A Data-Driven Approach to Modeling Choice

Vivek F. Farias **Srikanth Jagabathula** **Devavrat Shah***

Abstract

We visit the following fundamental problem: For a 'generic' model of consumer choice (namely, distributions over preference lists) and a limited amount of data on how consumers actually make decisions (such as marginal preference information), how may one predict revenues from offering a particular assortment of choices? This problem is central to areas within operations research, marketing and econometrics. We present a framework to answer such questions and design a number of tractable algorithms (from a data and computational standpoint) for the same.

1 Introduction

Consider a seller who must pick from a universe of N products, $\mathcal{N}$, a subset $\mathcal{M}$ of products to offer to his customers. The ith product has price p_i. Given a probabilistic model of how customers make choices, $\mathbb{P}(\cdot|\cdot)$, where $\mathbb{P}(i|\mathcal{M})$ is the probability that a potential customer purchases product i when faced with options $\mathcal{M}$, the seller may solve

$$\max_{\mathcal{M}\subset\mathcal{N}} \sum_{i\in\mathcal{M}} p_i \mathbb{P}(i|\mathcal{M}). \tag{1}$$

In addition to being a potentially non-trivial optimization problem, one faces a far more fundamental obstacle here: specifying the 'choice' model $\mathbb{P}(\cdot|\cdot)$ is a difficult task and it is unlikely that a seller will have sufficient data to estimate a generic such model. Thus, simply predicting expected revenues, $R(\mathcal{M}) = \sum_{i\in\mathcal{M}} p_i \mathbb{P}(i|\mathcal{M})$, for a given offer set, $\mathcal{M}$, is a difficult task. This problem, and variants thereof, are central in the fields of marketing, operations research and econometrics. With a few exceptions, the typical approach to dealing with the challenge of specifying a choice model with limited data has been to make parametric assumptions on the choice model that allow for its estimation from a limited amount of data. This approach has a natural deficiency – the implicit assumptions made in specifying a parametric model of choice may not hold. Indeed, for one of the most commonly used parametric models in modern practice (the multinomial logit), it is a simple task to come up with a list of deficiencies, ranging from serious economic fallacies presumed by the model ([5]), to a lack of statistical fit to observed data for real-world problems ([1, 8]). These issues have led to a proliferation of increasingly arcane parametric choice models.

The present work considers the following question: given a limited amount of data on customer preferences and assuming only a 'generic' model of customer choice, what can one predict about expected revenues from a given set of products? We take as our 'generic' model of customer choice, the set of distributions over all possible customer preference lists (i.e. all possible permutations of $\mathcal{N}$). We will subsequently see that essentially all extant models of customer choice can be viewed as a special case of this generic model. We view 'data' as some linear transformation of the distribution specifying the choice model, yielding marginal information. Again, we will see that this view is consistent with reality.

*VF, DS are affiliated with ORC; VF with Sloan School of Management; SJ and DS with LIDS and Department of EECS at MIT. Emails: `vfarias, jskanth, devavrat@mit.edu`. The work was supported in part by NSF CAREER CNS 0546590.

Given these views, we first consider finding the 'simplest' choice model, consistent with the observed marginal data on customer preferences. Here we take as our notion of simple, a distribution over permutations of $\mathcal{N}$ with the *sparsest* support. We present two simple abstract properties that if satisfied by the 'true' choice model, allow us to solve the sparsest fit problem exactly via a simple combinatorial procedure (Theorem 2). In fact, the sparsest fit in this case coincides with the true model (Theorem 1). We present a generative family of choice models that illustrates when the two properties we posit may be expected to hold (see Theorem 3). More generally, when we may not anticipate the above abstract properties, we seek to find a 'worst-case' distribution consistent with the observed data in the sense that this distribution yields minimum expected revenues for a given offer set $\mathcal{M}$ while remaining consistent with the observed marginal data. This entails solving mathematical programs with as many variables as there are permutations ($N!$). In spite of this, we present a simple efficient procedure to solve this problem that is exact for certain interesting types of data and produces approximations (and computable error bounds) in general. Finally, we present a computational study illustrating the efficacy of our approach relative to a parametric technique on a real-world data set.

