitted for reasons of brevity. They are available from the authors, upon request.

Steps	Statistic	MD($\mathbf{I}$)			MD(W^*_{AV})			ML		
		$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$
$K = 1$	$\sqrt{n}$·Bias	0.04	0.02	0.01	0.04	0.02	0.01	0.04	0.02	0.01
	$\sqrt{n}$·SD	0.27	0.25	0.24	0.25	0.23	0.22	0.23	0.22	0.22
	n·MSE	0.07	0.06	0.06	0.06	0.05	0.05	0.06	0.05	0.05
$K = 2$	$\sqrt{n}$·Bias	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00
	$\sqrt{n}$·SD	0.26	0.25	0.24	0.24	0.23	0.22	0.22	0.22	0.22
	n·MSE	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05
$K = 3$	$\sqrt{n}$·Bias	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00
	$\sqrt{n}$·SD	0.26	0.25	0.24	0.24	0.23	0.22	0.22	0.22	0.22
	n·MSE	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05
$K = 10$	$\sqrt{n}$·Bias	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00
	$\sqrt{n}$·SD	0.26	0.25	0.24	0.24	0.23	0.22	0.22	0.22	0.22
	n·MSE	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05

Table 1: Monte Carlo results for $\theta_{u,2}$ under correct specification.

Steps	Statistic	MD($\mathbf{I}$)			MD(W^*_{AV})			ML		
		$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$
$K = 1$	$\sqrt{n}$·Bias	0.08	0.04	0.03	0.08	0.04	0.03	0.08	0.04	0.03
	$\sqrt{n}$·SD	0.27	0.25	0.24	0.25	0.23	0.22	0.24	0.23	0.22
	n·MSE	0.08	0.07	0.06	0.07	0.06	0.05	0.06	0.05	0.05
$K = 2$	$\sqrt{n}$·Bias	0.04	0.03	0.02	0.05	0.03	0.02	0.05	0.03	0.02
	$\sqrt{n}$·SD	0.27	0.25	0.24	0.24	0.23	0.22	0.23	0.22	0.22
	n·MSE	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05
$K = 3$	$\sqrt{n}$·Bias	0.04	0.03	0.02	0.05	0.03	0.02	0.05	0.03	0.02
	$\sqrt{n}$·SD	0.26	0.25	0.24	0.24	0.23	0.22	0.23	0.22	0.22
	n·MSE	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05
$K = 10$	$\sqrt{n}$·Bias	0.04	0.03	0.02	0.05	0.03	0.02	0.05	0.03	0.02
	$\sqrt{n}$·SD	0.26	0.25	0.24	0.24	0.23	0.22	0.23	0.22	0.22
	n·MSE	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05

Table 2: Monte Carlo results for $\theta_{u,2}$ under local specification with $\theta_{u,3} \propto n^{-1}$.

Tables 4 and 5 provide results under asymptotically overwhelming local misspecification, i.e., $\delta = 1/3$. According to our theoretical results, the presence of this local misspecification dramatically changes the asymptotic distribution of all estimators under consideration. In particular, these no longer converge at the regular $\sqrt{n}$-rate. In fact, at the said rate, the asymptotic bias is no longer bounded as Table 4 suggests. Once we scale the estimators appropriately (at the $n^{1/3}$-rate) all estimators under consideration should converge to an asymptotic distribution that is dominated by bias. Furthermore, our theoretical results indicate that the number of iterations K does not affect this asymptotic distribution. These two facts are clearly depicted in Table 5. In line with previous results, the MD estimator with identity weight matrix is slightly more efficient than the ML and minimum variance MD estimator.

6 Conclusion

Dynamic discrete choice models are typically estimated using heavily parametrized econometric frameworks, making them susceptible to model misspecification. This paper investigates how misspecification can affect the results of inference in these models. To keep the analysis tractable, this paper focuses on *local misspecification*, which is an asymptotic device in which the mistake in the specification vanishes as the sample size

Steps	Statistic	MD($\mathbf{I}$)			MD(W^*_{AV})			ML		
		$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$
$K = 1$	$\sqrt{n}$·Bias	0.53	0.51	0.51	0.56	0.54	0.54	0.57	0.55	0.55
	$\sqrt{n}$·SD	0.34	0.29	0.27	0.31	0.26	0.25	0.28	0.25	0.24
	n·MSE	0.39	0.34	0.33	0.41	0.36	0.36	0.40	0.36	0.36
$K = 2$	$\sqrt{n}$·Bias	0.49	0.49	0.50	0.52	0.53	0.54	0.54	0.54	0.54
	$\sqrt{n}$·SD	0.33	0.28	0.27	0.30	0.26	0.25	0.27	0.25	0.24
	n·MSE	0.35	0.32	0.32	0.36	0.35	0.35	0.36	0.35	0.35
$K = 3$	$\sqrt{n}$·Bias	0.49	0.49	0.50	0.52	0.53	0.54	0.54	0.54	0.54
	$\sqrt{n}$·SD	0.33	0.28	0.27	0.30	0.26	0.25	0.27	0.25	0.24
	n·MSE	0.35	0.32	0.32	0.36	0.35	0.35	0.36	0.35	0.35
$K = 10$	$\sqrt{n}$·Bias	0.49	0.49	0.50	0.52	0.53	0.54	0.54	0.54	0.54
	$\sqrt{n}$·SD	0.33	0.28	0.27	0.30	0.26	0.25	0.27	0.25	0.24
	n·MSE	0.35	0.32	0.32	0.36	0.35	0.35	0.36	0.35	0.35

Table 3: Monte Carlo results for $\theta_{u,2}$ under local specification with $\theta_{u,3} \propto n^{-1/2}$.

Steps	Statistic	MD($\mathbf{I}$)			MD(W^*_{AV})			ML		
		$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$
$K = 1$	$\sqrt{n}$·Bias	1.10	1.26	1.44	1.18	1.36	1.56	1.24	1.42	1.61
	$\sqrt{n}$·SD	0.46	0.37	0.33	0.41	0.32	0.29	0.36	0.31	0.29
	n·MSE	1.43	1.73	2.19	1.55	1.95	2.50	1.67	2.13	2.69
$K = 2$	$\sqrt{n}$·Bias	1.06	1.25	1.44	1.14	1.35	1.55	1.21	1.41	1.61
	$\sqrt{n}$·SD	0.45	0.36	0.32	0.40	0.32	0.29	0.35	0.31	0.28
	n·MSE	1.32	1.70	2.18	1.46	1.92	2.49	1.58	2.09	2.66
$K = 3$	$\sqrt{n}$·Bias	1.06	1.25	1.44	1.14	1.35	1.55	1.21	1.41	1.61
	$\sqrt{n}$·SD	0.45	0.36	0.32	0.40	0.32	0.29	0.35	0.31	0.28
	n·MSE	1.32	1.70	2.18	1.45	1.92	2.49	1.58	2.09	2.66
$K = 10$	$\sqrt{n}$·Bias	1.06	1.25	1.44	1.14	1.35	1.55	1.21	1.41	1.61
	$\sqrt{n}$·SD	0.45	0.36	0.32	0.40	0.32	0.29	0.35	0.31	0.28
	n·MSE	1.32	1.70	2.18	1.45	1.92	2.49	1.58	2.09	2.66

Table 4: Monte Carlo results for $\theta_{u,2}$ under local specification with $\theta_{u,3} \propto n^{-1/3}$ using the regular scaling.

Steps	Statistic	MD($\mathbf{I}$)			MD(W^*_{AV})			ML		
		$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$	$n = 200$	$n = 500$	$n = 1,000$
$K = 1$	$n^{1/3}$·Bias	0.46	0.45	0.46	0.49	0.48	0.49	0.51	0.51	0.51
	$n^{1/3}$·SD	0.19	0.13	0.10	0.17	0.11	0.09	0.15	0.11	0.09
	$n^{2/3}$·MSE	0.24	0.22	0.22	0.27	0.25	0.25	0.29	0.27	0.27
$K = 2$	$n^{1/3}$·Bias	0.44	0.44	0.46	0.47	0.48	0.49	0.50	0.50	0.51
	$n^{1/3}$·SD	0.19	0.13	0.10	0.17	0.11	0.09	0.14	0.11	0.09
	$n^{2/3}$·MSE	0.23	0.21	0.22	0.25	0.24	0.25	0.27	0.26	0.27
$K = 3$	$n^{1/3}$·Bias	0.44	0.44	0.46	0.47	0.48	0.49	0.50	0.50	0.51
	$n^{1/3}$·SD	0.19	0.13	0.10	0.16	0.11	0.09	0.14	0.11	0.09
	$n^{2/3}$·MSE	0.23	0.21	0.22	0.25	0.24	0.25	0.27	0.26	0.27
$K = 10$	$n^{1/3}$·Bias	0.44	0.44	0.46	0.47	0.48	0.49	0.50	0.50	0.51
	$n^{1/3}$·SD	0.19	0.13	0.10	0.16	0.11	0.09	0.14	0.11	0.09
	$n^{2/3}$·MSE	0.23	0.21	0.22	0.25	0.24	0.25	0.27	0.26	0.27

Table 5: Monte Carlo results for $\theta_{u,2}$ under local specification with $\theta_{u,3} \propto n^{-1/3}$ using the correct scaling.

diverges. This device allows us to approximate the effects of small amounts of misspecification in a manner that is amenable to asymptotic analysis. However, we impose no restrictions on the rate at which these errors vanish.

We consider a general class of two-stage estimators based on the K-step sequential policy function iteration algorithm, where K denotes the number of iterations employed in the estimation. By appropriate choice of the criterion function, this class captures the K-step maximum likelihood estimators (K-ML) and the K-step minimum distance estimators (K-MD). Special cases of our framework are Rust (1987)'s nested fixed point estimator, Hotz and Miller (1993)'s conditional choice probability estimator, Aguirregabiria and Mira (2002)'s pseudo-likelihood estimator, and Pesendorfer and Schmidt-Dengler (2008)'s asymptotic least squares estimator.

