MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

A Non-parametric Approach to Modeling Choice with Limited Data

Vivek F. Farias
MIT Sloan, vivekf@mit.edu

Srikanth Jagabathula
NYU Stern, sjagabat@stern.nyu.edu

Devavrat Shah
EECS, MIT, devavrat@mit.edu

Choice models are today ubiquitous across a range of applications in operations and marketing. Real world implementations of many of these models face the formidable stumbling block of simply identifying the 'right' model of choice to use. Since models of choice are inherently high dimensional objects, the typical approach to dealing with this problem is positing, a-priori, a parametric model that one believes adequately captures choice behavior. This approach can be substantially sub-optimal in scenarios where one cares about using the choice model learned to make fine-grained predictions; one must contend with the risks of mis-specification and over/under-fitting.

Thus motivated, we visit the following 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 information about these distributions), how may one predict revenues from offering a particular assortment of choices? An outcome of our investigation is a *non-parametric* approach in which the data automatically selects the 'right' choice model for revenue predictions. The approach is practical. Using a data set consisting of automobile sales transaction data from a major US automaker, our method demonstrates a 20% improvement in prediction accuracy over state-of-the art benchmark models, which can result in a 10% increase in revenues from optimizing the offer set. We also address a number of theoretical issues, among them a qualitative examination of the choice models implicitly learned by the approach. We believe that this paper takes a step towards 'automating' the crucial task of choice model selection.

1. Introduction

A problem of central interest to operations managers is to predict the revenues or sales from offering a particular assortment of products to customers using historical sales data. As one can imagine, such predictions form crucial inputs to several important business decisions, both operational and otherwise. A classical example of such a decision problem is that of assortment planning: deciding the "optimal" assortment of products to offer customers with a view to maximizing expected revenues (or some related objective) subject to various constraints (e.g. limited display or shelf space). A number of variants of this problem, both static and dynamic, arise in essentially every facet of revenue management. Such problems are seen as crucial revenue management tasks and needless to say, accurate revenue or sales predictions fundamentally impact how well we can perform such tasks.

Why might these crucial predictions be difficult to make? Consider the task of predicting expected sales rates from offering a particular set of products to customers. In industry jargon, this is referred to as 'conversion-rate', and is defined as the probability of converting an arriving customer into a purchasing customer. Predicting the conversion-rate for an offer set is difficult because the probability of purchase of each product depends on all the products on offer. This is due to substitution behavior,

where an arriving customer potentially substitutes an unavailable product with an available one. Due to substitution, the sales observed for a product may be viewed as a combination of its 'primary' demand and additional demand due to substitution. *Customer Choice Models* have been used to model this behavior with success. At an abstract level, a choice model can be thought of as a conditional probability distribution that for any offer set yields the probability that an arriving customer purchases a given product in that set.

There is *vast* literature spanning marketing, economics, and psychology devoted to the construction of parametric choice models and their estimation from data. In the literature that studies the sorts of revenue management decision problems we alluded to above, such models are typically assumed *given*. The implicit understanding is that a complete prescription for these decision problems will require fitting the "right" parametric choice model to data, so as to make accurate revenue or sales predictions. This is a complex task. Apart from the fact that one can never be sure that the chosen parametric structure is a "good" representation of the underlying ground truth, parametric models are prone to over-fitting and under-fitting issues. Once a structure is *fixed*, one does not glean new structural information from data. This is a serious issue in practice because although a simple model (such as the multinomial logit (MNL) model) may make practically unreasonable assumptions (such as the so-called "IIA" assumption), fitting a more complex model can lead to worse performance due to over-fitting – and one can never be sure.

In this paper, we propose a *non-parametric, data-driven* approach to making revenue or sales predictions that affords the revenue manager the opportunity to avoid the challenging task of fitting an appropriate parametric choice model to historical data. Our approach views choice models *generically*, namely as distributions over rankings (or preference lists) of products. As shall be seen subsequently, this view subsumes essentially all extant choice models. Further, this view yields a *nonparametric* approach to choice modeling where the revenue manager does not need to think about the appropriate parametric structure for his problem, or the tradeoff between model parsimony and the risk of over-fitting. Rather, through the use of a nonparametric approach, our goal is to offload as much of this burden as possible to the data itself.

1.1. Contributions

As mentioned above, we consider entirely generic models of choice, specified as a distribution over all possible rankings (or preference lists) of products. Our view of data is aligned with what one typically has available in reality – namely, sales rates of products in an assortment, for some set of product assortments. This is a general view of choice modeling. Our main contribution is to make this view operational, yielding a *data-driven, nonparametric* approach. Specifically, we make the following contributions in the context of this general setup:

- **Revenue Predictions:** As mentioned above, accurate revenue or sales predictions form core inputs for a number of important revenue/ inventory management problems. Available sales data will typically be insufficient to fully specify a generic model of choice of the type we consider. We therefore seek to identify the *set* of generic choice models consistent with available sales data. Given the need to make a revenue or sales prediction on a heretofore unseen assortment, we then offer the *worst-case* expected revenue possible for that assortment assuming that the true model lies in the set of models found to be consistent with observed sales data. Such an approach makes no a-priori structural assumptions on the choice model, and has the appealing feature that as more data becomes available, the predictions will improve, by narrowing down the set of consistent models. This simple philosophy dictates challenging computational problems; for instance, the sets we compute are computationally unwieldy and, at first glance, highly intractable. Nonetheless, we successfully develop several simple algorithms of increasing sophistication to address these problems.
- **Empirical Evaluation:** We conducted an empirical study to gauge the practical value of our approach, both in terms of the absolute quality of the predictions produced, and also relative to using alternative parametric approaches. We describe the results of two such studies:

(i) *Simulation Study:* The purpose of our simulation study is to demonstrate that the robust approach can effectively capture model structure consistent with a number of different parametric models and produce good revenue predictions. The general setup in this study was as follows: We use a parametric model to generate synthetic transaction data. We then use this data in conjunction with our revenue prediction procedure to predict expected revenues over a swathe of offer sets. Our experimental design permits us to compare these predictions to the corresponding 'ground truth'. The parametric families we considered included the multinomial logit (MNL), nested logit (NL), and mixture of multinomial logit (MMNL) models. In order to 'stress-test' our approach, we conducted experiments over a wide range of parameter regimes for these generative parametric choice models, including some that were fit to DVD sales data from Amazon.com. The predictions produced are remarkably accurate.

(ii) *Empirical Study with Sales Data from a Major US Automaker:* The purpose of our empirical study is two-fold: (1) to demonstrate how our setup can be applied with real-world data, and (2) to pit the robust method in a "horse-race" against the MNL and MMNL parametric families of models. For the case study, we used sales data collected daily at the dealership level over 2009 to 2010 for a range of small SUVs offered by a major US automaker for a dealership zone in the Midwest. We used a portion of this sales data as 'training' data. We made this data available to our robust approach, as well as in the fitting of an MNL model and an MMNL model. We tested the quality of 'conversion-rate' predictions (i.e. a prediction of the sales rate given the assortment of models on the

lot) using the robust approach and the incumbent parametric approaches on the remainder of the data. We conducted a series of experiments by varying the amount of training data made available to the approaches. We conclude that (a) the robust method improves on the accuracy of either of the parametric methods by about 20% (this is large) in all cases and (b) unlike the parametric models, the robust method is apparently not susceptible to under-fitting and over-fitting issues. In fact, we see that the performance of the MMNL model relative to the MNL model deteriorates as the amount of training data available decreases due to over-fitting. Improved forecast accuracy improves the decisions made. For instance, a 20% improvment in forecast accuracy can result in a 10% increase in revenues from optimizing the offer set.

- **Descriptive Analysis:** In making revenue predictions, we did not need to concern ourselves with the choice model implicitly assumed by our prediction procedure. This fact notwithstanding, it is natural to consider criteria for selecting choice models consistent with the observed data that are independent of any decision context. Thus motivated, we consider the natural task of finding the *simplest* choice model consistent with the observed data. As in much of contemporary high dimensional statistics, we employ *sparsity*[1] as our measure of simplicity. To begin, we use the sparsest fit criterion to obtain a characterization of the choice models implicitly used by the robust revenue prediction approach. Loosely speaking, we show that the choice model implicitly used by the robust approach is essentially the sparsest model (Theorem 1) and the complexity of the model (as measured by its sparsity) scales with the "amount" of data. This provides an explanation for the immunity of the robust approach to over/under fitting as observed in our case study. Second, we characterize the family of choice models that can be identified only from observed marginal data via the sparsest fit criterion (Theorems 2 and 3). Our characterization formalizes the notion that the complexity of the models that can be identified via the sparsest fit criterion scales with the "amount" of data at hand.

1.2. Relevant Literature

The study of choice models and their applications spans a vast literature across multiple fields including at least Marketing, Operations and Economics. In disciplines such as marketing learning a choice model is an interesting goal unto itself given that it is frequently the case that a researcher wishes to uncover "why" a particular decision was made. Within operations, the goal is frequently more application oriented with the choice model being explicitly used as a predictive tool within some larger decision model. Since our goals are aligned with the latter direction, our literature review focuses predominantly on OM; we briefly touch on key work in Marketing. We note that our consideration of 'sparsity' as an appropriate non-parametric model selection criterion is closely

[1] By sparsity we refer to the number of rank lists or, in effect, customer types, assumed to occur with positive probability in the population.

related to the burgeoning statistical area of compressive sensing; we discuss those connections in a later Section.

The vast majority of decision models encountered in operations have traditionally ignored substitution behavior (and thereby choice modeling) altogether. Within airline RM, this is referred to as the "independent demand" model (see Talluri and van Ryzin (2004b)). Over the years, several studies have demonstrated the improvements that could be obtained by incorporating choice behavior into operations models. For example, within airline RM, the simulation studies conducted by Belobaba and Hopperstad (1999) on the well known passenger origin and destination simulator (PODS) suggested the value of corrections to the independent demand model; more recently, Ratliff et al. (2008a) and Vulcano et al. (2010) have demonstrated valuable average revenue improvements from using MNL choice-based RM approaches using real airline market data. Following such studies, there has been a significant amount of research in the areas of inventory management and RM attempting to incorporate choice behavior into operations models.

The bulk of the research on choice modeling in both the areas has been optimization related. That is to say, most of the work has focused on devising optimal decisions *given* a choice model. Talluri and van Ryzin (2004a), Gallego et al. (2006), van Ryzin and Vulcano (2008), Mahajan and van Ryzin (1999), Goyal et al. (2009) are all papers in this vein. Kök et al. (2008) provides an excellent overview of the state-of-the-art in assortment optimization. Rusmevichientong et al. (2008) consider the multinomial logit (MNL) model and provide an efficient algorithm for the static assortment optimization problem and propose an efficient policy for the dynamic optimization problem. A follow on paper, Rusmevichientong and Topaloglu (2011), considers the same optimization problem but where the mean utilities in the MNL model are allowed to lie in some arbitrary uncertainty set. Saure and Zeevi (2009) propose an alternative approach for the dynamic assortment optimization problem under a general random utility model.

The majority of the work above focuses on optimization issues given a choice model. Paper such as Talluri and van Ryzin (2004a) discuss optimization problems with *general* choice models, and as such our revenue estimation procedure fits in perfectly there. In most cases, however, the choice model is assumed to be given and of the MNL type. Papers such as Saure and Zeevi (2009) and Rusmevichientong and Topaloglu (2011) loosen this requirement by allowing some amount of *parametric* uncertainty. In particular, Saure and Zeevi (2009) assume unknown mean utilities and learn these utilities, while the optimization schemes in Rusmevichientong and Topaloglu (2011) require knowledge of mean utilities only within an interval. In both cases, the structure of the model (effectively, MNL) is *fixed* up front.

The MNL model is by far the most popular choice model studied and applied in OM. The origins of the MNL model date all the way back to the Plackett-Luce model, proposed independently by Luce (1959) and Plackett (1975). Before becoming popular in the area of OM, the MNL model found widespread use in the areas of transportation (see seminal works of McFadden (1980), Ben-Akiva and Lerman (1985)) and marketing (starting with the seminal work of Guadagni and Little (1983), which paved the way for choice modeling using scanner panel data). See Wierenga (2008), Chandukala et al. (2008) for a detailed overview of choice modeling in the area of Marketing. The MNL model is popular because its structure makes it tractable both in terms of estimating its parameters and solving decision problems. However, the tractability of the MNL model comes at a cost: it is incapable of capturing any heterogeneity in substitution patterns across products (see Debreu (1960)) and suffers from Independent of Irrelevant Alternatives (IIA) property (see Ben-Akiva and Lerman (1985)), both of which limit its practical applicability.

Of course, these issues with the MNL model are well recognized, and far more sophisticated models of choice have been suggested in the literature (see, for instance, Ben-Akiva and Lerman (1985), Anderson et al. (1992)); the price one pays is that the more sophisticated models may not be easily identified from sales data and are prone to over-fitting. It must be noted that an exception to the above state of affairs is the paper by Rusmevichientong et al. (2006) that considers a general nonparametric model of choice similar to the one considered here in the context of an assortment pricing problem. The caveat is that the approach considered requires access to samples of entire customer preference lists which are unlikely to be available in many practical applications.

Our goal relative to all of the above work is to *eliminate* the need for structural assumptions and thereby, the associated risks as well. We provide a means of going directly from raw sales transaction data to revenue or sales estimates for a given offer set. While this does not represent the entirety of what can be done with a choice model, it represent a valuable application, at least within the operational problems discussed.

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 customer is associated with a permutation (or ranking) σ of the products in $\mathcal{N}$; the customer prefers product i to product j if and only if $\sigma(i) < \sigma(j)$. A customer will be presented with a set of alternatives $\mathcal{M} \subset \mathcal{N}$; any set of alternatives will, by convention, be understood to include the no-purchase alternative i.e. the 0th product. The customer will subsequently choose to purchase her single most preferred product among those in $\mathcal{M}$. In particular, she purchases

$$\arg\min_{i \in \mathcal{M}} \sigma(i). \tag{1}$$

It is quickly seen that the above structural assumption is consistent with structural assumptions made in commonly encountered choice models including the multinomial logit, nested multinomial logit, or more general random utility models. Those models make many additional structural assumptions which may or may not be reasonable for the application at hand. Viewed in a different light, basic results from the theory of social preferences dictate that the structural assumptions implicit in our model are no more restrictive than assuming that the customer in question is endowed with a utility function over alternatives and chooses an alternative that maximizes her utility from among those available. Our model of the customer is thus general [2].

2.1. Choice Model

In order to make useful predictions on customer behavior that might, for instance, guide the selection of a set $\mathcal{M}$ to offer for sale, one must specify a choice model. A general choice model is effectively a conditional probability distribution $\mathbb{P}(\cdot|\cdot) : \mathcal{N} \times 2^{\mathcal{N}} \to [0,1]$, that yields the probability of purchase of a particular product in $\mathcal{N}$ given the set of alternatives available to the customer.

We will assume essentially the most general model for $\mathbb{P}(\cdot|\cdot)$. In particular, we assume that there exists a distribution $\lambda : S_N \to [0,1]$ over the set of all possible permutations S_N. Recall here that S_N is effectively the set of all possible customer types since every customer is associated with a permutation which uniquely determines her choice behavior. The distribution λ defines our choice model as follows: Define the set

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

$\mathcal{S}_j(\mathcal{M})$ is simply the set of all customer types that would purchase product j when the offer set is $\mathcal{M}$. Our choice model is then given by

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

Not surprisingly, as mentioned above, the above model subsumes essentially any model of choice one might concoct: in particular, all we have assumed is that at a *given* point in time a customer possess *rational* (transitive) (see Mas-Colell et al. (1995)) preferences over all alternatives [3], and that a particular customer will purchase her most preferred product from the offered set according to these preferences; a given customer sampled at different times may well have a distinct set of preferences.

[2] As opposed to associating a customer with a fixed σ, one may also associate customers with distributions over permutations. This latter formalism is superfluous for our purposes.

[3] Note however that the customer need not be aware of these preferences; from (1), it is evident that the customer need only be aware of his preferences for elements of the offer set.

2.2. Data

The class of choice models we will work with is quite general and imposes a minimal number of behavioral assumptions on customers a-priori. That said the data available to calibrate such a model will typically be limited in the sense that a modeler will have sales rate information for a potentially small collection of assortments. Ignoring the difficulties of such a calibration problem for now, we posit a general notion of what we will mean by observable 'data'. The abstract notion we posit will quickly be seen as relevant to data one might obtain from sales information.

We 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. Typically we anticipate $m \ll N!$ signifying, for example, the fact that we have sales information for only a limited number of assortments.

We now show how the type of data available in practice can be cast in the form of $y = A\lambda$. In the retail context, historical data about customer purchase behavior is available in the form of observed sales transactions from a set of displayed assortments. In particular, one typically has information about observed sales for a sequence of test assortments say $\mathcal{M}_1, \mathcal{M}_2 \ldots, \mathcal{M}_L$. For each of the assortments $\mathcal{M}_\ell$ and products i, the sales data provides the fraction of purchasing customers who purchased product i when the displayed assortment was $\mathcal{M}_\ell$. Given this data, we claim that it can written as a linear combination of λ for an appropriate choice of matrix A.

To see this, it is instructive to start with a simple special case that we call 'comparison data.' Specifically, suppose that we have access to data that provides us with the information about the fraction of customers that prefer poduct i to product j, for all pairs of products i and j. For this case, the partial information vector y may be indexed by i, j with $0 \le i, j \le N-1; i \ne j$. For each i, j, y_{ij} denotes the fraction of customers that prefer product i to 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 $\sigma(i) < \sigma(j)$. It is important to note here that we have introduced the 'comparison data' for the simplicity of exposition and (as will become apparent later) for theoretical considerations. The way we have defined it, 'comparison data' is in fact *not* readily available in practice: for starters, there is typically censoring because of which we can only observe the fraction of customers who prefer i to both j and 0 when we offer the pair of products i, j. In addition, practical applications do not typically provide sales information about all the $\binom{N}{2}$ possible pairs of products. Nevertheless, 'comparison data' provides a simple yet non-trivial example that makes our setup concrete.