Our main contribution is thus a novel approach to modeling customer choice given limited data. The approach we propose is complemented with efficient, implementable algorithms. These algorithms yield subroutines that make non-parametric revenue predictions for any given offer set (i.e. predict $R(\mathcal{M})$ for any $\mathcal{M}$) given limited data. Such subroutines could then be used in conjunction with generic set-function optimization heuristics to solve (1).

Relevant Literature: There is a vast body of literature on the parametric modeling of customer choice; a seminal paper in this regard is [10]. See also [14] and references therein for an overview of the area with an emphasis on applications. There is a stream of research (eg. [6]) on estimating and optimizing (parametric) choice models when products possess measurable attributes that are the sole influencers of choice; we do not assume the availability of such attributes and thus do not consider this situation here. A non-parametric approach to choice modeling is considered by [12]; that work studies a somewhat distinct pricing problem, and assumes the availability of a specific type of rich observable data. Fitting a sparsest model to observable data has recently become of great interest in the area of compressive sensing in signal processing [3, 7], and in the design of sketches for streaming algorithms [2, 4]. This work focuses on deriving precise conditions on the support size of the true model, which, when satisfied, guarantee that the sparsest solution is indeed the true solution. However, these prior methods do not apply in the present context (see [9]); therefore, we take a distinct approach to the problem in this paper.

2 The Choice Model and Problem Formulations

We consider a universe of N products, $\mathcal{N} = \{0, 1, 2, \ldots, N-1\}$. We assume that the 0th product in $\mathcal{N}$ corresponds to the 'outside' or 'no-purchase' option. A consumer is associated with a permutation σ of the elements of $\mathcal{N}$; the customer prefers product i to product j iff $\sigma(i) < \sigma(j)$. Given that the customer is faced with a set of alternatives $\mathcal{M} \subset \mathcal{N}$, she chooses to purchase her single most preferred product among those in $\mathcal{M}$. In particular, she purchases $\operatorname{argmin}_{i \in \mathcal{M}} \sigma(i)$.

Choice Model: We take as our model of customer choice a distribution, $\lambda : S_N \to [0, 1]$, over all possible permutations (i.e. the set of all permutations S_N). Define the set

$$\mathcal{S}_j(\mathcal{M}) = \{\sigma \in S_N : \sigma(j) < \sigma(i), \forall i \in \mathcal{M}, i \neq j\}$$

as the set of all customer types that would result in a purchase of j when the offer set is $\mathcal{M}$. Our choice model is thus

$$\mathbb{P}(j|\mathcal{M}) = \sum_{\sigma \in \mathcal{S}_j(\mathcal{M})} \lambda(\sigma) \triangleq \lambda^j(\mathcal{M}).$$

This model subsumes a vast body of extant parametric choice models.

Revenues: We associate every product in $\mathcal{N}$ with a retail price p_j. Of course, $p_0 = 0$. The expected revenues to a retailer from offering a set of products $\mathcal{M}$ to his customers under our choice model is thus given by $R(\mathcal{M}) = \sum_{j \in \mathcal{M}} p_j \lambda^j(\mathcal{M})$.

Data: A seller will have limited data with which to estimate λ. We simply assume that the data observed by the seller is given by an m-dimensional 'partial information' vector $y = A\lambda$, where $A \in \{0,1\}^{m \times N!}$ makes precise the relationship between the observed data and the underlying choice model. For the purposes of illustration, we consider the following concrete examples of data vectors y:

Ranking Data: This data represents the fraction of customers that rank a given product i as their rth choice. Here the partial information vector y is indexed by i, r with $0 \leq i, r \leq N$. For each i, r, y_{ri} denotes the probability that product i is ranked at position r. The matrix A is thus in $\{0,1\}^{N^2 \times N!}$. For a column of A corresponding to the permuation σ, $A(\sigma)$, we will thus have $A(\sigma)_{ri} = 1$ iff $\sigma(i) = r$.

Comparison Data: This data represents the fraction of customers that prefer a given product i to a product j. The partial information vector y is indexed by i, j with $0 \leq i, j \leq N; i \neq j$. For each i, j, $y_{i,j}$ denotes the probability that product i is preferred to product j. The matrix A is thus in $\{0,1\}^{N(N-1) \times N!}$. A column of A, $A(\sigma)$, will thus have $A(\sigma)_{ij} = 1$ if and only if $\sigma(i) < \sigma(j)$.

