What's Your Choice? Learning the Mixed Multi-nomial Logit Model

Ammar Ammar
ammar@mit.edu

Sewoong Oh
swoh@illinois.edu

Devavrat Shah
devavrat@mit.edu

Luis Filipe Voloch
voloch@mit.edu

ABSTRACT

Computing a ranking over choices using consumer data gathered from a heterogenous population has become an indispensable module for any modern consumer information system, e.g. Yelp, Netflix, Amazon and app-stores like Google play. In such applications, a ranking or recommendation algorithm needs to extract meaningful information from noisy data *accurately* and in a *scalable* manner. A principled approach to resolve this challenge requires a *model* that connects observations to recommendation decisions and a tractable inference algorithm utilizing this model. To that end, we abstract the preference data generated by consumers as noisy, partial realizations of their innate preferences, i.e. orderings or permutations over choices. Inspired by the seminal works of Samuelson (cf. *axiom of revealed preferences*) and that of McFadden (cf. discrete choice models for transportation), we model the population's innate preferences as a mixture of the so called Multi-nomial Logit (MMNL) model. Under this model, the recommendation problem boils down to (a) learning the MMNL model from population data, (b) finding am MNL component within the mixture that closely represents the revealed preferences of the consumer at hand, and (c) recommending other choices to her/him that are ranked high according to thus found component. In this work, we address the problem of learning MMNL model from partial preferences. We identify fundamental limitations of *any* algorithm to learn such a model as well as provide conditions under which, a simple, data-driven (non-parametric) algorithm learns the model effectively. The proposed algorithm has a pleasant similarity to the standard *collaborative filtering* for scalar (or star) ratings, but in the domain of permutations. This work advances the state-of-art in the domain of learning distribution over permutations (cf. [2]) as well as in the context of learning mixture distributions (cf. [4]).

Model and Problem Statement:
This section explains the problem of learning the mixed Multinomial Logit (MMNL) model. In particular, it provides a definition of the model, a procedure for generating the data from this model, and a lower-bound that reflects the difficulty in learning the model.

SIGMETRICS'14, June 16–20, 2014, Austin, Texas, USA.
ACM 978-1-4503-2789-3/14/06.
http://dx.doi.org/10.1145/2591971.2592020.

The MNL Model: the Multi-nomial Logit (MNL) model is a probability distribution over permutations. For a set of n items, this model is specified by a set of weight parameters $w_1, \ldots, w_n$. Given a set of items, C, the probability of choosing an item $i \in C$ is

$$\mathbb{P}_C(i; w) = \frac{w_i}{\sum_{j \in C} w_j} \ . \tag{1}$$

and the probability of a permutation σ has the form

$$\mathbb{P}(\sigma; w) = \prod_{j=1}^{n} \frac{w_{\sigma^{-1}(j)}}{w_{\sigma^{-1}(j)} + w_{\sigma^{-1}(j+1)} + \cdots + w_{\sigma^{-1}(n)}} \ . \tag{2}$$

This expression suggests the following (sequential) sampling procedure for generating σ: choose the first item, $\sigma^{-1}(1)$, from the set of all n items, then choose the second item, $\sigma^{-1}(2)$, from the remaining $n-1$ items, and so on, for n steps. Furthermore, to sample a partial permutation of length l items, terminate this procedure after l steps.

The MMNL Model: the mixed MNL (MMNL) model is a probability mixture of several MNL models. The probability assigned to a permutation σ under this model takes the form

$$\mathbb{P}(\sigma) = \sum_{k=1}^{K} \alpha_k \cdot \mathbb{P}(\sigma; w^k) \ , \tag{3}$$

where K denotes the number of components in the mixture, $\{\alpha_k\}$ the prior of the mixture (i.e., $\alpha_k \geq 0$ and $\sum_{k=1}^{K} \alpha_k = 1$), and $w^k \in \mathbb{R}^n$ the parameter vector of the k-th MNL component. To sample a full, or partial, permutation from this model, one can start by sampling an integer $k \in \{1, \ldots, K\}$ using the prior $\{\alpha_k\}$. The desired sample can then be obtained from the corresponding MNL, $\mathbb{P}(\cdot; w^k)$, as outlined previously.

Problem Statement: given a MMNL model over n items, with K mixing components, and data generated from this model, the problem of interest is that of learning the parameters α_k and w^k for all the components. The data available for this purpose consists of N IID samples in the form of partial permutations of length l, where $l < n$.

Drawing on previous work (e.g. [5][3][1]), the learning problem can be decoupled into two stages. First, identify the label of the MNL component that generated each sample. Once the labels have been identified, the samples associated with a given label can be used to estimate the parameters, α_k and w^k. That said, note $\{\alpha_k\}$ can be estimated in a straight-forward fashion from the label assignment. Given one such assignment, one only needs to estimate the parameter vectors w^k. To that end, several algorithms have been proposed (e.g. [3] and [5]). With that in mind, the problem of

learning the MMNL model reduces to that of computing a 'correct' label assignment for the data. We consider this problem in the next section.

Main Results

Intuitively, to solve the label assignment problem one has to be able to distinguish between samples coming from different MNL components, if at all possible. To this end, we provides two concrete results: one negative, in the form of a lower bound, and one positive in the form of a set of sufficient conditions under which we are able to learn the MMNL model.

A Lower Bound: the richness of the MMNL family presents us with some difficulties. One such difficulty is demonstrated by the following result; it establishes the impossibility of distinguishing members of a certain family of 'limiting' MMNL models using samples of length l for certain values of l.

