Demand Modeling in the Presence of Unobserved Lost Sales

Shivaram Subramanian
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, subshiva@us.ibm.com

Pavithra Harsha
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, pharsha@us.ibm.com

We present an integrated optimization approach to parameter estimation for discrete choice demand models where data for one or more choice alternatives are censored. We employ a mixed-integer program (MIP) to jointly determine the prediction parameters associated with the customer arrival rate and their substitutive choices. This integrated approach enables us to recover proven, (near-) global optimal parameter values with respect to the chosen loss-minimization objective function, thereby overcoming a limitation of prior approaches that employ multi-start heuristic procedures and terminate without providing precise information on the solution quality. The imputations are done endogenously in the MIP by estimating optimal values for the probabilities of the unobserved choices being selected. Under mild assumptions, we prove that the approach is asymptotically consistent. Partial information, user acceptance criteria, model selection, and regularization techniques can be incorporated to enhance practical efficacy. We test the method on simulated and real data, and present results for a variety of single- and multi-item demand prediction scenarios, and for learning the unobserved market shares of competitors.

Key words: Discrete choice model, Lost sales imputation, Utility-preference : Estimation, Statistics : Censoring, Programming : Integer : Applications

History:

1. Introduction

Demand forecasting is an important part of business planning across a wide variety of industries. An estimate of future demand forms a fundamental input for a variety of downstream decision making operations, such as manufacturing, inventory planning, scheduling and pricing. In today's price-transparent and omni-channel world, customers have a wide variety of choices in front of them prior to finalizing their purchase. In this environment, it is important for sellers to not only forecast demand for ones own products or services, but to obtain an integral '360-degree' view of future demand potential, in order to support optimal decision making in the presence of competition. A system having such a capability can accurately predict how the future market-place evolves over time, which includes the proportion of customers that are likely to walk away from the seller's product assortment, as well as the market-share that is at risk of being lost to one or more competitors. Typically, such systems require a variety of data, including the purchase of

expensive syndicated data regarding competitor sales, as well as installation of in-store video and internet based analytical systems, etc., to record lost sales opportunities. However, the hardware and software infrastructure to acquire such uncensored data is not yet available to a vast majority of sellers and service providers, and it may simply not be possible in other settings. The demand forecasting methods presented in this work are designed to provide a seller a panoramic view of future demand using the available aggregate-censored data.

Discrete choice models are one of the most widely used demand functions to model consumer choice in marketing, economics, supply chain and revenue management (Ben-Akiva and Lerman 1985, Train 2009), with successful implementations in a variety of industries (e.g., Guadagni and Little 1983 Ratliff et al. 2008, Vulcano et al. 2010, Subramanian and Sherali 2010). Discrete choice demand models have a convenient hierarchical structure, which first models the market size of interested customers, and then the market share of specific customer choices amongst alternatives. The demand for a customer choice is calculated as the product of the market-size and its corresponding market share. Hence, they can be can be used to predict demand and demand potential in a variety of customer choice scenarios that arise in practice. This includes, for example, the (1) binary choice setting (buy vs. no-buy); (2) multiple substitutive choice setting where there are at least three choices for a customer (no-buy, vs. buying one of multiple competing substitutes in the assortment, e.g., store brand vs. national-brand soda) where understanding the relative attractiveness of the competing substitutes is important; and, (3) non-substitutive positively correlated settings where demand of one item is positively correlated with the other items in the assortment (e.g., different sizes of a T-shirt). Their structural form makes them a viable demand modeling alternative for downstream pricing and assortment optimization applications where some computationally tractable and efficient algorithms have be developed (e.g., Talluri and Van Ryzin 2004, Rusmevichientong et al. 2010, Keller et al. 2014).

In many applications it is quite common for retailers or service providers to have no information regarding customers who were interested in a product but did not purchase, let alone the fraction of those customers who simply did not buy the product at all, or purchased the same item (or a substitute) from a competitor. Out-of-stock situations also result in similar data censoring. Calibrating the discrete choice models to accurately predict future demand can be quite challenging when complete historical data regarding all the customer choices is unavailable. Sometimes, partial information in the form of sample data, or aggregated statistics may be available, which have to be incorporated to improve predictive accuracy. This particularly challenging prediction problem has received attention in recent years and our paper proposes novel methods using constrained optimization to solve it effectively in practice.

Contributions: We present an integrated optimization approach to parameter estimation and missing data imputation for calibrating discrete choice demand models where historical data regarding one or more choice alternatives are censored. In particular, we consider two distinct cases: (1) single unobserved no-purchase option; and (2) multiple unobserved substitutive options, which additionally embeds the loss to one or more competitors having observable attributes. In both cases, we jointly determine the prediction parameters associated with the customer arrival rate and their substitutive choices. Our method is based on an effective mixed-integer program (MIP) that minimizes a measure of the total error between predicted and observed (or imputed) quantities, both with respect to the arrival rates as well as the market share ratios. The imputations are done endogenously in the MIP by estimating optimal values for the probabilities of the unobserved choices being selected. The integrated single step approach enables us to recover proven, (near-) global optimal prediction parameter values with respect to the loss-minimization objective under consideration. Furthermore, partial information, several user acceptance criteria, model selection and regularization techniques can be incorporated to enhance the efficacy of the proposed method in practice.

We theoretically prove and empirically demonstrate using examples that the integrated optimization approach (in both aforementioned cases) leads to asymptotically consistent estimators of all the parameters under mild assumptions. We present several numerical experiments to highlight the flexibility of the model and also, present demand calibration results in two real world applications. First, we present win probability estimation results on a real-world request-for-quote (RFQ) data set. Next, we use publicly available hotel data set parameters to simulate a realistic market environment and present results in the context of multi-item assortment demand estimation, and the learning of unobserved market shares of competitors. The proposed MIP method has been observed to converge rapidly and is simple to implement and maintain, while being highly flexible.

Besides the results presented in the paper, we have commercially deployed a proprietary version of the method to predict omni-channel weekly demand for retailers for non-perishable and perishable items in pricing applications. We have also implemented the forecasting methods for predicting dynamic time-of-day electricity demand in a smart grid application, and monthly demand for chemical and food manufacturers.

Organization: The rest of the paper is organized as follows. We discuss background and related prior art in the Section 1.1. In Section 2, we describe the discrete choice demand model. We present the proposed integrated optimization approach for the single unobserved no purchase option in Section 3 and the multiple unobserved substitutive options in Section 4. In each of these sections, we also present the respective results on consistency and experiments on simulated data sets. In Section 5 and Section 6 we present the calibration results for two real-world applications, RFQ and hotel settings respectively.

1.1. Background and related literature review

Complete information about all customer choices: The most widely used method to estimate discrete choice models under complete information is the maximum likelihood estimation (MLE) method (Domencich and McFadden 1975, Ben-Akiva and Lerman 1985) and in its simplest form, the likelihood of a choice(s) given the assortment of choices is maximized. A commonly used alternative is the Berkson's method (Berkson 1953, Ben-Akiva and Lerman 1985). This is a linear regression method where the error in a specific non-linear ratio transformation of the choice function resulting in a linear expression in the covariates is minimized. The observational data for the MLE method can be as granular as the actual response of an individual and the associated covariate vector, whereas the Berkson's method requires grouped data in terms of either counts or proportions which consists of the response of a group of individuals all of whom have the same covariates. For most parametric choice models, both problems are relatively easy to solve numerically using different optimization algorithms, yet have the same theoretical asymptotic properties (Greene 2011). As we will see next, under incomplete information, most methods have focused on the extensions of the MLE method. In contrast, our work extends the Berkson's method.

Incomplete information about customer choices: Talluri and Van Ryzin (2004), in their work on choice based revenue management, develop an MLE based estimation method using an expectation maximization (EM) approach (Dempster et al. 1977) to jointly estimate a constant mean arrival rate and the parameters of the choice model based on sales transaction data and unobserved no-purchases. The method iterates between estimating an expected lost sales in each period, and maximizing the resulting expected value of the log-likelihood function. Although the number of iterations required to obtain accurate parameter estimates can result in prohibitively large run times, EM approaches are attractive because of their relative ease of implementation and wide applicability. In general, an EM method provably converges to stationary point only (Wu 1983). This means, the solution can be a local maximum, saddle point, or the local minimum of the likelihood function. Therefore, EM methods are rerun with multiple starting points, further increasing run time.

We refer to Vulcano et al. (2010) and Newman et al. (2014) for a variety of simulations that study the empirical performance of the EM algorithm, and the former paper for an implementation of the EM algorithm on real airline booking data. Kök and Fisher (2007) and Vulcano et al. (2012) review various demand estimation approaches for assortment planning using the EM method.

Vulcano et al. (2012) extend and improve the performance of the EM algorithm for the incomplete data likelihood function with aggregate data assuming that the arrivals follow a non-homogeneous (time-dependent) Poisson distribution. For the same setting, the Minorization-Maximization (MM)

approach of Abdallah and Vulcano (2016) provably converges to a stationary point and further enhances the run-time performance of MLE based methods. For the pure assortment case in the absence of covariates, the authors state sufficient conditions for identifiability (and hence consistency) of the choice parameters. The mean arrival rate parameters are also consistent if they are purely assortment dependent, and time-independent.

Newman et al. (2014) propose a computationally fast two step estimation method by decomposing and optimizing the incomplete data likelihood function assuming that the arrivals are modeled using a homogeneous Poisson distribution. A nonconvex optimization problem is required to be solved in the second step, and a multi-start procedure is employed to overcome local optima, with a local maximum guarantee at best. Talluri (2009) also proposes a two-step method wherein a risk-ratio error minimization is instead adopted. As a remark, the two-step methods cannot be directly implemented in the binary-choice (buy vs. no-buy) setting, as the first step requires at least two observed purchasing options.

The prior MLE based methods, in the presence of covariates, employ multi-start heuristic procedures to overcome local maxima when solving the resultant non-convex MLE problem. In comparison, the method we propose in the paper, adopts a single step MIP approach that solves a non-convex loss minimization problem to provable optimality even in finite data settings. Another drawback of MLE based methods in this setting is that arrivals are modeled as a Poisson random variable, independent of covariates. For example, traffic to a store increases during holidays and campaign periods. Here, potentially both sales and lost sales increases compared to non-holiday periods. This situation cannot be adequately captured by the aforementioned arrival rate model and requires the incorporation of covariates in order to capture such key trends and obtain a better fit with the data. Lastly, our paper also estimates the unobserved market-shares of competitors when the unobserved lost sales is attributable to no-purchase, or a purchase from one of the competitors. We propose a method to disambiguate the different lost sales using the covariates of the competitor (e.g., their price), which are assumed to be known. There is no prior work that we are aware of that considers disambiguation in discrete choice model estimation using censored data. Based on the above discussion and our contributions, we summarize the key differences of our method from the above methods in Table 1.

Although our paper focuses only on estimation of discrete choice demand models, papers by Haensel and Koole (2011), van Ryzin and Vulcano (2011), Farias et al. (2013) are a few examples among several works that have studied demand prediction problems with censored data for other classes of demand models, which has also been an active area of research.

Property	Proposed LM method	Referenced MLE based methods
Number of observed choices	1 or more	EM, MM: 1 or more; Two step: at least 2
Number of unobserved choices	1 or more	At most 1
Data aggregation	Required	Classical EM: not required; Other EM, MM and two step: required
Probabilistic assumptions	None	Required
Covariates for arrivals	Modeled	Not modeled
Optimization solution approach	Single step MIP; (Near-) global optimal solution with a certificate of optimality	At least two steps or iterative; With share covariates: Multi-start heuristic procedures; Local maxima guarantee; No share covariates: EM and MM converge to global optimal.
Asymptotic consistency	Proved	No share covariates: EM and MM methods are consistent

Table 1 Comparing the proposed loss minimization (LM) method with respect to the referenced MLE based methods to calibrate discrete choice models in an incomplete data setting.