Top Set Data: This data refers to a concatenation of the "Comparison Data" and information on the fraction of customers who have a given product i as their topmost choice for each i. Thus $A^\top = [A_1^\top A_2^\top]$ where A_1 is simply the A matrix for comparison data, and $A_2 \in \{0,1\}^{N \times N!}$ has $A_2(\sigma)_i = 1$ iff $\sigma(i) = 1$.

Many other types of data vectors consistent with the above view are possible; all we anticipate is that the dimension of the observed data m is substantially smaller than $N!$. We are now in a position to formulate the questions broached in the previous section precisely:

"Simplest" Model: In finding the simplest choice model consistent with the observed data we attempt to solve:

$$\text{minimize} \|\lambda\|_0 \quad \text{subject to} \quad A\lambda = y, \quad \mathbf{1}^\top \lambda = 1, \quad \lambda \geq 0. \tag{2}$$

Robust Approach: For a given offer set $\mathcal{M}$, and data vector y, what are the minimal expected revenues we might expect from $\mathcal{M}$ consistent with the observed data? To answer this question, we attempt to solve :

$$\underset{\lambda}{\text{minimize}} \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \quad \text{subject to} \quad A\lambda = y, \quad \mathbf{1}^\top \lambda = 1, \quad \lambda \geq 0. \tag{3}$$

3 Estimating Sparse Choice Models

Here we consider finding the sparsest model consistent with the observed data (i.e. problem (2)). We face two questions: (a) Why is sparsity an interesting criterion? (b) Is there an efficient procedure to solve the program in (2)? We begin by identifying two simple conditions that define a class of choice models (i.e. a class of distributions λ). Assuming that the 'true' underlying model λ belongs to this class, we prove that the sparsest model (i.e the solution to (2)) is in fact this true model. This answers the first question. We then propose a simple procedure inspired by [9] that correctly solves the program in (2) assuming these conditions. It is difficult to expect the program in (2) to recover the true solution in general (see [9] for a justification). Nonetheless, we show that the conditions we impose are not overly restrictive: we prove that a "sufficiently" sparse model generated uniformly at random from the set of all possible choice models satisfies the two conditions with a high probability.

Before we describe the conditions we impose on the true underlying distribution, we introduce some notation. Let λ denote the true underlying distribution, and let K denote the support size, $\|\lambda\|_0$. Let $\sigma_1, \sigma_2, \ldots, \sigma_K$ denote the permutations in the support, i.e, $\lambda(\sigma_i) \neq 0$ for $1 \leq i \leq K$, and $\lambda(\sigma) = 0$ for all $\sigma \neq \sigma_i, 1 \leq i \leq K$. Recall that y is of dimension m and we index its elements by d. The two conditions we impose are as follows:

Signature Condition: For every permutation σ_i in the support, there exists a $d(i) \in \{1, 2, \ldots, m\}$ such that $A(\sigma_i)_{d(i)} = 1$ and $A(\sigma_j)_{d(i)} \neq 0$, for every $j \neq i$ and $1 \leq i, j \leq K$. In other words, for each permutation σ_i in the support, $y_{d(i)}$ serves as its 'signature'.

Linear Independence Condition: $\sum_{i=1}^{K} c_i \lambda(\sigma_i) \neq 0$, for any $c_i \in \mathbb{Z}$ and $|c_i| \leq C$, where $\mathbb{Z}$ denotes the set of integers and C is a sufficiently large number $\geq K$. This condition is satisfied with probability 1 if $[\lambda_1 \lambda_2 \ldots \lambda_K]^\top$ is drawn uniformly from K-dim simplex.

When the two conditions are satisfied, the sparsest solution is indeed the true solution as stated in the following theorem:

Theorem 1. *Suppose we are given $y = A\lambda$ and λ satisfies the "Signature" condition and the "Linear Independence" condition. Then, λ is the unique solution to the program in* (2).

