Quantitative Economics 10 (2019), 67–103 1759-7331/20190067

Inference in dynamic discrete choice problems under local misspecification

Federico A. Bugni
Department of Economics, Duke University

Takuya Ura
Department of Economics, University of California, Davis

Single-agent dynamic discrete choice models are typically estimated using heavily parametrized econometric frameworks, making them susceptible to model misspecification. This paper investigates how misspecification affects the results of inference in these models. Specifically, we consider a local misspecification framework in which specification errors are assumed to vanish at an arbitrary and unknown rate with the sample size. Relative to global misspecification, the local misspecification analysis has two important advantages. First, it yields tractable and general results. Second, it allows us to focus on parameters with structural interpretation, instead of "pseudo-true" parameters.

We consider a general class of two-step estimators based on the K-stage sequential policy function iteration algorithm, where K denotes the number of iterations employed in the estimation. This class includes 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 the asymptotic distribution and even the rate of convergence of these estimators. In principle, one might expect that the effect of the local misspecification could change with the number of iterations K. One of our main findings is that this is not the case, that is, the effect of local misspecification is invariant to K. In practice, this means that researchers cannot eliminate or even alleviate problems of model misspecification by choosing K.

Keywords. Single-agent dynamic discrete choice models, estimation, inference, misspecification, local misspecification.

JEL classification. C13, C61, C73.

Federico A. Bugni: federico.bugni@duke.edu
Takuya Ura: takura@ucdavis.edu

We thank the three anonymous referees for comments and suggestions that have significantly improved this paper. We are also grateful for helpful discussions with Victor Aguirregabiria, Peter Arcidiacono, Joe Hotz, Shakeeb Khan, Matt Masten, Arnaud Maurel, Jia Li, and the participants at the Duke Microeconometrics Reading Group and the Yale Econometrics Lunch Group. Any and all errors are our own. The research of the first author was supported by NIH Grant 40-4153-00-0-85-399 and NSF Grant SES-1729280.

 https://doi.org/10.3982/QE917

1. Introduction

This paper investigates the effect of model misspecification on inference in single-agent 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, several methods 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. To the best of our knowledge, our paper is among the first ones to investigate the effect of misspecification on inference in these types of models.

In this paper, we propose a local misspecification approach in dynamic discrete choice models. By local misspecification, we mean that the econometric model is allowed to be misspecified, but the amount of misspecification vanishes as the sample size increases. Local misspecification is an asymptotic device that can provide concrete conclusions in the presence of misspecification while keeping the analysis tractable. As with any other asymptotic device, local misspecification is just an approximation to a finite sample situation (in this case, with misspecification) 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 the mistakes in the specification of the model are small. Second, there are multiple available estimation methods, and their performance under misspecification is not well understood. We believe their relative performance under local misspecification is a relevant comparison criterion. Finally, while our approach to misspecification is admittedly "local" in that it is assumed to vanish, we allow the rate at which this occurs to be completely arbitrary. In particular, we allow the misspecification to disappear at a faster, equal, or even slower rate than the regular parametric convergence rate of $\sqrt{n}$. While this rate will affect the asymptotic properties of the estimators under consideration, it will not alter the main qualitative conclusions of our paper.

We consider a class of two-step estimators based on the K-stage 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-stage maximum likelihood type estimators (K-ML) and the K-stage minimum distance type estimators (K-MD). This class includes most of the previously mentioned estimators as special cases.

Our main theoretical contribution is to characterize the asymptotic distribution of two-step K-MD and K-ML estimators under local misspecification. We show that local

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

misspecification can affect the asymptotic distribution and even the rate of convergence of these estimators. We are particularly interested in the asymptotic behavior of these estimators as we vary the number of iterations K.

We obtain three main results. Our first result is related to the asymptotic behavior of K-ML estimators. Under correct specification, Aguirregabiria and Mira (2002) proved that the asymptotic distribution of K-ML estimators is invariant to K. Under local misspecification, however, one might reasonably expect a different result. Intuitively, every stage of the policy function iteration algorithm brings the estimator closer to imposing the fixed-point/equilibrium conditions implied by the model. Given that the model is incorrectly specified, one might then conjecture that increasing the number of iterations would result in an estimator of inferior quality (e.g., more bias).[2] Our first main result is to show that this intuition is incorrect. We formally show that K-ML estimators are asymptotically equivalent for all K. Our second result is to show an analogous result for K-MD estimators, that is, given the choice of weight matrix, K-MD estimators are asymptotically equivalent for all K. If we combine these findings, we can conclude that the researcher cannot eliminate or even alleviate a problem of model misspecification by choosing the number of iterations K. Additional iterations are computationally costly and produce *no change* in asymptotic efficiency. Thus, from a practical viewpoint, we recommend using either the 1-ML or the 1-MD estimator.

Finally, our third result is to compare K-MD and K-ML estimators in terms of asymptotic mean squared error. We show that an optimally-weighted K-MD estimator depends on the unknown asymptotic bias and is thus generally unfeasible. In turn, the feasible K-MD estimator with a weight matrix that minimizes asymptotic variance could have an asymptotic mean squared error that is higher or lower than that of the K-ML estimator or the K-MD estimator with identity weight matrix. In other words, given a particular choice of the number of iterations K (e.g., $K = 1$), the presence of local misspecification implies that we cannot make clear-cut recommendations regarding the weight matrix for the K-MD estimator, and how this compares with the K-ML estimator.

From a technical viewpoint, our analysis exploits a distinctive feature of single-agent dynamic discrete choice problems known as the "zero Jacobian property." This property was used by Aguirregabiria and Mira (2002) under correct specification to obtain their results for K-ML estimators. One of our technical contributions is to use this property under local misspecification to derive analogous results for both K-ML and K-MD estimators.

As we have explained, this paper uses a local approach to the problem of model misspecification. In practice, researchers typically specify econometric models that may contain nonvanishing errors, that is, global misspecification. Relative to the global misspecification analysis, our local misspecification approach has two important advantages. First, allowing for global misspecification in our dynamic discrete choice model typically makes the problem intractable, and generally valid results are thus hard to obtain. In contrast, the local misspecification yields concrete and general conclusions. Second, recall that the literature has produced several estimation methods to estimate

[2]We are grateful to an anonymous referee for suggesting this interpretation.

the structural parameter of interest in a dynamic discrete choice problem. Under global misspecification, the different estimators typically converge in probability to different pseudo-true parameters which may or may not be related to the true structural parameter. This makes the results hard to interpret and compare. In contrast, under local misspecification, these different estimators are shown to consistently estimate the true structural parameter value. We can then compare their robustness to misspecification via their asymptotic distributions.

This paper relates to a vast literature on inference under model misspecification. White (1982, 1996) considered the problem of maximum likelihood estimation under global misspecification. Newey (1985a,b) and Tauchen (1985) investigated the power properties of the model specification tests under local misspecification. More recently, Schorfheide (2005) considered a locally misspecified vector autoregression process and proposed an information criterion for the lag length in the autoregression model. Bugni, Canay, and Guggenberger (2012) compared inference methods in partially identified moment (in)equality models that are locally misspecified. Kitamura, Otsu, and Evdokimov (2013) considered a class of estimators that are robust to local misspecification in the context of moment condition models. None of the previously mentioned references consider inference in dynamic discrete choice problems, whose specific features are central to the results in this paper. Few references explore the issue of misspecification in dynamic discrete choice problems. For example, Norets and Takahashi (2013) considered surjective dynamic discrete choice models and show that these are necessarily correctly specified. Last, Chernozhukov et al. (2016) considered two-step estimators in a dynamic discrete choice model with a locally misspecified first step, and proposed estimators that are robust to this issue. In contrast, this paper allows both steps to be locally misspecified.

The remainder of the paper is structured as follows. Section 2 describes the dynamic discrete choice model and introduces the possibility of its local misspecification. Section 3 develops a general result for two-step K-stage estimators under high-level conditions. Section 4 applies the general result to K-ML estimators (Section 4.1) and K-MD 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 $s_1, s_2 \in \mathbb{N}$, $\mathbf{0}_{s_1 \times s_2}$ and $\mathbf{1}_{s_1 \times s_2}$ denote a $(s_1 \times s_2)$-dimensional matrix composed of zeros and ones, respectively, and $\mathbf{I}_{s_1 \times s_2}$ denotes $(s_1 \times s_2)$-dimensional matrix equal to the left upper block of the $(\max(s_1, s_2) \times \max(s_1, s_2))$-dimensional identity matrix. We use $\|\cdot\|$ to denote the Euclidean norm. For any s-dimensional column vector V, $\text{diag}\{V\}$ is the $(s \times s)$-dimensional matrix with V as its diagonal. 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 $(|S_1| \times |S_2|)$-dimensional column vector equal to the vectorization of $\{\{M(s_1, s_2)\}_{s_1=1}^{|S_1|}\}_{s_2=1}^{|S_2|}$. For any differentiable matrix function $F(y) : \mathbb{R}^{a\times b} \to \mathbb{R}^{c\times d}$, $\partial F(y)/\partial y \in \mathbb{R}^{ac\times bd}$ denotes the usual matrix of derivatives. Finally, "w.p.a.1" abbreviates "with probability approaching one."

2. Setup

Section 2.1 describes the dynamic discrete choice model assumed by the researcher. This paper allows this model to be incorrectly specified. Section 2.2 describes the nature of the model misspecification.

2.1 *The econometric model*

An economic agent is assumed to behave 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 $s_t = (x_t, \varepsilon_t)$ 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 $\varepsilon_t \in \mathbb{R}^{|A|}$ represents an action-specific state vector only observed by the agent.

The agent's future state variables $(x_{t+1}, \varepsilon_{t+1})$ are assumed to follow a Markov transition probability density $d\Pr(x_{t+1}, \varepsilon_{t+1}|x_t, \varepsilon_t, a_t)$ that satisfies:

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

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

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

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

where $u_{\theta_u}(\cdot)$ is nonstochastic component of the current utility with parameter θ_u, and $\varepsilon_t(a_t)$ denotes the a_tth 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. Also, we denote $\theta = (\alpha, \theta_f) \in \Theta \equiv \Theta_\alpha \times \Theta_f$ with $\alpha \equiv (\theta_u, \theta_g) \in \Theta_\alpha$.

Following Aguirregabiria and Mira (2002), we impose the following regularity conditions on the primitive elements of the econometric model.

Assumption 1. *For every $\theta \in \Theta$, assume that*:

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

(b) *$\varepsilon = \{\varepsilon(a)\}_{a \in A}$ has full support,*

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

[3]This follows Aguirregabiria and Mira (2002, footnote 12) and Magnac and Thesmar (2002).

By Blackwell (1965)'s theorem and its generalization by Rust (1988), the optimal decision rule is stationary and Markovian, that is, the time subscript can be dropped. Furthermore, the optimal value function V_θ is the unique solution of the following Bellman equation:

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

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

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

which is the unique solution of the smoothed Bellman equation:

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

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 an action in A, $P_\theta(|A||x) = 1 - \sum_{a\in\tilde{A}} P_\theta(a|x)$ for all $x \in X$. Thus, the vector of model-implied conditional choice probabilities (CCPs) is completely characterized by $P_\theta \equiv \{P_\theta(a|x)\}_{(a,x)\in\tilde{A}\times X}$ with $\tilde{A} \equiv \{1,\ldots,|A|-1\}$. For the remainder of the paper, we use $\Theta_P \subset [0,1]^{|\tilde{A}\times X|}$ to denote the parameter space for the vector of 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. By utility maximization, the vector of CCPs is determined by the following equation:

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

which can be succinctly represented as follows:

$$P_\theta = \Lambda_\theta\left(\{V_\theta(x)\}_{x\in X}\right). \quad (2.3)$$

Also, notice that Equation (2.2) can be rewritten as

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

where $E_\theta[\varepsilon(a)|x,a]$ denotes the expectation of the unobservable $\varepsilon(a)$ 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 between the CCPs and the (normalized) smoothed value function. The inverse of this mapping allows us to re-express $\{E_\theta[\varepsilon(a)|x,a]\}_{(a,x)\in A\times X}$ as a function of the vector of CCPs. By combining this with Equation (2.4), we can express $\{V_\theta(x)\}_{x\in X}$ as a function of P_θ. An explicit formula for

such function is provided in Aguirregabiria and Mira (2002, Equation (8)), which we succinctly express as follows:

$$\{V_\theta(x)\}_{x\in X} = \varphi_\theta(P_\theta). \tag{2.5}$$

By combining Equations (2.3) and (2.5), we obtain the following fixed-point representation of the vector of CCPs:

$$P_\theta = \Psi_\theta(P_\theta), \tag{2.6}$$

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 CCPs, optimal or not. For any arbitrary $P \equiv \{P(a|x)\}_{(a,x)\in \tilde{A}\times X}$, $\Psi_\theta(P)$ provides the current optimal CCPs of an agent whose future behavior is according to P.