2. Demand functions based on discrete choice models

Discrete choice demand models anchored in utility theory are one of the commonly used demand functions to model consumer choice. They generalize the well-known *multinomial logit* (MNL) and the *multiplicative competitive interaction* (MCI) demand models (McFadden 1974, Urban 1969). In a setting with an assortment of purchasing choices and a no-purchase choice, we use the discrete choice demand functions to model the demand of a choice by estimating its market share. Suppose M is the set of purchasing choices indexed by m and $\emptyset$ denotes the no-purchase choice. The demand for a choice $m \in M$ or the choice $\emptyset$ at time t is given by:

$$D_{mt}(\mathbf{Y}_t, \mathbf{X}_t, S_t) = \text{Market Size at time } t \ * \ \begin{matrix}\text{Market Share} \\ \text{of item } m \text{ at time } t\end{matrix} \tag{2.1}$$

$$= \quad \tau(\mathbf{Y}_t) \ \frac{f_m(\mathbf{X}_{mt})}{1 + \sum_{m' \in S_t} f_{m'}(\mathbf{X}_{m't})}, \textit{and} \tag{2.2}$$

$$D_{\emptyset t}(\mathbf{Y}_t, \mathbf{X}_t, S_t) = \quad \tau(\mathbf{Y}_t) \ \frac{1)}{1 + \sum_{m' \in S_t} f_{m'}(\mathbf{X}_{m't})}, \tag{2.3}$$

where

- $\tau(\mathbf{Y}_t)$ is the market size model that outputs the measure of consumers interested in any of choices, including the no-purchase option, as a function of a vector of the market size attributes $\mathbf{Y}_t$ at time t,
- $f_m(\mathbf{X}_{mt})$ is the attraction model of choice m as a function of the vector of attributes $\mathbf{X}_{mt}$ for choice m at time t,
- $\mathbf{X}_t$ is the matrix of attributes where row m corresponds to $\mathbf{X}_{mt}$, and
- $S_t \subset M$ is the set of purchasing choices that are available at time t.

Note that we interchangeably use the term covariates and attributes through the paper, to refer to regressors or feature vectors in an estimation problem.

In continuous time settings, market size is measured and referred to also as an arrival rate. The size function aims to capture the impact of higher (assortment) level attributes that influence the arriving traffic, which include temporal effects (seasonality, assortment popularity or freshness, trend and holiday) or the marketing effects, and is independent of factors that impact the purchase of a specific choice. The market size function can be modeled as a linear, exponential or a power function of the attributes $\mathbf{Y}_t$ (for example, for an exponential function, $\tau(\mathbf{Y}_t) = e^{\gamma^T \mathbf{Y}_t}$) whose prediction coefficients ($\boldsymbol{\gamma}$, in the example) have to be estimated.

Market share, also commonly referred to as the purchase/choice probability, models how consumers choose between alternatives. Market share of any single alternative is its relative attractiveness and thus dependent on the attraction models of all choices (including no purchase which has a default attraction value taken as 1.0). Attraction models are based on a utility concept, where the utility is composed of an observable component and an unobservable random component. The observable part for any purchasing choice m is typically a function of the attributes $\mathbf{X}_{mt}$ at time t that depend on the specific choice m. Different attraction models can be derived depending on the assumptions of the unobservable random component. For example, when the noise for all choices are independent and identically distributed Gumbel (type 1 extreme value) and the observable components are linear, i.e., $\boldsymbol{\beta}_m^T \mathbf{X}_{mt}$, the MNL demand model is derived. The attraction model in the case of the MNL model is $f_m(X_{mt}) = e^{\boldsymbol{\beta}_m^T \mathbf{X}_{mt}}$, in the case of the MCI model is $f_m(\mathbf{X}_{mt}) = \prod_k X_{mtk}^{\beta_{mk}}$, and in the case of a linear attraction model is $f_m(\mathbf{X}_{mt}) = \boldsymbol{\beta}_m^T \mathbf{X}_{mt}$ where $\boldsymbol{\beta}_m$ are the prediction coefficients that need to be estimated.

A discrete choice function is practically convenient because of its parsimony in the number of coefficients to be estimated. In particular, the number of coefficients in the discrete choice demand model is $O(M)$ for M purchasing choices as opposed to $O(M^2)$ in demand models such as linear, exponential or power-law (Reibstein and Gatignon 1984, Berry 1994). Discrete choice models also enable efficient information sharing across choices in sparse data settings (because of their top down approach of estimation) compared to alternatives that employ separate models for each purchase choice (a bottom-up approach).

3. Estimation of censored data discrete choice models with a single unobserved no-purchase option

Consider a seller managing an assortment of $m \in M$ substitutable purchasing options. The no-purchase option $\emptyset$ is always available to the consumers. The seller varies the availability of the purchasing options, denoted by S_t, and its attributes (e.g., prices) over time. The seller only observes

the total sales transactions for each of the available purchasing options in M over time but cannot observe lost sales, which is the number of arriving customers choosing the no-purchase option $\emptyset$. Our goal is to use historical sales observations, the corresponding attributes, and availability data to predict the demand for each of the $m \in M$ purchasing options, as well as the unobserved lost sales. We present the single step MIP based estimation method in Section 3.1, prove that it is a consistent estimator in Section 3.2, discuss model enhancements and extensions in Section 3.3 and finally present computational experiments with simulated data in Section 3.4.

3.1. Proposed estimation method

The following table describes the notation used.

Indices	
t, m, k,	a period, a choice, a lost share value
Sets	
$\mathcal{T}, M$,	all periods (time windows), all purchasing choices
S_t,	all available purchasing choices in period t, $S_t \subset M$
K,	all lost market share values chosen between $[\epsilon, 1-\epsilon]$, where ϵ is a small number greater than 0
Parameters	
$\bar{s}_{mt}$,	observed sales for choice m in period t
$\bar{\mathbf{X}}_{mt}$,	row vector of attribute values for choice m in period t
$\bar{\mathbf{Y}}_t$	row vector of attribute values for market size in period t
f_k,	lost market share value corresponding to index k
$\bar{\lambda}_{kt}$,	lost sales due to no-purchase in period t, and equals $\frac{f_k}{1-f_k}\sum_{m\in S_t}\bar{s}_{mt}$
$\bar{\theta}_{kt}$,	market size in period t, and equals $\frac{1}{1-f_k}\sum_{m\in S_t}\bar{s}_{mt}$
Decision Variables	
$\boldsymbol{\beta}_m$,	row vector of coefficients for the attributes of choice m ($\bar{\mathbf{X}}_{mt}$), modeled as continuous variables
$\boldsymbol{\gamma}$,	row vector of coefficients for attributes of market size ($\bar{\mathbf{Y}}_t$), modeled as continuous variables
z_{kt},	probability that the lost market share in period t is f_k. These auxiliary decision variables are modeled as a Special-Ordered-Set Type 2 (SOS2) variables wherein at most two adjacent members (assuming the f_k's are ordered) can be non-zero.

We assume that when $m \notin S_t$, $\bar{s}_{mt} = 0$ and that $\bar{s}_{mt} > 0$ otherwise. The latter assumption is reasonable in most instances because we are working with aggregated data over a period t. In fact, we can simply discard observations for choices with zero observed sales and later in the section, we discuss the reason it does not impact the identifiability of the model. We also assume that the attribute vectors $\bar{\mathbf{Y}}_t$ and $\bar{\mathbf{X}}_{mt}$ incorporate the constant 1 to retrieve the intercept coefficients.

Formulation: Without loss of generality, for exposition, we present the optimization formulation for an exponential market size function and a MNL share model. Generalizations to other discrete choice models are discussed later in this section. We aim to perform the following fit:

$$\bar{s}_{mt} \sim e^{\boldsymbol{\gamma}^T \bar{\mathbf{Y}}_t} \frac{e^{\boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt}}}{1 + \sum_{m' \in S_t} e^{\boldsymbol{\beta}_{m'}^T \bar{\mathbf{X}}_{m't}}} \qquad \forall m \in S_t, t \in \mathcal{T}, \tag{3.1}$$

where the notation T refers to the transpose of the vector. Suppose we know the lost sales, denoted by λ_t, for each period, we obtain an additional set of equations to fit:

$$\lambda_t \sim e^{\gamma^T \bar{\mathbf{Y}}_t} \frac{1}{1+\sum_{m' \in S_t} e^{\boldsymbol{\beta}_{m'}^T \bar{\mathbf{X}}_{m't}}} \qquad \forall t \in \mathcal{T}. \tag{3.2}$$

As Eqs. (3.1–3.2) are non-linear equations, we can rewrite them as follows:

$$\ln(\bar{s}_{mt}) - \ln(\lambda_t) \sim \boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt} \qquad \forall m \in S_t, t \in \mathcal{T}, \text{ and} \tag{3.3}$$

$$\ln\left(\sum_{m \in S_t} \bar{s}_{mt} + \lambda_t\right) \sim \boldsymbol{\gamma}^T \bar{\mathbf{Y}}_t \qquad \forall t \in \mathcal{T}. \tag{3.4}$$

The first equation employs the log-ratio transformations similar to the Berkson's estimator and the second is just the sum of the equations for each t across all choice $m \in M$. We now employ the piecewise linear (PWL) transformation using the auxiliary decision variable, z_{kt}, to replace the unknown lost sales, λ_t, in particular, its functions:

$$\ln(\bar{s}_{mt}) - \sum_{k \in K} \ln\left(\bar{\lambda}_{kt}\right) z_{kt} \sim \boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt} \qquad \forall m \in S_t, t \in \mathcal{T}, \text{ and} \tag{3.5}$$

$$\sum_{k \in K} \ln\left(\bar{\theta}_{kt}\right) z_{kt} \sim \boldsymbol{\gamma}^T \bar{\mathbf{Y}}_t \qquad \forall t \in \mathcal{T}. \tag{3.6}$$

Our goal is therefore to identify the optimal values for the prediction parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and the lost share variables z_{kt} for each period that minimizes the total error between the predicted and observed (or imputed) quantities. To achieve this goal, we formulate an integrated optimization approach to estimate all variables jointly, as shown below. We employ a general loss function denoted by $\mathcal{L}[.]$ where $\mathcal{L}[.] : \mathbb{R} \to \mathbb{R}^+$ and refer to it as the loss minimization (LM) model:

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma},z} \sum_{t \in \mathcal{T}} \sum_{m \in S_t} \mathcal{L}\left[\ln(\bar{s}_{mt}) - \sum_{k \in K} \ln(\bar{\lambda}_{kt}) z_{kt} - \boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt}\right] + \sum_{t \in \mathcal{T}} \mathcal{L}\left[\sum_{k \in K} \ln(\bar{\theta}_{kt}) z_{kt} - \boldsymbol{\gamma}^T \bar{\mathbf{Y}}_t\right] \tag{LM}$$

$$\sum_{k \in K} z_{kt} = 1 \qquad \forall\ t \in \mathcal{T} \tag{3.7}$$

$$z_{kt} \geq 0 \qquad \forall\ t \in \mathcal{T}, k \in K \tag{3.8}$$

$$\{z_{kt}\ \forall\ k \in K\} \in \text{ SOS2} \qquad \forall\ t \in \mathcal{T}. \tag{3.9}$$

The objective function in the LM model minimizes the error between the left-hand and the right-hand side of Eqs. (3.5–3.6) using a loss function $\mathcal{L}[.]$.[1] In this paper, we work with non-negative continuous convex loss functions where $\mathcal{L}[0] = 0$ and $\mathcal{L}[x] \to \infty$ if $x \to \pm\infty$. In particular, we focus on the L1 or the L2 norms. The constraints (3.7–3.9) impose the SOS2 nature of the lost share

[1] The chosen objective function directly optimizes the relative accuracy metric (log ratio of predicted to actuals) which is a commonly used criterion for model selection to avoid bias in under-prediction (Tofallis 2015).

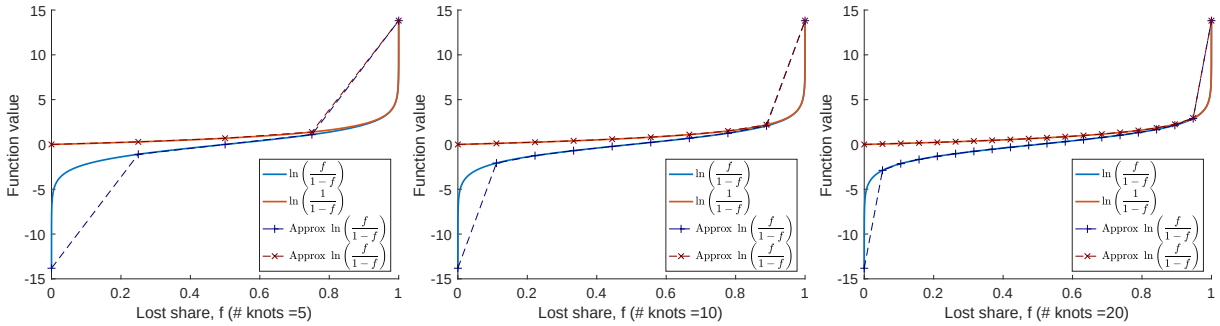

Figure 1 **True and approximate value of the projected lost sales and market size per unit of observed sales as a function of the lost share, f, for varying number of knots uniformly chosen in** $[10^{-6}, 1-10^{-6}]$.

variables. We do not expand on the form of the SOS2 constraints as all standard commercial optimization packages allows this level of specification and manage these variables internally. If we adopt the L1 loss function, the objective function can be linearized (see chapter 1 in Bertsimas and Tsitsiklis 1997). The resultant formulation is a MIP because of the binary variables required to model the SOS2 conditions in the presence of non-convexity. Instead if we adopt an L2 loss function, the problem reduces to a mixed-integer convex quadratic program. These resultant mathematical formulations can be efficiently solved in practice by standard optimization software packages like IBM-ILOG CPLEX to produce a (near-) global optimal solution within a user-specified optimality tolerance (e.g., with 1% of optimality).