The proof of Theorem 1 is given in the appendix. Next we describe the algorithm we propose for recovery. The algorithm takes y and A and as the input and outputs λ_i (denotes $\lambda(\sigma_i)$) and $A(\sigma_i)$ for every permutation σ_i in the support. The algorithm assumes the observed values y_d are sorted. Therefore, without loss of generality, assume that $y_1 < y_2 < \ldots < y_m$. Then, the algorithm is as follows:

Algorithm:

Initialization: $\lambda_0 = 0, k(0) = 0$ and $A(\sigma_i)_d = 0, 1 \leq i \leq K$ and $1 \leq d \leq m$.
for $d = 1$ to m
  if $y_d = \sum_{i \in T} \lambda_i$ for some $T \subseteq \{1, \ldots, k(d-1)\}$
    $k(d) = k(d-1), \qquad A(\sigma_i)_d = 1 \quad \forall \quad i \in T$
  else
    $k(d) = k(d-1) + 1, \qquad \lambda_{k(d)} = y_d, \qquad A(\sigma_{k(d)})_d = 1,$
  end if
end for
Output $K = k(m)$ and $(\lambda_i, A(\sigma_i)), 1 \leq i \leq K$.

Now, we have the following theorem:

Theorem 2. *Suppose we are given $y = A\lambda$ and λ satisfies the "signature" and the "linear independence" conditions. Then, the above described algorithm recovers λ.*

Theorem 2 is proved in the appendix. The algorithm we have described either succeeds in finding a valid λ or else determines that the two properties are not satisfied. We now show that the conditions we have imposed do not restrict the class of plausible models severely. For this, we show that models drawn from the following generative model satisfy the conditions with high probability.

Generative Model. Given K and an interval $[a, b]$ on the positive real line, we generate a choice model λ as follows: choose K permutations, $\sigma_1, \sigma_2, \ldots, \sigma_K$, uniformly at random with replacement, choose K numbers uniformly at random from the interval $[a, b]$, normalize the numbers so that they sum to 1, and assign them to the permutations σ_i, $1 \leq i \leq K$. For all other permutations $\sigma \neq \sigma_i$, $\lambda(\sigma) = 0$. Note that, since we are choosing permutations in the support with replacement, there could be repetitions. However, for large N and $K \ll N!$, this happens with a vanishing probability.

Depending on the observed data, we characterize values of sparsity K for which distributions generated by the above generative model can be recovered with a high probability. Specifically, we have the following theorem for the three forms of observed data mentioned in Section 2. The proof may be found in the appendix.

Theorem 3. *Suppose λ is a choice model of support size K drawn from the generative model. Then, λ satisfies the "signature" and "linear independence" conditions with probability $1 - o(1)$ as $N \to \infty$ provided $K = O(N)$ for ranking data, $K = o(log\, N)$ for comparison data, and $K = o(\sqrt{N})$ for the top set data.*

Of course, in general, the underlying choice model may not satisfy the two conditions we have posited or be exactly recoverable from the observed data. In order to deal with this more general scenario, we next propose an approach that implicitly identifies a 'worst-case' distribution consistent with the observed data.

4 Robust Revenue Estimates Consistent with Data

In this section, we propose a general algorithm for the solution of program (3). This LP has $N!$ variables and is clearly not amenable to direct solution; hence we consider its dual. In preparation for taking the dual, let $\mathcal{A}_j(\mathcal{M}) \triangleq \{A(\sigma) : \sigma \in \mathcal{S}_j(\mathcal{M})\}$, where, recall that, $\mathcal{S}_j(\mathcal{M})$ denotes the set of all permutations that result in the purchase of $j \in \mathcal{M}$ when offered the assortment $\mathcal{M}$. Since $S_N = \cup_{j\in\mathcal{M}}\mathcal{S}_j(\mathcal{M})$, we have implicitly specified a partition of the columns of the matrix A. Armed with this notation, the dual of (3) is:

$$\text{(4)} \quad \underset{\alpha,\nu}{\text{maximize}} \quad \alpha^\top y + \nu \quad \text{subject to} \quad \max_{x^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top x^j + \nu\right) \le p_j, \quad \text{for each } \ j \in \mathcal{M}.$$

Our solution procedure will rely on an effective representation of the sets $\mathcal{A}_j(\mathcal{M})$.

