Pairwise Choice Markov Chains

Stephen Ragain
Management Science & Engineering
Stanford University
Stanford, CA 94305
sragain@stanford.edu

Johan Ugander
Management Science & Engineering
Stanford University
Stanford, CA 94305
jugander@stanford.edu

Abstract

As datasets capturing human choices grow in richness and scale—particularly in online domains—there is an increasing need for choice models that escape traditional choice-theoretic axioms such as regularity, stochastic transitivity, and Luce's choice axiom. In this work we introduce the Pairwise Choice Markov Chain (PCMC) model of discrete choice, an inferentially tractable model that does not assume any of the above axioms while still satisfying the foundational axiom of *uniform expansion*, a considerably weaker assumption than Luce's choice axiom. We show that the PCMC model significantly outperforms both the Multinomial Logit (MNL) model and a mixed MNL (MMNL) model in prediction tasks on both synthetic and empirical datasets known to exhibit violations of Luce's axiom. Our analysis also synthesizes several recent observations connecting the Multinomial Logit model and Markov chains; the PCMC model retains the Multinomial Logit model as a special case.

1 Introduction

Discrete choice models describe and predict decisions between distinct alternatives. Traditional applications include consumer purchasing decisions, choices of schooling or employment, and commuter choices for modes of transportation among available options. Early models of probabilistic discrete choice, including the well known Thurstone Case V model [29] and Bradley-Terry-Luce (BTL) model [7], were developed and refined under diverse strict assumptions about human decision making. As complex individual choices become increasingly mediated by engineered and learned platforms—from online shopping to web browser clicking to interactions with recommendation systems—there is a pressing need for flexible models capable of describing and predicting nuanced choice behavior.

Luce's choice axiom, popularly known as the *independence of irrelevant alternatives* (IIA), is arguably the most storied assumption in choice theory [20]. The axiom consists of two statements, applied to each subset of alternatives S within a broader universe U. Let $p_{aS} = \Pr(a \text{ chosen from } S)$ for any $S \subseteq U$, and in a slight abuse of notation let $p_{ab} = \Pr(a \text{ chosen from } \{a, b\})$ when there are only two elements. Luce's axiom is then that: (i) if $p_{ab} = 0$ then $p_{aS} = 0$ for all S containing a and b, (ii) the probability of choosing a from U conditioned on the choice lying in S is equal to p_{aS}.

The BTL model, which defines $p_{ab} = \gamma_a/(\gamma_a + \gamma_b)$ for latent "quality" parameters $\gamma_i > 0$, satisfies the axiom while Thurstone's Case V model does not [1]. Soon after its introduction, the BTL model was generalized from pairwise choices to choices from larger sets [4]. The resulting Multinomal Logit (MNL) model again employs quality parameters $\gamma_i \geq 0$ for each $i \in U$ and defines p_{iS}, the probability of choosing i from $S \subseteq U$, proportional to γ_i for all $i \in S$. Any model that satisfies Luce's choice axiom is equivalent to some MNL model [21].

One consequence of Luce's choice axiom is *strict stochastic transitivity* between alternatives: if $p_{ab} \geq 0.5$ and $p_{bc} \geq 0.5$, then $p_{ac} \geq \max(p_{ab}, p_{bc})$. A possibly undesirable consequence of strict stochastic transitivity is the necessity of a total order across all elements. But note that strict

stochastic transitivity does not imply the choice axiom; Thurstone's model exhibits strict stochastic transitivity.

Many choice theorists and empiricists, including Luce, have noted that the choice axiom and stochastic transitivity are strong assumptions that do not hold for empirical choice data [10, 14, 15, 28, 30]. A range of discrete choice models striving to escape the confines of the choice axiom have emerged over the years. The most popular of these models have been Elimination by Aspects [31], mixed MNL (MMNL) [6], and nested MNL [24]. Inference is practically difficult for all three of these models [17, 25]. Additionally, Elimination by Aspects and the MMNL model also both exhibit the rigid property of regularity, defined below.

A broad, important class of models in the study of discrete choice is the class of *random utility models* (RUMs) [4, 22]. A RUM affiliates with each $i \in U$ a random variable X_i and defines for each subset $S \subseteq U$ the probability $\Pr(i \text{ chosen from } S) = \Pr(X_i \geq X_j, \forall j \in S)$. An *independent RUM* has independent X_i. RUMs assume neither choice axiom nor stochastic transitivity. Thurstone's Case V model and the BTL model are both independent RUMs; the Elimination by Aspects and MMNL models are both RUMs. A major result by McFadden and Train establishes that for any RUM there exists a MMNL model that can approximate the choice probabilities of that RUM to within an arbitrary error [25], a strong result about the generality of MMNL models. The nested MNL model, meanwhile, is not a RUM.

Although RUMs need not exhibit stochastic transitivity, they still exhibit the weaker property of *regularity*: for any choice sets A, B where $A \subseteq B$, $p_{xA} \geq p_{xB}$. Regularity may at first seem intuitively pleasing, but it prevents models from expressing framing effects [14] and other empirical observations from modern behavior economics [30]. This rigidity motivates us to contribute a new model of discrete choice that escapes historically common assumptions while still furnishing enough structure to be inferentially tractable.

The present work. In this work we introduce a conceptually simple and inferentially tractable model of discrete choice that we call the PCMC model. The parameters of the PCMC model are the off-diagonal entries of a rate matrix Q indexed by U. The PCMC model affiliates each subset S of the alternatives with a continuous time Markov chain (CTMC) on S with transition rate matrix Q_S, whose off-diagonal entries are entries of Q indexed by pairs of items in S. The model defines p_{iS}, the selection probability of alternative $i \in S$, as the probability mass of alternative $i \in S$ of the stationary distribution of the CTMC on S.

The transition rates of these CTMCs can be interpreted as measures of preferences between pairs of alternatives. Special cases of the model use pairwise choice probabilities as transition rates, and as a result the PCMC model extends arbitrary models of pairwise choice to models of set-wise choice. Indeed, we show that when the matrix Q is parameterized with the pairwise selection probabilities of a BTL pairwise choice model, the PCMC model reduces to an MNL model. Recent parameterizations of non-transitive pairwise probabilities such as the Blade-Chest model [8] can be usefully employed to reduce the number of free parameters of the PCMC model.