Under the current assumptions, the policy function mapping has several properties that are central to the results of this paper.

LEMMA 2.1. *Under Assumption 1, Ψ_θ satisfies the following properties*:

(a) *Ψ_θ has a unique fixed-point P_θ,*

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

(c) *The Jacobian matrix of Ψ_θ with respect to P is zero at P_θ.*

Following the literature on estimation of dynamic discrete choice models, the researcher estimates $\theta = (\alpha, \theta_f)$ using a two-step procedure. 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 to estimate α. The following assumption ensures that the model is identified.

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

(a) *θ_f is identified by f_{θ_f}, that is, $f_{\theta_{f,a}} = f_{\theta_{f,b}}$ implies $\theta_{f,a} = \theta_{f,b}$,*

(b) *α is identified by the fixed point condition $\Psi_{(\alpha,\theta_f)}(P) = P$ for any $(\theta_f, P) \in \Theta_f \times \Theta_P$, that is, $\forall \theta_f \in \Theta_f$, $\Psi_{(\alpha_a,\theta_f)}(P) = P$ 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. *Under Assumptions 1–2,*

(a) *P_θ is continuously differentiable,*

(b) *$\partial P_\theta/\partial\theta = \partial\Psi_\theta(P_\theta)/\partial\theta$,*

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

Lemmas 2.1 and 2.2 are well-known results under correct specification. At the risk of being repetitive, we include these in the paper for two reasons. First, we note that these properties belong to the econometric model, regardless of whether it is correctly specified or not. Second, the results in the paper will repeatedly make reference to these properties.

Thus far, we have described how the model specifies two conditional distributions: the CCPs and the transition probabilities. The remaining element of the specification is the marginal distribution of the state variables, which is left completely unspecified.

2.2 *Local misspecification*

We now describe the true data generating process (DGP), denoted by $\Pi_n^*(a,x,x')$, and explain its relationship to the econometric model in Section 2.1. Hereafter, a superscript with asterisk denotes true value.

By definition, the DGP is the product of the transition probability, the CCPs, and the marginal distribution of the state variable, that is, for all $(a,x,x') \in A \times X \times X$,

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

where

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

For the same reason as before, $P_n^*(|A||x) = 1 - \sum_{a \in \tilde{A}} P_n^*(a|x)$ for all $x \in X$. Thus, the vector of true CCPs is completely characterized by $P_n^* \equiv \{P_n^*(a|x)\}_{(a,x) \in \tilde{A} \times X} \in \Theta_P$.

Section 2.1 specifies $P_\theta(a|x)$ as the econometric model for $P_n^*(a|x)$ and $f_{\theta_f}(x'|a,x)$ as the econometric model for $f_n^*(x'|a,x)$. This paper allows the econometric model to be misspecified, that is,

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

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

ASSUMPTION 3. *The model is locally misspecified in the following sense*:

(a) *The sequence of DGPs* $\{\Pi_n^*\}_{n\geq 1}$ *with* $\Pi_n^* \equiv \{\Pi_n^*(a,x,x')\}_{(a,x,x')\in A\times X\times X}$ *converges to a limiting DGP* $\Pi^* \equiv \{\Pi^*(a,x,x')\}_{(a,x,x')\in A\times X\times X}$ *in the following manner*:

$$n^{\delta}\left(\Pi_n^* - \Pi^*\right) \to B_{\Pi^*} \in \mathbb{R}^{A\times X\times X},$$

where $\delta > 0$ *is an unknown parameter to the researcher.*

(b) *The econometric model is correctly specified in the limit*:

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

where $P^* \equiv \lim_{n\to\infty} P_n^*$ *denotes the limiting vector of CCPs and* $f^* \equiv \lim_{n\to\infty} f_n^*$ *denotes the limiting transition probabilities.*

Assumption 3 describes the effect of the local misspecification on the distribution of the data. This high-level condition represents a situation in which the underlying structural features of the econometric model are "close" to the true ones. We now provide three empirically relevant illustrations that can produce this high-level condition.

As our first illustration, suppose that the researcher incorrectly specifies one of the functional forms in the econometric model. For instance, he could specify that the utility function under action $a = 1$ is a linear function of the state variable x when, in reality, this function is quadratic:

$$u_\theta(x, a = 1) = \theta_1 + \theta_2 x + \tau_n x^2.$$

The coefficient of the quadratic term, τ_n, controls the degree of misspecification. In particular, $\tau_n \to 0$ implies that, in the limit, the econometric model is correctly specified.

As a second illustration, suppose that the data are generated by the presence of unobserved heterogeneity along the lines of Arcidiacono and Miller (2011).[4] Specifically, assume that the sample is composed of two types of agents, A and B. Both types of agents behave exactly according to the model but they differ in the parameter values of any of the structural functions (e.g., the utility function). If we use $\tau_n \in (0,1)$ to denote the proportion of agents of type B in the population, the observed true CCP P_n^* is the mixture of CCPs of agents of type A and B with weights $(1-\tau_n)$ and τ_n, respectively. The model is then misspecified in the sense that it presumes a homogenous sample. As in the first illustration, $\tau_n \to 0$ implies that the model is correctly specified in the limit.

As a third illustration, suppose that the agent exhibits small departures from the predicted rational behavior according to the econometric model.[5] Given state variables (x, ε), our model predicts that the agent deterministically chooses the action a that maximizes expected discounted utility, that is,

$$P_n^*(a|x,\varepsilon) = 1\left[a = \arg\max_{\tilde{a}\in A}\left(u_{\theta_u}(x,\tilde{a}) + \beta \sum_{x'\in X} V_\theta(x') f_{\theta_f}(x'|x,\tilde{a}) + \varepsilon(\tilde{a})\right)\right].$$

[4]We thank an anonymous referee for suggesting this second illustration.

[5]We thank the Editor for providing this third illustration.

Instead, suppose that the agent stochastically chooses actions according to a multinomial distribution with choice probabilities that are increasing in the action-specific expected discounted utility. For example, given (x, ε), the agent chooses action $a \in A$ with probability:

$$P_n^*(a|x,\varepsilon) = \frac{\exp\left[\left(u_{\theta_u}(x,a) + \beta \sum_{x' \in X} V_\theta(x') f_{\theta_f}(x'|x,a) + \varepsilon(a)\right)/\tau_n\right]}{\sum_{\tilde{a} \in A}\left[\exp\left(u_{\theta_u}(x,a) + \beta \sum_{x' \in X} V_\theta(x') f_{\theta_f}(x'|x,a) + \varepsilon(\tilde{a})\right)/\tau_n\right]}. \tag{2.10}$$

The parameter $\tau_n \geq 0$ controls the degree of departure from rational behavior. Once again, note that $\tau_n \to 0$ implies that the model is correctly specified in the limit. Finally, note that the CCPs $P_n^*(a|x)$ follow from integrating ε out from $P_n^*(a|x,\varepsilon)$.

Despite being very different from a conceptual viewpoint, these three illustrations can all be framed in terms of Assumption 3. First, these examples generate a discrepancy between the model-implied CCPs P_θ and the true CCPs P_n^*, that is, $\|P_n^* - P_\theta\| > 0$ for all $\theta \in \Theta$, that is, Equation (2.9) follows. Second, in all cases, the parameter τ_n determines the amount model misspecification. If this parameter is close to zero, a continuity argument implies that the model-implied CCPs should be "close" to the true ones. In particular, if $\tau_n = O(n^{-\delta})$ for some $\delta > 0$ and if the model CCPs are sufficiently smooth, it follows that:

(a) $n^\delta(P_n^* - P^*) \to C \in \mathbb{R}^{\tilde{A} \times X}$,

(b) $P^* = P_{(\alpha^*, \theta_f^*)}$ for some $(\alpha^*, \theta_f^*) \in \Theta$.

Assumption 3 then follows from this and the correct specification of the transition probabilities. We note in passing that the first illustration is used as the framework for our Monte Carlo simulations.

Assumption 3(a) requires that the local misspecification vanishes at a rate of n^δ for some $\delta > 0$. This rate depends on the difference between the model-implied CCPs and the true CCPs and, thus, it is unknown to the researcher. Our framework allows this rate to be faster, equal, or even slower than the parametric rate $\sqrt{n}$. This is more general than the typical local misspecification framework that usually restricts to $\delta \geq 1/2$ (e.g., see Newey (1985a,b), Tauchen (1985), Bugni, Canay, and Guggenberger (2012)).

Under these conditions, Theorem 2.1 demonstrates that there is a unique true limiting parameter value (α^*, θ_f^*). As we later show, the estimators considered in this paper will converge in probability to this parameter value despite the (local) misspecification.

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

3. General result for two-step K-stage estimators

This paper considers two-step estimators based on the K-stage sequential policy function iteration (PI) algorithm developed by Aguirregabiria and Mira (2002).[6] For any $K \in \mathbb{N}$, this estimator is defined as follows:

- Stage 1: Estimate θ_f^* with a first-step estimator, denoted by $\hat{\theta}_{f_n}$. Also, estimate P^* with the initial or 0-stage estimator of the CCPs, denoted by $\hat{P}_n^0$.
- Stage 2: Estimate α^* with $\hat{\alpha}_n^K$, computed using the following algorithm. Initialize $k = 1$ and then:

(a) Compute:

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

where $Q_n : \Theta_\alpha \times \Theta_f \times \Theta_P \to \mathbb{R}$ is the sample objective function. If $k = K$, exit the algorithm. If $k < K$, go to (b).

(b) Estimate P^* with the k-stage estimator of the CCPs, given by

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

Then increase k by one unit and return to (a).

For any $K \in \mathbb{N}$, we estimate $\theta^* = (\theta_f^*, \alpha^*)$ with $\hat{\theta}_n^K \equiv (\hat{\theta}_{f_n}, \hat{\alpha}_n^K)$. This algorithm leaves several aspects of the estimation method unspecified: the estimators $\hat{\theta}_{f_n}$ and $\hat{P}_n^0$ and the sample criterion function Q_n. Our strategy in this section is to produce a general result without specifying these objects and based on high-level conditions. In Section 4, we apply this general result to concrete estimators used in practice.

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

Assumption 5. *Let $\mathcal{N}$ denote an arbitrary small neighborhood of (α, θ_f, P) around $(\alpha^*, \theta_f^*, P^*)$. Then there is a limiting function $Q_\infty : \Theta_\alpha \times \Theta_f \times \Theta_P \to \mathbb{R}$ such that*:

(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) *For any $\lambda \in \{\alpha, \theta_f, P\}$, $\partial^2 Q_n(\alpha, \theta_f, P)/\partial\alpha\partial\lambda'$ is a continuous function for all $(\alpha, \theta_f, P) \in \mathcal{N}$ w.p.a.*1.