4.1 A Canonical Representation of $\mathcal{A}_j(\mathcal{M})$ and its Application

We assume that every set $\mathcal{S}_j(\mathcal{M})$ can be expressed as a disjoint union of D_j sets. We denote the dth such set by $\mathcal{S}_{jd}(\mathcal{M})$ and let $\mathcal{A}_{jd}(\mathcal{M})$ be the corresponding set of columns. Consider the convex hull of the set $\mathcal{A}_{jd}(\mathcal{M})$, $\text{conv}\{\mathcal{A}_{jd}(\mathcal{M})\} \triangleq \bar{\mathcal{A}}_{jd}(\mathcal{M})$. $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ is by definition a polytope contained in the m-dimensional unit cube, $[0,1]^m$. In other words,

$$\text{(5)} \qquad \bar{\mathcal{A}}_{jd}(\mathcal{M}) = \{x^{jd} : A_1^{jd}\, x^{jd} \ge b_1^{jd}, \quad A_2^{jd}\, x^{jd} = b_2^{jd}, \quad A_3^{jd}\, x^{jd} \le b_3^{jd}, \quad x^{jd} \ge 0.\}$$

for appropriately defined $A^{jd}_{\cdot}, b^{jd}_{\cdot}$. By a canonical representation of $\mathcal{A}_j(\mathcal{M})$, we will thus understand a partition of $\mathcal{S}_j(\mathcal{M})$ and a polyhedral representation of the columns corresponding to every set in the partition as given by (5). Ignoring the problem of actually obtaining this representation for now, we assume access to a canonical representation and present a simple program whose size is polynomial in the size of this representation that is equivalent to (3), (4). For simplicity of notation, we assume that each of the polytopes $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ is in standard form, i.e. $\bar{\mathcal{A}}_{jd}(\mathcal{M}) = \{x^{jd} : A^{jd}\, x^{jd} = b^{jd}, \quad x^{jd} \ge 0.\}$. Now since an affine function is always optimized at the vertices of a polytope, we know:

$$\max_{x^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top x^j + \nu\right) = \max_{d, x^{jd} \in \bar{\mathcal{A}}_{jd}(\mathcal{M})} \left(\alpha^\top x^{jd} + \nu\right).$$

We have thus reduced (4) to a 'robust' LP. By strong duality we have:

$$\text{(6)} \quad \max_{x^{jd} \in \bar{\mathcal{A}}_{jd}(\mathcal{M})} \left(\alpha^\top x^{jd} + \nu\right) \triangleq \begin{array}{ll} \underset{x^{jd}}{\text{maximize}} & \alpha^\top x^{jd} + \nu \\ \text{subject to} & A^{jd}\, x^{jd} = b^{jd} \\ & x^{jd} \ge 0. \end{array} = \begin{array}{ll} \underset{\gamma^{jd}}{\text{minimize}} & {b^{jd}}^\top \gamma^{jd} + \nu \\ \text{subject to} & {\gamma^{jd}}^\top A^{jd} \ge \alpha \end{array}$$

We have thus established the following useful equality:

$$\left\{\alpha, \nu : \max_{x^j \in \bar{\mathcal{A}}_j(\mathcal{M})} \left(\alpha^\top x^j + \nu\right) \le p_j\right\} = \left\{\alpha, \nu : {b^{jd}}^\top \gamma^{jd} + \nu \le p_j, {\gamma^{jd}}^\top A^{jd} \ge \alpha, d = 1, 2, \ldots, D_j\right\}.$$

It follows that solving (3) is equivalent to the following LP whose complexity is polynomial in the description of our canonical representation:

$$\text{(7)} \qquad \begin{array}{lll} \underset{\alpha,\nu}{\text{maximize}} & \alpha^\top y + \nu & \\ \text{subject to} & {b^{jd}}^\top \gamma^{jd} + \nu \le p_j & \text{for all } j \in \mathcal{M}, d = 1, 2, \ldots, D_j \\ & {\gamma^{jd}}^\top A^{jd} \ge \alpha & \text{for all } j \in \mathcal{M}, d = 1, 2, \ldots, D_j. \end{array}$$

Our ability to solve (7) relies on our ability to produce an efficient canonical representation of $\mathcal{S}_j(\mathcal{M})$. In what follows, we first consider an example where such a representation is readily available, and then consider the general case.