We establish that local misspecification can affect asymptotic bias, asymptotic variance, and even the rate of convergence of these estimators. However, our main finding is that the effect of the local misspecification is invariant to the number of iterations K. In particular, (a) all K-ML estimators are asymptotically equivalent for all values of K and (b) given the choice of weight matrix, all K-MD estimators are asymptotically equivalent for all values of K. In practice, this means that the choice of the number of iterations K should not be guided by concerns of model misspecification.

Under correct specification, the comparison between K-MD and K-ML estimators in terms of asymptotic mean squared error yields a clear-cut recommendation. Under local misspecification, this is no longer the case. In particular, local misspecification can introduce an unknown asymptotic bias, which greatly complicates the comparison. In particular, an optimally weighted K-MD estimator depends on the unknown asymptotic bias and is thus unfeasible. In turn, feasible K-MD estimators could have an asymptotic mean squared error that is higher or lower than that of the K-ML estimators.

A Appendix

A.1 Additional notation

Throughout this appendix, "s.t." abbreviates "such that", and "RHS" and " LHS" abbreviate "right hand side" and "left hand side", respectively. Furthermore, "LLN" refers to the strong law of large numbers, "CLT" refers to the central limit theorem, and "CMT" refers to the continuous mapping theorem.

Given a DGP Π_n^*, Eq. (2.8) defines transition probabilities f_n^*, CCPs P_n^*, and marginal probability distribution of states m_n^*. The probability of actions and states J_n^* can also be defined analogously:

$$J_n^*(a,x) \quad \equiv \quad \sum_{\tilde{x}' \in X} \Pi_n^*(a,x,\tilde{x}') \quad \forall (a,x) \in AX.$$

The limiting DGP f^*, transition probabilities f^*, CCPs P^* are defined in Assumption 3. The limiting probability of actions and states J^* and marginal probability distribution of states m^* can also be defined analogously:

$$\begin{aligned} J^*(a,x) \quad &\equiv \quad \sum_{\tilde{x}' \in X} \Pi^*(a,x,\tilde{x}') \quad \forall (a,x) \in AX, \\ m^*(x) \quad &\equiv \quad \sum_{(\tilde{a},\tilde{x}') \in AX} \Pi^*(\tilde{a},x,\tilde{x}') \quad \forall x \in X, \end{aligned}$$

The sample analogue DGP $\hat{\Pi}_n$, transition probabilities $\hat{f}_n$, CCPs $\hat{P}_n$ were defined in Eq. (4.1). The sample analogue probability of actions and states $\hat{J}_n$ and marginal probability distribution of states $\hat{m}_n$ can also be defined

analogously:

$$\begin{aligned}\hat{J}_n(a,x) &\equiv \sum_{\tilde{x}' \in X} \hat{\Pi}_n(a,x,\tilde{x}') \quad \forall (a,x) \in AX, \\ \hat{m}_n(x) &\equiv \sum_{(\tilde{a},\tilde{x})' \in AX} \hat{\Pi}_n(\tilde{a},x,\tilde{x}') \quad \forall x \in X.\end{aligned}$$

Finally, we now define a matrix $\Sigma \in \mathbb{R}^{|AX|\times|\tilde{A}X|}$ that appears repeatedly in our formal arguments and given by:

$$\Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0}_{|\tilde{A}|\times|A|} & \cdots & \mathbf{0}_{|\tilde{A}|\times|A|} \\ \mathbf{0}_{|\tilde{A}|\times|A|} & \Sigma_2 & \mathbf{0}_{|\tilde{A}|\times|A|} & \mathbf{0}_{|\tilde{A}|\times|A|} \\ \vdots & \mathbf{0}_{|\tilde{A}|\times|A|} & \ddots & \mathbf{0}_{|\tilde{A}|\times|A|} \\ \mathbf{0}_{|\tilde{A}|\times|A|} & \mathbf{0}_{|\tilde{A}|\times|A|} & \mathbf{0}_{|\tilde{A}|\times|A|} & \Sigma_{|X|} \end{bmatrix}, \tag{A.1}$$

where

$$\Sigma_x \equiv \frac{\mathbf{I}_{|\tilde{A}|\times A} - \{P^*(\tilde{a}|x)\}_{\tilde{a}\in\tilde{A}}\mathbf{1}_{1\times A}}{\sum_{a\in A} J^*(a,x)} \quad \forall x \in X. \tag{A.2}$$

A.2 Proofs of theorems

Proof of Theorem 2.1. Since $\Theta = \Theta_\alpha \times \Theta_f$ is compact and $||(P_{(\alpha,\theta_f)} - P_n^*), (f_{\theta_f} - f_n^*)||$ is a continuous function of (α, θ_f), the arguments in Royden (1988, pages 193-195) implies that $\exists(\alpha^*, \theta_f^*) \in \Theta$ that minimizes $||(P_{(\alpha,\theta_f)} - P_n^*), (f_{\theta_f} - f_n^*)||$. By Assumption 3(b), this minimum value is zero, i.e., $\exists(\alpha^*, \theta_f^*) \in \Theta$ s.t. $||(P_{(\alpha^*,\theta_f^*)} - P^*), (f_{\theta_f^*} - f^*)|| = 0$ or, equivalently, $P_{(\alpha^*,\theta_f^*)} = P^*$ and $f_{\theta_f^*} = f^*$.

Now suppose that this also occurs for $(\tilde{\theta}_f, \tilde{\alpha}) \in \Theta$. We now show that $(\theta_f^*, \alpha^*) = (\tilde{\theta}_f, \tilde{\alpha})$. By triangle inequality $||f_{\theta_f^*} - f_{\tilde{\theta}_f}|| \leq ||f_{\theta_f^*} - f^*|| + ||f_{\tilde{\theta}_f} - f^*||$ and since θ_f^* and $\tilde{\theta}_f$ both satisfy $||f_{\theta_f} - f^*|| = 0$, we conclude that $||f_{\theta_f^*} - f_{\tilde{\theta}_f}|| = 0$ and so $f_{\theta_f^*} = f_{\tilde{\theta}_f}$. By Assumption 2, this implies that $\theta_f^* = \tilde{\theta}_f$. Using this and by repeating the previous argument with $P_{(\alpha,\theta_f^*)}$ instead of f_{θ_f}, we conclude that $\alpha^* = \tilde{\alpha}$. □

Proof of Theorem 3.1. Throughout this proof, let $\mathcal{N}_\alpha$ denote an arbitrarily neighborhood of α^* that results from projecting $\mathcal{N}$ onto its α-coordinate.

Part 1. Fix $K \geq 1$ arbitrarily. We prove the result by assuming that:

$$n^{\min\{\delta,1/2\}}(\hat{P}_n^{K-1} - P^*) = O_{P_n}(1). \tag{A.3}$$

By definition, $\hat{\alpha}_n^K = \arg\max_{\alpha\in\Theta_\alpha} Q_n(\alpha)$ with $Q_n(\alpha) \equiv Q_n(\alpha, \hat{\theta}_{f,n}, \hat{P}_n^{K-1})$ and $\hat{P}_n^{K-1} \equiv \Psi_{(\hat{\alpha}_n^{K-1}, \hat{\theta}_{f,n})}(\hat{P}_n^{K-2})$ for $K > 1$ and $\hat{P}_n^{K-1} \equiv \hat{P}_n^0$ for $K = 1$. Provided that we check its conditions, the result follows from Theorem A.2.

Under Assumptions 5(a)-(c) and 6, Theorem A.1 implies that $\hat{\alpha}_n^K = \alpha^* + o_{P_n}(1)$, i.e., condition (a) in Theorem A.2. Assumption 4 is condition (b) in Theorem A.2.

Assumption 6(a), Eq. (A.3) imply that $(\alpha^*, \hat{\theta}_{f,n}, \hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1. In turn, these and $\hat{\alpha}_n^K = \alpha^* + o_{P_n}(1)$ imply that $(\hat{\alpha}_n^K, \hat{\theta}_{f,n}, \hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1. These results will be used repeatedly throughout the rest of this proof.

Assumption 5(c) and $(\hat{\alpha}_n^K, \hat{\theta}_{f,n}, \hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1 implies condition (c) in Theorem A.2.

We now verify condition (d) in Theorem A.2. Assumptions 5(d)-(f) and $(\hat{\alpha}_n^K, \hat{\theta}_{f,n}, \hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1 imply that

the following derivation holds w.p.a.1.

$$
\begin{aligned}
&n^{\min\{\delta,1/2\}}\frac{\partial Q_n(\alpha^*)}{\partial \alpha} = n^{\min\{\delta,1/2\}}\frac{\partial Q_n(\alpha^*,\hat{\theta}_{f,n},\hat{P}_n^{K-1})}{\partial \alpha} \\
&= n^{\min\{\delta,1/2\}}\frac{\partial Q_n(\alpha^*,\theta_f^*,P^*)}{\partial \alpha} + \frac{\partial^2 Q_n(\alpha^*,\tilde{\theta}_{f,n},\tilde{P}_n)}{\partial \alpha \partial \theta_f'} n^{\min\{\delta,1/2\}}(\hat{\theta}_{f,n}-\theta_f^*) + \frac{\partial^2 Q_n(\alpha^*,\tilde{\theta}_{f,n},\tilde{P}_n)}{\partial \alpha \partial P'} n^{\min\{\delta,1/2\}}(\hat{P}_n^{K-1}-P^*) \\
&= n^{\min\{\delta,1/2\}}[\frac{\partial Q_n(\alpha^*,\theta_f^*,P^*)}{\partial \alpha} + \frac{\partial^2 Q_\infty(\alpha^*,\theta_f^*,P^*)}{\partial \alpha \partial \theta_f'}(\hat{\theta}_{f,n}-\theta_f^*)] + o_{P_n}(1),
\end{aligned}
$$

where $(\tilde{\theta}_{f,n},\tilde{P}_n)$ is some sequence between $(\hat{\theta}_{f,n},\hat{P}_n^{K-1})$ and (θ_f^*,P^*). From this and Assumption 6(a),

$$
n^{\min\{\delta,1/2\}}\frac{\partial Q_n(\alpha^*)}{\partial \alpha} \xrightarrow{d} \zeta_1 + \frac{\partial^2 Q_\infty(\alpha^*,\theta_f^*,P^*)}{\partial \alpha \partial \theta_f'}\zeta_2.
$$

By denoting the RHS random variable Z, this is exactly condition (d) in Theorem A.2.

