A Discrete Choice Model for Subset Selection

Austin R. Benson*
Cornell University
Ithaca, New York
arb@cs.cornell.edu

Ravi Kumar
Google Inc.
Mountain View, California
ravi.k53@gmail.com

Andrew Tomkins
Google Inc.
Mountain View, California
atomkins@gmail.com

ABSTRACT

Multinomial logistic regression is a classical technique for modeling how individuals choose an item from a finite set of alternatives. This methodology is a workhorse in both discrete choice theory and machine learning. However, it is unclear how to generalize multinomial logistic regression to *subset selection*, allowing the choice of more than one item at a time. We present a new model for subset selection derived from the perspective of random utility maximization in discrete choice theory. In our model, the quality of a subset is determined by the quality of its elements, plus an optional correction. Given a budget on the number of subsets that may receive correction, we develop a framework for learning the quality scores for each item, the choice of subsets, and the correction for each subset. We show that, given the subsets to receive correction, we can efficiently and optimally learn the remaining model parameters jointly. We show further that learning the optimal subsets is both NP-hard and non-submodular, but there are efficient heuristics that perform well in practice. We combine these pieces to provide an overall learning solution and apply it to subset prediction tasks. We find that with reasonably-sized budgets, there are significant gains in average per-choice likelihood ranging from 7% to 8x depending on the dataset and also substantial improvements over a determinantal point process model.

ACM Reference Format:
Austin R. Benson, Ravi Kumar, and Andrew Tomkins. 2018. A Discrete Choice Model for Subset Selection. In *WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining, February 5–9, 2018, Marina Del Rey, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3159652.3159702

1 INTRODUCTION

Human decision making frequently falls within the classical framework of *discrete choice*, in which an individual selects a single item from an available slate of alternatives. The items might be a set of kid-safe family sedans for possible purchase, or modes of transportation to take to work. The available alternatives may differ between individuals or over time—e.g., walking to work might not be feasible for some and a carpool may only be available certain days of the week. There is a rich history in analyzing these types of choices [2, 8, 13, 32], but typical models predict the likelihood of choosing a *single* item. Often, one chooses a *subset* of the available alternatives, e.g., a subset of items to purchase from a supermarket, or a subset of songs to form a playlist.

Of course, this is a strict extension of traditional discrete choice, in which subsets are restricted to singletons. The motivation is clear, as higher-order interactions between items are well known to have a significant effect on decision making. See, for example, research on product bundling in marketing [31] or co-purchase modeling in e-commerce recommender systems [28].

Generalizing discrete choice models to subset choice brings a number of challenges. Naively, one could consider the subsets themselves to be the items and then employ the classical models of discrete choice. However, with this approach, the size of the model grows exponentially in the number of alternatives, and the underlying structure of item correlations becomes lost in the identities of the subsets. On the other hand, one might model the selection of a subset as the independent selection of individual items. While this approach is computationally feasible, it does not capture any of the higher-order interactions between items.

Here we develop a sparse model that interpolates between a "full" model that considers the available items to be subsets and a "separable" model that assumes the presence of an item in a subset has no effect on the likelihood of selecting any other item. Our model is rooted in random utility theory so that subset choice probabilities may be interpreted as the strategy of an individual trying to maximize her utility, where the utility of each subset is a random variable expressed as the sum of a utility value (a constant) and a random error. Intuitively, our model says: (i) every item has a fixed base utility, (ii) a subset's base utility is the sum of the utilities of the items in the subset, and (iii) up to k subsets may receive an additional corrective utility, either positive or negative. Corrective utility might represent complementary value, e.g., the utility of the subset {`chips`, `guacamole`} might be larger than the sum of the individual utilities. On the other hand, while the set {`milk`, `beer`} might be frequently purchased, the utility of the subset could be well-described by the utilities of the individual items. Our model can also capture negative utility corrections. For example, `pork sausage` and `vegan tofu sausage` each carry utility, but empirically individuals may be unlikely to purchase them together.

Given a budget k, our algorithmic challenge is two-fold: (i) identify k special subsets that will most benefit from corrective utility and (ii) given these subsets, find the base (item-level) and corrective (subset-level) utilities that maximize the likelihood of the data. From the utilities, we can then determine the probability that an individual will select a given subset from a given set of alternatives through discrete choice theory. We learn the utilities (and subsequently the selection probabilities) under two different assumptions. In the first, the slate of alternatives, or *choice set*, is "universal" and common to all choices. This is largely true in the case of grocery shopping or

*Most of this work was done while the author was at Google, Mountain View, CA.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02.
https://doi.org/10.1145/3159652.3159702

"market basket" data—individuals may choose any subset of items from the food in the store, modulo issues of availability. In this setting, we prove that given the k subsets receiving corrective utility, the optimal selection probabilities for universal slates have a closed form. Under the second and more general assumption, the choice set is "variable" and may change with each instance of choice in the data. For example, clicks on the youtube.com homepage depend on the slate of options available there, which changes frequently. In this case, there is no closed form for the selection probabilities that maximize likelihood of the data, but we can compute them efficiently via a convex optimization problem formulation, after we identify the subsets that will receive corrective utility.

For both universal and variable choice sets, we show that the problem of finding the best k subsets is NP-hard, and furthermore, the optimization problem is non-submodular. Thus, we present several heuristics for identifying good subsets and evaluate them on several real-world datasets from a variety of domains. We find that using our sparse model with these heuristics provides substantial likelihood gains over the baseline separable model while maintaining the same computational feasibility. In addition, our sparse model often out-predicts determinantal point processes that have been used for modeling subset creation [15], while also taking orders of magnitude less time to train. We also find that for some universal choice sets, the corrective utilities are always positive. We show that under this condition, our discrete choice models have an alternate representation as a mixture of two independent multinomial logits. Such mixtures have been studied heavily in discrete choice [33], where the structure is typically set by the data modeler in advance, rather than learned from data, as in our model.

To summarize, our contributions are the following:
(i) We develop a new discrete choice model for subset choice from the perspective of random utility theory, where the utility of subsets is the sum of the utility of the items and some special subsets receive corrective utility. The model parameter complexity is parameterized by the number of subsets receiving corrective utility and is close to that of multinomial logistic regression. In addition, the learned values are highly interpretable.
(ii) We analyze several structural properties of the model and give algorithms to learn the model parameters. We prove that the subset selection probabilities have a closed form for universal choice sets and a convex optimization form for variable choice sets. We also prove that it is NP-hard to find the best k subsets to receive corrective utility.
(iii) We find that over a collection of six datasets with universal choice sets, introduction of the corrective utilities improves likelihood on a held out test set in all cases. Our core metric is the geometric mean likelihood of selecting each subset. Under this metric, two of our datasets show significant gains of 5x or more; three more show between between 20% and 40% lift; and the remaining one over 7%, all with just a few corrective utilities. We also see improvements of 5–8% in two datasets with variable choice set, again with just a few corrective utilities.

Implementations of our algorithms and the datasets used in our experiments are available at

https://github.com/arbenson/discrete-subset-choice.

