Revenue Estimation under General Choice: Theoretical Guarantees and a Case Study

Vivek Farias • Srikanth Jagabathula • Devavrat Shah *

1. Introduction

The central operational problem faced by managers is the prediction of revenues from various offer sets using the available historical sales data. Such revenue predictions form crucial inputs to assortment planning: deciding the “optimal” offer set with respect to a given objective (e.g., revenue) subject to various constraints (e.g., limited display or shelf space). Clearly, the offer set significantly impacts the choices made by the customers and, hence, the revenues and the margins seen by the seller. Therefore, it is the key decision problem, for instance, faced by retailers and has several variants depending on the desired objective and the imposed constraints. In all these variants, the key building block towards making such decisions is the procedure that estimates the revenue for a given offer set based on the historical sales data.

Estimation of the revenues of a given offer set primarily requires two components: (1) prices of the products in the offer set and (2) the chance that a given product in the offer set is purchased by an incoming customer when the offer set is on display. Given prices of products, the revenue estimation boils down to predicting the share (fraction of sales) of each product in the offer set. This is challenging because of substitution behavior, where customers substitute an unavailable product with an available one. This makes the share observed for a product a combination of the primary demand for the product and the additional demand observed due to substitution. Clearly, capturing these substitution effects effectively is crucial to producing accurate revenue estimates.

Naturally, since revenue estimation is a question of great importance in operations, there is a vast and exciting literature to address the challenge of predicting the shares for different products. Specifically, the so-called customer choice models have proved to be effective in capturing the substitution behavior of customers to produce accurate revenue estimates. Simply put, a choice model is a conditional distribution that gives the fraction of sales for each product in every offer set i.e., the chance that a given product in the offer set is purchased by a typical customer when the offer set is on display. The traditional approach widely popular in the Operations Management (OM) literature to using choice models to generate revenue predictions is as follows: (1) Posit a parametric, structural choice model; (2) Fit the model to the historical sales data, and (3) Predict revenues using the learned model.

Parametric models have gained widespread use because their structure makes them tractable both in terms of estimating model parameters from limited data and solving the relevant decision problems. However, they suffer from the usual limitations of parametric approaches. In particular, in many practical situations choosing the appropriate tractable structure is far from straightforward. This choice typically involves several subjective judgments, usually from domain experts. The requirement of having expensive expert input for such parametric choice model makes them infeasible for many modern applications that either involve a large number (thousands) of product categories (e.g., in the context of e-businesses like Amazon.com) or highly dynamic environment with products changing continually (e.g., high fashion industry).

In order to overcome the difficulties of parametric approaches, recently there have been efforts to consider nonparametric approaches to choice modeling. Particularly, in a recent work [Farias et al., 2009], the authors consider a very generic model of choice, which is a distribution over preference lists of products. Under such a generic model and given sales data for a few assortments, the authors derive conditions on model recoverability and design a revenue estimation subroutine. In addition, the authors conduct extensive simulation studies in which they demonstrate that the revenue estimation subroutine produces accurate revenue estimates when sales data comes from wide range of parametric structures. In essence, the study shows that the subroutine can generate accurate revenue estimates while being completely agnostic to the underlying choice structure.

Our main contribution in this work is to extend the work in [Farias et al., 2009] in two important ways: (1) we derive explicit (theoretical) performance guarantees for the estimates produced by the revenue estimation subroutine when the underlying choice model is either a multinomial logit (MNL) model or a mixed-multinomial logit (MMNL) model; (2) we demonstrate the robustness of the revenue estimation subroutine in a real-world application through a case study on the revenue estimation problem faced by a

*VF is affiliated with MIT Sloan and SJ and DS with EECS, MIT. email: {vivekf, jskanth, devavrat}@mit.edu
major US automaker. This combination of theoretical guarantees combined with a real-world case study offer the following sets of important operations insights:

1. There is a clear trade-off between the complexity of the underlying choice model structure and the amount of information in the available data to obtain accurate revenue estimations: for instance, they confirm the intuitive relation that the “amount” of data required to obtain the same level accuracy increases with the complexity of the underlying choice structure.

2. The theoretical performance guarantee we derive are readily computable from the data available. Thus, we can readily determine how good (or bad) this revenue estimation is using the observed data itself!

3. Our guarantees provide guidance on the “amount” of data that needs to be collected for the desired level of accuracy. This is extremely valuable from managerial perspective.

Our case study with the sales data from a major US automaker establishes the superiority of this revenue estimation procedure. Specifically, it handily beats the performance of a benchmark MNL model fit to the same data. We note that the choice of the MNL model as the benchmark parametric approach was motivated by the following considerations: (a) it is the most popular approach in practice, and (b) while in principle a more complicated model can be fit to the data, it is far from desirable because of the several subjective decisions it entails. Next we give a brief description of the choice model and the revenue estimation subroutine introduced in Farias et al. [2009] followed by a summary of our results.