Our PCMC model can be thought of as building upon the observation underlying the recently introduced Iterative Luce Spectral Ranking (I-LSR) procedure for efficiently finding the maximum likelihood estimate for parameters of MNL models [23]. The analysis of I-LSR is precisely analyzing a PCMC model in the special case where the matrix Q has been parameterized by BTL. In that case the stationary distribution of the chain is found to satisfy the stationary conditions of the MNL likelihood function, establishing a strong connection between MNL models and Markov chains. The PCMC model generalizes that connection.

Other recent connections between the MNL model and Markov chains include the work on Rank-Centrality [26], which employs a discrete time Markov chain for inference in the place of I-LSR's continuous time chain, in the special case where all data are pairwise comparisons.

Separate recent work has contributed a different discrete time Markov chain model of "choice substitution" capable of approximating any RUM [3], a related problem but one with a strong focus on ordered preferences. Lastly, recent work by Kumar et al. explores conditions under which a probability distribution over discrete items can be expressed as the stationary distribution of a discrete time Markov chain with "score" functions similar to the "quality" parameters in an MNL model [19].

The PCMC model is not a RUM, and in general does not exhibit stochastic transitivity, regularity, or the choice axiom. We find that the PCMC model does, however, obey the lesser known but fundamental axiom of *uniform expansion*, a weakened version of Luce's choice axiom proposed by Yellott that implies the choice axiom for independent RUMs [32]. In this work we define a convenient structural property termed *contractibility*, for which uniform expansion is a special case, and we show that the PCMC model exhibits contractibility. Of the models mentioned above, only Elimination by Aspects exhibits uniform expansion without being an independent RUM. Elimination by Aspects obeys regularity, which the PCMC model does not; as such, the PCMC model is uniquely positioned in the literature of axiomatic discrete choice, minimally satisfying uniform expansion without the other aforementioned axioms.

After presenting the model and its properties, we investigate choice predictions from our model on two empirical choice datasets as well as diverse synthetic datasets. The empirical choice datasets concern transportation choices made on commuting and shopping trips in San Francisco. Inference on synthetic data shows that PCMC is competitive with MNL when Luce's choice axiom holds, while PCMC outperforms MNL when the axiom does not hold. More significantly, for both of the empirical datasets we find that a learned PCMC model predicts empirical choices significantly better than a learned MNL model.

2 The PCMC model

A Pairwise Choice Markov Chain (PCMC) model defines the selection probability p_{iS}, the probability of choosing i from $S \subseteq U$, as the probability mass on alternative $i \in S$ of the stationary distribution of a continuous time Markov chain (CTMC) on the set of alternatives S. The model's parameters are the off-diagonal entries q_{ij} of rate matrix Q indexed by pairs of elements in U. See Figure 1 for a diagram. We impose the constraint $q_{ij} + q_{ji} \geq 1$ for all pairs (i, j), which ensures irreducibility of the chain for all S.
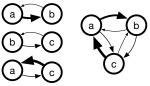

Figure 1: Markov chains on choice sets $\{a, b\}$, $\{a, c\}$, and $\{b, c\}$, where line thicknesses denote transition rates. The chain on the choice set $\{a, b, c\}$ is assembled using the same rates.

Given a query set $S \subseteq U$, we construct Q_S by restricting the rows and columns of Q to elements in S and setting $q_{ii} = -\sum_{j \in S \setminus i} q_{ij}$ for each $i \in S$. Let $\pi_S = \{\pi_S(i)\}_{i \in S}$ be the stationary distribution of the corresponding CTMC on S, and let $\pi_S(A) = \sum_{x \in A} \pi_S(x)$. We define the choice probability $p_{iS} := \pi_S(i)$, and now show that the PCMC model is well defined.

Proposition 1. *The choice probabilities p_{iS} are well defined for all $i \in S$, all $S \subseteq U$ of a finite U.*

Proof. We need only to show that there is a single closed communicating class. Because S is finite, there must be at least one closed communicating class. Suppose the chain had more than one closed communicating class and that $i \in S$ and $j \in S$ were in different closed communicating classes. But $q_{ij} + q_{ji} \geq 1$, so at least one of q_{ij} and q_{ji} is strictly positive and the chain can switch communicating classes through the transition with strictly positive rate, a contradiction. $\square$

While the support of π_S is the single closed communicating class, S may have transient states corresponding to alternatives with selection probability 0. Note that irreducibility argument needs only that $q_{ij} + q_{ji}$ be positive, not necessarily at least 1 as imposed in the model definition. One could simply constrain $q_{ij} + q_{ji} \geq \epsilon$ for some positive ϵ. However, multiplying all entries of Q by some $c > 0$ does not affect the stationary distribution of the corresponding CTMC, so multiplication by $1/\epsilon$ gives a Q with the same selection probabilities.

In the subsections that follow, we develop key properties of the model. We begin by showing how assigning Q according a Bradley-Terry-Luce (BTL) pairwise model results in the PCMC model being equivalent to BTL's canonical extension, the Multinomial Logit (MNL) set-wise model. We then construct a Q for which the PCMC model is neither regular nor a RUM.

2.1 Multinomial Logit from Bradley-Terry-Luce

We now observe that the Multinomial Logit (MNL) model, also called the Plackett-Luce model, is precisely a PCMC model with a matrix Q consisting of pairwise BTL probabilities. Recall that the BTL model assumes the existence of latent "quality" parameters $\gamma_i > 0$ for $i \in U$ with $p_{ij} = \gamma_i / (\gamma_i + \gamma_j), \forall i, j \in U$ and that the MNL generalization defines $p_{iS} \propto \gamma_i, \forall i \in S$ for each $S \subseteq U$.