More realistically, we have sales transaction data from a set of displayed assortments $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_L$. In this case, denoting by $y_{i\ell}$ the fraction of customers purchasing product i when assortment $\mathcal{M}_\ell$ is on offer, our partial information vector, $y \in [0,1]^{N.L}$, may thus be indexed by i, ℓ

with $0 \leq i \leq N-1, 1 \leq \ell \leq L$. The matrix A is then in $\{0,1\}^{N.L \times N!}$. For a column of A corresponding to the permutation σ, $A(\sigma)$, we will then have $A(\sigma)_{i\ell} = 1$ iff $i \in \mathcal{M}_\ell$ and $\sigma(i) < \sigma(j)$ for all products j in assortment $\mathcal{M}_\ell \cup \{0\}$.

Finally, we emphasize that the idea of viewing partial information as $y = A\lambda$ is a very powerful one. It captures several different types of interesting partial information in addition to the transactional data as described above. This becomes important both from a theoretical standpoint and from the standpoint of other applications. Although, we won't explore any other application contexts, we discuss two other types of partial information called the 'ranking data' and 'top-set data' for our theoretical analysis in Section 6.

2.3. Incorporating Choice In Decision Models: A Revenue Estimation Black-Box

While modeling choice is useful for a variety of reasons, we are largely motivated by decision models for OM problems that benefit from the incorporation of a choice model. In many of these models, the fundamental feature impacted by the choice model is a 'revenue function' that measures revenue rates corresponding to a particular assortment of products offered to customers. Concrete examples include static assortment management, network revenue management under choice and inventory management assuming substitution.

We formalize this revenue function. We associate every product in $\mathcal{N}$ with a retail price p_j. Of course, $p_0 = 0$. The revenue function, $R(\mathcal{M})$, determines expected revenues to a retailer from offering a set of products $\mathcal{M}$ to his customers. Under our choice model this is given by:

$$R(\mathcal{M}) = \sum_{j \in \mathcal{M}} p_j \lambda^j(\mathcal{M}).$$

The function $R(\cdot)$ is a fundamental building block for all of the OM problems described above, so that we view the problem of estimating $R(\cdot)$ as our central motivating problem. The above specification is general, and we will refer to *any* linear functional of the type above as a revenue function. As another useful example of such a functional, consider setting $p_j = 1$ for all $j > 0$ (i.e. all products other than the no-purchase option). In this case, the revenue function, $R(\mathcal{M})$ yields the probability an arriving customer will purchase some product in $\mathcal{M}$; i.e. the 'conversion rate' under assortment $\mathcal{M}$.

Given a 'black-box' that is capable of producing estimates of $R(\cdot)$ using some limited corpus of data, one may then hope to use such a black box for making assortment decisions over time in the context of the OM problems of the type discussed in the introduction.

2.4. Problem Formulations

Imagine we have a corpus of transaction data, summarized by an appropriate data vector y as described in Section 2.2. Our goal is to use *just* this data to make predictions about the revenue rate (i.e the expected revenues garnered from a random customer) for some given assortment, say

$\mathcal{M}$, that has never been encountered in past data. We propose accomplishing this by solving the following program:

$$\begin{array}{ll} \underset{\lambda}{\text{minimize}} & R(\mathcal{M}) \\ \text{subject to} & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \tag{2}$$

In particular, the optimal value of this program will constitute our prediction for the revenue rate. In words, the feasible region of this program describes the set of all choice models consistent with the observed data y. The optimal objective value consequently corresponds to the *minimum* revenues possible for the assortment $\mathcal{M}$ under any choice model consistent with the observed data. Since the family of choice models we considered was *generic* this prediction relies on simply the data and basic economic assumptions on the customer that are tacitly assumed in essentially any choice model.

The philosophy underlying the above program can be put to other uses. For instance, one might seek to recover a choice model itself from the available data. In a parametric world, one would consider a suitably small, fixed family of models within which a unique model would best explain (but not necessarily be consistent with) the available data. It is highly unlikely that available data will determine a unique model in the *general* family of models we consider here. Our non-parametric setting thus requires an appropriate selection criterion. A natural criterion is to seek the 'simplest' choice model that is consistent with the observed data. There are many notions of what one might consider simple. One criterion that enjoys widespread use in high-dimensional statistics is sparsity. In particular, we may consider finding a choice model λ consistent with the observed data, that has minimal support, $\|\lambda\|_0 \triangleq |\{\lambda(\sigma) \,:\, \lambda(\sigma) \neq 0\}|$. In other words, we might seek to explain observed purchasing behavior by presuming as small a number of modes of customer choice behavior as possible (where we associate a 'mode' of choice with a ranking of products). More formally, we might seek to solve:

$$\begin{array}{ll} \underset{\lambda}{\text{minimize}} & \|\lambda\|_0 \\ \text{subject to} & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \tag{3}$$

Sections 3, 4 and 5 are focused on providing procedures to solve the program (2), and on examining the quality of the predictions produced on simulated data and actual transaction data respectively. Section 6 will discuss algorithmic and interesting descriptive issues pertaining to (3).

3. Revenue Predictions: Computation

In the previous section we formulated the task of computing revenue predictions via a non-parametric model of choice and any available data as the mathematical program (2), which we repeat below, in

a slightly different form for clarity:

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

The above mathematical program is a linear program in the variables λ. Interpreting the program in words, the constraints $A\lambda = y$ ensure that any λ assumed in making a revenue estimate is *consistent* with the observed data. Other than this consistency requirement, writing the probability that a customer purchases $j \in \mathcal{M}, \mathbb{P}(j|\mathcal{M})$, as the quantity $\lambda_j(\mathcal{M}) \triangleq \sum_{\sigma\in\mathcal{S}_j(\mathcal{M})} \lambda(\sigma)$ assumes that the choice model satisfies the basic structure laid out in Section 2.1. We make no other assumptions outside of these, and ask for the lowest expected revenues possible for $\mathcal{M}$ under *any* choice model satisfying these requirements.

Thus, while the assumptions implicit in making a revenue estimate are something that the user need not think about, the two natural questions that arise are:

1. How does one solve this conceptually simple program in practice given that the program involves an intractable number of variables?
2. Even if one did succeed in solving such a program are the revenue predictions produced useful or are they too loose to be of practical value?

This section will focus on the first question. In practical applications such a procedure would need to be integrated into a larger decision problem and so it is useful to understand the computational details which we present at a high level in this section. The second, 'so what' question will be the subject of the next two sections where we will examine the performance of the scheme on simulated transaction data, and finally on a real world sales prediction problem using real data. Finally, we will examine an interesting property enjoyed by the choice models implicitly assumed in making the predictions in this scheme in Section 6.

3.1. The Dual to the Robust Problem

At a high level our approach to solving (2) will be to consider the dual of that program and then derive efficient exact or approximate descriptions to the feasible regions of these programs. We begin by considering the dual program to (2). In preparation for taking the dual, let us define

$$\mathcal{A}_j(\mathcal{M}) \triangleq \{A(\sigma) : \sigma \in \mathcal{S}_j(\mathcal{M})\},$$

where recall that $\mathcal{S}_j(\mathcal{M}) = \{\sigma \in S_N : \sigma(j) < \sigma(i), \forall i \in \mathcal{M}, i \neq j\}$ denotes the set of all permutations that result in the purchase of $j \in \mathcal{M}$ when the offered assortment is $\mathcal{M}$. Since $S_N = \cup_{j\in\mathcal{M}} \mathcal{S}_j(\mathcal{M})$ and

$\mathcal{S}_j(\mathcal{M}) \cap \mathcal{S}_i(\mathcal{M}) = \emptyset$ for $i \neq j$, we have implicitly specified a partition of the columns of the matrix A. Armed with this notation, the dual of (2) is:

$$\begin{array}{l} \underset{\alpha,\nu}{\text{maximize}}\ \alpha^\top y + \nu \\ \text{subject to} \underset{x^j \in \mathcal{A}_j(\mathcal{M})}{\max}\ (\alpha^\top x^j + \nu) \leq p_j, \ \text{ for each } \ j \in \mathcal{M}. \end{array} \tag{4}$$

where α and ν are dual variables corresponding respectively to the data consistency constraints $A\lambda = y$ and the requirement that λ is a probability distribution (i.e. $\mathbf{1}^\top \lambda = 1$) respectively. Of course, this program has a potentially intractable number of constraints. We explore two approaches to solving the dual:

1. An extremely simple to implement approach that relies on sampling constraints in the dual that will, in general produce approximate solutions that are upper bounds to the optimal solution of our robust estimation problem.
2. An approach that relies on producing effective representations of the sets $\mathcal{A}_j(\mathcal{M})$, so that each of the constraints $\max_{x^j \in \mathcal{A}_j(\mathcal{M})} (\alpha^\top x^j + \nu) \leq p_j$, can be expressed efficiently.This approach is slightly more complex to implement but in return can be used to sequentially produce tighter approximations to the robust estimation problem. In certain special cases, this approach is provably efficient and optimal.

3.2. The First Approach: Constraint Sampling

The following is an extremely simple to implement approach to approximately solve the problem (4):

1. Select a distribution over permutations, ψ.
2. Sample n permutations according to the distribution. Call this set of permutation $\hat{\mathcal{S}}$.
3. Solve the program:

$$\begin{array}{l} \underset{\alpha,\nu}{\text{maximize}}\ \alpha^\top y + \nu \\ \text{subject to } \alpha^\top A(\sigma) + \nu \leq p_j, \ \text{ for each } \ j \in \mathcal{M}, \sigma \in \hat{\mathcal{S}} \end{array} \tag{5}$$

Observe that (5) is essentially a 'sampled' version of the problem (4), wherein constraints of that problem have been sampled according to the distribution ψ and is consequently a relaxation of that problem. A solution to (5) is consequently an upper bound to the optimal solution to (4).

The question of whether the solutions thus obtained provide meaningful approximations to (4) is partially addressed by recent theory developed by Calafiore and Campi (2005). In particular, it has been shown that for a problem with m variables and given $n = O((1/\varepsilon)(m\ln(1/\varepsilon) + \ln(1/\delta))$ samples, we must have that with probability at least $1-\delta$ the following holds: An optimal solution to (5) violates at most an ϵ fraction of constraints of the problem (4) under the measure ψ. Hence, given a number of samples that scales only with the number of variables (and is independent of the number of constraints in (4), one can produce an solution to (4) that satisfies all but a small fraction of

constraints. The theory does not provide any guarantees on how far the optimal cost of the relaxed problem is from the optimal cost of the original problem.

The heuristic nature of this approach notwithstanding, it is extremely simple to implement, and in the experiments conducted in the next section, provided close to optimal solutions.

3.3. The Second Approach: Efficient Representations of $\mathcal{A}_j(\mathcal{M})$

We describe here one notion of an efficient representation of the sets $\mathcal{A}_j(\mathcal{M})$, and assuming we have such a representation, we describe how one may solve (4) efficiently. We will deal with the issue of actually coming up with these efficient representations in Appendix B, where we will develop an efficient representation for ranking data and demonstrate a generic procedure to sequentially produce such representations.

Let us 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 of A. Consider the convex hull of the set $\mathcal{A}_{jd}(\mathcal{M})$, $\text{conv}\{\mathcal{A}_{jd}(\mathcal{M})\} \triangleq \bar{\mathcal{A}}_{jd}(\mathcal{M})$. Recalling that $A \in \{0,1\}^{m\times N}$, $\mathcal{A}_{jd}(\mathcal{M}) \subset \{0,1\}^m$. $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ is thus a polytope contained in the m-dimensional unit cube, $[0,1]^m$. In other words,

$$\bar{\mathcal{A}}_{jd}(\mathcal{M}) = \{x^{jd} : A_1^{jd}\, x^{jd} \geq b_1^{jd}, \quad A_2^{jd}\, x^{jd} = b_2^{jd}, \quad A_3^{jd}\, x^{jd} \leq b_3^{jd}, \quad x^{jd} \in \mathbb{R}_+^m\} \tag{6}$$

for some matrices $A_\cdot^{jd}$ and vectors $b_\cdot^{jd}$. 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 (6). If the number of partitions as well as the polyhedral description of each set of the partition given by (6) is polynomial in the input size, we will regard the canonical representation as efficient. Of course, there is no guarantee that an efficient representation of this type exists; clearly, this must rely on the nature of our partial information i.e. the structure of the matrix A. Even if an efficient representation did exist, it remains unclear whether we can identify it. Ignoring these issues for now, we will in the remainder of this section demonstrate how given a representation of the type (6), one may solve (4) in time polynomial in the size of the representation.

For simplicity of notation, in what follows we assume that each polytope $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ is in standard form,

$$\bar{\mathcal{A}}_{jd}(\mathcal{M}) = \{x^{jd} : A^{jd}\, x^{jd} = b^{jd}, \quad x^{jd} \geq 0.\}.$$

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. Now, by strong duality we have:

$$\begin{array}{ll} \underset{x^{jd}}{\text{maximize}} & \alpha^\top x^{jd} + \nu \\ \text{subject to} & A^{jd}\, x^{jd} = b^{jd} \\ & x^{jd} \geq 0. \end{array} \quad \equiv \quad \begin{array}{ll} \underset{\gamma^{jd}}{\text{minimize}} & {b^{jd}}^\top \gamma^{jd} + \nu \\ \text{subject to} & {\gamma^{jd}}^\top A^{jd} \geq \alpha \end{array} \tag{7}$$

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 (2) is equivalent to the following LP whose complexity is polynomial in the description of our canonical representation:

$$\begin{array}{ll} \underset{\alpha,\nu}{\text{maximize}} & \alpha^\top y + \nu \\ \text{subject to} & b^{jd\top}\gamma^{jd} + \nu \le p_j \quad \text{for all } j \in \mathcal{M}, d = 1,2,\ldots,D_j \\ & \gamma^{jd\top} A^{jd} \ge \alpha \qquad \text{for all } j \in \mathcal{M}, d = 1,2,\ldots,D_j. \end{array} \tag{8}$$

As discussed, our ability to solve (8) relies on our ability to produce an efficient canonical representation of $\mathcal{S}_j(\mathcal{M})$ of the type (6). In Appendix B, we first consider the case of ranking data, where an efficient such representation may be produced. We then illustrate a method that produces a sequence of 'outer-approximations' to (6) for general types of data, and thereby allows us to produce a sequence of improving lower bounding approximations to our robust revenue estimation problem, (2). This provides a general procedure to address the task of solving (4), or equivalently, (2).

We end this section with a brief note on noise. So far we have assumed that the choice probabilities y_i obtained from historical data are known exactly and fit the model exactly so that there exists a choice model λ such that $y = A\lambda$. This is of course hardly the case in practice. Specifically, there are two sources of errors. First is the finite sample error caused due the fact that the choice probabilities y_i can only be estimated through a sample average of finitely many samples; depending on the number of samples available, there is uncertainty in the estimate of y_i. Secondly, there could be model misfit errors. That is, even if the choice probabilities y_i are known exactly, our choice model may not be an exact fit, making the set of equalities $y = A\lambda$ and λ a distribution infeasible. In order to overcome these issues, we incorporate these errors in making our predictions. Specifically, we assume that we are given an uncertainty region $\mathcal{E}$ constructed from the data such that there exists a choice model λ with $A\lambda \in \mathcal{E}$. The uncertainty region $\mathcal{E}$ may either be an uncertainty ellipsoid or a 'box' derived from sample averages of the associated choice probabilities and the corresponding confidence intervals. Given such an uncertainty region $\mathcal{E}$, we predict revenues by solving the following LP:

$$\begin{array}{ll} \underset{\lambda, y \in \mathcal{E}}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \ge 0. \end{array}$$

Provided $\mathcal{E}$ is convex, this program is essentially no harder to solve than the variant of the problem we have discussed and similar methods to those developed in this section apply. It is clear from the convex program that if the uncertainty region $\mathcal{E}$ is 'too small', then the constraint set will become infeasible. On the other hand, if $\mathcal{E}$ is 'too large', then we would have a poor fit to the data and

conservative revenue predictions. In order to balance the extremes, in our empirical analyses, we choose the 'smallest' uncertainty region such that the constraint set becomes feasible. Precise details of how we do that are provided in Section 5.

4. Revenue Predictions: Data Driven Computational Study

In this section, we describe the results of an extensive simulation study, the main purpose of which is to demonstrate that the robust approach can capture various underlying parametric structures and produce good revenue predictions. For this study, we pick a range of random utility parametric structures used extensively in current modeling practice.

The broad experimental procedure we followed is the following:

1. Pick a structural model. This may be a model derived from real-world data or a purely synthetic model.
2. Use this structural model to simulate sales for a set of test assortments. This simulates a data set that a practitioner likely has access to.
3. Use this transaction data to estimate marginal information y, and use y to implement the robust approach.
4. Use the implemented robust approach to predict revenues for a distinct set of assortments, and compare the predictions to the *true* revenues computed using the 'ground-truth' structural model chosen for benchmarking in step 1.

Notice that the above experimental procedure lets us isolate the impact of structural errors from that of finite sample errors. Specifically, our goal is to understand how well the robust approach captures the underlying choice structure. For this purpose, we ignore any estimation errors in data by using the 'ground-truth' parametric model to compute the *exact* values of any choice probabilities and revenues required for comparison. Therefore, if the robust approach has good performance across an interesting spectrum of structural models that are believed to be good fits to data observed in practice, we can conclude that the robust approach is likely to offer accurate revenue predictions with no additional information about structure across a wide-range of problems encountered in practice.