2 SUBSET CHOICE AND RANDOM UTILITY THEORY

We now develop our model for subset selection. We begin with a quick summary of *random utility theory*, a "rational actor" model of human decision making. In this theory, an individual's value for every possible choice is drawn from some pre-specified joint distribution. The user then behaves rationally by selecting the option of highest utility. Our models (and multinomial logit) belong to the class of random utility models and hence lie within a principled framework of decision making. Stated another way, for all the models we consider, the predicted probability of selecting a particular item or subset from a set of alternatives is always the likelihood of a rational user making this choice, under an appropriate joint utility distribution from which the rational user draws values.

The remainder of this section proceeds as follows. In Section 2.1 we review the multinomial logit model from the perspective of random utility theory. We then develop our model for size-2 subsets in Section 2.2 and extend to arbitrary subsets in Section 2.3. With the model fully described, Section 3 gives the learning algorithms and results for the case in which the choice set is always the entire universe of items, and Section 4 extends to choice sets.

2.1 Background on multinomial logit

We briefly derive the classical multinomial logit model from random utility maximization (see, e.g., the book by Train [33] for a thorough treatment). In this setting, choice theory stipulates the selection of a *single* element from a set C of choices, where C is a subset of some universe $\mathcal{U}$. The utility of choice $i \in C$ is a random variable

$$U_i = V_i + \epsilon_i, \tag{1}$$

where V_i is the inherent *quality* of item i (a constant), and the ϵ_i's are i.i.d. following a Gumbel distribution.

If individuals select the highest-utility option under this random model, an algebraic manipulation shows that the probability of selecting choice i from the choice set C is

$$\Pr[\text{select } i \mid C] = \frac{e^{V_i}}{\sum_{j \in C} e^{V_j}}. \tag{2}$$

In other words, the probability of selecting an item is proportional to the exponential of the base utility of the item, e^{V_i}, *independent* of the choice set C.

In the datasets we consider in this paper, there are many instances of selections made from choice sets. If the choice sets are $C_1, \ldots, C_N$, and $i \in C_r$, then the random utility of choosing item i in choice set C_r is

$$U_i^{(r)} = V_i + \epsilon_{ir}, \tag{3}$$

where the ϵ_{ir}'s are i.i.d. Gumbel (cf. Equation (1)). In other words, the errors are i.i.d. random across different choice sets. However, the choice probability (Equation (2)) remains the same. For the remainder of the paper, we drop the notational dependence on the choice set identifier r when it is clear from context.

While this formulation comes from discrete choice theory, multinomial logistic regression is also a widespread tool in machine learning. In this domain, one usually has a fixed choice set C of classes, a coefficient vector β_i for each class i, a feature vector x for each sample, and then assumes a linear form for the quality of each class, i.e., $V_i = \beta_i^T x$. Here, the probability of labeling the

sample as class i (or in our language, "choosing" class i as the appropriate choice for the context encoded by the sample's features) is proportional to $e^{\beta_i^T x}$, and the interest is in learning the regression coefficients β_i. In this paper, we focus on learning the inherent qualities V_i and not the regression coefficients, although our models can accommodate that structure.

2.2 Fixed-size subset choice models

We are interested in extending the traditional multinomial logit model to a random utility model of subset selection. Instead of a choice being an element $i \in C$, a choice will be a subset $S \subset C$. Technically, we consider S to be a multiset, so selections may be duplicated in S. This provides additional modeling flexibility for common situations in Web data such as multiple clicks on a link.

We first consider the case of $|S| = 2$ and generalize later to larger cardinalities. We begin by presenting the *full model* that captures arbitrary interactions between elements by assigning a general utility to each of the exponentially many possible subsets. We contrast this with the *separable model* in which each item has a utility, and each subset delivers the summed utilities of its members; this model is highly efficient but captures no higher-order interactions. We then present our *sparse model* that is between these two extremes.

Full model. One simple model for subset choice is to assume a multinomial logit with an expanded space of choices, i.e., the space of subsets. In this case, the random utility of selecting a subset $S = \{i, j\}$ is

$$U_{ij} = V_{ij} + \epsilon_{ij}, \tag{4}$$

where the ϵ_{ij}'s are i.i.d. from a Gumbel distribution. In the utility maximization setting,

$$\Pr[\text{select } \{i, j\} \mid C] \propto e^{V_{ij}}. \tag{5}$$

In the case when $|S| > 2$, we again specify a unique quality V_S for each subset of that size.

This model has the advantage that it captures arbitrary higher-order structure in the utilities. However, the model grows exponentially in the size of the subsets considered, and thus is impractical for modern large-scale datasets.

Separable model. An alternative model for subset choice assumes that the utilities of elements in a subset separate additively. In this model, the presence of any given item in the subset has no effect on the remainder of the subset.

Observation 2.1. *Let the random utility of the set $\{i, j\}$ be*

$$U_{ij} = V_i + V_j + \epsilon_{ij}, \tag{6}$$

where ϵ_{ij}'s are i.i.d. Gumbel-distributed. Then the probability of any item i belonging to the subset, conditioned on item j belonging to the subset, is the same for all j.

To see this, if we condition on $j \in S$ under the utility maximization framework, we get a term independent of j:

$$\Pr[\text{select } \{i, j\} \mid C,\ j] = \frac{e^{V_i + V_j}}{\sum_{k \in C} e^{V_j + V_k}} = \frac{e^{V_i}}{\sum_{k \in C} e^{V_k}}. \tag{7}$$

The model generalizes to subsets S of larger size in the same way—the inherent subset qualities just separate additively as in Equation (6). The model has the advantage that the number of parameters is linear in the number items (i.e., linear in $|\mathcal{U}|$). However, the separable assumption on the utility is quite strong as it ignores higher-order interactions between the items in the subset.

Sparse model. In this work, we focus on a new model that interpolates between the full and separable models. This is a sparse model that takes the separable model as a baseline and adds sparse corrections for several subsets. We again derive this model from random utility theory and again begin with subsets of cardinality 2. Let H be a (small) set of pairs of indices, representing subsets to receive utility correction. The random utility U_{ij} is

$$U_{ij} = \begin{cases} V_i + V_j + \epsilon_{ij} & \{i, j\} \notin H \\ V_i + V_j + W_{ij} + \epsilon_{ij} & \{i, j\} \in H, \end{cases} \tag{8}$$

where the ϵ_{ij}'s are i.i.d. Gumbel, V_i and V_j are constants representing the individual utilities of item i and j, and W_{ij} is a constant representing the additional utility of the subset $\{i, j\}$.

Intuitively, we follow the separable model for most cases but allow ourselves the flexibility to correct the utility for special pairs of indices in H. This sparse model is a special case of both the separable model (by setting H to be empty) and the full model (by setting H to be all pairs). We emphasize that this model is still a multinomial logit—the probability of choosing a set is proportional to the exponential of the constant term in the random utility. However, we have imposed a certain structure on these constants.

The subset utilities given above may be normalized to convenient probabilities as shown in the following observation. We use this representation throughout the paper.