Proposition 2. *Let γ be the parameters of a BTL model on U. For $q_{ji} = \frac{\gamma_i}{\gamma_i + \gamma_j}$, the PCMC probabilities p_{iS} are consistent with an MNL model on S with parameters γ.*

Proof. We aim to show that $\pi_S = \frac{\gamma}{||\gamma||_1}$ is a stationary distribution of the PCMC chain: $\pi_S^T Q_S = 0$. We have:

$$(\pi_S^T Q_S)_i = \frac{1}{||\gamma||_1} \left(\sum_{j \neq i} \gamma_j q_{ji} - \gamma_i (\sum_{j \neq i} q_{ji}) \right) = \frac{\gamma_i}{||\gamma||_1} \left(\sum_{j \neq i} \frac{\gamma_j}{\gamma_i + \gamma_j} - \sum_{j \neq i} \frac{\gamma_j}{\gamma_i + \gamma_j} \right) = 0, \ \forall i.$$

Thus π_S is always the stationary distribution of the chain, and we know by Proposition 1 that it is unique. It follows that $p_{iS} \propto \gamma_i$ for all $i \in S$, as desired. $\square$

Other parameterizations of Q, which can be used for parameter reduction or to extend arbitrary models for pairwise choice, are explored section 1 of the Supplementary material.

2.2 A counterexample to regularity

The regularity property stipulates that for any $S' \subset S$, the probability of selecting a from S' is at least the probability of selecting a from S. All RUMs exhibit regularity because $S' \subseteq S$ implies $\Pr(X_i = \max_{j \in S'} X_j) \geq \Pr(X_i = \max_{j \in S} X_j)$. We now construct a simple PCMC model which does not exhibit regularity, and is thus not a RUM.

Consider $U = \{r, p, s\}$ corresponding to a rock-paper-scissors-like stochastic game where each pairwise matchup has the same win probability $\alpha > \frac{1}{2}$. Constructing a PCMC model where the transition rate from i to j is α if j beats i in rock-paper-scissors yields the rate matrix

$$Q = \begin{bmatrix} -1 & 1-\alpha & \alpha \\ \alpha & -1 & 1-\alpha \\ 1-\alpha & \alpha & -1 \end{bmatrix}.$$

We see that for pairs of objects, the PCMC model returns the same probabilities as the pairwise game, i.e. $p_{ij} = \alpha$ when i beats j in rock-paper-scissors, as $p_{ij} = q_{ji}$ when $q_{ij} + q_{ji} = 1$. Regardless of how the probability α is chosen, however, it is always the case that $p_{rU} = p_{pU} = p_{sU} = 1/3$. It follows that regularity does not hold for $\alpha > 2/3$.

We view the PCMC model's lack of regularity is a positive trait in the sense that empirical choice phenomena such as framing effects and asymmetric dominance violate regularity [14], and the PCMC model is rare in its ability to model such choices. Deriving necessary and sufficient conditions on Q for a PCMC model to be a RUM, analogous to known characterization theorems for RUMs [11] and known sufficient conditions for nested MNL models to be RUMs [5], is an interesting open challenge.

3 Properties

While we have demonstrated already that the PCMC model avoids several restrictive properties that are often inconsistent with empirical choice data, we demonstrate in this section that the PCMC model still exhibits deep structure in the form of contractibility, which implies uniform expansion. Inspired by a thought experiment that was posed as an early challenge to the choice axiom, we define the property of contractibility to handle notions of similarity between elements. We demonstrate that the PCMC model exhibits contractibility, which gracefully handles this thought experiment.

3.1 Uniform expansion

Yellott [32] introduced *uniform expansion* as a weaker condition than Luce's choice axiom, but one that implies the choice axiom in the context of any independent RUM. Yellott posed the axiom of invariance to uniform expansion in the context of "copies" of elements which are "identical." In the context of our model, such copies would have identical transition rates to alternatives:

Definition 1 (Copies). *For i, j in $S \subseteq U$, we say that i and j are copies if for all $k \in S - i - j$, $q_{ik} = q_{jk}$ and $q_{ij} = q_{ji}$.*

Yellott's introduction to uniform expansion asks the reader to consider an offer of a choice of beverage from k identical cups of coffee, k identical cups of tea, and k identical glasses of milk. Yellott contends that the probability the reader chooses a type of beverage (e.g. coffee) in this scenario should be the same as if they were only shown one cup of each beverage type, regardless of $k \geq 1$.

Definition 2 (Uniform Expansion). *Consider a choice between n elements in a set $S_1 = \{i_{11}, \ldots, i_{n1}\}$, and another choice from a set S_k containing k copies of each of the n elements: $S_k = \{i_{11}, \ldots, i_{1k}, i_{21}, \ldots, i_{2k}, \ldots, i_{n1}, \ldots, i_{nk}\}$. The axiom of uniform expansion states that for each $m = 1, \ldots, n$ and all $k \geq 1$:*

$$p_{i_{m1}S_1} = \sum_{j=1}^{k} p_{i_{mj}S_k}.$$

We will show that the PCMC model always exhibits a more general property of contractibility, of which uniform expansion is a special case; it thus always exhibits uniform expansion.

Yellott showed that for any independent RUM with $|U| \geq 3$ the double-exponential distribution family is the only family of independent distributions that exhibit uniform expansion for all $k \geq 1$, and that Thurstone's model based on the Gaussian distribution family in particular does not exhibit uniform expansion.

While uniform expansion seems natural in many discrete choice contexts, it should be regarded with some skepticism in applications that model competitions. Sports matches or races are often modeled using RUMs, where the winner of a competition can be modeled as the competitor with the best draw from their random variable. If a competitor has a performance distribution with a heavy upper tail (so that their wins come from occasional "good days"), uniform expansion would not hold. This observation relates to recent work on team performance and selection [16], where non-invariance under uniform expansion plays a key role.

3.2 Contractibility

In a book review of Luce's early work on the choice axiom, Debreu [10] considers a hypothetical choice between three musical recordings: one of Beethoven's eighth symphony conducted by X, another of Beethoven's eighth symphony conducted by Y, and one of Debussy quartet conducted by Z. We will call these options B_1, B_2, and D respectively. When compared to D, Debreu argues that B_1 and B_2 are indistinguishable in the sense that $p_{DB_1} = p_{DB_2}$. However, someone may prefer B_1 over B_2 in the sense that $p_{B_1B_2} > 0.5$. This is impossible under a BTL model, in which $p_{DB_1} = p_{DB_2}$ implies that $\gamma_{B_1} = \gamma_{B_2}$ and in turn $p_{B_1B_2} = 0.5$.