A key insight in the proposed method is that the imputations are done endogenously in the optimization problem because when the lost sales fractions are specified, all the unknowns (market size and all shares) are fully specified and can be used for estimation of the parameters. As these lost sales fractions are modeled as SOS2 variables, the search is practically on the entire space of the PWL approximations of the projected lost sales, $\ln\left(\frac{f}{1-f}\right)$, and projected size, $\ln\left(\frac{1}{1-f}\right)$, for every unit of observed sales.

Fig. 1 plots the true (solid line) and approximate value (dotted lines) of these functions for varying number of knots of the PWL segments uniformly chosen between $[10^{-6}, 1-10^{-6}]$. As the number of knots increases, we obtain better approximations. As can be seen, the slope of these functions rapidly increases when the lost sales shares get closer to 0 or 1 and is relatively invariant in the mid range. In Section 3.4 we computationally show that the number of knots directly influences the tradeoff between run time and solution quality. Given this, a judicious selection of knot locations can be beneficial for the application under consideration.

3.2. Consistency of the proposed estimator

Under complete information with known lost shares, Berkson's estimator optimizes the L2 norm of the first term only of problem LM's objective function to estimate the $\boldsymbol{\beta}_m$ $\forall m \in M$. With

sufficient number of time buckets having linearly independent covariates, the estimator is known to be consistent (Cox 1970) as the number of data points in of each time bucket tends to infinity.

For consistency, we work with an arrival rate instead of a market size, because we analyze the optimization solution quality as the number of arrivals in a time bucket tends to infinity. We use the notation N to denote the number of unit time intervals present within any single aggregate time bucket t and for simplicity, assume it is the same for all t. Thus, in the LM model, θ_{kt} is modified to $\frac{\theta_{kt}}{N}$ to compute the feature vectors of an arrival rate as opposed to market size. For finite data, we conveniently set $N=1$ to identify market size. Arrival rate estimation is not considered as part of the Berkson's estimator but it is easy to conclude from basic statistical theory that it is also consistent as $N \to \infty$. The reason is we are averaging an increasingly large number of arrivals (per unit time) in a time bucket and it is a consistent estimator of an average true arrival rate, $e^{\boldsymbol{\gamma}^T \mathbf{Y}_t}$.

In the censored data setting, a joint optimization of both the first and the second term is required. For example, in the single purchase choice case, for any feasible value of the share prediction parameters $\boldsymbol{\beta}_m$, a feasible value for the lost share variable z_{kt} for each $t \in \mathcal{T}$ can be trivially chosen to reduce the error in the first term close to zero. Therefore, the second term is required to break the ties among all feasible alternatives that minimize the error in the first term. Similarly, for multiple choices, there remains one degree of freedom that requires the second term for model identification. This joint estimation is the main intuition underlying the consistency of the proposed estimator, which we prove in the theorem below. We begin by considering the continuous version of LM model:

$$Z^C(N, \boldsymbol{\beta}^N, \boldsymbol{\gamma}^N, \mathbf{f}^N) = \min_{\boldsymbol{\beta}_m, \boldsymbol{\gamma}, f_t \in (0,1)} Z^C(N, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{f}) \qquad \textbf{(LM-C)}$$

where

$$Z^C(N, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{f}) = \sum_{t \in \mathcal{T}} \sum_{m \in S_t} \mathcal{L}\left[\ln\left(\frac{\bar{s}_{mt}(1-f_t)}{f_t \sum_{m' \in S_t} \bar{s}_{m't}}\right) - \boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt}\right] + \sum_{t \in \mathcal{T}} \mathcal{L}\left[\ln\left(\frac{\sum_{m \in S_t} \bar{s}_{mt}}{(1-f_t)N}\right) - \boldsymbol{\gamma}^T \bar{\mathbf{Y}}_t\right].$$

Here, the notation N is overloaded to refer to a sample of the observed sales data under consideration. Before proving the main result on consistency, we make two assumptions.

ASSUMPTION 1. *For a given data set, suppose we have to identify the k_m unknown parameters for each choice $m \in M$ and l arrival rate parameters then we assume that the data set with covariate or assortment variations time buckets, satisfies the following:*

1. *There exists sets $K_m \subset \mathcal{T}$ of time buckets for each choice $m \in M$ and such that $|K_m| = k_m$ with linearly independent $\bar{\mathbf{X}}_{mt}$ for all $t \in K_m$.*
2. *The set $L = \cup_{m \in M} \mathcal{T} \setminus K_m$ of time buckets is such that $|L| \geq l$ with at least l linearly independent $\bar{\mathbf{Y}}_t$ across $t \in L$.*

This is a mild and natural necessary assumption on the minimum number of time buckets having linearly independent covariates for identifiability of the model given the number of unknowns. In fact, this is the reason we can discard share terms in the LM objective function for choices with zero observed sales for a time period. Note that because of the joint estimation of size and share terms, the size term has to be included whenever a share term for any choice is present.

ASSUMPTION 2. *For at least one choice of sets K_m $m \in M$ (described above), the following system of equations has a unique solution for $\boldsymbol{\gamma}$:*

$$w_t = \sum_{t' \in K_m} \eta_{t'}^m w_{t'} \qquad \forall\ m \in S_t, t \in \mathcal{T} \backslash K_m. \tag{3.10}$$

Here $\boldsymbol{\eta}^m$ satisfies $\bar{\mathbf{X}}_{mt} = \sum_{t' \in K_m} \eta_{t'}^m \bar{\mathbf{X}}_{mt'}$, and

$$w_t = \ln \left[\left(e^{\left(\gamma - \gamma^*\right)^T \bar{\mathbf{Y}}_t} - 1 \right) \left(\sum_{m \in S_t} e^{\boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}} + 1 \right) + 1 \right],$$

where $\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* \rangle$ are the true parameters. Note that $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ is always a feasible solution.

Eq. (3.10) are the resultant limiting non-linear equations for identifying unknown true parameter $\boldsymbol{\gamma}^*$ as $N \to \infty$. The assumption ensures that the unique solution is guaranteed using the chosen covariates and assortments across time buckets $\mathcal{T}$ for the true parameters. In most settings, the true parameters are unknown and the choice of the covariates cannot often be designed, yet the assumption is relatively easy to satisfy. *This is because, in general, $w_{t'}$ cannot be basis functions for generating w_t for any value of the true parameters $\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* \rangle$, given the choice of $\boldsymbol{\eta}^m$ vector for all choices m. So, even with a small number of time buckets and variation of covariates and assortments $(\bar{\mathbf{X}}_t, \bar{\mathbf{Y}}_t, S_t)$ across the time buckets, Eq. (3.10) will hold only if $\boldsymbol{\gamma} = \boldsymbol{\gamma}*$ and hence the assumption will hold.* We are unable to further simplify and quantify the precise condition to prescribe, as Eq. (3.10) is non-linear and so we state it as an assumption in the paper. Allowing $|\mathcal{T}| \to \infty$ guarantees the above assumption, while simulation experiments show that, a small number of time buckets (even just l) having a reasonable level of variation across the covariates or assortments is enough to identify the parameters.

In certain settings even the presence of certain types of time buckets guarantees assumption 2 for any $\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* \rangle$. For a pure assortment problem with no covariates for both choice and arrival rate, the assumption implies every choice must be offered at least once in the data and that there is at least one pair of time buckets that offer a common choice where the sum of the true attraction functions are not identical, i.e., $\sum_{m \in S_{t_1}} e^{\beta_m^*} \neq \sum_{m' \in S_{t_2}} e^{\beta_{m'}^*}$ where $S_{t_1} \cap S_{t_2} \neq \emptyset$ for some $t_1 \neq t_2 \in \mathcal{T}$. With unknown and potentially identical β_m^* across choices, this condition can be easily guaranteed if S_{t_1} is a strict subset of S_{t_2} or vice-versa.

THEOREM 1. *The following statements are true as $N \to \infty$:*

1. *The true parameters $\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* \rangle$ are asymptotically optimal to the LM-C problem, i.e.,*

$$\lim_{N\to\infty} Z^C\left(N, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{f}^*\left(\boldsymbol{\beta}^*\right)\right) = 0, \tag{3.11}$$

   where $\mathbf{f}^*(\boldsymbol{\beta}^*)$ is a vector of $\left(1 + \sum_{m' \in S_t} e^{\boldsymbol{\beta}^{*T}_{m'} \bar{\mathbf{X}}_{m't}}\right)^{-1}$ $\forall t \in \mathcal{T}$.
2. *The asymptotic optimal objective of the LM-C problem is also optimal, i.e.,*

$$\lim_{N\to\infty} Z^C\left(N, \boldsymbol{\beta}^N, \boldsymbol{\gamma}^N, \mathbf{f}^N\right) = 0. \tag{3.12}$$

3. *Under assumptions 1 and 2, the estimates $\boldsymbol{\beta}_m$ $\forall m \in M$, $\boldsymbol{\gamma}$ and f_t $\forall t \in \mathcal{T}$ using the LM-C problem are consistent with the true values as $N \to \infty$, i.e.,*

$$\lim_{N\to\infty} \left\langle \boldsymbol{\beta}^N, \boldsymbol{\gamma}^N, \mathbf{f}^N \right\rangle = \left\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{f}^* \right\rangle. \tag{3.13}$$

The proof of the theorem is provided in Appendix A. The first part of the theorem shows that the true parameters are optimal in the limit, meaning they achieve zero error to the LM-C problem. The second part ensures that, in the limit, the optimal solution to the LM-C problem achieves zero error and hence optimal as well. Therefore, with these parts, we can conclude that if any optimal solution converges to a unique solution in the limit then it coincides with the true parameters. The existence as well as the uniqueness property is shown in the third part of the theorem. This part of the proof heavily depends on the integrated optimization of both the share and size terms in the objective and reducing the limiting problem to a system of non-linear equations of the form Eq. (3.10) whose uniqueness is guaranteed by the assumptions 1 and 2, or by assuming sufficient variation in covariates and assortments across several time buckets.

To summarize, Theorem 1 proves that the estimator using the LM-C problem, which is the continuous version of the proposed integrated optimization method, is asymptotically optimal in the limit and in other words, consistent. This asymptotic property is the same as that achieved by the Berkson's estimator and the MLE based methods in complete data settings and using certain MLE based methods (e.g., Abdallah and Vulcano 2016) in special cases in censored data settings.

REMARK 1. *Assume that the true lost share in any time bucket is such that $f^* \in [\epsilon, 1-\epsilon]$ for a sufficiently small $\epsilon > 0$. Now suppose the PWL lost function approximates functions $\ln\left(\frac{1-f}{f}\right)$ and $\ln\left(\frac{1}{1-f}\right)$ with an accuracy α in the range $f \in [\epsilon, 1-\epsilon]$ and the Lipschitz constant for the loss function $\mathcal{L}[.]$ is ψ (e.g., ψ is 1 for L1 norm and 2 for L2 norm). Then $|Z - Z^C| \leq \psi\alpha$ where Z^C and Z are the objective functions of the LM-C and LM problems respectively.*

Because α is a function of the number and spread of the PWL segments that are chosen in the LM problem, it can be refined to approximate the LM-C problem arbitrarily closely. So, together with the theorem, *the proposed integrated approach (LM estimator) is consistent if both the number of arrivals in a time bucket as well as the number PWL segments go to infinity.* Even though the proofs are in the asymptotic limit, the finite data performance of the estimator using only a few knots (e.g., 10 to 20), is relatively good and discussed in Sections 3.4 and 6 respectively.