Observation 2.2. *Under the random utility model of Equation* (8) *the probability of selecting $S = \{i, j\}$ is proportional to*

$$p_{ij} = \begin{cases} \gamma p_i p_j & \{i, j\} \notin H \\ \gamma p_i p_j + q_{ij} & \{i, j\} \in H, \end{cases} \tag{9}$$

for some parameters p_i, q_{ij}, and γ, where $\sum_i p_i = 1$, $p_i \geq 0$, $\gamma \geq 0$, and $\sum_{\{i,j\}} p_{ij} = 1$.

To get this representation, set $p_i = e^{V_i} / \sum_j e^{V_j}$, $q'_{ij} = e^{V_i} e^{V_j} e^{W_{ij}} - p_i p_j$ if $\{i, j\} \in H$, $q'_{ij} = 0$ if $\{i, j\} \notin H$, and $\gamma = 1/(\sum_{\{i,j\}} (p_i p_j + q'_{ij}))$. Under utility maximization,

$$\Pr[\text{select } \{i, j\} \mid C] \propto e^{V_i} e^{V_j} e^{W_{ij}} = p_i p_j + q'_{ij} \propto \gamma (p_i p_j + q'_{ij}),$$

if $\{i, j\} \in H$. Finally, set $q_{ij} = \gamma q'_{ij}$ to get Equation (9). If $|S| > 2$, then the parameterization is again a mixture of the separable and full models: $p_S = \gamma \prod_{i \in S} p_i + q_S$ for $S \in H$.

At this point, Equation (9) provides a model for the likelihood of choosing a subset if we know the size of the subset to be selected. In the next subsection, we provide a complete model that does not rely on this conditioning.

2.3 Variable-size sparse subset choice model

We now remove the restriction that choices are a fixed size. In this case, our model becomes a mixture of multinomial logits, where each component of the mixture is a model for choosing subsets of a fixed size. Importantly, the utilities from the mixture components overlap so that the utility for each subset is the sum of utilities of the individual items in the subset plus a possible corrective term.

Following this mixture intuition, we select a subset by first choosing the size of the selection (the cardinality of S) and then choosing the subset S itself, conditioned on its size. Formally, given a subset

S of cardinality k, we say that the probability of selecting S from a choice set C is

$$\frac{z_k}{z_1 + \cdots + z_{|C|}} \cdot \Pr[\text{select } S \mid C, \text{ size-}k \text{ selection}], \tag{10}$$

where z_k is the probability of choosing a size-k set, $\sum_k z_k = 1$, (here, this is normalized over the possible selection size probabilities given C). Such an approach is common in other variable-size choice models such as approval voting [10, 24], and we will show in Section 3 that it makes learning model parameters simple.

Given the size of S, we model the selection using the sparse model already outlined in Section 2.2 with the crucial assumption that *the individual item qualities V_i are the same regardless of the size of S*. In other words, baseline item qualities are the same regardless of the size of the set in which they appear. However, we learn a parameter γ_k for each subset size k and H may contain multisets of various sizes. For example, if $H = \{\{i, j\}, \{i, j, k\}\}$, then

$\Pr[S = \{i, j\} \mid \text{size-2 choice}] \propto \gamma_2 p_i p_j + q_{ij}$

$\Pr[S = \{i, j, k\} \mid \text{size-3 choice}] \propto \gamma_3 p_i p_j p_k + q_{ijk}$.

Our formulation permits size-1 selections (and our datasets contain such selections), but we restrict H to contain multisets of size at least two, since the correction term is meant to model higher-order effects. Finally, we note that the variable-size sparse subset model is a random utility model as it is a mixture of logits [33].

3 UNIVERSAL CHOICE SETS

We first study the sparse model with universal choice sets, i.e., all subset selections are made from the same slate of alternatives, which is the universe of items $\mathcal{U}$. This is common in a variety of domains, including "market basket" data where individuals buy a collection of items from a store. We show that, given the set of corrections H, the maximum likelihood estimator for the sparse model has a simple closed form. However, we also show that finding the optimal set H is NP-hard. We compare several heuristics for choosing H on a variety of real-world datasets and find substantial likelihood improvements with the sparse model. These heuristics also lead to models that are better predictors on our datasets compared to more involved probabilistic models for subset selection, namely determinantal point processes.

3.1 Optimizing model parameters

We now assume that the set H is given and learn the model parameters in two parts: (i) the mixture probabilities z_k of Equation (10) and (ii) the subset selection probabilities of Observation 2.2. We assume that our data is a list of subsets $S_1, S_2, \ldots, S_m$ representing the selections from the universe of items $\mathcal{U}$ and that there are $n = |\mathcal{U}|$ items. We use maximum likelihood estimation (MLE) for optimizing the model parameters.

Mixture probabilities. It is straightforward to find the MLE for z_k, the probabilities of choosing a subset of a particular size. Since the choice set is universal, $\sum_k z_k = 1$. Let N_S be the number of times that the set S was chosen in the data. Then the log-likelihood given the data is:

$$\begin{aligned}&\textstyle\sum_k \sum_{|S|=k} N_S \log(z_k \Pr[S \mid \text{size-}k \text{ choice}])\\ &= \textstyle\sum_k (\sum_{|S|=k} N_S) \log z_k + \sum_k \sum_{|S|=k} N_S \log \Pr[S \mid \text{size-}k \text{ choice}].\end{aligned}$$

Under the constraint that $\sum_k z_k = 1$, it is easy to see that the MLE for z_k is simply the fraction of times a size-k subset was selected.

Subset selection probabilities. We now seek to maximize likelihood by optimizing the probabilities of choosing a subset for a fixed-size selection. For simplicity, we derive this for the case when $|S| = 2$. Let N_{ij} be the number of times that set $S = \{i, j\}$ was selected in the data. Then, the likelihood maximization problem is

$$\begin{aligned}\underset{p,q,\gamma}{\text{maximize}} \quad & \textstyle\sum_{\{i,j\}\in H} N_{ij} \log(\gamma p_i p_j + q_{ij}) + \sum_{\{i,j\}\notin H} N_{ij} \log(\gamma p_i p_j) \\ \text{subject to} \quad & \textstyle\sum_{i=1}^n p_i = 1,\ p_i \geq 0, 1 \leq i \leq n \\ & \textstyle\sum_{\{i,j\}} \gamma p_i p_j + \sum_{\{i,j\}\in H} q_{ij} = 1,\ \gamma \geq 0.\end{aligned} \tag{11}$$

Here, the likelihood formulation uses the structure of Equation (9). Theorem 3.1 below provides a simple closed-form solution for p, q, and γ that maximizes the likelihood. Essentially, using Equation (9) to codify the random utility model of Equation (8), the optimal parameters take an intuitive form: the p_i are empirical frequencies of item appearances for subset choices not in H and then γ and q_{ij} adjust the probability of subset choices in H to match their empirical frequency of selection. This generalizes the single-item logit model, where the exponentials of the maximum likelihood utilities match the empirical frequency of item selection.