To address contexts in which elements compare identically to alternatives but not each other (e.g. B_1 and B_2), we introduce *contractible partitions* that group these similar alternatives into sets. We then show that when a PCMC model contains a contractible partition, the relative probabilities of selecting from one of these partitions is independent from how comparisons are made between alternatives in the same set. Our contractible partition definition can be viewed as akin to (but distinct from) *nests* in nested MNL models [24].

Definition 3 (Contractible Partition). *A partition of U into non-empty sets $A_1, \ldots, A_k$ is a contractible partition if $q_{a_i a_j} = \lambda_{ij}$ for all $a_i \in A_i, a_j \in A_j$ for some $\Lambda = \{\lambda_{ij}\}$ for $i, j \in \{1, \ldots, k\}$.*

Proposition 3. *For a given Λ, let $A_1, \ldots, A_k$ be a contractible partition for two PCMC models on U represented by Q, Q' with stationary distributions π, π'. Then for any A_i:*

$$\sum_{j \in A_i} p_{jU} = \sum_{j \in A_i} p'_{jU}, \tag{1}$$

or equivalently, $\pi(A_i) = \pi'(A_i)$.

Proof. Suppose Q has contractible partition $A_1, \ldots, A_k$ with respect to Λ. If we decompose the balance equations (i.e. each row of $\pi^T Q = 0$), for $x \in A_1$ WLOG we obtain:

$$\pi(x) \left(\sum_{y \in A_1 \setminus x} q_{xy} + \sum_{i=2}^{k} \sum_{a_i \in A_i} q_{xa_i} \right) = \sum_{y \in A_1 \setminus x} \pi(y) q_{yx} + \sum_{i=2}^{k} \sum_{a_i \in A_i} \pi(a_i) q_{a_i x}. \tag{2}$$

5

Noting that for $a_i \in A_i$ and $a_j \in A_j$, $q_{a_i a_j} = \lambda_{ij}$, (2) can be rewritten:

$$\pi(x)\left(\sum_{y \in A_1 \setminus x} q_{xy}\right) + \pi(x)\sum_{i=2}^{k}|A_i|\lambda_{i1} = \sum_{y \in A_1 \setminus x}\pi(y)q_{yx} + \sum_{i=2}^{k}\pi(A_i)\lambda_{i1}.$$

Summing over $x \in A_1$ then gives

$$\sum_{x \in A_1}\pi(x)\left(\sum_{y \in A_1 \setminus x} q_{xy}\right) + \pi(A_1)\sum_{i=2}^{k}|A_i|\lambda_{i1} = \sum_{x \in A_1}\sum_{y \in A_1 \setminus x}\pi(y)q_{yx} + |A_1|\sum_{i=2}^{k}\pi(A_i)\lambda_{i1}.$$

The leftmost term of each side is equal, so we have

$$\pi(A_1) = \frac{|A_1|\sum_{i=2}^{k}\pi(A_i)\lambda_{i1}}{\sum_{i=2}|A_i|\lambda_{1i}}, \tag{3}$$

which makes $\pi(A_1)$ the solution to global balance equations for a different continuous time Markov chain with the states $\{A_1, \ldots, A_k\}$ and transition rate $\tilde{q}_{A_i A_j} = |A_j|\lambda_{ij}$ between state A_i and A_j, and $\tilde{q}_{A_i A_i} = -\sum_{j \neq i}\tilde{q}_{A_i A_j}$. Now $q_{a_i a_j} + q_{a_j a_i} \geq 1$ implies $\lambda_{ij} + \lambda_{ji} \geq 1$. Combining this observation with $|A_i| > 0$ shows (as with the proof of Proposition 1) that this chain is irreducible and thus that $\{\pi(A_i)\}_{i=1}^{k}$ are well-defined. Furthermore, because $\tilde{Q}$ is determined entirely by Λ and $|A_1|, \ldots, |A_k|$, we have that $\tilde{Q} = \tilde{Q}'$, and thus that $\pi(A_i) = \pi'(A_i), \forall i$ regardless of how Q and Q' may differ, completing the proof. $\qquad\square$

The intuition is that we can "contract" each A_i to a single "type" because the probability of choosing an element of A_i is independent of the pairwise probabilities between elements within the sets. The above proposition and the contractibility of a PCMC model on all uniformly expanded sets implies that all PCMC models exhibit uniform expansion.

Proposition 4. *Any PCMC model exhibits uniform expansion.*

Proof. We translate the problem of uniform expansion into the language of contractibility. Let U_1 be the universe of unique items $i_{11}, i_{21}, \ldots, i_{n1}$, and let U_k be a universe containing k copies of each item in U_1. Let i_{mj} denote the jth copy of the mth item in U_1. Thus $U_k = \cup_{m=1}^{n} \cup_{j=1}^{k} i_{mj}$.

Let Q be the transition rate matrix of the CTMC on U_1. We construct a contractible partition of U_k into the n sets, each containing the k copies of some item in U_1. Thus $A_m = \cup_{j=1}^{k} i_{mj}$. By the definition of copies, that $\{A_m\}_{m=1}^{n}$ is a contractible partition of U_k with $\Lambda = Q$. Noting $|A_m| = k$ for all m in Equation (3) above results in $\{\pi(A_m)\}_{m=1}^{n}$ being the solution to $\pi^T Q = \pi^T \Lambda = 0$. Thus $p_{i_m U_1} = \pi(A_m) = \sum_{j=1}^{k} p_{i_{mj} U_k}$ for each m, showing that the model exhibits uniform expansion. $\qquad\square$

We end this section by noting that every PCMC model has a trivial contractible partition into singletons. Detection and exploitation of Q's non-trivial contractible partitions (or appropriately defined "nearly contractible partitions") are interesting open research directions.

4 Inference and prediction

Our ultimate goal in formulating this model is to make predictions: using past choices from diverse subsets $S \subseteq U$ to predict future choices. In this section we first give the log-likelihood function $\log \mathcal{L}(Q; \mathcal{C})$ of the rate matrix Q given a choice data collection of the form $\mathcal{C} = \{(i_k, S_k)\}_{k=1}^{n}$, where $i_k \in S_k$ was the item chosen from S_k. We then investigate the ability of a learned PCMC model to make choice predictions on empirical data, benchmarked against learned MNL and MMNL models, and interpret the inferred model parameters $\hat{Q}$. Let $C_{iS}(\mathcal{C}) = |\{(i_k, S_k) \in \mathcal{C} : i_k = i, S_k = S\}|$ denote the number of times in the data that i was chosen out of set S for each $S \subseteq U$, and let $C_S(\mathcal{C}) = |\{(i_k, S_k) \in \mathcal{C} : S_k = S\}|$ be the number of times that S was the choice set for each $S \subseteq U$.