3.3. Model enhancements and extensions

Model enhancements: Partial information, user acceptance criteria, model selection, and regularization are easy to incorporate within the proposed method and we describe a few below:

1. *Partial or aggregate lost sales information*: Sometimes partial or aggregate loss sales information may be available from independent data companies such as Nielson and these can be included as constraints in the optimization formulation. Sample lost-share data can be employed to set up confidence-interval based restrictions to bound the range of the z_{kt} variables. For example, suppose a retailer knows that their market share is around $60 \pm 5\%$ over a duration of a month. This can be used to impose a constraint that says $35 \leq \sum_{k \in K, t \in \mathcal{T}_{month}} \frac{z_{kt}}{|\mathcal{T}_{month}|} \leq 45$. Prior information can also be used to optimize the placement and number of PWL knots. For example, in e-commerce, partial lost sales information obtained from abandoned shopping cart data can be gainfully incorporated within the MIP model to improve both runtime and solution quality.
2. *Model selection, regularization, sign-constraints, prior values and weighted optimization*: Constraints on prediction parameters are very useful in an automated demand estimation environment. For example, in a pricing application one expects to guarantee negative price coefficients. It is easy to incorporate such restrictions as well as prior values of prediction parameters into the MIP via bounds and constraints. Weighted or iterative weighted error minimization is useful in managing heteroskedasticity in the data (Theil 1970, Greene 2011), while ridge, lasso or elastic net penalties that are often employed in generalized linear regression techniques (James et al. 2013) are also easy to implement. An explicit constraint to identify the best subset of features as motivated in (Bertsimas et al. 2016) can also be included.

Model extensions: We describe a few different extensions of the proposed LM model below.

1. *Alternative discrete choice models:* We can extend the optimization method easily to other commonly employed discrete choice models. For example, for the MCI and linear attraction demand models, market share terms in the objective of the LM model can be replaced as follows:

$$\mathcal{L}\left[\ln(\bar{s}_{mt}) - \sum_{k \in K} \ln(\bar{\lambda}_{kt}) z_{kt} - \boldsymbol{\beta}_m^T \ln(\mathbf{X}_{mt})\right] \qquad \text{for MCI and} \tag{3.14}$$

$$\mathcal{L}\left[\sum_{k \in K} f_m^{-1}\left(\frac{\bar{s}_{mt}}{\bar{\lambda}_{kt}}\right) z_{kt} - \boldsymbol{\beta}_m^T \mathbf{X}_{mt}\right] \qquad \text{for generalized linear models.} \tag{3.15}$$

Here, $\ln(\bar{\mathbf{X}}_{mt})$ refers to a vector with the logarithm of each term of $\mathbf{X}_{mt}$, and generalized linear models employ attraction functions $f_m(\mathbf{X}_{mt})$ of the form $f_m(\boldsymbol{\beta}_m^T \mathbf{X}_{mt})$.

2. *Non-substitutive and positively correlated demands:* Consider the case where demand of one item is positively correlated with that of other items in the assortment (e.g., different sizes of a T-shirt). The customer buys their preferred item, if it is available, or walks away without making a purchase. This is the case when the items are neither substitutive nor complementary and can be represented using a hierarchical size-share demand model, with covariates for arrival rate but none for the share terms can be used, for example:

$$D_{mt}(\mathbf{Y}_t, S_t) = \begin{cases} \tau(\mathbf{Y}_t) \frac{\beta_m}{\sum_{m' \in M} \beta_{m'}}, & \forall\ m \in S_t. \\ 0 & \text{o.w.} \end{cases} \tag{3.16}$$

The market share terms are commonly referred to as size profiles. In the fashion retail industry, an accurate estimate of size profile is a critical input in designing good quality pre-packs (pre-determined quantities of different sizes and/or colors or styles of apparel boxed together) which are then allocated to the different stores, thereby reducing supply-chain complexity. Inaccurate size-profile estimates drive up supply chain costs. Out-of-stock scenarios are quite common during the life of the items that result in censored historical data. Eqs. (3.5–3.6) can be appropriately modified to derive a formulation very similar to the LM model.

3. *Tangential cutting planes to improve model performance:* Recall from Section 3.1 and Fig. 1 that the proposed method approximates the functions $\ln\left(\frac{f}{1-f}\right)$, and $\ln\left(\frac{1}{1-f}\right)$ which are the projected lost sales and market size per unit observed sales. We approximated these functions using a PWL approximation using knots $f_k \in (0,1)\ \forall k \in K$. Observe that these functions are composed of two modular functions $g(f) = \ln f$ and $h(f) = \ln(1-f)$ because $\ln\left(\frac{f}{1-f}\right) = g(f) - h(f)$ and $\ln\left(\frac{1}{1-f}\right) = -h(f)$. We can write the following valid inequalities for the functions $g(f)$ and $h(f)$ noting their concavity in $(0,1)$.

$$\ln(f_k) \le g(f) \le \ln(f_k) + \frac{1}{f_k}(f - f_k) \qquad \forall k \in K \tag{3.17}$$

$$\ln(1-f_k) \le h(f) \le \ln(1-f_k) - \frac{1}{1-f_k}(f - f_k) \qquad \forall k \in K \tag{3.18}$$

The LHS of the inequalities ensure that the functions $g(f), h(f)$ are above or equal to the PWL approximation we have in the problem. The RHS of the inequalities ensure that they are within the tangent approximations. Based on Fig. 1, it may be more beneficial to locate these tangents nearer the extremes of the boundary $(0,1)$. Along with the valid inequalities, the LM model can be modified by replacing the lost sales and the market size projection terms in the objective function by $g(f)$ and $h(f)$ respectively (appropriately also multiplying by the observed sales).

3.4. Computational experiments with simulated data

In this section, we study various computational aspects of the proposed approach in order to highlight some of its strengths and provide guidelines for solving the LM problem in practice. These experiments are done using simulated data, where the underlying model parameters are known, in order to analyze the performance of the proposed method in a controlled environment. We assume a MNL model where the utility of the purchasing choices are modeled using an intercept and a choice-dependent price coefficient, i.e., $f_j(p) = e^{\beta_0^j + \beta_1^j p^j}$ for choice j. In all experiments, a probabilisitic number of arrivals are generated (Poisson-distributed, unless stated otherwise) and the purchase choice of each arrival is simulated using the choice probabilities. The mean market size (same as arrival rate) and the share parameters are estimated using the LM model using only the observed sales data. We do not make any prior assumptions on the lost-share range. We employ the L1 loss function and use either 10 or 20 uniformly spaced knots in $[10^{-6}, 1-10^{-6}]$, unless stated otherwise. The absolute error deviation objective produces parameter values that tend to be relatively more robust, i.e. less sensitive to extreme data observations in the training data set and hence preferred. We report on computational times and evaluate the performance of the model in predicting observed sales and the unobserved lost sales as we vary a key model control parameters. We quantify the model-fit performance using the weighted mean absolute percentage error (WMAPE) metric that is commonly used by practitioners.

$$WMAPE = \frac{\sum_t |\text{predicted value}(t) - \text{actual value}(t)|}{\sum_t \text{actual value}(t)} * 100 \tag{3.19}$$

1. **Empirical bias and consistency:** In this experiment, we analyze the impact of increasing arrival rate on the quality of estimation of a three-parameter binary choice model. The mean arrival rates over 50 time buckets are varied from 5 to 20,480 by repeatedly doubling it across trials. The parameters of the single purchase choice are set to $\beta_0 = 3$ and $\beta_1 = -0.03$. The prices for each bucket are uniformly generated between $[0, 200]$ resulting in win rates between 40 to 60% on average for an instance. We generate 30 distinct instances for each mean arrival rate and estimate β_0, β_1 and γ using the LM model and 20 PWL knots.

The box plot of the estimated parameters for the different arrival rates are presented in Fig. 2. The ratio of the estimated to the true γ values are presented to enable comparison across different arrival rates. The true values of the parameters are marked with the dotted line as a reference. In the box plot, the line in the middle of the box is the median and the lower and upper edges of the box represent the 25th and 75th percentiles with the whiskers extending to the extreme points excluding outliers marked with a + sign. We observe that the parameters converge to the true values relatively quickly (no more than 40 arrivals, and within the box after no more than 20 arrivals) with lower biases in the parameter estimates achieved at higher arrival rates.

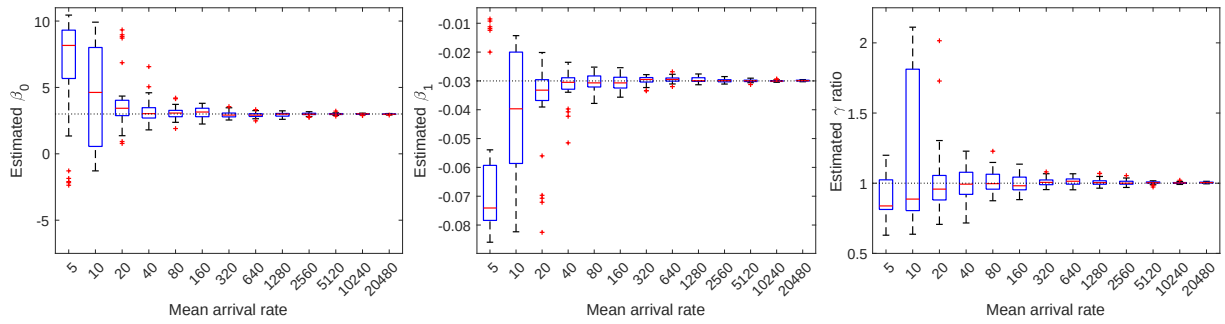

Figure 2 **Box plots of the estimated parameters β_0, β_1 and (normalized) γ with 30 distinct instances as a function of increasing mean arrival rate. The true values are marked with dotted lines.**

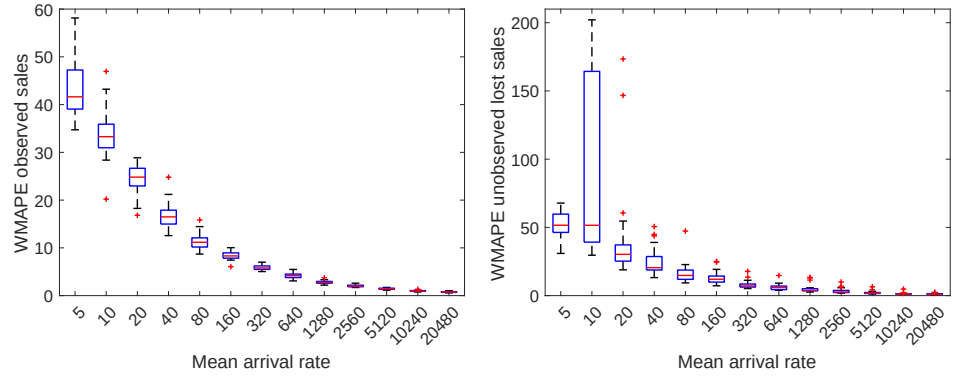

Figure 3 **Average WMPAE of the observed and unobserved sales as a function of the increasing arrival rates.**

In Fig. 3 we present the model fit WMAPEs for the observed and unobserved sales. We notice that the observed sales has a lower WMAPE than the unobserved lost sales across trials, lower by 4 percentage points even with a 20,480 arrival rate. The results in this section provide an empirical substantiation to the results in Theorem 1 and Remark 1.

