APPROXIMATING CHOICE DATA BY DISCRETE CHOICE MODELS

By

Haoge Chang, Yusuke Narita, Kota Saito

May 2024

COWLES FOUNDATION DISCUSSION PAPER NO. 2392

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Approximating Choice Data by Discrete Choice Models

Haoge Chang, Yusuke Narita, and Kota Saito*

May 28, 2024

Abstract

We obtain a necessary and sufficient condition under which random-coefficient discrete choice models, such as mixed-logit models, are rich enough to approximate any nonparametric random utility models arbitrarily well across choice sets. The condition turns out to be the affine-independence of the set of characteristic vectors. When the condition fails, resulting in some random utility models that cannot be closely approximated, we identify preferences and substitution patterns that are challenging to approximate accurately. We also propose algorithms to quantify the magnitude of approximation errors.

Keywords: Discrete choice, stochastic choice, mixed logit, random coefficients, finite mixture.

*Chang: Microsoft Research: haogechang@microsoft.com; Narita: Yale University: yusuke.narita@yale.edu; Saito: Caltech: saito@caltech.edu A part of this paper was first presented at the University of Tokyo on July 29, 2017. This paper subsumes parts of "Axiomatizations of the Mixed Logit Model" by Saito (The paper is available at `http://www.hss.caltech.edu/content/axiomatizations-mixed-logit-model`). We would like to thank Hiroki Saruya, Richard Gong, and Haruki Kono for their help as RAs. We appreciate the valuable discussions with Victor Aguirregabiria, Brendan Beare, Steven Berry, Vivek Bhattacharya, Federico A. Bugni, Giovanni Compiani, Colin Cameron, Alfred Galichon, Doignon Jean-Paul, Ariel Pakes, John Rust, Phil Haile, Hidehiko Ichimura, Jay Lu, Rosa Matzkin, Sunjog Misra, Whitney K. Newey, Robert Porter, Matt Shum, Satoru Takahashi, Takuya Ura, and Yi Xin. Satoru Takahashi and Jay Lu read some versions of the manuscript and offered helpful comments. We appreciate the insightful comments made by Victor Aguirregabiria at the ASSA meetings in January 2022. Saito acknowledges the financial support of the NSF through grants SES-1919263.

1 Introduction

Random-coefficient discrete choice models are workhorse models in many empirical applications. These models are commonly used to approximate various preferences and capture rich substitution patterns. However, the exact degree of flexibility and the limitations of the random-coefficient models have not been fully understood.[1] In this paper, we obtain a necessary and sufficient condition under which random coefficient models can approximate the choice behavior generated by any nonparametric random utility model arbitrarily well. Our results suggest that some widely-used empirical models may not be flexible enough to capture important economic quantities, such as substitution patterns. In such instances, we introduce methods to pinpoint preferences that are particularly difficult to approximate with precision. Additionally, we propose algorithms designed to quantify the magnitude of these approximation errors. Our results help researchers assess if the models can capture choice behaviors relevant to the research problem at hand.

We consider the standard setup (e.g. in Train (2009)). Let J be the set of all alternatives. For each alternative j, $x_j \in \mathbf{R}^K$ is the vector of characteristics of alternative j, where K is the number of explanatory variables. With an *additive random utility model (ARUM)*, the choice probability of an alternative j in a choice set $D \subset J$ is given by $\rho(D, j) = \mu(\{\varepsilon | \beta \cdot x_j + \eta_j + \varepsilon_j > \beta \cdot x_l + \eta_l + \varepsilon_l \ \forall l \in D \setminus \{j\}\})$, where β is a deterministic vector capturing an agent's preferences, η is a vector of *fixed effects*, which captures unobserved characteristics of alternatives, and ε is a random utility shock that follows a probability measure μ.[2] The class of the ARUMs is general and includes the probit, logit, and nested-logit models as special cases. The *random-coefficient* version of the ARUM is defined as follows. The choice probability is given by

$$\rho(D, j) = \int \mu(\{\varepsilon | \beta \cdot x_j + \eta_j + \varepsilon_j > \beta \cdot x_l + \eta_l + \varepsilon_l, \forall l \in D \setminus \{j\}\}) dm(\beta), \tag{1}$$

where m is a probability measure over β. In the standard interpretation, the distribution

[1] An important early work in this direction is McFadden and Train (2000). See the section on related literature for details.

[2] In this paper, we assume that μ is absolutely continuous with respect to the Lebesgue measure and the support is convex.

m captures the heterogeneity of preferences among the population of agents.[3] When μ is an iid type-I extreme-value distribution, then ρ reduces to a *mixed-logit model*, which is one of the widely used random-coefficient models.

Given the popularity of the random-coefficient ARAMs, it is important to understand its flexibility and limitations. For this purpose, we obtain a necessary and sufficient condition under which the random-coefficient ARUMs are rich enough to approximate any choice probabilities generated by nonparametric random utility models arbitrarily well across choice sets. For the approximation target, we choose the random utility models, which are defined as probability measures over strict preference rankings over alternatives. We choose this class of models because it is the most general and agnostic class of models assuming individuals' rational behavior. We study approximation *across choice sets* because many questions of interest (such as substitution patterns) rely on analyzing behaviors across choice sets. Moreover, choice data across choice sets are widespread in the fields of marketing, empirical industrial organization, environmental economics, and political science.

In our main theorem (Theorem 1), we state that the necessary and sufficient condition is the *affine-independence* of the set $\{x_j \in \mathbf{R}^K | j \in J\}$.[4] To interpret the condition, consider a typical setup, in which a researcher first fixes one probability measure μ over the shock ε; then the researcher estimates the distribution m over coefficients β and the fixed effects η after observing a dataset. If the affine-independence condition is satisfied, then the researcher should be able to approximate any given dataset by using some random-coefficient ARUMs arbitrarily well across choice sets. On the other hand, if the affine-independence condition is violated, there exists a dataset generated by a random utility model that cannot be approximated arbitrarily well by any random-coefficient ARUM, no matter which random-coefficient distribution m as well as fixed effects η we use. The affine-independence condition is easy to test: the condition is generically equivalent to a further simpler condition: $K \geq |J| - 1$, where $|J|$ is the number of alternatives and K is

[3]Notice that the roles of β and η are different. The probability measure m is only on β but not on η.

[4]A set $Y \equiv \{y_1, \ldots, y_n\}$ is *affinely independent* if for any $y_i \in Y$, there exists no real number $\{\mu_j\}_{j \neq i}$ such that $y_i = \sum_{j \neq i} \mu_j y_j$ and $\sum_{j \neq i} \mu_j = 1$. A set $Y \equiv \{y_1, \ldots, y_n\}$ is affinely independent if and only if $\{y_2 - y_1, \ldots, y_n - y_1\}$ is linearly independent, where y_1 can be replaced by any y_i.

the number of characteristics observed for each alternative.

In many empirical papers, researchers use the mixed-logit models that are linear in the original characteristics and do not contain additional terms such as polynomials. We call such models *linear mixed-logit models.* In these papers, we often observe a deviation from the condition $K \geq |J| - 1$, resulting in the violation of the affine-independence condition. This means that the linear mixed-logit models may not be rich enough to approximate the true substitution pattern arbitrarily well across subsets of J, no matter how one chooses the distribution m and the fixed effects η.

When the affine-independence condition is violated, our theorem shows that there exists random utility model that cannot be approximated arbitrarily well; moreover, our result implies that this happens because there are strict preference rankings that cannot be approximated arbitrarily well. Given this result, we introduce a tractable method to identify such strict preference rankings: this method can be efficiently implemented through straightforward linear programmings. To further quantify the flexibility of the random-coefficient ARUMs under researchers' consideration, we calculate the approximation error to degenerate stochastic choice funcition corresponding to the strict preference rankings that are challenging to approximate precisely. We also calculate the maximal substitution patterns allowed in the class of random-coefficient ARUMs. To calculate these quantities, we introduce two algorithms. One algorithm is a variant of the greedy algorithm proposed in Barron et al. (2008). The other algorithm is the EM (Expectation-Maximization) algorithm drawn from Dempster et al. (1977). The outputs of the algorithms can help researchers identify choice behaviors that are failed to be captured, and researchers can assess whether these behaviors are empirically relevant.

We apply our theorem and the two algorithms to a dataset of fishing-site choices (Thomson and Crooke, 1991). In the dataset, there are four alternatives (i.e., $|J| = 4$) and two characteristic variables (i.e., $K = 2$): price p_j and a quality measure q_j of each fishing site. We find that the affine-independence condition is violated with the original characteristics (i.e., $x_j = (p_j, q_j)$) because $K = 2 \ngeq 3 = |J| - 1$. By using our methods, we find that half of the preferences cannot be approximated arbitrarily well. With our two algorithms, we measure the approximation errors to these preferences by the class of linear mixed-logit models. Regardless of the algorithm used, we find that the approximation errors are large. Specifically, the choice probabilities predicted by the closest linear mixed-logit model sometimes deviate from the true ones by over 70 percentage points. Moreover, we identify substitution patterns that cannot be captured well.[5] We find that the class of linear mixed-logit models limits the largest substitution pattern from one alternative to another to be at most 12 percentage points, no matter how the parameters of the linear mixed-logit models are chosen.

The structure of the paper is as follows. In Section 2, we introduce the models underpinning our analysis. Section 3 presents the key theorems of the paper. A sketch of the proof is detailed in Section 4. In Section 5, we elaborate on the methodologies employed for measuring approximation errors. Finally, Section 6 applies our theoretical framework to an empirical context, utilizing a real-world dataset for illustration.

Related Literature

The work most closely related to our paper are Dagsvik (1994) and especially McFadden and Train (2000), who show that any given (nonparametric) continuous random utility model can be approximated arbitrarily well by a mixed-logit model.[6] Nevertheless, there are important differences to note. In particular, our result holds for a much more general class of random-coefficient ARUMs, including but not confined to mixed-logit models. Second, our result is not only sufficient but also necessary. This is crucial given our purpose of clarifying the exact extent of flexibility and limitations of the random-coefficient ARUMs. Moreover, through our condition, our results provide a tight bound on how many parameters we need for an arbitrarily good approximation. Third, the setup of McFadden and Train (2000) and our setup differ in that McFadden and Train (2000) focus on the case where the set of characteristics is continuous. Hence, neither result implies the other.

A recent paper by Lu and Saito (2021) also studies the extent to which the approximation of a continuous random utility model (i.e., pure characteristics model) is possible by using

[5] In this context, we define the substitution pattern as the largest increase of the choice probabilities of one alternative when another alternative becomes unavailable.

[6] Our result is consistent with their result: heuristically speaking, the result by McFadden and Train (2000) corresponds to the case when the researcher use arbitrarily higher order polynomials (i.e., $K \to \infty$), which satisfies our condition that $K \geq |J| - 1$.

mixed-logit models.

Another paper closely related to ours is Norets and Takahashi (2013). They study whether ARUMs can represent any stochastic choice (i.e., market shares). However, their analysis focuses on a fixed choice set while our paper studies approximation across various choice sets. Other related papers also focus on a fixed choice set. In particular, Berry (1994) provides an earlier and classical inversion result useful for representing any stochastic choice on a given choice set. Athey and Imbens (2007) investigate how a rich specification of the unobserved components is needed to represent any stochastic choice function in a fixed choice set.

Our analysis shares some of its spirit with the growing literature that identifies and estimates flexible discrete choice models under minimal assumptions. See, for example, Berry and Haile (2014), Compiani (2022), and Tebaldi et al. (2023). Despite the similarity in the spirit, our problem is different from the standard econometric problems. We are not concerned with statistical estimation or identification problems, i.e. recovering model parameters in either sampled or population setting. In contrast, our primary goal is to explore the limitations of common modeling strategies within the discrete choice literature: our work focuses on a specification or approximation question rather than identification, estimation, or inference.

In the decision theory literature, there are two interpretations of stochastic choice. The first interpretation is based on the observation that even a single agent may make stochastic choices, as observed in recent experiments (see Agranov and Ortoleva (2017)). The second interpretation suggests that stochasticity arises from unobserved heterogeneity among a population of agents, as typically assumed in the empirical literature. Although our paper aligns with this latter perspective, we know of no research that directly relates to our papers.

Historically, logit models and random utility models have been analyzed extensively ever since Luce (1959) and Block and Marschak (1960). Recent studies, such as those by Apesteguia and Ballester (2018) and Frick et al. (2019), highlight the distinctions in choice behavior between random utility models and logit models.

Furthermore, a few recent studies examine the substitution property in discrete choice analysis. Horan and Adam (2023) discusses the substitution patterns captured by random utility models. Allen and Rehbeck (2020) analyze aggregate complementarity in latent utility models used in discrete choice.

2 Model

2.1 Setup

Set of alternatives: The set of all alternatives is denoted by J. J is assumed to be finite.

Choice sets: Let $\mathcal{D} \subset 2^J \setminus \{\emptyset\}$ be the set of choice sets. Notice that $\mathcal{D}$ can be a proper subset of $2^J \setminus \emptyset$. Unless otherwise noted, throughout the paper we assume that $\mathcal{D}$ contains all binary and ternary choice sets: $\{j, l\} \in \mathcal{D}$ and $\{j, l, r\} \in \mathcal{D}$ for any $j, l, r \in J$. In a part of the paper (i.e., Section 3.1), however, we drop this assumption and assume that $\mathcal{D} = \{J\}$ when we consider the case in which the researcher's purpose is fitting a model to the observed choice probabilities from the single choice set.

The set $\mathcal{D}$ may contain both observed choice sets as well as hypothetical choice sets the researcher is interested in. For example, even when the researcher observes consumers' choices only over $\{\text{train}, \text{bus}, \text{car}\}$, he may also be interested in choices over $\{\text{train}, \text{bus}\}$, $\{\text{train}, \text{car}\}$, and $\{\text{bus}, \text{car}\}$ to learn the consumers' substitution pattern.

Explanatory variables: An alternative $j \in J$ is described by a real vector $x_j \in \mathbf{R}^K$ of explanatory variables, where K is the number of the explanatory variables. For instance, if an alternative j is a consumption good, the alternative may be described by its price p_j and its quality index q_j; in that case $x_j = (p_j, q_j)$. Moreover, the researcher can include functions of original characteristics in x_j. Empirical applications often include higher order polynomials as well as splines or wavelets (Chen, 2007). For example, with the original characteristics (p_j, q_j) of alternative j, the researcher may include higher order polynomials such as $p_j^2, q_j^2, p_j q_j$ in the characteristic vector x_j and can make the number K of characteristic vectors larger. If the researcher includes all terms, then $x_j = (p_j, q_j, p_j^2, q_j^2, p_j q_j)$ and $K = 5$.

Stochastic choice function: A function $\rho : \mathcal{D} \times J \to [0, 1]$ is called a *stochastic choice function* if $\sum_{j \in D} \rho(D, j) = 1$ and $\rho(D, j) = 0$ for any $j \notin D$. The set of stochastic choice functions is denoted by $\mathcal{P}$. For each $(D, j) \in \mathcal{D} \times J$, the number $\rho(D, j)$ is interpreted as the probability that an alternative j is chosen from a choice set D. In the context of discrete choice analysis, for example, $\rho(D, j)$ can be interpreted as the market share of product j in a market in which the set of available products is D. In such cases, we interpret the stochastic choice function ρ as aggregate choice probabilities across individuals.

Rankings: Let Π be the set of bijections between J and $\{1, \ldots, |J|\}$, where $|J|$ is the number of elements of J. For any element $\pi \in \Pi$, if $\pi(j) = i$, then we interpret alternative j to be the $|J| + 1 - i$-th best element of J with respect to π. If $\pi(j) > \pi(l)$, then j is better than l with respect to π. An element π of Π is called a *ranking* over J. A ranking describes an agent's strict preference relation.[7]

For all $(D, j) \in \mathcal{D} \times J$ such that $j \in D$, if $\pi(j) > \pi(l)$ for all $l \in D \setminus \{j\}$, then we often write $\pi(j) \geq \pi(D)$. There are $|J|!$ elements in Π.

2.2 Models

We denote the set of probability measures over Π by $\Delta(\Pi)$. Since Π is finite, $\Delta(\Pi) = \left\{(\nu_1, \ldots, \nu_{|\Pi|}) \in \mathbf{R}_+^{|\Pi|} \,\middle|\, \sum_{i=1}^{|\Pi|} \nu_i = 1\right\}$, where $\mathbf{R}_+$ is the set of nonnegative real numbers.

We now introduce the definition of random utility models:

Definition 1. A stochastic choice function ρ is called a *random utility model* if there exists a probability measure $\nu \in \Delta(\Pi)$ such that for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then

$$\rho(D, j) = \nu(\{\pi \in \Pi | \pi(j) \geq \pi(D)\}).$$

The set of random utility models is denoted by $\mathcal{P}_r$.[8]

Notice that when $\mathcal{D} = \{J\}$, the restriction of random utility is vacuous: any stochastic choice function is a random utility model (i.e., $\mathcal{P}_r = \mathcal{P}$).[9]

[7] Alternatively, a ranking can be defined as a binary relation that is complete, transitive and irreflexive. The relation is often called a linear order.

[8] While the function above is often called a random ranking function, a random utility model is often defined differently by using the existence of a probability measure μ over utilities such that for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then $\rho(D, j) = \mu(\{u \in \mathbf{R}^J | u(x) \geq u(D)\})$. Block and Marschak (1960)'s Theorem 3.1 proves that the two definitions are equivalent.

[9] To see this, observe that $\mathcal{P}_r \subset \mathcal{P}$ by definition. We show the converse. For any $j \in J$, let $\pi_j \in \Pi$ such that $\pi_j(j) > \pi_j(l)$ for all $l \in D \setminus \{j\}$. Then $\rho^{\pi_j}(l) = 1_j(l)$ for any $l \in J$, where $1_j(l) = 1$ if $l = j$ and $1_j(l) = 0$ if

In certain scenarios, researchers might want to exclude rankings deemed unreasonable and restrict the set of rankings. We consider such a case in Section A in the appendix.

In both theoretical and empirical literature, modeling assumptions are imposed to approximate the random utility models. Two important classes are defined as follows. First we introduce one definition.

Definition 2. A Borel probability measure μ on the Borel σ-algebra of $\mathbf{R}^{|J|}$ is said to be a *standard probability measure* if μ is absolutely continuous with respect to the Lebesgue measure and the support is convex.[10] Let $\mathcal{M}$ be the set of all standard probability measures.

In the following, we denote the inner product of two vectors x and y by $x \cdot y$.

Definition 3. Let $\eta \in \mathbf{R}^{|J|}$ be a real vector. A stochastic choice function ρ is called a *random-coefficient additive-random utility model (random-coefficient ARUM) with fixed effects* η if there exist a standard probability measure μ and a Borel probability measure m such that for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then

$$\rho(D, j) = \int \mu(\{\varepsilon | \beta \cdot x_j + \eta_j + \varepsilon_j > \beta \cdot x_l + \eta_l + \varepsilon_l \forall l \in D \setminus \{j\}\}) dm(\beta).$$

The random vector $(\varepsilon_j)_{j \in J} \in \mathbf{R}^{|J|}$ follows the distribution μ. When the support of m has only one point, the stochastic choice function ρ is called an *additive-random utility model (ARUM) with fixed effects* η: for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then

$$\rho(D, j) = \mu(\{\varepsilon | \beta \cdot x_j + \eta_j + \varepsilon_j > \beta \cdot x_l + \eta_l + \varepsilon_l \text{ for all } l \in D \setminus \{j\}\}).$$

The set of random-coefficient ARUMs is denoted by $\mathcal{P}_{ra}(\eta|\mu)$ and the set of ARUMs by $\mathcal{P}_a(\eta|\mu)$. When the context makes clear which standard probability measure μ we consider, we do not specify the standard probability measure μ.

$l \neq j$. (For the definition of ρ^π, see definition (4) in Section 3.3.) For any $\rho \in \mathcal{P}$, define $\rho' = \sum_{j \in J} \rho(j) \rho^{\pi_j}$. Then, $\rho' \in \mathcal{P}_r$ and $\rho'(j) = \rho(j)$ for any $j \in J$, as desired. Hence, $\mathcal{P} \subset \mathcal{P}_r$. In general, we have $\mathcal{P}_r \subsetneq \mathcal{P}$ and the random utility models have some testable implication. For example, when $\mathcal{D} = 2^J \setminus \emptyset$, random utility models are characterized by the non-negativity of the Block-Marschak polynomials.

[10] Remember that the support supp.μ is defined as $\{\varepsilon \in \mathbf{R}^{|J|} | \mu(N_\varepsilon) > 0$ for any open neighborhood N_ε of $\varepsilon\}$.

The term $\beta \cdot x_j$ is the *systematic part* of the utility of alternative j captured by the observed characteristics x_j. The vector β captures preferences of an agent and the distribution m over coefficients β describes the heterogeneity of preferences among the population of the agents. The constant η_j is called a *fixed effect* that captures the utility of alternative j from the unobserved characteristics; ε_j is the shock to the utility of alternative j.

Almost all probability measures used in practice are standard. For a mixed-logit model, μ is an iid extreme-value type-I distribution; for a probit model, μ is the multivariate standard normal distribution. Note that in most empirical applications of these models, the mixing distribution m is a parametric distribution like a multivariate normal distribution. In our case, the mixing distributions of the random coefficients do not come from a particular parametric family.

For the exposition later, we define the *mixed-logit models* formally as follows:

Definition 4. Let $\eta \in \mathbf{R}^{|J|}$ be a real vector. A stochastic choice function ρ is called a *mixed-logit model with fixed effects* η if there exists a Borel probability measure m such that for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then

$$\rho(D, j) = \int \frac{\exp(\beta \cdot x_j + \eta_j)}{\sum_{l \in D} \exp(\beta \cdot x_l + \eta_l)} dm(\beta). \tag{2}$$

The set of mixed-logit models with fixed effects η is denoted by $\mathcal{P}_{ml}(\eta)$. When m puts the unit mass on a particular β in (2), then ρ is called a *logit model.* The set of logit models with fixed effects η is denoted by $\mathcal{P}_l(\eta)$.