4.1 Maximum likelihood

For each $S \subseteq U$, $i \in S$, recall that $p_{iS}(Q)$ is the probability that i is selected from set S as a function of the rate matrix Q. After dropping all additive constants, the log-likelihood of Q given the data $\mathcal{C}$ (derived from the probability mass function of the multinomial distribution) is:

$$\log \mathcal{L}(Q; \mathcal{C}) = \sum_{S \subseteq U} \sum_{i \in S} C_{iS}(\mathcal{C}) \log(p_{iS}(Q)).$$

Recall that for the PCMC model, $p_{iS}(Q) = \pi_S(i)$, where π_S is the stationary distribution for a CTMC with rate matrix Q_S, i.e. $\pi_S^T Q_S = 0$ and $\sum_{i \in S} \pi_S(i) = 1$. There is no general closed form expression for $p_{iS}(Q)$. The implicit definition also makes it difficult to derive gradients for $\log \mathcal{L}$ with respect to the parameters q_{ij}. We employ SLSQP [27] to maximize $\log \mathcal{L}(Q; \mathcal{C})$, which is non-concave in general. For more information on the optimization techniques used in this section, see the Supplementary Materials.

4.2 Empirical data results

We evaluate our inference procedure on two empirical choice datasets, `SFwork` and `SFshop`, collected from a survey of transportation preferences around the San Francisco Bay Area [18]. The `SFshop` dataset contains 3,157 observations each consisting of a choice set of transportation alternatives available to individuals traveling to and returning from a shopping center, as well as a choice from that choice set. The `SFwork` dataset, meanwhile, contains 5,029 observations consisting of commuting options and the choice made on a given commute. Basic statistics describing the choice set sizes and the number of times each pair of alternatives appear in the same choice set appear in the Supplementary Materials[1].

We train our model on observations $T_{\text{train}} \subset \mathcal{C}$ and evaluate on a test set $T_{\text{test}} \subset \mathcal{C}$ via

$$\text{Error}(T_{\text{train}}; T_{\text{test}}) = \frac{1}{|T_{\text{test}}|} \sum_{(i,S) \in T_{\text{test}}} \sum_{j \in S} |p_{jS}(\hat{Q}(T_{\text{train}})) - \tilde{p}_{iS}(T_{\text{test}})|, \tag{4}$$

where $\hat{Q}(T_{\text{train}})$ is the estimate for Q obtained from the observations in T_{train} and $\tilde{p}_{iS}(T_{\text{test}}) = C_{iS}(T_{\text{test}})/C_S(T_{\text{test}})$ is the empirical probability of i was selected from S among observations in T_{test}. Note that $\text{Error}(T_{\text{train}}; T_{\text{test}})$ is the expected ℓ_1-norm of the difference between the empirical distribution and the inferred distribution on a choice set drawn uniformly at random from the observations in T_{test}. We applied small amounts of additive smoothing to each dataset.

We compare our PCMC model against both an MNL model trained using Iterative Luce Spectral Ranking (I-LSR) [23] and a more flexible MMNL model. We used a discrete mixture of k MNL models (with $O(kn)$ parameters), choosing k so that the MMNL model had strictly more parameters than the PCMC model on each data set. For details on how the MMNL model was trained, see the Supplementary Materials.

Figure 3 shows $\text{Error}(T_{\text{train}}; T_{\text{test}})$ on the `SFwork` data as the learning procedure is applied to increasing amounts of data. The results are averaged over 1,000 different permutations of the data with a 75/25 train/test split employed for each permutation. We show the error on the testing data as we train with increasing proportions of the training data. A similar figure for `SFshop` data appears in the Supplementary Materials.

We see that our model is better equipped to learn from and make predictions in both datasets, and when using all of the training data, we observe an error reduction of 36.2% and 46.5% compared to MNL and 24.4% and 31.7% compared to MMNL on `SFwork` and `SFshop` respectively.

Figure 3 also gives two different heat maps of $\hat{Q}$ for the `SFwork` data, showing the relative rates $\hat{q}_{ij}/\hat{q}_{ji}$ between pairs of items as well as how the total rate $\hat{q}_{ij} + \hat{q}_{ji}$ between pairs compares to total rates between other pairs. The index ordering of each matrix follows the estimated selection probabilities of the PCMC model on the full set of the alternatives for that dataset. The ordered options for `SFwork` are: (1) driving alone, (2) sharing a ride with one other person, (3) walking,

---

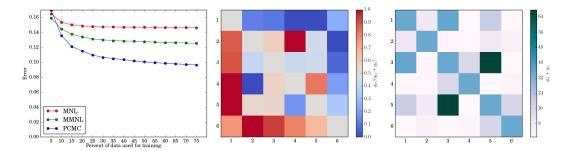[1]Data and code available here: https://github.com/sragain/pcmc-nips

Figure 2: Prediction error on a 25% holdout of the `SFwork` data for the PCMC, MNL, and MMNL models. PCMC sees improvements of 34.1% and 23.1% in prediction error over MNL and MMNL, respectively, when training on 75% of the data.

(4) public transit, (5) biking, and (6) carpooling with at least two others. Numerical values for the entries of $\hat{Q}$ for both datasets appear in the Supplementary Materials.

The inferred pairwise selection probabilities are $\hat{p}_{ij} = \hat{q}_{ji}/(\hat{q}_{ji} + \hat{q}_{ij})$. Constructing a tournament graph on the alternatives where $(i,j) \in E$ if $\hat{p}_{ij} \geq 0.5$, cyclic triplets are then length-3 cycles in the tournament. A bound due to Harary and Moser [13] establishes that the maximum number of cyclic triples on a tournament graph on n nodes is 8 when $n = 6$ and 20 when $n = 8$. According to our learned model, the choices exhibit 2 out of a maximum 8 cyclic triplets in the `SFwork` data and 6 out of a maximum 20 cyclic triplets for the `SFshop` data.