Theorem 1. *For any non-negative integer i, there exist pairs of 'limiting' MMNL models with $n = 2^{i+1}$ items and $k = 2^i$ mixing components, and identical mixing distributions, where the datasets generated by both models using samples of length $l = 2i + 1$ are identical in distribution (i.e., data cannot be used to distinguish between the two mixtures).*

With this in mind, we present an algorithm for computing a label assignment for the available data points. Roughly, the algorithm partitions the data set into subsets, where the data points in each subset originate from the same MNL component. The proposed algorithm consists of two steps: a preprocessing step and a clustering step. The preprocessing step produces a *sample graph* where the nodes correspond to the data samples, and the presence (absence) of an edge reflects whether a pair of data samples share a common (or different) MNL origin. The clustering step takes this graph as an input and produces the desired partition. Finally, we provide a set of sufficient conditions, under which the computed partition is guaranteed to be consistent with the underlying MMNL model.

The Sample Graph: given a set of partial permutations $\{\sigma^1, \ldots, \sigma^N\}$ (of length l), our algorithm constructs a graph on N, with adjacency matrix $W = [W_{ij}]$ given by:

$$W_{ij} = \begin{cases} 0, & \text{if overlap}(\sigma^i, \sigma^j) < 1/2 \text{ OR affinity}(\sigma^i, \sigma^j) < 1/2 \\ 1, & \text{otherwise,} \end{cases}$$

where $\text{overlap}(\sigma^i, \sigma^j)$ denotes the fraction of items shared among the l items of σ^i and σ^j. As for $\text{affinity}(\sigma^i, \sigma^j)$, it is defined as

$$\text{affinity}(\sigma^i, \sigma^j) = \sum_{k,l \in S, k \neq l} \sigma^i_{kl} \cdot \sigma^j_{kl} \Big/ \binom{|S|}{2}$$

where S is the set of items where σ^i and σ^j overlap, and σ_{kl} is a $+1/-1$ indicator that item k is preferred to item l.

Clustering: given the *sample graph* produced in the preprocessing step, we present the following algorithm to compute a partition of the nodes:

Algorithm 1 Common Neighbors

Input: Symmetric adjacency matrix $W \in \{0,1\}^{N \times N}$

1: Initialize clustering $C = \{\{\sigma_i\}\}$.
2: **for** $(u, v) \in A$ **do**
3: **if** $|N(u) \cap N(v)| > \frac{1}{2} \min\{|N(u)|, |N(v)|\}$ **then**
4: perform operation $UNION(u, v)$
5: **end if**
6: **end for**

The algorithm uses the Union-find data structure for partitions. We begin with each node in its set, and use the union operation whenever their neighbors overlap by more than half. Upon termination, the algorithm creates a clustering of the samples.

Sufficient Conditions We would like to provide the guarantee that with high probability nodes will be clustered together if and only if they come from the same component in the mixture distribution. Theorem 2, provides one such guarantee in the form of sufficient conditions on the underlying MMNL model and the length, l, of the partial permutation. In order to state these conditions, we use the following definitions.

Definition 1. *(Geometric MNL Model)*
Let $\mathbb{P}(.; w)$ be an MNL model over n items, and let the permutation π be the ordering of the the parameters $w_{\pi_1}, w_{\pi_2}, \ldots, w_{\pi_n}$. We call this MNL model geometric if

$$\frac{w_{\pi_i}}{w_{\pi_{i+1}}} \geq a$$

for some constant $a > 1$ and for each $i < n$. Further, we use $\mathbb{P}(\sigma; \pi, a)$ to shorthand notation for this kinds of models.

Definition 2. *(Uniform Geometric MNL Model with K components)*
Let $\mathbb{P}(.; w^1, \ldots, w^K, \{\alpha_k\})$ denote an MMNL model over n items, with K mixing components, $\mathbb{P}(.; w^k)$ with $k \in \{1, \ldots, K\}$, and mixing distribution $\{\alpha^k\}$. We call this model a uniform geometric MMNL model if

(i) Each component $\mathbb{P}(.; w^k) = \mathbb{P}(.; \pi^k, a^k)$ is a geometric MNL (as per definition 1).

(ii) Each permutation π^k is drawn uniformly at random from the set of all permutations of length n items.

(iii) The mixing distribution $\{\alpha_k\}$ is a discrete uniform distribution (i.e., $\alpha_k = 1/K \quad \forall k$) .

We can now state a guarantee for the Algorithm 1 with the following theorem.

Theorem 2. *Given a* uniform geometric MMNL *model over n items with K (fixed) components, and N samples, of length l drawn, from this distribution. For $l = \Theta(\log n)$, and $N = \Theta(\text{poly} \log n)$, algorithm 1 produces a correct partition with probability at least $1 - O(\frac{\text{poly} \log n}{n})$.*

Note that this theorem requires a fixed number of components K, whereas theorem 1 relies on K growing on the order of $\Theta(n)$, with $l = \Theta(\log n)$ in both cases.

References

[1] A. P. Dempster, N. M. Laird, D. B. Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

[2] V. F. Farias, S. Jagabathula, and D. Shah. A nonparametric approach to modeling choice with limited data. *Management Science*, 59(2):305–322, 2013.

[3] D. McFadden. Conditional logit analysis of qualitative choice behavior. 1973.

[4] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[5] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *NIPS*, pages 2483–2491, 2012.