Canonical Representation for Ranking Data: Recall the definition of *ranking data* from Section 2. Consider partitioning $\mathcal{S}_j(\mathcal{M})$ into N sets wherein the dth set is given by

$\mathcal{S}_{jd}(\mathcal{M}) = \{\sigma \in \mathcal{S}_j(\mathcal{M}) \colon \sigma(j) = d\}$. It is not difficult to show that the set $\mathcal{A}_{jd}(\mathcal{M})$ is equal to the set of all vectors x^{jd} in $\{0,1\}^N$ satisfying:

$$
\begin{array}{ll}
\sum_{i=0}^{N-1} x^{jd}_{ri} = 1 & \text{for } 0 \le i \le N-1 \\
\sum_{r=0}^{N-1} x^{jd}_{ri} = 1 & \text{for } 0 \le r \le N-1 \\
x^{jd}_{ri} \in \{0,1\} & \text{for } 0 \le i, r \le N-1. \\
x^{jd}_{dj} = 1 & \\
x^{jd}_{d'i} = 0 & \text{for all } i \in \mathcal{M}, i \ne j \text{and } 0 \le d' < d.
\end{array} \tag{8}
$$

Our goal is, of course, to find a description for $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ of the type (5). Now consider replacing the third (integrality) constraint in (8) with simply the non-negativity constraint $x^{jd}_{ri} \ge 0$. It is clear that the resulting polytope contains $\bar{\mathcal{A}}_{jd}(\mathcal{M})$. In addition, one may show that the resulting *polytope* has *integral* vertices since it is simply a matching polytope with some variables forced to be integers, so that in fact the polytope is precisely $\bar{\mathcal{A}}_{jd}(\mathcal{M})$, and we have our canonical representation. Further, notice that this representation yields an *efficient* algorithm to solve (3) via (7)!

4.2 Computing a Canonical Representation: Comparison Data

Recall the definition of *comparison data* from Section 2. We use this data as an example to illustrate a general procedure for computing a canonical representation. Consider $\mathcal{S}_j(\mathcal{M})$. It is not difficult to see that the corresponding set of columns $\mathcal{A}_j(\mathcal{M})$ is equal to the set of vectors in $\{0,1\}^{(N-1)N}$ satisfying the following constraints:

$$
\begin{array}{ll}
x^j_{il} \ge x^j_{ik} + x^j_{kl} - 1 & \text{for all } \ i,k,l \in \mathcal{N}, i \ne k \ne l \\
x^j_{ik} + x^j_{ki} = 1 & \text{for all } i,k \in \mathcal{N}, i \ne k \\
x^j_{ji} = 1 & \text{for all } i \in \mathcal{M}, i \ne j \\
x^j_{ik} \in \{0,1\} & \text{for all } i,k \in \mathcal{N}, i \ne k
\end{array} \tag{9}
$$

Briefly, the second constraint follows since for any $i,k, i \ne k$, either $\sigma(i) > \sigma(k)$ or else $\sigma(i) < \sigma(k)$. The first constraint enforces transitivity: $\sigma(i) < \sigma(k)$ and $\sigma(k) < \sigma(l)$ together imply $\sigma(i) < \sigma(l)$. The third constraint enforces that all $\sigma \in S_j(\mathcal{M})$ must satisfy $\sigma(j) < \sigma(i)$ for all $i \in \mathcal{M}$. Now consider the polytope obtained by relaxing the fourth (integrality) constraint to simply $x^j_{ik} \ge 0$. Call this polytope $\bar{\mathcal{A}}^o_j(\mathcal{M})$. Of course, we must have $\bar{\mathcal{A}}^o_j(\mathcal{M}) \supseteq \bar{\mathcal{A}}_j(\mathcal{M})$. Unlike the case of ranking data, however, $\bar{\mathcal{A}}^o_j(\mathcal{M})$ can in fact be shown to be *non*-integral, so that $\bar{\mathcal{A}}^o_j(\mathcal{M}) \ne \bar{\mathcal{A}}_j(\mathcal{M})$ in general. In this case we resort to the following procedure.

[1.] Solve (7) using the representation of $\bar{\mathcal{A}}^o_j(\mathcal{M})$ in place of $\bar{\mathcal{A}}_j(\mathcal{M})$. This yields a lower bound on (3) since $\bar{\mathcal{A}}^o_j(\mathcal{M}) \supset \bar{\mathcal{A}}_j(\mathcal{M})$. Call the corresponding solution $\alpha_{(1)}, \nu_{(1)}$.

[2.] Solve the optimization problem $\max \alpha^\top_{(1)} x^j$ subject to $x^j \in \bar{\mathcal{A}}^o_j(\mathcal{M})$ for each j. If the optimal solution $\hat{x}^j$ is integral for each j, then stop; the solution computed in the first step is in fact optimal.