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

(e) *$\partial^2 Q_\infty(\alpha, \theta_f, P)/\partial\alpha\partial\alpha'$ is a continuous function and nonsingular at $(\alpha^*, \theta_f^*, P^*)$.*

(f) *For any $\lambda \in \{\alpha, \theta_f\}$, $\partial^2 Q_\infty(\alpha^*, \theta_f^*, P^*)/\partial\lambda\partial P' = \mathbf{0}_{d_\lambda \times |\tilde{A} \times X|}$.*

[6] Results for single-step estimators are easy to derive from our analysis by considering the special case in which the entire parameter vector is estimated on the second step.

ASSUMPTION 6. *The following results hold*:

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

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

The second step of the K-stage PI algorithm is an iterative version of an extremum estimator. Assumption 4 is a standard assumption for extremum estimators, which allows us to rely on the first order conditions of the optimization in Equation (3.1). With the exception of Assumption 5(f), Assumption 5 is composed of the usual regularity conditions for extremum estimators under a drifting sequence of DGPs. Assumption 5(f) is critical to establish the main result in this section and we will show that it is a consequence of the zero Jacobian property proved in Lemma 2.1(d). Assumption 6 requires that certain random variables converge in distribution or are bounded in probability. The rate of convergence for these variables is $n^{\min\{1/2,\delta\}}$, that is, the slowest rate between the local misspecification and the regular rate for parametric estimation. Assumption 6(a) does not specify the distribution of ζ, as this is not required to establish the general result in this section. For the estimators that we consider in Section 4, ζ is shown to be a multivariate normal random variable with possibly nonzero mean. Finally, note that Assumptions 5(f), (d), and 6(b) use the subscript p_n to refer to a drifting sequence of DGPs. This is necessary in our paper to handle the presence of the local misspecification.

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

THEOREM 3.1 (General Result). *Suppose Assumptions 1–6. For any* $K \geq 1$,

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

The result reveals two important features of the asymptotic distribution of our estimators. First, the local misspecification vanishing at the rate of n^δ causes the estimator to converge to the true limiting structural parameter at a rate of $n^{\min\{1/2,\delta\}}$. Second, the asymptotic distribution is invariant to the number of iterations K. In fact, the first equality in Theorem 3.1 implies that changes in K are asymptotically irrelevant. This result holds regardless of the rate of local misspecification δ. The invariance of the asymptotic distribution to K is one of the main findings of this paper. The intuition of this result is

as follows. By evaluating Equation (3.1) at $k = K$, we find that

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

This equation reveals that local misspecification can affect the asymptotic distribution of $\hat{\alpha}_n^K$ via three channels: the sample criterion function Q_n, the first-step estimator $\hat{\theta}_{f,n}$, and the $(K-1)$-stage estimator of the CCPs $\hat{P}_n^{K-1}$. As the notation shows, the third channel depends explicitly on K, and its asymptotic effect could potentially change with every iteration. Theorem 3.1 indicates that this is not the case, and we now provide some intuition. Our proof reveals that the zero Jacobian property (embodied in Assumption 5(f)) effectively erases the accumulated effect of the model misspecification from all the preceding iterations. Our formal arguments also show that the effect of the model misspecification on the last iteration is invariant with the number of iterations. From these two observations, the invariance result follows.

4. Applications of the general result

We now apply Theorem 3.1 to classes of estimators used in practice. Section 4.1 considers K-ML estimation and Section 4.2 considers K-MD estimation.

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

Assumption 7. *For each $n \in \mathbb{N}$, $\{(a_i, x_i, x_i')\}_{i \le n}$ 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 DGP, the CCPs, and the transition probabilities, that is, for all $(a, x, x') \in A \times X \times X$,

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

We now consider a general framework for the preliminary estimators in the algorithm.

Assumption 8. *For $\hat{\Pi}_n \equiv \{\hat{\Pi}_n(a, x, x')\}_{(a,x,x') \in A \times X \times X}$, assume that*

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

where the function $G : \mathbb{R}^{|A\times X\times X|} \to \mathbb{R}^{d_{\theta_f}} \times \Theta_P$ *is continuously differentiable at* Π^* *and* $(\theta_f^*, P^*) = G(\Pi^*)$.

Assumption 8 is very mild. By the identification result in Theorem 2.1 and Assumption 7, it is reasonable to presume that the researcher estimates (θ_f^*, P^*) using a smooth function of the sample analogue estimator of the DGP. In fact, Assumption 8 is automatically satisfied if we use a sample analogue estimator of the CCPs, that is, $\hat{P}_n^0 = \hat{P}_n$, and a nonparametric model for the transition probability that is estimated by sample analogues, that is, $\theta_f \equiv \{f(x'|a,x)\}_{(a,x,x')\in A\times X\times X}$ and $\hat{\theta}_{f,n} = \hat{f}_n$.[7]

4.1 *K-ML estimators*

We now specialize the general result to 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, that is,

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

To derive the asymptotic distribution of the K-ML estimator, we impose the following regularity conditions.

Assumption 9. *Ψ satisfies the following properties*:

(a) $\Psi_\theta(P)(a|x) \in (0,1)$ *for any* $(a,x) \in \tilde{A}\times X$.

(b) $\Psi_\theta(P)$ *is twice continuously differentiable in* θ *and* P.

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

Assumption 9 are connected with the requirements in Assumption 5 and are standard in the literature. Assumption 9(a)–(b) are identical to Aguirregabiria and Mira (2002, conditions (b)–(c) of Proposition 4). Assumption 9(c) is connected to the nonsingularity requirement in Assumption 5(e). Since the α has been assumed to be identified by $\Psi_\theta(P) = P$ (and, thus, locally identified by it), Assumption 9(c) 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 (K-ML). *Suppose Assumptions 1–4 and 7–9. Then, for any* $K, \tilde{K} \geq 1$,

$$\begin{aligned} &n^{\min\{1/2,\delta\}}\big(\hat{\alpha}_n^{K\text{-ML}} - \alpha^*\big) \\ &\quad = n^{\min\{1/2,\delta\}}\big(\hat{\alpha}_n^{\tilde{K}\text{-ML}} - \alpha^*\big) + o_{p_n}(1) \\ &\quad \xrightarrow{d} Y_{\mathrm{ML}} \times \Delta \times N\big(B_{\Pi^*} \times 1[\delta \leq 1/2], \big(\mathrm{diag}\big(\Pi^*\big) - \Pi^*\Pi^{*\prime}\big) \times 1[\delta \geq 1/2]\big) \end{aligned}$$

[7] One practical problem with the sample analogue estimators is that they are undefined if $\sum_{x'\in X}\hat{\Pi}_n(a,x,x') = 0$ for any $(a,x) \in A\times X$. For a discussion of this, see Hotz et al. (1994) and Pesendorfer and Schmidt-Dengler (2008, p. 914). Of course, this is only a problem in small samples and does not affect the validity of our asymptotic arguments.

where B_{Π^*} and Π^* are as in Assumption 3, and Υ_{ML} and Δ are the following matrices:

$$\begin{aligned}\Upsilon_{\mathrm{ML}} &\equiv \left(\frac{\partial P'_{\theta^*}}{\partial \alpha}\Phi\frac{\partial P_{\theta^*}}{\partial \alpha'}\right)^{-1}\frac{\partial P'_{\theta^*}}{\partial \alpha}\Phi\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta'_f}\right] \in \mathbb{R}^{d_\alpha \times (|A\times X|+d_{\theta_f})},\\ \Delta &\equiv \begin{bmatrix}\mathbf{I}_{|A\times X|\times|A\times X|} & \mathbf{I}_{|A\times X|\times|A\times X|} & \cdots & \mathbf{I}_{|A\times X|\times|A\times X|}\\ & \dfrac{\partial G_1(\Pi^*)}{\partial \Pi^{*\prime}} & & \end{bmatrix} \in \mathbb{R}^{(|A\times X|+d_{\theta_f})\times|A\times X\times X|},\end{aligned} \tag{4.2}$$

where G_1 denotes the first component of G in Assumption 8, that is, $\hat{\theta}_{f,n} \equiv G_1(\hat{\Pi}_n)$,

$$\begin{aligned}\Phi &\equiv \begin{bmatrix}\Phi_1 & \mathbf{0}_{|\tilde{A}|\times|\tilde{A}|} & \cdots & \mathbf{0}_{|\tilde{A}|\times|\tilde{A}|}\\ \mathbf{0}_{|\tilde{A}|\times|\tilde{A}|} & \Phi_2 & \cdots & \vdots\\ \vdots & \ddots & \ddots & \mathbf{0}_{|\tilde{A}|\times|\tilde{A}|}\\ \mathbf{0}_{|\tilde{A}|\times|\tilde{A}|} & \cdots & \mathbf{0}_{|\tilde{A}|\times|\tilde{A}|} & \Phi_{|X|}\end{bmatrix} \in \mathbb{R}^{|\tilde{A}\times X|\times|\tilde{A}\times X|},\\ \Sigma &\equiv \begin{bmatrix}\Sigma_1 & \mathbf{0}_{|\tilde{A}|\times|A|} & \cdots & \mathbf{0}_{|\tilde{A}|\times|A|}\\ \mathbf{0}_{|\tilde{A}|\times|A|} & \Sigma_2 & \cdots & \vdots\\ \vdots & \ddots & \ddots & \mathbf{0}_{|\tilde{A}|\times|A|}\\ \mathbf{0}_{|\tilde{A}|\times|A|} & \cdots & \mathbf{0}_{|\tilde{A}|\times|A|} & \Sigma_{|X|}\end{bmatrix} \in \mathbb{R}^{|\tilde{A}\times X|\times|A\times X|},\end{aligned} \tag{4.3}$$

and, finally, for all $x \in X$,

$$\begin{aligned}\Phi_x &\equiv m^*(x)\left[\mathrm{diag}\left\{\left\{1/P^*(a|x)\right\}_{a\in\tilde{A}}\right\} + \mathbf{1}_{|\tilde{A}|\times|\tilde{A}|}/\left(1-\sum_{a\in\tilde{A}}P^*(a|x)\right)\right] \in \mathbb{R}^{|\tilde{A}|\times|\tilde{A}|},\\ \Sigma_x &\equiv \left[\mathbf{I}_{|\tilde{A}|\times|A|} - \left\{P^*(a|x)\right\}_{a\in\tilde{A}} \times \mathbf{1}_{1\times|A|}\right]/m^*(x) \in \mathbb{R}^{|\tilde{A}|\times|A|},\\ m^*(x) &\equiv \sum_{(a,x')\in A\times X}\Pi^*(a,x,x') \in \mathbb{R}.\end{aligned}$$

As discussed in Theorem 3.1, Theorem 4.1 shows that the K-ML estimator converges at a rate of $n^{\min\{1/2,\delta\}}$ and that changes in K have a no effect on its asymptotic distribution. Theorem 4.1 also reveals that the asymptotic distribution is normal with asymptotic bias and variance given by

$$\begin{aligned}\mathrm{AB}_{\mathrm{ML}} &= \Upsilon_{\mathrm{ML}} \times \Delta \times B_{\Pi^*} \times 1[\delta \le 1/2],\\ \mathrm{AV}_{\mathrm{ML}} &= \Upsilon_{\mathrm{ML}} \times \Delta \times \left(\mathrm{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\right) \times \Delta' \times \Upsilon'_{\mathrm{ML}} \times 1[\delta \ge 1/2].\end{aligned}$$

In the case of $\delta > 1/2$, the local misspecification is irrelevant relative to sampling error and has no effect on the rate of convergence or the asymptotic distribution. A very different situation occurs when $\delta < 1/2$. In this case, the local misspecification is overwhelming relative to sampling error and dominates the asymptotic distribution. The rate of convergence of the estimator is n^δ and, at this rate, the asymptotic distribution is a pure bias term. Finally, we have a knife-edge case with $\delta = 1/2$. The rate of convergence

is the usual parametric rate $\sqrt{n}$, but the asymptotic distribution can be biased. In consequence, an adequate characterization of the estimator's precision is the asymptotic mean squared error:

$$Y_{\mathrm{ML}} \times \Delta \times \left(\mathrm{diag}(\Pi^*) - \Pi^*\Pi^{*\prime} + B_{\Pi^*}B_{\Pi^*}'\right) \times \Delta' \times Y_{\mathrm{ML}}'.$$

The K-ML estimator is a "partial" ML estimator in the sense that it plugs in the first-step estimator into the second step. In other words, it is not a "full" maximum likelihood estimator with respect to the entire parameter vector $\theta = (\alpha, \theta_f)$. 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 can be more efficient than the K-ML estimator, even in the absence of local misspecification.