Consider any arbitrary $\alpha \in \mathcal{N}_\alpha$ and so $(\alpha,\hat{\theta}_{f,n},\hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1. By this and Assumptions 5(c)-(e) imply that the following derivation holds w.p.a.1.

$$
\frac{\partial^2 Q_n(\alpha)}{\partial \alpha \partial \alpha'} = \frac{\partial^2 Q_n(\alpha,\hat{\theta}_{f,n},\hat{P}_n^{K-1})}{\partial \alpha \partial \alpha'} = \frac{\partial^2 Q_\infty(\alpha,\hat{\theta}_{f_n},\hat{P}_n^{K-1})}{\partial \alpha \partial \alpha'} + o_{P_n}(1) = \frac{\partial^2 Q_\infty(\alpha,\theta_f^*,P^*)}{\partial \alpha \partial \alpha'} + o_{P_n}(1),
$$

where convergence is uniform in $\alpha \in \mathcal{N}_\alpha$. By denoting $H(\alpha)$ the first expression on the RHS, this verifies conditions (e)-(f) in Theorem A.2.

Theorem A.2 then implies that:

$$
n^{\min\{\delta,1/2\}}(\hat{\alpha}_n^K-\alpha^*) = A_1 n^{\min\{\delta,1/2\}}\frac{\partial Q_n(\alpha^*,\theta_f^*,P^*)}{\partial \alpha'} + A_2 n^{\min\{\delta,1/2\}}(\hat{\theta}_{f,n}-\theta_f^*) + o_{P_n}(1), \tag{A.4}
$$

with

$$
A_1 \equiv (\frac{\partial^2 Q_\infty(\alpha,\theta_f^*,P^*)}{\partial \alpha \partial \alpha'})^{-1} \quad \text{and} \quad A_2 \equiv (\frac{\partial^2 Q_\infty(\alpha,\theta_f^*,P^*)}{\partial \alpha \partial \alpha'})^{-1}\frac{\partial^2 Q_\infty(\alpha^*,\theta_f^*,P^*)}{\partial \alpha \partial \theta_f'}.
$$

Part 2. The objective of this part is to show Eq. (A.3) holds for all $K \geq 1$. We prove the result by induction.

We begin with the initial step. For $K=1$, the result holds by Assumption 6(b). In addition, part 1 implies that Eq. (A.4) with $K=1$.

We now verify the inductive step. Suppose that for some $K \geq 1$, Eqs. (A.3)-(A.4) hold. Based on this, we show that Eqs. (A.3)-(A.4) hold with K replaced by $K+1$. Consider the following argument. By inductive assumption, $(\hat{\alpha}_n^K,\hat{\theta}_{f,n},\hat{P}_n^{K-1}) = (\alpha^*,\theta_f^*,P^*)+o_{P_n}(1)$ and so $(\hat{\alpha}_n^K,\hat{\theta}_{f,n},\hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1. Then, the following derivation holds by exploiting the properties in Lemma 2.1.

$$
\begin{aligned}
&n^{\min\{\delta,1/2\}}(\hat{P}_n^K-P^*) = n^{\min\{\delta,1/2\}}(\Psi_{(\hat{\alpha}_n^K,\hat{\theta}_{f,n})}(\hat{P}_n^{K-1}) - \Psi_{(\alpha^*,\theta_f^*)}(P^*)) \\
&= n^{\min\{\delta,1/2\}}[\frac{\partial \Psi_{(\hat{\alpha}_n^K,\hat{\theta}_{f,n})}(\hat{P}_n^{K-1})}{\partial \alpha'}(\hat{\alpha}_n^K-\alpha^*) + \frac{\partial \Psi_{(\tilde{\alpha}_n^K,\tilde{\theta}_{f,n})}(\tilde{P}_n^{K-1})}{\partial \theta_f'}(\hat{\theta}_{f,n}-\theta^*) + \frac{\partial \Psi_{(\tilde{\alpha}_n^K,\tilde{\theta}_{f,n})}(\tilde{P}_n^{K-1})}{\partial P'}(\hat{P}_n^{K-1}-P^*)] \\
&= \frac{\partial \Psi_{(\alpha^*,\theta_f^*)}(P^*)}{\partial \alpha'} n^{\min\{\delta,1/2\}}(\hat{\alpha}_n^K-\alpha) + \frac{\partial \Psi_{(\alpha^*,\theta_f^*)}(P^*)}{\partial \theta_f'} n^{\min\{\delta,1/2\}}(\hat{\theta}_{f,n}-\theta^*) + o_{P_n}(1),
\end{aligned}
$$

where $(\tilde{\alpha}_n,\tilde{\theta}_{f,n},\tilde{P}_n)$ is some sequence between $(\hat{\alpha}_n^K,\hat{\theta}_{f,n},\hat{P}_n^{K-1})$ and $(\alpha^*,\theta_f^*,P^*)$, and the first equality uses that $P^* = \Psi_{(\alpha^*,\theta_f^*)}(P^*)$, the second equality holds by Lemma 2.1(c), and the final equality holds by Lemma 2.1(c)-(d) and $n^{\min\{\delta,1/2\}}((\hat{\alpha}_n^K,\hat{\theta}_{f,n},\hat{P}_n^{K-1})-(\alpha^*,\theta_f^*,P^*)) = O_{P_n}(1)$. From this, we conclude that $n^{\min\{\delta,1/2\}}(\hat{P}_n^K-P^*) = O_{P_n}(1)$, i.e., Eq. (A.3) holds with K replaced by $K+1$. In turn, this and part 1 then implies that Eq. (A.4) holds with K replaced by $K+1$. This concludes the inductive step and the proof. □

Proof of Theorem 4.1. This result is a corollary of Theorem 3.1 and Lemma A.6. To apply Theorem 3.1, we need to verify Assumptions 5-6. We anticipate that $Q_\infty^{ML}(\theta,P) = \sum_{(a,x)\in AX} J^*(a,x)\ln\Psi_\theta(P)(a|x)$.

Part 1: Verify Assumption 5.

Condition (a). First, notice that $\hat{J}_n - J^* = o_{P_n}(1)$ and $\Psi_\theta(P)(a|x) > 0$ for all $(\theta,P)\in\Theta\times\Theta_P$ implies that $Q_n^{ML}(\theta,P) - Q_\infty^{ML}(\theta,P) = o_{P_n}(1)$. Furthermore, notice that

$$\begin{aligned}\sup_{(\theta,P)\in\Theta\times\Theta_P}|Q_n^{ML}(\theta,P) - Q_\infty^{ML}(\theta,P)| &= \sup_{(\theta,P)\in\Theta\times\Theta_P}|\sum_{(a,x)\in AX}(\hat{J}_n(a,x) - J^*(a,x))\ln\Psi_\theta(P)(a|x)| \\ &\leq \sum_{(a,x)\in AX}|\hat{J}_n(a,x) - J^*(a,x)||\ln[\min_{(a,x)\in AX}\inf_{(\theta,P)\in\Theta\times\Theta_P}\Psi_\theta(P)(a|x)]|\end{aligned}$$

Since $\Psi_\theta(P)(a|x) > 0$ for all $(\theta,P)\in\Theta\times\Theta_P$ and all $(a,x)\in AX$, $\Psi_\theta(P)(a|x):\Theta\times\Theta_P\to\mathbb{R}$ is continuous in (θ,P) for all $(a,x)\in AX$, and $\Theta\times\Theta_P$ is compact, this implies that $\min_{(a,x)\in AX}\inf_{(\theta,P)\in\Theta\times\Theta_P}\Psi_\theta(P)(a|x) > 0$. From this and $\hat{J}_n - J^* = o_{P_n}(1)$, we conclude that $\sup_{(\theta,P)\in\Theta\times\Theta_P}|Q_n^{ML}(\theta,P) - Q_\infty^{ML}(\theta,P)| = o_{P_n}(1)$. Second, the previous arguments imply that $Q_\infty^{ML}(\theta,P):\Theta\times\Theta_P\to\mathbb{R}$ is continuous in (θ,P). In turn, since $\Theta\times\Theta_P$ is compact it follows that $Q_\infty^{ML}(\theta,P):\Theta\times\Theta_P\to\mathbb{R}$ is uniformly continuous in (θ,P). Third, $(\hat{\theta}_{f,n},\tilde{P}_n) - (\theta_f^*,P^*) = o_{P_n}(1)$, where $\tilde{P}_n$ is the arbitrary sequence in condition (a). By combining these with Gourieroux and Monfort (1995, Lemma 24.1), the result follows.

Condition (b). This result is a consequence of Assumption 2 and the information inequality (e.g. White (1996, Theorem 2.3)).

Condition (c). Since $\Psi_\theta(P)(a|x) > 0$ for all $(\theta,P)\in\Theta\times\Theta_P$ and all $(a,x)\in AX$, and $\Psi_\theta(P)(a|x):\Theta\times\Theta_P\to\mathbb{R}$ is twice continuously differentiable in (θ,P) for all $(a,x)\in AX$, we conclude that $\ln\Psi_\theta(P)(a|x):\Theta\times\Theta_P\to\mathbb{R}$ is twice continuously differentiable in (θ,P) for all $(a,x)\in AX$. From here, the result follows.