4.1. Benchmark Models and Nature of Synthetic Data

The above procedure generates data sets using a variety of 'ground truth' structural models. We pick the following 'random utility' models as benchmarks. A self-contained and compact exposition on the foundations of each of the benchmark models below may be found in the appendix.

Multinomial logit family (MNL): For this family, we have:

$$\mathbb{P}(j|\mathcal{M}) = w_j / \sum_{i \in \mathcal{M}} w_i.$$

where the w_i are the parameters specifying the models. See Appendix C.1 for more details.

Nested logit family (NL): This model is a first attempt at overcoming the 'independence of irrelevant alternatives' effect, a shortcoming of the MNL model. For this family, the universe of products is partitioned into L mutually exclusive subsets, or 'nests', denoted by $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_L$ such that

$$\mathcal{N} = \bigcup_{\ell=1}^{L} \mathcal{N}_\ell \qquad \text{and} \qquad \mathcal{N}_\ell \cap \mathcal{N}_m = \emptyset, \text{ for } m \neq \ell.$$

This model takes the form:

$$\mathbb{P}(j|\mathcal{M}) = \mathbb{P}(\mathcal{N}_\ell|\mathcal{M})\,\mathbb{P}(j|\mathcal{N}_\ell, \mathcal{M}) = \frac{(w(\ell,\mathcal{M}))^\rho}{\sum_{m=1}^{L}(w(m,\mathcal{M}))^\rho}\ \frac{w_j}{w(\ell,\mathcal{M})}. \tag{9}$$

where $\rho < 1$ is a certain scale parameter, and

$$w(\ell,\mathcal{M}) \stackrel{\text{def}}{=} \alpha_\ell w_0 + \sum_{i \in (\mathcal{N}_\ell \cap \mathcal{M}) \setminus \{0\}} w_i.$$

Here α_ℓ is the parameter capturing the level of membership of the no-purchase option in nest ℓ and satisfies, $\sum_{\ell=1}^{L} \alpha_\ell^\rho = 1, \quad \alpha_\ell \geq 0, \text{ for } \ell = 1, 2, \ldots, L$. In cases when $\alpha_\ell < 1$ for all ℓ, the family is called the *Cross nested logit (CNL)* family. For a more detailed description including the corresponding random utility function and bibliographic details, see Appendix C.2

Mixed multinomial logit family (MMNL): This model accounts specifically for customer heterogeneity. In its most common form, the model reduces to:

$$\mathbb{P}(j|\mathcal{M}) = \int \frac{\exp\{\beta^T x_j\}}{\sum_{i \in \mathcal{M}} \exp\{\beta^T x_i\}} G(d\beta;\theta).$$

where x_j is a vector of observed attributes for the jth product, and $G(\cdot,\theta)$ is a distribution parameterized by θ selected by the econometrician that describes heterogeneity in taste. Because the coefficients β are assumed to be random (unlike for the MNL model), this model is often also termed in the literature the random coefficients multinomial logit (RC-MNL) model. For this paper we restrict ourselves to Gaussian MMNL models in which we assume that β has a multivariate Gaussian distribution. For a more detailed description including the corresponding random utility function and bibliographic details, see Appendix C.3.

Transaction Data Generated: Having selected (and specified) a structural model from the above list, we generated sales transactions as follows:

1. Fix an assortment of two products, i, j.
2. Compute the values of $P(i|\{i,j,0\}), P(j|\{i,j,0\})$ using the chosen parametric model.
3. Repeat the above procedure for all pairs, $\{i,j\}$, and single item sets, $\{i\}$.

The above data is succinctly summarized as an $N^2 - N$ dimensional data vector y, where $y_{i,j} = P(i|\{i,j,0\})$ for $0 \leq i, j \leq N-1$, $i \neq j$. Given the above data, the precise specialization of the robust estimation problem (2) that we solve may be found in Appendix B.3.

4.2. Experiments Conducted

With the above setup we conducted two broad sets of experiments. In the first set of experiments, we picked specific models from the MNL, CNL, and MMNL model classes; the MNL model was constructed using DVD shopping cart data from Amazon.com, and the CNL and MMNL models were obtained through slight 'perturbations' of the MNL model. In order to avoid any artifacts associated with specific models, in the second set of experiments, we conducted 'stress tests' by generating a number of instances of models from each of the MNL, CNL, and MMNL models classes. We next present the details of the two sets of experiments.

The Amazon Model: We considered an MNL model fit to Amazon.com DVD sales data collected between 1 July 2005 to 30 September 2005 [4] ,where an individual customer's utility for a given DVD, j is given by:

$$U_j = \theta_0 + \theta_1 x_{j,1} + \theta_2 x_{j,2} + \xi_j;^{5}$$

here $x_{j,1}$ is the the price of the package j divided by the number of physical discs it contains, and $x_{j,2}$ is the total number of helpful votes received by product j and ξ_j is a standard Gumbel. The model fit to the data has $\theta_0 = -4.31$, $\theta_1 = -0.038$ and $\theta_2 = 3.54 \times 10^{-5}$. See Table 2 for the attribute values taken by the 15 products we used for our experiments. We will abbreviate this model AMZN for future reference.

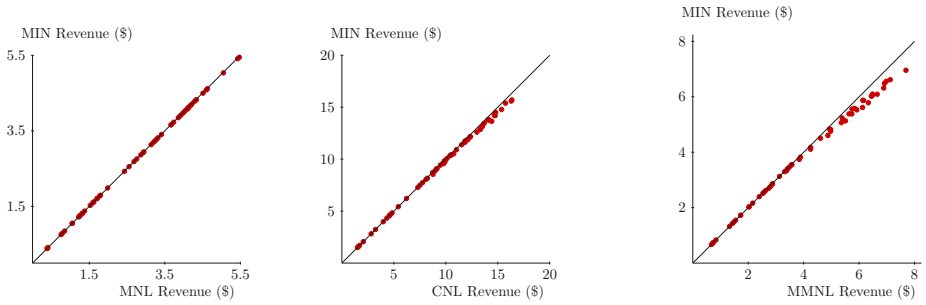
We also considered the following synthetic perturbations of the AMZN model:

1. AMZN-CNL: We derived a CNL model from the original AMZN model by partitioning the products into 4 nests with the first nest containing products 1 to 5, the second nest containing products 6 to 9, the third containing products 10 to 13, and the last containing products 14 and 15. We choose $\rho = 0.5$. We assigned the no-purchase option to every nest with nest membership parameter $\alpha_\ell = (1/4)^{(1/\rho)} = 1/16$.
2. AMZN-MMNL: We derived an MMNL model from the original AMZN model by replacing each θ_i parameter with the random quantity $\beta_i = (1+\eta_{i,j})\theta_i$, for $i = 0, 1, 2$ with $\eta_{i,j}$ a customer specific random variable distributed as a zero mean normal random variable with standard deviation $s = 0.25$. Put differently, we assumed that the random coefficients β_i are independent and have a Gaussian distribution with mean θ_i and standard deviation $s\theta_i$.

Figure 1 shows the results of the generic experiment for each of the three models above. Each experiment queries the robust estimate on sixty randomly drawn assortments of sizes between one and seven and compares these estimates to those under the respective true model for each case.

[4] The specifics of this model were shared with us by the authors of Rusmevichientong et al. (2008).

[5] The corresponding weights w_j are given by $w_j = \exp(\theta_0 + \theta_1 x_{j,1} + \theta_3 x_{j,2})$.



(a) AMZN (b) AMNN-CNL (c) AMZN-MMNL

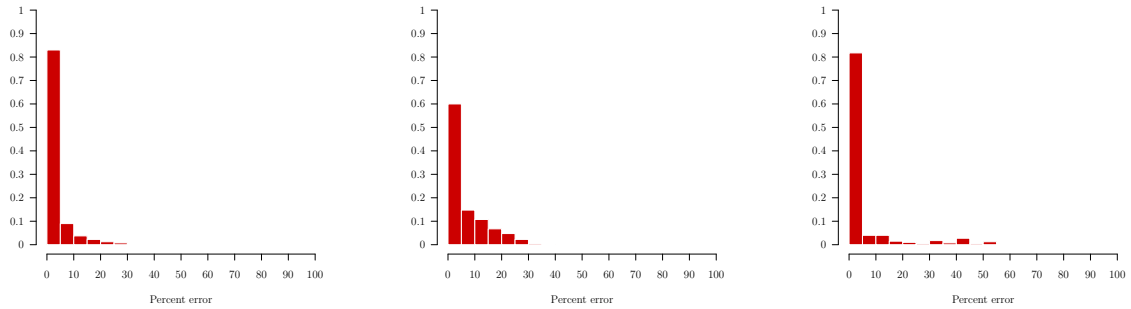
Figure 1 Robust revenue predictions (MIN) vs. true revenues for the AMZN, AMZN-CNL and AMZN-MMNL models. Each of the 60 points in a plot corresponds to (true revenue, MIN) for a randomly drawn assortment.

Synthetic Model Experiments: The above experiments considered structurally diverse models, each for a *specific* set of parameters. Are the conclusions suggested by Figure 1 artifacts of the set of parameters? To assuage this concern, we performed 'stress' tests by considering each structural model in turn, and for each model generating a number of instances of the model by drawing the relevant parameters from a generative family. For each structural model, we considered the following generative families of parameters:

1. MNL Random Family: 20 randomly generated models on 15 products, each generated by drawing mean utilities, $\ln w_j$, uniformly between -5 and 5.
2. CNL Random Family: We maintained the nests, selection of ρ and α_j as in the AMZN-CNL model. We generated 20 distinct CNL models, each generated by drawing $\ln w_j$ uniformly between -5 and 5.
3. MMNL Random Family: We preserved the basic nature of the AMZN-MMNL model. We considered 20 randomly generated MMNL models. Each model differs in the distribution of the parameter vector β. The random coefficients β_j in each case are defined as follows: $\beta_j = (1+\eta_{i,j})\theta_j$ where $\eta_{i,j}$ is a $N(\mu_j, 0.25)$ random variable. Each of the 20 models corresponds to a single draw of $\mu_j, j = 0,1,2$ form the uniform distribution on $[-1,1]$.

For each of the 60 structural model instances described above, we randomly generated 20 offer sets of sizes between 1 and 7. For a given offer set $\mathcal{M}$, we queried the robust procedure and compared the revenue estimate produced to the true revenue for that offer set; we can compute the latter quantity theoretically. In particular, we measured the relative error, $\varepsilon(\mathcal{M}) \stackrel{\text{def}}{=} \frac{R^{\text{true}}(\mathcal{M}) - R^{\text{MIN}}(\mathcal{M})}{R^{\text{MIN}}(\mathcal{M})}$. The three histograms in Figure 2 below represent distributions of relative error for the three generative families described above. Each histogram consists of 400 test points; a given test point corresponds to one of the 20 randomly generated structural models in the relevant family, and a random assortment.

In the above 'stress' tests, we kept the standard deviation of the coefficients β_i in the MMNL models as $s\theta_i$ with the multiplier s fixed at 0.25. The standard deviation of the coefficients in the



(a) MNL-Rand (b) CNL-Rand (c) MMNL-Rand

Figure 2 Relative error across multiple instances of the MNL, CNL and MMNL structural models.

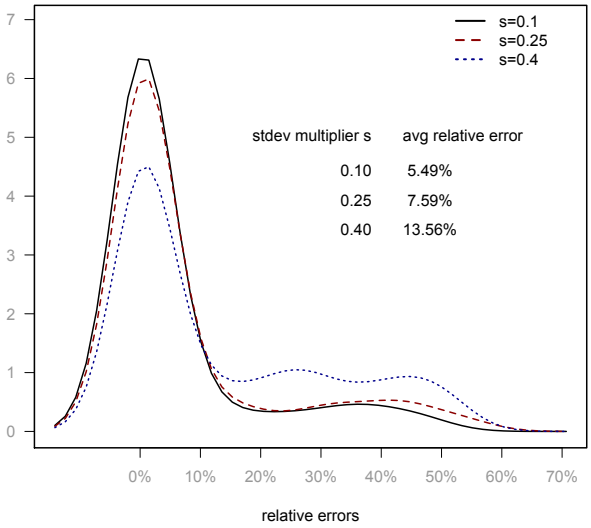


Figure 3 The accuracy of robust revenue predictions deteriorates with increase in model complexity, measured in terms of the standard deviation $s\theta_i$ of the normally distributed coefficients β_i in the MMNL model. s is the standard deviation multiplier taking three different values: 0.1, 0.25, and 0.4. The densities were estimated through kernel density estimation. The density estimates go below zero as a result of smoothing.

MMNL model can be treated as a measure of the heterogeneity or the "complexity" of the model. Naturally, if we keep the "amount" of transaction data fixed and increase the standard deviation – and hence the complexity of the underlying model – we expect the accuracy of robust estimates to deteriorate. To give a sense of the sensitivity of the accuracy of robust revenue predictions to changes in the standard deviation of coefficients, we repeated the above stress tests with the MMNL model class for three different values of the multiplier s: 0.1, 0.25, and 0.4. Figure 3 shows the comparison of the density plots of relative errors for the three cases.

We draw the following broad conclusion from the above experiments:

- Given limited marginal information for distributions over permutations λ arising from a number of commonly used structural models of choice, the robust approach effectively captures diverse parametric structures and provides close revenue predictions under range of practically relevant parametric models.
- With the type of marginal information *y fixed*, the accuracy of robust revenue predictions deteriorates (albeit mildly) as the complexity of the underlying model increases; this is evidenced by the deterioration of robust performance as we go from the MNL to the MMNL model class, and similarly as we increase the standard deviation of the coefficients for the MMNL model while keeping the 'amount' of data fixed.
- The design of our experiments allows us to conclude that in the event that a given structural model among the types used in our experiments predicts revenue rates accurately, the robust approach is likely to be just as good *without* knowledge of the relevant structure. In the event that the structural model used is a poor fit, the robust approach will continue to provide meaningful guarantees on revenues under the mild condition that it is tested in an environment where the distribution generating sales is no different from the distribution used to collect marginal information.

5. Revenue Predictions: Case Study with a Major US Automaker

In this section, we present the results of a case study conducted using sales transaction data from the dealer network of a major US automaker. Our goal in this study is to use historical transaction data to predict the sales rate or 'conversion rate' for any given offer set of automobiles on a dealer lot. This conversion-rate is defined as the probability of converting an arriving customer into a purchasing customer. The purpose of the case study is two-fold: (1) To demonstrate how the prediction methods developed in this paper can be applied in the real-world and the quality of the predictions they offer in an absolute sense, and (2) To pit the robust method for revenue predictions in a 'horse-race' against parametric approaches based on the MNL and MMNL families of choice models. In order to test the performance of these approaches in different regimes of calibration data, we carried out cross-validations with varying 'amounts' of training/calibration data. The results of the experiments conducted as part of the case study provide us with the evidence to draw two main conclusions:

1. The robust method predicts conversion rates more accurately than either of the parametric methods. In our case study, the improvement in accuracy was about 20% across all regimes of calibration data.
2. Unlike the parametric methods we study, the robust approach is apparently *not* susceptible to over-fitting and under-fitting.

The 20% improvement in accuracy is substantial. The second conclusion has important implications as well: In practice, it is often difficult to ascertain whether the data available is "sufficient" to fit

the model at hand. As a result, parametric structures are prone to over-fitting or under-fitting. The robust approach, on the other hand, *automatically* scales the complexity of the underlying model class with data available, so in principle one should be able to avoid these issues. This is borne out by the case study. In the remainder of this section we describe the experimental setup and then present the evidence to support the above conclusions.

5.1. Setup

We collect data comprising purchase transactions of a specific range of small SUVs offered by a major US automaker over 16 months. The data is collected at the dealership level (i.e the finest level possible) for a network of dealers in the Midwest. Each transaction contains information about the date of sale, the identity of the SUV sold, and the identity of the other SUVs on the dealership lot at the time of sale. Here by 'identity' we mean a unique model identifier that collectively identifies a package of features, color and invoice price point. We make the assumption that purchase behavior within the zone can be described by a single distribution over preference lists. To ensure the validity of this assumption, we restrict attention to a specific dealership zone, defined as the collection of dealerships within an appropriately defined geographical area with relatively homogeneous demographic features. Our data consisted of sales information on 14 distinct SUV identities (as described above), which we term products.

Data: To describe the data and our methods precisely, we introduce some notation. We let $\mathscr{M}^{\text{training}}$ and $\mathscr{M}^{\text{test}}$ respectively denote the set of assortments used as part of 'training' and 'test' data. For any given assortment, $\mathcal{M}$ and product $i \in \mathcal{M}$, define $C_{i\mathcal{M}}$ as the total number of sales of product i across *all* dealerships over the data collection period, such that the assortment on offer at the time of sale was $\mathcal{M}$. Similarly $C_{0\mathcal{M}}$ denotes the number of customers who purchased nothing when $\mathcal{M}$ was on offer. Notice that we do not have access to this latter quantity; we will describe how it is estimated momentarily. Finally, let T^{training} denote the set of tuples $(i, \mathcal{M})$ such that $\mathcal{M} \in \mathscr{M}^{\text{training}}$. We observed a total of $M \triangleq |\mathscr{M}^{\text{training}}| + |\mathscr{M}^{\text{test}}| = 203$ distinct assortments (or subsets) of the 14 products in the dataset, where each assortment $\mathcal{M}_i$, $i = 1, 2, \ldots, M$, was on offer at some point at some dealership in the dealership zone.