We give a few remarks on the models. First, we consider the models with fixed effects given their popularity in empirical applications. Fixed effects are used frequently to capture the unobserved characteristics of alternatives (Berry et al., 1995). As we are interested in approximating choice data for the same population, we consider the case where fixed effects do not change across choice sets. While there is a common presumption that utilizing fixed effects enables us to depict any behavior generated by the random utility model, we demonstrate that this may not hold when there are multiple choice sets. Each of our results is stated with and without fixed effects.

Second, in the models above, following McFadden and Train (2000) as well as many other papers in discrete choice analysis, we write the systematic part of utility by $\beta \cdot x_j$. Although we have a linear structure in the explanatory variables x_j, it is important to remember that x_j can be a function of original characteristics such as higher order polynomials of original characteristics as used in the proof of McFadden and Train (2000). Since any continuous function can be approximated by polynomials, our models are general enough to allow for flexible systematic parts of utility functions.

Finally, we review essential mathematical concepts. A *polytope* is a convex hull of finitely many points. The closure of a set C is denoted by cl.C with respect to the standard finite dimensional Euclidean topology. The *affine hull* of a set C is the smallest affine set that contains C, and it is denoted by aff.C. The convex hull of a set C is denoted by co.C. The *relative interior* of a convex set C is the interior of C in the relative topology with respect to aff.C. The relative interior of C is denoted by rint.C.

3 Main Result

To state the main result of the paper, we review a basic concept in geometry: A set $\{x_j \in \mathbf{R}^K | j \in J\}$ is *affinely independent* if no x_j can be written as an affine combination of the other elements $\{x_l\}_{l \neq j}$. Formally, for any $j \in J$, there exists no real numbers $\{\alpha_l\}_{l \in J \setminus \{j\}}$ such that $x_j = \sum_{l \in J \setminus \{j\}} \alpha_l x_l$ and $\sum_{l \in J \setminus \{j\}} \alpha_l = 1$.[11]

Theorem 1. *(i) Let μ be any standard probability measure. If the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is affinely independent, then any random utility model can be approximated arbitrarily well by a random-coefficient ARUM; moreover, the approximation can be done without fixed effects (i.e., $\eta = 0$). That is,*

$$\forall \rho \in \mathcal{P}_r \ \exists \{\rho_n\}_{n=1}^{\infty} \subset \mathcal{P}_{ra}(0|\mu), \forall D \in \mathcal{D} \text{ and } \forall j \in D, [\lim_{n \to \infty} \rho_n(D, j) = \rho(D, j)]$$

(ii) If the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is not affinely independent, then there exists a random utility model that cannot be approximated arbitrarily well by any random-coefficient ARUM with any sequence of fixed effects and with any standard probability measure

[11]It is easy to see that a set $\{x_j \in \mathbf{R}^K | j \in J\}$ is affinely independent if and only if $\{x_l - x_j\}_{l \in J \setminus \{j\}}$ is linearly independent for any j.

μ. That is,

$$\exists \rho \in \mathcal{P}_r \, \forall \mu \in \mathcal{M}, \rho \notin cl. \bigcup_{\eta \in \mathbf{R}^{|J|}} \mathcal{P}_{ra}(\eta|\mu).$$

We provide a sketch of the proof in Section 4 and the formal proof in the appendix.

To interpret the main theorem, consider a typical practice where the shock distribution μ over ε is fixed a priori by a researcher; then he or she estimates the distribution m over coefficients β as well as the fixed effects $\eta \in \mathbf{R}^K$ after seeing a dataset ρ generated from a random utility model.

Statement (i) shows that if the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is affinely independent, then the researcher should be able to approximate the given dataset ρ arbitrarily well across choice sets $D \in \mathcal{D}$ by choosing an appropriate distribution m over coefficients β. This direction builds upon the classical result of McFadden and Train (2000).

Statement (ii) shows that if the affine-independence condition fails, then there exists a random utility model that cannot be approximated arbitrarily well by any random-coefficient ARUM, no matter how the researcher changes the distribution m over coefficients β as well as the fixed effects η, given the arbitrary chosen standard probability measure μ over ε. This direction is novel, which has not been studied in McFadden and Train (2000). Moreover, note that the result allows a general class of utility-shock distributions, not confining to the type-I extreme-value distribution or a particular choice of mixing distributions. For example, the approximation is impossible using any mixed-logit model nor any random-coefficient probit-model. In Propositions 2 and 3 in Section 3.3, we will give examples of the random utility models that cannot be approximated arbitrarily well.

The affine-independence condition can be simplified further to a generically equivalent condition. To see this, remember the following basic facts: (i) if $|J| > K + 1$, then $\{x_j \in \mathbf{R}^K | j \in J\}$ is not affinely independent; (ii) if $|J| \leq K+1$, then the set is generically affinely independent.[12] Given these observations, Theorem 1 implies the following corollary:[13]

[12]This is a standard concept of genericity in the literature of discrete geometry. Even if the set is not affinely independent, as long as $|J| \leq K + 1$, for any $\varepsilon > 0$, there exists an affinely independent set X', obtained from X by moving each point by a distance of at most ε (see Section 3 of Matousek (2013)).

[13]One caveat of the result is that even though the generic condition holds, the original condition of the affine independence may not hold when explanatory variables include zeroes and ones. In that case, one should check

Corollary 1. *Let K be the number of explanatory variables and $|J|$ be the number of alternatives.*

(i) If $K \geq |J| - 1$, then the statements in Theorem 1 (i) hold generically.

(ii) If $K < |J| - 1$, then the statements in Theorem 1 (ii) hold.

We now mention a few remarks on the results. First, to understand the results correctly, it is crucial to understand the space that we consider for the approximation. In Theorem 1, we study approximation across coordinates (D, x) such that $x \in D \in \mathcal{D}$, which has very high dimensionality. On the other hand, in Proposition 1 of the following section, we consider a fixed choice set $\mathcal{D} = \{J\}$. Thus the underlying dimension is much smaller.

Second, as mentioned earlier, to increase the number K of explanatory variables, researchers may include additional terms such as higher order polynomials (McFadden and Train (2000)) as well as splines or wavelets (Chen, 2007). In their proof, McFadden and Train (2000) use higher order polynomials of arbitrarily high degrees to approximate continuous random utility model. In particular, in their construction, x_j is a vector of monomials of any degree of original characteristics $(y_{j1}, \cdots, y_{jn})$:

$$x_j = (\underbrace{y_{j1}, \cdots, y_{jn}}_{\text{1st order terms}}, \underbrace{y_{j1}^2, \cdots, y_{jn}^2}_{\text{2nd order terms}}, y_{j1}^3, \cdots, y_{jn}^3, ...) \in \mathbf{R}^K,$$

where $K \to \infty$. Their result is thus consistent with the sufficiency part of our result: we proved that $K \geq |J| - 1$ is sufficient in our setup.

Finally, in the theorem, following Mcfadden and Train (2000), we consider all possible random utility models (i.e., probability distributions over all rankings) as the prediction target mainly for simplicity. As mentioned after Definition 1, in some cases, the researchers may want to restrict the set of rankings by excluding those deemed unreasonable. In Section A of the appendix, we provide a necessary and sufficient condition for the approximation of the restricted random utility model.

---

the affine-independence of $\{x_j \in \mathbf{R}^K | j \in J\}$, rather than the generic condition.

3.1 Additional Result for Single Choice Set Case

In the following, we provide a supplemental result for the case in which $\mathcal{D} = \{J\}$. Such a case corresponds to a situation in which the researcher is interested only in the observed choice probabilities (i.e., market shares) on a single set J (but not on its subsets).

Proposition 1. *Assume that $\mathcal{D} = \{J\}$.*

(i) Let μ be any standard probability measure. If the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is convex-independent (i.e., if $x_j \notin co.\{x_l | l \in J \setminus \{j\}\}$ for any $j \in J$), then any random utility model can be approximated arbitrarily well by a random-coefficient ARUM; moreover, the approximation can be done without fixed effects (i.e., $\eta = 0$). That is,

$$\forall \mu \in \mathcal{M},\ \forall \rho \in \mathcal{P}_r,\ \exists \{\rho_n\}_{n=1}^{\infty} \subset \mathcal{P}_{ra}(0|\mu),\ \forall j \in J\ \lim_{n \to \infty} \rho_n(J, j) = \rho(J, j).$$

(ii) (a) If the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is not convex-independent, then there exists a random utility model that cannot be approximated arbitrarily well by any random-coefficient ARUM with any standard probability measure μ and without fixed effects (i.e., $\eta = 0$). That is, $\exists \rho \in \mathcal{P}_r,\ \forall \mu \in \mathcal{M},\ \rho \notin cl.\mathcal{P}_{ra}(0|\mu)$.

(b) However, if fixed effects are used, any random utility model can be approximated arbitrarily well by an ARUM with any standard probability measure μ.

Note that the convex-independence condition is weaker than the affine-independence condition. This makes sense because the convex-independence condition guarantees the approximation only on the single choice set (i.e., $\{J\}$), while the affine-independence condition guarantees the approximation across all subsets $D \in \mathcal{D}$ of J (including J itself).

The implications of Theorem 1 and Proposition 1 are similar. One important difference arises when the conditions (i.e., the affine-independence condition in Theorem 1 and the convex-independence condition in Proposition 1) are violated. In both cases, there exists a random utility model that cannot be approximated arbitrarily well *without* using fixed effects. However, as stated in Proposition 1 (ii)(b), if fixed effects are used, any random utility model can be approximated arbitrarily well on one particular choice set J. (This result directly follows from Norets and Takahashi (2013).) This is in contrast to Theorem

1 (ii), which claims that there exists a random utility model that cannot be approximated arbitrarily well even using fixed effects across choice sets $\mathcal{D}$.[14]

Unlike the affine-independence, the convex-independence does not restrict the number of elements in a convex-independent set.[15] So there exists no counterpart of Corollary 1.

3.2 Implications of Theorems and Propositions

In this section, we mention the implications of the theorem and the proposition to the empirical literature. Many empirical papers use the mixed logit models that are linear in original characteristics and do not contain additional terms such as polynomials. We call such models *linear mixed-logit models*. In the papers, the convex-independence condition is usually satisfied. That is, it is often the case that any alternative x_j lies outside the convex hull co.$(\{x_l | l \in J \setminus \{j\})$ of the other alternatives. In fact, we will see this is the case in a real dataset in Section 6.

On the other hand, the condition that $K \geq |J| - 1$ is frequently violated in various contexts, which in turn results in the breach of the affine-independence condition. Remember that $|J|$ is the number of alternatives and K is the number of characteristics. There are many choice situations in which $|J|$ is very large such as choices of groceries, hospitals, cars, schools, or restaurants etc. In such a dataset, the condition is likely to be violated. This means that the class of linear mixed-logit models is rich enough to describe the choice data from a single choice set J; however, the class of the models may not be rich enough to approximate the true substitution pattern across choice sets, no matter how one chooses parameters and fixed effects.[16]

3.3 Preferences that Cannot be Approximated Well

When the affine-independence condition is violated, there exist random utility models that cannot be approximated arbitrarily well. Our results in Section 4 show hat this happens because the ARUMs cannot approximate some rankings arbitralily well. In the following,

[14]This difference originates from the fact that we require approximation on all $D \in \mathcal{D}$ in Theorem 1, while in Proposition 1, we require approximation only on J.

[15]For example, in three-dimensional space (x, y, z), consider a circumference of radius one whose origin is $(0, 0, 1)$ on a hyperplane of $z = 1$. The number of points on the circumference is a continuum. However, the set of points on the circumference is convex-independent.

[16]By substitution patterns, in general, we mean how choice probabilities change in different choice sets. In Section 6.2.1, we provide a more specific definition of substitution patterns.

we propose a method to identify such rankings that are difficult to approximate precisely. The following definition is crucial:

Definition 5. A ranking $\pi \in \Pi$ is *representable in choice sets* $\mathcal{D}$ if there exists a real vector β such that, for all $D \in \mathcal{D}$ and $j \in D$,

$$\pi(j) > \pi(l) \text{ for all } l \in D \setminus \{j\} \text{ if and only if } \beta \cdot x_j > \beta \cdot x_l \text{ for all } l \in D \setminus \{j\}. \quad (3)$$

If π is not representable in $\mathcal{D}$, we say that π is *unrepresentable in* $\mathcal{D}$.

The proposition below demonstrates that the failure of the affine-independence condition leads to the existence of unrepresentable rankings, which are exactly the rankings challenging to approximate accurately. To determine the representability of a specific ranking π, one can employ linear programming techniques.[17] In Section 6, utilizing a real dataset, we identify such unrepresentable rankings, as detailed in Table 1 of that section. The identification of these rankings enables researchers to evaluate which substitution patterns are challenging to capture within their models.

Notice that the requirement (3) depends both on the specification of the characteristic vectors $\{x_j\}_{j \in J}$ as well as the set $\mathcal{D}$ of choice sets. The requirement (3) becomes less restrictive as the characteristic vector x_j becomes longer because it becomes easier to find the desired β with additional characteristic variables; the requirement (3) becomes more restrictive as the set $\mathcal{D}$ of choice sets becomes richer, simply because the number of inequalities to be satisfied becomes larger. As mentioned earlier, we usually consider the general choice sets $\mathcal{D}$, while in some places, we assume a simpler case in which $\mathcal{D} = \{J\}$. In the rest of this section, we consider the general $\mathcal{D}$. When there is no risk of confusion, we will simply say that a ranking $\pi \in \Pi$ is representable without specifying choice sets $\mathcal{D}$.

To show the proposition, for each ranking $\pi \in \Pi$, define

$$\rho^{\pi}(D, j) = \begin{cases} 1 & \text{if } \pi(j) > \pi(l) \text{ for all } l \in D \setminus j; \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

[17] The ranking $\pi(1) > \pi(2) > ... > \pi(J)$ is representable if and only if the linear programming problem: $\max_{\beta \in \mathbf{R}^K, t \in \mathbf{R}} t$ subject to $\beta \cdot (x_j - x_{j+1}) \geq t$ for each $j = 1, ..., J-1$ has the optimal value ∞. If the ranking is unrepresentable, the problem has the optimal value 0.

The function ρ^π gives probability one to the best alternative x in a choice set D according to the strict ranking π.

Proposition 2. *The set $\{x_j \in \mathbf{R}^K | j \in J\}$ is not affinely independent if and only if there exists a ranking $\pi \in \Pi$ that is not representable in $\mathcal{D}$. For any unrepresentable ranking π and any standard probability measure μ, there exists a neighborhood U of ρ^π such that any random utility model that belongs to U cannot be approximated arbitrarily well by any random-coefficient ARUM without fixed effects.*

Remember that so far we have assumed no fixed effects (i.e., $\eta = 0$). In the following, we consider the case with fixed effects. As we explain in the next section, by using fixed effects, we can approximate any ρ^π. However, when the affine-independence condition fails, there still exist some random utility models that cannot be approximated well. The next proposition provides such random utility models:

Definition 6. For any ranking $\pi \in \Pi$, define $\pi^- \in \Pi$ such that $\pi(j) > \pi(l)$ if and only if $\pi^-(l) > \pi^-(j)$ for any $j, l \in J$. The ranking π^- is called the *reverse* ranking of π.

Note that if a ranking π is representable, then π^- is also representable. The next proposition shows that when the affine-independence condition fails, approximating a mixture of ρ^π and ρ^{π^-} is impossible even using fixed effects when π is not representable.

Proposition 3. *Suppose that $\{x_j \in \mathbf{R}^K | j \in J\}$ is not affinely independent. For any unrepresentable ranking $\pi \in \Pi$, any standard probability measure μ, and any $\alpha \in (0, 1)$, there exists a neighborhood U of $\alpha\rho^\pi + (1-\alpha)\rho^{\pi^-}$ such that any random utility model that belongs to U cannot be approximated arbitrarily well by any random-coefficient ARUM with any sequence of fixed effects and the probability measure μ.*

4 Sketch of Proof

In this section, we provide a proof sketch. Readers who are not interested in the proofs may skip this section and go directly to the empirical sections (Sections 5 and 6).

We prove Theorem 1 and Proposition 1 by using the five lemmas below. As byproducts, we obtain Propositions 2 and 3.[18]

Lemma 1 states a general condition for approximating random utility models with random coefficient ARUMs. Lemmas 2 and 3 translate the condition to a condition on the dimension of characteristics, which is easy to check in practice. Lemma 4 provides a class of random utility models that is hard to approximate even with the help of fixed effects. Lemma 5 states an important geometric insight that appears in the proof of Lemma 4.

We first consider models without fixed effects (i.e., $\eta = 0$). The following fact is elementary but fundamental:

Observation: *The set $\mathcal{P}_r$ of random utility models is a polytope, that is, $\mathcal{P}_r = co.\{\rho^\pi | \pi \in \Pi\}$.*

The observation holds because for any random utility model $\rho \in \mathcal{P}_r$, we have $\rho = \sum_{\pi \in \Pi} \nu(\pi)\rho^\pi$, where ν is the probability measure on Π rationalizing the random utility model. The hexagons in Figure 2 below illustrate the polytope.[19]

Lemma 1. *$\mathcal{P}_r \subset cl.\mathcal{P}_{ra}(0|\mu)$ if and only if $\rho^\pi \in cl.\mathcal{P}_a(0|\mu)$ for any $\pi \in \Pi$.*

Lemma 1 gives a necessary and sufficient condition under which any random utility models can be approximated arbitrarily well by random coefficient ARUMs (*i.e.*, $\mathcal{P}_{ra}(0|\mu)$) without fixed effects (i.e., $\eta = 0$). The condition is that $\rho^\pi \in \mathrm{cl}.\mathcal{P}_a(0|\mu)$ for any $\pi \in \Pi$, which means that ρ^π can be approximated by a sequence of ARUMs without fixed effects (i.e., a sequence of elements of $\mathcal{P}_a(0|\mu)$).

The next lemma makes it easier for us to check the conditions of Lemma 1.

Lemma 2. *For any ranking $\pi \in \Pi$, the following statements hold:*

1. *If π is representable in $\mathcal{D}$, then for any $\mu \in \mathcal{M}$, $\rho^\pi \in cl.\mathcal{P}_a(0|\mu)$.*

[18]Proposition 2 follows from Corollary 2-(ii) and Lemma 3, and Proposition 3 follows from Lemma 3 and Lemma 4.

[19]Although the geometric intuition is useful, it is important to notice that the figure oversimplifies the reality since the number (i.e., $|J|!$) of vertices and the dimension of a random utility model can be very large. To see why the dimension of a random utility model can be very large, remember that it assigns a number for each pair of $(D, j) \in \mathcal{D} \times J$. As mentioned, we calculate the dimension later in Proposition 5 in Section A.2 of the appendix.

2. *If π is not representable in $\mathcal{D}$, then there exists no standard probability measure μ such that $\{\rho^{\pi}, \rho^{\pi^-}\} \in cl.\mathcal{P}_{ra}(0|\eta)$.*

Lemmas 1 and 2 imply the following:

Corollary 2.

(i) Let μ be any standard probability measure. If any ranking $\pi \in \Pi$ is representable in $\mathcal{D}$, then any random utility model can be approximated arbitrarily well by a random-coefficient ARUM. Moreover, the approximation can be done without fixed effects (i.e., $\eta = 0$.) That is, $\mathcal{P}_r \subset cl.\mathcal{P}_{ra}(0|\mu)$.

(ii) If some ranking $\pi \in \Pi$ is not representable in $\mathcal{D}$, then there exists a random utility model that cannot be approximated arbitrarily well by any random-coefficient ARUM without fixed effects. That is, $\mathcal{P}_r \not\subset cl.\mathcal{P}_{ra}(0|\mu)$.

Corollary 2 offers a testable condition for determining if the random-coefficient ARUMs, without fixed effects, can adequately approximate any random utility model.[20]

As mentioned earlier, checking the representability of a particular ranking is easy. However, checking the representability of *all* rankings may be computationally demanding. This is because the number of rankings equals $|J|!$ and can be large. To overcome this problem, we obtain a simpler necessary and sufficient condition for any ranking $\pi \in \Pi$ to be representable:

Lemma 3.

1. *Any ranking is representable in $\mathcal{D}$ if and only if the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is affinely independent.*
2. *Any ranking is representable in $\{J\}$ if and only if the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is convex-independent.*

To understand Lemma 3 (1) geometrically, see Figure 1. In the figure, we assume that there are two original characteristic variables, say (p_j, q_j) for each alternative $j \in J$. In Figure 1 (a) and (b), we consider the models with the original characteristics (i.e.,

[20]In Section A of the appendix, we provide a generalization of Corollary 2 for the case in which a researcher wishes to omit certain rankings from their analysis and restrict the set of possible rankings.

$K = 2$ and $x_j = (p_j, q_j)$ for each $j \in J$). In Figure 1 (a), the set $\{x_1, x_2, x_3\}$ is affinely independent. Thus, by Lemma 3 (1) (the "if" part), any ranking is representable. For example, the ranking $\pi(1) > \pi(2) > \pi(3)$ is representable by $\beta \in \mathbf{R}^2$, which defines the parallel hyperplanes (indifference curves) in Figure 1 (a). On the other hand, in Figure 1 (b), the set $\{x_1, x_2, x_3, x_4\}$ is not affinely independent. The ranking $\pi(1) > \pi(4) > \pi(3) > \pi(2)$ is not representable. As the figure shows, no matter how one chooses $\beta \in \mathbf{R}^2$ and draws parallel hyperplanes as indifference curves, it does not hold that $\beta \cdot x_1 > \beta \cdot x_4 > \beta \cdot x_3 > \beta \cdot x_2$. The existence of such an unrepresentable ranking is implied by the "only if" part of Lemma 3 (1).[21]

If we use ellipses as indifference curves, however, we can represent the ranking $\pi(1) > \pi(4) > \pi(3) > \pi(2)$ as in Figure 1 (c).[22] The existence of such curves is again implied by the "if" part of Lemma 3 (1) since ellipses can be defined with the quadratic polynomials $\beta \cdot x_j$ with $x_j = (p_j, q_j, p_j^2, q_j^2, p_j q_j)$. Moreover the generic condition with quadratic polynomials is satisfied (i.e., $K = 5 \geq 3 = |J| - 1$) in this example.[23]

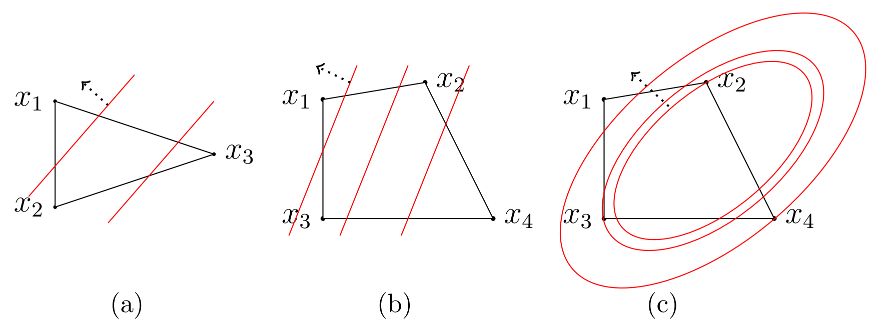

Figure 1: Illustration of the affine-independence condition.