Condition (d). In the verification of condition (c), we have shown that $Q_n^{ML}(\theta,P):\Theta\times\Theta_P\to\mathbb{R}$ is twice continuously differentiable in $(\theta,P)\in\Theta\times\Theta_P$. By the same argument, $Q_\infty^{ML}(\theta,P):\Theta\times\Theta_P\to\mathbb{R}$ is also twice continuously differentiable in $(\theta,P)\in\Theta\times\Theta_P$. Then, by direct computation,

$$\begin{aligned}\sup_{(\theta,P)\in\mathcal{N}}|\frac{\partial^2 Q_n^{ML}(\alpha,\theta_f,P)}{\partial\lambda'\partial\alpha} - \frac{\partial^2 Q_\infty^{ML}(\alpha,\theta_f,P)}{\partial\lambda'\partial\alpha}| &= \sup_{(\theta,P)\in\mathcal{N}}|\sum_{(a,x)\in AX}(J_n^*(a,x) - J^*(a,x))M_{\theta,P}(a,x)| \\ &\leq \sum_{(a,x)\in AX}|J_n^*(a,x) - J^*(a,x)||\max_{(a,x)\in AX}\sup_{(\theta,P)\in\Theta\times\Theta_P}|M_{\theta,P}(a,x)|,\end{aligned}$$

with $M_{\theta,P}(a,x)$ defined by:

$$M_{\theta,P}(a,x) \equiv \frac{-1}{(\Psi_\theta(P)(a|x))^2}\frac{\partial\Psi_\theta(P)(a|x)}{\partial\lambda'}\frac{\partial\Psi_\theta(P)(a|x)}{\partial\alpha} + \frac{1}{\Psi_\theta(P)(a|x)}\frac{\partial^2\Psi_\theta(P)(a|x)}{\partial\lambda'\partial\alpha}.$$

Since $\Psi_\theta(P)(a|x) > 0$ for all $(\theta,P)\in\Theta\times\Theta_P$ and all $(a,x)\in AX$, $\Psi_\theta(P)(a|x):\Theta\times\Theta_P\to\mathbb{R}$ is continuous in (θ,P) for all $(a,x)\in AX$, and $\Theta\times\Theta_P$ is compact, this implies that $\inf_{(\theta,P)\in\Theta\times\Theta_P}\Psi_\theta(P)(a|x) > 0$ for all $(a,x)\in AX$. From this and that $\Psi_\theta(P)$ is twice continuously differentiable in (θ,P), we conclude that $M_{\theta,P}(a|x)$ is continuous in (θ,P) for all $(a,x)\in AX$. Since $\Theta\times\Theta_P$ is compact, $\max_{(a,x)\in AX}\sup_{(\theta,P)\in\Theta\times\Theta_P}|M_{\theta,P}(a,x)| < \infty$. From this and $\hat{J}_n - J^* = o_{P_n}(1)$, the result follows.

Condition (e). By direct computation, for any $\lambda\in(\alpha,\theta_f,P)$,

$$\frac{\partial^2 Q_\infty^{ML}(\alpha,\theta_f,P)}{\partial\lambda\partial\alpha'} = \sum_{(a,x)\in AX}J^*(a,x)\{\frac{-1}{(\Psi_\theta(P)(a,x))^2}\frac{\partial\Psi_\theta(P)(a|x)}{\partial\lambda}\frac{\partial\Psi_\theta(P)(a|x)}{\partial\alpha'} + \frac{1}{\Psi_\theta(P)(a|x)}\frac{\partial^2\Psi_\theta(P)(a|x)}{\partial\lambda\partial\alpha'}\}. \tag{A.5}$$

This function is continuous and, when evaluated at $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$, we obtain:

$$\begin{aligned}
\frac{\partial^2 Q_\infty^{ML}(\alpha^*, \theta_f^*, P^*)}{\partial\lambda\partial\alpha'} &= \sum_{(a,x)\in AX} J^*(a,x)\{\frac{-1}{(P_{\theta^*}(a|x))^2}\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda}\frac{\partial P_{\theta^*}(a|x)}{\partial\alpha'} + \frac{1}{P_{\theta^*}(a|x)}\frac{\partial^2\Psi_{\theta^*}(P^*)(a|x)}{\partial\lambda\partial\alpha'}\} \\
&= -\sum_{(a,x)\in AX} J^*(a,x)\frac{\partial \ln P_{\theta^*}(a|x)}{\partial\lambda}\frac{\partial \ln P_{\theta^*}(a|x)}{\partial\alpha'} + \sum_{x\in X} m^*(x)\sum_{a\in A}\frac{\partial^2\Psi_{\theta^*}(P^*)(a|x)}{\partial\lambda\partial\alpha'} \\
&= -E_{J^*}[\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\alpha}'] = -\frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\frac{\partial P_{\theta^*}}{\partial\alpha}' \qquad \text{(A.6)}
\end{aligned}$$

where the first equality uses $\partial\Psi_{\theta^*}(P^*)/\partial\alpha = \partial P_{\theta^*}/\partial\alpha$, and $P_{\theta^*} = P^*$, the second equality uses $J^*(a,x) = P^*(a|x)m^*(x)$, the third equality interchanges summation and differentiation and uses that $\sum_{a\in A}\Psi_{\theta^*}(P^*)(a|x) = 1$ for all $x \in X$, and the final equality in the third line uses Lemma A.6. To verify the result, it suffices to consider the last expression with $\lambda = \alpha$. Since $(\Sigma\Omega_{J,J}\Sigma')^{-1}$ is a non-singular matrix and $\partial P_{\theta^*}/\partial\alpha$ has full rank matrix by Assumption 9, we conclude that the expression is square, symmetric, and negative definite, and, consequently, it must be non-singular.

Condition (f). If we focus Eq. (A.5) on $\lambda = P$ and evaluate at $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$,

$$\begin{aligned}
&\frac{\partial^2 Q_\infty^{ML}(\alpha^*, \theta_f^*, P^*)}{\partial P\partial\alpha'} \\
&= \sum_{(a,x)\in AX} J^*(a,x)\{\frac{-1}{(\Psi_{\theta^*}(P^*)(a,x))^2}\frac{\partial\Psi_{\theta^*}(P^*)(a,x)}{\partial P}\frac{\partial\Psi_{\theta^*}(P^*)(a,x)}{\partial\alpha'} + \frac{1}{\Psi_\theta(P)(a,x)}\frac{\partial^2\Psi_{\theta^*}(P^*)(a,x)}{\partial P\partial\alpha'}\} \\
&= \sum_{(a,x)\in AX} J^*(a,x)\{\frac{-1}{(\Psi_{\theta^*}(P^*)(a,x))^2}\frac{\partial\Psi_{\theta^*}(P_{\theta^*})(a,x)}{\partial P}\frac{\partial\Psi_{\theta^*}(P^*)(a,x)}{\partial\alpha'} + \frac{1}{\Psi_\theta(P)(a,x)}\frac{\partial}{\partial\alpha'}(\frac{\partial\Psi_{\theta^*}(P_{\theta^*})(a,x)}{\partial P})\}.
\end{aligned}$$

where the second line uses $P_{\theta^*} = P^*$ and Young's theorem. Since the Jacobian matrix of Ψ_{θ^*} with respect to P is zero at P_{θ^*}, the result follows.

Part 2: Verify Assumption 6.

Assumption 6(b) holds by Lemma A.3. To verify Assumption 6(a), consider the following argument. By the verification of Assumption 5(c)-(d), Q_n^{ML} and Q_∞^{ML} are twice continuously differentiable in $(\theta, P) \in \mathcal{N}$. By direct computation,

$$\begin{aligned}
\frac{\partial Q_n^{ML}(\alpha^*, \theta_f^*, P^*)}{\partial\alpha} &= \sum_{(a,x)\in AX}\hat{J}_n(a,x)\frac{1}{\Psi_{\theta^*}(P^*)(a,x)}\frac{\partial\Psi_{\theta^*}(P^*)(a,x)}{\partial\alpha} \\
&= \frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\alpha}\hat{J}_n = \frac{\partial P_{\theta^*}}{\partial\alpha}(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma\hat{J}_n,
\end{aligned}$$

where the last equality uses that $\partial\Psi_{\theta^*}(P^*)/\partial\alpha = \partial P_{\theta^*}/\partial\alpha$ and Lemma A.5. Also, by using an analogous argument but applied to the population,

$$\frac{\partial Q_\infty^{ML}(\alpha^*, \theta_f^*, P^*)}{\partial\alpha} = \frac{\partial P_{\theta^*}}{\partial\alpha}(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma J^* = \mathbf{0},$$

where the equality to zero holds by Lemma A.5. By combining these results, we conclude that:

$$n^{\min\{\delta,1/2\}}\begin{bmatrix}\partial Q_n^{ML}(\alpha^*, \theta_f^*, P^*)/\partial\alpha \\ (\hat{\theta}_{f,n} - \theta_f^*)\end{bmatrix} = \begin{bmatrix}\frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma & \mathbf{0}_{|AX|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times|AX|} & \mathbf{I}_{d_{\theta_f}}\end{bmatrix} n^{\min\{\delta,1/2\}}\begin{pmatrix}\hat{J}_n - J^* \\ \hat{\theta}_{f,n} - \theta_f^*\end{pmatrix}.$$

From this and Lemma A.1, we conclude that the desired result holds with $[\zeta_1, \zeta_2]'$ distributed according to:

$$N\left(\begin{pmatrix} \frac{\partial P_{\theta^*}}{\partial \lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma B_J \\ B_{\theta_f} \end{pmatrix}, \begin{pmatrix} \frac{\partial P_{\theta^*}}{\partial \lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma\Omega_{J,J}\Sigma'(\Sigma\Omega_{J,J}\Sigma')^{-1}\frac{\partial P_{\theta^*}}{\partial \lambda}' & \frac{\partial P_{\theta^*}}{\partial \lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma\Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f}\Sigma'(\Sigma\Omega_{J,J}\Sigma')^{-1}\frac{\partial P_{\theta^*}}{\partial \lambda}' & \Omega_{\theta_f,\theta_f} \end{pmatrix}\right). \tag{A.7}$$

This completes the verification of Assumptions 5-6 and so Theorem 3.1 applies. The specific formula for the asymptotic distribution relies on the expressions in Eqs. (A.6)-(A.7). □

Proof of Theorem 4.2. This result is a corollary of Theorem 3.1. To complete the proof, we need to verify Assumptions 5-6. We anticipate that $Q_\infty^{MD}(\theta, P) = -[P^* - \Psi_\theta(P)]'W^*[P^* - \Psi_\theta(P)]$.