Demand Untruncation: As discussed above, $C_{0\mathcal{M}}$ is unavailable since we do not observe arriving customers that do not purchase. This issue impacts choice modeling irrespective of whether one chooses the non-parametric approach adopted here of any of the extant parametric approaches. We follow a strategy that is common in practice when one has access to the rich data we do here. In more data limited scenarios more sophisticated techniques can be applied; see for instance

Talluri and van Ryzin (2004a), Vulcano et al. (2010, 2008), Ratliff et al. (2008b). Importantly, our estimates of $C_{0\mathcal{M}}$ will be common to our robust revenue prediction approach and the incumbent parametric approaches we study. Central to the above task is estimating the number of customers that considered assortment $\mathcal{M}$. In particular, given this estimate, we can compute $C_{0\mathcal{M}}$ as

$$C_{0\mathcal{M}} = (\text{num of customer arrivals when } \mathcal{M} \text{ is on offer}) - \sum_{j\in\mathcal{M}} C_{j\mathcal{M}},$$

For a given dealership this is the number of arriving customers over days when $\mathcal{M}$ was on offer at that dealership; we then simply sum this figure over all dealerships. In particular:

$$\text{num of customer arrivals when } \mathcal{M} \text{ was on offer } = \sum_d \alpha_d \,\text{days}_d(\mathcal{M}), \tag{10}$$

where α_d denotes the average number of customers arriving *daily* at dealership d and $\text{days}_d(\mathcal{M})$ denotes the number of days for which $\mathcal{M}$ was on offer at dealership d. There are a number of ways of estimating α_d; in fact this is the focus of the untruncation literature alluded to above. As mentioned we adopt a strategy that is common in practice when one has access to historical sales at a retailer: Assume that α_d is proportional to total sales at the dealer over the year preceding the year in which the data was collected. We estimate the proportionality constant using cross-validation on our training data. It is worth noting that this scheme of estimating α_d implicitly assumes a coarse relationship between arrivals and sales in the preceding year. This relationship is assumed *only* for the purposes of estimating α_d.

Robust method Given $\mathscr{M}^{\text{training}}$ and an assortment $\mathcal{M} \in \mathscr{M}^{\text{test}}$, the conversion-rate of $\mathcal{M}$ is predicted by the robust approach by solving the following LP:

$$\begin{array}{ll} \underset{\lambda}{\text{minimize}} & \sum_{j\in\mathcal{M}} \mathbb{P}_\lambda(j|\mathcal{M}) \\ \text{subject to} & a_{i\mathcal{M}} \le \mathbb{P}_\lambda(i|\mathcal{M}) \le b_{i\mathcal{M}}, \quad \forall (i,\mathcal{M}) \in T^{\text{training}} \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \ge 0, \end{array} \tag{11}$$

where recall that $\mathbb{P}_\lambda(i|\mathcal{M}) = \sum_{\sigma\in\mathcal{S}_i(\mathcal{M})} \lambda(\sigma)$ with $\mathcal{S}_i(\mathcal{M})$ denoting the set $\{\sigma\colon \sigma(i) < \sigma(j) \forall j \in \mathcal{M}, i \ne j\}$ and $[a_{i\mathcal{M}}, b_{i\mathcal{M}}]$ denotes the interval to which $\mathbb{P}_\lambda(i|\mathcal{M})$ belongs. We obtained an approximate solution to the LP in (11) by taking its dual and using the approach of constraint sampling as described in Section 3.2. The LP (11) is a slight modification of the LP in (2) in that the prices p_i are all set to 1 and the equalities $y_t = \mathbb{P}_\lambda(\mathcal{M})$ are changed to inequalities to account for finite sample errors in the data. Setting all the prices to 1 has the effect of computing the conversion-rate for the assortment. For each tuple $(i,\mathcal{M}) \in T^{\text{training}}$, we computed the left and

right end-points as $a_{i\mathcal{M}} = y_{i\mathcal{M}}(1 - z\varepsilon_{i\mathcal{M}})$ and $b_{i\mathcal{M}} = y_{i\mathcal{M}}(1 + z\varepsilon_{i\mathcal{M}})$, where

$$\varepsilon_{i\mathcal{M}} = \sqrt{\frac{1 - y_{i\mathcal{M}}}{C_{i\mathcal{M}}}} \quad \text{and} \quad y_{i\mathcal{M}} = \frac{C_{i\mathcal{M}}}{C_{0\mathcal{M}} + \sum_{i\in\mathcal{M}} C_{i\mathcal{M}}}.$$

Here, $\hat{y_{i\mathcal{M}}}\varepsilon_{i\mathcal{M}}$ is the standard error, and z is a constant multiplier that determines the width of the confidence interval. Different values of z give us approximate confidence intervals for $\mathbb{P}_\lambda(i|\mathcal{M})$. For our experiments, we had set z to be 3.15, which corresponded to the smallest value of z for which (11) was feasible; incidentally, this value of z also corresponds to approximate 99.8% confidence interval for $\mathbb{P}_\lambda(i|\mathcal{M})$.

Parametric methods: As benchmarks, we fitted an MNL as well as an MMNL model given the $C_{i\mathcal{M}}$ and $C_{0\mathcal{M}}$ estimates described above, For the **MNL model**, we assumed the following specific random utility structure: $U_i = V_i + \xi_i$, $i = 0, 1, 2, \ldots, N$, where V_i is the mean utility and ξ_i are i.i.d. Gumbel distributed with location parameter 0 and scale parameter 1, and $N = 14$ is the number of products. For the **MMNL model** we assumed the following specific random utility structure: $U_i = V_i + \beta x_i + \xi_i$, $i = 0, 1, 2, \ldots, 14$, where as before V_i denotes the mean utility and $N = 14$ the number of products, ξ_i are i.i.d. Gumbel with location parameter 0 and scale parameter 1, x_i are dummy features with $x_0 = 0$ and $x_i = 1$ for $i > 0$, and β is Gaussian with mean 0 and variance s^2.

For both models, we used the training data $\hat{y}_{i\mathcal{M}}$, for all $(i, \mathcal{M}) \in T^{\text{training}}$ to determine the Maximum-Likelihood (ML) estimates of the parameters. Specifically, fixing V_0 to 0, we used BIOGEME Bierlaire (2003, 2008) to estimate $V_i, i > 0$, and s.

5.2. Experiments and results

We now describe the experiments we conducted and present the results we obtained. In order to test the predictive performance of the robust, the MNL, and the MMNL methods, we carried out k-fold cross-validations with $k = 2, 5, 10$. In k-fold cross-validation (see Mosteller and Tukey (1987)), we arbitrarily partition the collection of assortments $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M$ into k partitions of about equal size, except may be the last partition. Then, using $k - 1$ partitions as training data to calibrate the methods, we test their performance on the k^{th} partition. We repeat this process k times with each of the k partitions used as test data exactly once. This repetition ensures that each assortment is tested at least once. Note that as k decreases, the number of training assortments decreases resulting in more limited data scenarios. Such limited data scenarios are of course of great practical interest.

We measure the prediction accuracy of the methods using the relative error metric. In particular, letting $\hat{y}(\mathcal{M})$ denote the conversion-rate prediction for test assortment $\mathcal{M}$, the incurred relative error is defined as $|\hat{y}(\mathcal{M}) - y(\mathcal{M})|/y(\mathcal{M})$, where

$$y(\mathcal{M}) := \frac{\text{num of customers who purchase a product when } \mathcal{M} \text{ is on offer}}{\text{num of customer arrivals when } \mathcal{M} \text{ was on offer}}.$$

In the case of the parametric approaches, $\hat{y}(\mathcal{M})$ is computed using the choice model fit to the training data. In the case of the robust approach, we solve an appropriate mathematical program. A detailed description of how $\hat{y}(\mathcal{M})$ is determined by each method is given in the appendix.

We now present the results of the experiments. Figure 4 shows the comparison of the relative errors of the three methods from k-fold cross-validations for $k = 10, 5, 2$. Table 1 shows the mean relative error percentages of the three methods and the percent improvement in mean relative error achieved by the robust method over the MNL and MMNL methods for the three calibration data regimes of $k = 10, 5, 2$. It is clear from the definition of k-fold cross-validation that as k decreases, the "amount" of calibration data decreases, or equivalently calibration data sparsity increases. Such sparse calibration data regimes are of course of great practical interest.

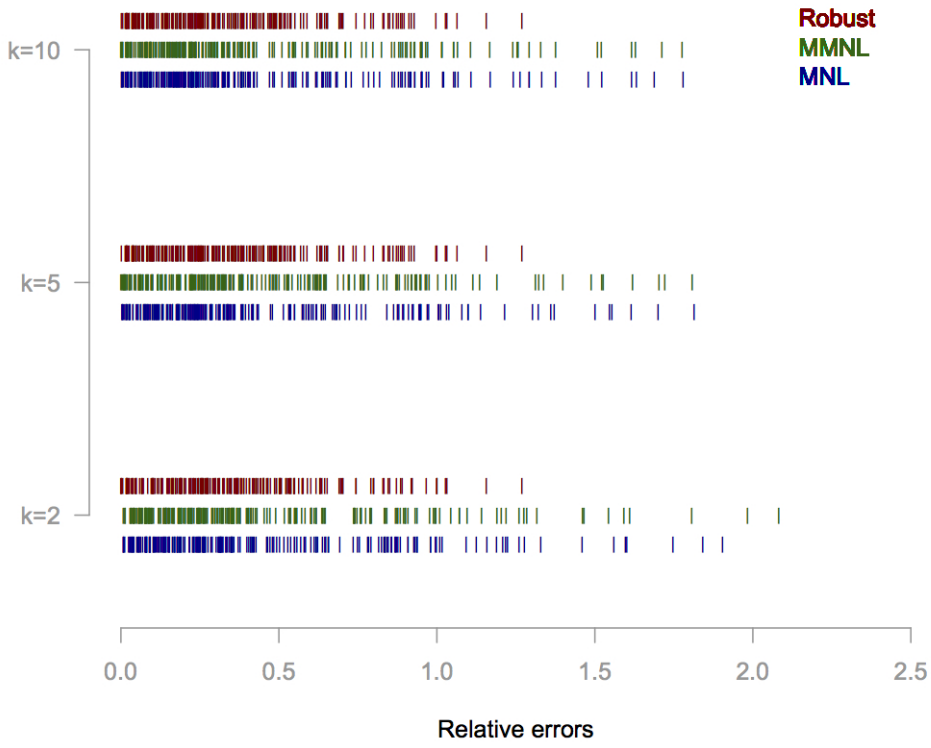


Figure 4 Robust method outperforms both MNL and MMNL methods in conversion-rate predictions across various calibration data regimes. The figure compares relative errors of the three methods in k-fold cross-validations for $k = 10, 5, 2$. Each point corresponds to the relative error for a particular test assortment.

The immediate conclusion we draw from the results is that the prediction accuracy of the robust method is better than those of both MNL and MMNL methods in all calibration data regimes. In

Table 1 Mean relative errors in percentages of different methods

k	MNL	MMNL	Robust	Percent Improvement over MNL	Percent Improvement over MMNL
10	43.43	43.39	34.79	19.89%	19.80%
5	43.25	45.73	35.79	17.23%	21.62%
2	45.65	46.61	36.83	19.33%	20.99%

particular, using the robust method results in close to 20% improvement in prediction accuracy over the MNL and MMNL methods. We also note that while the prediction accuracy of the more complex MMNL method is marginally better than that of the MNL method in the high calibration-data regime of $k = 10$, it quickly becomes worse as the amount of calibration data available decreases. This behavior is a consequence of over-fitting caused due to the complexity of the MMNL model. The performance of the robust method, on the other hand, remains stable across the different regimes of calibration-data.

6. The Sparsest Choice Model Consistent with Data

In making revenue predictions, we did not need to concern ourselves with the choice model implicitly assumed by our prediction procedure. This fact notwithstanding, it is natural to consider criteria for selecting choice models consistent with the observed data that are independent of any decision context. Thus motivated, we consider the natural task of finding the *simplest* choice model consistent with the observed data. As in much of contemporary high dimensional statistics (see for example, Candes et al. (2006), Cormode and Muthukrishnan (2006)), we employ *sparsity* as our measure of simplicity. Sparse models use as few preference lists as possible to explain observed substitutions, and have provided a great deal of tractability in multiple applications (see for example van Ryzin and Vulcano (2008)). Our goal in this section is to first understand the choice models implicitly assumed by the robust procedure through the lens of the sparsity criterion, and second, to understand the discriminative power of this criterion.

Towards the above goal, we begin by characterizing choice models implicitly used by the robust approach in terms of their sparsity. Loosely speaking, we establish that the choice model implicitly used by the robust approach is indeed simple or sparse. In particular, such choice models have sparsity within at most one of the sparsity of the sparsest model consistent with the data. As such, we see that the choice model implicitly selected by our robust revenue prediction procedure is, in essence, the sparsest choice model consistent with the data. From a descriptive perspective, this establishes the appealing fact that simplicity or sparsity is a natural property possessed by all choice models used in making robust revenue predictions. We also establish that the sparsity of the choice model used by the robust approach scales with the dimension of the data vector y thereby establishing that the complexity of the model used by the robust approach scales with the "amount" of data available.

This provides a potential explanation for the immunity of the robust approach to over/under fitting issues, as evidenced in our case study.

Next, we turn to understanding the discriminative power of the sparsest fit criterion. Towards this end, we describe a family of choice models that can be uniquely identified from the given marginal data using the sparsest fit criterion. We intuitively expect the complexity of identifiable models to scale with the "amount" of data that is available. We formalize this intuition by presenting for various types of data, conditions on the model generating the data under which identification is possible. These conditions characterize families of choice models that can be identified in terms of their sparsity and formalize the scaling between the complexity of a model class and the "amount" of data needed to identify it.

6.1. Revenue Prediction and Sparse Models

We now provide a characterization of the choice models implicitly used by the robust procedure through the lens of model sparsity. As mentioned above, loosely speaking, we can establish that the choice models selected implicitly via our revenue estimation procedure are, in essence, close to the sparsest model consistent with the observed data. In other words, the robust approach implicitly uses the simplest models consistent with observed data to predict revenues.

To state our result formally, let us define the set $\mathcal{Y}$ as the set of all possible data vectors, namely the convex hull of the columns of the matrix A. For some $y \in \mathcal{Y}$ and an arbitrary offer set, $\mathcal{M}$, let $\lambda^{\min}(y)$ be an optimal *basic feasible* solution to the program used in our revenue estimation procedure, namely, (2). Moreover, let, $\lambda^{\text{sparse}}(y)$ be the *sparsest* choice model consistent with the data vector y; i.e. $\lambda^{\text{sparse}}(y)$ is an optimal solution to (3). We then have that with probability one, the sparsity (i.e. the number of rank lists with positive mass) under $\lambda^{\min}(y)$ is close to that of $\lambda^{\text{sparse}}(y)$. In particular, we have:

THEOREM 1. *For any distribution over $\mathcal{Y}$ that is absolutely continuous with respect to Lebesgue measure on $\mathcal{Y}$, we have with probability 1, that:*

$$0 \leq \|\lambda^{\min}(y)\|_0 - \|\lambda^{\text{sparse}}(y)\|_0 \leq 1$$

Theorem 1 establishes that if K were the support size of the sparsest distribution consistent with y, the sparsity of the choice model used by our revenue estimation procedure is either K or $K+1$ for "almost all" data vectors y. As such, this establishes that the choice model implicitly employed by the robust procedure is essentially also the sparsest model consistent with the observed data.

In addition the proof of the theorem reveals that the sparsity of the robust choice model consistent with the observed data is either[6] m or $m+1$ for almost all data vectors y of dimension m. This

[6] Here, we assume that matrix A has full row rank.

yields yet another valuable insight into the choice models implicit in our revenue predictions – the complexity of these models, as measured by their sparsity, grows with the amount of observed data. As such, we see that the complexity of the choice model implicitly employed by the robust procedure scales automatically with the amount of available data, as one would desire from a non-parametric scheme. This provides a potential explanation for the robust procedures' lack of susceptibility to the over-fitting observed for the MMNL model in our empirical study.

6.2. Identifiable Families of Choice Models

We now consider the family of choice models that can be identified via the sparsest fit criterion. For that, we present two abstract conditions that, if satisfied by the choice model generating the data y, guarantee that the optimal solution to (3) is unique, and in fact, equal to the choice model generating the data.

Before we describe the conditions, we introduce some notation. As before, 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 are:

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 $A(\sigma_j)_{d(i)} = 0$, 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 the K-dim simplex, or for that matter, any distribution on the K-dim simplex with a density.

When the two conditions above are satisfied by a choice model, this choice model can be recovered from observed data as the solution to problem (3). Specifically, we have:

THEOREM 2. *Suppose we are given $y = A\lambda$ and λ satisfies the signature and linear independence conditions. Then, λ is the unique solution to the program in* (3).

The proof of Theorem 2 is given in Appendix A.2. The proof is constructive in that it describes an efficient scheme to determine the underlying choice model. Thus, the theorem establishes that whenever the underlying choice model satisfies the signature and linear independence conditions, it can identified using an efficient scheme as the optimal solution to the program in (3). We next characterize a family of choice models that satisfy the signature and linear independence conditions. Specifically, we show that *essentially all* choice model with sparsity $K(N)$ satisfy these two conditions

as long as $K(N)$ scales as $\log N, \sqrt{N}$, and N for comparison, top-set, and ranking data respectively. To capture this notion of 'essentially' all choice models, we introduce a natural generative model. It then remains to understand how restrictive these values of $K(N)$ are, which we discuss subsequently.

A 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[7], choose K numbers uniformly at random from the interval $[a,b]$, normalize the numbers so that they sum to 1[8], and assign them to the permutations σ_i, $1 \le i \le K$. For all other permutations $\sigma \neq \sigma_i$, $\lambda(\sigma) = 0$.