Remark 4.1. Theorem 4.1 applies to any $K \in \mathbb{N}$ but does not extend to $K \to \infty$. Under some additional conditions, however, Aguirregabiria and Mira (2002, Proposition 3) showed that if the K-ML estimator converges as $K \to \infty$, it will do so to a solution of Rust (1987)'s nested fixed- point estimator. As noted in Aguirregabiria and Mira (2002, footnote 16), this result presumes the convergence of the K-ML estimator as $K \to \infty$, which has not been shown in the literature.

4.2 *K-MD estimators*

To specialize the general result to K-MD estimators, we set the sample objective function Q_n to

$$Q_n^{\mathrm{MD}}(\alpha, \theta_f, P) \equiv -\left[\hat{P}_n - \Psi_{(\alpha,\theta_f)}(P)\right]'\hat{W}_n\left[\hat{P}_n - \Psi_{(\alpha,\theta_f)}(P)\right], \tag{4.4}$$

where $\hat{P}_n$ is the sample frequency estimator of the vector of CCPs in Equation (4.1) and $\hat{W}_n \in \mathbb{R}^{|\tilde{A}\times X|\times|\tilde{A}\times X|}$ is the weight matrix. In the special case of $K = 1$ and $\hat{P}_n^0 = \hat{P}_n$, the K-MD estimator coincides with the estimators considered in Hotz and Miller (1993) and Pesendorfer and Schmidt-Dengler (2008).[8] We impose the following condition regarding the weight matrix.

Assumption 10. *$\hat{W}_n = W^* + o_{p_n}(1)$, where $W^* \in \mathbb{R}^{|\tilde{A}\times X|\times|\tilde{A}\times X|}$ is positive definite and symmetric.*

In principle, we could generalize Assumption 10 by allowing the weight matrix to be a function of the parameters of the problem. Similar results would then follow from longer arguments.

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

[8]To be precise, Pesendorfer and Schmidt-Dengler (2008, Equations (18)–(19)) consider a sample criterion function that allows $\hat{P}_n$ in Equation (4.4) to differ from the sample frequency estimator. We could also incorporate this feature in our setup at the expense of using longer arguments.

THEOREM 4.2 (K-MD). *Suppose Assumptions 1–4 and 7–10. Then, for any $K, \tilde{K} \geq 1$,*

$$\begin{aligned} &n^{\min\{1/2,\delta\}}\big(\hat{\alpha}_n^{K\text{-MD}} - \alpha^*\big) \\ &\quad = n^{\min\{1/2,\delta\}}\big(\hat{\alpha}_n^{\tilde{K}\text{-MD}} - \alpha^*\big) + o_{p_n}(1) \\ &\quad \xrightarrow{d} Y_{\text{MD}}(W^*) \times \Delta \times N\big(B_{\Pi^*} \times 1[\delta \leq 1/2], \big(\text{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\big) \times 1[\delta \geq 1/2]\big) \end{aligned}$$

where B_{Π^*} and Π^* are as in Assumption 3, $Y_{\text{MD}}(W^*)$ is the following matrix:

$$Y_{\text{MD}}(W^*) \equiv \left(\frac{\partial P_{\theta^*}'}{\partial \alpha} W^* \frac{\partial P_{\theta^*}}{\partial \alpha'}\right)^{-1} \frac{\partial P_{\theta^*}'}{\partial \alpha} W^* \left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right] \in \mathbb{R}^{d_\alpha \times (|A\times X| + d_{\theta_f})},$$

with Δ is as in Equation (4.2) and Σ is as in Equation (4.3).

REMARK 4.2. The asymptotic distribution of the K-ML estimator is a special case of that of the K-MD estimator with $W^* = \Phi$.

As discussed in previous theorems, Theorem 4.2 shows that the K-MD estimator converges at a rate of $n^{\min\{1/2,\delta\}}$ and that changes in K have no effect on its asymptotic distribution. Theorem 4.2 also reveals that the asymptotic distribution is normal with asymptotic bias and variance given by

$$\begin{aligned} \text{AB}_{\text{MD}}(W^*) &= Y_{\text{MD}}(W^*) \times \Delta \times B_{\Pi^*} \times 1[\delta \leq 1/2], \\ \text{AV}_{\text{MD}}(W^*) &= Y_{\text{MD}}(W^*) \times \Delta \times \big(\text{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\big) \times \Delta' \times Y_{\text{MD}}(W^*)' \times 1[\delta \geq 1/2]. \end{aligned} \tag{4.5}$$

In the knife-edge case with $\delta = 1/2$, the asymptotic variance and bias can coexist. In consequence, an adequate characterization of the estimator's precision is the asymptotic mean squared error:

$$Y_{\text{MD}}(W^*) \times \Delta \times \big(\text{diag}(\Pi^*) - \Pi^*\Pi^{*\prime} + B_{\Pi^*}B_{\Pi^*}'\big) \times \Delta' \times Y_{\text{MD}}(W^*)'. \tag{4.6}$$

We now briefly discuss the optimality in the choice of W^* in K-MD estimation. First, consider the case in which local misspecification is asymptotically irrelevant, that is, $\delta > 1/2$. In this case, the K-MD estimator presents no asymptotic bias and the asymptotic variance and mean squared error coincide. Provided that relevant matrices are nonsingular, standard arguments in GMM estimation imply that the minimum asymptotic variance and mean squared error among K-MD estimators are both equal to

$$\left(\frac{\partial P_{\theta^*}}{\partial \alpha'}' \left[\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right] \Delta\big(\text{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\big)\Delta' \left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right]'\right]^{-1} \frac{\partial P_{\theta^*}}{\partial \alpha'}\right)^{-1}.$$

This minimum can be achieved by the following "feasible" choice of limiting weight matrix:

$$W_{\text{AV}}^* \equiv \left[\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right] \Delta\big(\text{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\big)\Delta' \left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right]'\right]^{-1}. \tag{4.7}$$

We say that Equation (4.7) is a feasible choice because it can be consistently estimated. As pointed out in Remark 4.2, the K-ML estimator has the same asymptotic distribution as the K-MD estimator with $W^*_{\text{ML}} = \Phi$. Except under special conditions on the econometric model (e.g. $\partial P_{\theta^*}/\partial \theta_f' = \mathbf{0}_{|\tilde{A}\times X|\times d_{\theta_f}}$), the K-ML estimator is not necessarily optimal among the K-MD estimators.

Next, consider the knife-edge case in which local misspecification vanishes at the rate of sampling error, that is, $\delta = 1/2$. Once again, standard arguments in GMM estimation imply that the minimum asymptotic mean squared error among all K-MD estimators is

$$\left(\frac{\partial P'_{\theta^*}}{\partial \alpha}\left[\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right]\Delta\big(\text{diag}(\Pi^*) - \Pi^*\Pi^{*\prime} + B_{\Pi^*}B'_{\Pi^*}\big)\Delta'\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right]'\right]^{-1}\frac{\partial P_{\theta^*}}{\partial \alpha'}\right)^{-1}.$$

According to McFadden and Newey (1994, p. 2165), any limiting weight matrix W^* that minimizes the asymptotic mean squared error should satisfy the following condition: for some matrix $C \in \mathbb{R}^{|\tilde{A}\times X|\times|\tilde{A}\times X|}$,

$$\begin{aligned}\frac{\partial P_{\theta^*}}{\partial \alpha'}W^* = C\frac{\partial P_{\theta^*}}{\partial \alpha'}\Bigg[\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right]\Delta\big(\text{diag}(\Pi^*) - \Pi^*\Pi^{*\prime} + B_{\Pi^*}B'_{\Pi^*}\big)\\ \times\Delta'\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right]'\Bigg]^{-1}.\end{aligned} \tag{4.8}$$

In other words, Equation (4.8) characterizes the class of optimal limiting weight matrices. A simple example of an optimal limiting weight matrix is

$$W^*_{AMSE} \equiv \left[\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right]\Delta\big(\text{diag}(\Pi^*) - \Pi^*\Pi^{*\prime} + B_{\Pi^*}B'_{\Pi^*}\big)\Delta'\left[\Sigma \quad -\frac{\partial P_{\theta^*}}{\partial \theta_f'}\right]'\right]^{-1}. \tag{4.9}$$

By Equation (4.8), the consistent estimation of any optimal limiting weight matrix requires the consistent estimation of the asymptotic bias. Unfortunately, the researcher is not aware of this feature of the population distribution and, in this sense, estimating the optimal limiting weight matrix under local misspecification is infeasible in practice. In particular, note that the feasible limiting weight matrix W^*_{AV} that minimizes asymptotic variance could fail to be optimal (i.e., Equation (4.8) may not be satisfied when $W^* = W^*_{\text{AV}}$).[9]

Finally, we could consider the case in which local misspecification is asymptotically overwhelming, that is, $\delta < 1/2$. In this case, the asymptotic distribution collapses to the pure bias term in Equation (4.5), and the asymptotic mean squared error coincides with the square of the asymptotic bias. As in the case with $\delta = 1/2$, minimizing asymptotic mean squared error is infeasible in the sense that it depends on the unknown asymptotic bias. Also, any feasible choice of limiting weight matrix such as W^*_{AV} could fail to be optimal.

[9]This is the case in some of our Monte Carlo simulations, in which using W^*_{AV} produces an asymptotic mean squared error that is larger than that the one obtained by using $\hat{W}_n = \mathbf{I}_{|\tilde{A}\times X|\times|\tilde{A}\times X|}$.

5. Monte Carlo simulations

This section investigates the finite sample performance of the two-step estimators considered in previous sections under local misspecification.

We simulate data using the classical bus engine replacement problem studied by Rust (1987). In each period $t = 1, \ldots, T \equiv \infty$, the bus owner decides 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 $a \in A = \{1, 2\}$, where $a = 2$ represents replacing the engine, $a = 1$ represents not replacing the engine, and the current engine mileage is denoted by $x \in X \equiv \{1, \ldots, 20\}$.

The researcher assumes that all individuals in his sample have the same deterministic part of the utility (profit) function, given by

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

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 errors are i.i.d. extreme value type I, independent of x, that is,

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

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

$$\begin{aligned} f_{\theta_f}\big(x'|x, a\big) = (1 - \theta_f) \times 1\big[a = 1, x' = \min\big\{x + 1, |X|\big\}\big] \\ + \theta_f \times 1\big[a = 1, x' = x\big] + 1\big[a = 2, x' = 1\big], \end{aligned} \tag{5.3}$$

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 $\alpha = \theta_u \in \Theta_\alpha = \Theta_u$.

The researcher correctly specified the error distribution and state transition probabilities, which satisfy Equation (5.3) with $\theta_f = 0.25$. Unfortunately, he does not correctly specify the utility function. The correct utility function is as follows:

$$u_{\theta_u,n}(x, a) = -\theta_{u,1} \times 1[a = 2] - \theta_{u,2} \times 1[a = 1]x + \tau_n \times 1[a = 1]x^2, \tag{5.4}$$

with $\theta_{u,1} = 1$, $\theta_{u,2} = 0.05$, and $\tau_n = -0.025n^{-\delta}$ with $\delta \in \{1/3, 1/2, 1\}$. Notice that this form of model misspecification is analogous to the first illustration discussed in Section 2.2.[10] By the arguments in Section 2, the true CCPs P_n^* are determined by the true error distribution (Equation (5.2)), true state transition probabilities (Equation (5.3) with $\theta_f = 0.25$), and the true utility function (Equation (5.4)). Our choices of δ include a case in which the local misspecification is asymptotically irrelevant ($\delta = 1$), one case in which local misspecification is the knife-edge case ($\delta = 1/2$), and one case in which the local misspecification is overwhelming ($\delta = 1/3$). In addition, we also consider a case in which the econometric model is correctly specified.