Lemma 3 (2) is more straightforward. To see this, notice that when $\mathcal{D} = \{J\}$, any $\pi \in \Pi$ is representable in $\{J\}$ if and only if, for any $j \in J$, there exists β such that $\beta \cdot x_j > \beta \cdot x_l$ for all $l \in J \setminus j$, which means that J is convex-independent. By using Lemmas 1, 2, and 3, we obtain statement (i) of Theorem 1 and Proposition 1.

[21] The slope of the "indifference" line must be steeper than the slope of the line segment (x_4, x_2) (because $\pi(4) > \pi(2)$) and less steep than the slope of the line segment (x_4, x_1) (because $\pi(1) > \pi(4)$), which together imply that $\beta \cdot x_4 > \beta \cdot x_3$.

[22] As the radius of the ellipses becomes larger, the ranking becomes better.

[23] In fact, we verified that the affine-independence condition is satisfied with quadratic polynomials.

Remember that so far we have assumed no fixed effects (i.e., $\eta = 0$). In the following, we analyze the extent to which random utility models can be approximated arbitrarily well by using fixed effects. In particular, we show that if the affine independence condition fails then there exists a random utility model that cannot be approximated arbitrarily well even with using fixed effects.

First, we will see the usefulness of the fixed effects. It is easy to observe that when $\mathcal{D} = \{J\}$, any stochastic choice can be approximated arbitrarily well by using fixed effects.[24] Even for general $\mathcal{D}$, the following holds:

Observation: *For any ranking π, there exists an ARUM with fixed effects that can approximate ρ^π (i.e., vertices of polytope) arbitrarily well.*[25]

However, this is not enough to approximate any random utility model arbitrarily well across choice sets. As an illustration, consider two fixed effects, η_1 and η_2, and see Figure 2 below. Remember that in the heuristic figure, the hexagon represents $\mathcal{P}_r = \text{co.}\{\rho^\pi | \pi \in \Pi\}$. The two convex sets in the hexagon correspond to $\mathcal{P}_{ra}(\eta_1|\mu)$ and $\mathcal{P}_{ra}(\eta_2|\mu)$ shaded pink and blue, respectively.[26] Notice that all vertices in the figure can be approximated arbitrarily well by elements of $\mathcal{P}_{ra}(\eta_1|\mu)$ or $\mathcal{P}_{ra}(\eta_2|\mu)$. However, some areas of the hexagon are not covered by either $\mathcal{P}_{ra}(\eta_1|\mu)$ or $\mathcal{P}_{ra}(\eta_2|\mu)$.

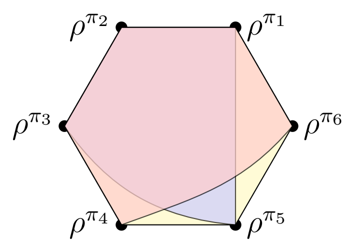

Figure 2: Illustration of $\mathcal{P}_{ra}(\eta_1|\mu)$ and $\mathcal{P}_{ra}(\eta_2|\mu)$

In reality, the problem is more complicated since we need to consider the union of all possible values of fixed effects, and thus the union of the continuum of convex sets

[24] This is intuitive since we can choose $|J|$ parameters (i.e., $(\eta_j)_{j\in J}$) to fit $|J|$ data points (i.e., $(\rho(J,j))_{j\in J}$).

[25] To see this, fix π and choose $\eta \in \mathbf{R}^J$ such that $\eta_j > \eta_l$ if and only if $\pi(j) > \pi(l)$. Then, it can be shown that an ARUM ρ_n defined by $\rho_n(D,j) = \mu(\{\varepsilon | n\eta_j + \varepsilon_j \geq n\eta_l + \varepsilon_l \text{ for all } l \in D \setminus \{j\}\})$ converges to $\rho^\pi(D,j)$ as $n \to \infty$.

[26] To see this remember that an element of $\mathcal{P}_{ra}(\eta|\mu)$ belongs to $\mathcal{P}_r$ and the set $\mathcal{P}_{ra}(\eta|\mu)$ is convex.

$\mathcal{P}_{ra}(\eta|\mu)$ across all values of $\eta \in \mathbf{R}^{|J|}$.[27] Moreover, we need to consider all possible standard probability measure $\mu \in \mathcal{M}$. Nevertheless, Lemma 4 provides a clear answer and states that if there exists a unrepresentable ranking, then there exists a class of random utility models that cannot be approximated arbitrarily well, no matter which fixed effects and probability distribution we use.

Lemma 4. *Let μ be a standard probability measure. For any $\alpha \in (0,1)$ and any ranking π that is not representable, there exists a neighborhood U of $\alpha\rho^{\pi} + (1-\alpha)\rho^{\pi^-}$ such that any random utility model that belongs to U cannot be approximated arbitrarily well by any random-coefficient ARUM with any fixed effects.*

To prove the lemma, we need to prove the following two statements: (a) any strict convex combination between ρ^{π} and ρ^{π^-} cannot be approximated arbitrarily well by a *degenerate* ARUM with fixed effects; and (b) moreover it cannot be approximated arbitrarily well by a *nondegenerate* random-coefficient ARUM even with any fixed effects. We prove statement (a) in the appendix. To show statement (b), we introduce the following concept:

Definition 7. The two rankings π and π' are *adjacent* if there exists $t \in \mathbf{R}^{|\mathcal{D}|\times|J|}$ and $a \in \mathbf{R}$ such that (i) $\rho^{\pi} \cdot t = \rho^{\pi'} \cdot t = a$, and (ii) for any $\hat{\pi} \in \Pi$, if $\pi \neq \hat{\pi} \neq \pi'$, then $\rho^{\hat{\pi}} \cdot t > a$.[28]

For example, in Figure 2, ρ^{π_1} and ρ^{π_6} as well as ρ^{π_i} and $\rho^{\pi_{i+1}}$ for each $i \leq 5$ are adjacent and no other pairs are adjacent. Since π and π^- are reversed with each other, ρ^{π} and ρ^{π^-} seem very different. It turns out, however, that they are adjacent:[29]

Lemma 5. *For any ranking $\pi \in \Pi$, ρ^{π} and ρ^{π^-} are adjacent.*

The characterization of adjacency of vertices for the case $\mathcal{D} = 2^J \setminus \emptyset$ appears in Doignon and Saito (2023). Lemma 5 holds even for the case in which $\mathcal{D} \neq 2^J \setminus \emptyset$ as long as $\mathcal{D}$

[27]More formally, the difficulty arises because the set of all ARUMs with probability measure μ and with any fixed effects (i.e., $\bigcup_{\eta \in \mathbf{R}^J} \mathcal{P}_{ra}(\eta|\mu)$) may not be convex, although given η, each set $\mathcal{P}_{ra}(\eta|\mu)$ is convex. This is because mixtures can be taken only over β but not over η. Thus approximating vertices is not enough for approximation over the polytope of random utility models.

[28]$t \in \mathbf{R}^{|\mathcal{D}|\times|J|}$ is a vector that gives a real number for each pair of $(D,j) \in \mathcal{D} \times J$. For any $\rho \in \mathcal{P}$, $\rho \cdot t = \sum_{(D,j)\in D\times J} \rho(D,j)t(D,j)$.

[29]Our discussions with Jean-Paul Doignon and Haruki Kono were very helpful for obtaining this result.

contains all binary and trinary sets.[30] The lemma allows us to complete the proof of Lemma 4 as follows. If π is not representable, then π^- is also not representable. Although fixed effects are powerful enough to approximate each vertex ρ^π, we will prove that it is not powerful enough to approximate both ρ^π and ρ^{π^-} by using the same fixed effects, intuitively because ρ^π and ρ^{π^-} are reversed. Thus, no strict convex combination of ρ^π and ρ^{π^-} can be approximated arbitrarily well by the random-coefficient ARUMs with standard probability measure μ, no matter which fixed effects we use. Notice that this conclusion does *not* follow if ρ^π and ρ^{π^-} are not adjacent since a strict convex combination of ρ^π and ρ^{π^-} may be represented in a different way. This proves statement (b) and thus, Lemma 4. Lemmas 1, 2, 3, and 4 prove statement (ii) of Theorem 1, as the proof in the appendix formalizes.

5 Measuring Approximation Errors

Propositions 2 and 3 in Section 3.3 show that the approximation errors to some random utility models may not be negligible when the affine-independence condition fails. In this section, we provide a way to quantify the approximation errors. We first define the distance function as follows: For any $\hat{\rho}, \rho \in \mathcal{P}_r$, define

$$d(\hat{\rho}, \rho) \equiv \sqrt{\frac{\sum_{D\in\mathcal{D}} \sum_{j\in D} \left(\rho(D,j) - \hat{\rho}(D,j)\right)^2}{|\mathcal{D}|}}.$$

In our analysis, $\hat{\rho}$ is a given random utility model; ρ is a random-coefficient ARUM by which we approximate $\hat{\rho}$. We divide the norm by $\sqrt{|\mathcal{D}|}$ to make the distance independent from the number of choice sets.[31] Notice that the maximal distance is 2. For example, $d(\rho^\pi, \rho^{\pi^-}) = 2$ for any ranking π.

Given an approximation target $\hat{\rho} \in \mathcal{P}_r$ and a standard probability measure μ, when

[30] We are grateful to Haruki Kono for pointing out this fact.

[31] $d(\hat{\rho}, \rho)$ can be written as $\|\rho - \hat{\rho}\|/\sqrt{|\mathcal{D}|}$, where $\|\cdot\|$ is the Euclidean norm (i.e., $l2$ norm). One can consider an alternative distance function based on l_1 norm as follows: $d_1(\hat{\rho}, \rho) \equiv (\sum_{(j,D)\in J\times\mathcal{D}} |\hat{\rho}(D,j) - \rho(D,j)|)/|\mathcal{D}|$. Since $\sqrt{\sum_{(j,D)\in J\times\mathcal{D}} (\hat{\rho}(D,j) - \rho(D,j))^2} \leq \sum_{(j,D)\in J\times\mathcal{D}} |\hat{\rho}(D,j) - \rho(D,j)|$, we can show that $d(\hat{\rho}, \rho)/\sqrt{|\mathcal{D}|} \leq d_1(\hat{\rho}, \rho)$ for any ρ and $\hat{\rho}$. So our approximation error divided by $\sqrt{|\mathcal{D}|}$ will provide a lower bound of an alternative approximation error measured by d_1. We use our distance function d rather than d_1 because of the convexity of d is useful for constructing an algorithm.

researchers use random-coefficient ARUMs with fixed effects η, the approximation error is defined as:

$$\inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}, \rho). \tag{5}$$

We call (5) the *approximation error to* $\hat{\rho}$ by random-coefficient ARUMs with fixed effects η. Proposition 2 shows that the approximation error to ρ^{π} is nonzero if π is not representable. In Section 6 below, we find that the approximation error can be large for some unrepresentable rankings in a real dataset.

Given $\hat{\rho}$, we propose two algorithms to solve (5) and compute the approximation errors. The first is the standard EM (Expectation-Maximization) algorithm (Dempster et al., 1977) to estimate the best possible finite-mixture logit model. It is known, however, that the EM algorithm may not converge to the global optimum. To alleviate this concern, we propose a second greedy algorithm inspired by Barron et al. (2008). This algorithm solves a sequence of optimization problems to converge to the global optimal solution. The structure of the random-coefficient RUMs is important for the proof. We provide explanations of these algorithms in Section A of the online appendix.

6 Application to Data

In this section, we measure approximation errors with and without fixed effects by using a dataset on fishing-site choices from Thomson and Crooke (1991).[32] The dataset has been used by Herriges and Kling (1999) and Cameron and Trivedi (2005) (p.464).

In the dataset, 1182 individuals choose among 4 fishing modes, namely, $J = \{\text{beach}, \text{boat}, \text{charter}, \text{pier}\}$, which denote fishing from the beach, a private boat, a charter boat or a pier, respectively. Each alternative $j \in J$ is described by a vector of two characteristics (p_j, q_j). The first characteristic p_j is the fishing mode j's price. The other characteristic q_j is the *catch rate*, defined as a per-hour-fished basis for major species by fishing mode j.[33]

Our empirical analysis concentrates primarily on mixed-logit models. This focus is

[32] The dataset is taken directly from the R package 'mlogit' by Croissant (2020).

[33] In the original study, the values of p_j and q_j depend on each individual. For our analysis, we aggregate them by taking the average over individuals.

motivated by the widespread application and acceptance of these models in contemporary research. In particular, we consider two specifications of mixed-logit models. The first one is the linear mixed-logit models, the mixed logit models that are linear in the original characteristics (i.e., $x_j = (p_j, q_j)$ for each $j \in J$). The second one is the mixed-logit model defined with quadratic polynomials (i.e., $x_j = (p_j, q_j, p_j^2, q_j^2, p_j q_j)$ for each $j \in J$). We call the model *quadratic mixed-logit model.*

We assume that $\mathcal{D} = 2^J \setminus \emptyset$. Propositions 4 and 5 in Section A.2 of the online appendix imply that in order to obtain the best approximating random-coefficient model to the observed choice probabilities, it is sufficient to consider finite mixture models with at most $(\dim \mathcal{P}_r) + 1 = 1 + \sum_{D \in \mathcal{D}} (|D| - 1) = 18$ mixtures. See Section A.2 of the online appendix for the details.

6.1 Application of Theorem 1

The dataset contains four alternatives (i.e., $|J| = 4$). If we use original characteristics as explanatory variables (i.e., $x_j = (p_j, q_j)$), then $K = 2$ and the condition in Corollary 1 is violated (i.e., $K = 2 \not\geq 3 = |J| - 1$); thus the set $\{(p_j, q_j) \in \mathbf{R}^2 | j \in J\}$ is not affinely independent. Thus, by Theorem 1, the linear mixed-logit models with fixed effects is not flexible enough to approximate some random utility models. This observation motivates us to compute approximation errors of the linear logit models without fixed effects (in Section 6.2) and the errors with fixed effects (in Section 6.3).

On the other hand, with quadratic polynomials, the generic condition for representability in Corollary 1 is satisfied, since $K = 5 \geq 3 = |J| - 1$.[34] In fact, we verified that $\{(p_j, q_j, p_j^2, q_j^2, p_j q_j) \in \mathbf{R}^5 | j \in J\}$ is affinely independent. Thus, by Theorem 1, the quadratic mixed-logit models are flexible enough to approximate any random utility model. This theoretical implication is also numerically verified below.

6.2 Approximation Errors without Fixed Effects

In this section, we detail approximation errors without fixed effects. We say that a ranking π is *linearly representable* if π is representable in $\mathcal{D}$ with $x_j = (p_j, q_j)$; π is *linearly unrepresentable* if π is not linearly representable.

[34] Given the increasing number of characteristic variables, a natural concern is the overfitting problem. We investigate this concern in Section B.1 of the online appendix.

Since the affine-independence condition fails in the dataset with $x_j = (p_j, q_j)$, it follows from Proposition 2 that there exists a ranking π that is linearly unrepresentable. Thus the corresponding deterministic choice functions ρ^π cannot be approximated by any linear mixed-logit model without fixed effects. Since there are four alternatives, there are twenty four rankings. Among them, we find that twelve rankings are not representable, and thus cannot be approximated arbitrarily well by linear mixed-logit models, as shown in Table 1.

Table 1: Approximation errors to ρ^π

Ranking π	Linear mixed-logit		Quadratic mixed-logit	
	Greedy	EM	Greedy	EM
(1)	(2)	(3)	(4)	(5)
Linearly Unrepresentable Rankings				
$\pi(1) > \pi(2) > \pi(3) > \pi(4)$	0.723	0.753	0.000	0.000
$\pi(1) > \pi(2) > \pi(4) > \pi(3)$	0.670	0.700	0.000	0.000
$\pi(1) > \pi(3) > \pi(2) > \pi(4)$	0.425	0.381	0.000	0.000
$\pi(1) > \pi(4) > \pi(2) > \pi(3)$	0.418	0.547	0.000	0.000
$\pi(2) > \pi(1) > \pi(3) > \pi(4)$	0.458	0.488	0.000	0.000
$\pi(2) > \pi(1) > \pi(4) > \pi(3)$	0.391	0.408	0.000	0.000
$\pi(3) > \pi(2) > \pi(4) > \pi(1)$	0.302	0.318	0.000	0.000
$\pi(3) > \pi(4) > \pi(1) > \pi(2)$	0.401	0.425	0.000	0.000
$\pi(3) > \pi(4) > \pi(2) > \pi(1)$	0.494	0.531	0.000	0.000
$\pi(4) > \pi(2) > \pi(3) > \pi(1)$	0.375	0.381	0.000	0.000
$\pi(4) > \pi(3) > \pi(1) > \pi(2)$	0.521	0.514	0.000	0.000
$\pi(4) > \pi(3) > \pi(2) > \pi(1)$	0.604	0.614	0.000	0.000
Linearly Representable Rankings	0.000	0.000	0.000	0.000

Notes: The numbers in the table show the approximation errors for each ρ^π, where each ranking π is defined in Column (1). Alternative numbers 1, 2, 3, 4 denote beach, boat, charter, and pier, respectively. For each ranking, columns (2) and (3) show the approximation errors of the linear mixed-logit models computed by the greedy algorithm and the EM algorithm, respectively. Columns (4) and (5) show the approximation errors of the quadratic mixed-logit models calculated by each algorithm. All numbers are rounded to three decimal places. For the greedy algorithm we set the number of iterations to 1000. For the EM algorithm we set the number of random initial points to 10. The greedy algorithm sometimes produces larger approximation errors than the EM algorithm, which is possible with finitely many steps.

The table shows the approximation errors of the linear or quadratic mixed-logit models. We calculated the errors by the greedy algorithm and the EM algorithm. In both algorithms, approximation errors for representable rankings π shown in the bottom row of the table are always zero, as the theorem predicts. On the other hand, the approximation errors for unrepresentable rankings π are almost always larger than 0.4, which means that even the best possible linear mixed-logit model deviates from the corresponding choice probabilities ρ^{π} by 40 percentage points or more on average. Some errors are much larger. For example, the approximation errors of the two rankings $\pi(1) > \pi(2) > \pi(3) > \pi(4)$ and $\pi(1) > \pi(2) > \pi(4) > \pi(3)$ by the linear models are more than 0.67. This suggests that the substitution from alternative 1 to 2 would be difficult to capture. To see this notice that these two rankings are the only rankings in which alternative 1 is the best and alternative 2 is the second-best. See the next subsection for more detail.

Note that the approximation errors by quadratic mixed-logit models are zero, as shown in columns (4) and (5) in the table. This finding is also consistent with the theorem.

6.2.1 Maximal Substitution

Our attention now turns to substitution patterns. We quantify how flexible the linear mixed-logit models are, by measuring the maximal substitution patterns that can be generated by the models. Specifically, for two alternatives j and l, we calculate the following quantity:

$$\sup_{\rho \in \mathcal{P}_{ra}(0|\mu)} \Big(\rho(J \setminus \{j\}, l) - \rho(J, l) \Big) \tag{6}$$

Since ρ is a random utility model, we have $\rho(J \setminus \{j\}, l) \geq \rho(J, l)$.[35] Thus the quantity in (6) can be any nonnegative number that is less than or equal to one. The quantity $\rho(J \setminus \{j\}, l) - \rho(J, l)$ describes how consumers would substitute to alternative l if alternative j becomes unavailable. The supremum of such quantities captures the *maximal substitution pattern* that can be generated by mixed-logit models without fixed effects.[36] We use the greedy algorithm to solve (6), as detailed in section A.4 of the online appendix.[37] Conlon and Mortimer (2021) analyze similar concepts called *diversion ratios*, which measure the fraction of consumers who switch their choices from alternative j to l after the price of alternative j increases marginally. Our measure (6) corresponds to the limit of *diversion*

[35] The property is called monotonicity or regularity.

[36] The quantity in (6) is always 1 when we can choose fixed effects freely. This is because we can always choose fixed effects large enough to approximate degenerate preferences.

[37] We do not use the EM algorithm as it cannot be easily transformed to solve the problem in (6). We note that we can replace the sup in (6) with inf, for which we calculate the *minimal substitution pattern*. This problem can also be solved by the greedy algorithm. We omit the detail since the dataset satisfies the convex independence, so the minimal substitution pattern is 0.

ratios in which the price of alternative j increases to the infinity.[38]

Table 2: Maximal substitution of the linear mixed-logit models

j (drop) \ l (choose)	1	2	3	4
1	-	0.120	0.998	0.998
2	0.317	-	1.000	0.997
3	0.998	1.000	-	0.286
4	0.994	0.998	0.137	-

Notes: The numbers in the table show the value of (6) for each $j, l \in \{1, 2, 3, 4\}$ s.t. $j \neq l$. Alternative numbers 1, 2, 3, 4 denote beach, boat, charter, pier, respectively. All numbers are rounded to three decimal places.