Additional evaluations of predictive performance across a range of synthetic datasets appear in the Supplementary Materials. The majority of datasets in the literature on discrete choice focus on pairwise comparisons or ranked lists, where lists inherently assume transitivity and the independence of irrelevant alternatives. The `SFwork` and `SFshop` datasets are rare examples of public datasets that genuinely study choices from sets larger than pairs.

5   Conclusion

We introduce a Pairwise Choice Markov Chain (PCMC) model of discrete choice which defines selection probabilities according to the stationary distributions of continuous time Markov chains on alternatives. The model parameters are the transition rates between pairs of alternatives.

In general the PCMC model is not a random utility model (RUM), and maintains broad flexibility by eschewing the implications of Luce's choice axiom, stochastic transitivity, and regularity. Despite this flexibility, we demonstrate that the PCMC model exhibits desirable structure by fulfilling uniform expansion, a property previously found only in the Multinomial Logit (MNL) model and the intractable Elimination by Aspects model.

We also introduce the notion of contractibility, a property motivated by thought experiments instrumental in moving choice theory beyond the choice axiom, for which Yellott's axiom of uniform expansion is a special case. Our work demonstrates that the PCMC model exhibits contractibility, which implies uniform expansion. We also showed that the PCMC model offers straightforward inference through maximum likelihood estimation, and that a learned PCMC model predicts empirical choice data with a significantly higher fidelity than both MNL and MMNL models.

The flexibility and tractability of the PCMC model opens up many compelling research directions. First, what necessary and sufficient conditions on the matrix Q guarantee that a PCMC model is a RUM [11]? The efficacy of the PCMC model suggests exploring other effective parameterizations for Q, including developing inferential methods which exploit contractibility. There are also open computational questions, such as streamlining the likelihood maximization using gradients of the implicit function definitions. Very recently, learning results for nested MNL models have shown favorable query complexity under an oracle model [2], and a comparison of our PCMC model with these approaches to learning nested MNL models is important future work.

8

Ravi Kumar for noticing an irregularity in the previously published version of Figure 4 that helped us identify a minor bug. Figures 3 and 4 have been fixed in this arxiv version.

References

[1] E. Adams and S. Messick. An axiomatic formulation and generalization of successive intervals scaling. *Psychometrika*, 23(4):355–368, 1958.

[2] A. R. Benson, R. Kumar, and A. Tomkins. On the relevance of irrelevant alternatives. In *WWW*, 2016.

[3] J. Blanchet, G. Gallego, and V. Goyal. A markov chain approximation to choice modeling. In *EC*, 2013.

[4] H. D. Block and J. Marschak. Random orderings and stochastic theories of responses. *Contributions to Probability and Statistics*, 2:97–132, 1960.

[5] A. Börsch-Supan. On the compatibility of nested logit models with utility maximization. *Journal of Econometrics*, 43(3):373–388, 1990.

[6] J H Boyd and R E Mellman. The effect of fuel economy standards on the us automotive market: an hedonic demand analysis. *Transportation Research Part A: General*, 14(5-6):367–378, 1980.

[7] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[8] S. Chen and T. Joachims. Modeling intransitivity in matchup and comparison data. In *WSDM*, 2016.

[9] Shuo Chen and Thorsten Joachims. Predicting matchups and preferences in context. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 775–784, New York, NY, USA, 2016. ACM.

[10] G. Debreu. Review of individual choice behavior: A theoretical analysis. *American Economic Review*, 1960.

[11] J.-C. Falmagne. A representation theorem for finite random scale systems. *J. Math. Psych.*, 18(1):52–72, 1978.

[12] P. E. Green, A. M. Krieger, and Y. Wind. Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, 31(3_supplement):S56–S73, 2001.

[13] Frank Harary and Leo Moser. The theory of round robin tournaments. *The American Mathematical Monthly*, 73(3):231–246, 1966.

[14] J. Huber, J. W. Payne, and C. Puto. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, pages 90–98, 1982.

[15] S. Ieong, N. Mishra, and O. Sheffet. Predicting preference flips in commerce search. In *ICML*, 2012.

[16] J. Kleinberg and M. Raghu. Team performance with test scores. In *EC*, pages 511–528, 2015.

[17] R. Kohli and K. Jedidi. Error theory for elimination by aspects. *Operations Research*, 63(3):512–526, 2015.

[18] F. S Koppelman and C. Bhat. A self instructing course in mode choice modeling: multinomial and nested logit models. *US Department of Transportation, Federal Transit Administration*, 31, 2006.

[19] Ravi Kumar, Andrew Tomkins, Sergei Vassilvitskii, and Erik Vee. Inverting a steady-state. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 359–368. ACM, 2015.

[20] R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

[21] R. D. Luce. The choice axiom after twenty years. *J. Math. Psych.*, 15(3):215–233, 1977.

[22] C. F. Manski. The structure of random utility models. *Theory and Decision*, 8(3):229–254, 1977.

[23] L. Maystre and M. Grossglauser. Fast and accurate inference of plackett–luce models. In *NIPS*, 2015.

[24] D. McFadden. Econometric models for probabilistic choice among products. *Journal of Business*, pages S13–S29, 1980.

[25] D. McFadden, K. Train, et al. Mixed mnl models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 2000.

[26] S. Negahban, S. Oh, and D. Shah. Rank centrality: Ranking from pair-wise comparisons. *arXiv preprint arXiv:1209.1688v4*, 2015.

[27] J. Nocedal and S. J. Wright. Numerical optimization. 2006.

[28] I. Simonson and A. Tversky. Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29(3):281, 1992.

[29] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

[30] J. S. Trueblood, S. D. Brown, A. Heathcote, and J. R. Busemeyer. Not just for consumers context effects are fundamental to decision making. *Psychological Science*, 24(6):901–908, 2013.

[31] A. Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281, 1972.

[32] J. I. Yellott. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *J. Math. Psych.*, 15(2):109–144, 1977.