[10] Section 2.2 provides two other examples of local misspecification. We have also conducted Monte Carlo simulations based on these designs. For the sake of brevity, these are presented in the Supplemental Material (see Bugni anf Ura (2019)).

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 $n \in \{200, 500, 1000\}$. We generate marginal observations of the state variables according to the following distribution:

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

Together with previous elements, this determines the true joint DGP Π_n^* according to Equation (2.7).

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-step K-stage PI algorithm described in Sections 3–4. In the first step, the researcher estimates P^* and θ_f^* using preliminary estimators $\hat{P}_n^0 = \hat{P}_n$ and

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

In the second step, the researcher estimates $(\theta_{u,1}, \theta_{u,2})$ using the K-stage policy function iteration algorithm using criterion function Q_n 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 two limiting weight matrices: identity (i.e., $W^* = \mathbf{I}_{|\tilde{A}\times X|\times|\tilde{A}\times X|}$) and asymptotic variance minimizer (i.e., $W^* = W_{\mathrm{AV}}^*$[12]). We show results for number of stages $K \in \{1, 2, 3, 10\}$.[13]

We now describe simulation results for the estimator of $\theta_{u,2}$. We focus on $\theta_{u,2}$ because we consider the linear coefficient of the utility function to be more interesting than the constant coefficient.[14] Table 1 describes results under correct specification. As expected, all estimators appear to converge to a distribution with zero mean and finite variance. Also as expected, the number of iterations K does not seem to affect the bias or the variance of the estimators under consideration. Similar to Aguirregabiria and Mira (2002), we detect small differences between the results with $K = 1$ and those with $K > 1$, especially for the smallest sample size. This effect tends to vanish as the sample size increases. These findings could be rationalized by the higher-order analysis in Kasahara and Shimotsu (2008). The K-ML estimator and the K-MD estimator with $W^* = W_{\mathrm{AV}}^*$ are similar and more efficient than the K-MD estimator with $W^* = \mathbf{I}_{|\tilde{A}\times X|\times|\tilde{A}\times X|}$.

Table 2 provides results under asymptotically irrelevant local misspecification, that is, $\delta = 1$. According to our theoretical results, the asymptotic behavior of all estimators

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

[12] This matrix is calculated using numerical derivates and Monte Carlo integration with a sample size that is significantly larger than those used in the actual Monte Carlo simulations.

[13] 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 and are available from the authors upon request.

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

TABLE 1. Simulation results under correct specification, that is, $\tau_n = 0$.

		K-MD($\mathbf{I}_{\lvert\tilde{A}\times X\rvert\times\lvert\tilde{A}\times X\rvert}$)			K-MD(W^*_{AV})			K-ML		
K	Statistic	$n=200$	$n=500$	$n=1000$	$n=200$	$n=500$	$n=1000$	$n=200$	$n=500$	$n=1000$
1	$\sqrt{n}$ Bias	0.07	0.02	0.01	0.06	0.02	0.01	0.06	0.02	0.01
	$\sqrt{n}$ SD	0.25	0.25	0.24	0.23	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
2	$\sqrt{n}$ Bias	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00
	$\sqrt{n}$ SD	0.24	0.25	0.24	0.23	0.23	0.22	0.22	0.22	0.22
	n MSE	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
3	$\sqrt{n}$ Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sqrt{n}$ SD	0.24	0.25	0.24	0.23	0.23	0.22	0.22	0.22	0.22
	n MSE	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
10	$\sqrt{n}$ Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sqrt{n}$ SD	0.25	0.25	0.24	0.23	0.23	0.22	0.22	0.22	0.22
	n MSE	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05

TABLE 2. Simulation results under local misspecification with $\tau_n \propto n^{-1}$.

		K-MD($\mathbf{I}_{\lvert\tilde{A}\times X\rvert\times\lvert\tilde{A}\times X\rvert}$)			K-MD(W^*_{AV})			K-ML		
K	Statistic	$n=200$	$n=500$	$n=1000$	$n=200$	$n=500$	$n=1000$	$n=200$	$n=500$	$n=1000$
1	$\sqrt{n}$ Bias	0.11	0.04	0.03	0.10	0.04	0.03	0.09	0.04	0.03
	$\sqrt{n}$ SD	0.25	0.25	0.24	0.24	0.23	0.22	0.23	0.23	0.22
	n MSE	0.07	0.06	0.06	0.07	0.06	0.05	0.06	0.05	0.05
2	$\sqrt{n}$ Bias	0.04	0.03	0.02	0.04	0.03	0.02	0.04	0.03	0.02
	$\sqrt{n}$ SD	0.25	0.25	0.24	0.23	0.23	0.22	0.22	0.22	0.22
	n MSE	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
3	$\sqrt{n}$ Bias	0.04	0.03	0.02	0.04	0.03	0.02	0.04	0.03	0.02
	$\sqrt{n}$ SD	0.25	0.25	0.24	0.23	0.23	0.22	0.22	0.22	0.22
	n MSE	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
10	$\sqrt{n}$ Bias	0.04	0.03	0.02	0.04	0.03	0.02	0.04	0.03	0.02
	$\sqrt{n}$ SD	0.25	0.25	0.24	0.23	0.23	0.22	0.22	0.22	0.22
	n MSE	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05

should be identical to the correctly specified model. This is 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, that is, $\delta = 1/2$. According to our theoretical results, this should produce an asymptotic distribution that has nonzero bias and is not affected by the number of iterations K. By and large, these predictions are confirmed in our simulations. Given the presence of asymptotic bias, we now evaluate efficiency using the mean squared error. In this simulation design, the K-MD estimator with $W^* = \mathbf{I}_{\lvert\tilde{A}\times X\rvert\times\lvert\tilde{A}\times X\rvert}$ is now also slightly more efficient than the K-ML estimator or the K-MD estimator with $W^* = W^*_{\text{AV}}$. This is also consistent with our theory: while K-MD estimator with $W^* = \mathbf{I}_{\lvert\tilde{A}\times X\rvert\times\lvert\tilde{A}\times X\rvert}$ has more

TABLE 3. Simulation results under local misspecification with $\tau_n \propto n^{-1/2}$.

		K-MD($\mathbf{I}_{\lvert\tilde{A}\times X\rvert\times\lvert\tilde{A}\times X\rvert}$)			K-MD(W^*_{AV})			K-ML		
K	Statistic	$n=200$	$n=500$	$n=1000$	$n=200$	$n=500$	$n=1000$	$n=200$	$n=500$	$n=1000$
1	$\sqrt{n}$ Bias	0.50	0.46	0.46	0.51	0.49	0.50	0.52	0.50	0.50
	$\sqrt{n}$ SD	0.28	0.28	0.27	0.26	0.26	0.24	0.25	0.25	0.24
	n MSE	0.33	0.29	0.28	0.33	0.31	0.30	0.33	0.31	0.31
2	$\sqrt{n}$ Bias	0.42	0.44	0.45	0.46	0.48	0.49	0.47	0.49	0.49
	$\sqrt{n}$ SD	0.29	0.28	0.26	0.26	0.25	0.24	0.24	0.24	0.24
	n MSE	0.26	0.27	0.28	0.27	0.29	0.30	0.28	0.30	0.30
3	$\sqrt{n}$ Bias	0.42	0.44	0.45	0.46	0.48	0.49	0.47	0.49	0.49
	$\sqrt{n}$ SD	0.28	0.28	0.26	0.26	0.25	0.24	0.24	0.24	0.24
	n MSE	0.26	0.27	0.28	0.27	0.29	0.30	0.28	0.30	0.30
10	$\sqrt{n}$ Bias	0.42	0.44	0.45	0.46	0.48	0.49	0.47	0.49	0.49
	$\sqrt{n}$ SD	0.29	0.28	0.26	0.26	0.25	0.24	0.24	0.24	0.24
	n MSE	0.26	0.27	0.28	0.27	0.29	0.30	0.28	0.30	0.30

TABLE 4. Simulation results under local misspecification with $\tau_n \propto n^{-1/3}$ and using the correct scaling.

		K-MD($\mathbf{I}_{\lvert\tilde{A}\times X\rvert\times\lvert\tilde{A}\times X\rvert}$)			K-MD(W^*_{AV})			K-ML		
K	Statistic	$n=200$	$n=500$	$n=1000$	$n=200$	$n=500$	$n=1000$	$n=200$	$n=500$	$n=1000$
1	$n^{1/3}$ Bias	0.41	0.40	0.41	0.43	0.44	0.45	0.45	0.46	0.46
	$n^{1/3}$ SD	0.14	0.12	0.10	0.13	0.10	0.09	0.12	0.10	0.09
	$n^{2/3}$ MSE	0.18	0.18	0.18	0.20	0.20	0.21	0.22	0.22	0.22
2	$n^{1/3}$ Bias	0.37	0.40	0.41	0.40	0.43	0.45	0.43	0.45	0.46
	$n^{1/3}$ SD	0.14	0.12	0.10	0.12	0.10	0.09	0.11	0.10	0.09
	$n^{2/3}$ MSE	0.16	0.17	0.18	0.18	0.20	0.21	0.20	0.21	0.22
3	$n^{1/3}$ Bias	0.37	0.40	0.41	0.40	0.43	0.45	0.43	0.45	0.46
	$n^{1/3}$ SD	0.14	0.12	0.10	0.12	0.10	0.09	0.11	0.10	0.09
	$n^{2/3}$ MSE	0.16	0.17	0.18	0.18	0.20	0.21	0.20	0.21	0.22
10	$n^{1/3}$ Bias	0.37	0.40	0.41	0.40	0.43	0.45	0.43	0.45	0.46
	$n^{1/3}$ SD	0.14	0.12	0.10	0.12	0.10	0.09	0.11	0.10	0.09
	$n^{2/3}$ MSE	0.16	0.17	0.18	0.18	0.20	0.21	0.20	0.21	0.22

variance that the other two, it also appears to have less bias, resulting in less overall mean squared error.

Table 4 provides results under asymptotically overwhelming local misspecification, that is, $\delta = 1/3$. According to our theoretical results, the presence of this local misspecification changes dramatically the asymptotic distribution of all estimators. In particular, these no longer converge at the regular $\sqrt{n}$-rate, but rather at $n^{1/3}$-rate. In fact, at $\sqrt{n}$-rate, the asymptotic bias is no longer bounded (see Table S1 in the Supplemental Material). Once we scale the estimators at the appropriate $n^{1/3}$-rate, they converge to an asymptotic distribution dominated by the bias. Furthermore, our theoretical results in-

dicate that the number of iterations K does not affect this asymptotic distribution. These predictions are clearly depicted in Table 4. In line with the results in Table 3, the K-MD estimator with $W^* = \mathbf{I}_{|\tilde{A}\times X|\times|\tilde{A}\times X|}$ is slightly more efficient than the K-ML estimator and the K-MD estimator with $W^* = W^*_{\text{AV}}$.

6. Conclusion

Single-agent 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 inference results in these models. This paper considers a *local misspecification* framework, which is an asymptotic device in which the mistake in the specification vanishes as the sample size diverges. In this paper, we impose no restrictions on the rate at which these specification errors disappear. Relative to global misspecification, the local misspecification analysis has two important advantages. First, it yields tractable and general results. Second, it allows us to focus on parameters with structural interpretation, instead of "pseudo-true" parameters.

We consider a general class of two-step estimators based on the K-stage 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 includes 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 the asymptotic distribution and even the rate of convergence of these estimators. In principle, one might expect that the effect of the local misspecification could change with the number of iterations K. The main finding in the paper is that this is not the case, that is, the effect of local misspecification is invariant to K. In particular, (a) K-ML estimators are asymptotically equivalent and (b) given the choice of the weight matrix, K-MD estimators are asymptotically equivalent. In practice, this means that researchers cannot eliminate or even alleviate problems of model misspecification by choosing K. Additional iterations are computationally costly and produce *no change* in asymptotic efficiency.