Table 2 shows the values of maximal substitution between the two alternatives j and l. Some numbers in the table are close to one, which implies that the linear mixed-logit models are rich enough to capture flexible substitution from j to l. Some other numbers are smaller. In particular, the maximal substitution between alternative 1 (i.e., beach) and 2 (i.e., private boat) as well as the substitution between alternative 3 (i.e., charter) and 4 (i.e., pier) are at most 0.3. In fact, the maximal substitution from 1 to 2 is 0.12. This means that no matter how the parameters of a linear mixed-logit model are chosen, the maximal substitution from alternative 1 to alternative 2 is very limited.

This finding aligns with the result presented in Table 1, where we observe substantial approximation errors for the two specific rankings: $\pi(1) > \pi(2) > \pi(3) > \pi(4)$ and $\pi(1) > \pi(2) > \pi(4) > \pi(3)$ are very large. In this way, identifying preferences that are hard to approximate with precision helps researchers in evaluating whether their models successfully capture relevant economic behaviors such as substitution patterns.

6.3 Approximation Errors with Fixed Effects

What remains to be explored is the approximation errors with fixed effects. By using fixed effects, we can approximate ρ^π for any ranking π. By Proposition 3, however, for each unrepresentable ranking π and each $\alpha \in (0, 1)$, any random utility model in a neighborhood of $\alpha\rho^\pi + (1-\alpha)\rho^{\pi^-}$ cannot be approximated arbitrarily well by the linear

[38]Precisely speaking, our measure corresponds to the limit of the numerator of the diversion ratio. Some recent papers study the substitution and complementarity property in the discrete choice analysis. See Horan and Adam (2023) and Allen and Rehbeck (2020).

Table 3: Approximation errors to random utility models $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$

Ranking π	Linear mixed-logit		Quadratic mixed-logit	
	Greedy	EM	Greedy	EM
(1)	(2)	(3)	(4)	(5)
Linearly unrepresentable rankings				
$\pi(1) > \pi(2) > \pi(3) > \pi(4)$	0.229	0.255	0.000	0.000
$\pi(1) > \pi(2) > \pi(4) > \pi(3)$	0.229	0.250	0.000	0.000
$\pi(1) > \pi(3) > \pi(2) > \pi(4)$	0.163	0.217	0.000	0.000
$\pi(1) > \pi(4) > \pi(2) > \pi(3)$	0.163	0.173	0.000	0.000
$\pi(2) > \pi(1) > \pi(3) > \pi(4)$	0.192	0.238	0.000	0.000
$\pi(2) > \pi(1) > \pi(4) > \pi(3)$	0.192	0.198	0.000	0.000
Linearly representable rankings	0.000	0.000	0.000	0.000

Notes: The numbers in the table show the approximation errors to $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$, where π is defined in Column (1). All numbers are rounded to three decimal places. For the greedy algorithm we set the number of iterations to 1000. For the EM algorithm we set the number of random initial points to 10.

mixed-logit models with fixed effects. In Table 3, we show the approximation error to $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$ for each unrepresentable ranking.

In both algorithms, the approximation errors to $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$ are always around 0.2 if π is not representable. This means that even the best possible linear mixed-logit model deviates from $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$ by 20 percentage points or more on average.

On the other hand, the approximation errors to $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$ are almost zero if π is representable, as the theorem predicts. Also, the approximation errors by quadratic mixed-logit models are also almost zero, as the theorem again predicts.

7 Concluding Remark

In Section 6, we applied our theorem and algorithms to a real dataset. The results summarized in Tables 1-3 demonstrate how the affine-independence condition and its generic condition $K \geq |J|-1$ serve as straightforward indicators for evaluating the efficacy of random-coefficient models in approximating random utility models.

When the affine condition is not met, as observed in our dataset utilizing the linear mixed-logit model, our methodology enables the quantification of approximation errors. This quantification is detailed in Tables 1 and 3, where errors are often significant. Furthermore, our analysis extends to assessing the limitations in the substitution patterns generated by random coefficient models, as described in Table 2. Additionally, consistent with the predictions of our theorem, we confirm that all approximation errors are zero (up to rounding errors) when employing quadratic models. We believe that these results provide researchers with useful tools for determining the extent to which a given model accurately reflects choice behaviors.

Appendix

A Generalization: Restricting Possible Rankings

As highlighted earlier, there are situations where researchers might deem certain rankings as unreasonable, leading them to constrain the possible set of rankings to $\hat{\Pi} \subset \Pi$. To accommodate this case, define $\mathcal{P}_r(\hat{\Pi}) \equiv \text{co.}\{\rho^\pi | \pi \in \hat{\Pi}\}$. We call an element of $\mathcal{P}_r(\hat{\Pi})$ *a random utility model on* $\hat{\Pi}$.

Following McFadden and Train (2000), in our main results (Theorem 1 and Proposition 1), we considered that case where $\hat{\Pi} = \Pi$ for simplicity. In this section, we provide a necessary and sufficient condition for the approximation of the random utility models on $\hat{\Pi}$.

Since the set $\mathcal{P}_r(\hat{\Pi})$ of random utility models on $\hat{\Pi}$ is also a polytope, we can generalize Lemma 1 for $\mathcal{P}_r(\hat{\Pi})$ by simply changing Π to $\hat{\Pi}$. That is, we have the following result:

Lemma 6. *Let $\hat{\Pi} \subset \Pi$. Then $\mathcal{P}_r(\hat{\Pi}) \subset cl.\mathcal{P}_{ra}(0|\mu)$ if and only if $\rho^\pi \in cl.\mathcal{P}_a(0|\mu)$ for any $\pi \in \hat{\Pi}$.*

This result together with Lemma 2 implies that the following generalization of Corollary 2.

Corollary 3. *Let $\hat{\Pi} \subset \Pi$.*

(i) Let $\mu \in \mathcal{M}$. If any ranking $\pi \in \hat{\Pi}$ is representable in $\mathcal{D}$, then any random utility model on $\hat{\Pi}$ can be approximated arbitrarily well by a random-coefficient ARUM. Moreover, the approximation can be done without fixed effects (i.e., $\eta = 0$). That is, $\mathcal{P}_r(\hat{\Pi}) \subset cl.\mathcal{P}_{ra}(0|\mu)$.

(ii) If some ranking $\pi \in \hat{\Pi}$ is not representable, then there exists a random utility model

on $\hat{\Pi}$ that cannot be approximated arbitrarily well by any random-coefficient ARUM without fixed effects with any $\mu \in \mathcal{M}$. That is, $\mathcal{P}_r(\hat{\Pi}) \not\subset cl.\mathcal{P}_{ra}(0|\mu)$ for any $\mu \in \mathcal{M}$.

Corollary 3 offers a testable condition for determining if the random-coefficient ARUMs, without fixed effects, can adequately approximate any random utility model on $\hat{\Pi} \subset \Pi$. Should the researcher wish to omit certain rankings from their analysis, this Corollary may prove more beneficial than both Theorem 1 and Proposition 1.

B Proofs

In Corollary 4 of Section A.2 of the online appendix, we prove that $\mathcal{P}_{ra}(\eta|\mu) = \text{co}.\mathcal{P}_a(\eta|\mu)$ for any $\mu \in \mathcal{M}$ and $\eta \in \mathbf{R}^{|J|}$, where co. denotes the convex hull. We use this results throughout the proof.

B.1 Proof of Theorem 1

The affine independence of $\{x_j|j \in J\}$ implies that all rankings are representable (by Lemma 3) and thus can be approximated by a sequence of ARUMs (by Lemma 2 (1)). Applying Lemma 1 proves Theorem 1-(i).

When the affine-independence condition fails, Lemma 3 (1) implies a ranking and its reverse ranking are not representable. Lemma 4 gives examples of random utility models that cannot be approximated by random-coefficient ARUMs no matter how one chooses fixed effects.

B.2 Proof of Proposition 1

Lemma 1,2, and 3 (2) imply statement (i) and parts of statement (ii) of Proposition 1. The last statement of statement (ii) can be proved as follows. Consider any stochastic choice function ρ on {J}. Then there exists a sequence of stochastic choice functions $\{\rho_n\}$ such that $\rho_n \to \rho$ and $\rho_n(J, j) > 0$ for any $j \in J$. Fix $\mu \in \mathcal{M}$. Note that our assumption of the convexity of the support implies the connectedness. By Corollary 1 of Norets and Takahashi (2013), ρ_n can be represented as the ARUMs.

B.3 Proof of Lemma 1 and 6

We prove Lemma 6, which implies Lemma 1 with $\hat{\Pi} = \Pi$.

We first prove the if direction. Suppose that $\rho^\pi \in \text{cl}.\mathcal{P}_a(0|\mu)$ for all $\pi \in \hat{\Pi}$. Then $\text{co}.\{\rho^\pi|\pi \in \hat{\Pi}\} \subset \text{co.cl}.\mathcal{P}_a(0|\mu)$. By the fact that $\mathcal{P}_a(0|\mu)$ is bounded, it follows from Theorem 17.2 of Rockafellar (2015) that $\text{cl.co}.\mathcal{P}_a(0|\mu) = \text{co.cl}.\mathcal{P}_a(0|\mu)$. Thus, $\mathcal{P}_r(\hat{\Pi}) \subset \text{cl.co}.\mathcal{P}_a(0|\mu) \subset \text{cl}.\mathcal{P}_{ra}(0|\mu)$. (We used the fact that $\mathcal{P}_{ra}(0|\mu) = \text{co}.\mathcal{P}_a(0|\mu)$ for any $\mu \in \mathcal{M}$.[39])

We now prove the only-if direction. By assumption, $\mathcal{P}_r(\hat{\Pi}) \subset \text{cl.co}.\mathcal{P}_a(0|\mu) = \text{co.cl}.\mathcal{P}_a(0|\mu)$, where the equality holds by the previous argument. Since $\mathcal{P}_r(\hat{\Pi}) \subset \text{co.cl}.\mathcal{P}_a(0|\mu)$, for any $\pi \in \hat{\Pi}$, there exist positive numbers $\{\lambda_i\}_{i=1}^m$ and $\rho^i \in \text{cl}.\mathcal{P}_a(0|\mu)$ for all $i \in \{1, \dots, m\}$ such that $\sum_{i=1}^m \lambda_i = 1$ and $\sum_{i=1}^m \lambda_i \rho^i = \rho^\pi$. Note that $\rho^i \in \text{cl}.\mathcal{P}_a(0|\mu) \subset \mathcal{P}_r(\Pi)$ for each i because $\mathcal{P}_r(\Pi)$ is compact. Since ρ^π is a vertex of $\mathcal{P}_r(\Pi)$, thus ρ^π is an exposed point.[40] Hence, $\sum_{i=1}^m \lambda_i \rho^i = \rho^\pi$ implies $\rho^i = \rho^\pi$ for all i. This means that $\rho^\pi \in \text{cl}.\mathcal{P}_a(0|\mu)$ for all $\pi \in \hat{\Pi}$.

B.4 Proof of Lemma 2

Lemma 2 (1) is easy to prove.

Step 1: For any $\pi \in \Pi$ and any $\mu \in \mathcal{M}$, if a ranking π is representable, then there exists a sequence $\{\rho_n\}$ of $\mathcal{P}_a(0|\mu)$ such that $\rho_n \to \rho^\pi$.

Proof. Assume that a ranking π is representable. This implies that if j is the dominating alternative in D, then $\beta \cdot x_j - \max_{l \in D \setminus \{j\}} \beta \cdot x_l > 0$. If j is dominated by another alternative in D, then $\beta \cdot x_j - \max_{l \in D \setminus \{j\}} \beta \cdot x_l < 0$.

Let ρ_n be the sequence of ARUMs with coefficient $n\beta$. For any positive integer n and any $(D, j) \in \mathcal{D} \times J$ such that $j \in D$, note

$$\rho_n(D, j) \geq \int 1\Big\{n\left(\beta \cdot x_j - \max_{l \in D \setminus j} \beta \cdot x_l\right) \geq \max_{l \in D \setminus j} \varepsilon_l - \varepsilon_j\Big\} d\mu,$$

[39] See Corollary 4 of Section A.2 of the online appendix.

[40] A point of a convex set is an exposed point if there is a supporting hyperplane which contains no other points of the set (Rockafellar (2015), Page 162)

where $1\{\cdot\}$ is the indicator function.[41] By the dominated convergence theorem,

$$\lim_{n\to\infty} \rho_n(D,j) \geq \int \lim_n 1\Big\{n\left(\beta \cdot x_j - \max_{l\in D\setminus j} \beta \cdot x_l\right) \geq \max_{l\in D\setminus j} \varepsilon_l - \varepsilon_j\Big\} d\mu = 1.$$

Since elements for each choice set are nonnegative and sum to 1. This implies $\rho_n \to \rho^\pi$. $\square$

We prove the contrapositive of Lemma 2 (2). We first show that the existence of a converging sequence of ARUMs to ρ^π implies the representability of π (Step 2). Step 3 shows that such a converging sequence of ARUMs exists if a sequence of random coefficient ARUMs converges to ρ^π.

Step 2: If there exists a $\mu \in \mathcal{M}$ and sequences $\{\rho_n\}$ and $\{\rho_n'\}$ of $\mathcal{P}_a(0|\mu)$ such that $\rho_n \to \rho^\pi$ and $\rho_n' \to \rho^{\pi^-}$, then π and π^- are representable.[42]

Proof. Let β_n and β_n' be the coefficient vectors of ρ_n and ρ_n', respectively. We argue that for n large enough, $\beta_n - \beta_n'$ can represent the ranking π.

Consider a binary choice set $\{j, l\}$. Define $\gamma_n \equiv \beta_n \cdot (x_j - x_l)$ and $\gamma_n' \equiv \beta_n' \cdot (x_l - x_j)$. Without loss of generality, assume $\pi(j) > \pi(l)$. Note that γ_n and γ_n' must be bounded below.[43] There are two cases.

Case 1: Consider the case where at least one of γ_n or γ_n' is unbounded above. Sine both of them are bounded below, $\gamma_n + \gamma_n'$ is unbounded above, then there exists N_{jl} such that for any $n > N_{jl}$, $(\beta_n - \beta_n') \cdot (x_j - x_l) = \gamma_n + \gamma_n' > 0$.

Case 2: Both γ_n or γ_n' are bounded above (and below). Thus there exist convergent

[41]Note that $\rho_n(D,j) \equiv \mu(\{\varepsilon | n\beta \cdot x_j + \varepsilon_j \geq \max_{l\in D\setminus j}\{n\beta \cdot x_l + \varepsilon_l\}\}) \geq \mu(\{\varepsilon | n\beta \cdot x_j + \varepsilon_j \geq \max_{l\in D\setminus j} n\beta \cdot x_l + \max_{l\in D\setminus j} \varepsilon_l\}) = \mu(\{\varepsilon | n(\beta \cdot x_j - \max_{l\in D\setminus j} \beta \cdot x_l) \geq \max_{l\in D\setminus j} \varepsilon_l - \varepsilon_j\}) = \int 1\{n(\beta \cdot x_j - \max_{l\in D\setminus j} \beta \cdot x_l) \geq \max_{l\in D\setminus j} \varepsilon_l - \varepsilon_j\} d\mu$.

[42]The statement that "if $\rho_n \to \rho^\pi$ then π is representable" is incorrect. Suppose that $\varepsilon_j > \varepsilon_l$ a.s. if and only if $\pi(j) > \pi(l)$. In this case, even if π is not representable, it is possible that $\rho_n = \rho^\pi$ for sufficiently large n. Note also that π is representable if and only if π^- is representable.

[43]This can be proved by contradiction. Firstly, note that $\varepsilon_l - \varepsilon_j$ is a well-defined random variable by definition. Hence, it is tight: for each $\delta \in (0,1)$, there exists a positive number N_δ such that $\mu\left(\varepsilon | \varepsilon_l - \varepsilon_j \in (-N_\delta, N_\delta)^c\right) < \delta$. Now, if $\{\gamma_n\}_{n=1}$ is not bounded below, we can choose $\delta < 1$ and find a subsequence $\{\gamma_{n_k}\}_{k=1}$ such that $\gamma_{n_k} < -N_\delta$ for all k. On this subsequence we have $\rho_{n_k}(\{j,l\}, j) = \mu\big(\varepsilon | \gamma_n \geq \varepsilon_l - \varepsilon_j\big) \leq \mu\big(\varepsilon | -N_\delta > \varepsilon_l - \varepsilon_j\big) \leq \delta < 1$. Clearly, $\rho_{n_k}(\{j,l\}, j)$ does not converge to $1 = \rho^\pi(\{j,l\}, j)$, a contradiction. Thus we must have that $\{\gamma_n\}_{n=1}$ is bounded below.

subsequences $\{\gamma_{n_k}\}$ and $\{\gamma'_{n_k}\}$. By the dominated convergence theorem

$$1 = \lim_{n_k} \rho_{n_k}(\{j,l\},j) = \mu(\varepsilon | \lim_{n_k} \gamma_{n_k} > \varepsilon_l - \varepsilon_j),\ 1 = \lim_{n_k} \rho'_{n_k}(\{j,l\},l) = \mu(\varepsilon | \varepsilon_l - \varepsilon_j > -\lim_{n_k} \gamma'_{n_k}).$$

We drop the equalities by the absolute continuity of μ. This implies that $\lim_{n_k} \beta_{n_k} \cdot (x_j - x_l) = \lim_{n_k} \gamma_{n_k} > -\lim_{n_k} \gamma'_{n_k} = -\lim_{n_k} \beta'_{n_k} \cdot (x_l - x_j)$ so that $\lim_{n_k}(\beta_{n_k} - \beta'_{n_k}) \cdot (x_j - x_l) > 0$. It follows that there exists N_{jl} such that for any $n_k > N_{jl}$, $\lim_{n_k}(\beta_{n_k} - \beta'_{n_k}) \cdot (x_j - x_l) > 0$.

Finally, although N_{jl} depends on a particular binary choice set, we have a finite number of binary choice sets. Thus, if necessary, we can consider a subsequence $\{(\beta_m - \beta'_m) \cdot (x_j - x_l)\}$ that works for all $j, l \in J$. Taking the maximum N^* of N_{jl} among all binary choice sets, we proved that π is representable. This also implies that π^- is representable. □

In the following, we again use the fact that $\mathcal{P}_{ra}(0|\mu) = \text{co}.\mathcal{P}_a(0|\mu)$ for any $\mu \in \mathcal{M}$.[44]

Step 3: If there exists $\mu \in \mathcal{M}$ and a sequence $\{\rho_n\}$ of $\text{co}.\mathcal{P}_a(0|\mu)$ such that $\rho_n \to \rho^\pi$, then there exists a sequence $\{\rho'_n\}$ of $\mathcal{P}_a(0|\mu)$ such that $\rho'_n \to \rho^\pi$.

Proof. Fix $\pi \in \Pi$. Suppose that there exists a sequence ρ_n of $\text{co}.\mathcal{P}_a(0|\mu)$ such that $\rho_n \to \rho^\pi$ as $n \to \infty$. Let $M = \dim \text{co}.\mathcal{P}_a(0|\mu)$. Then for each ρ_n, by Caratheodory's theorem, there exist $\{\rho_n^i\}_{i=1}^{M+1} \subset \mathcal{P}_a(0|\mu)$ and nonnegative numbers $\{\alpha_n^i\}_{i=1}^{M+1}$ such that $\rho_n = \sum_{i=1}^{M+1} \alpha_n^i \rho_n^i$ and $\sum_{i=1}^{M+1} \alpha_n^i = 1$. Denote $(\alpha_n^i)_{i=1}^{M+1}$ by α_n. Then α_n belongs to a compact set (i.e., M-dimensional simplex). There exists a convergent subsequence $\{\alpha_{n'}\}$. Thus $\rho'_n \equiv \sum_{i=1}^{M+1} \alpha_{n'}^i \rho_{n'}^i$ is a subsequence of $\{\rho_n\}$. For each i, let α_*^i be the limit of $\{\alpha_{n'}^i\}$. Since $\sum_{i=1}^{M+1} \alpha_{n'}^i = 1$ for all n', we have $\sum_{i=1}^{M+1} \alpha_*^i = 1$, so that there must exist i^* such that $\alpha_*^{i^*} \neq 0$.

In the following, we will show that $\rho_{n'}^{i^*} \to \rho^\pi$ as $n' \to \infty$. To show the claim, we prove that if $\rho_{n'}^{i^*} \not\to \rho^\pi$, then $\alpha_{n'}^{i^*} \to 0$, which is a contradiction. Assume that $\rho_{n'}^{i^*} \not\to \rho^\pi$. Then there exist $D \in \mathcal{D}$, $j \in D$, and $\varepsilon > 0$ such that for any integer N there exists $n' > N$ such that $|\rho_n^{i^*}(D,j) - \rho^\pi(D,j)| > \varepsilon$. This implies that for any N there exists $n' > N$ such that $\left| \sum_{i=1}^{M+1} \alpha_{n'}^i \rho_{n'}^i(D,j) - \rho^\pi(D,j) \right| = \sum_{i=1}^{M+1} \alpha_{n'}^i |\rho_{n'}^i(D,j) - \rho^\pi(D,j)| \geq \alpha_{n'}^{i^*} \varepsilon$, where the first equality holds because if $\pi(j) \geq \pi(D)$ then $\rho_{n'}^i(D,j) - \rho^\pi(D,j) \leq 0$ for all i; if $\pi(j) \geq \pi(D)$ does not hold (i.e., j is worse than another alternative $i \in D$ with respect to

[44] See Corollary 4 of Section A.2 of the online appendix.

π), then $\rho^i_{n'}(D,j) - \rho^\pi(D,j) \geq 0$ for all i. Since $\sum_{i=1}^{M+1} \alpha^i_{n'}\rho^i_{n'}(D,j) \to \rho^\pi(D,j)$, it must hold that $\alpha^{i^*}_{n'} \to 0$. □

Steps above show that if there exists a sequence $\{\rho_n\}$ of co.$\mathcal{P}_a(0|\mu)$ such that $\rho_n \to \rho^\pi$, then π is representable. The contraposition of this statement is the second statement of Lemma 2.