Depending on the observed data, we characterize values of sparsity $K = K(N)$ up to which distributions generated by the above generative model can be recovered with a high probability. We derive the sparsity bound for three different types of partial information: comparison data, top-set data, and ranking data. We introduced 'comparison data' above in Section 2. We define ranking and top-set data as follows:

- *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 $0 \le i, r \le N$. For each i, r, y_{ri} is thus the fraction of customers that rank product i at position r. The matrix A is then in $\{0,1\}^{N^2 \times N!}$. For a column of A corresponding to the permutation σ, $A(\sigma)$, we will thus have $A(\sigma)_{ri} = 1$ iff $\sigma(i) = r$.
- *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 $A_2(\sigma)_i = 1$ if and only if $\sigma(i) = 1$.

With the above definitions, we can now state the following result.

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 $K = o(\log N)$ for comparison data, $K = o(\sqrt{N})$ for the top set data, and $K = O(N)$ for ranking data.*

Theorem 3 above implies that essentially all choice models of sparsity $\log N$ (and higher) can be recovered from the types of observed data discussed in the theorem. A natural question that arises at this juncture is what a reasonable value of $K(N)$ might be. To give a sense of this, we provide the following approximation result: a good approximation to *any* choice model for the purposes of revenue

[7] Though replacement makes repetitions likely, for large N and $K \ll \sqrt{N!}$, they happen with a vanishing probability.
[8] Any distribution with a density on the K-dim simplex may be picked; we picked uniform for concreteness.

estimation is obtained by a sparse choice model with support scaling as $\log N$. Specifically, let us restrict ourselves to offer sets that are 'small', i.e. bounded by a constant $|\mathcal{M}| \leq C$; this is legitimate from an operational perspective and in line with many of the applications we have described. We now show that *any* customer choice model can be well-approximated by a choice model with *sparse* support for the purpose of evaluating revenue of any offer set $\mathcal{M}$ of size upto C. In particular, we have:

THEOREM 4. *Let λ be an arbitrary given choice model. Then, there exists a choice model $\hat{\lambda}$ with support $O\left(\frac{2C^2 p_{\max}^2}{\varepsilon^2}\left(\log 2C + C\log N\right)\right)$ such that*

$$\max_{\mathcal{M}:|\mathcal{M}|\leq C}\left|R(\mathcal{M}) - \sum_{j\in\mathcal{M}} p_j \hat{\lambda}_j(\mathcal{M})\right| \leq \varepsilon$$

The proof is provided in Appendix A.3. Along with Theorem 3, the above result establishes the potential generality of the signature and linear independence conditions.

In summary, this section visited the issues of explicitly selecting a choice model consistent with the observed data. This is in contrast to our work thus far, which has been simply making revenue predictions. We showed that the robust procedure we used in making revenue predictions may also be seen to yield what is essentially the sparsest choice model consistent with the observed data. Finally, by presenting a family of models for which the sparsest fit to the observed data was unique, and studying the properties of this unique solution, we were able to delineate a data-dependent family of choice models for which the sparsest fit criterion actually yields identification. This formalized the intuitive notion that the complexity of the choice model that can be recovered scales with the "amount" of data that is available.

7. Conclusion and Potential Future Directions

This paper presented a new approach to the problem of using historical sales data to predict expected sales / revenues from offering a particular assortment of products. We depart from traditional parametric approaches to choice modeling in that we assume little more than a weak form of customer rationality; the family of choice models we focus on is essentially the most general family of choice models one may consider. In spite of this generality, we have presented schemes that succeed in producing accurate sales / revenue predictions. We complemented those schemes with extensive empirical studies using both simulated and real-world data, which demonstrated the power of our approach in producing accurate revenue predictions without being prone to over and under fitting. We believe that these schemes are particularly valuable from the standpoint of incorporating models of choice in decision models frequently encountered in operations management. Our schemes are efficient from a computational standpoint and raise the possibility of an entirely 'data-driven' approach to the

modeling of choice for use in those applications. We also discussed some ideas on the problem of identifying sparse or simple models that are consistent with the available marginal information.

With that said, this work cannot be expected to present a panacea for choice modeling problems. In particular, one merit of a structural/ parametric modeling approach to modeling choice is the ability to extrapolate. That is to say, a non-parametric approach such as ours can start making useful predictions about the interactions of a particular product with other products only once *some* data related to that product is observed. With a structural model, one can hope to say useful things about products never seen before. The decision of whether a structural modeling approach is relevant to the problem at hand or whether the approach we offer is a viable alternative thus merits a careful consideration of the context. Of course, as we have discussed earlier, resorting to a parametric approach will typically require expert input on underlying product features that 'matter', and is thus difficult to automate on a large scale. In addition, while our modeling approach is very general, it does not account for competition (see Berry et al. (1995)) or inter-temporal choice behavior (time-dependence in the presence of anticipated discounts or end of the season clearances; see Li et al. (2011))

We believe this paper presents a starting point for a number of research directions. There are numerous directions to pursue from an applications perspective:

1. The focus of this paper has been the estimation of the revenue function $R(\mathcal{M})$ with the rationale that it forms a core subroutine in essentially any revenue optimization problem seeking to optimize revenues in the face of customer choice. A number of generic algorithms (such as local search) can potentially be used in conjunction with the subroutine we provide to solve such optimization problems. It would be interesting to study such a procedure in the context of problems such as network revenue optimization in the presence of customer choice.
2. Having learned a choice model that consists of a distribution over a small number of rank lists, there are a number of qualitative insights one might hope to draw. For instance, using fairly standard statistical machinery, one might hope to ask for the product features that most influence choice from among thousands of potential features by understanding which of these features best rationalize the rank lists learned. In a different direction, one may use the distribution learned as a 'prior', and given further interactions with a given customer infer a distribution specialized to that customer via Bayes rule. This is effectively a means to accomplishing 'collaborative filtering'.

There are also interesting directions to pursue from a theoretical perspective: First, extending our understanding of the limits of identification. In particular, it would be useful to characterize the limits of recoverability for additional families of observable data beyond those discussed in Theorem

3. Second, Theorem 4 points to the existence of sparse approximations to generic choice models. Can we compute such approximations for any choice model but with limited data? Finally, the robust approach in Section 3 presents us with a family of difficult optimization problems for which the present work has presented a generic optimization scheme that is in the spirit of cutting plane approaches. An alternative to this is the development of strong relaxations that yield uniform approximation guarantees (in the spirit of the approximation algorithms literature).

References

S. P. Anderson, A. De Palma, and J. F. Thisse. *Discrete choice theory of product differentiation*. MIT press, Cambridge, MA, 1992.

P. P. Belobaba and C. Hopperstad. Boeing/MIT simulation study: PODS results update. In *1999 AGIFORS Reservations and Yield Management Study Group Symposium, April*, pages 27–30, 1999.

M. E. Ben-Akiva. *Structure of passenger travel demand models*. PhD thesis, Department of Civil Engineering, MIT, 1973.

M. E. Ben-Akiva and S. R. Lerman. *Discrete choice analysis: theory and application to travel demand*. CMIT press, Cambridge, MA, 1985.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

M. Bierlaire. BIOGEME: a free package for the estimation of discrete choice models. In *Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland*, 2003.

M. Bierlaire. An introduction to BIOGEME Version 1.7. 2008.

G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman Rev. Ser. A*, 5:147–151, 1946.

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

G. Calafiore and M. C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.

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.

N. S. Cardell and F. C. Dunbar. Measuring the societal impacts of automobile downsizing. *Transportation Research Part A: General*, 14(5-6):423 – 434, 1980.

S.R. Chandukala, J. Kim, and T. Otter. *Choice Models in Marketing*. Now Publishers Inc, 2008.

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

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

G. Gallego, G. Iyengar, R. Phillips, and A. Dubey. Managing flexible products on a network. Working Paper, 2006.

V. Goyal, R. Levi, and D. Segev. Near-optimal algorithms for the assortment planning problem under dynamic substitution and stochastic demand. Submitted, June 2009.

P. M. Guadagni and J. D. C. Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238, 1983.

D. A. Hensher and W. H. Greene. The mixed logit model: the state of practice. *Transportation*, 30(2): 133–176, 2003.

S. Jagabathula and D. Shah. Inferring rankings under constrained sensing. In *NIPS*, 2008.

A. G. Kök, M. L. Fisher, and R. Vaidyanathan. Assortment planning: Review of literature and industry practice. *Retail Supply Chain Management*, pages 1–55, 2008.

J. Li, N. Granados, and S. Netessine. Are consumers strategic? structural estimation from the air-travel industry. *INSEAD Working Paper No. 2011/104/TOM/ACGRE*, 2011.

R.D. Luce. *Individual choice behavior: A theoretical analysis*. Wiley, New York, 1959.

S. Mahajan and G. J. van Ryzin. On the relationship between inventory costs and variety benefits in retail assortments. *Management Science*, 45(11):1496–1509, 1999.

V. Marzano and A. Papola. On the covariance structure of the cross-nested logit model. *Transportation Research Part B: Methodological*, 42(2):83 – 98, 2008.

A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

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

D. McFadden and K. Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15 (5):447–470, September 2000.

E. H. McKinney. Generalized birthday problem. *American Mathematical Monthly*, pages 385–387, 1966.

F. Mosteller and J. Tukey. Data analysis, including statistics. *The Collected Works of John W. Tukey: Philosophy and principles of data analysis, 1965-1986*, page 601, 1987.

RL Plackett. The analysis of permutations. *Applied Statistics*, 24(2):193–202, 1975. ISSN 0035-9254.

R. M. Ratliff, V. Rao, C. P. Narayan, and K. Yellepeddi. A multi-flight recapture heuristic for estimating unconstrained demand from airline bookings. *Journal of Revenue and Pricing Management*, 7(2): 153–171, 2008a.

R.M. Ratliff, B.V. Rao, C.P. Narayan, and K. Yellepeddi. A multi-flight recapture heuristic for estimating unconstrained demand from airline bookings. *Journal of Revenue & Pricing Management*, 7(2):153–171, 2008b.

P. Rusmevichientong and H. Topaloglu. Robust assortment optimization in revenue management under the multinomial logit choice model. *To appear in Operations Research*, 2011.

P. Rusmevichientong, B. Van Roy, and P. Glynn. A nonparametric approach to multiproduct pricing. *Operations Research*, 54(1), 2006.

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

D. Saure and A. Zeevi. Optimal dynamic assortment planning. Columbia GSB Working Paper, 2009.

K. A. Small. A discrete choice model for ordered alternatives. *Econometrica: Journal of the Econometric Society*, 55(2):409–424, 1987.

K. Talluri and G. J. van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004a.

K. T. Talluri and G. J. van Ryzin. *The Theory and Practice of Revenue Management*. Springer Science+Business Media, 2004b.

G. J. van Ryzin and G. Vulcano. Computing virtual nesting controls for network revenue management under customer choice behavior. *Manufacturing & Service Operations Management*, 10(3):448–467, 2008.

J. von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. *In Contributions to the theory of games*, 2, 1953.

P. Vovsha. Cross-nested logit model: an application to mode choice in the Tel-Aviv metropolitan area. *Transportation Research Record*, 1607:6–15, 1997.

G. Vulcano, G. J. van Ryzin, and R. M. Ratliff. Estimating primary demand for substitutable products from sales transaction data. Working Paper, 2008.

G. Vulcano, G. van Ryzin, and W. Chaar. Om practice—choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management*, 12(3):371–392, 2010.

B. Wierenga. *Handbook of marketing decision models*. Springer Verlag, 2008.

Appendix

A. Proofs for Section 6

A.1. Proof of Theorem 1

The result of Theorem 1 follows immediately from the following lemma, which we prove below.

LEMMA 1. *Let d denote the column rank of matrix A and $\mathcal{Y}$ denote the convex hull of the columns of A. Then, it must be that y belongs to a $d-1$ dimensional subspace, $\|\lambda^{\min}(y)\|_0 \le d+1$, and*

$$\mathrm{vol}_{d-1}(\mathcal{Y}^{\mathrm{sparse}}) = \mathrm{vol}_{d-1}(\mathcal{Y}),$$

where $\mathcal{Y}^{\mathrm{sparse}} \subset \mathcal{Y}$ denotes the set of all data vectors such that

$$\|\lambda^{\min}(y)\|_0 \le d+1 \text{ and } \|\lambda^{\mathrm{sparse}}(y)\|_0 \ge d$$

and $\mathrm{vol}_{d-1}(S)$ denotes the $d-1$ dimensional volume of a set S of points.

Proof of Lemma 1 We prove this lemma in two parts: (1) $\mathcal{Y}$ belongs to a $d-1$ dimensional subspace and $\|\lambda^{\min}(y)\|_0 \le d+1$ for all $y \in \mathcal{Y}$, and (2) $\mathrm{vol}_{d-1}(\mathcal{Y} \setminus \mathcal{Y}^{\mathrm{sparse}}) = 0$.

To prove the first part, note that any data vector $y \in \mathcal{Y}$ belongs to $d-1$ dimensional subspace because A has a d dimensional range space and y belongs to the intersection of the range space of A and the hyperplane $\sum_\sigma \lambda_\sigma = 1$. Let $\tilde{A}$ denote the augmented matrix, which is obtained by augmenting the last row of matrix A with a row of all 1s. Similarly, let $\tilde{y}$ denote the vector obtained by augmenting vector y with 1. The equality constraints of (2) can now be written as $\tilde{y} = \tilde{A}\lambda$, $\lambda \ge 0$. Since A has rank d, the rank of $\tilde{A}$ will be at most $d+1$. Therefore, for any data vector $y \in \mathcal{Y}$, an optimal BFS solution to (2) must be such that

$$\|\lambda^{\min}(y)\|_0 \le d+1, \quad \forall\, y \in \mathcal{Y}. \tag{12}$$

Coming to the second part of the proof, for any $r \le d-1$, let $\mathcal{Y}_r$ denote the set of all data vectors that can be written as a convex combination of at most r columns of matrix A. Let L denote the number of columns of A of size at most r, and let $S_1, S_2, \ldots, S_L$ denote the corresponding subsets of columns of A of size at most r. Then, it is easy to see that $\mathcal{Y}_r$ can be written as the union of disjoint subsets $\mathcal{Y}_r = \cup_{i=1}^{L} \mathcal{Y}_{ri}$, where for each i, $\mathcal{Y}_{ri}$ denotes the set of data vectors that can be written as the convex combination of the columns in subset S_i. For each i, since $\mathcal{Y}_{ri}$ is a polytope residing in $r-1 \le d-2$ dimensional space, it must follow that $\mathrm{vol}_{d-1}(\mathcal{Y}_{ri}) = 0$. Since L is finite, it follows that $\mathrm{vol}_{d-1}(\mathcal{Y}_r) = 0$. Therefore, we can conclude that

$$\mathrm{vol}_{d-1}\left(y \in \mathcal{Y} \colon \|\lambda^{\mathrm{sparse}}(y)\|_0 \le d-1\right) = 0 \tag{13}$$

The result of the lemma now follows from (12) and (13).

A.2. Proof of Theorem 2

Before we prove Theorem 2, we propose a simple combinatorial algorithm that recovers the model λ whenever λ satisfies the signature and linear independence conditions; we make use of this algorithm in the proof of the theorem.

The algorithm recovers λ when the signature and linear independence conditions are satisfied. If the conditions are not satisfied, the algorithm provides a certificate to that effect. The algorithm takes y as an explicit input with the prior knowledge of the structure of A as an auxiliary input. It's aim is to produce λ. In particular, the algorithm outputs the sparsity of λ, $K = \|\lambda\|_0$, permutations $\sigma_1, \ldots, \sigma_K$ so that $\lambda(\sigma_i) \ne 0$, $1 \le i \le K$ and the values $\lambda(\sigma_i)$, $1 \le i \le K$. Without loss of generality, assume that the values $y_1, \ldots, y_m$ are sorted with $y_1 \le \cdots \le y_m$ and further that $\lambda(\sigma_1) \le \lambda(\sigma_2) \le \ldots \lambda(\sigma_K)$.

Before we describe the algorithm, we observe the implication of the two conditions. The *Linear Independence* condition says that for any two non-empty distinct subsets $S, S' \subset \{1, \ldots, K\}$, $S \ne S'$,

$$\sum_{i \in S} \lambda(\sigma_i) \ne \sum_{j \in S'} \lambda(\sigma_j).$$

This means that if we know all $\lambda(\sigma_i)$, $1 \le i \le K$ and since we know $y_d, 1 \le d \le m$, then we can recover $A(\sigma_i)_d, i = 1, 2, \ldots, K$ as the unique solution to $y_d = \sum_{i=1}^{K} A(\sigma_i)_d \lambda(\sigma_i)$ in $\{0,1\}^K$.

Therefore, the non-triviality lies in finding K and $\lambda(\sigma_i)$, $1 \le i \le K$. This issue is resolved by use of the *Signature* condition in conjunction with the above described properties in an appropriate recursive manner.

Specifically, recall that the *Signature* condition implies that for each σ_i for which $\lambda(\sigma_i) \neq 0$, there exists d such that $y_d = \lambda(\sigma_i)$. By *Linear Independence*, it follows that all $\lambda(\sigma_i)$s are distinct and hence by our assumption

$$\lambda(\sigma_1) < \lambda(\sigma_2) < \cdots < \lambda(\sigma_K).$$

Therefore, it must be that the smallest value, y_1 equals $\lambda(\sigma_1)$. Moreover, $A(\sigma_1)_1 = 1$ and $A(\sigma_i)_1 = 0$ for all $i \neq 1$. Next, if $y_2 = y_1$ then it must be that $A(\sigma_1)_2 = 1$ and $A(\sigma_i)_2 = 0$ for all $i \neq 1$. We continue in this fashion until we reach a d' such that $y_{d'-1} = y_1$ but $y_{d'} > y_1$. Using similar reasoning it can be argued that $y_{d'} = \lambda(\sigma_2)$, $A(\sigma_2)_{d'} = 1$ and $A(\sigma_i)_{d'} = 0$ for all $i \neq 2$. Continuing in this fashion and repeating essentially the above argument with appropriate modifications leads to recovery of the sparsity K, the corresponding $\lambda(\sigma_i)$ and $A(\sigma_i)$ for $1 \leq i \leq K$. The complete procedural description of the algorithm is given below.