Part 1: Verify the conditions in Assumption 5.

Condition (a). First, we show that $\sup_{(\theta,P)\in\Theta\times\Theta_P} |Q_n^{MD}(\theta, P) - Q_\infty^{MD}(\theta, P)| = o_{P_n}(1)$. Consider the following argument:

$$\begin{aligned} \sup_{(\theta,P)\in\Theta\times\Theta_P} |Q_n^{MD}(\theta, P) - Q_\infty^{MD}(\theta, P)| &= \sup_{(\theta,P)\in\Theta\times\Theta_P} \left| \begin{array}{c} -(\hat{P}_n - P^*)'\hat{W}_n[\hat{P}_n - \Psi_\theta(P)] \\ -(P^* - \Psi_\theta(P))'[\hat{W}_n - W^*][\hat{P}_n - \Psi_\theta(P)] \\ -(P^* - \Psi_\theta(P))'W^*(\hat{P}_n - P^*) \end{array} \right| \\ &\leq \|\hat{P}_n - P^*\|[\|\hat{W}_n - W^*\| + 2\|W^*\|] + \|\hat{W}_n - W^*\| \end{aligned}$$

Second, since $\Psi_\theta(P)(a|x) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ, P) for all (a, x), it follows that $Q_\infty^{MD}(\theta, P) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ, P). In turn, since $\Theta \times \Theta_P$ is compact it follows that $Q_\infty^{MD}(\theta, P) : \Theta \times \Theta_P \to \mathbb{R}$ is uniformly continuous in (θ, P). Third, $(\hat{\theta}_{f,n}, \tilde{P}_n) - (\theta_f^*, P^*) = o_{P_n}(1)$, where $\tilde{P}_n$ is the arbitrary sequence in condition (a). By combining these with Gourieroux and Monfort (1995, Lemma 24.1), the result follows.

Condition (b). $Q_\infty^{MD}(\alpha, \theta_f^*, P^*) = -[P^* - \Psi_{(\alpha,\theta_f^*)}(P^*)]'W^*[P^* - \Psi_{(\alpha,\theta_f^*)}(P^*)]$ is uniquely maximized at α^*. First, notice that $\Psi_{(\alpha^*,\theta_f^*)}(P^*) = P^*$ and so $Q_\infty^{MD}(\alpha^*, \theta_f^*, P^*) = 0$. Second, consider any $\tilde{\alpha} \in \Theta_\alpha \backslash \alpha^*$. By the identification assumption, $\Psi_{(\tilde{\alpha},\theta_f^*)}(P^*) \neq \Psi_{(\alpha^*,\theta_f^*)}(P^*) = P^*$. Since W^* is positive definite, $Q_\infty^{MD}(\tilde{\alpha}, \theta_f^*, P^*) > 0$.

Condition (c). This result follows from the fact that $\Psi_\theta(P)(a|x) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in (θ, P) for all $(a, x) \in AX$.

Condition (d). By the same argument as in the verification of condition (c), it follows that $Q_\infty^{MD}(\theta, P) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in (θ, P). Since $\Psi_\theta(P)(a|x)$ is twice continuously differentiable in (θ, P) for all (a, x), we conclude that $\partial\Psi_\theta(P)(a, x)/\partial\lambda$ and $\partial\Psi_\theta(P)(a, x)/(\partial\lambda'\partial\alpha)$ are continuous in (θ, P) for $\lambda \in \{\theta, P\}$ for all (a, x). From this and the fact that $\Theta \times \Theta_P$ is compact, we conclude that $\max_{(a,x)\in AX} \sup_{(\theta,P)\in\Theta\times\Theta_P} \|\partial\Psi_\theta(P)(a, x)/\partial\lambda\| < \infty$ and $\max_{(a,x)\in AX} \sup_{(\theta,P)\in\Theta\times\Theta_P} \|\partial\Psi_\theta(P)(a, x)/\partial\lambda'\partial\alpha\| < \infty$. From this, $\hat{P}_n - P^* = o_{P_n}(1)$, and $\hat{W}_n - W^* = o_{P_n}(1)$, the result follows.

Condition (e). By direct computation, for any $\lambda \in (\alpha, \theta_f, P)$,

$$\frac{\partial^2 Q_\infty^{MD}(\alpha, \theta_f, P)}{\partial\lambda\partial\alpha'} = 2[\frac{\partial}{\partial\lambda}\frac{\Psi_\theta(P)}{\partial\alpha}W^*(P^* - \Psi_\theta(P)) - \frac{\partial\Psi_\theta(P)}{\partial\alpha}W^*\frac{\partial\Psi_\theta(P)}{\partial\lambda}']. \tag{A.8}$$

This function is continuous and, when evaluated at $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$, we obtain:

$$\frac{\partial Q_\infty^{MD}(\alpha^*, \theta_f^*, P^*)}{\partial\lambda\partial\alpha'} = -2\frac{\partial\Psi_{\theta^*}(P^*)}{\partial\alpha}W^*\frac{\partial\Psi_{\theta^*}(P^*)}{\partial\lambda}' = -2\frac{\partial P_{\theta^*}}{\partial\alpha}W^*\frac{\partial P_{\theta^*}}{\partial\lambda}'$$

where the first line uses that $P^* = \Psi_{\theta^*}(P^*)$ and $\partial\Psi_{\theta^*}(P^*)/\partial\alpha = \partial P_{\theta^*}/\partial\alpha$. To verify the result, it suffices to consider the last expression with $\lambda = \alpha$. By assumption, this expression is square, symmetric, and negative definite, and, consequently, it must be non-singular.

Condition (f). If we focus Eq. (A.8) on $\lambda = P$ and evaluate at $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$,

$$\frac{\partial^2 Q_\infty^{MD}(\alpha^*, \theta_f^*, P^*)}{\partial P \partial \alpha'} = -2 \frac{\Psi_{\theta^*}(P^*)}{\partial \alpha} W^* \frac{\Psi_{\theta^*}(P^*)'}{\partial P} = 0.$$

where the we have used that the Jacobian matrix of Ψ_{θ^*} with respect to P is zero at $P_{\theta^*} = P^*$, the result follows.

Part 2: Verify the conditions in Assumption 6.

Assumption 6(b) holds by Lemma A.3. To verify Assumption 6(a), consider the following argument. By the verification of conditions (c) and (d), Q_n^{MD} and Q_∞^{MD} are twice continuously differentiable in $(\theta, P) \in \mathcal{N}$. By direct computation,

$$\frac{\partial Q_n^{MD}(\alpha^*, \theta_f^*, P^*)}{\partial \alpha} = 2 \frac{\Psi_{\theta^*}(P^*)}{\partial \alpha} \hat{W}_n [\hat{P}_n - \Psi_{\theta^*}(P^*)] = 2 \frac{\partial P_\theta^*}{\partial \alpha} W^* [\hat{P}_n - P^*] + o_{P_n}(1),$$

where the last equality uses that $\Psi_{\theta^*}(P^*) = P^*$, $\partial \Psi_{\theta^*}(P^*)/\partial \alpha = \partial P_{\theta^*}/\partial \alpha$, $\hat{P}_n - P^* = o_{P_n}(1)$, and $\hat{W}_n - W^* = o_{P_n}(1)$. We then conclude that:

$$n^{\min\{\delta,1/2\}} \begin{bmatrix} \partial Q_n^{MD}(\alpha^*, \theta_f^*, P^*)/\partial \alpha \\ (\hat{\theta}_{f,n} - \theta_f^*) \end{bmatrix} = \begin{bmatrix} 2 \frac{\partial P_\theta^*}{\partial \alpha} W^* & \mathbf{0}_{|AX| \times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f} \times |AX|} & \mathbf{I}_{d_{\theta_f}} \end{bmatrix} n^{\min\{\delta,1/2\}} \begin{bmatrix} (\hat{P}_n - P^*) \\ (\hat{\theta}_{f,n} - \theta_f^*) \end{bmatrix} + o_{P_n}(1).$$

From this and Lemma A.2, we conclude that the desired result holds with $[\zeta_1, \zeta_2]'$ distributed according to:

$$N\left(\begin{pmatrix} 2 \frac{\partial P_\theta^*}{\partial \alpha} W^* \Sigma B_J \\ B_{\theta_f} \end{pmatrix}, \begin{pmatrix} 4 \frac{\partial P_\theta^*}{\partial \alpha} W^* \Sigma \Omega_{J,J} \Sigma' W^* \frac{\partial P_\theta^*}{\partial \alpha}' & 2 \frac{\partial P_\theta^*}{\partial \alpha} W^* \Sigma \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} \Sigma' W^{*\prime} \frac{\partial P_\theta^*}{\partial \alpha}' & \Omega_{\theta_f,\theta_f} \end{pmatrix} \right). \tag{A.9}$$

This completes the verification of Assumptions 5-6 and so Theorem 3.1 applies. The specific formula for the asymptotic distribution relies on the expressions in Eqs. (A.8)-(A.9). □

A.3 Proofs of lemmas

Proof of Lemma 2.1. The econometric model imposes all assumptions in Aguirregabiria and Mira (2002, Sections 2-3). Thus, Parts (a)-(b) follow from Aguirregabiria and Mira (2002, Proposition 1), Part (d) follows from Aguirregabiria and Mira (2002, Proposition 2), and Parts (c) and (e) are a corollary of the discussion in Aguirregabiria and Mira (2002, Page 1532). □