B.5 Proof of Corollary 2 and 3

We prove Corollary 3, which implies Corollary 2 with $\hat{\Pi} = \Pi$. Statement (i) follows from Lemma 1 and Lemma 2 (1). To Prove Statement (ii), suppose that there exists a ranking $\pi \in \hat{\Pi}$ that is not representable. Then, by Lemma 2 (1), we have that for any $\mu \in \mathcal{M}$, $\{\rho^\pi, \rho^{\pi^-}\} \notin \text{cl}.\mathcal{P}_{ra}(0|\mu)$.

B.6 Proof of Lemma 3

B.6.1 Proof of Statement (1)

We use the following lemma:

Lemma 7. *Let A be an $r \times n$ real matrix, B be an $l \times n$ real matrix, and E be an real $m \times n$ matrix. Exactly one of the following alternatives is true.*

1. *There is $u \in \mathbf{R}^n$ such that $Au = 0$, $Bu \geq 0$, $Eu \gg 0$.*
2. *There is $\theta \in \mathbf{R}^r$, $\eta \in \mathbf{R}^l$, and $\lambda \in \mathbf{R}^m$ such that $\theta A + \eta B + \lambda E = 0$, $\lambda > 0$ and $\eta \geq 0$,*

where $\gg 0$ means all entries are positive, > 0 means all entries are nonnegative and positive for some entry, and $\geq$ means all entries are nonnegative.

See Theorem 1.6.1 of Stoer and Witzgall (2012) for the proof.

For simplicity of notation, let $J = \{1, 2, \ldots, |J|\}$. For any ranking $\pi \in \Pi$, by relabeling J if necessary, we assume that $\pi(i) > \pi(i+1)$ for all $i \leq |J| - 1$ without loss of generality. We label the following condition as Condition $(*)$: if $\lambda_1 x_1 + \sum_{i=2}^{|J|-1}(\lambda_i - \lambda_{i-1})x_i - \lambda_{|J|-1}x_{|J|} = 0$ and $\lambda_i \geq 0$ for all $i \in \{1, \ldots, |J|-1\}$, then $\lambda_i = 0$ for all $i \in \{1, \ldots, |J|-1\}$.

Step 1: For each $\pi \in \Pi$, Condition $(*)$ holds if and only if π is representable.

Proof. Since $\mathcal{D}$ contains all binary sets, $\pi \in \Pi$ is representable if and only if there exists β such that for any $j, l \in J$, $\pi(j) > \pi(l) \Leftrightarrow \beta \cdot x_j > \beta \cdot x_l$. Fix $\pi \in \Pi$.

$$
\begin{aligned}
&\exists\beta\ \left[x_1 \cdot \beta > x_2 \cdot \beta > \cdots > x_{|J|-1} \cdot \beta > x_{|J|} \cdot \beta\right] \\
&\Longleftrightarrow \exists\beta\ \left[(x_1 - x_2) \cdot \beta > 0, \ldots, (x_{|J|-1} - x_{|J|}) \cdot \beta > 0\right] \Longleftrightarrow \exists\beta\ [E\beta \gg 0] \\
&\Longleftrightarrow \nexists\lambda \in \mathbf{R}^{|J|-1}\ [\lambda > 0,\ \lambda E = 0] \Longleftrightarrow \nexists\lambda \in \mathbf{R}^{|J|-1}\ \left[\lambda > 0, \textstyle\sum_{i=1}^{|J|-1} \lambda_i(x_i - x_{i+1}) = 0\right] \\
&\Longleftrightarrow \nexists\lambda \in \mathbf{R}^{|J|-1}\ \left[\lambda > 0, \lambda_1 x_1 + \textstyle\sum_{i=2}^{|J|-1}(\lambda_i - \lambda_{i-1})x_i - \lambda_{|J|-1}x_{|J|} = 0\right] \\
&\Longleftrightarrow \text{Condition}(*),
\end{aligned}
$$

where $\lambda \equiv (\lambda_1, ..., \lambda_{|J|-1})$ and the third equivalence is obtained by using Lemma 7 with $A, B = 0$ and $E \equiv (x_1 - x_2; \ldots; x_{|J|-1} - x_{|J|}) \in \mathbb{R}^{(|J|-1)\times K}$. $\square$

Step 2: The set $\{x_j | j \in J\}$ is affinely independent if and only if Condition $(*)$ holds for any $\pi \in \Pi$.

Proof. We first show that the only if part. Fix any $\pi \in \Pi$. Without loss of generality assume that $\pi(i) > \pi(i+1)$ for all $i \le |J|-1$. Suppose that $\lambda_1 x_1 + \sum_{i=2}^{|J|-1}(\lambda_i - \lambda_{i-1})x_i - \lambda_{|J|-1}x_{|J|} = 0$ and $\lambda_i \ge 0$ for all i. Then, $\lambda_1 x_1 + \sum_{i=2}^{|J|-1}(\lambda_i - \lambda_{i-1})x_i - \lambda_{|J|-1}x_{|J|} = 0$. Define $\mu_1 = \lambda_1$, $\mu_i = \lambda_i - \lambda_{i-1}$ for all $i \in \{2, \ldots, |J|-1\}$, and $\mu_{|J|} = -\lambda_{|J|-1}$. Then $\sum_{i=1}^{|J|} \mu_i x_i = 0$, and, $\sum_{i=1}^{|J|} \mu_i = \lambda_1 + \sum_{i=2}^{|J|-1}(\lambda_i - \lambda_{i-1}) - \lambda_{|J|-1} = 0$. If $\{x_j | j \in J\}$ is affinely independent, then $\mu_i = 0$ for all $i \in \{1, \ldots, |J|\}$. Hence, $\lambda_i = 0$ for all $i \in \{1, \ldots, |J|-1\}$. This implies Condition $(*)$.

Next we show the if part. Choose any real numbers $\{\mu_i\}_{i=1}^{|J|}$ such that $\sum_{i=1}^{|J|} \mu_i x_i = 0$ and $\sum_{i=1}^{|J|} \mu_i = 0$. Order μ_i by its value so that (after relabelling J) we have $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_{|J|}$. Let $\mu \equiv (\mu_1, ..., \mu_{|J|})$. Define $\lambda_1 = \mu_1$ and $\lambda_i = \sum_{j=1}^{i} \mu_j$ for all $i \in \{2, \ldots, |J|-1\}$.

First we show that $\lambda_i \ge 0$ for all $i \in \{1, \ldots, |J|-1\}$. Suppose by way of contradiction that $\lambda_i < 0$ for some $i \in \{1, \ldots, |J|-1\}$. Then $\mu_i < 0$ because $\mu_1 \ge \cdots \ge \mu_i$. Since $0 > \mu_i \ge \mu_j$ for all $j \ge i$, we have $\sum_{j=i+1}^{|J|} \mu_j < 0$. It follows that $\sum_{j=1}^{|J|} \mu_j = \lambda_i + \sum_{j=i+1}^{|J|} \mu_j < 0$. This contradicts that $\sum_{j=1}^{|J|} \mu_j = 0$. Therefore, $\lambda_i \ge 0$ for all $i \in \{1, \ldots, |J|-1\}$.

Now we show $\mu = 0$. Notice that $\lambda_1 x_1 + \sum_{i=2}^{|J|-1}(\lambda_i - \lambda_{i-1})x_i - \lambda_{|J|-1}x_{|J|} = \mu_1 x_1 + \sum_{i=2}^{|J|-1} \mu_i x_i - \sum_{i=1}^{|J|-1} \mu_i x_{|J|} = \sum_{i=1}^{|J|} \mu_i x_i = 0$, where the second to the last equality holds because $\sum_{i=1}^{|J|} \mu_i = 0$. Therefore, by Condition $(*)$, $\lambda_i = 0$ for all $i \in \{1, \dots, |J|-1\}$. This implies $\mu = 0$. $\square$

B.6.2 Proof of Statement (2)

For any $j \in J$, $x_j \notin \text{co.}(\{x_l | l \in J \setminus \{j\}\}) \Leftrightarrow x_j$ is an extreme point of $\text{co.}(\{x_l | l \in J\}) \Leftrightarrow$ x_j is an exposed point of $\text{co.}(\{x_l | l \in J\}) \Leftrightarrow \exists \beta \; \forall l \in J \setminus \{j\} [\beta \cdot x_j > \beta \cdot x_l] \Leftrightarrow$ all rankings π on $\{J\}$, whose best alternative is j, is representable. The first and third equivalences are by the definitions of extreme points and exposed points, respectively, while the second equivalence is by the fact that $\text{co.}(\{x_l | l \in J\})$ is a polytope.

B.7 Proof of Lemma 5

In this section, we prove Lemma 5. We will prove Lemma 4 later by using Lemma 7-9.

For each positive integer n, define $J_n = \{1, \dots, n\}$. We prove the lemma by an induction on n. Let Π_n be the set of all rankings on J_n.

Let π_n be a ranking over J_n such that $\pi_n(i) > \pi_n(i+1)$ for any $i \leq n-1$. We will prove that ρ^{π_n} and $\rho^{\pi_n^-}$ are adjacent. In particular, we will find $t_n \in \mathbf{R}^{|\mathcal{D}_n| \times |J_n|}$ such that $\rho^{\pi_n} \cdot t_n = \rho^{\pi_n^-} \cdot t_n = 0$, and $\rho^{\sigma_n} \cdot t_n > 0$ for any $\sigma_n \in \Pi_n \setminus \{\pi_n, \pi_n^-\}$.

Induction Base: Let us consider the case of $n = 3$. (Remember that all binary and ternary choice sets are in $\mathcal{D}_n$. The cases for $n = 1$ and $n = 2$ are trivial.) WLOG we consider $\pi(1) > \pi(2) > \pi(3)$ and its reverse. For $b > a > 0$, let $t_3(\{1,2\}, 1) = a$, $t_3(\{2,3\}, 2) = -b$, $t_3(\{1,3\}, 1) = b - a$, and $t_3(\{1,2,3\}, 2) = a + b$. For all other $(D, j) \in \mathcal{D} \times J$, $t_3(D, j) = 0$. A direct calculation shows that $\rho^{\pi_3} \cdot t_3 = \rho^{\pi_3^-} \cdot t_3 = 0$, and $\rho^{\sigma_3} \cdot t_3 > 0$ for any $\sigma_3 \in \Pi_3 \setminus \{\pi_3, \pi_3^-\}$.

Assume that $n \geq 4$. For each i such that $3 \leq i \leq n-1$, we define a set of sets $\mathcal{D}_i \subset 2^{J_i} \setminus \emptyset$ such that (i) $\mathcal{D}_i \subset \mathcal{D}_{i+1}$ and $\mathcal{D}_n = \mathcal{D}$; (ii) for each i, $\{j, l\} \in \mathcal{D}_i$ and $\{j, l, r\} \in \mathcal{D}_i$ for any $j, l, r \in J_i$.

Induction Step: WLOG, let π_{n-1} be the ranking over J_{n-1} such that $\pi_{n-1}(i) > \pi_{n-1}(i+1)$ for any $i \leq n-2$. By the induction hypothesis there exists $t_{n-1} \in \mathbf{R}^{|\mathcal{D}_{n-1}| \times |J_{n-1}|}$ such that $\rho^{\pi_{n-1}} \cdot t_{n-1} = \rho^{\pi_{n-1}^-} \cdot t_{n-1} = 0$, and $\rho^{\sigma_{n-1}} \cdot t_{n-1} > 0$ for $\sigma_{n-1} \in \Pi_{n-1} \setminus \{\pi_{n-1}, \pi_{n-1}^-\}$. Choose a positive number ε_{n-1} such that $0 < \varepsilon_{n-1} < \min_{\sigma_{n-1} \in \Pi_{n-1} \setminus \{\pi_{n-1}, \pi_{n-1}^-\}} \rho^{\sigma_{n-1}} \cdot t_{n-1}$.

We define $t_n \in \mathbf{R}^{|\mathcal{D}_n|\times|J_n|}$ as follows: For each $(D,x) \in \mathcal{D}_n \times J_n$

$$t_n(D,j) = \begin{cases} t_{n-1}(D,j) & \text{if } (D,x) \in (\mathcal{D}_{n-1} \times J_{n-1}) \setminus \{(\{1,2\},1)\}, \\ t_{n-1}(D,j) + \varepsilon_{n-1} & \text{if } (D,j) = (\{1,2\},1), \\ -\varepsilon_{n-1}/(n-1) & \text{if } (D,j) = (\{i,n\},i) \text{ for some } i \in \{1,\cdots,n-1\}, \\ 2\varepsilon_{n-1} & \text{if } (D,j) = (\{n-2,n-1,n\},n-1), \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that $\rho^{\pi_n} \cdot t_n = \rho^{\pi_n^-} \cdot t_n = 0$. Fix $\sigma_n \in \Pi_n \setminus \{\pi_n, \pi_n^-\}$. Let $j \in \{1,\cdots,n\}$ be such that the element n is jth best element in σ_n. There exists $\sigma_{n-1} \in \Pi_{n-1}$ such that the ranking σ_n can be written as $(\sigma_{n-1}^{-1}(n-1),\cdots,\sigma_{n-1}^{-1}(n-(j-1)), n, \sigma_{n-1}^{-1}(n-j),\cdots,\sigma_{n-1}^{-1}(1))$ in decreasing order of the ranking if $2 \le j \le n-1$. (If $j=1$, then n is the best element in σ_n. If $j=n$, then n is the worst element in σ_n.)

First notice that by the definition of t_n and $\rho^{\sigma_{n-1}} = \rho^{\sigma_n}$ on $\{1,2\}$, $\rho^{\sigma_n} \cdot t_n = \rho^{\sigma_{n-1}} \cdot t_{n-1} + \varepsilon_{n-1}\rho^{\sigma_{n-1}}(\{1,2\},1) - \dfrac{\varepsilon_{n-1}}{n-1}(j-1) + 2\varepsilon_{n-1}\rho^{\sigma_n}(\{n-2,n-1,n\},n-1)$, where the second term of the right hand side follows since in σ_n, there are $j-1$ elements that are better than n.

Case 1: $\sigma_{n-1} = \pi_{n-1}$. Note that $\rho^{\sigma_{n-1}}(\{1,2\},1) = \rho^{\pi_{n-1}}(\{1,2\},1) = 1$, and also $\rho^{\sigma_n}(\{n-2,n-1,n\},n-1) = 0$. Thus, $\rho^{\sigma_n} \cdot t_n = 0 + \varepsilon_{n-1} - \dfrac{\varepsilon_{n-1}}{n-1}(j-1) + 0 > 0$, where the last inequality holds because $j < n$. (If $j = n$, then $\sigma_{n-1} = \pi_{n-1}$ implies that $\sigma_n = \pi_n$.)

Case 2: $\sigma_{n-1} = \pi_{n-1}^-$. Note that $\rho^{\sigma_{n-1}} \cdot t_{n-1} = \rho^{\pi_{n-1}^-} \cdot t_{n-1} = 0$ and $\rho^{\sigma_{n-1}}(\{1,2\},1) = \rho^{\pi_{n-1}^-}(\{1,2\},1) = 0$. Note also that $\rho^{\sigma_n}(\{n-2,n-1,n\},n-1) = 1$ because $n-1$ is the best element in σ_n (except the case in which n is the best element in σ_n and $\sigma_{n-1} = \pi_{n-1}^-$, then the ranking in σ_n coincides with π_n^-). Thus, $\rho^{\sigma_n} \cdot t_n = 0 + 0 - \dfrac{\varepsilon_{n-1}}{n-1}(j-1) + 2\varepsilon_{n-1} > 0$.

Case 3: $\sigma_{n-1} \notin \{\pi_{n-1}, \pi_{n-1}^-\}$. Thus, $\rho^{\sigma_n} \cdot t_n > \varepsilon_{n-1} - \dfrac{\varepsilon_{n-1}}{n-1}(j-1) > 0$, where the first inequality holds by $\rho^{\sigma_{n-1}} \cdot t_{n-1} > \varepsilon_{n-1}$ and the second inequality holds by $j \le n$.

B.8 Proof of Lemma 4

To prove the lemma, we prove the following lemmas. Fix a ranking π that is not representable. For any $\alpha \in (0,1)$, define $\rho_\alpha^\pi \equiv \alpha\rho^\pi + (1-\alpha)\rho^{\pi^-}$. We first will show statement

(a) mentioned after Lemma 4 in Section 4.

Lemma 8. *Let $\mu \in \mathcal{M}$. For any $\alpha \in (0,1)$, $\rho^{\pi}_{\alpha} \notin cl. \bigcup_{\eta} \mathcal{P}_a(\eta|\mu)$.*

Proof. Choose any $j, l, r \in J$ such that $\pi(j) > \pi(l) > \pi(r)$. Suppose by contradiction that $\rho^{\pi}_{\alpha} \in \text{cl.} \bigcup_{\eta} \mathcal{P}_a(\eta|\mu)$. This implies that there exists a sequence $\{\rho_n\}_{n=1}^{\infty} \subset \bigcup_{\eta} \mathcal{P}_a(\eta|\mu)$ converging to ρ^{π}_{α}. Let (η_n, β_n) be the fixed-effects-coefficients pair corresponding to ρ_n. Define $\gamma_{n,jl} \equiv \beta_n \cdot (x_j - x_l) + \eta_{nj} - \eta_{nl} l$. Define $\gamma_{n,jr}$, $\gamma_{n,lr}$, $\gamma_{n,lj}$, $\gamma_{n,rl}$ and $\gamma_{n,rj}$ similarly. First consider the sequence $\{\gamma_{n,jl}\}_{n=1}^{\infty}$.

Step 1: The sequence $\{\gamma_{n,jl}\}_{n=1}^{\infty}$ is uniformly bounded.

Proof. We prove this by contradiction. Firstly, note that $\varepsilon_l - \varepsilon_j$ is a tight random variable: for each $\delta > 0$, there exists a positive number N_δ such that $\mu\left(\varepsilon | \varepsilon_l - \varepsilon_j \in (-N_\delta, N_\delta)^c\right) < \delta$. Now, if $\{\gamma_{n,jl}\}_{n=1}^{\infty}$ is not a bounded sequence, we choose $\delta = \min\{\frac{\alpha}{2}, 1 - \alpha - \varepsilon\}$, for some $0 < \varepsilon < 1 - \alpha$, and find a subsequence $\{\gamma_{n_k,jl}\}_{k=1}^{\infty}$ such that $|\gamma_{n_k,jl}| > N_\delta$. On this subsequence we either have $\rho_{n_k}(\{j,l\}, j) = \mu\big(\varepsilon | \gamma_{n_k,jl} > \varepsilon_l - \varepsilon_j\big) \geq \mu\big(\varepsilon | N_\delta > \varepsilon_l - \varepsilon_j\big) \geq 1 - \delta \geq \alpha + \varepsilon > \alpha$ or $\rho_{n_k}(\{j,l\}, j) = \mu\big(\varepsilon | \gamma_{n_k,jl} > \varepsilon_l - \varepsilon_j\big) \leq \mu\big(\varepsilon | -N_\delta > \varepsilon_l - \varepsilon_j\big) \leq \delta \leq \frac{\alpha}{2} < \alpha$. Clearly, $\rho_{n_k}(\{j,l\}, j)$ does not converge to $\alpha = \rho^{\pi}_{\alpha}(\{j,l\}, j)$. We reach a contradiction and thus$\{\gamma_{n,jl}\}_{n=1}^{\infty}$ must be uniformly bounded. $\square$

Similar conclusions hold for $\gamma_{n,jr}$, $\gamma_{n,lr}$, $\gamma_{n,lj}$, $\gamma_{n,rl}$, $\gamma_{n,rj}$. Given Step 1, we can select convergent subsequences $\{(\gamma_{n_k,jl}, \gamma_{n_k,jr}, \gamma_{n_k,lr}, \gamma_{n_k,lj}, \gamma_{n_k,rj}, \gamma_{n_k,rl})\}_{k \in \mathbf{N}}$. We denote the limits as $(\gamma^*_{jl}, \gamma^*_{jr}, \gamma^*_{lr}, \gamma^*_{lj}, \gamma^*_{rj}, \gamma^*_{rl})$. We consider the corresponding stochastic choice functions ρ_{n_k}. Note that, by definition, $\lim_{n_k} \rho_{n_k} = \rho^{\pi}_{\alpha}$.

For any $s, t \in \{j, l, r\}$ and a nonnegative number n, define $E_{n,st} = \{\varepsilon | \gamma_{n,st} > \varepsilon_t - \varepsilon_s\}$, $E_{st} = \{\varepsilon | \gamma^*_{st} \geq \varepsilon_t \varepsilon_s\}$, and $E'_{st} = \{\varepsilon | \gamma^*_{st} > \varepsilon_t - \varepsilon_t\}$. Since $\mu \in \mathcal{M}$ is absolutely continuous with respect to the Lebesgue measure, $\mu\{\varepsilon | \gamma^*_{st} = \varepsilon(t) - \varepsilon(s)\} = 0$. Thus $\mu(E_{st}) = \mu(E'_{st})$.

Step 2: (i) $E_{jl} = E_{jr}$ and $E_{rl} = E_{rj}$ up to a measure zero set; (ii) $\mu(E_{jl} \cap E_{jr}) = \alpha$ and $\mu(E_{rl} \cap E_{rj}) = 1 - \alpha$.