[3.] Let $\hat{x}^j_{ik}$ be a non-integral variable. Partition $\mathcal{S}_j(\mathcal{M})$ on this variable - i.e. define $\mathcal{S}_{j1}(\mathcal{M}) = \{\sigma : \sigma \in \mathcal{S}_j(\mathcal{M}), \sigma(i) < \sigma(k)\}$ and $\mathcal{S}_{j2}(\mathcal{M}) = \{\sigma : \sigma \in \mathcal{S}_j(\mathcal{M}), \sigma(i) > \sigma(k)\}$. Define outer-approximations to $\bar{\mathcal{A}}_{j1}(\mathcal{M})$ and $\bar{\mathcal{A}}_{j2}(\mathcal{M})$ as the projection of $\bar{\mathcal{A}}^o_j(\mathcal{M})$ on $x^j_{ik} = 1$ and $x^j_{ik} = 0$ respectively. Go to step 1.

The above procedure is finite, but the size of the LP we solve at each iteration doubles. Nonetheless, each iteration produces a lower bound to (3) whose quality is easily measured (for instance, by solving the maximization version of (3) using the same procedure), and this quality improves with each iteration. In our computational experiments with a related type of data, it sufficed to stop after a single iteration.

5 An Empirical Evaluation of the Approach

We have presented simple sub-routines to estimate the revenues $R(\mathcal{M})$ from a particular offer set $\mathcal{M}$, given marginal preference data y. These sub-routines are effectively 'non-parametric' and can form the basis of a procedure that solves the revenue optimization problem posed in the introduction. Here we seek to contrast this approach with a commonly used parametric approach. We consider two types of observable data: ranking data and a 'censored' version of the comparison data which gives us for every pair of products $i, j, \neq 0$, the fraction of customers that prefer i to j, and in addition prefer i to 0 (i.e. not buying). The latter type of data is quite realistic.

The parametric recipe we consider is the following: One fits a *Multinomial Logit* (MNL) model to the observable data and picks an optimal offer set by evaluating $R(\mathcal{M}) = \sum_{j \in \mathcal{M}} p_j \mathbb{P}(j|\mathcal{M})$ assuming $\mathbb{P}(\cdot|\mathcal{M})$ follows the estimated model. The MNL is a commonly used parametric model that associates with each product i in $\mathcal{N}$ a positive scalar w_i; $w_0 = 1$ by convention. The model assumes $\mathbb{P}(i|\mathcal{M}) = w_i / \sum_{j \in \mathcal{M}} w_j$. In place of making this parametric assumption, we could instead evaluate $R(\mathcal{M})$ using the robust sub-routine developed in the previous section and pick $\mathcal{M}$ to maximize this conservative estimate. It is clear that if the MNL model is a poor fit to the true choice model, $\mathbb{P}$, our robust approach is likely to outperform the parametric approach substantially. Instead, what we focus on here is what happens if the MNL model is a *perfect* fit to the true choice model. In this case, the parametric approach is the best possible. How sub-optimal is our non-parametric approach here?

We consider an MNL model on $N = 25$ products. The model and prices were specified using customer utilities for Amazon.com's highest selling DVDs (and their prices) during a 3-month period from 1 July 2005 to 30 September 2005 estimated by [13] [1]. We generate synthetic observed data (of both the ranking type and the comparison type) *according* to this fitted MNL model. This represents a scenario where the fitted MNL is a perfect descriptor of reality. We conduct the following experiments:

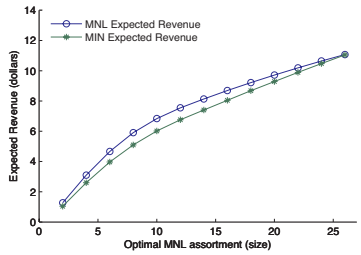
Quality of Revenue Predictions: For each type of observable data we compute our estimate of the minimum value that $R(\mathcal{M})$ can take on, consistent with that data, by solving (3). We compare this with the value of $R(\mathcal{M})$ predicted under the MNL model (which in this case, is exact). Figures 1(b) and 1(d) compare these two quantities for a set of randomly chosen subsets $\mathcal{M}$ of the 25 potential DVD's assuming ranking data and the censored comparison data respectively. In both cases, our procedure produces excellent predictions of expected revenue without making the assumptions on $\mathbb{P}(\cdot|\cdot)$ inherent in the MNL model.