THEOREM 3.1. *Let $p^D_{ij} = N_{ij} / \sum_{\{k,l\}} N_{kl}$ be the empirical probability of observing set $\{i, j\}$ in the data. The MLE for Equation* (11) *is:*
(i) the p_i's are proportional to the number of times item i is selected in any set $\{i, j\} \notin H$, i.e., $p_i \propto \sum_{j:\{i,j\}\notin H} N_{ij}$;
(ii) $\gamma = (1 - \sum_{\{i,j\}\in H} p^D_{ij}) / (\sum_{\{i,j\}\notin H} p_i p_j)$; and
(iii) given p and γ, q is set to match the empirical distribution of $\{i, j\}$, i.e., $\gamma p_i p_j + q_{ij} = p^D_{ij}$.

PROOF. Suppose p and γ are fixed. Dividing Equation (11) by the constant $\sum_{ij} N_{ij}$, the likelihood maximization problem becomes

$$\begin{aligned}\underset{q}{\text{maximize}} \quad & \textstyle\sum_{\{i,j\}\in H} p^D_{ij} \log(\gamma p_i p_j + q_{ij}) \\ \text{subject to} \quad & \textstyle\sum_{\{i,j\}\in H} q_{ij} = 1 - \gamma \sum_{\{i,j\}} p_i p_j.\end{aligned}$$

Setting the gradient of the Lagrangian to zero (with Lagrange multiplier λ) gives $p^D_{ij}/(\gamma p_i p_j + q_{ij}) - \lambda = 0 \rightarrow \gamma p_i p_j + q_{ij} = p^D_{ij}/\lambda$.

Let $R = \sum_{\{i,j\}\in H} p_i p_j$, $M = \sum_{\{i,j\}\in H} p^D_{ij}$, and $C = \sum_{\{i,j\}} p_i p_j$. Then the constraints give

$$\lambda = M/(1 - \gamma(C - R)). \tag{12}$$

Now suppose p is fixed in Equation (11). Plugging Equation (12) into the objective function over γ gives (up to constant)

$$\textstyle\sum_{\{i,j\}\in H} p^D_{ij} \log(1/\lambda) + \sum_{\{i,j\}\notin H} p^D_{ij} \log(\gamma) \tag{13}$$

$$= -M \log M + M \log(1 - \gamma(C - R)) + (1 - M) \log \gamma \tag{14}$$

Equation (14) is maximized for $\gamma = (1 - M)/(C - R) \geq 0$. For this value of γ, $\lambda = 1$ and $\gamma p_i p_j + q_{ij} = p^D_{ij}$. This proves parts (ii) and (iii) of the theorem.

For constants γ and q, the log-likelihood is (up to a constant)

$$\textstyle\sum_{\{i,j\}\notin H} N_{ij} \log p_i p_j = \sum_i \left(\sum_{j:\{i,j\}\notin H} N_{ij}\right) \log p_i,$$

which is maximized with $p_i \propto \sum_{j:\{i,j\}\notin H} N_{ij}$.

Finally, we check that our set probabilities sum to 1:

$\sum_{\{i,j\}} \gamma p_i p_j + \sum_{\{i,j\}\in H} q_{ij} = \sum_{\{i,j\}\notin H} \gamma p_i p_j + \sum_{\{i,j\}\in H} \gamma p_i p_j + q_{ij}$

$= \gamma(C-R) + M = \frac{1-M}{C-R}(C-R) + M = 1.$ □

Theorem 3.1 can be generalized for choice sets with greater than two elements; the proof remains the same. For an arbitrary subset S with $|S| = k$, we set p_S^D to be the empirical probability of observing set S among all size-k choice sets. We then set γ_k and q according to the same formulas (ii) and (iii) in the statement of Theorem 3.1, but with size-k sets instead of size-2 sets.

At this point, if we are given the set H of corrections, it is straightforward to compute the MLE. Next, we deal with the issue of constructing a set H of a fixed size to maximize likelihood.

3.2 Constructing H

Theorem 3.1 says that once we know the subsets receiving probability corrections from the separable model, finding the parameters that maximize the likelihood of the data is easy. We are now interested in algorithms for finding H. Unfortunately, this problem is difficult in general.

Proposition 3.2. *Finding the set H with $|H| = k$ that maximizes the likelihood of the sparse model is NP-hard.*

Proof. We reduce from maximum 3-dimensional matching (3DM): given a subset $T \subset X \times Y \times Z$ and an integer k, the decision problem is to determine if there exists $M \subset T$, $|M| = k$, such that no two tuples in M share an element. We construct an instance of the sparse model maximum likelihood problem as follows: for every $(x, y, z) \in T$, create one subset choice of the subset $\{x, y, z\}$. Let $d = |T| + 1$ and let d_a be the number of times element a shows up in T ($a \in X \cup Y \cup Z$) and create $d - d_a$ choices of sets $\{a\}$. With this construction, each item is selected exactly d times.

Set the budget for the size of H to be k. We claim that if there is a 3DM of size k, then such a matching is the optimal H. Let c_a be the number of times element a shows up in a subset in H. For any H, applying (the generalization of) Theorem 3.1 says that the log-likelihood is

$k \log(1/|T|) + (|T| - k) \log \gamma + \sum_a (d - c_a) \log \frac{d-c_a}{W}$,

where $W = \sum_a d - c_a$ is the normalization constant and

$$\gamma = \frac{1 - k/|T|}{\sum_{\{x,y,z\}} p_x p_y p_z - \sum_{\{x,y,z\} \in H} p_x p_y p_z} \tag{15}$$

The normalization constant W and the numerator of Equation (15) are independent of H. Thus, the likelihood is maximized when the denominator of Equation (15) is minimized. We have that

$$-\sum_{\{x,y,z\} \in H} p_x p_y p_z = -\sum_{\{x,y,z\} \in H} \frac{(d-c_x)(d-c_y)(d-c_z)}{W^3},$$

which is minimized when $c_x = c_y = c_z = 1$ for all $\{x, y, z\} \in H$. This occurs if and only if the subsets in H form a 3DM.

Next, let $L = \sum_{S=\{x,y,z\}:|S|=3} p_x p_y p_z$, $K = \sum_{S=\{x,x,y\}:|S|=2} p_x^2 p_y$, and $D = \sum_{S=\{x,x,x\}} p_x^3$. Then $6L + 3K + D = \sum_x \sum_y \sum_z p_x p_y p_z = 1$, and $\sum_{\{x,y,z\}} p_i p_j p_k = L + K + D = \frac{1+3K+5D}{6}$. We claim that this term is minimized with a 3DM. Note that $K + D = \sum_x \sum_y p_x^2 p_y = \sum_x p_x^2 \sum_y p_y = \sum_x p_x^2 = \frac{1}{W^2} \sum_x (d - c_x)^2$. Under the constraint that k subsets are in H, this summation is minimized when H is a 3DM so that $0 \le c_x \le 1$. Next, note that $D = \sum_x p_x^3 = \frac{1}{W^3} \sum_x (d - c_x)^3$, which again is minimized with a 3DM so that $0 \le c_x \le 1$.

Table 1: Subset choice datasets with universal choice sets. The z_k are the fraction of selections that are size-k subsets.