Proof of Lemma 2.2. Part (a) follows from $P_\theta = \Psi_\theta(P_\theta)$ and that $\Psi_\theta(P)(a|x) > 0$ for any $(a, x) \in AX$ and any θ and P. Part (b) follows from Rust (1988, Pages 1015-6). Part (c) follows from $P_\theta = \Psi_\theta(P_\theta)$ and $\partial \Psi_\theta(P)/\partial P = \mathbf{0}$ at $P = P_\theta$. Part (d) follows from the following argument. Suppose that $\exists \theta_f \in \Theta_f$ and $\exists \alpha_a, \alpha_b \in \Theta_\alpha$ s.t. $P_{(\alpha_a,\theta_f)} = P_{(\alpha_b,\theta_f)}$. Then, $P_\theta = \Psi_\theta(P_\theta)$ implies that $\Psi_{(\alpha_a,\theta_f)}(P) = P$ and $\Psi_{(\alpha_b,\theta_f)}(P) = P$ for $P = P_{(\alpha_a,\theta_f)} = P_{(\alpha_b,\theta_f)}$. In turn, since this condition identifies α, we conclude that $\alpha_a = \alpha_b$. □

Lemma A.1. *Assume Assumptions 3-7. Then,*

$$n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{J}_n - J^* \\ \hat{\theta}_{f,n} - \theta_f^* \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} B_J \\ B_{\theta_f} \end{pmatrix} \times 1[\delta \leq 1/2], \begin{pmatrix} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{ff} \end{pmatrix} \times 1[\delta \geq 1/2] \right), \tag{A.10}$$

where

$$\begin{pmatrix} B_J \\ B_{\theta_f} \end{pmatrix} \equiv \Delta B_{\Pi^*},$$
$$\begin{pmatrix} \Omega_{J,J} & \Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f} & \Omega_{ff} \end{pmatrix} \equiv \Delta(diag(\Pi^*) - \Pi^*\Pi^{*\prime})\Delta', \tag{A.11}$$

with

$$\Delta \equiv \begin{pmatrix} \{\{1[(a,x) = (\tilde{a},\tilde{x})]\}'_{(a,x,x')\in AX^2}\}_{(\tilde{a},\tilde{x})\in AX}, \\ DG_{\theta_f}(\Pi^*) \end{pmatrix},$$

and $G_{\theta_f}(\Pi^*)$ denotes the θ_f-component of G in Assumption 8, and DG_{θ_f} denotes its gradient.

Proof. Under Assumptions 3 and 7, the triangular array CLT (e.g. Davidson (1994, page 369)) implies that:

$$\sqrt{n}(\hat{\Pi}_n - \Pi^*_n) \xrightarrow{d} N(0, diag(\Pi^*) - \Pi^*\Pi^{*\prime}).$$

If this is combined with Assumptions 3, we conclude that:

$$n^{\min\{\delta,1/2\}}(\hat{\Pi}_n - \Pi^*) \xrightarrow{d} N(B_{\Pi^*} \times 1[\delta \leq 1/2], (diag(\Pi^*) - \Pi^*\Pi^{*\prime}) \times 1[\delta \geq 1/2]). \tag{A.12}$$

Consider the following argument.

$$\begin{aligned} n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{J}_n - J^* \\ \hat{\theta}_{f,n} - \theta^*_f \end{pmatrix} &= n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{J}_n - J^* \\ G_{n,\theta_f}(\hat{\Pi}_n) - G_{\theta_f}(\Pi^*) \end{pmatrix} \\ &= n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{J}_n - J^* \\ G_{\theta_f}(\hat{\Pi}_n) - G_{\theta_f}(\Pi^*) \end{pmatrix} + o_{P_n}(1) \\ &= n^{\min\{\delta,1/2\}}(F(\hat{\Pi}_n) - F(\Pi^*)) + o_{P_n}(1) \end{aligned}$$

where the first two equalities follow from Assumption 8, where G_{n,θ_f} denotes the θ_f-component of G_n, and the last equality follows from defining the function $F : \mathbb{R}^{|AX^2|} \to \mathbb{R}^{|AX|+d_{\theta_f}}$ as follows. For coordinates $j = 1, \ldots, |AX|$ where j represents coordinate $(a,x) \in AX$, $F_j(z) \equiv \sum_{\tilde{x}'\in X} z_{(a,x,\tilde{x}')}$, and for coordinates $j = |AX|+1, \ldots, |AX| + d_{\theta_f}$, $F_j(z) \equiv G_{\theta_f,j}(z)$. By definition, $F(\hat{\Pi}_n) \equiv (\hat{J}_n, G_{\theta_f}(\hat{\Pi}_n))$ and $F(\Pi^*) \equiv (J^*, G_{\theta_f}(\Pi^*))$. By Assumptions 3 and Eq. (A.12), $\hat{\Pi}_n$,belongs to any arbitrarily small neighborhood of Π^* w.p.a.1. Eq. (A.10) holds by the delta method provided that F is continuously differentiable for any Π in a neighborhood of Π^* and its gradient at Π^* is equal to Δ, as we now verify. For any coordinate $j > |AX|$, Assumption 8 implies that G_{θ_f} is differentiable at Π^*. For any coordinate $j \leq |AX|$ that represents a certain $(a,x) \in AX$, the gradient is given by $\partial F_j(z)/\partial z_{(\tilde{a},\tilde{x},\tilde{x}')} = 1[(a,x) = (\tilde{a},\tilde{x})]$. □

Lemma A.2. *Assume Assumptions 3-7. Then,*

$$n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{P}_n - P^* \\ \hat{\theta}_{f,n} - \theta^*_f \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \Sigma B_J \\ B_{\theta_f} \end{pmatrix} \times 1[\delta \leq 1/2], \begin{pmatrix} \Sigma\Omega_{J,J}\Sigma' & \Sigma\Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f}\Sigma' & \Omega_{\theta_f,\theta_f} \end{pmatrix} \times 1[\delta \geq 1/2] \right), \tag{A.13}$$

where

$$\begin{pmatrix} \Sigma B_J \\ B_{\theta_f} \end{pmatrix} \equiv \begin{bmatrix} \Sigma & \mathbf{0}_{|\tilde{A}X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times|\tilde{A}X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}} \end{bmatrix} \Delta B_{\Pi^*},$$

$$\begin{pmatrix} \Sigma\Omega_{J,J}\Sigma' & \Sigma\Omega_{J,\theta_f} \\ \Omega'_{J,\theta_f}\Sigma' & \Omega_{\theta_f,\theta_f} \end{pmatrix} \equiv \begin{bmatrix} \Sigma & \mathbf{0}_{|\tilde{A}X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times|\tilde{A}X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}} \end{bmatrix} \Delta(diag(\Pi^*) - \Pi^*\Pi^{*\prime})\Delta' \begin{bmatrix} \Sigma & \mathbf{0}_{|\tilde{A}X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times|\tilde{A}X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}} \end{bmatrix}',$$

with Σ *as in Eq.* (A.1),

$$\Delta \equiv \begin{pmatrix} \{\{1[(a,x) = (\tilde{a},\tilde{x})]\}'_{(a,x,x')\in AX^2}\}_{(\tilde{a},\tilde{x})\in AX}, \\ DG_{\theta_f}(\Pi^*) \end{pmatrix},$$

and $G_{\theta_f}(\Pi^*)$ *denotes the* θ_f*-component of* G *in Assumption 8, and* DG_{θ_f} *denotes its gradient.*

Proof. This result follows from Lemma A.1 and the delta method. Let $F : \mathbb{R}^{|AX|+d_{\theta_f}} \to \mathbb{R}^{|\tilde{A}X|+d_{\theta_f}}$ be defined as follows. For coordinates $j = 1,\ldots,|\tilde{A}X|$ where j represents coordinate $(a,x) \in \tilde{A}X$, $F_j(z) \equiv z_{(a,x)}/\sum_{a\in A} z_{(\tilde{a},x)}$, and for $j = |\tilde{A}X|+1,\ldots,|\tilde{A}X|+d_{\theta_f}$, $F_j(z) = z_j$. By definition, $F((\hat{J}_n,\hat{\theta}_{f,n})) \equiv (\hat{P}_n,\hat{\theta}_{f,n})$ and $F((J^*,\theta_f^*)) \equiv (P^*,\theta_f^*)$. Eq. (A.13) holds by the delta method provided that F is continuously differentiable, as we now verify.

For any coordinate $j > |AX|$, the function F_j is trivially differentiable. For any coordinate $j \leq |AX|$ that represents a certain $(a,x) \in AX$ and any $(\tilde{a},\tilde{x}) \in \tilde{A}X$, the gradient is given by:

$$\begin{aligned} \frac{\partial F_{(a,x)}(z)}{\partial z_{(\tilde{a},\tilde{x})}} &= 1[x=\tilde{x}]\left[\left(\frac{\sum_{\check{a}\in A} z_{(\check{a},\tilde{x})} - z_{(a,\tilde{x})}}{(\sum_{\check{a}\in A} z_{(\check{a},\tilde{x})})^2}\right)1[a=\tilde{a}] + \left(\frac{-z_{(a,\tilde{x})}}{(\sum_{\check{a}\in A} z_{(\check{a},\tilde{x})})^2}\right)1[a\neq\tilde{a}]\right] \\ &= \frac{1[x=\tilde{x}]}{(\sum_{\check{a}\in A} z_{(\check{a},\tilde{x})})}[1[a=\tilde{a}] - \frac{z_{(\tilde{a},\tilde{x})}}{(\sum_{\check{a}\in A} z_{(\check{a},\tilde{x})})}], \end{aligned}$$

provided that $\sum_{\check{a}\in A} z_{(\check{a},\tilde{x})} > 0$. Since $\sum_{\tilde{a}\in A} J^*(\tilde{a},x) > 0$ for all $x \in X$, we then conclude that F is continuously differentiable at (J^*,θ_f^*). By combining the information from all coordinates, it follows that the gradient is given by:

$$DF(J^*,\theta_f^*) = \begin{bmatrix} \Sigma & \mathbf{0}_{|\tilde{A}X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times|\tilde{A}X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}} \end{bmatrix}.$$