2. Model

We provide a brief description of the model and the revenue estimation subroutine described in Farias et al. [2009].

Model. Suppose \( N \) is the universe of \( N \) products. For an offer set \( M \) subset of \( N \), for any product \( i \in M \), a choice model specifies the chance or probability that \( i \) is purchased when \( M \) is offered. We model customer choice decisions as follows: each customer has a preference list of all the products. When offered an assortment of products, the customer purchases her most preferred of the offered products. The customer choice model is determined by a distribution over preference lists, seen as follows. Assume that there are \( K \) customer types in the population with each customer type associated with a preference list. The choice model is completely specified by the distribution that assigns to each preference list a weight equal the fraction of customers that belong to the corresponding customer type. Now, the probability that a product \( i \in M \) is purchased when offered \( M \) is equal to the sum of the weights of all the preference lists that result in the purchase of \( i \) when offered \( M \). Essentially any reasonable ‘parametric’ model for customer decisions can be cast in this form.

Data. In order to estimate this choice model, we assume we have access to sales data for a few offer sets. This data can be interpreted as ‘marginal’ information about the distribution over preference lists. For example, given sales data for every pair of products \( i, j \), we can cast it as the probability that \( i \) (resp. \( j \)) is purchased given the offer set \( \{i, j\} \). More generally, the sales data from larger offer sets can be cast cast as marginal information about our distribution over preference lists.

Revenue estimation. Assuming that the price of each product is fixed (but otherwise completely arbitrary), given a distribution over preference lists, one can immediately compute the expected revenue. However, we have access to only marginal information and several distributions will be consistent with the given marginals. Each consistent distribution yields a revenue estimate and choose the ‘worst-case’ revenue estimate. In other words, of all the revenue estimates corresponding to distributions consistent with the given data, use the minimum possible revenue as the estimate and, clearly, in reality the revenue is no less (or worse) than this.

---

1The customer however need only be aware of her preferences over the offered products.
2We implicitly assume here that a ‘no-purchase’ option is always available.
3Note that these probabilities don’t add up to 1 because customers have the ‘no-purchase’ option.
The guarantee $f_k(M)$ under an MNL model for offer set $M$ with $|M| = C$ is, roughly speaking, equal to the shaded area under the curve in the weight profile of the offer set. If the weights fall as $\sim e^{-c_i}$ for some constant $c$ and product $i$, $1 \leq i \leq C$, then the error bound $\sim e^{-k}$. This is indeed the kind of scaling we observed for an MNL model fit to the DVD sales data from Amazon.com.

3. Main results

We now present our main results. **Theoretical performance guarantees.** We derive guarantees for the revenue estimates when the underlying choice model is an MNL model or an MMNL model. We derive the guarantees for the case when we have access to what we call $k^{th}$ order marginal information: purchase probabilities of all products in every offer set of size at most $k$. We derive bounds for the relative error i.e., the absolute difference between the estimated revenue and the true revenue as a percentage of the true revenue. Specifically, for any offer set $M$, we show that the relative error is bounded above by $f_k(M)$, where $f_k(M)$ depends on the “amount” of information $k$ and the model parameters of the products in the offer set. $f_k$ decreases as $k$ increases and for the type of applications common in practice, this decrease can be shown to be rapid (see Figure 1). In addition, for an MMNL model we can show that for a given $k$, $f_k(M)$ increases with the increases in the heterogeneity.

**Case study.** We tested the performance of the subroutine using car sales data from a major US automaker. In particular, we used sales data of the Fusion and Escape models from the dealership zone around Dearborn, MI; the zone contains 31 dealerships. For each dealership, we have access to transaction data over a sales horizon of nine months (from August 2009 to April 2010). For each transaction, we know the car that was purchased, the price that was paid, and the set of other cars in the lot at the time of purchase. We aggregated the data by grouping the cars according to their “Rapid Spec”\(^4\). This resulted in 21 different products (10 Fusion and 11 Escape) and a total of 15 different offer sets of the 21 products that were in the lots at different points of time. Hence, we have access to purchase data of each of the products in 15 different assortments. In order to test the performance of the subroutine, we carried out cross-validation. Specifically, we considered every possible partition of the 15 offer sets into training set of 10 and test set of 5 offer sets; clearly, this yields a total of $\binom{15}{5} = 3003$ partitions. For each partition, we used the training data as the input and estimated the revenues of the 5 test offer sets using the subroutine. For each of the estimates we then computed the relative errors. This resulted in a total of 1200 relative errors. As a benchmark, we repeated the above steps with an MNL model. The histograms of the relative errors are shown in Figure 2. It is evident from the comparison of the histograms that the subroutine outperforms the MNL model.

References


\(^4\)Each “Rapid Spec” corresponds to a set of features of the car like the types of tires, rooftops, electronics system, etc.