Dataset	#items = $\lvert\mathcal{U}\rvert$	#choices	z_1	z_2	z_3	z_4	z_5
Bakery	50	67,488	0.05	0.20	0.37	0.25	0.13
WalmartItems	183	16,698	0.51	0.45	0.03	0.01	0.00
WalmartDepts	66	120,973	0.33	0.28	0.17	0.13	0.10
Kosarak	2,605	505,217	0.27	0.30	0.23	0.14	0.07
Instacart	9,544	806,662	0.19	0.21	0.21	0.21	0.19
LastfmGenres	413	643,982	0.52	0.21	0.12	0.08	0.06

Finally, we account for likelihood due to the selection of subsets not in H, which is (up to constant) $\sum_x \frac{(d-c_x)}{W} \log \frac{d-c_a}{W} = E(p)$, where $p_x = (d - c_x)/W$ and E is the entropy function. Under the constraint $\sum_x c_x = 3k$, entropy is maximized when there are exactly $3k$ elements x for which $c_x = 1$. Again, this occurs if and only if H is a 3DM. □

In addition to being NP-hard, the problem is non-submodular (formalized below), which rules out easy algorithmic approaches.

Observation 3.3. *Let $L(H)$ be the log-likelihood of the MLE. The function $f : H \to L(H)$ is non-submodular.*

To see this, consider a dataset consisting of three subset selections: $\{a, b\}$, $\{a, c\}$, and $\{d, e\}$. Using Theorem 3.1, some simple calculations show that $L(\{\{a, b\}, \{a, c\}\}) + L(\emptyset) > L(\{\{a, b\}\}) + L(\{\{a, c\}\})$.

We leave the design of approximation algorithms to future work. Instead, we develop a few simple heuristics for constructing H that still lead to substantial likelihood gains in our experiments on several real-world datasets in the next section. We briefly describe the heuristics below.

Frequency heuristic. With the frequency heuristic, we put the most frequently selected sets S in H. An advantage of this heuristic is that we can use established algorithms from frequent itemset mining to find the most frequently occurring subsets [1, 20, 27].

Lift heuristic. The frequency heuristic ignores the fact that items in frequently selected subsets may already be frequently occurring and their likelihood accounted for by high utility of those items. With the lift heuristic (analogous to the lift score in association rule mining), we select the subsets S maximizing $N_S / \prod_{i \in S} N_i$, where N_S is the number of times that the set is selected, and N_i is the number of times that item i is selected in any set.

Normalized lift heuristic. With the normalized lift heuristic, we "normalize" the lift by the frequency of the subset in the data. This corresponds to selecting the subsets S maximizing $N_S^2 / \prod_{i \in S} N_i$.

3.3 Data

We collected six universal choice set datasets from a variety of domains. We describe them below and provide summary statistics in Table 1.

Bakery. This dataset is comprised of the receipts from purchases by patrons of a bakery.[1] The items on each receipt form a selected subset. The elements of each selected subset are distinct.

WalmartItems and WalmartDepts. These datasets are constructed from individual shopping trips at Walmart.[2] For each trip,

[1] https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery

[2] https://www.kaggle.com/c/walmart-recruiting-trip-type-classification

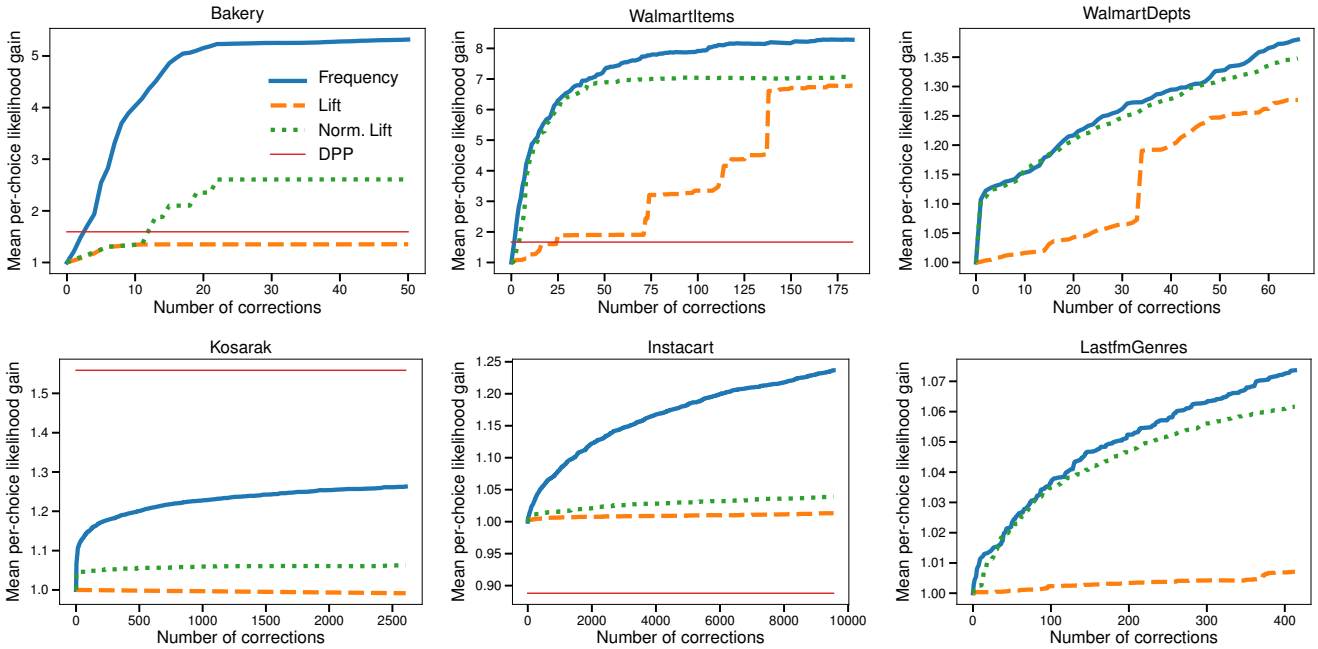

Figure 1: Mean per-choice likelihood improvements over the separable model as a function of the number of corrections in the sparse model (i.e., $|H|$) for several datasets with universal choice sets along with the relative likelihood of a determinantal point process (DPP) model. DPPs do not model subset selections with repeated items, so there are no DPP results on WalmartDepts and LastfmGenres, which contain such multiset selections. Our sparse model (i) provides substantial improvements over the separable model with only a modest increase in the number of model parameters and (ii) out-performs DPPs with at most a few corrections on Bakery, WalmartItems, and Instacart.

the items purchased are the selected subset. Using department identifiers for the items, we also construct a separate dataset where the selected subset consists of all departments from which the shopper made a purchase. Selections in WalmartDepts may contain repeats (multiple purchases from the same department), but the elements of a subsets selected in WalmartItems are distinct.

Kosarak. The Kosarak dataset is derived from publicly available anonymized clickstreams from a Hungarian online news portal.[3] The de-duplicated set of links visited by a user in a given session is the selected subset, so no subset selections contain repeated items.

Instacart. Instacart is an online same-day grocery delivery service, where users order groceries through a web application. This dataset consists of a sample of orders from users living in the United States [14].[4] We consider the de-duplicated items in a user's order to be a subset choice.