Proof. By Fatou's lemma i) $\alpha = \rho^{\pi}_{\alpha}(\{j,l,r\}, j) = \limsup \rho_n(\{j,l,r\}, j) = \limsup$ $\mu(E_{n,jl} \cap E_{n,jr}) \leq \mu(\limsup(E_{n,jl} \cap E_{n,jr})) \leq \mu(E_{jl} \cap E_{jr})$, and, (ii) $\alpha = \rho^{\pi}_{\alpha}(\{j,l\}, j) = \liminf \mu(E_{n,jl}) \geq \mu(\liminf E_{n,jl})$. By definition, $\liminf E_{n,jl} \equiv \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} E_{i,jl}$, and we have $E'_{jl} \subset \liminf E_{n,jl} \subset E_{jl}$. Since $\mu(E'_{jl}) = \mu(E_{jl})$, it follows that $\alpha \geq \mu(\liminf E_{n,jl}) =$

$\mu(E_{jl})$. In the same way, we have $\alpha \geq \mu(E_{jr})$. Thus we have $\mu(E_{jl} \cap E_{jr}) \geq \alpha \geq \mu(E_{jl}) = \mu(E_{jr})$. It follows that $E_{jl} = E_{jr}$ up to a measure zero set and $\mu(E_{jl} \cap E_{jr}) = \alpha$. By symmetry, we obtain $E_{rl} = E_{rj}$ up to a measure zero set and $\mu(E_{rl} \cap E_{rj}) = 1 - \alpha$. $\square$

Define a column vector $\varepsilon = (\varepsilon_j, \varepsilon_l, \varepsilon_r) \in \mathbf{R}^3$. Define Ω to be the support of μ projected onto the coordinates $(\varepsilon_j, \varepsilon_l, \varepsilon_r)$. Define μ_Ω to be the measure μ restricted on $(\varepsilon_j, \varepsilon_l, \varepsilon_r)$: for any Borel measurable set S on $\mathbf{R}^3$, $\mu_\Omega(\{(\varepsilon_j, \varepsilon_l, \varepsilon_r) \in S\}) = \mu(S \times \mathbf{R}^{|J|-3})$. Further define $A = \{(\varepsilon_j, \varepsilon_l, \varepsilon_r) | U\varepsilon \geq c\}$ and $B = \{(\varepsilon_j, \varepsilon_l, \varepsilon_r) | U\varepsilon \leq c\}$, where

$$U = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \quad c = \begin{bmatrix} \gamma^*_{lj} \\ \gamma^*_{rl} \end{bmatrix}. \tag{7}$$

Step 3: $\mu_\Omega(A) = \alpha$, $\mu_\Omega(B) = 1 - \alpha$, and $\mu_\Omega(A \cup B) = 1$.

Proof. By Step 2, $\mu(E_{jl} \cap E_{jr}) = \alpha$ and $\mu(E_{rl} \cap E_{rj}) = 1 - \alpha$. Remember E_{st} is the event that s is chosen over t in the binary set $\{s, t\}$. Notice $E_{jl} \cap E_{jr}$ and $E_{rl} \cap E_{rj}$ have measure zero intersections by the transitivity of rankings, so the two events partition the probability space (ignoring measure zero events).

Notice that $\alpha = \rho^\pi_\alpha(\{r, l\}, r) = \mu(E_{rl})$. Since the event E_{jr} is incompatible with the event $E_{rl} \cap E_{rj}$ up to a measure zero set, E_{jr} must completely lie within the event $E_{jl} \cap E_{jr}$. Moreover, since $\mu(E_{jr}) = \alpha = \mu(E_{jl} \cap E_{jr})$, the event E_{jr} coincides with the event $E_{jl} \cap E_{jr}$ (ignoring measure zero events). Finally, notice the event A is the intersection of $E_{jl} \cap E_{jr}$ and E_{jr}. Thus, $\mu_\Omega(A) = \alpha$. In a similar way, we can show that $\mu_\Omega(B) = 1 - \alpha$. Thus we have $\mu_\Omega(A \cup B) = 1$.[45] $\square$

Remember $\Omega \subset A \cup B$.[46] Define $A' \equiv \{\varepsilon | U\varepsilon > c\}$ and $B' \equiv \{\varepsilon | U\varepsilon < c\}$. Notice that the sets A' and B' are two disjoint sets determined by two half spaces.[47]

Step 4: There exists $\varepsilon_a \in A' \cap \Omega$ and $\varepsilon_b \in B' \cap \Omega$ such that $\varepsilon_\lambda \equiv \lambda\varepsilon_a + (1-\lambda)\varepsilon_b \notin A \cup B$

[45] First we can show that notice that $1 - \alpha = \rho^\pi_\alpha(\{j, l\}, l) = \mu(E_{lj})$. Since the event E_{lj} is not compatible with the event $E_{jl} \cap E_{jr}$, E_{lj} must coincide with the event $E_{rl} \cap E_{rj}$ (ignoring measure zero events). Finally notice that the event B is the intersection of $E_{rl} \cap E_{rj}$ and E_{lj}, up to a measure zero events. Thus $\mu_\Omega(B) = 1 - \alpha$.

[46] If not, there exists a point $x \notin A \cup B$ and $\delta > 0$ such that the ball centers around x with radius δ, $B_\delta(x)$, satisfies $\mu(B_\delta(x)) > 0$ and $B_\delta(x) \cap (A \cup B) = \emptyset$. This is a contradiction since $\mu_\Omega(B_\delta(x)) + \mu_\Omega(A) + \mu_\Omega(B) > 1$.

[47] Consider two hyperplanes $H_1 \equiv \{(\varepsilon_j, \varepsilon_l, \varepsilon_r) | \varepsilon_j - \varepsilon_l = \gamma^*_{lj}\}$ and $H_2 \equiv \{(\varepsilon_j, \varepsilon_l, \varepsilon_r) | \varepsilon_l - \varepsilon_r = \gamma^*_{rl}\}$. Notice that the set A' is the intersection of two half spaces $H_1^+ \cap H_2^+$; similarly, $B' = H_1^- \cap H_2^-$, where $H_1^+ \equiv \{(\varepsilon_j, \varepsilon_l, \varepsilon_r) | \varepsilon_j - \varepsilon_l > \gamma^*_{lj}\}$ and $H_1^- \equiv \{(\varepsilon_j, \varepsilon_l, \varepsilon_r) | \varepsilon_j - \varepsilon_l < \gamma^*_{lj}\}$. ($H_2^+$ and H_2^- can be defined in a similar way.)

for some $\lambda \in [0, 1]$.

Proof. We prove by contradiction. For any $(\varepsilon_a, \varepsilon_b) \in A' \cap \Omega \times B' \cap \Omega$, define $\varepsilon_\lambda \equiv \lambda\varepsilon_a + (1-\lambda)\varepsilon_b \in A \cup B$ for all $\lambda \in [0, 1]$. Consider the line segment $\{\varepsilon_\lambda\}_{\lambda\in[0,1]}$ and denote it by $l(\varepsilon_a, \varepsilon_b)$. The line segment $l(\varepsilon_a, \varepsilon_b)$ must intercept with the line $A \cap B = \{\varepsilon\,|U\varepsilon = c\}$ for any $(\varepsilon_a, \varepsilon_b) \in (A' \cap \Omega) \times (B' \cap \Omega)$. This is because A' and B' are two disjoint sets.

We choose two arbitrary pairs $(\varepsilon_a^*, \varepsilon_b^*)$ and $(\varepsilon_{a'}^*, \varepsilon_b^*)$ such that ε_a^*, $\varepsilon_{a'}^*$ and ε_b^* are not collinear.[48] The points ε_a^*, $\varepsilon_{a'}^*$ and ε_b^* together determine a hyperplane. Denote the hyperplane by H. Notice that the line $A \cap B$ belongs to H because $l(\varepsilon_a^*, \varepsilon_b^*)$ and $l(\varepsilon_{a'}^*, \varepsilon_b^*)$ intercept with $A \cap B$ at two different points, say ε_1 and ε_2 . Since $l(\varepsilon_1, \varepsilon_2) \subset H$, we have $A \cap B \subset H$ because $l(\varepsilon_1, \varepsilon_2) \subset A \cap B$ and the affine hull of $l(\varepsilon_1, \varepsilon_2)$ is $A \cap B$ and H is affine.

Secondly, we show that $B' \cap \Omega$ belongs to H. To show this choose any $\varepsilon_{b'} \in B' \cap \Omega$. Denote the intersection of $l(\varepsilon_a^*, \varepsilon_{b'})$ and $A \cap B$ by $\hat{\varepsilon}$. By the above argument $\hat{\varepsilon} \in H$ because $\hat{\varepsilon} \in A \cap B$. Moreover $\varepsilon_a^* \in H$. Thus $\varepsilon_{b'} \in H$ because $\varepsilon_{b'} = \varepsilon_a^* + \lambda(\varepsilon_a^* - \hat{\varepsilon})$ for some $\lambda \in \mathbf{R}$.

In the same way, we can show that $A' \cap \Omega \subset H$. Since $A \cap B \subset H$, $A' \cap \Omega \subset H$, and $B' \cap \Omega \subset H$, we have $A \cap \Omega \subset H$ and $B \cap \Omega \subset H$. Since $\Omega \subset A \cup B$ as noted before Step 4, we have $\Omega \subset H$. It follows that the support of μ is contained in $H \times \mathbf{R}^{|J|-3}$. The set $H \times \mathbf{R}^{|J|-3}$ has the Lebesgue measure 0. By the absolute continuity of μ, this implies that μ has zero total measure. This is a contradiction because μ is a probability measure. $\square$

Using the ε_a and ε_b shown to exist in Step 4, and the assumption that supp.μ is convex, we have $\varepsilon_\lambda \in \text{supp.}\mu \cap (A \cup B)^c$ for some $\lambda \in [0, 1]$. Moreover, $(A \cup B)^c$ is open. It follows from the definition of the support that there exists $r > 0$ such that the ball centers around ε_λ with radius r, $B_r(\varepsilon_\lambda)$, satisfies $\mu(B_r(\varepsilon_\lambda)) > 0$ and $B_r(\varepsilon_\lambda) \not\subset A \cup B$. This contradicts with $\mu(A \cup B) = 1$. $\blacksquare$

Given Lemma 8, in order to prove Lemma 4, it suffices to show that even with using mixtures, it is impossible to approximate ρ_α^π. We need two more lemma.

Lemma 9. *(i)For any $\rho \in cl.\bigcup_\eta \mathcal{P}_a(\eta|\mu)$, if $\rho \notin \{\rho^\pi, \rho^{\pi^-}\}$ then $\rho \notin \{\rho_\alpha^\pi | \alpha \in [0, 1]\}$; (ii) Let (t, a) be as in Definition 7 with a pair (ρ^π, ρ^{π^-}) of adjacent rankings. For any*

[48] This is possible because if we cannot find such two pairs, it implies that $A \cap \Omega$ is a line and hence has μ-measure 0 by absolute continuity. This is a contradiction since $\mu_\Omega(A \cap \Omega) = \alpha$.

$\rho \in cl. \bigcup_\eta \mathcal{P}_a(\eta|\mu)$, *if* $\rho \notin \{\rho^\pi, \rho^{\pi^-}\}$ *then* $\rho \cdot t > a$.

Proof. To show (i), suppose by way of contradiction that $\rho \in \{\rho^\pi_\alpha | \alpha \in [0,1]\}$. Since $\rho \notin \{\rho^\pi, \rho^{\pi^-}\}$, $\rho = \rho^\pi_\alpha$ for some $\alpha \in (0,1)$. By Lemma 8, $\rho \notin \text{cl.} \bigcup_\eta \mathcal{P}_a(\eta|\mu)$, which is a contradiction.

Now we will show (ii) by using (i). Since $\rho \in \mathcal{P}_r$, it can be written as a convex combination of ρ^πs: $\rho = \sum_{\sigma \in \Pi} \mu(\sigma)\rho^\sigma = \mu(\pi)\rho^\pi + \mu(\pi^-)\rho^{\pi^-} + \sum_{\sigma \in \Pi \setminus \{\pi, \pi^-\}} \mu(\sigma)\rho^\sigma$. By (i), $\rho \in \text{cl.} \bigcup_\eta \mathcal{P}_a(\eta|\mu)$ and $\rho \notin \{\rho^\pi, \rho^{\pi^-}\}$ implies that $\rho \notin \{\rho^\pi_\alpha | \alpha \in [0,1]\}$. Thus, we must have one of weights $\mu(\sigma)$s in the third term is positive. By definition for any $\sigma \notin \{\pi, \pi^-\}$, $\rho^\sigma \cdot t > a = \rho^\pi \cdot t = \rho^{\pi^-} \cdot t$. Thus we can conclude that $\rho \cdot t > a$. ■

Lemma 10. *Let* $\mu \in \mathcal{M}$. *If there exists a sequence of random-coefficient ARUMs with fixed effects* η_n *converging to* ρ^π_α *for some* $\alpha \in (0,1)$, *then there exists two sequences of ARUMs with fixed effects* η_n *that converge to* ρ^π *and* ρ^{π^-}, *respectively.*

Proof. Since $\bigcup_\eta \mathcal{P}_{ra}(\eta|\mu) = \bigcup_\eta co.\mathcal{P}_a(\eta|\mu)$ by Corollary 4 in the online appendix, there exists $\sum_{i=1}^{M+1} \mu_{ni}\rho_{ni} \to \rho^\pi_\alpha$, where $M = dim\mathcal{P}_r + 1$ (allowing $\mu_{ni} = 0$ for some i). Since the sets of weights and random utility models are compact, we can extract converging subsequences such that for all i, $\mu_{n_k i} \to \mu^*_i$ and $\rho_{n_k i} \to \rho^*_i$ as $n_k \to \infty$. Thus, $\sum_i \mu_{n_k i}\rho_{n_k i} \to \sum_i \mu^*_i \rho^*_i = \rho^\pi_\alpha$. Moreover, since $\rho_{n_k i} \in \bigcup_\eta \mathcal{P}_a(\eta|\mu)$, we have $\rho^*_i \in \text{cl.} \bigcup_\eta \mathcal{P}_a(\eta|\mu)$.

We argue that there exist some i, j such that $\rho^*_i = \rho^\pi$ and $\rho^*_j = \rho^{\pi^-}$. Suppose, by contradiction and without loss of generality, that $\rho^*_i \neq \rho^\pi$ for any i.[49]

Let (t, a) be as in Definition 7 with a pair (ρ^π, ρ^{π^-}) of adjacent rankings.

We will consider two cases.

Case 1: $\rho^*_i \neq \rho^{\pi^-}$ for any i. For all i, $\rho^*_i \in \text{cl.} \bigcup_\eta \mathcal{P}_a(\eta|\mu)$ and $\rho^*_i \notin \{\rho^\pi, \rho^{\pi^-}\}$. Then, by Lemma 9 (ii), $\rho^*_i \cdot t > a$ for all i. Thus, $\left(\sum_i \mu^*(i)\rho^*_i\right) \cdot t = \sum_i \mu^*(i)\rho^*_i \cdot t > a$. On the other hand by Definition 7, $\left(\sum_i \mu^*(i)\rho^*_i\right) \cdot t = \rho^\pi_\alpha \cdot t = a$. This is a contradiction.

Case 2: $\rho^*_i = \rho^{\pi^-}$ for some i. Define $I = \{i \in \{1, \dots, M+1\} | \rho^*_i = \rho^{\pi^-}\}$. First notice that there exists $i \in \{1, \dots, M+1\} \setminus I$ such that $\mu^*(i) > 0$. (If such i does not exist, then $\rho^{\pi^-} = \sum_i \mu^*(i)\rho^*_i = \rho^\pi_\alpha$, which contradicts with $\alpha \notin \{0,1\}$.) Then, $a = \rho^\pi_\alpha \cdot t = \sum_i \mu^*(i)\rho^*_i \cdot t = \sum_{i \in I} \mu^*(i)\rho^{\pi^-} \cdot t + \sum_{i \notin I} \mu^*(i)\rho^*_i \cdot t > a$. ■

[49] The proof for the other case is exactly the same after changing ρ^{π^-} to ρ^π and ρ^π to ρ^{π^-}.

B.8.1 Main Proof of Lemma 4 by Using Lemma 8, 9, 10

As mentioned, given Lemma 8, it suffices to show that even with using mixtures, it is impossible to approximate ρ^{π}_{α}.

Let π be a ranking that is not representable. By Lemma 5, ρ^{π} and ρ^{π^-} are adjacent. Now suppose by way of contradiction that there exists a sequence of random-coefficient ARUMs with fixed effects that approximates ρ^{π}_{α} for some $\alpha \in (0,1)$. Then by Lemma 10, there exist two sequences $\{\rho_n\}$ and $\{\rho'_n\}$ such that (i) $\rho_n \to \rho^{\pi}$ and (ii) $\rho'_n \to \rho^{\pi^-}$. Let η_n be the sequence of fixed effects, β_n the sequence of coefficients associated with ρ_n, and β'_n the sequence of coefficients associated with ρ'_n.

Given (i), by exactly the same argument as Step 2 of Lemma 2, we can prove that there exists a large positive integer N_1 such that for any $n \geq N_1$, we have $\beta_n \cdot x_j + \eta_n(j) > \beta_n \cdot x_l + \eta_n(l)$ for any $j, l \in J$ such that $\pi(j) > \pi(l)$. Similarly by (ii), there exists a large N_2 such that for any $n \geq N_2$, we have $\beta'_n \cdot x_j + \eta_n(j) > \beta'_n \cdot x_l + \eta_n(l)$ for any $j, l \in J$ such that $\pi^-(j) > \pi^-(l)$.

Fix any $j, l \in J$ such that $\pi(j) > \pi(l)$. Fix any number $n_{jl} \geq \max\{N_1, N_2\}$. Then for nay $n \geq n_{jl}$, we have $\beta_n \cdot x_j + \eta_n(j) > \beta_n \cdot x_l + \eta_n(l)$. Since $\pi^-(l) > \pi^-(j)$, we have $-\beta'_n \cdot x_j - \eta_n(j) > -\beta'_n \cdot x_l - \eta_n(l)$. Summing the two inequalities, we have $(\beta_n - \beta'_n) \cdot x_j > (\beta_n - \beta'_n) \cdot x_l$. Because the number of binary choice sets is finite, we can find $n^* > n_{jl}$ for any $j, l \in J$ with $\pi(j) > \pi(l)$, such that $(\beta_{n^*} - \beta'_{n^*}) \cdot x_j > (\beta_{n^*} - \beta'_{n^*}) \cdot x_l$. This contradicts with the fact that π is not representable.

References

Agranov, M. and P. Ortoleva (2017): "Stochastic choice and preferences for randomization," *Journal of Political Economy*, 125, 40–68.

Allen, R. and J. Rehbeck (2020): "Hicksian Complementarity and Perturbed Utility Models," *Economic Theory Bulletin*, 8, 245–261.

Apesteguia, J. and M. Ballester (2018): "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences," *Journal of Political Economy*, 126, 74–106.

Athey, S. and G. W. Imbens (2007): "Discrete Choice Models with Multiple Unobserved Choice Characteristics," *International Economic Review*, 48, 1159–1192.

Barron, A. R., A. Cohen, W. Dahmen, and R. A. DeVore (2008): "Approximation and Learning by Greedy Algorithms," *Annals of Statistics*, 36, 64–94.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 841–890.

Berry, S. T. (1994): "Estimating Discrete-choice Models of Product Differentiation," *RAND Journal of Economics*, 242–262.

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Block, H. D. and J. Marschak (1960): "Random Orderings and Stochastic Theories of Responses," *Contributions to Probability and Statistics*, 2, 97–132.

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics: Methods and Applications*, Cambridge University Press.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models," *Handbook of Econometrics*, 6, 5549–5632.

Compiani, G. (2022): "Market Counterfactuals and the Specification of Multiproduct Demand: A Nonparametric Approach," *Quantitative Economics*, 13, 545–591.

Conlon, C. and J. H. Mortimer (2021): "Empirical Properties of Diversion Ratios," *RAND Journal of Economics*, 52, 693–726.

Croissant, Y. (2020): "Estimation of Random Utility Models in R: The mlogit Package," *Journal of Statistical Software*, 95, 1–41.

Dagsvik, J. K. (1994): "Discrete and Continuous Choice, Max-Stable Processes, and Independence from Irrelevant Attributes," *Econometrica*, 1179–1205.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1–22.

Doignon, J.-P. and K. Saito (2023): "Adjacencies on Random Ordering Polytopes and Flow Polytopes," *Journal of Mathematical Psychology*, 114, 102768.

Frick, M., R. Iijima, and T. Strzalecki (2019): "Dynamic Random Utility," *Econometrica*, 87, 1941–2002.

Ghalanos, A. and S. Theussl (2015): *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, r package version 1.16.

Herriges, J. A. and C. L. Kling (1999): "Nonlinear Income Effects in Random Utility Models," *Review of Economics and Statistics*, 81, 62–72.

Horan, S. and A. Adam (2023): "Slutsky Patterns of Substitution in Discrete Choice," Working Paper.

Lu, J. and K. Saito (2021): "Mixed Logit and Pure Characteristic Models," Working Paper.

Luce, D. (1959): *Individual Choice Behavior*, New York: Wiley.

Matousek, J. (2013): *Lectures on Discrete Geometry*, vol. 212, Springer Science & Business Media.

McFadden, D. and K. Train (2000): "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 447–470.

Norets, A. and S. Takahashi (2013): "On the Surjectivity of the Mapping between Utilities and Choice Probabilities," *Quantitative Economics*, 4, 149–155.

Rockafellar, R. T. (2015): *Convex Analysis*, Princeton University Press.

Stoer, J. and C. Witzgall (2012): *Convexity and Optimization in Finite Dimensions I*, vol. 163, Springer Science & Business Media.

Tebaldi, P., A. Torgovitsky, and H. Yang (2023): "Nonparametric Estimates of Demand in the California Health Insurance Exchange," *Econometrica*, 91, 107–146.

Thomson, C. J. and S. J. Crooke (1991): "Results of the Southern California Sportfish Economic Survey," *NOAA Technical Memorandum NMFS*.

Train, K. E. (2009): *Discrete Choice Methods with Simulation*, Cambridge University Press.

Ye, Y. (1987): "Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming," Ph.D. thesis, Department of ESS, Stanford University.