By application of the delta method, the desired result follows. □

Lemma A.3. *Assume Assumptions 3-7. Then,*

$$n^{\min\{\delta,1/2\}}(\hat{P}_n^0 - P^*) \xrightarrow{d} N(B_P \times 1[\delta \leq 1/2], \Omega_P \times 1[\delta \geq 1/2]),$$

where $B_P \equiv DG_P(\Pi^*)B_{\Pi^*}$, $\Omega_P \equiv DG_P(\Pi^*)(diag(\Pi^*) - \Pi^*\Pi^{*\prime})DG_P(\Pi^*)'$, *and* G_P *denotes the* P*-component of* G *in Assumption 8, and* DG_P *denotes its gradient.*

Proof. This proof is analogous to that of Lemma A.1 and is therefore omitted. □

Lemma A.4. *Assume a non-parametric model for the first stage, i.e.,* $\theta_f \equiv \{f(x'|a,x)\}_{(a,x,x')\in AX^2}$. *Then, the preliminary estimators* $(\hat{\theta}_{f,n},\hat{P}_n^0) = (\hat{f}_n,\hat{P}_n)$ *satisfy Assumption 8.*

Proof. By definition, $(\hat{\theta}_{f,n},\hat{P}_n^0)$ satisfies $(\hat{\theta}_{f,n},\hat{P}_n^0) = G(\hat{\Pi}_n)$ with $G : \mathbb{R}^{|AX^2|} \to \mathbb{R}^{|AX^2|} \times \mathbb{R}^{|AX|}$ defined as follows $G_j(\Pi) \equiv \sum_{\tilde{x}'\in X} \Pi(a,x,\tilde{x}')/\sum_{(\check{a},\tilde{x}')\in AX} \Pi(\check{a},x,\tilde{x}')$ for $j = 1,\ldots,|AX^2|$ and $G_j(\Pi) \equiv \Pi(a,x,x')/\sum_{\tilde{x}'\in X} \Pi(a,x,\tilde{x}')$ for $j = |AX^2|+1,\ldots,|AX^2|+|AX|$. Also, let us define $\mathcal{N}_{\Pi^*} \equiv \{\Pi \in \mathbb{R}^{|AX^2|} : \Pi(a,x,x') \geq \eta/2\}$ for $\eta \equiv \inf_{(a,x,x')\in AX^2} \Pi^*(a,x,x') > 0$. Notice that this implies that $\Pi^* \in \mathcal{N}_{\Pi^*}$.

Assumption 8(a) is automatically satisfied because G is a constant function. Assumption 8(b) follows from the fact that $\Pi(a,x,x') \geq \eta/2 > 0$ for all $\Pi \in \mathcal{N}_{\Pi^*}$. Finally, Assumption 8(c) follows from the definition of G as it implies that $G(\Pi^*) = (f^*, P^*) = (\theta_f^*, P^*)$. □

Lemma A.5. *The following algebraic results hold:*

(a) $\Sigma J^* = \mathbf{0}_{|\tilde{A}X|}$,

(b) $\Sigma B_J = \{B_J(a,x)/m^*(x)\}_{(a,x)\in\tilde{A}X}$,

(c) $\Sigma diag\{J^*\}\Sigma' = diag\{\Omega_{P,P}(x)/m^*(x)\}_{x\in X}$, *where for every* $x \in X$,

$$\Omega_{P,P}(x) \equiv [diag\{P^*(a|x)\}_{a\in\tilde{A}} - \{P^*(a|x)\}_{a\in\tilde{A}}\{P^*(a|x)\}'_{a\in\tilde{A}}]$$

and so

$$\Omega_{P,P}^{-1}(x) \equiv [diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}-1|\times|\tilde{A}-1|}].$$

(d) $(\Sigma\Omega_{J,J}\Sigma')^{-1} = diag\{m^*(x)[diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}|\times|\tilde{A}|}]\}_{x\in X}$,

(e) $(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma = diag\{[diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\}, -P^*(|A||x)^{-1}\mathbf{1}_{|\tilde{A}|\times 1}]\}_{x\in X}$.

Proof. Part (a). Notice that:

$$\begin{aligned}
\Sigma J^* &= diag\{\Sigma_x\}_{x\in X} J^* = diag\{\Sigma_x J^*(\cdot,x)\}_{x\in X} \\
&= diag\{\frac{1}{m^*(x)}[\mathbf{I}_{|\tilde{A}|\times|A|}J^*(\cdot,x) - \{P^*(a|x)\}_{a\in\tilde{A}}\mathbf{1}_{1\times|A|}J^*(\cdot,x)]\}_{x\in X} \\
&= diag\{\frac{1}{m^*(x)}[\{J^*(a,x)\}_{a\in\tilde{A}} - \{P^*(\tilde{a}|x)\}_{\tilde{a}\in\tilde{A}}m^*(x)]\}_{x\in X} \\
&= diag\{\frac{1}{m^*(x)}\mathbf{0}_{|\tilde{A}|}\}_{x\in X} = \mathbf{0}_{|\tilde{A}|\times|X|},
\end{aligned}$$

where we have used that $\mathbf{I}_{|\tilde{A}|\times|A|}J^*(\cdot,x) = \{J^*(a,x)\}_{a\in\tilde{A}}$, $\mathbf{1}_{1\times|A|}J^*(\cdot,x) = m^*(x)$, and $P^*(a|x)m^*(x) = J^*(a,x)$.

Part (b). Notice that:

$$\begin{aligned}
\Sigma B_J &= diag\{\Sigma_x\}_{x\in X} B_J = diag\{\Sigma_x B_J(\cdot,x)\}_{x\in X} \\
&= diag\{\frac{1}{\pi^*(x)}[\mathbf{I}_{|\tilde{A}|\times|A|}B_J(\cdot,x) - \{P^*(\tilde{a},x)\}_{\tilde{a}\in\tilde{A}}\mathbf{1}_{1\times|A|}B_J(\cdot,x)]\}_{x\in X} \\
&= \{B_J(a,x)/\pi^*(x)\}_{(a,x)\in\tilde{A}X},
\end{aligned}$$

Part (c). The first display is the result of algebraic derivations. The second display follows from Seber (2008, result 15.5).

Part (d). Consider the following argument. By definition, $\Omega_{J,J} \equiv diag\{J^*\} - J^*J^{*\prime}$. This and previous parts imply that $\Sigma\Omega_{J,J}\Sigma' = \Sigma diag\{J^*\}\Sigma' = diag\{\Omega_{PP}(x)/m^*(x)\}_{x\in X}$. Notice then that $diag\{\Omega_{PP}(x)/m^*(x)\}_{x\in X}$ is a block-diagonal matrix and each block is invertible. Then, by elementary properties of block-diagonal matrices, the result follows.

Part (e). Notice that:

$$(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma = diag\{m^*(x)\Omega_{PP}^{-1}(x)\}_{x\in X} diag\{\Sigma_x\}_{x\in X} = diag\{m^*(x)\Omega_{PP}^{-1}(x)\Sigma_x\}_{x\in X},$$

where:

$$\begin{aligned} m^*(x)\Omega_{PP}^{-1}(x)\Sigma_x &= [diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}-1|\times|\tilde{A}-1|}][\mathbf{I}_{|\tilde{A}|\times|A|} - \{P^*(\tilde{a}|x)\}_{\tilde{a}\in\tilde{A}}\mathbf{1}_{1\times|A|}] \\ &= \left\{ \begin{array}{c} [diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}-1|\times|\tilde{A}-1|}, \mathbf{0}_{\tilde{A}\times 1}] \\ -\mathbf{1}_{|\tilde{A}|\times|A|} + (1-(1/P^*(|A||x)))\mathbf{1}_{|\tilde{A}|\times|A|} \end{array} \right\} \\ &= [diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}-1|\times|\tilde{A}-1|}]. \end{aligned}$$

This completes the step and the proof. □

Lemma A.6. *For any $\lambda, \tilde{\lambda} \in \{\theta_f, \alpha\}$, the following algebraic results hold:*

$$\begin{aligned} \frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda} &= \frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma, \\ E_{J^*}[\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}'] &= \frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\frac{\partial P_{\theta^*}}{\partial\tilde{\lambda}}'. \end{aligned}$$

Proof. Before deriving the results, consider some preliminary observations. Notice that $\sum_{a\in A} P_{\theta^*}(a|x) = 1$ and so $\partial P_{\theta^*}(|A||x)/\partial\lambda' = -\sum_{a\in\tilde{A}}\partial P_{\theta^*}(a|x)/\partial\lambda'$. Also, notice that $P^*(a|x) = P_{\theta^*}(a|x)$ and so $(\partial P_{\theta^*}(a|x)/\partial\lambda)(1/P^*(a|x)) = \partial \ln P_{\theta^*}(a|x)/\partial\lambda$.

For the first result, consider the following derivation:

$$\begin{aligned} \frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\Sigma &= [\frac{\partial\{P_{\theta^*}(a|x)\}_{(a,x)\in\tilde{A}X}}{\partial\lambda}]diag\{[diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\}, -P^*(|A||x)^{-1}\mathbf{1}_{|\tilde{A}|\times 1}]\}_{x\in X} \\ &= \{[\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{a\in\tilde{A}}}{\partial\lambda}, -\sum_{a\in\tilde{A}}\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda}\frac{1}{P^*(|A||x)}]\}_{x\in X} \\ &= \{\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{a\in A}}{\partial\lambda}\}_{x\in X} = \frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}, \end{aligned}$$

where the first equality uses Lemma A.5(e) and rest of the equalities use the preliminary observations.