2. **Model sensitivity to the granularity of piecewise-linearization:** In this experiment, we study the influence of the number of PWL knots on the tradeoff between predictive performance and run times. We generate 30 random model instances of a three-parameter binary choice model with a mean arrival rate of 5000 over 20 time buckets. Across instances, β_0 and β_1 are uniformly distributed in $[0.5, 2.5]$ and $[-0.04, -0.02]$ respectively. The prices for each bucket in an instance are uniformly distributed between $[0, 200]$ resulting in a wide range of win rates, between $[0.1, 99.9]$ percent across the different time buckets for the various instances under consideration. The LM model is used to estimate γ, β_0 and β_1 for the same 30 instances while varying the number of PWL knots. Fig. 4 presents the average model fit WMAPEs of the observed sales and unobserved lost sales as well as the average run times across the instances as a function of the number of knots. The key observation is that the performance of the estimation significantly improves with the initial increase in the number of PWL segments employed. Further increase in the number of

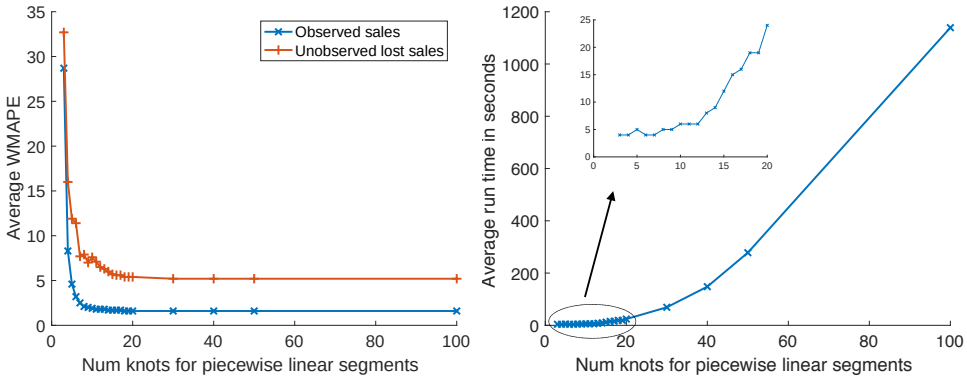

Figure 4 **Average WMAPEs and run times as a function of the number of piecewise linear lost sales segments.**

knots yields diminishing returns, with no discernible improvement beyond 20 knots. On the other hand, we observe the run times growing exponentially, highlighting the tradeoff between predictive performance and run times. Employing fewer, judiciously located knots further reduces run times in practice without sacrificing solution quality. Computational experiments exploring this tradeoff additionally with tangential cuts for the same 30 instances are presented in Appendix B.

3. **Model sensitivity to the optimality tolerance setting:** In this experiment, we study the model sensitivity to the user-tunable optimality tolerance parameter available in most commercial MIP solution packages that sets a relative tolerance on the gap between the best feasible integer objective value (upper bound) and the best lower bound (LP relaxation).

We choose three arrival rate settings to generate simulated data, which for the sake of comparison, we denote as sparse, medium and dense, having 100, 1000, and 10,000 arrivals per time bucket respectively, across 50 time buckets. We model 5 purchasing choices with an average win rate of 30%, 10% and 4-6% for the remaining 3 choices. The average loss rate is 40%, ranging between 10-80% across time buckets. For each arrival rate, we generate 100 random instances and estimate the parameters for these instances using the LM model, which is run with different optimality tolerance levels using 10 PWL knots.

Table 2 summarizes the results. We observe that a tighter optimality tolerance leads to an improvement in predictive performance but yields diminishing returns beyond a 5% range, suggesting that near-optimal solutions are likely to be adequate in practice. Dense data sets always result in a better solution quality compared to sparser data sets, and sometimes even with a more relaxed optimality tolerance. This result highlights the benefit of increasing aggregation whenever possible, and its impact on consistency. Interestingly, the improvement in solution quality is steeper for unobserved lost sales over observed sales. In fact, the observed sales WMAPE does not change at all for dense arrivals, while the corresponding improvement is small for the sparse and medium instances. This is due to the pareto-optimality of the choice parameters for any lost share values

Optimality	Sparse (100)			Medium (1000)			Dense (10,000)		
Tolerance	WMAPE		Runtime	WMAPE		Runtime	WMAPE		Runtime
	Unobs	Obs	(in secs)	Unobs	Obs	(in secs)	Unobs	Obs	(in secs)
40%	2.99	0.49	11	0.93	0.33	15	0.29	0.22	56
20%	1.55	0.51	13	0.37	0.26	29	0.18	0.22	64
10%	0.92	0.52	14	0.2	0.25	36	0.13	0.22	71
5%	0.52	0.47	26	0.15	0.24	51	0.12	0.22	75
1%	0.31	0.43	746	0.14	0.24	885	0.12	0.22	137

Table 2 Impact of optimality tolerance on the average WMAPEs and run times

chosen, while the lost shares themselves are being refined as the optimality tolerance is tightened. Therefore, as a note, while a tighter optimality tolerance may have an immediately visible impact on the performance of observed data, it can result in better model identification, which leads to better performance in forecasting applications.

From a run time perspective, although the sparser instances converge faster, the run times for the dense data sets are lesser for a given performance level. As the optimality tightens from 5% to 1%, we observe that the corresponding run time increases by a factor of 2-10 for dense to medium instances, while the achieved WMAPE quality remains steady, suggesting that most of the CPU time is spent by the solver trying to generate the certificate of optimality. A strength of the proposed optimization based approach is that we can achieve a good balance between run time and the desired solution quality in practice by choosing an appropriate optimality tolerance.

4. **Model sensitivity to different loss functions:** Here we compare the achieved solution quality and run times when using L1 and L2 loss functions. We use an optimality tolerance of 1% and employ the same data generation process of experiment 3. Table 3 summarizes the results of the study. From the perspective of achieved WMAPE, L1 loss functions seem to outperform the L2 loss function except in very sparse instances. We also note that L2 is on-par with L1 for very dense arrival rates. This is possible because WMAPE is a first order metric like L1 and in general, these trends depend on the choice of the metric of evaluation. From a run time perspective, we notice that the L2 loss function, interestingly, is an order of magnitude faster (even though it results in a second order MIP) than the L1 loss function, unless the arrival rate is very high. Methods that combine L1 and L2 using side constraints and other risk mitigating loss functions, can be adopted in practice, depending on the application and solution requirements.
5. **Estimating arrival rate covariates:** In Fig. 5, we provide illustrative examples of various types of mean market size trends that can be modeled using the proposed approach without any distributional assumptions on the arrivals. In each example, we estimate the unobserved mean values of the market size, and the share parameters of the observed single choice sales. We plot the

Arrival Rate	L1			L2		
	WMAPE		Runtime	WMAPE		Runtime
	Unobs	Obs	(in secs)	Unobs	Obs	(in secs)
10	1	0.53	384	**0.91**	**0.47**	**70**
25	**0.57**	0.59	596	0.93	0.55	**39**
100	0.31	**0.43**	746	0.67	0.56	**54**
1000	**0.14**	**0.24**	885	0.17	0.26	**144**
10000	**0.12**	**0.22**	**137**	**0.12**	**0.22**	150

Table 3 Impact of the choice of loss functions on the average WMAPEs and run times

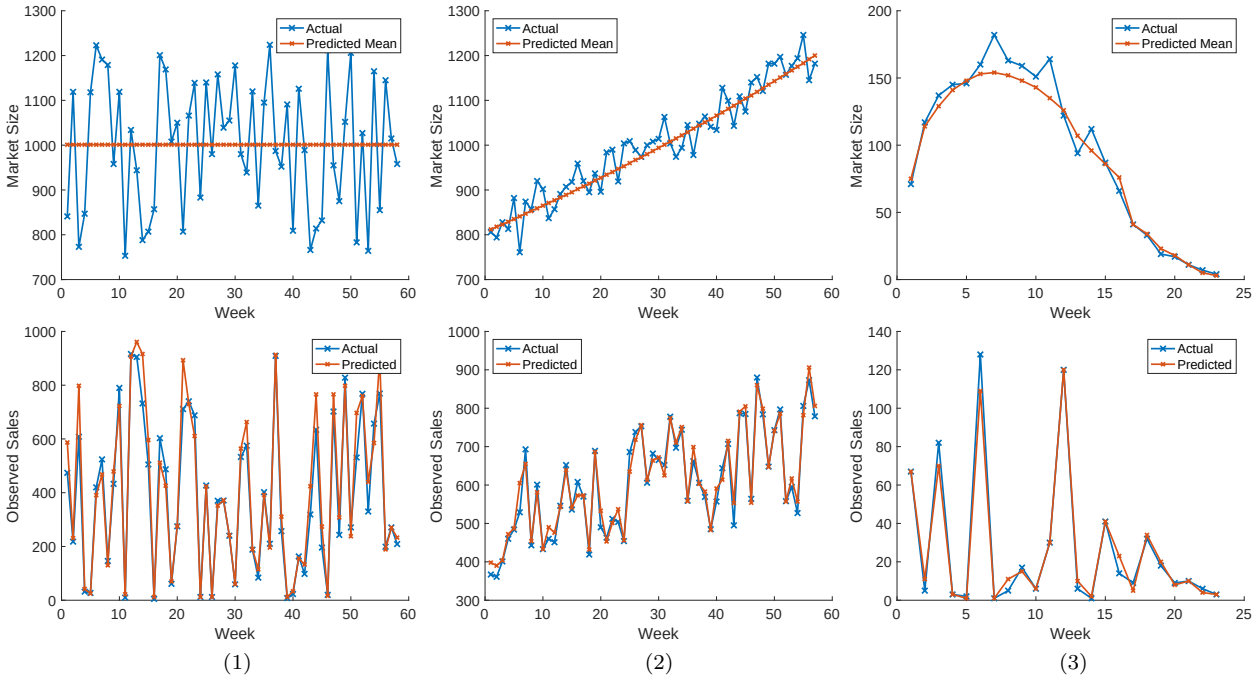

Figure 5 Mean unobserved distribution-free market size and observed sales estimates when the true market size has (1) a uniform distribution and constant arrival rate, (2) a Poisson distribution with a linear mean arrival rate and (3) a Poisson distribution with a product life cycle mean arrival rate.

(simulated) actual and (mean) predicted market sizes as well as the fitted observed sales for three simulated instances: (1) Uniform distribution with a constant arrival rate; (2) Poisson distribution with linear time dependent mean arrival rate, i.e., $\gamma_0 + \gamma_1 Y_t$ where $Y_t = \frac{t}{|\mathcal{T}|}$; and, (3) Poisson distribution with a product life cycle mean arrival rate, i.e., $\gamma_o (Y_t)^{\gamma_1} (1 - Y_t)^{\gamma_2}$ where $Y_t = \frac{t}{|\mathcal{T}|}$. In settings (1) and (3), we work with the log-linear measure of error in the LM model for the market size term but switch to a linear measure of error for setting (2). The plots highlight the wide applicability of the methodology, in particular, to general settings beyond Poisson distributions and constant arrival rates, a typical assumption in many earlier papers.

4. Estimating competitor parameters: scenarios with more than one unobserved substitutive choice

In this section, we present a method that enables a seller to learn the market shares and attractiveness of their competitors by analyzing their own censored historical sales data and available competitor product attribute data. Consider a setting with multiple unobserved substitutive options, including no-purchase. With the emergence of big-data technologies, sellers can continually track data streams such as competitive assortment, prices and promotions that affect the seller's demand. A heuristic approach is to incorporate these effects as additional factors within the seller's attraction model and treat the entire lost sales as just one unobserved choice. However, this approach provides little information about how much and when sales opportunities are lost to key competitors. To address this limitation, we propose an alternative estimation approach that models lost sales to competition as additional substitutive choices and treat it different from lost sales due to no-purchase. Essentially, this requires the total unobserved sales by the seller to be disambiguated based on the competitor's attributes (prices, promotions, holidays and events) that are observed.

In addition to the no-purchase option $\emptyset$, let $M_u \subset M$ be the set of competitor purchasing options choices that are unobserved by the retailer. The other choices are fully observed. For simplicity of exposition, we assume that all the options are always available, noting that our method extends to include assortment changes as discussed in the previous section using availability sets S_t. Let α_{mt} denote the the proportion of the total unobserved lost share that is attributed to choice $m \in M_u \cup \emptyset$ in period t. We impute the sales for choice m, $D_{mt} = \alpha_{mt} \sum_{k\in K} \bar{\lambda}_{kt} z_{kt}$ We model α_{mt} as a PWL function over the set of possible discrete values $\bar{\alpha}_{mjt} \in (0,1)$, i.e., $\alpha_{mt} = \sum_{j\in J} \bar{\alpha}_{mjt} w_{mjt}$ where w_{mjt} is the probability that α_{mt} is at the discrete value $\bar{\alpha}_{mjt}$. We now linearize this combined expression ($D_{mt} = \left[\sum_{j\in J} \bar{\alpha}_{mjt} w_{mjt}\right] \sum_{k\in K} \bar{\lambda}_{kt} z_{kt}$) for imputed sales and substitute $Q_{mt} = \ln D_{mt}$ to obtain constraint (4.2) in the formulation below. We now formalize the full model:

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{z},\mathbf{w},Q} \quad \sum_{t\in\mathcal{T}} \sum_{m\in M} \mathcal{L}\left[Q_{mt} - Q_{\emptyset t} - \boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt}\right] + \sum_{t\in\mathcal{T}} \mathcal{L}\left[\sum_{k\in K} \ln(\bar{\theta}_{kt}) z_{kt} - \boldsymbol{\gamma}^T \bar{\mathbf{Y}}_t\right] \qquad \textbf{(Comp-LM)}$$

$$Q_{mt} = \ln(\bar{s}_{mt}) \qquad \forall\, t \in \mathcal{T}, m \in M \backslash M_u \qquad (4.1)$$

$$Q_{mt} - \sum_{k\in K} \ln(\bar{\lambda}_{kt}) z_{kt} - \sum_{j\in J} \ln(\bar{\alpha}_{mjt}) w_{mjt} = 0 \qquad \forall\, t \in \mathcal{T}, m \in M_u \cup \emptyset \qquad (4.2)$$

$$\sum_{m\in M_u\cup\emptyset} \sum_{j\in J} \bar{\alpha}_{mjt} w_{mjt} = 1 \qquad \forall\, t \in \mathcal{T} \qquad (4.3)$$

$$\sum_{j\in J} w_{mjt} = 1 \qquad \forall\, t \in \mathcal{T}, m \in M_u \cup \emptyset \qquad (4.4)$$

$$\sum_{k\in K} z_{kt} = 1 \qquad \forall\, t \in \mathcal{T} \qquad (4.5)$$

$$\{w_{mjt} \;\forall\; j \in J\} \in \text{SOS2} \qquad \forall\, t \in \mathcal{T}, m \in M_u \cup \emptyset \qquad (4.6)$$

$$\{z_{kt} \ \forall \ k \in K\} \in \ \text{SOS2} \qquad \forall \ t \in \mathcal{T} \tag{4.7}$$

$$w_{mjt}, z_{kt} \geq 0. \tag{4.8}$$

Disambiguation constraint (4.2) delineates the share of each unobserved choice based on its observed attributes, while constraints (4.3–4.4) ensures that these proportions sum to 1.0 for every time period. We model the w variables as SOS2 variables, similar to the z variables. The resultant optimization instances for popular loss functions like L1 and L2 can be directly solved by standard optimization software packages.