For Online Publication

Online Appendix for "Approximating Choice Data by Discrete Choice Models"

Haoge Chang, Yusuke Narita, and Kota Saito

A Algorithms to Compute Approximation Errors

A.1 EM Algorithm

To compute approximation errors in Table 1, we fit a finite-mixture logit model to each deterministic ranking by the method of maximum likelihood. The data input is the stochastic choice function $\widehat{\rho}(D, j)$ and characteristics of each alternative j. We choose the number of mixtures, $M = 18$, according to the theoretical upper bound suggested by Corollary 4 in section A.2. Given the number of mixtures, the model has two sets of parameters: (1) mixture weights $\{\lambda_i\}_{i=1}^M$ and (2) coefficients for each mixture $\{\beta_i\}_{i=1}^M$. The log-likelihood function of a finite mixture model with M mixtures is

$$\mathcal{L}(\{\lambda_i\}_{i=1}^M, \{\beta_i\}_{i=1}^M) \equiv \sum_{D \in \mathcal{D}} \sum_{j \in D} \widehat{\rho}(D, j) \log \sum_{i=1}^M \lambda_i \frac{\exp(\beta_i \cdot x_j)}{\sum_{l \in D} \exp(\beta_i \cdot x_l)}.$$

We estimate the parameters by the EM algorithm (Dempster et al., 1977; Train, 2009). We implement the algorithm according to Chapter 14 in Train (2009). We terminate the algorithm when the change of the implied l_2 distance between the estimated choice probability and the target choice probability becomes smaller than $\frac{1}{10^6}$ between two successive runs.

Our use of the maximum likelihood method with the EM algorithm is motivated by the following observation: if the affine-independence condition is satisfied and the target choice probability is an interior random utility model $\hat{\rho} \in \text{rint}.\mathcal{P}_r$, then the model that maximizes the likelihood will yield a perfect fit to the target probability. Maximum likelihood method therefore minimizes the approximation error metric in (5).

To see this, notice that under the affine-independence condition, Proposition 4 and

5 in the next section imply that any interior random utility model can be represented by a finite mixture of logit models with $M = \sum_{D\in\mathcal{D}}(|D|-1)+1$ mixtures. That is, if $M = \sum_{D\in\mathcal{D}}(|D|-1)+1$, there exists a set of parameters $\{(\beta_i^*, \lambda_i^*)\}_{i=1}^M$ such that $\sum_{i=1}^M \lambda_i^* \dfrac{\exp(\beta_i^*\cdot x_j)}{\sum_{l\in D}\exp(\beta_i^*\cdot x_l)} = \hat{\rho}(D,j)$ for any $D\in\mathcal{D}$, $j\in D$. The choice probability generated by this set of parameters yields a perfect fit of the target probability. Hence this set of parameters maximizes the likelihood.[50]

A.2 Bounding the Number of Mixtures in the EM Algorithm

This section provides a theoretical result that upper-bounds the number of mixtures required for best possible approximation. Results in this section stipulate the number of mixtures to be used in the EM algorithm.

The first proposition (Proposition 4) implies that any random-coefficient ARUMs can be represented as a finite mixture of ARUMs. The second proposition (Proposition 5) gives us an upper bound on the number of mixtures required by calculating the dimension of the set of the random utility models. To prove Proposition 4, we need the following lemma:

Lemma 11. *Let K be a fixed integer. For any bounded Borel set $C\subset\mathbf{R}^K$, let $\Delta(C)$ denote the set of Borel probability measures over C.*[51] *Then, $co.C = \left\{\int xdm(x) | m\in\Delta(C)\right\}$, where $\int xdm(x)$ denotes K-dimensional vector whose l-th element is $\int x(l)dm(x)$ for any $l\in\{1,\dots,K\}$.*

The proof is in the next section.[52] Recall the definition of $\mathcal{P}_{ra}(\eta|\mu)$ and $\mathcal{P}_a(\eta|\mu)$ from

[50]Recall that for any other stochastic choice function ρ, the likelihood $\sum_{D\in\mathcal{D}}\sum_{j\in D}\hat{\rho}(D,j)\log(\rho(D,j))$. $\hat{\rho}$ maximizes the likelihood since

$$\begin{aligned} & \textstyle\sum_{D\in\mathcal{D}}\sum_{j\in D}\hat{\rho}(D,j)(\log(\hat{\rho}(D,j))-\log(\rho(D,j))) = \sum_{D\in\mathcal{D}}\sum_{j\in D}\hat{\rho}(D,j)\log\frac{\hat{\rho}(D,j)}{\rho(D,j)} \\ = & -\textstyle\sum_{D\in\mathcal{D}}\sum_{j\in D}\hat{\rho}(D,j)\log\frac{\rho(D,j)}{\hat{\rho}(D,j)} \geq -\sum_{D\in\mathcal{D}}\sum_{j\in D}\hat{\rho}(D,j)(\frac{\rho(D,j)}{\hat{\rho}(D,j)}-1) \\ = & -\textstyle\sum_{D\in\mathcal{D}}\sum_{j\in D}\rho(D,j)+\sum_{D\in\mathcal{D}}\sum_{j\in D}\hat{\rho}(D,j) = -|\mathcal{D}|+|\mathcal{D}| = 0, \end{aligned}$$

where we use the fact $-\log(x)\geq -(x-1)$ when $x\geq 0$.

[51]In particular, $m(C)=1$ for any $m\in\Delta(C)$.

[52]We note that the result is not true in an infinite dimensional space. To see this, let $\{e_i\}_{i=1}^\infty$ be the base of

Definition 3. By using Lemma 11, we can show the following results that we use in the main body of the paper.

Proposition 4. *For any* (η, μ), $\mathcal{P}_{ra}(\eta|\mu) = co.\mathcal{P}_a(\eta|\mu) = \left\{ \sum_{m=1}^{\dim \mathcal{P}_r + 1} \lambda_m \rho_m \middle| \rho_m \in \mathcal{P}_a(\eta|\mu), \lambda_m \geq 0, \forall m = 1, \ldots, \dim \mathcal{P}_r + 1, \sum_{m=1}^{\dim \mathcal{P}_r+1} \lambda_m = 1 \right\}$.

Proof. The second equality follows from the Caratheodory's theorem. To show the first equality, fix (η, μ). Since $\text{co}.\mathcal{P}_a(\eta|\mu) \subset \mathcal{P}_{ra}(\eta|\mu)$ because a discrete probability measure is a Borel probability measure.

To show $\mathcal{P}_{ra}(\eta|\mu) \subset \text{co}.\mathcal{P}_a(\eta|\mu)$, let $\mathcal{P}_a(\eta|\mu) = C$ in Lemma 11, then we have $\text{co}.\mathcal{P}_a(\eta|\mu) = \left\{ \int \rho dm(\rho) | m \in \Delta(\mathcal{P}_a(\eta|\mu)) \right\}$.[53] Thus, it suffices to show $\mathcal{P}_{ra}(\eta|\mu) \subset \left\{ \int \rho dm(\rho) | m \in \Delta(\mathcal{P}_a(\eta|\mu)) \right\}$. To prove this define a mapping $F : \mathbf{R}^K \to \mathcal{P}_a(\eta|\mu) \subset \mathbf{R}^{|\mathcal{D}| \times J}$ by $F(\beta)(D, j) = \mu(\varepsilon | \beta \cdot x_j + \eta_j + \varepsilon_j > \beta \cdot x_l + \eta_l + \varepsilon_l \forall l \in D \setminus j)$. The mapping is continuous by the dominated convergence theorem given the fact that μ is absolutely continuous with respect to the Lesgesgue measure. Fix $m \in \Delta(\mathbf{R}^K)$ to show $\int \rho_\beta dm(\beta) \in \left\{ \int \rho dm(\rho) | m \in \Delta(\mathcal{P}_a(\eta|\mu)) \right\}$, where $\rho_\beta \in \mathcal{P}_a(\eta|\mu)$. For any every Borel set $C \subset \mathcal{P}_a(\eta|\mu)$, define $\hat{m}(C) = m(F^{-1}(C))$, where $F^{-1}(C)$ is a Borel set because F is continuous. Then, $\hat{m} \in \Delta(\mathcal{P}_a(\eta|\mu))$; moreover, we have $\int \rho_\beta dm(\beta) = \int \rho d\hat{m}(\rho)$, as desired. $\square$

We now calculate the number $\dim \mathcal{P}_r$. We note that the set $\mathcal{P}_r$ defined in Definition 1 is associated with a set of choice sets $\mathcal{D}$.

Proposition 5. $\dim \mathcal{P}_r = \sum_{D \in \mathcal{D}} (|D| - 1)$.

The proof is in a latter section. Propositions 4 and 5 imply that in order to obtain the best approximating random-coefficient model to the observed choice probabilities, it is sufficient to consider finite mixture models with at most $1 + \sum_{D \in \mathcal{D}} (|D| - 1)$ mixtures. For example, in Section 6, we analyze a choice data with $|J| = 4$ with $\mathcal{D} = 2^J \setminus \emptyset$. These results imply that it is enough to consider the finite mixture models with at most 18 mixtures if one considers the whole choice sets.

---

the infinite dimensional real space. Define $C = \{e_i\}_{i=1}^{\infty}$. Define a measure m on C such that $m(e_i) = (1/2)^i$ for each i. Then, $\sum_{i=1}^{\infty} m(e_i) = 1$, so that m is a probability measure on C. $\int x dm$ cannot be represented as any finite mixture of elements of C. For any $y \in \text{co}.C$, there exists i such that $y(e_i) = 0$.

[53] To see that $\mathcal{P}_a(\eta|\mu)$ is a Borel set, fix a standard distribution μ. The set $\mathcal{P}_a(\eta|\mu)$ is the image of the continuous mapping F from $\mathbf{R}^K$ to $\mathbf{R}^{|\mathcal{D}| \times J}$ defined in the proof. In particular, $\mathcal{P}_a(\eta|\mu)$ is the countable union of closed images of continuous mapping of compact cubes. Thus $\mathcal{P}_a(\eta|\mu)$ is a Borel set.

A.2.1 Proof of Lemma 11

By definition, co.C is a subset of $\{\int xdm(x)|m \in \Delta(C)\}$. To show the reverse direction, we first establish a relaxed statement: $\left\{\int xdm(x)|m \in \Delta(C)\right\} \subset$ cl.co.C. Suppose by way of contradiction that $\int xdm(x) \notin$ cl.co.C for some $m \in \Delta(C)$. By the separating hyperplane theorem (Corollary 11.4.2 of Rockafellar (2015)), there exist a $t \in \mathbf{R}^K \setminus \{0\}$ and $\alpha \in \mathbf{R}$ such that $(\int xdm(x)) \cdot t = \alpha > x \cdot t$ for any $x \in$ cl.co.C. This is a contradiction because $\alpha = (\int xdm(x)) \cdot t = \int (x \cdot t)dm(x) < \int \alpha dm(x) = \alpha$.

We now prove $\left\{\int xdm(x)|m \in \Delta(C)\right\} \subset$ co.C by an induction on the dimension of co.C.

Induction Base: If $\dim$ co.$C = 1$, there exist $y = \inf\{x|x \in$ co.$C\}$ and $z = \sup\{x|x \in$ co.$C\}$ such that co.C can be represented as the line segment between y and z. We consider the case where the line segment does not contain both y and z. Proofs for the other cases are similar. For any $x \in C$, there exists a weighting function $\alpha(x) = \frac{z-x}{z-y} \in (0,1)$ such that $x = \alpha(x)y + (1-\alpha(x))z$. The function α is bounded, nonegative, and continuous in x. Hence it is measurable and integrable.

Choose any $m \in \Delta(C)$. By the monotone convergence theorem, there exists l and u such that $y < l < u < z$ and $m(\{x|x \in (l,u)\}) \geq 1 - \varepsilon$ for some $\varepsilon < 1$.[54] Note that $\alpha(x)$ is uniformly bounded away from 0 and 1 for $x \in (l,u)$. Then it follows that $\int \alpha dm = \int \alpha(x)dm(x)$ exists and $0 < \int \alpha(x)dm(x) < 1$. Then, $\int xdm(x) = \int \alpha(x)dm(x) \times y + \int(1-\alpha(x))dm(x) \times z \in$ co.C, as desired.

Let $l \geq 2$ be an integer.

Induction Step: Suppose that $\left\{\int xdm(x)|m \in \Delta(C)\right\} \subset$ co.C holds for any C such that $\dim C \leq l$. Then it holds for any C such that $\dim C = l+1$.

To prove the step, choose any $m \in \Delta(C)$. We have $\int xdm(x) \in$ cl.co.C. First consider the case where $\int xdm(x) \in$ rint.cl.co.C. Since rint.cl.co.C = rint.co.C (by Theorem 6.3 of Rockafellar (2015)), we have $\int xdm(x) \in$ co.C, as desired.

Next consider the case where $\int xdm(x) \notin$ rint.cl.co.C. Then, $\int xdm(x) \in \partial$cl.co.C and by the supporting hyperplane theorem (Corollary 11.4.2 of Rockafellar (2015)), there exists a supporting hyperplane H of cl.co.C at $\int xdm(x)$. There exist $t \in \mathbf{R}^K \setminus \{0\}$ and

[54]Note here we use the fact that $m(C) = 1$ and $m(\{y\}) = m(\{z\}) = 0$

$\alpha \in \mathbf{R}$ such that $H = \{x | x \cdot t = \alpha\}$ and $\int x dm(x) \cdot t = \alpha > x \cdot t$ for any $x \in \text{cl.co.}C \cap H^c$. This implies that $m(H) = 1$ and hence $m(H \cap C) = 1$. Since H is a supporting hyperplane and cl.co.$C \not\subset H$, we obtain $\dim(H \cap \text{aff.}C) \leq l$. Hence, $\dim(H \cap C) \leq l$. Therefore, the induction hypothesis shows that $\int x dm(x) \in \text{co.}(H \cap C) \subset \text{co.}C$, as desired.

A.2.2 Proof of Proposition 5

To prove Proposition 5, we prove two lemmas. Lemma 12 is a technical lemma that facilitates the characterization of the affine hull of random utility polytopes in Lemma 13. Dimension of a set is defined as the dimension of its affine hull, and Proposition 5 follows.

Lemma 12. *For any $t \in \mathbf{R}^{|\mathcal{D}| \times |J|}$, $\rho^\pi \cdot t = \rho^{\pi'} \cdot t$ for all $\pi, \pi' \in \Pi$ if and only if $t(D, j) = t(D, l)$ for all $D \in \mathcal{D}$ and $j, l \in D$.*[55]

Proof. To prove the if part, assume $t(D, j) = t(D, l)$ for all $D \in \mathcal{D}$ and $j, l \in D$. Define $t(D) = t(D, j)$ for any $j \in D$. Then for any $\pi \in \Pi$, $\rho^\pi \cdot t = \sum_{D \in \mathcal{D}} \sum_{j \in D} \rho^\pi(D, j) t(D, j) = \sum_{D \in \mathcal{D}} t(D) \sum_{j \in D} \rho^\pi(D, j) = \sum_{D \in \mathcal{D}} t(D)$, completing the proof of the if part.

The only-if direction is trivially true when the set D contains only one element. We prove the only-if direction for the remaining D, $|D| \geq 2$, by induction. Consider the sets in $\mathcal{D}$ with a size greater or equal to 2. Let m be the smallest cardinality of the sets: $m \equiv \min\{|D| \big| D \in \mathcal{D}, |D| \geq 2\}$.[56]

Induction Base: For any $D \in \mathcal{D}$ such that $|D| = m$ and any $j, l \in D$, $t(D, j) = t(D, l)$.

Proof. To prove the claim, choose any $\pi, \pi' \in \Pi$ such that for $\pi(J \setminus D) > \pi(j) > \pi(l) > \pi(D \setminus \{j, l\})$, and $\pi'(J \setminus D) > \pi'(l) > \pi'(j) > \pi'(D \setminus \{j, l\})$

Note that the choice sets that have different choice probabilities under π and π' are those that contain both j and l, and have j dominate all elements in the set under π. The set of those sets can be written as $\mathcal{E} = \{E \big| E \in \mathcal{D}, \{j, l\} \subset E, \pi(j) \geq \pi(E)\}$. In particular for each $E \in \mathcal{E}$, $\rho^\pi(E, j) = 1$, $\rho^\pi(E, r) = 0$ for any $r \in E \setminus \{j\}$ and $\rho^{\pi'}(E, l) = 1$, $\rho^{\pi'}(E, r) = 0$ for any $r \in E \setminus \{l\}$.

[55] We are identifying each $\rho \in \mathcal{P}$ as an element of $\mathbf{R}^{\mathcal{D} \times J}$.

[56] Note we do not induct from $|D| = 1$ because $\mathcal{D}$ may not contain any set with one element.

The choice probabilities of other choice sets are the same for both π and π' because either only one of j or l is in the set, or there exists a element that dominates j and l under both π and π'.

Since $t \cdot \rho^{\pi} = t \cdot \rho^{\pi'}$,

$$0 = \sum_{(E,r)\in\mathcal{D}\times J} t(E,r)(\rho^{\pi}(E,r) - \rho^{\pi'}(E,r)) = \sum_{(E,r)\in\mathcal{E}\times J} t(E,r)((\rho^{\pi}(E,r) - \rho^{\pi'}(E,r))$$
$$= \sum_{E\in\mathcal{E}}(t(E,j) - t(E,l)) = t(D,j) - t(D,l),$$

where the last equality holds because if E contains both j and l, $\pi(j) \geq \pi(E)$, and has cardinality bigger and equal to $|D|$, then E must be equal to D. So $t(D,j) = t(D,l)$ and this completes the proof of the induction base. □

Let $k \geq m$.

Induction Step: Suppose that for any $D \in \mathcal{D}$ such that $|D| \leq k$ and any $j, l \in D$, $t(D,j) = t(D,l)$. Then the same claim holds for any $D \in \mathcal{D}$ such that $|D| = k+1$.

Proof. By the induction hypothesis, for any $E \in \mathcal{D}$, if $|E| \leq k$ then $t(E,j) = t(E,l)$ for any $j, l \in E$. Repeat the proof above by considering the choice sets $\mathcal{D} \setminus \{E | E \in \mathcal{D}, |E| \leq k\}$. □

■

Lemma 13. *The affine hull of $\mathcal{P}_r$ is $\mathcal{P}_{\pm} \equiv \big\{q \in \mathbf{R}^{|\mathcal{D}|\times|J|} \big| (i) \sum_{j\in D} q(D,j) = 1 \forall D \in \mathcal{D}$; (ii) $q(D,j) = 0\ \forall j \notin D \in \mathcal{D}\big\}$.*

Proof. The set $\mathcal{P}_{\pm}$ is affine. So it suffices to show that for any affine set A, if $\mathcal{P}_r \subset A$, then $\mathcal{P}_{\pm} \subset A$. For any affine set A, by Theorem 1.4 of Rockafellar (2015), it has a representation $A = \{q \in \mathbf{R}^{|\mathcal{D}|\times|J|} | Bq = b\}$, where B is a $L \times (|\mathcal{D}| \times |J|)$ matrix, b is a L-dimensional vector, and L is an arbitrary positive integer.

For any $l \in \{1, \ldots, L\}$, let $B_l(D,j)$ denote the $(l,(D,j))$ entry of B. Note that each column of B is associated with a $(D,j) \in \mathcal{D} \times J$. So $Bq = b$ means that for any row index $l \in \{1, \ldots, L\}$,

$$\sum_{D\in\mathcal{D}} \sum_{j\in J} B_l(D,j) q(D,j) = b_l. \tag{8}$$

By assuming $\mathcal{P}_r \subset A = \{q \in \mathbf{R}^{\mathcal{D}\times J} | Bq = b\}$, we will show that if q satisfies (i) and (ii), then (8) holds for any $l \in \{1, \ldots, L\}$.

Step 1: We show that $B_l(D, j) = B_l(D, r)$ for any $l \in \{1, \ldots, L\}$, $D \in \mathcal{D}$, and $j, r \in D$. For any $\pi \in \Pi$, $\rho^\pi \in \mathcal{P}_r \subset A = \{q \in \mathbf{R}^{\mathcal{D}\times J} | Bq = b\}$. Hence, (8) holds with $q = \rho^\pi$ for any $\pi \in \Pi$. Thus $\rho^\pi \cdot B_l = \rho^{\pi'} \cdot B_l$ for any $\pi, \pi' \in \Pi$ for any l. By Lemma 12, this implies that $B_l(D, j) = B_l(D, r)$ for any $D \in \mathcal{D}$, and $j, r \in D$.

By Step 1, we can define $B_l(D) = B_l(D, j)$ for any $j \in D$.

Step 2: If q satisfies property (i) and (ii), choose any $\pi \in \Pi$ and $l \in \{1, \ldots, L\}$. Since $\rho^\pi \in A$, then by (8),

$$b_l = \sum_{D\in\mathcal{D}} \sum_{r\in J} B_l(D, r)\rho^\pi(D, r) = \sum_{D\in\mathcal{D}} B_l(D), \tag{9}$$

where the second equality holds by $\rho^\pi(D, r) = 1$ if $\pi(r) \geq \pi(D)$ and $\rho^\pi(D, r) = 0$ otherwise. Finally, by using these equalities, for each $l \in \{1, \ldots, L\}$, we obtain the following equations:

$$\begin{aligned}
\textstyle\sum_{D\in\mathcal{D}} \sum_{r\in J} B_l(D, r)q(D, r) &= \textstyle\sum_{D\in\mathcal{D}} \sum_{r\in D} B_l(D, r)q(D, r) && \text{(by property (ii))} \\
&= \textstyle\sum_{D\in\mathcal{D}} \sum_{r\in D} B_l(D)q(D, r) && \text{(by Step 1)} \\
&= \textstyle\sum_{D\in\mathcal{D}} B_l(D) \sum_{r\in D} q(D, r) && \\
&= \textstyle\sum_{D\in\mathcal{D}} B_l(D) = b_l. && \text{(by property (i) and (9))}
\end{aligned}$$

This establishes that aff.$\mathcal{P}_r = \{q \in \mathbf{R}^{\mathcal{D}\times J} |\text{(i) and (ii)}\}$.

The equalities in (i) and (ii) are independent. The dimension of $\{q \in \mathbf{R}^{\mathcal{D}\times J} |\text{(ii)}\}$ is $\sum_{D\in\mathcal{D}} |D|$. The number of equalities of (i) is $|\mathcal{D}|$. Hence, the dimension of $\mathcal{P}_r$ is $(\sum_{D\in\mathcal{D}} |D|) - |\mathcal{D}| = \sum_{D\in\mathcal{D}}(|D| - 1)$.