For the second result, consider the following derivation:

$$\begin{aligned} &\frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{J,J}\Sigma')^{-1}\frac{\partial P_{\theta^*}}{\partial\tilde{\lambda}}' \\ &= \frac{\partial P_{\theta^*}}{\partial\lambda}diag\{m^*(x)[diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}-1|\times|\tilde{A}-1|}]\}_{x\in X}\frac{\partial P_{\theta^*}}{\partial\tilde{\lambda}}' \\ &= diag\{m^*(x)[diag\{\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda}\frac{1}{P^*(a|x)} + \sum_{a\in\tilde{A}}\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda}\frac{1}{P^*(|A||x)}\}_{a\in\tilde{A}}]'\}_{x\in X}\frac{\partial P_{\theta^*}}{\partial\tilde{\lambda}}' \\ &= diag\{m^*(x)[diag\{\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda}\frac{1}{P^*(a|x)} - \frac{\partial P^*(|A||x)}{\partial\lambda}\frac{1}{P^*(|A||x)}\}_{a\in\tilde{A}}]'\}_{x\in X}\frac{\partial P_{\theta^*}}{\partial\tilde{\lambda}}' \\ &= \sum_{x\in X} m^*(x)\{\sum_{a\in\tilde{A}}\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda}\frac{\partial P_{\theta^*}(a|x)}{\partial\tilde{\lambda}'}\frac{1}{P^*(a|x)} - \frac{\partial P^*(|A||x)}{\partial\lambda}\frac{1}{P^*(|A||x)}\sum_{a\in\tilde{A}}\frac{\partial P_{\theta^*}(\tilde{a}|x)}{\partial\tilde{\lambda}}\} \\ &= \sum_{x\in X} m^*(x)\sum_{a\in A}\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda}\frac{\partial P_{\theta^*}(a|x)}{\partial\tilde{\lambda}'}\frac{1}{P^*(a|x)} \\ &= \sum_{(a,x)\in AX} J^*(a,x)\frac{\partial \ln P_{\theta^*}(a|x)}{\partial\lambda}\frac{\partial \ln P_{\theta^*}(a|x)}{\partial\tilde{\lambda}'} = E_{J^*}[\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\tilde{\lambda}}'], \end{aligned}$$

where the first equality uses Lemma A.5(d) and rest of the equalities use the preliminary observations. □

Lemma A.7. *Let $\mathcal{W}$ the space of positive definite and symmetric matrices in $\mathbb{R}^{|AX|\times|AX|}$, and let $f:\mathcal{W}\to\mathbb{R}^{d_\alpha\times d_\alpha}$,*

$$f(W^*)=\Upsilon_{MD}(W^*)\times M\times\Upsilon_{MD}(W^*)'.$$

with $\Upsilon_{MD}(W^*)$ defined as in Lemma 4.2 and M is positive definite and symmetric matrix. Then,

$$W^*=\arg\min_{\tilde{W}^*\in\mathcal{W}} f(\tilde{W}^*)$$

if and only if W^* satisfies the following condition:

$$\frac{\partial P_{\theta^*}}{\partial\alpha}' W^* = \left\{ \begin{array}{c} (\frac{\partial P_{\theta^*}}{\partial\alpha}W^*\frac{\partial P_{\theta^*}}{\partial\alpha}')\left(\frac{\partial P_{\theta^*}}{\partial\alpha}\left[\begin{array}{cc}\Sigma & -\frac{\partial P_{\theta^*}}{\partial\theta_f}'\end{array}\right]M\left[\begin{array}{cc}\Sigma & -\frac{\partial P_{\theta^*}}{\partial\theta_f}'\end{array}\right]'\frac{\partial P_{\theta^*}}{\partial\alpha}'\right)^{-1} \\ \times\frac{\partial P_{\theta^*}}{\partial\alpha}'\left(\left[\begin{array}{cc}\Sigma & -\frac{\partial P_{\theta^*}}{\partial\theta_f}'\end{array}\right]M\left[\begin{array}{cc}\Sigma & -\frac{\partial P_{\theta^*}}{\partial\theta_f}'\end{array}\right]'\right)^{-1} \end{array} \right\},$$

Proof. This result follows directly from reinterpreting the argument in McFadden and Newey (1994, Page 2165). □

A.4 Review of results on extremum estimators

The purpose of this section is to prove the consistency and asymptotic normality of extremum estimators under certain regularity conditions. Relative to the standard results in the literature (e.g. McFadden and Newey (1994)), our arguments allow for (a) a rate of convergence that is not necessarily $\sqrt{n}$ and (b) sequences of data generating processes that change as a function of the sample size. Both of these modifications are important for our theoretical results. We omit the proofs for reasons of brevity but the formal arguments are available from the authors upon request.

Theorem A.1. *Assume the following:*

(a) $Q_n(\theta)$ converges uniformly in probability to $Q(\theta)$ along $\{P_n\}_{n\geq1}$, i.e., for any $\varepsilon>0$,

$$P_n(\sup_{\theta\in\Theta}|Q_n(\theta)-Q(\theta)|>\varepsilon)=o(1).$$

(b) $Q(\theta)$ is upper semi-continuous, i.e., for any $\{\theta_n\}_{n\geq1}$ with $\theta_n\to\tilde{\theta}$,

$$\limsup Q(\theta_n)\leq Q(\tilde{\theta}).$$

(c) $Q(\theta)$ is uniquely maximized at $\theta=\theta^*$.

Consider an estimator $\hat{\theta}_n\in\Theta$ that satisfies $Q_n(\hat{\theta}_n)=\max_{\theta\in\Theta}Q_n(\theta)$. Then, $\hat{\theta}_n=\theta^*+o_{P_n}(1)$.

Theorem A.2. *Consider an estimator $\hat{\theta}_n$ of a parameter θ^* that satisfies $\hat{\theta}_n\in\Theta$ and $Q_n(\hat{\theta}_n)=\max_{\theta\in\Theta}Q_n(\theta)$. Furthermore,*

(a) $\hat{\theta}_n=\theta^*+o_{P_n}(1)$,

(b) θ^* belongs to the interior of Θ,

(c) Q_n is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ^* w.p.a.1,

(d) For some $\delta>0$, $n^\delta\partial Q_n(\theta^*)/\partial\theta\xrightarrow{d}Z$ for some random variable Z along $\{P_n\}_{n\geq1}$,

(e) $\sup_{\theta\in\mathcal{N}}||\partial^2Q_n(\theta)/\partial\theta\partial\theta'-H(\theta)||=o_{P_n}(1)$ for some function $H:\mathcal{N}\to\mathbb{R}^{k\times k}$ that is continuous at θ^*,

(f) $H(\theta^*)$ is non-singular.

Then, $n^\delta(\hat{\theta}_n-\theta^*)=-H(\theta^*)^{-1}n^\delta\partial Q_n(\theta^*)/\partial\theta+o_{P_n}(1)\xrightarrow{d}-H(\theta^*)^{-1}Z$ along $\{P_n\}_{n\geq1}$.

References

Aguirregabiria, V. and P. Mira (2002): "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models," *Econometrica*, 70, 1519–1543.

——— (2010): "Dynamic Discrete Choice Structural Models: A Survey," *Journal of Econometrics*, 156, 38–67.

Arcidiacono, P. and P. B. Ellickson (2011): "Practical Methods for Estimation of Dynamic Discrete Choice Models," *Annual Review of Economics*, 3, 363–394.

Blackwell, D. (1965): "Discounted Dynamic Programming," *The Annals of Mathematical Statistics*, 36, 226–235.

Bugni, F. A., I. A. Canay, and P. Guggenberger (2012): "Distortions of Asymptotic Confidence Size in Locally Misspecified Moment Inequality Models," *Econometrica*, 80, 1741–1768.

Chesher, A. (1991): "The Effect of Measurement Error," *Biometrika*, 78, 451–462.

Davidson, J. (1994): *Stochastic Limit Theory*, Oxford University Press.

Gourieroux, C. and A. Monfort (1995): *Statistics and Econometric Models: Volume 2*, Cambridge University Press.

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

Hotz, J. V., R. T. A. Miller, S. Sanders, and J. Smith (1994): "A Simulation Estimator for Dynamic Models of Discrete Choice," *Review of Economics Studies*, 61, 265–289.

Kasahara, H. and K. Shimotsu (2008): "Pseudo-likelihood Estimation and Bootstrap Inference for Structural Discrete Markov Decision Models," *Journal of Econometrics*, 146, 92–106.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013): "Robustness, Infinitesimal Neighborhoods, and Moment Restrictions," *Econometrica*, 81, 1185–1201.

Magnac, T. and D. Thesmar (2002): "Identifying Dynamic Discrete Decision Processes," *Econometrica*, 70, 801–816.

McFadden, D. and W. K. Newey (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Elsevier, vol. 4 of *Handbook of Econometrics*, 2111–2245.

Newey, W. K. (1985a): "Generalized Method of Moments Specification Testing," *Journal of Econometrics*, 29, 229–256.

——— (1985b): "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, 5, 1047–1070.

Pesendorfer, M. and P. Schmidt-Dengler (2008): "Asymptotic Least Squares Estimators for Dynamic Games," *Review of Economic Studies*, 75, 901–928.

ROTHENBERG, T. J. (1971): "Identification in Parametric Models," *Econometrica*, 39, 577–591.

ROYDEN, H. L. (1988): *Real Analysis*, Prentice-Hall.

RUST, J. (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 55, 999–1033.

——— (1988): "Maximum Likelihood Estimation of Discrete Control Processes," *SIAM J. Control and Optimization*, 26, 1006–1024.

SCHORFHEIDE, F. (2005): "VAR Forecasting under Misspecification," *Journal of Econometrics*, 128, 99–136.

SEBER, G. A. F. (2008): *A Matrix Handbook for Statisticians*, John Wiley and Sons, Inc.

WHITE, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 681–700.

——— (1996): *Estimation, Inference and Specification Analysis*, Econometric Society Monographs No. 22, Cambridge University Press.