LastfmGenres. This dataset come from the listening behavior of users from the music streaming service Last.fm [6].[5] We break user behavior into sessions, where a new session is created if the user goes 20 minutes without starting a new song (this is the same procedure from our previous work [3]). We create subset choices by the genres of music played in the session, where genres are derived from user-provided tags for artists.[6] We assign an artist to the most commonly provided tag for that artist. Many subset selections contain repeated genres, corresponding to cases when a user listens to the same genre more than once in a session.

To unify structure across our data, we filter each dataset to contain subset selections of size at most 5 and only items that are selected at least 25 times. This filtering also focuses our attention on sets that are possible candidates for corrective utility, as larger sets with more than 5 items do not tend to appear repeatedly nor frequently.

3.4 Experiments

Likelihood improvements. We used the frequency, lift, and normalized lift heuristics to find correction sets H for each dataset for $|H| = 0, 1, 2, \ldots |\mathcal{U}|$ (see Table 1 for the values of $|\mathcal{U}|$). The case of $|H| = 0$ corresponds to the separable model and the case of $|H| = |\mathcal{U}|$ corresponds to the sparse model with twice the number of parameters as the separable model. Thus, all model sizes are linear in the number of items. For each value of $|H|$, we trained the model with 80% of the data and evaluated the mean per-choice likelihood gain on the remaining 20% of the data (the test data). Specifically, if LL_k is the log-likelihood on the test data when $|H| = k$, then we measured the relative improvement $r_k = e^{(LL_k - LL_0)/N}$, where N is the number of choices made in the test set.

The separable model (i.e., the case when $|H| = 0$) serves as one baseline, and we also compare against a determinantal point process (DPP) model. DPPs are probabilistic models for generating "diverse" subsets from a discrete set of objects [15]. We use the expectation maximization algorithm of Gillenwater et al. [12] to

[3] http://fimi.ua.ac.be/data
[4] https://www.instacart.com/datasets/grocery-shopping-2017
[5] http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html
[6] https://musicmachinery.com/2010/11/10/lastfm-artisttags2007

learn a DPP from the same training data as our sparse model. DPPs cannot model selected subsets with repeated items (i.e., multisets), and the WalmartDepts and LastfmGenres contain such subsets. Therefore, we do not evaluate DPPs for these two datasets.

Figure 1 shows r_k—the mean per-choice relative likelihood gain over the separable model—as a function of k for each of our datasets as well as the relative likelihood gain for the DPP (note that the DPP is just one model and does not depend on k). For all datasets, the likelihood gains from the frequency heuristic exhibit a sharp increase for small values of k and then gradually decay for larger values of k. In other words, a few sparse corrections can dramatically increase likelihood. In nearly all cases, the frequency heuristic outperforms the lift-based heuristics.

In all datasets, our sparse model provides substantial improvements over the separable model, often with just a few subsets in H receiving corrective probability. Our sparse model is also a much better predictor than DPPs in the Bakery, WalmartItems, and Instacart datasets, again with just a few subsets in H (the exception is Kosarak, where DPPs provide a performance gain over our sparse model). Our sparse model's performance is especially good for Bakery and WalmartItems, where we achieve 5x per-choice likelihood improvements with only a modest increase in the number of parameters in the model ($|H| < \frac{1}{2}|\mathcal{U}|$).

Running time. Using Theorem 3.1, the computational complexity of training the sparse subset choice model depends on (i) computing relative item frequencies, (ii) computing γ, and (iii) constructing H and computing the relative frequencies of subsets. Parts (i) and (ii) take linear time and space in the size of the data. The complexity of part (iii) is more involved and depends on the heuristic and distribution of subsets in the data.

Instead of a detailed complexity analysis, we report the time to train and test our sparse model with the frequency heuristic in Table 2 (times for the other heuristics were similar) as well as the time for the DPP model for our datasets. For our model, we measured the running time for the largest value of $|H|$ in Figure 1 (i.e., the biggest model). All experiments ran on a single thread of a 2.4 GHz Intel Xeon E7-4870 server. The software for learning our model was implemented in Julia,[7] and the software for DPP learning was implemented in MATLAB.[8] While the implementations are in different languages, our goal is to roughly compare the time it takes to use these models.

As shown in Table 2, training and testing our sparse model is at least two orders of magnitude faster than to train and test the DPP models. The EM algorithm for DPP training has a cubic dependence on the number of items $|\mathcal{U}|$ [12], so this performance discrepancy is not that surprising. Practically, the running times for our model are all quite modest, often completing in just a few seconds and taking fewer than 5 minutes for the largest dataset (Instacart).

Values of correction probabilities. Next, we analyze the actual values of the sparse corrections q. Specifically, we are interested in the sign of the corrections. Figure 2 shows that there are relatively few sets $S \in H$ for which $q_S < 0$, (regardless of the size of H) when using the frequency heuristic to construct H. In fact, for the Bakery, WalmartItems, and WalmartDepts datasets, the values

[7] https://github.com/arbenson/discrete-subset-choice
[8] https://code.google.com/archive/p/em-for-dpps

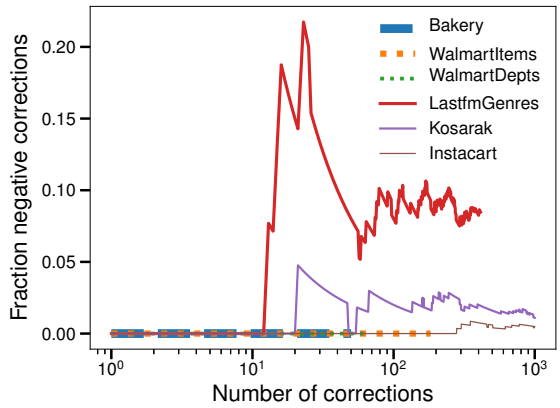

Figure 2: Fraction of subsets $S \in H$ with correction probability $q_S < 0$ as a function of the size of H when using the frequency heuristic to construct H. In many cases, the corrections are always positive, in which case our model has an alternative interpretation (Observation 3.4).

of q are always positive, and the same is true for nearly all sizes of H in the Instacart dataset. These results are consistent with prior work on supermarket purchase data showing that most co-purchase correlations are positive [9]. It turns out that our choice model has an alternative interpretation in the case when the correction values are all positive, as summarized in the following observation.

Observation 3.4. *If $q_S > 0$ for all $S \in H$, then the sparse model is a mixture of two logit models: (i) with probability $\alpha = \sum_{\{i,j\} \in H} q_{ij}$, choices follow the full model restricted to subsets in H and (ii) with probability $1 - \alpha$, choices follow the separable model.*

To see this, let $|S| = 2$ for each $S \in H$ for simplicity. Since $q_{ij}, \gamma, p_i > 0$ and $\gamma \sum_{\{i,j\}} p_i p_j + \sum_{\{i,j\} \in H} q_{ij} = 1$, $\alpha \in [0, 1]$. The probabilities for the full model are simply $\tilde{q}_{ij} = q_{ij}/\alpha$.