Sparsest Fit Algorithm:

Initialization: $k(1) = 1$, $d = 1$, $\lambda(\sigma_1) = y_1$ and $A(\sigma_1)_1 = 1$, $A(\sigma_1)_\ell = 0$, $2 \leq \ell \leq m$.
for $d = 2$ to m
  if $y_d = \sum_{i \in T} \lambda(\sigma_i)$ for some $T \subseteq \{1, \ldots, k(d-1)\}$
    $k(d) = k(d-1)$
    $A(\sigma_i)_d = 1 \quad \forall \quad i \in T$
  else
    $k(d) = k(d-1) + 1$
    $\lambda(\sigma_{k(d)}) = y_d$
    $A(\sigma_{k(d)})_d = 1$ and $A(\sigma_{k(d)})_\ell = 0$, for $1 \leq \ell \leq m, \ell \neq d$
  end if
end for
Output $K = k(m)$ and $(\lambda(\sigma_i), A(\sigma_i)), 1 \leq i \leq K$.

Now, we have the following theorem justifying the correctness of the above algorithm:

THEOREM 5. *Suppose we are given $y = A\lambda$ and λ satisfies the "Signature" and the "Linear Independence" conditions. Then, the Sparsest Fit algorithm recovers λ.*

We present the proof of Theorem 2 followed by the proof of Theorem 5.

Proof of Theorem 2 Suppose, to arrive at a contradiction, assume that there exists a distribution μ over the permutations such that $y = A\mu$ and $\|\mu\|_0 \leq \|\lambda\|_0$. Let $v_1, v_2, \ldots, v_K$ and $u_1, u_2, \ldots, u_L$ denote the values that λ and μ take on their respective supports. It follows from our assumption that $L \leq K$. In addition, since λ satisfies the "signature" condition, there exist $1 \leq d(i) \leq m$ such that $y_{d(i)} = v_i$, for all $1 \leq i \leq K$. Thus, since $y = A\mu$, for each $1 \leq i \leq K$, we can write $v_i = \sum_{j \in T(i)} u_j$, for some $T(i) \subseteq \{1, 2, \ldots, L\}$. Equivalently, we can write $v = Bu$, where B is a $0-1$ matrix of dimensions $K \times L$. Consequently, we can also write $\sum_{i=1}^{k} v_i = \sum_{j=1}^{L} \zeta_j u_j$, where ζ_j are integers. This now implies that $\sum_{j=1}^{L} u_j = \sum_{j=1}^{L} \zeta_j u_j$ since $\sum_{i=1}^{K} v_i = \sum_{j=1}^{L} u_j = 1$.

Now, there are two possibilities: either all the ζ_js are > 0 or some of them are equal to zero. In the first case, we prove that μ and λ are identical, and in the second case we arrive at a contradiction. In the case when $\zeta_j > 0$ for all $1 \leq j \leq L$, since $\sum_j u_j = \sum_j \zeta_j u_j$, it should follow that $\zeta_j = 1$ for all $1 \leq j \leq L$. Thus, since $L \leq K$, it should be that $L = K$ and $(u_1, u_2, \ldots, u_L)$ is some permutation of $(v_1, v_2, \ldots, v_K)$. By relabeling the u_js, if required, without loss of generality, we can say that $v_i = u_i$, for $1 \leq i \leq K$. We have now proved that the values of λ and μ are identical. In order to prove that they have identical supports, note that since $v_i = u_i$ and $y = A\lambda = A\mu$, μ must satisfy the "signature" and the "linear independence" conditions. Thus, the algorithm we proposed accurately recovers μ and λ from y. Since the input to the algorithm is only y, it follows that $\lambda = \mu$.

Now, suppose that $\zeta_j = 0$ for some j. Then, it follows that some of the columns in the B matrix are zeros. Removing those columns of B, we can write $v = \tilde{B}\tilde{u}$ where $\tilde{B}$ is B with the zero columns removed and $\tilde{u}$ is u with u_js such that $\zeta_j = 0$ removed. Let $\tilde{L}$ be the size of $\tilde{u}$. Since at least one column was removed $\tilde{L} < L \leq K$. The condition $\tilde{L} < K$ implies that the elements of vector v are not linearly independent i.e., we can find integers c_i such that $\sum_{i=1}^{K} c_i v_i = 0$. This is a contradiction, since this condition violates our "linear independence" assumption. The result of the theorem now follows.

Proof of Theorem 5 Let $\sigma_1, \sigma_2, \ldots, \sigma_K$ be the permutations in the support and $\lambda_1, \lambda_2, \ldots, \lambda_K$ be their corresponding probabilities. Since we assumed that λ satisfies the "signature" condition, for each $1 \le i \le K$, there exists a $d(i)$ such that $y_{d(i)} = \lambda_i$. In addition, the "linear independence" condition guarantees that the condition in the "if" statement of the algorithm is not satisfied whenever $d = d(i)$. To see why, suppose the condition in the "if" statement is true; then, we will have $\lambda_{d(i)} - \sum_{i \in T} \lambda_i = 0$. Since $d(i) \notin T$, this clearly violates the "linear independence" condition. Therefore, the algorithm correctly assigns values to each of the λ_is. We now prove that the $A(\sigma)$s that are returned by the algorithm do indeed correspond to the σ_is. For that, note that the condition in the "if" statement being true implies that y_d is a linear combination of a subset T of the set $\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$. Again, the "linear independence" condition guarantees that such a subset T, if exists, is unique. Thus, when the condition in the "if" statement is true, the only permutations with $A(\sigma)_d = 1$ are the ones in the set T. Similarly, when the condition in the "if" statement is false, then it follows from the "signature" and "linear independence" conditions that only for σ_i, $A(\sigma)_{d(i)} = 1$. From this, we conclude that the algorithm correctly finds the true underlying distribution.

A.3. Proof of Theorem 3

First, we note that, irrespective of the form of observed data, the choice model generated from the "generation model" satisfies the "linear independence" condition with probability 1. The reason is as follows: the values $\lambda(\sigma_i)$ obtained from the generation model are i.i.d uniformly distributed over the interval $[a, b]$. Therefore, the vector $(\lambda(\sigma_1), \lambda(\sigma_2), \ldots, \lambda(\sigma_K))$ corresponds to a point drawn uniformly at random from the hypercube $[a, b]^K$. In addition, the set of points that satisfy $\sum_{i=1}^{K} c_i \lambda(\sigma_i) = 0$ lie in a lower-dimensional space. Since c_is are bounded, there are only finitely many such sets of points. Thus, it follows that with probability 1, the choice model generated satisfies the "linear independence" condition.

The conditions under which the choice model satisfies the "signature" condition depends on the form of observed data. We consider each form separately.

1. Ranking Data: The bound of $K = O(n)$ directly follows from Lemma 2 of Jagabathula and Shah (2008).
2. Comparison Data: For each permutation σ, we truncate its corresponding column vector $A(\sigma)$ to a vector of length $N/2$ by restricting it to only the disjoint unordered pairs: $\{0,1\}, \{2,3\}, \ldots, \{N-2, N-1\}$. Denote the truncated binary vector by $A'(\sigma)$. Let $\tilde{A}$ denote the matrix A with each column $A(\sigma)$ truncated to $A'(\sigma)$. Clearly, since $\tilde{A}$ is just a truncated form of A, it is sufficient to prove that $\tilde{A}$ satisfies the "signature" condition.

   For brevity, let L denote $N/2$, and, given K permutations, let B denote the $L \times K$ matrix formed by restricting the matrix $\tilde{A}$ to the K permutations in the support. Then, it is easy to see that a set of K permutations satisfies the "signature" condition iff there exist K rows in B such that the $K \times K$ matrix formed by the K rows is a permutation matrix.

   Let $R_1, R_2, \ldots, R_J$ denote all the subsets of $\{1, 2, \ldots, m\}$ with cardinality K; clearly, $J = \binom{L}{K}$. In addition, let B^j denote the $K \times K$ matrix formed by the rows of B that are indexed by the elements of R_j. Now, for each $1 \le j \le J$, when we generate the matrix B by choosing K permutations uniformly at random, let $\mathcal{E}_j$ denote the event that the $K \times K$ matrix B^j is a permutation matrix and let $\mathcal{E}$ denote the event $\cup_j \mathcal{E}_j$. We want to prove that $\mathbb{P}(\mathcal{E}) \to 1$ as $N \to \infty$ as long as $K = o(\log N)$. Let X_j denote the indicator variable of the event $\mathcal{E}_j$, and X denote $\sum_j X_j$. Then, it is easy to see that $\mathbb{P}(X = 0) = \mathbb{P}((\mathcal{E})^c)$. Thus, we need to prove that $\mathbb{P}(X = 0) \to 0$ as $N \to \infty$ whenever $K = o(\log n)$. Now, note the following:

$$\mathrm{Var}(X) \ge (0 - \mathbb{E}[X])^2 \mathbb{P}(X = 0)$$

   It thus follows that $\mathbb{P}(X = 0) \le \mathrm{Var}(X)/(\mathbb{E}[X])^2$. We now evaluate $\mathbb{E}[X]$. Since X_js are indicator variables, $\mathbb{E}[X_j] = \mathbb{P}(X_j = 1) = \mathbb{P}(\mathcal{E}_j)$. In order to evaluate $\mathbb{P}(\mathcal{E}_j)$, we restrict our attention to the $K \times K$ matrix B^j. When we generate the entries of matrix B by choosing K permutations uniformly at random, all the elements of B will be i.i.d Be(1/2) i.e., uniform Bernoulli random variables. Therefore, there are 2^{K^2} possible configurations of B^j and each of them occurs with a probability $1/2^{K^2}$. Moreover, there are $K!$ possible $K \times K$ permutation matrices. Thus, $\mathbb{P}(\mathcal{E}_j) = K!/2^{K^2}$. Thus, we have:

$$\mathbb{E}[X] = \sum_{j=1}^{J} \mathbb{E}[X_j] = \sum_{j=1}^{J} \mathbb{P}(\mathcal{E}_j) = \frac{JK!}{2^{K^2}}. \tag{14}$$

Since $J = \binom{L}{K}$, it follows from Stirling's approximation that $J \geq L^K/(eK)^K$. Similarly, we can write $K! \geq K^K/e^K$. It now follows from (14) that

$$\mathbb{E}[X] \geq \frac{L^K}{e^K K^K} \frac{K^K}{e^K} \frac{1}{2^{K^2}} = \frac{L^K}{e^{2K} 2^{K^2}}. \tag{15}$$

We now evaluate $\mathrm{Var}(X)$. Let ρ denote $K!/2^{K^2}$. Then, $\mathbb{E}[X_j] = \rho$ for all $1 \leq j \leq J$. We can write,

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \sum_{i=1}^{J} \sum_{j=1}^{J} \mathbb{P}(X_i = 1, X_j = 1) - J^2\rho^2.$$

Suppose $|R_i \cap R_j| = r$. Then, the number of possible configurations of B^i and B^j is $2^{(2K-r)K}$ because, since there is an overlap of r rows, there are $2K - r$ distinct rows and, of course, K columns. Since all configurations occur with the same probability, it follows that each configuration occurs with a probability $1/2^{(2K-r)K}$, which can also be written as $2^{rK}\rho^2/(K!)^2$. Moreover, the number of configurations in which both B^i and B^j are permutation matrices is equal to $K!(K-r)!$, since, fixing the configuration of B^i will leave only $K - r$ rows of B^j to be fixed.

For a fixed R_i, we now count the number of subsets R_j such that $|R_i \cap R_j| = r$. We construct an R_j by first choosing r rows from R_i and then choosing the rest from $\{1, 2, \ldots, l\} \setminus R_i$. We can choose r rows from the subset R_i of K rows in $\binom{K}{r}$ ways, and the remaining $K - r$ rows in $\binom{L-K}{K-r}$ ways. Therefore, we can now write:

$$\begin{aligned}
\sum_{j=1}^{J} \mathbb{P}(X_i = 1, X_j = 1) &= \sum_{r=0}^{K} \binom{K}{r}\binom{L-K}{K-r} K!(K-r)! \frac{2^{rK}\rho^2}{(K!)^2} \\
&\leq \rho^2 \sum_{r=0}^{K} \binom{L}{K-r} \frac{2^{rK}}{r!}, \quad \text{Using } \binom{L-K}{K-r} \leq \binom{L}{K-r} \\
&= \binom{L}{K}\rho^2 + \rho^2 \sum_{r=1}^{K} \binom{L}{K-r} \frac{2^{rK}}{r!} \\
&\leq J\rho^2 + \rho^2 L^K \sum_{r=1}^{K} \left(\frac{e2^K}{L}\right)^r \frac{1}{r^r (K-r)^{K-r}}
\end{aligned}$$

The last inequality follows from Stirling's approximation: $\binom{L}{K-r} \leq (L/(K-r))^{K-r}$ and $r! \geq (r/e)^r$; in addition, we have used $J = \binom{L}{K}$. Now consider

$$\begin{aligned}
r^r (K-r)^{K-r} &= \exp\{r\log r + (K-r)\log(K-r)\} \\
&= \exp\{K\log K - KH(r/K)\} \\
&\geq \frac{K^K}{2^K}
\end{aligned}$$

where $H(x)$ is the Shannon entropy of the random variable distributed as $\mathrm{Be}(x)$, defined as $H(x) = -x\log x - (1-x)\log(1-x)$ for $0 < x < 1$. The last inequality follows from the fact that $H(x) \leq \log 2$ for all $0 < x < 1$. Putting everything together, we get

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=1}^{J} \left[\sum_{j=1}^{J} \mathbb{P}(X_i = 1, X_j = 1)\right] - \mathbb{E}[X]^2 \\
&\leq J\left[J\rho^2 + \rho^2 L^K \frac{2^K}{K^K} \sum_{r=1}^{K} \left(\frac{e2^K}{L}\right)^r\right] - J^2\rho^2 \\
&= \frac{J\rho^2 2^K L^K}{K^K} \sum_{r=1}^{K} \left(\frac{e2^K}{L}\right)^r
\end{aligned}$$

We can now write,

$$\begin{aligned}
\mathbb{P}(X=0) &\leq \frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2} \\
&\leq \frac{1}{J^2\rho^2}\frac{J\rho^2 2^K L^K}{K^K}\sum_{r=1}^{K}\left(\frac{e2^K}{L}\right)^r \\
&= \frac{1}{J}\,\frac{2^K L^K}{K^K}\,\frac{e2^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r \\
&\leq \frac{e^K K^K}{L^K}\,\frac{2^K L^K}{K^K}\,\frac{e2^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r, \quad \text{Using } J=\binom{L}{K}\leq\left(\frac{L}{eK}\right)^K \\
&= e\frac{(4e)^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r
\end{aligned}$$

It now follows that for $K = o(\log L/\log(4e))$, $\mathbb{P}(X=0)\to 0$ as $N\to\infty$. Since, by definition, $L = N/2$, this completes the proof of the theorem.

3. Top Set Data: For this type of data, note that it is sufficient to prove that $A^{(1)}$ satisfies the "signature" property with a high probability; therefore, we ignore the comparison data and focus only on the data corresponding to the fraction of customers that have product i as their top choice, for every product i. For brevity, we abuse the notation and denote $A^{(1)}$ by A and $y^{(1)}$ by y. Clearly, y is of length N and so is each column vector $A(\sigma)$. Every permutation σ ranks only one product in the first position. Hence, for every permutation σ, exactly one element of the column vector $A(\sigma)$ is 1 and the rest are zeros.

   In order to obtain a bound on the support size, we reduce this problem to a balls-and-bins setup. For that, imagine K balls being thrown uniformly at random into N bins. In our setup, the K balls correspond to the K permutations in the support and the N bins correspond to the N products. A ball is thrown into bin i provided the permutation corresponding to the ball ranks product i to position 1. Our "generation model" chooses permutations independently; hence, the balls are thrown independently. In addition, a permutation chosen uniformly at random ranks a given product i to position 1 with probability $1/N$. Therefore, each ball is thrown uniformly at random.

   In the balls-and-bins setup, the "signature" condition translates into all K balls falling into different bins. By "Birthday Paradox" McKinney (1966), the K balls falls into different bins with a high probability provided $K = o(\sqrt{N})$.

This finishes the proof of the theorem.

A.4. Proof of Theorem 4

To show existence of a choice model $\hat{\lambda}$ with sparse support, that approximates expected revenue of all offer sets of size at most C with respect to the true model, we shall utilize the probabilistic method. Specifically, consider M samples chosen as per the true choice model λ : let these be $\sigma_1,\ldots,\sigma_M$. Let $\hat{\lambda}$ be the empirical choice model (or distribution on permutations) induced by these M samples. We shall show that for M large enough (as claimed in the statement of Theorem 4), this empirical distribution $\hat{\lambda}$ satisfies the desired properties with positive probability. That is, there exists a distribution with sparse support that satisfies the desired property and hence implies Theorem 4.