Supplemental Materials

6   Parameterizations of the Q matrix

As observed in the main paper, if we parametrize Q with pairwise probabilities from a BTL model with parameters $\gamma = \{\gamma_i\}_{i=1}^n$, i.e. $q_{ij} = p_{ji} = \frac{\gamma_j}{\gamma_i + \gamma_j}$, the resulting PCMC model is equivalent to an MNL model with parameters γ. In this section we explore some other ways of parameterizing Q via a pairwise probability matrix P with entries p_{ij} and setting $Q = P^T$.

Blade-Chest models [8] are based on geometric d-dimensional embeddings of alternatives, where each alternative i is parameterized with a blade vector $b_i \in \mathbb{R}^d$ and a chest vector $c_i \in \mathbb{R}^d$, in addition to a BTL-like quality parameter $\gamma_i > 0$. The Blade-Chest model comes in two variations: the *Blade-Chest distance model*, where

$$p_{ij}(b, c, \gamma) = S(||b_i - c_j||_2^2 - ||b_j - c_i||_2^2 + \gamma_i - \gamma_j),$$

where $S(x) = (1 + \exp(-x))^{-1}$ is the sigmoid/logistic function, and the *Blade-Chest inner product model*, where

$$p_{ij}(b, c, \gamma) = S(b_i \cdot c_j - b_j \cdot c_i + \gamma_i - \gamma_j).$$

The quality parameters γ_i serve to connect the models to the BTL model, but do not meaningfully increase their expressiveness, so we disregard them in our use of the Blade-Chest model here. These two Blade-Chest models provide useful parameterizations of non-transitive pairwise probability matrices, $P(\theta)$, with $\theta = \{b, c\}$ consisting of the $2dn$ parameters of the "blade" and "chest" embeddings.

Another technique for parametrizing Q involves representing q_{ij} as a function of features of i and j, i.e. $q_{ij} = f(X_i, X_j; \theta)$ where X_i gives the salient features of i, and θ represents parameters that, for instance, give weights to these features. We can also formulate such analysis as a factoring $Q = W^T F - D$, where W is a weight matrix, F a feature matrix, and D a diagonal matrix ensuring that the row sums of Q are zero. We do not explore such parameterizations in this work, but merely highlight the potential to employ latent features of objects in a straight-forward manner, an approach closely related to conjoint analysis [12]. Such extensions would be similar to Chen and Joachims's work exploiting features of pairwise matchups by parametrizing the blades and chests as functions of those features [9].

7   Inference with synthetic data

We now evaluate our inference procedure's performance in three synthetic data regimes: (i) choice data generated from a PCMC model with q_{ij} drawn i.i.d. uniformly from $[0, 1]$, (ii) choice data generated for a simple MNL model with qualities γ drawn uniformly on the simplex, and (iii) choice data generated from a PCMC model with Q parameterized by a two-dimensional Blade-Chest distance model. In order to create a strongly non-transitive instance of the Blade-Chest distance model, we draw the blades b_i and chests c_i uniformly at i.i.d. points along the two-dimensional unit circle, naturally producing many triadic impasses.

The PCMC model's Q matrix has $n(n-1)$ parameters in general. When Q is parameterized according to BTL we have just n parameters, and when it is parameterized according to a Blade-Chest distance model in d dimensions, we have $2dn$ parameters.

We evaluate each parameterization (arbitrary, MNL, Blade-Chest distance) for each synthetic regime. We employ the Iterative Luce Spectral Ranking (I-LSR) algorithm to learn the model under the BTL parameterization of Q, where the PCMC model is equivalent to an MNL model. When the data is generated by MNL, we expect MNL to outperform inference under the general parameterization. When the data is generated by a non-MNL PCMC model, we expect MNL to exhibit restricted performance compared to a general parameterization, since the data is not generated by a model that MNL can capture.

7.1   Synthetic data results

We generate training data T_{train} and test data T_{test} from each model using 25 randomly chosen triplets as choice sets. We then follow the inferential procedure in the main paper to evaluare the inferential efficacy of each of the three models on data generated according to each.

Figure 3 shows our error performance as the data grows, averaged across 10 instances, for each data generating process and each inference parameterization. We generate 5000 samples, assign 1000 of these to be testing samples, and incrementally add the other 4000 samples to the training data, tracking error on the testing samples as we increase the size of the training data set.

The inference is applied to a set U with $n = 10$ objects, meaning that MNL has 20 parameters, the PCMC model with arbitrary Q has $n(n-1) = 90$ parameters, and the PCMC model with the Blade-Chest distance parameterization in $\mathbb{R}^2$ uses $2dn = 40$ parameters. Overall we examine 9 data–model pairs, trained sequentially in 5 episodes, averaged across 10 instances. The figure thus represents 450 trained models.
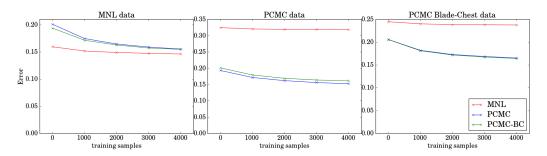
Figure 3: Learning error for PCMC, I-LSR, and Blade-Chest PCMC on synthetic data generated from (i) Arbitrary Q, (ii) MNL, and (iii) Blade-Chest models.

As expected, the inferred PCMC models outperform MNL on data exhibiting IIA violations, while the MNL model learns the MNL data better, though PCMC is not far behind. More significantly, the Blade-Chest parametrization of the PCMC model performs very similarly to the general PCMC model in all three scenarios, despite having far fewer parameters. This is promising in domains where the $O(n^2)$ parameters of a general PCMC model is infeasible but a $O(n)$ parameterization using a Blade-Chest representation may be feasible.

8   Additional empirical data results and analysis

8.1   Optimization and smoothing

The MNL models were trained using I-LSR, a specialized algorithm for training Multinomial Logit models. Meanwhile, the PCMC likelihood was optimized using SQSLP [**?**] while the Mixed MNL models were trained using L-BFGS-B [**?**], which are both general purpose optimization algorithms available as part of the `scipy.optimize.minimize` software package. The reason for the different choices is that L-BFGS-B does not support the linear constraints that are part of the PCMC model likelihood. We choose to use L-BFGS-B for the MMNL model because it outperformed SQSLP on `SFtravel` data, and we wanted to ensure that we were affording it the best possible chance to do well against the new model we contribute in this work.