The Comp-LM model integrates two layers of learning. It learns the unobserved share, and from this learned value, further disambiguates the no-purchase from the competitor purchase options using the relative variations in their attributes over time. Consequently, this problem is harder to solve compared to the no-competitor case. We observed that relatively more time buckets, with sufficient variation in the covariates (and/or assortments), are required to recover a solution of comparable quality, which in turn increases the computational time. This observation will be highlighted later in this section and Section 6 along with computational results. Below we show that even in this setting, the continuous version of the proposed approach results in consistent estimators after assumption 2 is extended as stated below.

ASSUMPTION 3. *For at least one choice of sets K_m $m \in M \backslash M_u$ as described in assumption 1 the following system of equations has a unique solution for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_{M \backslash M_u}$:*

$$w_t = \sum_{t' \in K_m} \eta_{t'}^m w_{t'} \ \forall t \in \mathcal{T} \backslash K_m, \ m \in M_u. \tag{4.9}$$

where $\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* \rangle$ are the true parameters. Here $\boldsymbol{\eta}^m$ satisfies $\bar{\mathbf{X}}_{mt} = \sum_{t' \in K_m} \eta_{t'}^m \bar{\mathbf{X}}_{mt'}$, and

$$w_t = \ln \left[\left(e^{(\gamma - \gamma^*)^T \bar{\mathbf{Y}}_t} - 1 \right) \left(\sum_{m \in S_t} e^{\boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}} + 1 \right) + 1 + \sum_{m \in M_u} e^{\boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}} \right] - \ln \left(1 + \sum_{m \in M_u} e^{\boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt}} \right),$$

where $\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* \rangle$ are the true parameters. Note that $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ and $\boldsymbol{\beta}_{M \backslash M_u} = \boldsymbol{\beta}^*_{M \backslash M_u}$ is always a feasible solution.

Eq. (4.9) are the resultant limiting non-linear equation for identifying unknown true parameter $\boldsymbol{\gamma}^*$ and $\boldsymbol{\beta}^*_{M \backslash M_u}$ as $N \to \infty$ and relatively complex compared to Eq. (3.10). Similar to what we stated in Section 3.2, we know that *in general, $w_{t'}$ cannot be basis functions for generating w_t for any value of the true parameters $\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* \rangle$, given the choice of $\boldsymbol{\eta}^m$ vector for all choices m. So, even with a small number of time buckets and variation of covariates and assortments ($\bar{\mathbf{X}}_t, \bar{\mathbf{Y}}_t, S_t$) across the time buckets, Eq. (4.9) will hold only if $\boldsymbol{\gamma} = \boldsymbol{\gamma}*$ and $\boldsymbol{\beta}_{M \backslash M_u} = \boldsymbol{\beta}^*_{M \backslash M_u}$ and hence the assumption will be true.* Eq. (4.9) highlights the higher degree of complexity of the non-linear equations we have

to satisfy and hence the need for additional time buckets with distinct covariates and assortments compared to the setting with no competitors in order to achieve a similar quality level.

We state the consistency theorem in this setting for completeness.

THEOREM 2. *Let Comp-LM-C be the continuous version of the Comp-LM problem and Z^{CC} be its objective function. Then the following statements are true as $N \to \infty$:*

1. *The true parameters $\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* \rangle$ are asymptotically optimal to the Comp-LM-C problem, i.e.,*

$$\lim_{N\to\infty} Z^{CC}\left(N, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{f}^*, \boldsymbol{\alpha}^*\right) = 0, \tag{4.10}$$

   where $\mathbf{f}^*$ is a vector of $\left(1+\sum_{m\in M_u} \boldsymbol{\beta}_m^{*T}\bar{\mathbf{X}}_{mt}\right)\left(1+\sum_{m\in M} \boldsymbol{\beta}_m^{*T}\bar{\mathbf{X}}_{mt}\right)^{-1}$ $\forall t \in \mathcal{T}$ and $\boldsymbol{\alpha}^*(\boldsymbol{\beta}^*)$ is a $|M_u| \times |\mathcal{T}|$ matrix of the form $\alpha^*_{mt} = \alpha^*_{\emptyset t} e^{\boldsymbol{\beta}_m^{*T}\bar{\mathbf{X}}_{mt}}$ $\forall m \in M_u$ and $\alpha^*_{\emptyset t} = \left(1+\sum_{m\in M_u} \boldsymbol{\beta}_m^{*T}\bar{\mathbf{X}}_{mt}\right)^{-1}$.

2. *The asymptotic optimal objective of the Comp-LM-C problem is also optimal, i.e.,*

$$\lim_{N\to\infty} Z^{CC}\left(N, \boldsymbol{\beta}^N, \boldsymbol{\gamma}^N, \mathbf{f}^N, \alpha^N\right) = 0. \tag{4.11}$$

3. *Under assumptions 1 and 3, the estimates $\boldsymbol{\beta}_m$ $\forall m \in M$, $\boldsymbol{\gamma}$, f_t $\forall t \in \mathcal{T}$ and α_{mt} $\forall m \in M_u \cup \emptyset, t \in \mathcal{T}$ using the Comp-LM-C problem are consistent with the true values as $N \to \infty$, i.e.,*

$$\lim_{N\to\infty} \left\langle \boldsymbol{\beta}^N, \boldsymbol{\gamma}^N, \mathbf{f}^N, \alpha^N \right\rangle = \left\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{f}^*, \alpha^* \right\rangle. \tag{4.12}$$

The proof of the above theorem is very similar to Theorem 1 with some differences, in particular, the derivation of the limiting Eq. (4.9), the proof of which is provided in Appendix C.

Empirical bias and consistency with hidden competitor choice: Similar to experiment 1 in Section 3.4, we estimate a five-parameter two purchase choice model with Poisson arrivals and study the impact of increasing arrival rate, where additionally, the sales of one of the choices is unobserved, along with the no-purchase option. We simulate the arrivals over 50 bins where the mean arrival rates are varied from 5 to 40,960 by repeatedly doubling it across experiments. The parameters of the two choices are set to $\beta_0^1 = 4$, $\beta_1^2 = -0.04$ $\beta_0^2 = 2$ and $\beta_1^2 = -0.02$. Here, the sales data for choice 2 (competitor's choice) is unobserved. The prices for each bucket are uniformly generated between $[0, 200]$ resulting in an average percentage win rate range of $[25, 50]$ and $[27, 42]$ for choice 1 and 2 respectively across instances. We generate 30 distinct instances for each mean arrival rate and estimate β_0^1, β_1^1, β_0^2, β_1^2 and γ using the Comp-LM model using 20 PWL knots for f and α values uniformly distributed in $[10^{-3}, 1-10^{-3}]$. Note that the absolute share of no-purchase is obtained as a product of the α and f and can be as small as 10^{-6}.

The box plot of the estimated parameters for the different arrival rates are presented in Fig. 6. We observe that the parameters converge to their true parameters values relatively quickly with lower biases in the parameter estimates achieved at higher arrival rates. The rate of convergence is

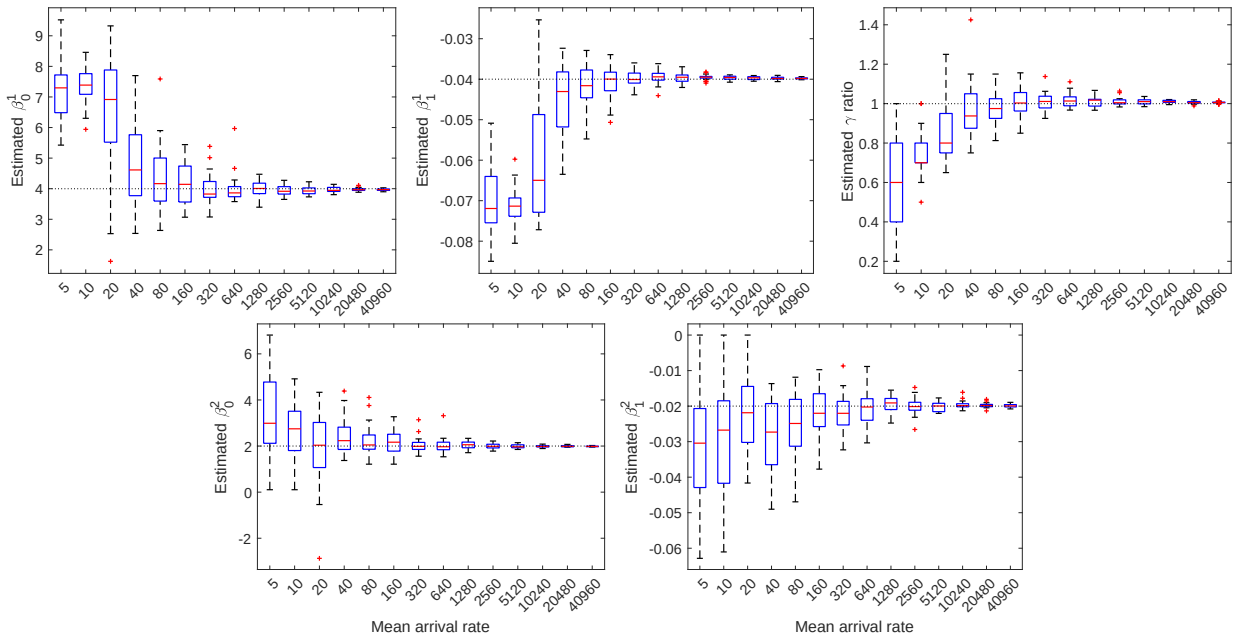

Figure 6 **Box plots of the estimated parameters $\beta_0^1, \beta_1^1, \beta_0^2, \beta_1^2$ and (normalized) γ with 30 distinct instances as a function of increasing mean arrival rate. The true values are marked with dotted lines. Here, the competitor sales (choice 2) is hidden in addition to lost sales.**

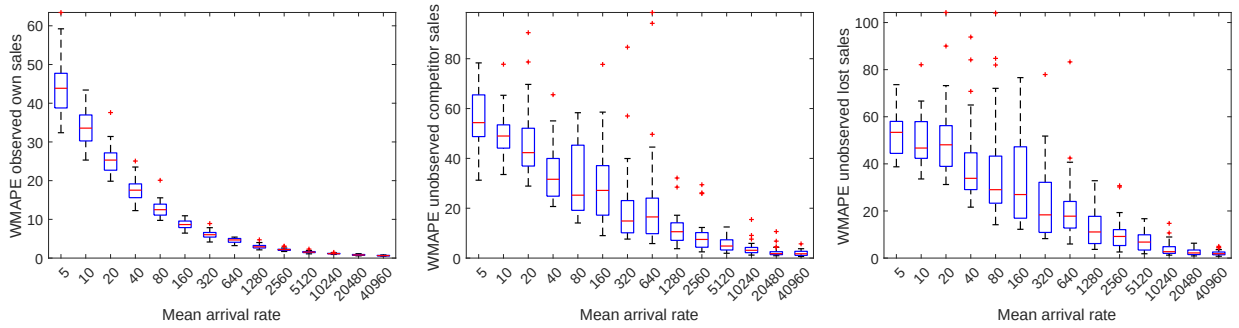

Figure 7 **Average WMPAE of the observed and unobserved sales as a function of the increasing arrival rates. Here, the competitor sales are hidden in addition to lost sales.**

slower than the no-competitor case, where an arrival rate of 10 was sufficient to contain the true parameters within the box, whereas this requirement increases to 40 in this setting. In Fig. 7, we present the model fit WMAPEs for the observed and unobserved sales to competitor as well as no-purchase. The observed sales WMAPE continues to outperform both the unobserved WMPAEs.