Thus, in many of our datasets, we can interpret user behavior as the following. First, the user decides whether or not to pick a subset from H. If the user picks from H, each set has its own utility and the user picks according to a multinomial logit. If the user does not pick from H, their set is filled with non-interacting items.

Finally, we examined the largest correction probabilities in the LastfmGenres dataset using the frequency-based algorithm. Table 3 lists the five sets with the most positive and most negative corrective values. We see that the most positive corrections account for repeat behavior—users listen to several indie, hip hop, or rock songs in the same session—whereas the negative corrections show unlikely combinations such as indie and metal.

4 VARIABLE CHOICE SETS

Next we consider datasets with subset choices where the available alternatives may change with each selection. In this setting, we no longer have the closed form of the maximum likelihood parameters provided by Theorem 3.1, but we show in Section 4.1 that the likelihood function remains concave, given the set H of subsets receiving corrective utility.

Table 2: Training and test time in seconds for (i) our sparse subset choice model with the frequency heuristic for the largest value of $|H|$ in Figure 1 and (ii) a determinantal point process (DPP) model, using EM [12]. DPPs cannot model selections with repeated items, so there are no results for WalmartDepts and LastfmGenres. In all cases, our sparse model is much faster.

	Bakery	WalmartItems	WalmartDepts	Kosarak	Instacart	LastfmGenres
Sparse subset choice	1s	0.2s	1s	26s	270s	7s
DPP	193s	12s	N/A	3,080s	67,054s	N/A

Table 3: The top five most positive and most negative correction probabilities q in LastfmGenres.

Most positive	q	Most negative	q
{indie, indie}	0.0301	{indie, metal}	-0.0015
{rock, indie}	0.0174	{indie, progressive_metal}	-0.0009
{hip_hop, hip_hop}	0.0123	{rock, rock, electronic}	-0.0007
{indie, indie, indie}	0.0119	{indie, industrial}	-0.0006
{rock, rock, rock}	0.0101	{metal, electronic}	-0.0005

4.1 Optimizing model parameters and constructing H

Again, we learn the model parameters in two parts: the mixture probabilities and the subset selection probabilities.

Mixture probabilities. There is no longer a closed form for the optimal mixture probabilities z_k that model how likely it is to choose a size-k set. The log-likelihood of choosing a size-k set S from choice set C is

$$\log \frac{z_k}{z_1+z_2+\ldots z_{|C|}} + \log \Pr[\text{select } S \mid \text{size-}k \text{ choice}]. \qquad (16)$$

Since $z_k \geq 0$, we may write $z_k = e^{Y_k}$ for some Y_k, and the log-likelihood function becomes concave. We learn the z_k with a standard gradient descent algorithm.

Subset selection probabilities. With variable choice sets, it is easier to derive the optimization of model parameters from the random utility perspective of Equation (8), rather than the probabilistic formulation of Observation 2.2 used for universal choice sets. Consider some choice set C and suppose that we know that the selected subset is of size 2. The likelihood of observing a choice $S = \{i, j\} \subset C$ is $e^{V_i+V_j+W_{ij}} / \sum_{\{k,l\}\subset C} e^{V_k+V_l+W_{kl}}$ with the understanding that $W_{kl} = 0$ if $\{k, l\} \notin H$. Let $S_r = \{i_r, j_r\} \subset C_r$ be the subset selections, $r = 1, \ldots, N$, where N is the total number of selections. Given H, maximizing the log-likelihood is the following optimization problem:

$$\begin{array}{ll} \underset{V,W}{\text{maximize}} & \sum_{r=1}^{N} V_{i_r} + V_{j_r} + W_{i_r j_r} - \log \sum_{\{k,l\}\subset C_r} e^{V_k+V_l+W_{kl}} \\ \text{subject to} & W_{ij} = 0, \text{ for } \{i, j\} \notin H. \end{array}$$

The optimization problem has a concave objective with trivial linear constraints. Thus, if we know the set of correction locations H, then we can efficiently find the optimal likelihood parameters (we use a simple gradient descent procedure). However, since the universal choice set data is a special case of variable choice sets, it is still NP-hard to find the optimal H by Proposition 3.2.

Constructing H. In our prior experiments, the frequency-based algorithm performed well in all cases, so we use that heuristic for our experiments here.

Table 4: Subset choice datasets with variable choice sets. Here, $|C|$ denotes the range of choice set sizes. The subset size probabilities z_k are learned from data (Section 4.1).

Dataset	#items	#choices	$\lvert C\rvert$	z_1	z_2	z_3	z_4	z_5
YcCats	20	134,057	2–10	0.26	0.31	0.23	0.12	0.08
YcItems	2,975	156,039	2–10	0.16	0.20	0.23	0.22	0.18

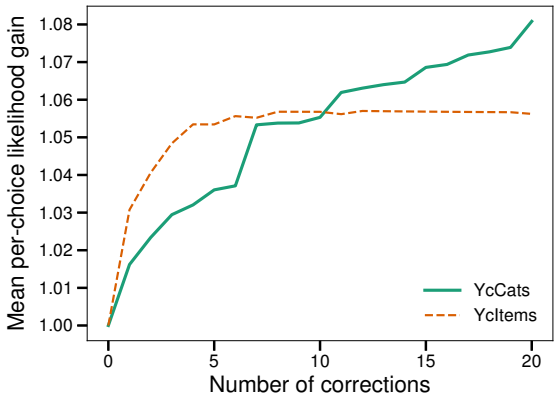

Figure 3: Mean per-choice likelihood improvements on the test set over the separable model as a function of the number of corrections in the sparse model (i.e., $|H|$) for datasets with variable choice sets.

4.2 Data

We constructed two subset choice datasets from user browsing sessions on an e-commerce Web site that contain click and purchase information (made available from YOOCHOOSE[9]). In YcItems, the variable choice set C is all items clicked on by a user in a given browsing session, and the subset selection is the set of items purchased in that session. Each item also has a category, and YcCats is constructed by considering choices at the category level. In both cases, the choice set varies for each browsing session. We filtered the datasets to consider subset selections of size at most 5 made from choice sets of size at most 10, where each item or category appears in at least 8 total choice sets and at least 4 subset selections. Table 4 provides summary statistics of the datasets.

4.3 Experiments

We used the frequency-based algorithm to find the correction sets H for each dataset for $|H| = 0, 1, 2, \ldots, 20$. Again, the case of $|H| = 0$ corresponds to the separable model. We trained the model on 80% of the data and evaluated the mean per-choice likelihood gain on the remaining 20% of the data (the same evaluation as for the datasets

[9] http://2015.recsyschallenge.com

with universal choice sets in Figure 1). There is no methodology for learning determinantal point processes with variable choice sets, so we do not compare against them here.

Figure 3 shows the mean per-choice likelihood improvements over the separable model. In both datasets, our model achieves over 5% likelihood gain with $|H| < 10$. With YcItems, we see the same pattern in likelihood gains as for the datasets with universal choice sets: a rapid increase for the first several corrections and then a leveling of the gains. For this dataset, $|H|$ is less than 1% the number of items, and we achieve substantial likelihood gains with an extremely sparse model.