The additive smoothing applied was $\alpha = 0.1$ for `SFwork` and $\alpha = 5$ for `SFshop`, where α is added to each C_{iS} appearing in the likelihood function. The major motivator for the additive smoothing is that SQSLP occasionally goes awry on some of the permutations of the data, maintaining high error after a bad step. Even as currently formulated, the mean error improvement is somewhat underestimating the efficacy of the PCMC model, as a few bad runs out of 1000 will skew the distribution. Bad runs are easy to identify using cross-validation. Additionally, the implementation of SQSLP would try to compute numerical gradients through function evaluation at points outside the feasible space, which sometimes involved $q_{ij} + q_{ji} = 0$, violating the irreducibility of the chain. Additive smoothing prevents these issues at the cost of some efficacy of the model. We find the improvement in error of the PCMC model to be even more significant in light of the optimization issues involved, and find that it reinforces development of PCMC training algorithms as an interesting research area. Refining the optimization routines for learning PCMC models should be seen as important future work.

8.2   Empirical results for `SFshop` data

Figure 4 analyzes the `SFshop` dataset[2], repeating the analysis of `SFwork` found in the main paper. The indexing of $\hat{Q}$ is again according to the estimated selection probabilities on the full set of alternatives, which were: (1) drive alone both directions, (2) share a ride with one person in both directions, (3) share a ride with one person in one direction and drive alone in the other direction, (4) walk, (5) share a ride with more than one person in both directions, (6) share a ride with one person in one direction and more than one in the other direction, (7) bike, and (8) take public transit. The MMNL model here mixes $k = 6$ models, giving it 48 parameters while PCMC has 56 and MNL has 8. When using the full training set, PCMC performs 31.3% better than MNL and 19.2% better than MMNL.

---

[2]May 2021: The version of Figure 4 that appeared in the NIPS 2016 conference proceedings incorrectly incorporated smoothing of the test data. Fixing this bug, the absolute performance of the models changes somewhat, but the relative performance and conclusions drawn remain unchanged.
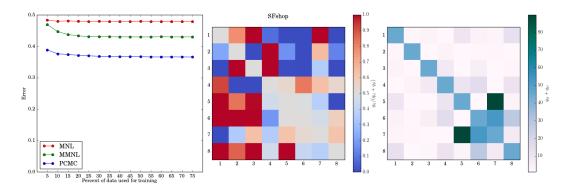
Figure 4: Prediction error on `SFshop` data for the PCMC, MNL, and MMNL models. There are improvements of 23.1% and 13.9% in prediction error over MNL and MMNL respectively when training on 75% of the data.

8.3 Inferred $\hat{Q}$ matrices

The numerical values of the learned $\hat{Q}$ matrices, trained 100% of the data, are given below.

$$\hat{Q}_{\text{work}} = \begin{bmatrix} -3.875 & 2.314 & 0.557 & 0. & 0. & 1.004 \\ 18.17 & -29.571 & 0.776 & 1.836 & 2.075 & 6.713 \\ 4.84 & 7.752 & -35.994 & 1.042 & 14.476 & 7.884 \\ 1. & 0.105 & 0.456 & -13.147 & 3.65 & 7.937 \\ 21.201 & 9.108 & 3.323 & 7.363 & -47.7 & 6.704 \\ 11.459 & 3.014 & 0.117 & 5.67 & 12.334 & -32.594 \end{bmatrix}$$

$$\hat{Q}_{\text{shop}} = \begin{bmatrix} -35.264 & 1. & 0. & 1. & 0. & 0. & 5.142 & 28.122 \\ 0. & -12.959 & 3.363 & 0. & 0. & 2.03 & 2.433 & 5.133 \\ 1.635 & 0. & -22.945 & 0.637 & 0.243 & 0. & 4.877 & 15.553 \\ 0. & 12.73 & 5.95 & -24.455 & 2.174 & 0. & 1. & 2.601 \\ 1. & 3.487 & 4.458 & 0.194 & -15.366 & 0. & 5.227 & 1. \\ 1. & 1.143 & 5.788 & 6.841 & 6.344 & -31.747 & 6.15 & 4.482 \\ 1.331 & 1.305 & 0.136 & 0. & 0.226 & 0. & -30.693 & 27.695 \\ 0. & 0. & 0.402 & 10.521 & 0. & 0. & 1.602 & -12.526 \end{bmatrix}$$

8.4 Count matrices

Here we present data about the frequency with which pairs of alternatives appear in the same choice set. Matrices A_{work} and A_{shop} have as their (i, j) entry the number of choice sets which contained both i and j in the `SFwork` and `SFshop` datasets respectively:

$$A_{\text{work}} = \begin{bmatrix} - & 1323 & 4755 & 3729 & 1658 & 4755 \\ 1323 & - & 1479 & 1395 & 797 & 1479 \\ 4755 & 1479 & - & 4003 & 1738 & 5029 \\ 3729 & 1395 & 4003 & - & 1611 & 4003 \\ 1658 & 797 & 1738 & 1611 & - & 1738 \\ 4755 & 1479 & 5029 & 4003 & 1738 & - \end{bmatrix}$$

$$A_{\text{shop}} = \begin{bmatrix} - & 3075 & 3075 & 3075 & 1844 & 3075 & 2069 & 2916 \\ 3075 & - & 3157 & 3157 & 1908 & 3157 & 2136 & 2997 \\ 3075 & 3157 & - & 3157 & 1908 & 3157 & 2136 & 2997 \\ 3075 & 3157 & 3157 & - & 1908 & 3157 & 2136 & 2997 \\ 1844 & 1908 & 1908 & 1908 & - & 1908 & 1908 & 1876 \\ 3075 & 3157 & 3157 & 3157 & 1908 & - & 2136 & 2997 \\ 2069 & 2136 & 2136 & 2136 & 1908 & 2136 & - & 2094 \\ 2916 & 2997 & 2997 & 2997 & 1876 & 2997 & 2094 & - \end{bmatrix}$$

The relative frequencies of choice sets of different sizes in the two datasets are given in Table 1.

13

Table 1:

$|S| =$	3	4	5	6	7	8
SFwork	948	1918	1461	702	-	-
SFshop	0	1	131	902	311	1812