■

Proposition 5 follows from Lemma 12 and Lemma 13.

A.3 Greedy Algorithm

It is known that the EM algorithm may not converge to the global optimum. To alleviate this concern, we propose a second greedy algorithm inspired by Barron et al. (2008).

This algorithm has the useful feature that, given the setup of our problem, it solves a sequence of optimization problems to converge to the global optimal solution (up to small approximation errors which can be made arbitrarily small by increasing the number of steps). The structure of the random-coefficient RUMs is important for the proof.

The algorithm takes a stochastic choice function $\hat{\rho}$ and a fixed effects vector η as input and returns a solution to (5). The algorithm is iterative: each step seeks to optimize based on the results of previous steps:

- **Step 1**: Given $\hat{\rho}$, choose ρ^1 such that $\rho^1 = \arg\inf_{\rho \in \text{cl.}\mathcal{P}_a(\eta|\mu)} ||\hat{\rho} - \rho||^2$.
- **Step n**, $n \geq 2$:
  - Consider a set of grids $\alpha_n = \{\frac{2}{i+1}\}_{i=1}^n$.
  - Find $(\alpha_n^*, \rho_n^*) = \arg\inf_{(\alpha,\rho) \in \alpha_n \times \text{cl.}\mathcal{P}_a(\eta|\mu)} ||\hat{\rho} - (1-\alpha)\rho^{n-1} - \alpha\rho||^2$.
  - Define $\rho^n = (1-\alpha_n^*)\rho^{n-1} + \alpha_n^*\rho_n^*$ and let $\rho^{out} = \rho^n$.
- Stop if a terminating criterion is reached.
- Return ρ^{out} at the final step.

The next proposition shows that the algorithm will converge for our problems in Section 5. Define $d(\rho, \hat{\rho})$ as in (5).

Proposition 6. *Let $\hat{\rho} \in \mathcal{P}$ be any stochastic function, $\eta \in \mathbf{R}^{|J|}$, and $\mu \in \mathcal{M}$. Define $d^* = \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\rho, \hat{\rho})$. Let n denote the number of steps and ρ^n denote the output after the completion of the n-th step of the algorithm. Then,*

$$d(\rho^n, \hat{\rho}) - d^* \leq \sqrt{\frac{8}{n+1}}. \tag{10}$$

For our implementation, the terminating criterion is when the number of steps taken reaches 1000. With 1000 steps, (10) implies the margin of error is within 0.09.[57] When we approximate $\hat{\rho}$ without fixed effects, we let $\eta = 0$. When we approximate $\hat{\rho}$ with fixed effects, we couple the algorithm with a grid of fixed effects to search for the minimum.

[57]This is calculated by noting $\sqrt{8/1001} \approx 0.09$.

Proof. Since the set $\text{cl.co.}\mathcal{P}_a(\eta|\mu)$ is compact and convex, $\rho^* = \arg\inf_{\rho\in\text{cl.co.}\mathcal{P}_a(\eta|\mu)} d(\rho,\hat{\rho}) = \arg\inf_{\rho\in\text{cl.co.}\mathcal{P}_a(\eta|\mu)} ||\rho-\hat{\rho}||_2^2$ exists and it can be written as a convex combination of elements of $\text{cl.}\mathcal{P}_a(\eta|\mu)$. By Caratheodory's theorem, it can be written as $\rho^* = \sum_{i=1}^M \lambda_i\rho_i$, where $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$ and $\rho_i \in \text{cl.}\mathcal{P}_a(\eta|\mu)$, $M = \dim \mathcal{P}_r + 1$. For each step n, define $E_n = \|\hat{\rho}-\rho^n\|^2 - \|\hat{\rho}-\rho^*\|^2$. For each step n, let α_n^* and ρ_n^* be the minimizers over the grids $\{\alpha_n\}$ and $\text{cl.}\mathcal{P}_a(\eta|\mu)$, respectively. Define $C = \sum_i \lambda_i\|\rho_i-\rho^*\|^2$ and $T = \max\{2E_1, 4C\}$. Note that E_1 can be upper bounded by $2||\hat{\rho}||^2 + 2||\rho^n||^2 \leq 4|\mathcal{D}|$ and similarly $C \leq 2|\mathcal{D}|$. Thus we can chooce $T = 8|\mathcal{D}|$.

Then, for each step n and let $\alpha_n = \frac{2}{n+1}$, we have the following:

$$\begin{aligned}
E_n &= \|\hat{\rho}-(1-\alpha_n^*)\rho^{n-1}-\alpha_n^*\rho_n^*\|^2 - \|\hat{\rho}-\rho^*\|^2 \\
&\leq \textstyle\sum_i \lambda_i\|\hat{\rho}-(1-\alpha_n)\rho^{n-1}-\alpha_n\rho_i\|^2 - \|\hat{\rho}-\rho^*\|^2 \\
&= \textstyle\sum_i \lambda_i\{(1-\alpha_n)^2\|\hat{\rho}-\rho^{n-1}\|^2 + 2\alpha_n(1-\alpha_n)\big((\hat{\rho}-\rho^{n-1})\cdot(\hat{\rho}-\rho_i)\big) + \alpha_n^2\|\hat{\rho}-\rho_i\|^2\} \\
&\quad -\|\hat{\rho}-\rho^*\|^2 \\
&= (1-\alpha_n)^2\|\hat{\rho}-\rho^{n-1}\|^2 + 2\alpha_n(1-\alpha_n)\big((\hat{\rho}-\rho^{n-1})\cdot(\hat{\rho}-\textstyle\sum_i\lambda_i\rho_i)\big) \\
&\quad +\alpha_n^2\textstyle\sum_i\lambda_i\|\hat{\rho}-\rho_i\|^2 - \|\hat{\rho}-\rho_*\|^2 \\
&\leq (1-\alpha_n)^2\|\hat{\rho}-\rho^{n-1}\|^2 + \alpha_n(1-\alpha_n)(\|\hat{\rho}-\rho^{n-1}\|^2 + \|\hat{\rho}-\rho^*\|^2) - \|\hat{\rho}-\rho^*\|^2 \\
&\quad +\alpha_n^2\textstyle\sum_i\lambda_i\|\hat{\rho}-\rho_i\|^2 \\
&\leq (1-\alpha_n)^2\|\hat{\rho}-\rho^{n-1}\|^2 + \alpha_n(1-\alpha_n)(\|\hat{\rho}-\rho^{n-1}\|^2 + \|\hat{\rho}-\rho^*\|^2) - \|\hat{\rho}-\rho^*\|^2 \\
&\quad +\alpha_n^2\textstyle\sum_i\lambda_i\|\rho_i-\rho^*\|^2 + \alpha_n^2\|\rho^*-\hat{\rho}\|^2 \\
&= (1-\alpha_n)\|\hat{\rho}-\rho^{n-1}\|^2 - (1-\alpha_n)\|\hat{\rho}-\rho^*\|^2 + \alpha_n^2\textstyle\sum_i\lambda_i\|\rho_i-\rho^*\|^2 \\
&= (1-\alpha_n)E_{n-1} + \alpha_n^2\textstyle\sum_i\lambda_i\|\rho_i-\rho^*\|^2.
\end{aligned}$$

Thus we have

$$E_n \leq (1-\alpha_n)E_{n-1} + C\alpha_n^2. \tag{11}$$

In the following, we will show $E_n \leq \frac{T}{n+1}$ for each n. We prove this by induction. The inequality holds with $n=1$. Fix n. Suppose $E_{n-1} \leq \frac{T}{n}$. By substituting $\alpha_n = \frac{2}{n+1}$ to (11), we have (i): $E_n \leq \frac{T}{n+1}$.[58] Let $d^n = d(\hat{\rho},\rho^n)$ and $d^* = d(\hat{\rho},\rho^*)$. Since $E_n = |\mathcal{D}|((d^n)^2-(d^*)^2)$, we have (ii): $(d^n)^2-(d^*)^2 \leq \frac{T'}{n+1}$, where $T' = \frac{T}{|\mathcal{D}|} = 8$. Then we have

[58] $E_n \leq \dfrac{n-1}{n+1}\dfrac{T}{n} + C\dfrac{4}{(n+1)^2} = \dfrac{(n^2-1)T+4Cn}{(n+1)^2 n} \leq \dfrac{(n^2-1)T+Tn}{(n+1)^2 n} \leq \dfrac{n^2T+Tn}{(n+1)^2 n} = \dfrac{Tn(n+1)}{(n+1)^2 n} = \dfrac{T}{n+1}$.

$(d^n - d^*)^2 \leq (d^n - d^*)(d^n - d^*) + (d^n - d^*)2d^* = (d^n - d^*)(d^n + d^*) \leq \frac{8}{n+1}$, where we use the fact that $d^n \geq d^*$ and $d^* \geq 0$. This implies $d^n - d^* \leq \sqrt{\frac{8}{n+1}}$.[59] ■

A.4 Calculating the Maximal Substitution in (6)

To calculate the maximal substitution (6), we consider the problem

$$\inf_{\rho \in \mathcal{P}_{ra}(0|\mu)} \left(\sum_{r \in J \setminus \{j,l\}} \rho(J \setminus \{j\}, r) + \rho(J, l) \right)^2, \tag{12}$$

which can be readily solved by the greedy algorithm.[60] Taking 1 minus the squared root of the minimized value in (12) gives the solution to the problem in (6). To see this, notice

$$\sup_{\rho \in \mathcal{P}_{ra}(0|\mu)} (\rho(J \setminus \{j\}, l) - \rho(J, l)) = 1 - \inf_{\rho \in \mathcal{P}_{ra}(0|\mu)} \left(\sum_{r \in J \setminus \{j,l\}} \rho(J \setminus \{j\}, r) + \rho(J, l) \right).$$

Because $\sum_{r \in J \setminus \{j,l\}} \rho(J \setminus \{j\}, r) + \rho(J, l)$ is nonnegative, minimizer(s) of $\sum_{r \in J \setminus \{j,l\}} \rho(J \setminus \{j\}, r) + \rho(J, l)$ is the same as the minimizer(s) of the problem when the criterion is squared. It can also be shown that the greedy algorithm converges to the global optimal solutions after solving a sequence of optimization problems.

We modify the greedy algorithm for maximal substitution as follows:

- **Step 1**: Choose ρ^1 as a solution of $\inf_{\rho \in \mathcal{P}_{ra}(0|\mu)} \left(\sum_{r \in J \setminus \{j,l\}} \rho(J \setminus \{j\}, r) + \rho(J, l) \right)^2$.
- **Step n**, $n \geq 2$:
  - Consider a set of grids $\alpha_n = \{\frac{2}{i+1}\}_{i=1}^n$.
  - Find (α_n^*, ρ_n^*) as a solution of

$$\inf_{(\alpha,\rho) \in \alpha_n \times \text{cl.}\mathcal{P}_{ml}(0)} \left((1-\alpha)\Big(\sum_{r \in J \setminus \{j,l\}} \rho^{n-1}(J \setminus \{j\}, r) + \rho^{n-1}(J, r) \Big) + \alpha \Big(\sum_{r \in J \setminus \{j,l\}} \rho(J \setminus \{j\}, r) + \rho(J, r) \Big) \right)^2.$$

  - Define $\rho^n = (1 - \alpha_n^*)\rho^{n-1} + \alpha_n^* \rho_n^*$ and let $\rho^{out} = \rho^n$.

[59] We comment that we can upper bound E_1 and C by the squared diameter of the random utility polytope. For example $C \leq \sum_i \lambda_i ||\rho_i - \rho^*||^2 \leq \sup_{\rho, \hat{\rho} \in \mathcal{P}_r} ||\rho - \hat{\rho}||_2^2 = 2 \times$ the number of choice sets. The extremum is achieved by selecting x to be a degenerate ranking and y its reverse ranking. Similarly we can bound E_1. Notice this implies $T = 8 \times$ (the number of choice sets).

[60] Note that the objective function can be viewed as a distance metric.

- Stop if a terminating criterion is reached.
- Return ρ^{out} at the final step.

B Additional Empirical Results

B.1 In-sample and Out-of-sample Fit

In this section, we evaluate in-sample and out-of-sample fit of our model. We show that our method performs better or equally well compared to standard methods, not only in terms of in-sample fit but also in terms of out-of-sample fit. We use the same fishing choice dataset used in Section 6 and predict choice probabilities using aggregated characteristics.

We estimate a random-coefficient logit model with arbitrary mixing distributions. In the dataset, we have four alternatives and we consider only one choice set $\mathcal{D} = \{J\}$. Thus by Proposition 5 and Corollary 4, it suffices to mix four logit models without fixed effects to represent any random utility model. We refer to a 4-mixture mixed-logit model as *our method* hereafter. We also estimate several standard models for comparison. They include 1) a multinomial logit model, 2) nested logit model with two nests (charter and the rest), 3) a nested logit model with two nests (boat and the rest), 4) a random coefficient logit model with a log-normal mixing distribution for each variable, 5) a multinomial logit model with alternative fixed effects, 6) and a random coefficient logit model with log-normal mixing distributions and alternative fixed effects. We detail the definition of each specification in Section B.2.

To evaluate in-sample and out-of-sample fits, we adopt the following strategy. We randomly divide individuals in the sample into a training sample and a test sample of equal sizes. Separately for the training and testing samples, we average individual choices and characteristics to obtain aggregate data on choice probabilities and characteristics. We then estimate the models using the training sample. The models are estimated by maximizing the log-likelihoods. That is, for each model, we solve the problem $\max_{\theta\in\Theta} \sum_{j=1}^{|J|} \hat{\rho}_j \log \rho(j|\theta)$, where j indexes fishing modes, θ is the parameter vector of the model, Θ denotes the set of possible parameter vectors, $\hat{\rho}_j$ is the observed market share for fishing mode j in the training data, and $\rho(j|\theta)$ is the model-predicted choice probability for fishing mode j with characteristic vector x_j. See Section B.2 for a likelihood expression for each model. For the standard models, we maximize the likelihoods with the nonlinear optimization package in R (Ghalanos and Theussl, 2015; Ye, 1987). For our model, we use the EM algorithm.[61]

To evaluate the in-sample fit performance, we compute the predicted choice probabilities in the training sample $\hat{\rho}_{train} \in \mathbf{R}^{|J|}$ and compare it with the observed choice probabilities in the training sample $\rho_{train} \in \mathbf{R}^{|J|}$.[62] For this comparison, we calculate the l_2 distance between the predicted choice probabilities and the aggregated observed choice probabilities $||\hat{\rho}_{train} - \rho_{train}||_2$. Similarly, to evaluate the out-of-sample performance, we compute the predicted choice probabilities using the testing sample $\hat{\rho}_{test} \in \mathbf{R}^{|J|}$ and compare it with the aggregated observed choice probabilities in the testing sample $\rho_{test} \in \mathbf{R}^{|J|}$. We use the l_2 metric $||\hat{\rho}_{test} - \rho_{test}||_2$ for this comparison as well.

We repeat this exercise with 50 random splits. The results for in-sample fits are reported in Table A.1. The results for out-of-sample fits are in Table A.2.

As expected, the in-sample fit of our model is perfect. Several standard models, especially those without fixed effects, exhibit imperfect in-sample fit. For example, the random coefficient logit model with the log normal distributions has the l_2 prediction error 0.038.

Table A.2 shows that the out-of-sample prediction error of our model is positive but small. Standard models without alternative fixed effects have out-of-sample prediction errors substantially larger than our model. The two alternative models with fixed effects have out-of-sample prediction errors comparable to ours. This result suggests that even without using the fixed effects, our model performs better or equally well compared to standard models in this simulation, not only in terms of in-sample fit but also in terms of out-of-sample fit.

[61] We prefer the EM algorithm over the greedy algorithm here because the EM algorithm is faster.

[62] We only consider the single choice set case in this simulation. So the choice probability vector has length $|J|$.

Table A.1: In-Sample Fit

Model	Choice probabilities				Prediction error
	Beach	Boat	Charter	Pier	
(1)	(2)	(3)	(4)	(5)	(6)
Our method	0.114 (0.009)	0.353 (0.015)	0.383 (0.017)	0.151 (0.010)	0.000 (0.000)
Multinomial logit	0.141 (0.007)	0.355 (0.015)	0.378 (0.017)	0.126 (0.007)	0.038 (0.009)
Nested logit (charter and others)	0.114 (0.010)	0.353 (0.015)	0.383 (0.017)	0.150 (0.011)	0.001 (0.002)
Nested logit (boat and others)	0.141 (0.007)	0.355 (0.015)	0.378 (0.017)	0.126 (0.007)	0.038 (0.009)
Mixed logit with log normal distribution	0.142 (0.008)	0.354 (0.015)	0.378 (0.017)	0.126 (0.007)	0.038 (0.009)
Multinomial logit with fixed effects	0.113 (0.009)	0.353 (0.015)	0.383 (0.017)	0.151 (0.010)	0.000 (0.000)
Mixed logit with log normal distribution and fixed effects	0.114 (0.009)	0.353 (0.015)	0.383 (0.017)	0.151 (0.010)	0.000 (0.000)

Notes: Table A.1 summarizes the in-sample fit of different models. The row "our method" presents choice probabilities predicted by the four-mixture mixed-logit model and the prediction error. The remaining rows present in-sample predicted choice probabilities and prediction errors obtained by standard models. In parentheses are standard deviations obtained by repeating the same analyses 50 times.

B.2 Definitions of Other Models

In each of the standard models used in our empirical section, the choice probability $\rho(J, j) \equiv \rho_j$ of alternative j from J is specified as follows:

- Multinomial logit: $\rho_j = \dfrac{\exp(\beta \cdot x_j)}{\sum_{j' \in J} \exp(\beta \cdot x_j{}')}$
- Nested logit (charter and others): the choice probability of alternative j that belongs to nest J_g is specified as

$$\rho_j = \frac{\exp(\beta \cdot x_j/\lambda)}{\sum_{j' \in J_g} \exp(\beta \cdot x_j{}'/\lambda)} \times \frac{\left[\sum_{j' \in J_g} \exp(\beta \cdot x_j{}'/\lambda)\right]^{\lambda}}{\sum_{g' \in G} \left[\sum_{j' \in J_{g'}} \exp(\beta \cdot x_j{}'/\lambda)\right]^{\lambda}}.$$

  The nest is defined by the partition $G = \{\{\text{charter}\}, \{\text{beach, boat, pier}\}\}$.

- Nested logit (boat and others): the nested logit model specified above, with the nest defined by $G = \{\{\text{boat}\}, \{\text{beach, charter, pier}\}\}$.

Table A.2: Out-Sample Fit

Model	Choice probabilities				Prediction error
	Beach	Boat	Charter	Pier	
(1)	(2)	(3)	(4)	(5)	(6)
Our method	0.114 (0.009)	0.353 (0.015)	0.383 (0.017)	0.151 (0.010)	0.049 (0.017)
Multinomial logit	0.143 (0.011)	0.353 (0.017)	0.377 (0.022)	0.127 (0.011)	0.058 (0.019)
Nested logit (charter and others)	0.115 (0.012)	0.350 (0.022)	0.383 (0.018)	0.152 (0.014)	0.050 (0.020)
Nested logit (boat and others)	0.143 (0.011)	0.353 (0.017)	0.377 (0.022)	0.127 (0.011)	0.058 (0.019)
Mixed logit with log normal distribution	0.143 (0.010)	0.352 (0.016)	0.377 (0.021)	0.127 (0.010)	0.058 (0.018)
Multinomial logit with fixed effects	0.116 (0.014)	0.350 (0.019)	0.380 (0.022)	0.154 (0.019)	0.048 (0.022)
Mixed logit with log normal distribution and fixed effects	0.113 (0.008)	0.353 (0.015)	0.383 (0.018)	0.151 (0.010)	0.048 (0.017)

Notes: Table A.2 summarizes the out-sample fit of different models. The row "our method" presents choice probabilities predicted by the four-mixture mixed-logit model and the prediction error (5). The remaining rows present out-of-sample predicted choice probabilities and prediction errors obtained by standard models. In parentheses are standard deviations obtained by repeating the same analyses 50 times.

- Mixed logit: $\rho_j = \int \frac{\exp(\beta \cdot x_j)}{\sum_{j' \in J} \exp(\beta \cdot x_j{}')} f(\beta) d\beta$ where f is the density of the distribution of random coefficients. We use independent log-normal distributions for each coefficient. To evaluate the integral, we randomly draw 100 realizations from the random coefficient distribution.
- Multinomial logit with fixed effects: the above multinomial logit model with x including dummies for each alternative (except for beach).
- Mixed logit with fixed effects: the random coefficient logit model with log normal distributions. We also include fixed effects for each alternative (except for beach). To evaluate the integral, we randomly draw 100 realizations from the random coefficient distribution.

C Notation

Notation	Meaning	Def.
J	the set of alternatives	2.1
D	a generic choice set	1
$\mathcal{D}$	the set of generic choice sets	2.1
$\mathcal{M}$	the set of standard probability measures	2.2
K	the dimension of explanatory variables	2.1
$\mathcal{P}$	the set of stochastic choice functions	2.1
$\mathcal{P}_r$	the set of random utility models	2.2
ρ	a generic stochastic choice function	2.1
Π	the set of rankings	2.1
co.C	the convex hull of a set C	2.2
rint.C	the relative interior of C	2.2
cl.C	the closure of C	2.2
$\mathcal{P}_{ra}(\eta\|\mu)$	the set of random-coefficient ARUMs with fixed effects η and a probability measure μ	2.2
$\mathcal{P}_a(\eta\|\mu)$	the set of ARUMs with fixed effects η and a probability measure μ	2.2
$\mathcal{P}_{ml}(\eta)$	the set of mixed logit models with fixed effects η	2.2
$\mathcal{P}_l(\eta)$	the set of logit models with fixed effects η	2.2

Notes: The Def. column indicates the (sub)section in which the definition of the notation appears.