5. Estimating average win-probability from real-life RFQ data

In this section, we apply the LM model to demonstrate the practical efficacy and viability of the proposed loss minimization models for decision support in RFQ settings. We analyze and present results for a real-life data set associated with RFQs for a leading information technology (IT) service provider that sells a number of different hardware products to its buyers in a B2B

environment. The buyers submit bids to the seller specifying the desired purchase quantity of each of the hardware products, and the seller, in return, specifies prices for each item. Depending on the prices offered, the buyer chooses to purchase or not to purchase some, or all the items in the RFQ. Typically, the seller balances the profit margin, the customer's willingness to pay, order size, inventory and supply-chain goals, market-strategy, availability of generic (OEM) alternatives, as well as long term sales objectives, in order to determine these prices. To support this complex tradeoff goal, a variety of product, customer, supply chain, and market attributes are tracked over time, and recorded with every RFQ. In this data set, we have historical RFQ data for four different hardware components recorded over a duration of 30 weeks. Each record shows whether a particular RFQ was won or lost, and is associated with a feature vector whose components include the order date (week), normalized price, order size, product quality, manufacturing cost, list price, product category and competitor price. Typically, this feature vector gradually evolves over time, and does not change significantly week-to-week.

A quick analysis of the purchase data revealed a decline in sales for all products over the last 30 weeks. In Fig. 8 we plot the normalized weekly total sales trend across all products. Our goal is to use only the historical data associated with wins (losses are assumed to be hidden) in order to discover the possible reasons for this decline by completing the following two tasks:

1. Predict the weekly arrival rate of RFQs and compare with the actual average arrival rate
2. Predict the average win rate over a 30-week period and compare with the actual win-rate

These tasks turn out to be practically important because we encountered several prospective clients in the B2B space (e.g., chemicals, food) who only recorded historical purchases (wins) in their database, and wanted to know whether their sales trends was due to natural seasonal variations that affected arrivals, or due to recent pricing changes that affected their win rate, or a combination of these two factors. Furthermore, by knowing the future arrival rate trend, one can accordingly adjust prices to achieve a relatively steady sales-rate, which may be beneficial to the up-stream manufacturing facility. Ideally, this data censoring in RFQ settings has to be avoided and is encouraged as a standard practice moving forward in these engagements.

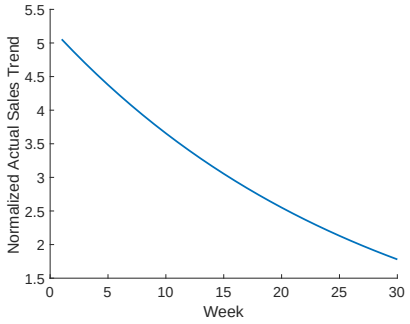

Figure 8 Weekly normalized total actual sales trend across all products.

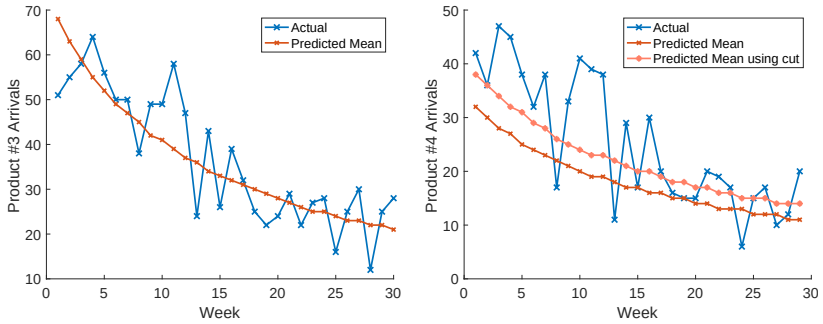

Figure 9 **Weekly actual censored arrivals plotted against its predicted mean.**

Product	Number of RFQs	Actual Win Rate	Predicted Win Rate
Product 1	1515	32.1	31.0
Product 2	1131	27.9	27.4
Product 3	1102	30.1	31.2
Product 4	735	28.3	36.6
Product 4 (with constrained win rate)	735	28.3	30.4

Table 4 **Average actual and predicted win rate in the RFQ application**

To achieve the two goals using the LM method, we aggregate individual winning quotes into weekly sales bins. Doing so captures the aggregate wins and average buyer response to the prevailing product attributes for a week. We use an exponential arrival rate model with week-index as a size attribute, and an MNL binary choice model with the remaining product specific attributes. We employ 10 PWL knots uniformly distributed in the range $[10^{-3}, 1-10^{-3}]$, and an optimality tolerance of 1%. The results below show that the LM method was able to produce practically useful predictions regarding the arrival rates and win rates.

In Fig. 9 we plot the mean weekly predicted arrivals against the censored actual values for two hardware products for illustration. The LM model was able to correctly identify a seasonal decline in weekly RFQ arrivals that affected all hardware components, thereby confirming that the observed negative trend in wins was less likely to be due to any adverse pricing recommendations by the application. Furthermore, in three of the four cases, we were able to learn the average win rate, see Table 4. The fourth product had relatively sparse purchase data, which impacted our prediction accuracy. In practical RFQ situations, customers were often able to provide an approximate estimate of average market-share. To simulate the effect of partial information, we used a one-week (first week) sample of win-loss data to generate a 90% confidence interval for the win-rate of the fourth product. The sample lost share was 79% out of 42 quotes and the corresponding 90% confidence interval based heuristic cut added to the MIP, restricted the average lost share values to a range [0.69,0.89] (i.e., $0.79 \pm 1.65\sqrt{.79(1-.79)/42}$). Re-solving the resultant

optimization model enabled us to improve the win rate by 6 percentage points from 36.6% to 30.4%. The cumulative CPU time for running the LM model across all four products was 129 seconds.

6. Simulation testing with publicly available hotel data set parameters

This experiment is based on a single large urban hotel data ("Hotel 1") that is made publicly available (Bodea et al. 2009). A choice model specification that closely represents the behavior of this real hotel data set with product attributes was created and used by Newman et al. (2014). Being a realistic simulation of a market environment, we used their parameters and simulation strategy to test the performance of the LM and Comp-LM methods.

Arrivals are Poisson distributed, and every arrival selects from a total of 9 choices: 8 room choices in the assortment offered at their respective price, and a no-buy option. There are three nested common price elasticity terms associated with 14-day ahead, 7-day ahead, and 1-day ahead purchases, respectively, with the earlier bookings being more price sensitive. In all, a total of 11 independent MNL parameters (9 intercept terms for each choice, with the no-purchase intercept set to 0, and 3 price effects), and the arrival rate, have to be estimated. The simulated historical data is in the form of 28-day booking curves for one year's worth of check-in days. The no-buy transactions are not observed. The goal of our experiment is to use the proposed LM method to identify the model parameters using aggregated purchase data. We also test the Comp-LM method on the same instances by hiding the sales data of one of the room choices.

Our data generation and aggregation process is as follows. We simulate Poisson arrivals for every booking day and record their choices, which are aggregated across 100 bins. Assuming an average of 40 transactions per day, we obtained approximately 40,000 ($\sim$40 x 360 x 28 x 0.1) purchases in all. For every bin, we generated random price vectors which are uniformly distributed between their minimum and maximum values as specified in Newman et al. (2014), and the resultant average lost share was approximately 90%. Every bin is associated with one of the 28 booking days. Doing so enables the variations in the willingness-to-pay for any particular booking day to be observed across at least 3 bins. For brevity, we do not simulate stock-outs and assume that all rooms are available for all booking days.

We formulate the LM (or the Comp-LM, equivalently) using the L1-norm, with $T = 100$ and $m = 8$. We employ 20 uniformly distributed knots with $\epsilon = 10^{-6}$ (or 10^{-3} for the Comp-LM) to generate the PWL model for the lost share. As in the rest of the paper, we do not use any information regarding the hidden market-share to customize the PWL knot locations. The resultant parameter estimation model is solved using a 1% optimality tolerance (5% for the Comp-LM). To assess the practical viability of our model in terms of achieving a good quality solution within a limited run time, we impose a 100K branch-and-bound node limit in the CPLEX solver on a 2.7 GHz 16GB RAM Windows notebook computer.

Parameter	True Value	LM Method		EM Method	
		Estimate	% Error	Best Estimate	% Error
β_{Nobuy}	0	0	–	0	–
β_{King1}	5.3	5.3292	0.6	5.2528	−0.9
β_{King3}	4.3465	4.3350	−0.3	4.2703	−1.8
β_{King4}	5.3488	5.3645	0.3	5.2974	−1.0
β_{Queen1}	3.9869	3.9809	−0.1	3.9309	−1.4
$\beta_{Special}$	4.2074	4.1994	−0.2	4.1445	−1.5
β_{Suite1}	7.6141	7.6175	0.0	7.5485	−0.9
β_{Suite2}	5.176	5.1741	0.0	5.0840	−1.8
β_{TwoDbl}	4.2262	4.2544	0.7	4.1651	−1.4
β_{price}	−0.01719	−0.01701	−1.1	−0.01683	−2.1
$\beta_{price,day\geq 1}$	−0.00361	−0.00367	1.6	−0.00368	1.9
$\beta_{price,day\geq 14}$	−0.00193	−0.00186	−3.7	−0.00191	−0.9
λ	40	38.424	−3.9	38.136	−4.7
Cum. Time (secs)	–	10 (single run)		8,642 (converged 13),	26,453 (all 20)
Cum. # Iterations	–	1		34,657 (converged 13),	104,657 (all 20)

Table 5 Estimated parameters of the proposed LM method and the EM algorithm for an instance.

In the first experiment, we compare the solution quality of the LM method against the EM method. We employed the EM algorithm described in Talluri and Van Ryzin (2004) with minor modifications to account for aggregated bins (see Vulcano et al. 2012 for the incomplete likelihood function used). The EM method was terminated when the absolute change in parameter values between successive iterations was less than 0.001, or if the iteration limit of 10,000 was reached. For a simulation instance, Table 5 summarizes the true and estimated parameter values and the cumulative run times for the two methods. Note that because we take the utility of the no-purchase option to be zero, we recalibrate the original parameter values from Newman et al. (2014) to reflect this setting.

We observe that, for the given model settings, the solution quality of the LM method is on par with that achieved by the EM algorithm. Recall that the EM method is well-known to converge to local optima and so a multi-start procedure was employed. We initialized the EM using 20 different starting points, and 13 of those instances converged before reaching the maximum iteration limit. We present the best parameter estimates for the EM algorithm, in terms of the likelihood value, which also happens to be the mode estimate (6 of the 20 runs). The cumulative run time for the EM required to produce this mode estimate is 7.4 hours, while it took 10 seconds for the LM method to achieve a solution having the same or better quality. Similar quality of the parameters using the two-step method for a different instance are reported in Table 1 of the Newman et al. (2014) paper. In contrast to the multi-start approaches, the main benefit of our MIP based implicit enumeration approach is that it returns a certificate of (near-)global optimality, which the earlier methods are unable to provide.