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 complicates this comparison. In the presence of asymptotic bias, the optimality of the estimator should be evaluated using the asymptotic mean squared error. We show that an optimally-weighted K-MD estimator depends on the unknown asymptotic bias and is thus generally unfeasible. In turn, the feasible K-MD estimator with a weight matrix that minimizes asymptotic variance could have an asymptotic mean squared error that is higher or lower than that of the K-ML estimator or the K-MD estimator with identity weight matrix.

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 the true DGP Π_n^*, Equation (2.8) defined transition probabilities f_n^*, CCPs P_n^*, and marginal distribution of states m_n^*. The unconditional probability of $(a, x) \in A \times X$ is analogously defined by

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

and $J_n^* \equiv \{J_n^*(a, x)\}_{(a,x)\in A\times X}$. The limiting DGP f^*, transition probabilities f^*, CCPs P^* were defined in Assumption 3. The other limiting objects are analogously defined by $J^* \equiv \lim_{n\to\infty} J_n^*$ and $m^* \equiv \lim_{n\to\infty} m_n^*$.

The sample analogue DGP $\hat{\Pi}_n$, transition probabilities $\hat{f}_n$, CCPs $\hat{P}_n$ were defined in Equation (4.1). The sample analogue marginal distribution of states $\hat{m}_n$ and unconditional probabilities $\hat{J}_n$ are analogously defined. For any $(a, x) \in A \times X$,

$$\hat{m}_n(x) \equiv \sum_{(a,\tilde{x}')\in A\times X} \hat{\Pi}_n(a, x, \tilde{x}'),$$

$$\hat{J}_n(a, x) \equiv \sum_{\tilde{x}' \in X} \hat{\Pi}_n(a, x, \tilde{x}'),$$

$\hat{m}_n \equiv \{\hat{m}_n(x)\}_{x\in X}$, and $\hat{J}_n \equiv \{\hat{J}_n(a, x)\}_{(a,x)\in A\times X}$.

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, pp. 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, that is, $\exists(\alpha^*, \theta_f^*) \in \Theta$ 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 $\|f_{\theta_f^*} - f_{\tilde{\theta}_f}\| \le \|f_{\theta_f^*} - f^*\| + \|f_{\tilde{\theta}_f} - f^*\|$ 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 $\theta_f^* = \tilde{\theta}_f$. 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. Without loss of generality, we can consider the neighborhood $\mathcal{N}$ to be "rectangular," in the sense that $\mathcal{N} = \mathcal{N}_\alpha \times \mathcal{N}_{\theta_f} \times \mathcal{N}_P$, where $\mathcal{N}_\lambda$ denotes the neighborhood of λ^* for any $\lambda \in \{\alpha, \theta_f, P\}$. This can always be achieved by replacing $\mathcal{N}$ with $\tilde{\mathcal{N}} \subseteq \mathcal{N}$ that has the desired structure. This proof will make repeated reference to $\mathcal{N}_\alpha$.

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

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

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 $K = 1$. The result then follows from Theorem A.2. To apply this result, we first check its conditions.

Condition (a). Under Assumptions 5(a)–(c) and 6, Theorem A.1 implies that $\hat{\alpha}_n^K = \alpha^* + o_{p_n}(1)$.

Condition (b). This is imposed in Assumption 4.

Assumption 6(a) and Equation (A.1) imply that $(\alpha^*, \hat{\theta}_{f,n}, \hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1. In turn, this 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.

Condition (c). This follows from Assumption 5(c) and $(\hat{\alpha}_n^K, \hat{\theta}_{f,n}, \hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1.

Condition (d). Assumptions 5(d)–(f), 6(b), 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:

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

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),

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

If we denote the RHS random variable by Z, condition (d) follows.

Condition (e)–(f). 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. This and Assumptions 5(c)–(e) imply that the following derivation holds w.p.a.1:

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

where convergence is uniform in $\alpha \in \mathcal{N}_\alpha$, that is, condition (e) follows. In addition, if we denote the first term on the RHS by $H(\alpha)$, condition (f) follows.

Under these conditions, Theorem A.2 then implies that

$$\begin{aligned} n^{\min\{\delta,1/2\}}\big(\hat{\alpha}_n^K - \alpha^*\big) &= A_1 n^{\min\{\delta,1/2\}} \frac{\partial Q_n\big(\alpha^*, \theta_f^*, P^*\big)}{\partial \alpha} \\ &\quad + A_2 n^{\min\{\delta,1/2\}}\big(\hat{\theta}_{f,n} - \theta_f^*\big) + o_{p_n}(1), \end{aligned} \tag{A.2}$$

with

$$\begin{aligned} A_1 &\equiv -\left(\frac{\partial^2 Q_\infty\big(\alpha, \theta_f^*, P^*\big)}{\partial \alpha \partial \alpha'}\right)^{-1}, \\ A_2 &\equiv -\left(\frac{\partial^2 Q_\infty\big(\alpha, \theta_f^*, P^*\big)}{\partial \alpha \partial \alpha'}\right)^{-1} \frac{\partial^2 Q_\infty\big(\alpha^*, \theta_f^*, P^*\big)}{\partial \alpha \partial \theta_f'}. \end{aligned}$$

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

Initial step. For $K = 1$, the result holds by Assumption 6(b). Also, part 1 implies that Equation (A.2) holds for $K = 1$.

Inductive step. The inductive assumption is that Equations (A.1)–(A.2) hold for some $K \geq 1$. Our goal is then to show that Equations (A.1)–(A.2) hold with K replaced by $K + 1$. 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:

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

where $(\tilde{\alpha}_n, \tilde{\theta}_{f,n}, \tilde{P}_n)$ is 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 $P^* = \Psi_{(\alpha^*, \theta_f^*)}(P^*)$, and the final equality holds by Lemma 2.1(c) 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 $n^{\min\{\delta,1/2\}}(\hat{P}_n^K - P^*) = O_{p_n}(1)$, that is, Equation (A.1) holds with K replaced by $K + 1$. In turn, this and part 1 then imply that Equation (A.2) 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.3. To apply Theorem 3.1, we first need to verify Assumptions 5–6. We anticipate that $Q_\infty^{\mathrm{ML}}(\theta, P) = \sum_{(a,x)\in A\times X} J^*(a,x)\ln\Psi_\theta(P)(a|x)$.

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 $(\theta, P) \in \Theta \times \Theta_P$ and $(a,x) \in A \times X$ implies that $Q_n^{\mathrm{ML}}(\theta, P) - Q_\infty^{\mathrm{ML}}(\theta, P) = o_{p_n}(1)$. Furthermore, notice that

$$\begin{aligned}
&\sup_{(\theta,P)\in\Theta\times\Theta_P} \left|Q_n^{\mathrm{ML}}(\theta,P) - Q_\infty^{\mathrm{ML}}(\theta,P)\right| \\
&\quad= \sup_{(\theta,P)\in\Theta\times\Theta_P} \left|\sum_{(a,x)\in A\times X} \left(\hat{J}_n(a,x) - J^*(a,x)\right)\ln\Psi_\theta(P)(a|x)\right| \\
&\quad\le \sum_{(a,x)\in A\times X} \left|\hat{J}_n(a,x) - J^*(a,x)\right| \times \left|\ln\Big(\min_{(a,x)\in A\times X}\inf_{(\theta,P)\in\Theta\times\Theta_P} \Psi_\theta(P)(a|x)\Big)\right|.
\end{aligned}$$

Since $\Psi_\theta(P)(a|x) > 0$ for all $(\theta, P) \in \Theta \times \Theta_P$ and $(a,x) \in A \times X$, $\Psi_\theta(P)(a|x) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ, P) for all $(a,x) \in A \times X$, and $\Theta \times \Theta_P$ is compact, $\min_{(a,x)\in A\times X}\inf_{(\theta,P)\in\Theta\times\Theta_P} \Psi_\theta(P)(a|x) > 0$. From this and $\hat{J}_n - J^* = o_{p_n}(1)$, we conclude that $\sup_{(\theta,P)\in\Theta\times\Theta_P} |Q_n^{\mathrm{ML}}(\theta,P) - Q_\infty^{\mathrm{ML}}(\theta,P)| = o_{p_n}(1)$. Second, previous arguments imply that $Q_\infty^{\mathrm{ML}}(\theta,P) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ, P). Since $\Theta \times \Theta_P$ is compact, it then follows that $Q_\infty^{\mathrm{ML}}(\theta,P) : \Theta \times \Theta_P \to \mathbb{R}$ 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 follows from 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 $(a,x) \in A \times X$, and $\Psi_\theta(P)(a|x) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in (θ, P) for all $(a,x) \in A \times X$, $\ln\Psi_\theta(P)(a|x) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in (θ, P) for all $(a,x) \in A \times X$. From here, the result follows.

Condition (d). By direct computation,

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

with

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

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

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

$$\begin{aligned}\frac{\partial^2 Q_\infty^{\mathrm{ML}}(\alpha, \theta_f, P)}{\partial\alpha\partial\lambda'} = \sum_{(a,x)\in A\times X} J^*(a, x)\Bigg[&\frac{-1}{\big(\Psi_\theta(P)(a, x)\big)^2}\frac{\partial\Psi_\theta(P)(a|x)}{\partial\alpha}\frac{\partial\Psi_\theta(P)(a|x)}{\partial\lambda'} \\ &+ \frac{1}{\Psi_\theta(P)(a|x)}\frac{\partial^2\Psi_\theta(P)(a|x)}{\partial\alpha\partial\lambda'}\Bigg].\end{aligned} \tag{A.3}$$

This function is continuous and if we evaluate it at $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$, we obtain

$$\begin{aligned}&\frac{\partial^2 Q_\infty^{\mathrm{ML}}\big(\alpha^*, \theta_f^*, P^*\big)}{\partial\alpha\partial\lambda'} \\ &\quad= \sum_{(a,x)\in A\times X} J^*(a, x)\Bigg[\frac{-1}{\big(P_{\theta^*}(a|x)\big)^2}\frac{\partial P_{\theta^*}(a|x)}{\partial\alpha}\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda'} \\ &\qquad+ \frac{1}{P_{\theta^*}(a|x)}\frac{\partial^2\Psi_{\theta^*}\big(P^*\big)(a|x)}{\partial\alpha\partial\lambda'}\Bigg] \\ &\quad= -\sum_{(a,x)\in A\times X} J^*(a, x)\frac{\partial\ln P_{\theta^*}(a|x)}{\partial\alpha}\frac{\partial\ln P_{\theta^*}(a|x)}{\partial\lambda'} \\ &\qquad+ \sum_{x\in X} m^*(x)\sum_{a\in A}\frac{\partial^2\Psi_{\theta^*}\big(P^*\big)(a|x)}{\partial\alpha\partial\lambda'} \\ &\quad= -\sum_{(a,x)\in A\times X} J^*(a, x)\frac{\partial\ln P_{\theta^*}(a|x)}{\partial\alpha}\frac{\partial\ln P_{\theta^*}(a|x)}{\partial\lambda} = -\frac{\partial P'_{\theta^*}}{\partial\alpha}\Phi\frac{\partial P_{\theta^*}}{\partial\lambda'},\end{aligned} \tag{A.4}$$

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.3. To verify the result, it suffices to consider the last expression with $\lambda = \alpha$. Since Φ is a nonsingular matrix and $\partial P_{\theta^*}/\partial\alpha$ has full rank matrix, we conclude that the expression is square, symmetric, and negative definite, and, consequently, it must be nonsingular.