To this end, consider an offer set $\mathcal{M}$ of size at most C. As noted earlier, the expected revenue $R(\mathcal{M})$ is given by

$$R(\mathcal{M}) = \sum_{j\in\mathcal{M}} p_j\lambda_j(\mathcal{M}),$$

where p_j is the price of product $j\in\mathcal{M}$ and $\lambda_j(\mathcal{M})$ is the probability of customer choosing j to purchase, i.e. $\lambda(\mathcal{S}_j(\mathcal{M}))$. We wish to show that $\hat{\lambda}_j(\mathcal{M}) = \hat{\lambda}(\mathcal{S}_j(\mathcal{M}))$ is good approximation of $\lambda_j(\mathcal{M})$, for all $j\in\mathcal{M}$ and for all $\mathcal{M}$ of size at most C. To show this, we shall use a combination of Chernoff/Hoeffding bound and union bound.

To this end, consider the given $\mathcal{M}$ and a fixed $j\in\mathcal{M}$. For $1\leq\ell\leq M$, define

$$X_\ell^j = \begin{cases} 1 & \text{if } \sigma_\ell\in\mathcal{S}_j(\mathcal{M}), \\ 0 & \text{otherwise.}\end{cases}$$

Then, $X_\ell^j,\ 1 \le \ell \le M$, are independent and identically distributed Bernoulli random variables with $\mathbb{P}(X_\ell^j = 1) = \lambda^j(\mathcal{M})$. By definition,

$$\hat{\lambda}^j(\mathcal{M}) = \frac{1}{M}\sum_{\ell=1}^{M} X_\ell^j. \tag{16}$$

Using (16) and Chernoff/Hoeffding bound for $\sum_{\ell=1}^{M} X_\ell^j$, it follows that for any $t > 0$,

$$\mathbb{P}\left(\left|\hat{\lambda}^j(\mathcal{M}) - \lambda^j(\mathcal{M})\right| > t\right) \le 2\exp\left(-\frac{t^2 M}{2}\right). \tag{17}$$

Let $p_{\max} = \max_{i=1}^{N} p_i$. By selecting, $t = \frac{\varepsilon}{C p_{\max}}$ in (17) we have

$$\mathbb{P}\left(\left|\hat{\lambda}^j(\mathcal{M}) - \lambda^j(\mathcal{M})\right| > \frac{\varepsilon}{C p_{\max}}\right) \le 2\exp\left(-\frac{\varepsilon^2 M}{2C^2 p_{\max}^2}\right). \tag{18}$$

Therefore, for the given $\mathcal{M}$ of size at most C, by union bound we have

$$\mathbb{P}\left(\left|\sum_{j\in\mathcal{M}} p_j \hat{\lambda}^j(\mathcal{M}) - \sum_{j\in\mathcal{M}} p_j \lambda^j(\mathcal{M})\right| > \varepsilon\right) \le 2C\exp\left(-\frac{\varepsilon^2 M}{2C^2 p_{\max}^2}\right). \tag{19}$$

There are at most N^C sets of size upto C. Therefore, by union bound and (19) it follows that

$$\mathbb{P}\left(\max_{\mathcal{M}:|\mathcal{M}|\le C}\left|R(\mathcal{M}) - \sum_{j\in\mathcal{M}} p_j \hat{\lambda}_j(\mathcal{M})\right| > \varepsilon\right) \le 2CN^C\exp\left(-\frac{\varepsilon^2 M}{2C^2 p_{\max}^2}\right). \tag{20}$$

For choice of M such that

$$M > \frac{2C^2 p_{\max}^2}{\varepsilon^2}\left(\log 2C + C\log N\right),$$

the right hand size of (20) becomes < 1. This establishes the desired result.

B. The Exact Approach to Solving the Robust Problem

Here we provide further details on the second approach described for the solution to the (dual of) the robust problem (4). In particular, we first consider the case of ranking data, where an efficient representation of the constraints in the dual may be produced. We then illustrate a method that produces a sequence of 'outer-approximations' to (6) for general types of data, and thereby allows us to produce a sequence of improving lower bounding approximations to our robust revenue estimation problem, (2). This provides a general procedure to address the task of solving (4), or equivalently, (2).

B.1. A Canonical Representation for Ranking Data

Recall the definition of *ranking data* from Section 2: This data yields the fraction of customers that rank a given product i as their rth choice. Thus, the partial information vector y is indexed by i, r with $0 \le i, r \le N$. For each i, r, y_{ri} denotes the probability that product i is ranked at position r. The matrix A is thus in $\{0,1\}^{N^2 \times N!}$ and for a column of A corresponding to the permutation σ, $A(\sigma)$, we will thus have $A(\sigma)_{ri} = 1$ iff $\sigma(i) = r$. We will now construct an efficient representation of the type (6) for this type of data.

Consider partitioning $\mathcal{S}_j(\mathcal{M})$ into $D_j = N$ sets wherein the dth set is given by

$$\mathcal{S}_{jd}(\mathcal{M}) = \{\sigma \in \mathcal{S}_j(\mathcal{M}) \colon \sigma(j) = d\}.$$

and define, as usual, $\mathcal{A}_{jd}(\mathcal{M}) = \{A(\sigma) : \sigma \in \mathcal{S}_{jd}(\mathcal{M})\}$. Thus, $\mathcal{A}_{jd}(\mathcal{M})$ is the set of columns of A whose corresponding permutations rank the jth product as the dth most preferred choice.

It is easily seen that the set $\mathcal{A}_{jd}(\mathcal{M})$ is equal to the set of all vectors x^{jd} in $\{0,1\}^{N^2}$ satisfying:

$$\begin{aligned}
&\sum_{i=0}^{N-1} x_{ri}^{jd} = 1 \quad \text{for } 0 \le r \le N-1\\
&\sum_{r=0}^{N-1} x_{ri}^{jd} = 1 \quad \text{for } 0 \le i \le N-1\\
&x_{ri}^{jd} \in \{0,1\} \quad \text{for } 0 \le i, r \le N-1.\\
&x_{dj}^{jd} = 1\\
&x_{d'i}^{jd} = 0 \qquad \text{for all } i \in \mathcal{M}, i \ne j \text{and } 0 \le d' < d.
\end{aligned} \tag{21}$$

The first three constraints in (21) enforce the fact that x^{jd} represents a valid permutation. The penultimate constraint requires that the permutation encoded by x^{jd}, say σ^{jd}, satisfies $\sigma^{jd}(j) = d$. The last constraint simply ensures that $\sigma^{jd} \in \mathcal{S}_j(\mathcal{M})$.

Our goal is, of course, to find a description for $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ of the type (6). Now consider replacing the third (integrality) constraint in (21)

$$x^{jd}_{ri} \in \{0,1\} \quad \text{for } 0 \le i, r \le N-1$$

with simply the non-negativity constraint

$$x^{jd}_{ri} \ge 0 \quad \text{for } 0 \le i, r \le N-1$$

We claim that the resulting polytope is precisely the convex hull of $\mathcal{A}_{jd}(\mathcal{M}), \bar{\mathcal{A}}_{jd}(\mathcal{M})$. To see this, we note that all feasible points for the resulting polytope satisfy the first, second, fourth and fifth constraint of (21). Further, the polytope is integral, being the projection of a matching polytope with some variables forced to be integers (Birkhoff (1946), von Neumann (1953)), so that any feasible solution must also satisfy the third constraint of (21). We consequently have an *efficient* canonical representation of the type (6), which via (8) yields, in turn, an efficient solution to our robust revenue estimation problem (2) for ranking data, which we now describe for completeness.

Let us define for convenience the set $\mathcal{V}(\mathcal{M}) = \{(j,d) : j \in \mathcal{M}, 0 \le d \le N-1\}$, and for each pair (j,d), the sets $\mathcal{B}(j,d,\mathcal{M}) = \{(i,d') : i \in \mathcal{M}, i \ne j, 0 \le d' < d\}$. Then, specializing (8) to the canonical representation just proposed, we have that the following simple program in the variables α, ν and $\gamma^{jd} \in \mathbb{R}^{2N}$ is, in fact, equivalent to (2) for ranking data:

$$\begin{array}{ll} \underset{\alpha,\nu}{\text{maximize}} & \alpha^\top y + \nu \\ \text{subject to} & \gamma^{jd}_i + \gamma^{jd}_{N+r} \ge \alpha_{ri} \qquad \text{for all } (j,d) \in \mathcal{V}(\mathcal{M}), (i,r) \notin \mathcal{B}(j,d,\mathcal{M}) \\ & \sum_{i \ne j} \gamma^{jd}_i + \sum_{r \ne d} \gamma^{jd}_{N+r} + \nu \le p_j - \alpha_{dj} \quad \text{for all } (j,d) \in \mathcal{V}(\mathcal{M}) \end{array} \tag{22}$$

B.2. Computing a Canonical Representation: The General Case

While it is typically quite easy to 'write down' a description of the sets $\mathcal{A}_{jd}(\mathcal{M})$ as all integer solutions to some set of linear inequalities (as we did for the case of ranking data), relaxing this integrality requirement will typically *not* yield the convex hull of $\mathcal{A}_{jd}(\mathcal{M})$. In this section we describe a procedure that starting with the former (easy to obtain) description, solves a sequence of linear programs that yield improving solutions. More formally, we assume a description of the sets $\mathcal{A}_{jd}(\mathcal{M})$ of the type

$$\mathcal{I}_{jd}(\mathcal{M}) = \{x^{jd} : A^{jd}_1 x^{jd} \ge b^{jd}_1, \quad A^{jd}_2 x^{jd} = b^{jd}_2, \quad A^{jd}_3 x^{jd} \le b^{jd}_3, \quad x^{jd} \in \{0,1\}^m\} \tag{23}$$

This is similar to (6), with the important exception that we now allow integrality constraints. Given a set $\mathcal{I}_{jd}(\mathcal{M})$ we let $\bar{\mathcal{I}}^0_{jd}(\mathcal{M})$ denote the polytope obtained by relaxing the requirement $x^{jd} \in \{0,1\}^m$ to simply $x^{jd} \ge 0$. In the case of ranking data, $\bar{\mathcal{I}}^o_{jd}(\mathcal{M}) = \text{conv}(\mathcal{I}_{jd}(\mathcal{M})) = \bar{\mathcal{A}}_{jd}(\mathcal{M})$ and we were done; we begin with an example where this is not the case.

Example 1. Recall the definition of *comparison data* from Section 2. In particular, this data yields the fraction of customers that prefer a given product i to a product j. The partial information vector y is thus indexed by i,j with $0 \le i,j \le N; i \ne j$ and 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 $\sigma(i) < \sigma(j)$.

Consider $\mathcal{S}_j(\mathcal{M})$, the set of all permutations that would result in a purchase of j assuming $\mathcal{M}$ is the set of offered products. 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:

$$\begin{array}{ll} x^j_{il} \ge x^j_{ik} + x^j_{kl} - 1 & \text{for all } i,k,l \in \mathcal{N}, i \ne k \ne l \\ x^j_{ik} + x^j_{ki} = 1 & \text{for all } i,k \in \mathcal{N}, i \ne k \\ x^j_{ji} = 1 & \text{for all } i \in \mathcal{M}, i \ne j \\ x^j_{ik} \in \{0,1\} & \text{for all } i,k \in \mathcal{N}, i \ne k \end{array} \tag{24}$$

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 $i \in \mathcal{M}$. Thus, (24) is a description of the type (23) with $D_j = 1$ for all j. Now consider the polytope $\bar{\mathcal{I}}^o_j(\mathcal{M})$ obtained by relaxing the fourth (integrality)

constraint to simply $x^j_{ik} \geq 0$. Of course, we must have $\bar{\mathcal{I}}^o_j(\mathcal{M}) \supseteq \text{conv}(\mathcal{I}_j(\mathcal{M})) = \text{conv}(\mathcal{A}_j(\mathcal{M}))$. Unlike the case of ranking data, however, $\bar{\mathcal{I}}^o_j(\mathcal{M})$ can in fact be shown to be *non*-integral [9], so that $\bar{\mathcal{I}}^o_j(\mathcal{M}) \neq \text{conv}(\mathcal{A}_j(\mathcal{M}))$ in general.

We next present a procedure that starting with a description of the form in (23), solves a sequence of linear programs each of which yield improving solutions to (2) along with bounds on the quality of the approximation:

1. Solve (8) using $\bar{\mathcal{I}}^o_{jd}(\mathcal{M})$ in place of $\text{conv}(\mathcal{I}_{jd}(\mathcal{M})) = \bar{\mathcal{A}}_{jd}(\mathcal{M})$. This yields a lower bound on (2) since $\bar{\mathcal{I}}^o_{jd}(\mathcal{M}) \supset \bar{\mathcal{A}}_{jd}(\mathcal{M})$. Call the corresponding solution $\alpha_{(1)}, \nu_{(1)}$.
2. Solve the optimization problem $\max \alpha^\top_{(1)} x^{jd}$ subject to $x^{jd} \in \bar{\mathcal{I}}^o_{jd}(\mathcal{M})$ for each pair (j,d). If the optimal solution $\hat{x}^{jd}$ is integral for each (j,d), then stop; the solution computed in the first step is in fact optimal.
3. Otherwise, let $\hat{x}^{jd}$ possess a non-integral component for some (j,d); say $\hat{x}^{jd}_c \in (0,1)$. Partition $\mathcal{A}_{jd}(\mathcal{M})$ on this variable - i.e. define

$$\mathcal{A}_{jd_0}(\mathcal{M}) = \{A(\sigma) : A(\sigma) \in \mathcal{A}_{jd}(\mathcal{M}), A(\sigma)_c = 0\}$$

and

$$\mathcal{A}_{jd_1}(\mathcal{M}) = \{A(\sigma) : A(\sigma) \in \mathcal{A}_{jd}(\mathcal{M}), A(\sigma)_c = 1\},$$

and let $\mathcal{I}_{jd_0}(\mathcal{M})$ and $\mathcal{I}_{jd_1}(\mathcal{M})$ represent the corresponding sets of linear inequalities with integer constraints (i.e. the projections of $\mathcal{I}_{jd}(\mathcal{M})$ obtained by restricting x^{jd}_c to be 0 and 1 respectively). Of, course, these sets remain of the form in (23). Replace $\mathcal{I}_{jd}(\mathcal{M})$ with $\mathcal{I}_{jd_0}(\mathcal{M})$ and $\mathcal{I}_{jd_1}(\mathcal{M})$ and go to step 1.

The above procedure is akin to a cutting plane method and is clearly finite, but the size of the LP we solve increases (by up to a factor of 2) at each iteration. Nonetheless, each iteration produces a lower bound to (2) whose quality is easily measured (for instance, by solving the maximization version of (2) using the same procedure, or by sampling constraints in the program (4) and solving the resulting program in order to produce an upper bound on (2)). Moreover, the quality of our solution improves with each iteration. In our computational experiments with a related type of data, it sufficed to stop after a single iteration of the above procedure.

B.3. Explicit LP solved for censored comparison data in Section 4

The LP we want to solve is

$$\begin{array}{ll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \tag{25}$$

For the 'censored' comparison data, the partial information vector is indexed by i,j with $0 \leq i,j \leq N-1$, $i \neq j$. For each i,j such that $i \neq 0$, y_{ij} denotes the fraction of customers that prefer product i to both products j and 0; in other words, y_{ij} denotes the fraction of customers that purchase product i when then offer set is $\{i,j,0\}$. Further, for each $j \neq 0$, y_{0j} denotes the fraction of customers who prefer the 'no-purchase' option to product j; in fact, y_{0j} is the fraction of customers who don't purchase anything when the set $\{j,0\}$ is on offer. The matrix A is then in $\{0,1\}^{N(N-1)}$, with the column of A corresponding to permutation σ, $A(\sigma)$, having $A(\sigma)_{ij} = 1$ if $\sigma(i) < \sigma(j)$ and $\sigma(i) < \sigma(0)$ for each $i \neq 0, j$, and $A(\sigma)_{0j} = 1$ if $\sigma(0) < \sigma(j)$ for $j \neq 0$, and $A(\sigma)_{ij} = 0$ otherwise.

For reasons that will become apparent soon, we modify the LP in (25) by replacing the constraint $A\lambda = y$ with $A\lambda \geq y$. It is now easy to see the following:

$$\begin{array}{ll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda \geq y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \leq \begin{array}{ll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \tag{26}$$

[9] for $N \geq 5$; the polytope can be shown to be integral for $N \leq 4$

We now take the dual of the modified LP. In order to do that, recall from section 3 that $\mathcal{S}_j(\mathcal{M}) = \{\sigma \in S_N : \sigma(j) < \sigma(i), \forall i \in \mathcal{M}, i \neq j\}$ denotes the set of all permutations that result in the purchase of the product $j \in \mathcal{M}$ when the offered assortment is $\mathcal{M}$. In addition, $\mathcal{A}_j(\mathcal{M})$ denotes the set $\{A(\sigma) : \sigma \in \mathcal{S}_j(\mathcal{M})\}$. Now, the dual of the modified LP is

$$\begin{array}{ll} \underset{\alpha,\nu}{\text{maximize}} & \alpha^\top y + \nu \\ \text{subject to} & \max\limits_{z^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, \text{ for each } j \in \mathcal{M} \\ & \alpha \geq 0. \end{array} \tag{27}$$

where α and ν are dual variables corresponding respectively to the data consistency constraints $A\lambda = y$ and the requirement that λ is a probability distribution (i.e. $\mathbf{1}^\top \lambda = 1$) respectively.