Quality of Optimal Solutions to Revenue Maximization Problems: For each type of observable data, we compute optimal offer sets $\mathcal{M}$ of varying capacities assuming the fitted MNL model and an optimization procedure described in [13]. We then evaluate the revenue predictions for these optimal offer sets by solving (3). Figures 1(a) and 1(c) plot these estimates for the two types of observable data. The gap between the 'MNL' and the 'MIN' curves is thus an upper bound on the expected revenue loss if one used our non-parametric procedure to pick an optimal offer set $\mathcal{M}$ over the parametric procedure (which in this setting is optimal). Again, we see that the revenue loss is surprisingly small.

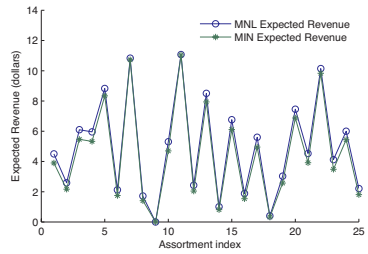
6 Conclusion and Potential Future Directions

We have presented a general framework that allows us to answer questions related to how consumers choose among alternatives using limited observable data and without making additional parametric assumptions. The approaches we have proposed are feasible from a data availability standpoint as well as a computational standpoint and provide a much needed non-parametric 'sub-routine' for the revenue optimization problems described at the outset. This paper also opens up the potential for a stream of future work.

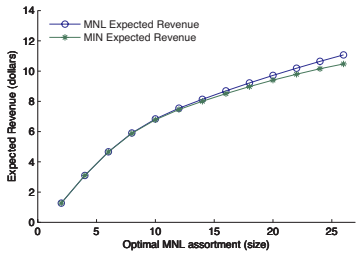
[1] The problem of optimizing over $\mathcal{M}$ is particularly relevant to Amazon.com given limited screen real-estate and cannibilization effects



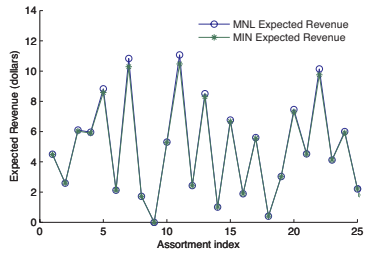
(a) Ranking Data: Optimal $\mathcal{M}$



(b) Ranking Data: Random $\mathcal{M}$



(c) Comparison Data: Optimal $\mathcal{M}$



(d) Comparison Data: Random $\mathcal{M}$

References

[1] K. Bartels, Y. Boztug, and M. M. Muller. Testing the multinomial logit model. Working Paper, 1999.

[2] R. Berinde, AC Gilbert, P. Indyk, H. Karloff, and MJ Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. *Preprint*, 2008.

[3] E.J. Candes, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8), 2006.

[4] G. Cormode and S. Muthukrishnan. Combinatorial algorithms for compressed sensing. *Lecture Notes in Computer Science*, 4056:280, 2006.

[5] G. Debreu. Review of r.d. luce, 'individual choice behavior: A theoretical analysis'. *American Economic Review*, 50:186–188, 1960.

[6] G. Dobson and S. Kalish. Positioning and pricing a product line. *Marketing Science*, 7(2):107–125, 1988.

[7] DL Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[8] J. L. Horowitz. Semiparametric estimation of a work-trip mode choice model. *Journal of Econometrics*, 58:49–70, 1993.

[9] S. Jagabathula and D. Shah. Inferring rankings under constrained sensing. In *NIPS*, pages 7–1, 2008.

[10] D. McFadden. Econometric models for probabiistic choice among products. *The Journal of Business*, 53(3):S13–S29, 1980.

[11] EH McKinney. Generalized birthday problem. *American Mathematical Monthly*, pages 385–387, 1966.

[12] P. Rusmevichientong, B. Van Roy, and P. Glynn. A nonparametric approach to multi-product pricing. *Operations Research*, 54(1), 2006.

[13] P. Rusmevichientong, ZJ Shen, and D.B. Shmoys. Dynamic Assortment Optimization with a Multinomial Logit Choice Model and Capacity Constraint. Technical report, Working Paper, 2008.

[14] Kalyan T. Talluri and Garrett J. van Ryzin. *The Theory and Practice of Revenue Management*. Springer Science+Business Media, 2004.