Condition (f). By Young's theorem and Equation (A.3) with $\lambda = P$ and $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$,

$$\begin{aligned}&\frac{\partial^2 Q_\infty^{\mathrm{ML}}\big(\alpha^*, \theta_f^*, P^*\big)}{\partial P\partial\alpha'} \\ &\quad= \sum_{(a,x)\in A\times X} J^*(a, x)\Bigg[\frac{-1}{\big(\Psi_{\theta^*}\big(P^*\big)(a, x)\big)^2}\frac{\partial\Psi_{\theta^*}\big(P^*\big)(a, x)}{\partial P}\frac{\partial\Psi_{\theta^*}\big(P^*\big)(a, x)}{\partial\alpha'}\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\Psi_\theta(P)(a,x)}\frac{\partial^2\Psi_{\theta^*}(P^*)(a,x)}{\partial P\partial\alpha'}\Bigg]\\
&=\sum_{(a,x)\in A\times X}J^*(a,x)\Bigg[\frac{-1}{\big(\Psi_{\theta^*}(P^*)(a,x)\big)^2}\frac{\partial\Psi_{\theta^*}(P_{\theta^*})(a,x)}{\partial P}\frac{\partial\Psi_{\theta^*}(P^*)(a,x)}{\partial\alpha'}\\
&\quad+\frac{1}{\Psi_\theta(P)(a,x)}\frac{\partial}{\partial\alpha'}\frac{\partial\Psi_{\theta^*}(P_{\theta^*})(a,x)}{\partial P}\Bigg]\\
&=\mathbf{0}_{|\tilde{A}\times X|\times d_\alpha},
\end{aligned}
$$

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

Part 2: Verify Assumption 6.

Assumption 6(b) holds as a corollary of Lemma A.2. To verify Assumption 6(a), consider the following argument. By direct computation,

$$
\begin{aligned}
\frac{\partial Q_n^{\mathrm{ML}}(\alpha^*,\theta_f^*,P^*)}{\partial\alpha}&=\sum_{(a,x)\in A\times X}\hat{J}_n(a,x)\frac{1}{\Psi_{\theta^*}(P^*)(a,x)}\frac{\partial\Psi_{\theta^*}(P^*)(a,x)}{\partial\alpha}\\
&=\frac{\partial\{\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in A\times X}\}'}{\partial\alpha}\hat{J}_n=\frac{\partial P'_{\theta^*}}{\partial\alpha}\Phi\Sigma\hat{J}_n,
\end{aligned}
\tag{A.5}
$$

where the second equality uses that $\partial\Psi_{\theta^*}(P^*)/\partial\alpha=\partial P_{\theta^*}/\partial\alpha$ and the last equality uses Lemma A.3. Also, by using an analogous argument applied to the population,

$$
\begin{aligned}
&\frac{\partial Q_\infty^{\mathrm{ML}}(\alpha^*,\theta_f^*,P^*)}{\partial\alpha}\\
&=\frac{\partial P'_{\theta^*}}{\partial\alpha}\Phi\Sigma J^*\\
&=\frac{\partial P'_{\theta^*}}{\partial\alpha}\Phi\big\{\big[\big[\mathbf{I}_{|\tilde{A}|\times|A|}-\{P^*(a|x)\}_{a\in\tilde{A}}\times\mathbf{1}_{1\times|A|}\big]/m(x)\big]\times\{J^*(a,x)\}_{a\in A}\big\}_{x\in X}\\
&=\frac{\partial P'_{\theta^*}}{\partial\alpha}\Phi\big\{\big[\mathbf{I}_{|\tilde{A}|\times|A|}-\{P^*(a|x)\}_{a\in\tilde{A}}\times\mathbf{1}_{1\times|A|}\big]\times\{P^*(a|x)\}_{a\in A}\big\}_{x\in X}\\
&=\frac{\partial P'_{\theta^*}}{\partial\alpha}\Phi\mathbf{0}_{|\tilde{A}|\times|X|}=\mathbf{0}_{|\tilde{A}|\times|X|},
\end{aligned}
\tag{A.6}
$$

where the third equality uses that $P^*(a|x)=J^*(a,x)/m^*(x)$ and the fourth equality uses that $\sum_{a\in A}P^*(a|x)=1$ for all $(a,x)\in A\times X$.

By combining Equations (A.5) and (A.6), we conclude that

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

From this and Lemma A.1, we conclude that the desired result holds with

$$\zeta \sim \begin{bmatrix} \dfrac{\partial P'_{\theta^*}}{\partial \alpha}\Phi\Sigma & \mathbf{0}_{|A\times X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times|A\times X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}} \end{bmatrix} \times \Delta N\big(B_{\Pi^*}\times 1[\delta\le 1/2], \big(\operatorname{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\big)\times 1[\delta\ge 1/2]\big). \tag{A.7}$$

This completes the verification of Assumptions 5–6 and so Theorem 3.1 applies. The specific formula for the asymptotic distribution relies on Equations (A.4) and (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 $Q_\infty^{\mathrm{MD}}(\theta,P) = -[P^* - \Psi_\theta(P)]'W^*[P^* - \Psi_\theta(P)]$.

Part 1: Verify the conditions in Assumption 5.

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

$$\begin{aligned} &\sup_{(\theta,P)\in\Theta\times\Theta_P}\big|Q_n^{\mathrm{MD}}(\theta,P) - Q_\infty^{\mathrm{MD}}(\theta,P)\big| \\ &= \sup_{(\theta,P)\in\Theta\times\Theta_P}\left|\begin{matrix} -\big(\hat P_n - P^*\big)'\hat W_n\big[\hat P_n - \Psi_\theta(P)\big] \\ -\big(P^* - \Psi_\theta(P)\big)'\big[\hat W_n - W^*\big]\big[\hat P_n - \Psi_\theta(P)\big] \\ -\big(P^* - \Psi_\theta(P)\big)'W^*\big(\hat P_n - P^*\big)\end{matrix}\right| \\ &\le \big\|\hat P_n - P^*\big\|\times\big(\big\|\hat W_n - W^*\big\| + 2\big\|W^*\big\|\big) + \big\|\hat W_n - W^*\big\|. \end{aligned}$$

Second, since $\Psi_\theta(P)(a|x):\Theta\times\Theta_P\to\mathbb{R}$ is continuous in (θ,P) for all (a,x), $Q_\infty^{\mathrm{MD}}(\theta,P):\Theta\times\Theta_P\to\mathbb{R}$ is continuous in (θ,P). In turn, since $\Theta\times\Theta_P$ is compact, $Q_\infty^{\mathrm{MD}}(\theta,P):\Theta\times\Theta_P\to\mathbb{R}$ 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). $Q_\infty^{\mathrm{MD}}(\alpha,\theta_f^*,P^*) = -[P^* - \Psi_{(\alpha,\theta_f^*)}(P^*)]'W^*[P^* - \Psi_{(\alpha,\theta_f^*)}(P^*)]$ is uniquely maximized at α^*. First, notice that $\Psi_{(\alpha^*,\theta_f^*)}(P^*) = P^*$ and so $Q_\infty^{\mathrm{MD}}(\alpha^*,\theta_f^*,P^*) = 0$. Second, consider any $\tilde\alpha\in\Theta_\alpha\backslash\alpha^*$. By the identification assumption, $\Psi_{(\tilde\alpha,\theta_f^*)}(P^*)\ne\Psi_{(\alpha^*,\theta_f^*)}(P^*) = P^*$. Since W^* is positive definite, $Q_\infty^{\mathrm{MD}}(\tilde\alpha,\theta_f^*,P^*) > 0$.

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 $(a,x)\in A\times X$.

Condition (d). By the same argument as in the verification of condition (c), $Q_\infty^{\mathrm{MD}}(\theta,P):\Theta\times\Theta_P\to\mathbb{R}$ is twice continuously differentiable in (θ,P). Since $\Psi_\theta(P)(a|x)$ is twice continuously differentiable in (θ,P) for all $(a,x)\in\tilde A\times X$, we conclude that $\partial\Psi_\theta(P)(a,x)/\partial\lambda$ and $\partial\Psi_\theta(P)(a,x)/\partial\alpha\partial\lambda'$ are continuous in (θ,P) for all $\lambda\in\{\theta,P\}$ and $(a,x)\in\tilde A\times X$. From this and the fact that $\Theta\times\Theta_P$ is compact,

$$\max_{(a,x)\in A\times X}\sup_{(\theta,P)\in\Theta\times\Theta_P}\|\partial\Psi_\theta(P)(a,x)/\partial\lambda\| < \infty \quad\text{and}$$

$$\max_{(a,x)\in A\times X}\sup_{(\theta,P)\in\Theta\times\Theta_P}\|\partial\Psi_\theta(P)(a,x)/\partial\alpha\partial\lambda'\| < \infty.$$

From this, $\hat P_n - P^* = o_{p_n}(1)$, and $\hat W_n - W^* = o_{p_n}(1)$, the desired result follows.

Condition (e). For any $\lambda \in \{\alpha, \theta_f, P\}$, direct computation shows that

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

This function is continuous and if we evaluate at $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$, we obtain

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

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 nonsingular.

Condition (f). By Young's theorem and Equation (A.8) with $\lambda = P$ and $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$,

$$\begin{aligned}\frac{\partial^2 Q_\infty^{\mathrm{MD}}(\alpha^*, \theta_f^*, P^*)}{\partial P \partial \alpha'} &= -2\frac{\partial \Psi_{\theta^*}(P^*)'}{\partial P} W^* \frac{\partial \Psi_{\theta^*}(P^*)}{\partial \alpha'} = -2\frac{\partial \Psi_{\theta^*}(P_{\theta^*})'}{\partial P} W^* \frac{\partial \Psi_{\theta^*}(P_{\theta^*})}{\partial \alpha'} \\ &= \mathbf{0}_{|\tilde{A}\times X|\times d_\alpha},\end{aligned}$$

where the last equality uses that the Jacobian matrix of Ψ_{θ^*} with respect to P is zero at $P_{\theta^*} = P^*$.

Part 2: Verify the conditions in Assumption 6.

Assumption 6(b) holds as a corollary of Lemma A.2. To verify Assumption 6(a), consider the following argument. By direct computation,

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

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

$$\begin{aligned}n^{\min\{\delta,1/2\}} \begin{bmatrix} \partial Q_n^{\mathrm{MD}}(\alpha^*, \theta_f^*, P^*)/\partial \alpha \\ \big(\hat{\theta}_{f,n} - \theta_f^*\big) \end{bmatrix} &= \begin{bmatrix} 2\dfrac{\partial P'_{\theta^*}}{\partial \alpha} W^* & \mathbf{0}_{|A\times X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times |A\times X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}} \end{bmatrix} n^{\min\{\delta,1/2\}} \begin{bmatrix} \big(\hat{P}_n - P^*\big) \\ \big(\hat{\theta}_{f,n} - \theta_f^*\big) \end{bmatrix} \\ &\quad + o_{p_n}(1).\end{aligned}$$

From this and Lemma A.2, we conclude that the desired result holds with

$$\begin{aligned}\zeta \sim &\begin{bmatrix} \dfrac{\partial P'_{\theta^*}}{\partial \alpha} W^* \Sigma & \mathbf{0}_{|A\times X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times |A\times X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}} \end{bmatrix} \\ &\times \Delta N\big(B_{\Pi^*} \times 1[\delta \le 1/2], \big(\mathrm{diag}\big(\Pi^*\big) - \Pi^*\Pi^{*\prime}\big) \times 1[\delta \ge 1/2]\big).\end{aligned} \tag{A.9}$$

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

A.3 *Proofs of lemmas*

PROOF OF LEMMA 2.1. This proof follows from Aguirregabiria and Mira (2002, Propositions 1–2). □

PROOF OF LEMMA 2.2. Parts (a)–(b) follow from Rust (1988, pp. 1015–6). Part (c) follows from combining Lemma 2.1 and Assumption 2. □

LEMMA A.1. *Suppose Assumptions* 3–7. *Then*

$$n^{\min\{\delta,1/2\}}\begin{pmatrix}\hat{J}_n - J^* \\ \hat{\theta}_{f,n} - \theta_f^*\end{pmatrix} \xrightarrow{d} \Delta \times N\big(B_{\Pi^*} \times 1[\delta \le 1/2], \big(\operatorname{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\big) \times 1[\delta \ge 1/2]\big), \tag{A.10}$$

with Δ *as in Equation* (4.2).