Now, consider the following representation of the set $\mathcal{A}_j(\mathcal{M})$, for a fixed j.

$$\begin{array}{ll} z^j_{ik} = \min\left\{x^j_{ik}, x^j_{i0}\right\} & \text{for all } i,k \in \mathcal{N}, i \neq k, i \neq 0 \\ z^j_{0k} = x^j_{0k} & \text{for all } k \in \mathcal{N}, k \neq 0 \\ z^j_{ik} \in \{0,1\} & \text{for all } i,k \in \mathcal{N}, i \neq k \\ x^j_{il} \geq x^j_{ik} + x^j_{kl} - 1 & \text{for all } i,k,l \in \mathcal{N}, i \neq k \neq l \\ x^j_{ik} + x^j_{ki} = 1 & \text{for all } i,k \in \mathcal{N}, i \neq k \\ x^j_{ji} = 1 & \text{for all } i \in \mathcal{M}, i \neq j \\ x^j_{ik} \in \{0,1\} & \text{for all } i,k \in \mathcal{N}, i \neq k \end{array} \tag{28}$$

The last four constraints are the same as the set of inequalities in (24), which correspond to the representation of the set $\mathcal{A}_j(\mathcal{M})$ for comparison data; thus, every point satisfying the set of last four constraints in (28) corresponds to a permutation $\sigma \in \mathcal{S}_j(\mathcal{M})$ such that $x^j_{ik} = 1$ if and only if $\sigma(i) < \sigma(k)$. We now claim that the set of points z^j that satisfy the constraints in (28) is equal to the set of vectors in $\mathcal{A}_j(\mathcal{M})$. To see that, note that $z^j_{ik} = 1$ if and only if the corresponding $x^j_{ik} = 1$ *and* $x^j_{i0} = 1$, for $i \neq 0$. This implies that $z^j_{ik} = 1$ if and only if i is preferred to k *and* i is preferred to 0. Similarly, $z^j_{0k} = 1$ if and only if $x^j_{0k} = 1$ i.e., 0 is preferred to k.

Let $\bar{\mathcal{I}}_j(\mathcal{M})$ denote the convex hull of the vectors in $\mathcal{A}_j(\mathcal{M})$, equivalently, of the vectors z^j satisfying the set of constraints in (28). Let $\bar{\mathcal{I}}^o_j(\mathcal{M})$ be the convex hull of the vectors z^j satisfying the constraints in (28) with the constraint $z^j_{ik} = \min\left\{x^j_{ik}, x^j_{i0}\right\}$ replaced by the constraints $z^j_{ik} \leq x^j_{ik}$ and $z^j_{ik} \leq x^j_{i0}$, and the constraint $z^j_{0k} = x^j_{0k}$ replaced by the constraint $z^j_{0k} \leq x^j_{0k}$. Finally, let $\bar{\mathcal{I}}^1_j(\mathcal{M})$ represent the polytope $\bar{\mathcal{I}}^0_j(\mathcal{M})$ with the integrality constraints relaxed to $z^j_{ik} \geq 0$ and $x^j_{ik} \geq 0$. We now have the following relationships:

$$\begin{aligned} \left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\} &= \left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}^o_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\} \\ &\supseteq \left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}^1_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\} \end{aligned} \tag{29}$$

The first equality follows because $\alpha \geq 0$ and, hence, at the optimal solution, $z^j_{ik} = 1$ if $x^j_{ik} = x^j_{i0} = 1$, and $z^j_{0k} = 1$ if $x^j_{0k} = 1$. It should be now clear that in order to establish this equality we considered the modified LP. The second relationship follows because of the relaxation of constraints. It now follows from (26), (27) and (29) that

$$\begin{array}{lll} \underset{\lambda}{\text{minimize}} & \sum\limits_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) & \underset{\alpha,\nu}{\text{maximize}}\ \alpha^\top y + \nu \\ \text{subject to} & A\lambda = y, & \geq \text{subject to} \max\limits_{z^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, \text{ for each } j \in \mathcal{M} \\ & \mathbf{1}^\top \lambda = 1, & \qquad \alpha \geq 0. \\ & \lambda \geq 0. & \underset{\alpha,\nu}{\text{maximize}}\ \alpha^\top y + \nu \\ & & \geq \text{subject to} \max\limits_{z^j \in \bar{\mathcal{I}}^1_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, \text{ for each } j \in \mathcal{M} \\ & & \qquad \alpha \geq 0. \end{array} \tag{30}$$

Using the procedure described in Section 3.3, we solve the last LP in (30) by taking the dual of the constraint in the LP. For convenience, we write out the program $\max_{z^j \in \mathcal{I}^1_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right)$ and the corresponding

dual variables we use for each of the constraints.

$$\begin{array}{lll}
\underset{z^j}{\text{maximize}} & \alpha^\top z^j + \nu & \\
\text{subject to} & & \text{Dual Variables} \\
z^j_{ik} - x^j_{ik} \le 0 & \text{for all } i,k \in \mathcal{N}, i \neq k & \Omega 1^j_{ik} \\
z^j_{ik} - x^j_{i0} \le 0 & \text{for all } i,k \in \mathcal{N}, i \neq k, i \neq 0 & \Omega 2^j_{ik} \\
x^j_{ik} + x^j_{kl} - x^j_{il} \le 1 & \text{for all } \ i,k,l \in \mathcal{N}, i \neq k \neq l & \Gamma^j_{ikl} \\
x^j_{ik} + x^j_{ki} = 1 & \text{for all } i,k \in \mathcal{N}, i < k & \Delta^j_{ik} \\
x^j_{ji} = 1 & \text{for all } i \in \mathcal{M}, i \neq j & \Theta^j_i \\
x^j_{ik}, z^j_{ik} \ge 0 & \text{for all } i,k \in \mathcal{N}, i \neq k &
\end{array} \tag{31}$$

Let P denote the set $\{(i,k)\colon i \neq k, 0 \le i,k \le N-1\}$, and T denote the set $\{(i,k,l)\colon i \neq k \neq l, 0 \le i,k,l \le N-1\}$. Moreover, let $g(a,b,k,j)$ denote $\sum_{k\in\mathcal{N},k\neq a,b} \Gamma^j_{abk} + \sum_{k\in\mathcal{N},k\neq a,b} \Gamma^j_{kab} - \sum_{k\in\mathcal{N},k\neq a,b} \Gamma^j_{akb}$. Then, the LP we solve is

$$\underset{\nu,\alpha}{\text{maximize}} \quad \nu + \textstyle\sum_{(i,k)\in P} \alpha_{ik} y_{ik}$$

subject to

$$\begin{array}{ll}
\displaystyle\sum_{(i,k,l)\in T} \Gamma^j_{ikl} + \sum_{(i,k)\in P, i<k} \Delta^j_{ik} + \sum_{i\in\mathcal{M}, i\neq j} \Theta^j_i \le p_j - \nu \ \forall\, j \in \mathcal{M} & \\
g(a,b,k,j) + \Delta^j_{ab} - \Omega 1^j_{ab} \ge 0 & \forall\ j \in \mathcal{M}, a,b \in \mathcal{N}, a < b;\ \text{if } a = j, b \notin \mathcal{M} \\
g(a,b,k,j) + \Delta^j_{ba} - \Omega 1^j_{ab} \ge 0 & \forall\ j \in \mathcal{M}, a,b \in \mathcal{N}, a > b, b \neq 0;\ \text{if } a = j, b \notin \mathcal{M} \\
g(a,b,k,j) + \Delta^j_{ab} + \Theta^j_b - \Omega 1^j_{ab} \ge 0 & \forall\ j \in \mathcal{M}, a = j, b \in \mathcal{M}, a < b \\
g(a,b,k,j) + \Delta^j_{ba} + \Theta^j_b - \Omega 1^j_{ab} \ge 0 & \forall\ j \in \mathcal{M}, a = j, b \in \mathcal{M}, a > b, b \neq 0 \\
g(a,b,k,j) + \Delta^j_{ba} - \sum_{k\in\mathcal{N},k\neq a} \Omega 2^j_{ak} \ge 0 & \forall\ j \in \mathcal{M}, a \in \mathcal{N}, a \neq j, b = 0 \\
g(a,b,k,j) + \Delta^j_{ba} + \Theta^j_b - \sum_{k\in\mathcal{N},k\neq a} \Omega 2^j_{ak} \ge 0 & \forall\ j \in \mathcal{M}, a = j, b = 0 \\
\Omega 1^j_{ab} + \Omega 2^j_{ab} \ge \alpha_{ab} & \forall\ j \in \mathcal{M}, a,b \in P, a \neq 0, b \neq 0 \\
\Omega 2^j_{ab} \ge \alpha_{ab} & \forall\ j \in \mathcal{M}, a \in \mathcal{N} \setminus \{0\}, b = 0 \\
\Omega 1^j_{ab} \ge \alpha_{ab} & \forall\ j \in \mathcal{M}, a = 0, b \in \mathcal{N} \setminus \{0\} \\
\alpha, \Gamma, \Omega 1, \Omega 2 \ge 0. &
\end{array} \tag{32}$$

C. An Overview of Common Structural Models

Here we give a brief overview of each the parametric choice models we compare our approach with. The descriptions we provide are brief and we refer an interested reader to Ben-Akiva and Lerman (1985) for more details.

C.1. Multinomial logit (MNL) family

The MNL model is a popular and most commonly used parametric model in economics, marketing and operations management (see Ben-Akiva and Lerman (1985), Anderson et al. (1992)) and is the canonical example of a *Random Utility Model*. In the MNL model, the utility of the customer from product j takes the form $U_j = V_j + \xi_j$, where V_j is the deterministic component and the error terms $\xi_0, \xi_1, \ldots, \xi_{N-1}$ are i.i.d. random variables having a Gumbel distribution with location parameter 0 and scale parameter 1. Since only differences of utilities matter, without loss of generality, it is assumed that the mean utility of the "no-purchase" option, $V_0 = 0$. Let w_j denote e^{μ_j}; then, according to the MNL model, the probability that product j is purchased from an assortment $\mathcal{M}$ is given by

$$\mathbb{P}(j|\mathcal{M}) = w_j / \sum_{i\in\mathcal{M}} w_i.$$

A major advantage of the MNL model is that it is analytically tractable. In particular, it has a closed form expression for the choice probabilities. However, it has several shortcomings. One of the major limitations of the MNL model is that it exhibits Independent of Irrelevant Alternatives (IIA) property i.e., the relative likelihood of the purchase of any two given product variants is independent of the other products on offer. This property may be undesirable in contexts where some product are 'more like' other products so that the

randomness in a given customers utility is potentially correlated across products. There are other – more complicated – variants that have been proposed to alleviate the IIA issue – the most popular being the NL model, which we describe next.

C.2. Nested logit family (NL)

The nested logit (NL) family of models, first derived by Ben-Akiva (1973), was designed to explicitly capture the presence of shared unobserved attributes among alternatives. In particular, the universe of products is partitioned into L mutually exclusive subsets called *nests* denoted by $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_L$ such that

$$\mathcal{N} = \bigcup_{\ell=1}^{L} \mathcal{N}_\ell \qquad \text{and} \qquad \mathcal{N}_\ell \cap \mathcal{N}_m = \emptyset, \text{ for } m \neq \ell.$$

The partitioning is such that products sharing unobserved attributes lie in the same nest. Each customer has utility U_j for product j given by $U_j = V_j + \xi_\ell + \xi_{j,\ell}$; here, ξ_ℓ is the error term shared by all the products in nest $\mathcal{N}_\ell$, and $\xi_{j,\ell}$ is the error term that is product specific and assumed to be i.i.d across different products. In this logit case, $\xi_{j,\ell}$ are assumed to be i.i.d standard Gumbel distributed with location parameter 0 and scale parameter 1. The nest specific error terms $\xi_1, \xi_2, \ldots, \xi_L$ are assumed i.i.d., distributed such that for each ℓ, j, $\xi_\ell + \xi_{j,\ell}$ is Gumbel distributed with location parameter 0 and scale parameter $\rho < 1$. The no-purchase option is assumed to be in a nest of its own. Let w_j denote e^{μ_j} and let

$$w(\ell, \mathcal{M}) \stackrel{\text{def}}{=} \sum_{i \in \mathcal{N}_\ell \cap \mathcal{M}} w_i.$$

Then, with the above assumptions on the error terms, it can be shown that (see Ben-Akiva and Lerman (1985)) the probability that product j is purchased when offered assortment $\mathcal{M}$ is

$$\mathbb{P}(j|\mathcal{M}) = \mathbb{P}(\mathcal{N}_\ell|\mathcal{M})\,\mathbb{P}(j|\mathcal{N}_\ell, \mathcal{M}) = \frac{(w(\ell, \mathcal{M}))^\rho}{\sum_{m=1}^{L} (w(m, \mathcal{M}))^\rho} \; \frac{w_j}{w(\ell, \mathcal{M})}. \tag{33}$$

Nested MNL models alleviate the issue of IIA exhibited by the MNL models. Further, they have a closed form expression for the choice probabilities, which makes them computationally tractable. However, these models still exhibit IIA property within a nest. Moreover, it is often a challenging task to partition the products into different nests. Further, the model is limited because it requires each product to belong to exactly one nest.

The fact that NL model requires each product to be placed in exactly one nest is a limitation in several applications where a particular product is correlated with products across nests; for example, the no-purchase option in our setup is clearly correlated with all the products. In order to overcome this problem the NL model was extended to a *Cross Nested Logit (CNL)* model where each product can belong to multiple nests. The name cross-nested seems to be due to Vovsha (1997) and Vovsha's model is similar to the Ordered GEV model proposed by Small (1987). For our experiments, we assume that the no-purchase option has membership in all the nests and all other products have membership in only one nest. For this formulation, the probability of purchase of product j is given by (33) (see Ben-Akiva and Lerman (1985)), where $w(\mathcal{M}, \ell)$ is now defined as

$$w(\ell, \mathcal{M}) \stackrel{\text{def}}{=} \alpha_\ell w_0 + \sum_{i \in (\mathcal{N}_\ell \cap \mathcal{M}) \setminus \{0\}} w_i.$$

Here α_ℓ is the parameter capturing the level of membership of the no-purchase option in nest ℓ. The following conditions are imposed on the parameters α_ℓ, $\ell = 1, 2, \ldots, L$

$$\sum_{\ell=1}^{L} \alpha_\ell^\rho = 1, \quad \alpha_\ell \geq 0, \text{ for } \ell = 1, 2, \ldots, L.$$

The first condition is a normalization condition that is imposed because it is not possible to identify all the parameters. In our setup, it is only natural to assume that the no-purchase option has equal membership in all nests. This assumption translates into assuming that $\alpha_\ell = (1/L)^{1/\rho}$, for all ℓ. Further, we note that we assume without loss of generality that $V_0 = 1$ since only differences between utilities matter.

While the CNL model overcomes the limitations of the NL model, it is less tractable and Marzano and Papola (2008) showed that it cannot capture all possible types of correlations among products. Further, both

Table 2 Relevant attributes of DVDs from Amazon.com data and mean utilities of MNL model fit by Rusmevichientong et al. (2008)

Product ID	Mean utility	Price (dollars)	Average price per disc (dollars)	Total number of helpful votes
1	-4.513	115.49	5.7747	462
2	-4.600	92.03	7.6694	20
3	-4.790	91.67	13.0955	496
4	-4.514	79.35	13.2256	8424
5	-4.311	77.94	6.4949	6924
6	-4.839	70.12	14.0242	98
7	-4.887	64.97	16.2423	1116
8	-4.757	49.95	12.4880	763
9	-4.552	48.97	6.9962	652
10	-4.594	46.12	7.6863	227
11	-4.552	45.53	6.5037	122
12	-3.589	45.45	11.3637	32541
13	-4.738	45.41	11.3523	69
14	-4.697	44.92	11.2292	1113
15	-4.706	42.94	10.7349	320

the MNL and NL models don't account for heterogeneity in customer tastes. The MMNL family of models, described next, explicitly account for heterogeneity in customer tastes.

C.3. Mixed multinomial logit (MMNL)[10] family

The Mixed multinomial logit (MMNL) family of models is the most general of the three parametric families we compare our approach with. In fact, it is considered to be the most widely used and the most promising of the discrete choice models currently available Hensher and Greene (2003). It was introduced by Boyd and Mellman (1980) and Cardell and Dunbar (1980). McFadden and Train (2000) show that under mild regularity conditions, an MMNL model can approximate arbitrarily closely the choice probabilities of *any* discrete choice model that belongs to the class of RUM models. This is a strong result showing that MMNL family of models is very rich and models can be constructed to capture various aspects of consumer choice. In particular, it also overcomes the IIA limitation of the MNL and nested MNL (within a nest) families of models. However, MMNL models are far less computationally tractable than both the MNL and nested MNL models. In particular, there is in general no closed form expression for choice probabilities, and thereby the estimation of these models requires the more complex simulation methods. In addition – and more importantly – while the MMNL family can in principle capture highly complex consumer choice behavior, appropriate choice of features and distributions of parameters is highly application dependent and is often very challenging Hensher and Greene (2003).

In this model, the utility of customer c from product j is given by $U_{c,j} = \beta_c^T x_j + \varepsilon_{c,j}$, where x_j is the vector of *observed* attributes of product j; β_c is the vector of regression coefficients that are *stochastic* and not fixed[11] to account for the *unobserved* effects that depend on the *observed explanatory variables*; and $\varepsilon_{c,j}$ is the stochastic term to account for the rest of the unobserved effects. In this logit context, it is assumed that the variables $\varepsilon_{c,j}$ are i.i.d across customers and products Gumbel distribution of location parameter 0 and scale parameter 1. The distribution chosen for β_c depends on the application at hand and the variance of the components of β_c accounts for the heterogeneity in customer tastes. Assuming that β has a distribution $G(\beta;\theta)$ parameterized by θ, probability that a particular product j is purchased from assortment $\mathcal{M}$ is

$$\mathbb{P}(j|\mathcal{M}) = \int \frac{\exp\{\beta^T x_j\}}{\sum_{i\in\mathcal{M}} \exp\{\beta^T x_i\}} G(d\beta;\theta).$$

[10] This family of models is also referred to in the literature as Random parameter logit (RPL), Kernel/Hybrid logit.

[11] This is unlike in the MNL model where the coefficients are assumed to be fixed, but unknown.