5 RELATED WORK

There are variable-size choice models for universal choice sets based on pairwise interactions and conditional distributions [26], correlated random utility errors [17], or a priori knowledge of utilities [11, 35]. Our model makes no assumptions on pairwise interactions and can handle arbitrary-order interactions. Also, the datasets used in our experiments contain orders of magnitude more items than experiments for pairwise interaction models in market basket data [5, 9]. Set prediction functions in neural networks [22, 37] and mutli-label classification methods more broadly [23, 34] are also relevant to the subset prediction problem. However, these methods are designed to predict a set of labels from features (e.g., predict several tags of an image), whereas our experiments predict new subsets given previous subset selections. The recently developed set embedding model offers a statistical approach to the subset choice problem with universal choice sets [25]. Unlike our models, neural networks and the embedding models are not known to carry a random utility maximization interpretation.

Another variant of subset choice is approval voting, where an individual selects all candidates that she approves [10, 24, 29]. In this case, there is an implicit assumption that only one alternative will ultimately be selected (only one candidate will win the election), whereas we deal with subsets that give complementary utility.

Lastly, our work fits into the context of several recent analyses on effectively learning discrete choice models [4, 16, 18, 19, 21, 38] as well as applications of discrete choice models to user behavior on the Web [7, 30, 36].

6 DISCUSSION

We developed a general random utility model for how individuals choose a subset of items from a given choice set and analyzed its structure in two contexts: (i) the choice set is universal and is the same for all selections and (ii) the choice set varies. In both cases, we prove that after identifying subsets receiving corrective probability within our model, we can efficiently find the optimal model parameters. However, we also showed that finding the best set of subsets to receive corrections is NP-hard. Approximation algorithms for coping with this issue are a direction for future work. Nevertheless, our sparse model provides substantial likelihood improvements in subset prediction compared to competing baselines.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules. In *VLDB*.

[2] Moshe E Ben-Akiva and Steven R Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press.

[3] Austin R. Benson, Ravi Kumar, and Andrew Tomkins. 2016. Modeling User Consumption Sequences. In *WWW*. 519–529.

[4] Austin R. Benson, Ravi Kumar, and Andrew Tomkins. 2016. On the Relevance of Irrelevant Alternatives. In *WWW*. 963–973.

[5] Yasemin Boztuğ and Lutz Hildebrandt. 2006. A market basket analysis conducted with a multivariate logit model. In *From Data and Information Analysis to Knowledge Engineering*. Springer, 558–565.

[6] O. Celma. 2010. *Music Recommendation and Discovery in the Long Tail*. Springer.

[7] Shuo Chen and Thorsten Joachims. 2016. Modeling intransitivity in matchup and comparison data. In *WSDM*. 227–236.

[8] David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* (1958), 215–242.

[9] Katrin Dippold and Harald Hruschka. 2013. Variable selection for market basket analysis. *Computational Statistics* 28, 2 (2013), 519–539.

[10] J-Cl Falmagne and Michael Regenwetter. 1996. A random utility model for approval voting. *Journal of Mathematical Psychology* 40, 2 (1996), 152–159.

[11] Peter C Fishburn and Irving H LaValle. 1996. Binary interactions and subset choice. *European Journal of Operational Research* 92, 1 (1996), 182–192.

[12] Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. 2014. Expectation-maximization for learning determinantal point processes. In *NIPS*.

[13] Nick Hanley, Robert E Wright, and Vic Adamowicz. 1998. Using choice experiments to value the environment. *Environ. Resour. Econ* 11, 3-4 (1998), 413–428.

[14] Instacart. 2017. The Instacart Online Grocery Shopping Dataset. https://www.instacart.com/datasets/grocery-shopping-2017. (2017).

[15] Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning* 5, 2–3 (2012), 123–286.

[16] Ravi Kumar, Andrew Tomkins, Sergei Vassilvitskii, and Erik Vee. 2015. Inverting a steady-state. In *WSDM*. 359–368.

[17] Puneet Manchanda, Asim Ansari, and Sunil Gupta. 1999. The "shopping basket": a model for multicategory purchase incidence decisions. *Marketing Science* 18, 2 (1999), 95–114.

[18] Lucas Maystre and Matthias Grossglauser. 2015. Fast and accurate inference of Plackett–Luce models. In *NIPS*.

[19] Lucas Maystre and Matthias Grossglauser. 2017. ChoiceRank: Identifying Preferences from Node Traffic in Networks. In *ICML*.

[20] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. 1995. An Effective Hash-based Algorithm for Mining Association Rules. In *SIGMOD*. 175–186.

[21] Stephen Ragain and Johan Ugander. 2016. Pairwise Choice Markov Chains. In *NIPS*.

[22] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. 2017. Deep learning with sets and point clouds. In *ICLR Workshop Track*.

[23] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning* 85, 3 (2011), 333–359.

[24] Michel Regenwetter and Bernard Grofman. 1998. Choosing subsets: a size-independent probabilistic model and the quest for a social welfare ordering. *Social Choice and Welfare* 15, 3 (1998), 423–443.

[25] Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *NIPS*.

[26] Gary J Russell and Ann Petersen. 2000. Analysis of cross category dependence in market basket selection. *Journal of Retailing* 76, 3 (2000), 367–392.

[27] Ashoka Savasere, Edward Omiecinski, and Shamkant B Navathe. 1995. An Efficient Algorithm for Mining Association Rules in Large Databases. In *VLDB*.

[28] J Ben Schafer, Joseph Konstan, and John Riedl. 1999. Recommender systems in e-commerce. In *EC*. 158–166.

[29] Nihar Shah, Dengyong Zhou, and Yuval Peres. 2015. Approval voting and incentives in crowdsourcing. In *ICML*.

[30] Or Sheffet, Nina Mishra, and Samuel Ieong. 2012. Predicting Consumer Behavior in Commerce Search. In *ICML*.

[31] Stefan Stremersch and Gerard J Tellis. 2002. Strategic bundling of products and prices: A new synthesis for marketing. *Journal of Marketing* 66, 1 (2002), 55–72.

[32] Kenneth Train. 1986. *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*. Vol. 10. MIT press.

[33] Kenneth E Train. 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press.

[34] Grigorios Tsoumakas and Ioannis Vlahavas. 2007. Random k-labelsets: An ensemble method for multilabel classification. In *ECML*. 406–417.

[35] Iris van Rooij, Ulrike Stege, and Helena Kadlec. 2005. Sources of complexity in subset choice. *Journal of Mathematical Psychology* 49, 2 (2005), 160–187.

[36] Shuang-Hong Yang, Bo Long, Alexander J Smola, Hongyuan Zha, and Zhaohui Zheng. 2011. Collaborative competitive filtering: learning recommender using context of user choice. In *SIGIR*. 295–304.

[37] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. 2017. Deep Sets. In *NIPS*.

[38] Danqing Zhang, Kimon Fountoulakis, Junyu Cao, Mogeng Yin, Michael Mahoney, and Alexei Pozdnoukhov. 2017. Social Discrete Choice Models. *arXiv* (2017).