PROOF. Under Assumption 7, the triangular array CLT (e.g., Davidson (1994, p. 369)) implies that

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

If we combine this with Assumption 3,

$$n^{\min\{\delta,1/2\}}\big(\hat{\Pi}_n - \Pi^*\big) \xrightarrow{d} N\big(B_{\Pi^*} \times 1[\delta \le 1/2], \big(\operatorname{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\big) \times 1[\delta \ge 1/2]\big). \tag{A.11}$$

Also, notice that

$$n^{\min\{\delta,1/2\}}\begin{pmatrix}\hat{J}_n - J^* \\ \hat{\theta}_{f,n} - \theta_f^*\end{pmatrix} = n^{\min\{\delta,1/2\}}\big(F(\hat{\Pi}_n) - F\big(\Pi^*\big)\big),$$

where $F : \mathbb{R}^{|A\times X\times X|} \to \mathbb{R}^{|A\times X|+d_{\theta_f}}$ is defined as follows. For coordinates $j \le |A \times X|$ where j represents the corresponding coordinate $(a, x) \in A \times X$, $F_j(z) \equiv \sum_{\tilde{x}'\in X} z_{(a,x,\tilde{x}')}$, and for coordinates $j > |A \times X|$, $F_j(z) \equiv G_{1,j}(z)$. By definition of G_1, $\hat{\theta}_{f,n} = G_1(\hat{\Pi}_n)$, and by Assumption 8, $\theta_f^* = G_1(\Pi^*)$ and F is continuously differentiable at Π^*. By direct computation, $\Delta = \partial F(\Pi^*)/\partial \Pi'$. Then the result follows from the delta method and Equation (A.11). □

LEMMA A.2. *Suppose Assumptions* 3–7. *Then*

$$n^{\min\{\delta,1/2\}}\begin{pmatrix}\hat{P}_n - P^* \\ \hat{\theta}_{f,n} - \theta_f^*\end{pmatrix} \xrightarrow{d} \begin{bmatrix}\Sigma & \mathbf{0}_{|\tilde{A}\times X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times|\tilde{A}\times X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}}\end{bmatrix} \times \Delta \times N\big(B_{\Pi^*}1[\delta \le 1/2], \big(\operatorname{diag}(\Pi^*) - \Pi^*\Pi^{*\prime}\big)1[\delta \ge 1/2]\big), \tag{A.12}$$

with Δ *as in Equation* (4.2) *and* Σ *as in Equation* (4.3).

PROOF. Let $F : \mathbb{R}^{|A\times X|+d_{\theta_f}} \to \mathbb{R}^{|\tilde{A}\times X|+d_{\theta_f}}$ be defined as follows. For coordinates $j \leq |\tilde{A}\times X|$ with j representing coordinate $(a,x) \in \tilde{A}\times X$, $F_j(z) \equiv z_{(a,x)}/\sum_{a\in A} z_{(\tilde{a},x)}$, and for $j > |\tilde{A}\times X|$, $F_j(z) = z_j$. Notice that $F((\hat{J}_n', \hat{\theta}_{f,n}')') \equiv (\hat{P}_n', \hat{\theta}_{f,n}')'$ and $F((J^{*\prime}, \theta_f^{*\prime})') \equiv (P^{*\prime}, \theta_f^{*\prime})'$ by definition of F. If we verify F is continuously differentiable at $z = (J^{*\prime}, \theta_f^{*\prime})'$ and

$$\frac{\partial F\big((J^{*\prime}, \theta_f^{*\prime})'\big)}{\partial z'} = \begin{bmatrix} \Sigma & \mathbf{0}_{|\tilde{A}\times X|\times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f}\times|\tilde{A}\times X|} & \mathbf{I}_{d_{\theta_f}\times d_{\theta_f}} \end{bmatrix}, \tag{A.13}$$

then the result follows from the delta method and Lemma A.2. We do this next.

Consider $(j,\check{j}) \in \{1,\ldots,|\tilde{A}\times X|\}\times\{1,\ldots,|A\times X|\}$ representing $(a,x)\in\tilde{A}\times X$ and $(\check{a},\check{x})\in A\times X$. For any $j > |\tilde{A}\times X|$, F_j is continuously differentiable and $\partial F_j(z)/\partial z_{\check{j}} = 1[j=\check{j}]$. For any $j \leq |\tilde{A}\times X|$,

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

provided that $\sum_{\tilde{a}\in A} z_{(\tilde{a},x)} > 0$. Since $\sum_{\tilde{a}\in A} J^*(\tilde{a},x) > 0$ for all $x\in X$, F is continuously differentiable at $((J^{*\prime}, \theta_f^{*\prime})')$. By combining the formula for the derivatives from all coordinates, Equation (A.13) follows. □

LEMMA A.3. *For any $\lambda, \tilde{\lambda} \in \{\theta_f, \alpha\}$, the following algebraic results hold*:

$$\frac{\partial\{\ln P_{\theta^*}(a|x)\}'_{(a,x)\in A\times X}}{\partial\lambda} = \frac{\partial P'_{\theta^*}}{\partial\lambda}\Phi\Sigma,$$

$$\sum_{(a,x)\in A\times X} J^*(a,x)\frac{\partial \ln P_{\theta^*}(a|x)}{\partial\lambda}\frac{\partial \ln P_{\theta^*}(a|x)}{\partial\tilde{\lambda}'} = \frac{\partial P'_{\theta^*}}{\partial\lambda}\Phi\frac{\partial P_{\theta^*}}{\partial\tilde{\lambda}'},$$

with Φ and Σ as in Equation (4.3).

PROOF. Before deriving the results, consider some preliminary observations. For any $\lambda\in\{\alpha,\theta_f,P\}$, $\sum_{a\in A} P_{\theta^*}(a|x) = 1$ and so $\partial P_{\theta^*}(|A||x)/\partial\lambda = -\sum_{a\in\tilde{A}}\partial P_{\theta^*}(a|x)/\partial\lambda$. Also, for any $\lambda\in\{\alpha,\theta_f,P\}$ and $(a,x)\in A\times X$, $P^*(a|x) = P_{\theta^*}(a|x)$ 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:

$$\frac{\partial P'_{\theta^*}}{\partial\lambda}\Phi\Sigma = \frac{\partial\{P_{\theta^*}(a|x)\}'_{(a,x)\in\tilde{A}\times X}}{\partial\lambda}\times\{\operatorname{diag}\{\Phi_x\Sigma_x\}_{x\in X}\}$$

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

where the last equality uses the preliminary observations.

For the second result, consider the following derivation:

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

where the last equality uses the preliminary observations. □

A.4 *Review of results on extremum estimators*

The purpose of this section is to state well-known results regarding the consistency and asymptotic normality of extremum estimators under certain regularity conditions. These results are referenced in our formal arguments. Relative to the standard versions in the literature (e.g., McFadden and Newey (1994)), our results allow for: (a) a rate of convergence that may differ from $\sqrt{n}$ and (b) a sequence of DGPs that may change with sample size. Both of these features are important for our theoretical results. We omit the proofs for reasons of brevity but theses 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* $\{p_n\}_{n\geq 1}$.

(b) $Q(\theta)$ *is upper semicontinuous, that is, for any* $\{\theta_n\}_{n\geq 1}$ *with* $\theta_n \to \tilde{\theta}$, $\limsup Q(\theta_n) \leq Q(\tilde{\theta})$.

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

Then $\hat{\theta}_n = \arg\max_{\theta\in\Theta} Q_n(\theta)$ *satisfies* $\hat{\theta}_n = \theta^* + o_{p_n}(1)$.

THEOREM A.2. *Consider an estimator* $\hat{\theta}_n$ *of a parameter* θ^* *s.t.* $\hat{\theta}_n = \arg\max_{\theta\in\Theta} Q_n(\theta)$. *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* $\{p_n\}_{n\geq 1}$,

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

(f) $H(\theta^*)$ *is nonsingular.*

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* $\{p_n\}_{n\geq 1}$.

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. [67, 68, 69, 71, 73, 77, 80, 82, 86, 89, 98]

Aguirregabiria, V. and P. Mira (2010), "Dynamic discrete choice structural models: A survey." *Journal of Econometrics*, 156, 38–67. [68]

Arcidiacono, P. and P. B. Ellickson (2011), "Practical methods for estimation of dynamic discrete choice models." *Annual Review of Economics*, 3, 363–394. [68]

Arcidiacono, P. and R. A. Miller (2011), "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity." *Econometrica*, 79, 1823–1867. [75]

Blackwell, D. (1965), "Discounted dynamic programming." *The Annals of Mathematical Statistics*, 36, 226–235. [72]

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. [70, 76]

Bugni, F. A., and T. Ura (2019), "Supplement to 'Inference in dynamic discrete choice problems under local misspecification'." *Quantitative Economics Supplemental Material*, 10, https://doi.org/10.3982/QE917. [85]

Chernozhukov, V., J. C. Escanciano, H. Ichimura, and W. K. Newey (2016), "Locally robust semiparametric estimation." Working paper. [70]

Davidson, J. (1994), *Stochastic Limit Theory*. Oxford University Press. [98]

Gourieroux, C. and A. Monfort (1995), *Statistics and Econometric Models*, Vol. 2. Cambridge University Press. [93, 96]

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. [67, 68, 72, 82, 89]

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. [80]

Kasahara, H. and K. Shimotsu (2008), "Pseudo-likelihood estimation and bootstrap inference for structural discrete Markov decision models." *Journal of Econometrics*, 146, 92–106. [86]

Kitamura, Y., T. Otsu, and K. Evdokimov (2013), "Robustness, infinitesimal neighborhoods, and moment restrictions." *Econometrica*, 81, 1185–1201. [70]

Magnac, T. and D. Thesmar (2002), "Identifying dynamic discrete decision processes." *Econometrica*, 70, 801–816. [71, 73]

McFadden, D. and W. K. Newey (1994), "Large sample estimation and hypothesis testing." In *Handbook of Econometrics* (R. F. Engle and D. L. McFadden, eds.), Handbook of Econometrics, Vol. 4, 2111–2245, Elsevier. [84, 100]

Newey, W. K. (1985a), "Generalized method of moments specification testing." *Journal of Econometrics*, 29, 229–256. [70, 76]

Newey, W. K. (1985b), "Maximum likelihood specification testing and conditional moment tests." *Econometrica*, 5, 1047–1070. [70, 76]

Norets, A. and S. Takahashi (2013), "On the surjectivity of the mapping between utilities and choice probabilities." *Quantitative Economics*, 4, 149–155. [70]

Pesendorfer, M. and P. Schmidt-Dengler (2008), "Asymptotic least squares estimators for dynamic games." *Review of Economic Studies*, 75, 901–928. [67, 68, 80, 82, 89]

Rothenberg, T. J. (1971), "Identification in parametric models." *Econometrica*, 39, 577–591. [80]

Royden, H. L. (1988), *Real Analysis*. Prentice-Hall. [90]

Rust, J. (1987), "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher." *Econometrica*, 55, 999–1033. [68, 82, 85]

Rust, J. (1988), "Maximum likelihood estimation of discrete control processes." *SIAM J. Control and Optimization*, 26, 1006–1024. [68, 72, 98]

Schorfheide, F. (2005), "VAR forecasting under misspecification." *Journal of Econometrics*, 128, 99–136. [70]

Tauchen, G. (1985), "Diagnosing testing and evaluation of maximum likelihood models." *Journal of Econometrics*, 30, 415–443. [70, 76]

White, H. (1982), "Maximum likelihood estimation of misspecified models." *Econometrica*, 50, 681–700. [70]

White, H. (1996), *Estimation, Inference and Specification Analysis*. Econometric Society Monographs, Vol. 22. Cambridge University Press. [70, 93]

Co-editor Christopher Taber handled this manuscript.

Manuscript received 10 July, 2017; final version accepted 28 April, 2018; available online 9 May, 2018.