Parameter	True value	Mean estimate	% Error in mean	Coeff. of variation	90% confidence interval
β_{Nobuy}	0	0	0.0	–	[0,0]
β_{King1}	5.3	5.3413	0.8	0.040	[5.0055 , 5.6578]
β_{King3}	4.3465	4.3858	0.9	0.046	[4.0539 , 4.7059]
β_{King4}	5.3488	5.3910	0.8	0.038	[5.0355 , 5.7048]
β_{Queen1}	3.9869	4.0251	1.0	0.051	[3.7147 , 4.3246]
$\beta_{Special}$	4.2074	4.2464	0.9	0.048	[3.8934 , 4.5407]
β_{Suite1}	7.6141	7.6643	0.7	0.031	[7.2831 , 8.0366]
β_{Suite2}	5.176	5.2071	0.6	0.042	[4.8347 , 5.5311]
β_{TwoDbl}	4.2262	4.2648	0.9	0.048	[3.9239 , 4.5916]
β_{price}	−0.01719	−0.01725	0.3	−0.021	[−0.01776 , −0.01660]
$\beta_{price,day\geq 1}1$	−0.00361	−0.00362	0.3	−0.064	[−0.00394 , −0.00327]
$\beta_{price,day\geq 14}$	−0.00193	−0.00188	−2.7	−0.029	[−0.00197 , −0.00179]
λ	40	40.28	0.7	0.134	[33.3 , 48.7]

Table 6 Empirical unbiasedness of the estimated parameters with the proposed LM method over 50 simulation instances. The average run time is 11.4 seconds.

In the second experiment, we analyze the empirical unbiasedness of the LM model in a realistic environment. We re-run the estimation method multiple times, using a different randomization seed each time to generate different data sets. In Table 6, we report the average parameter estimate and run times. We also calculate the coefficient of variation and a 95% confidence interval for each estimated MNL parameter, and the arrival rate. All parameters estimates are within 3% of the true value, with a coefficient of variation of less than 0.14. The average run time per instance is 11.4 seconds. These findings are positive from a practical perspective, particularly, given that we are solving an underlying non-convex optimization problem to near global optimality.

The third experiment, similar to the second, analyzes the empirical unbiasedness of the Comp-LM model, wherein additionally, the king4 room's sales history as well as the no-buy data are unavailable to the model. The results for the same instances generated in the previous experiment are reported in Table 7 under the columns titled '20 PWL'. We observe that all parameters exhibit higher deviations from their true values in comparison to the LM model, ranging between 4-7% for the attraction coefficients and 14% for the arrival rate. The run time is approximately 7 times more (78 seconds on average) despite using a relaxed optimality tolerance setting of 5%. This increase in run-time and empirical bias is indicative of the challenging nature of the Comp-LM's MIP. In Table 7, we also present results when we increase the arrival rate by a factor of 100 and employ 50 PWL knots. This was done in order to assess the impact of increased modeling accuracy and data on solution quality. We were able to improve the results of the Comp-LM model and bring it closer to the performance of the LM model, while doubling the run times to 137 seconds on average. From the coefficient of variation (CV) stand point, interestingly, the performance of the Comp-LM method with 20PWL is similar to the LM-method (less than .12) while it increases marginally for

Parameter	True value	20 PWL segments			50 PWL segments, 100X arrival rate		
		Mean estimate	% Error in mean	Coeff. of variation	Mean estimate	% Error in mean	Coeff. of variation
β_{Nobuy}	0	0	–	–	0	–	–
β_{King1}	5.3	5.5446	4.6	0.046	5.3634	1.2	0.054
β_{King3}	4.3465	4.5896	5.6	0.052	4.4102	1.5	0.065
β_{King4}	5.3488	5.5323	3.4	0.049	5.3741	0.5	0.035
β_{Queen1}	3.9869	4.2311	6.1	0.058	4.0508	1.6	0.071
$\beta_{Special}$	4.2074	4.4510	5.8	0.055	4.2709	1.5	0.067
β_{Suite1}	7.6141	7.8578	3.2	0.039	7.6769	0.8	0.040
β_{Suite2}	5.176	5.4156	4.6	0.050	5.2385	1.2	0.056
β_{TwoDbl}	4.2262	4.4719	5.8	0.056	4.2899	1.5	0.067
β_{price}	−0.01719	−0.01703	−0.9	−0.024	−0.01702	−1.0	−0.030
$\beta_{price,day\geq 1}$	−0.00361	−0.00386	6.9	−0.067	−0.00378	4.7	−0.152
$\beta_{price,day\geq 14}$	−0.00193	−0.00189	−1.9	−0.032	−0.00193	−0.1	−0.031
λ	40	34.6	−13.5	0.111	39.2	−2.0	0.236
Run time		78 secs			137 secs		

Table 7 **Empirical unbiasedness of the estimated parameters with the proposed Comp-LM method over 50 simulation instances where the purchase history of room king4 is also unobserved.**

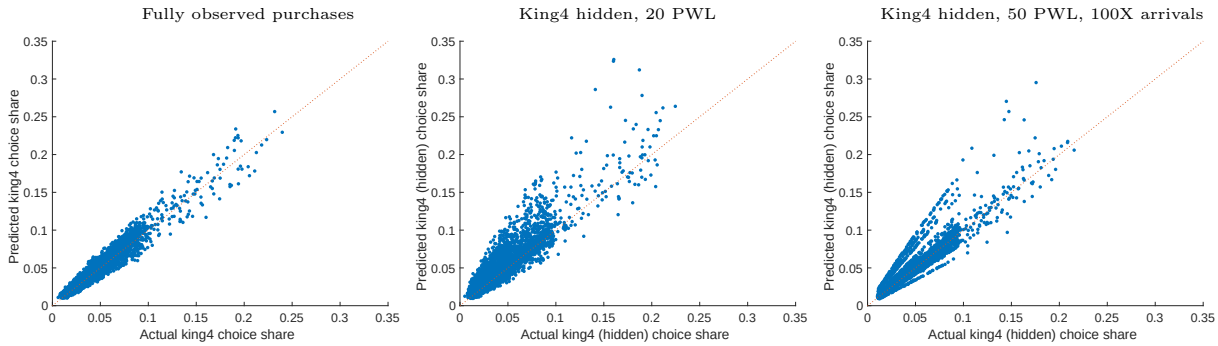

Figure 10 **Scatter plot of the predicted versus the actual market share of king4 room choice.**

the 50PWL case. This slight increase in variation may be due to the fact that the solution quality (achieved optimality gap) within the preset CPLEX node limit for the easier 20PWL instances was about 4 percentage points better than that achieved in the 50PWL case.

Despite the highly challenging nature of this problem due to the dual layer of learning involved, the performance of the model is practically sound. For example, Fig. 10 shows the ability of the LM and Comp-LM models to successfully predict the weekly market share of the king4 room choice compared to the actuals across all the instances under the various scenarios considered above. We notice that all points clustered around the 45 degree line, more so when king4 room's sales history was uncensored than in censored case. The average competitor market-share estimated by the Comp-LM model was 4.7% and 4.2% for the 20PWL and 50PWL cases respectively, compared to the actual value of 4.1%. The average estimated lost share was 83.9% and 85.6% respectively,

compared to the true value of 86.4%. Overall, the Comp-LM models results are encouraging from a prediction standpoint.

Appendix A: Proof of Theorem 1

Proof of part 1: We know that

$$\lim_{N\to\infty}\frac{\bar{s}_{mt}}{N}=e^{\boldsymbol{\gamma}^{*T}\bar{\mathbf{Y}}_t}\frac{e^{\boldsymbol{\beta}_m^{*T}\bar{\mathbf{X}}_{mt}}}{1+\sum_{m'\in S_t}e^{\boldsymbol{\beta}_{m'}^{*T}\bar{\mathbf{X}}_{m't}}}\qquad\forall m\in S_t,t\in\mathcal{T}\tag{A.1}$$

Substituting the true parameters $\langle\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\mathbf{f}^*(\boldsymbol{\beta}^*)\rangle$ and Eq. (A.1) in $Z^C(N,\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{f})$, proves this part of the theorem that the true parameters are asymptotically optimal to the LM-C problem

Proof of part 2: The existence of a finite minimum for the LM-C problem for any given data set N is always guaranteed because the objective function of the LM-C problem is (1) continuous convex with boundaries tending to positive infinity in the prediction parameters $\langle\boldsymbol{\beta},\boldsymbol{\gamma}\rangle$ for given lost share values, $f_t\ \forall t\in\mathcal{T}$ and (2) continuous and bounded below in the lost share values that are in a bounded range, i.e., $f_t\in(0,1)\ \forall t\in\mathcal{T}$, given the prediction parameters.

We also know that given data set N, the optimal solution of the LM-C problem is a lower bound to any other feasible solution, in particular when the decision variables are equal to their true parameter values:

$$0\le Z^C\left(N,\boldsymbol{\beta}^N,\boldsymbol{\gamma}^N,\mathbf{f}^N\right)\le Z^C\left(N,\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\mathbf{f}^*(\boldsymbol{\beta}^*)\right).\tag{A.2}$$

Therefore, taking the limit at $N\to\infty$ and combining it with the results of the above part (i.e., Eq. (3.11)), we get Eq. (3.12), proving this part of the theorem.

Proof of part 3: As $N\to\infty$, we have a infinite sequence of f_t^N in a bounded interval $(0,1)$, and there must be at least one converging subsequence of f_t^N. Consider the same converging subsequence of the $f_t^N\ \forall t\in\mathcal{T}$. Substituting this subsequence in the LM-C problem, reduces the LM-C problem to the complete information setting. We know that the first and second term in the complete information settings are consistent. Note that consistency for the first term is obtained from the consistency of the Berkson's estimator and the consistency of the second term is derived from standard statistical theory where we see observations (arrivals) in a bin tending to infinity, and hence the arrival rate converges to its true arrival rate for that bin. This in turn means, that we have a corresponding converging subsequence for $\langle\boldsymbol{\beta}^N,\boldsymbol{\gamma}^N\rangle$ for the LM-C problem.

Now consider these converging subsequences and say they converge as $N\to\infty$ to $\left\langle\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{f}_t\right\rangle$. We show that any set of subsequences converge to the true parameters $\langle\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,f_t^*\rangle$. This mean the limit exists and also proves that the estimates are consistent. We prove this by contradiction.

From part 2 of the theorem, i.e., Eq. (3.12), we know that every loss function term in the objective has to be equal to zero as the loss function is by definition non-negative. We now arrive at a system of equations, one for each loss function term, by substituting the converging subsequences of LM-C problem by $\left\langle\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{f}_t\right\rangle$ and also, substituting the observed sales with the true parameters $\langle\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\mathbf{f}^*(\boldsymbol{\beta}^*)\rangle$ in the limit using Eq. (A.1).

$$\ln\left(\frac{1-\hat{f}_t}{\hat{f}_t}\frac{f_t^*}{1-f_t^*}\right)=\left(\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_m^*\right)^T\bar{\mathbf{X}}_{mt},\qquad\forall\ m\in S_t,t\in\mathcal{T},\ \text{and}\tag{A.3}$$

$$\ln\left(\frac{1-f_t^*}{1-\hat{f}_t}\right)=\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\right)^T\bar{\mathbf{Y}}_t,\qquad\forall\ t\in\mathcal{T}.\tag{A.4}$$

where $f_t^* = \left(1 + \sum_{m \in S_t} e^{\boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}}\right)^{-1}$.

Our goal is to show that the above system of equations has a unique solution which are indeed the true parameters. First we eliminate $\hat{f}_t$ from the above equations to get:

$$\hat{\boldsymbol{\beta}}_m^T \bar{\mathbf{X}}_{mt} = \boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt} - \ln\left[\left(e^{(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T \bar{\mathbf{Y}}_t} - 1\right)\left(\sum_{m \in S_t} e^{\boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}} + 1\right) + 1\right] \qquad \forall\ m \in S_t, t \in \mathcal{T}. \tag{A.5}$$

The second term in the RHS varies with the assortment as well as the feature vector of the other choices, and even that of the arrival rate. We first show below that $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^*$. We begin by choosing $K = \sum_{m \in M} k_m$ equations from Eq. (A.5) that are linearly independent in the choice features, in particular, k_m per choice $m \in M$. This is possible because of assumption 1. We solve for the K unknown $\hat{\boldsymbol{\beta}}$ values and substitute it in the remaining equations to find the $\hat{\boldsymbol{\gamma}}$ vector. For ease of exposition, we introduce some notation.

We consider two types of matrices of $[\bar{\mathbf{X}}_{mt}]$ for the K equations each trying to simplify the representation of LHS and second term of the RHS in Eq. (A.5) respectively. We denote the matrices by $\mathbf{X}$ and $\tilde{\mathbf{X}}$. They are both $K \times K$ matrices where the rows correspond to the selected K equations and the columns correspond to the K feature vectors of all the M choices. In every row of $\mathbf{X}$, which corresponds to a specific equation of Eq. (A.5), the feature vector of all choices except the specific choice m that resulted in that equation are zero. In every row of $\tilde{\mathbf{X}}$, which corresponds to a specific equation of Eq. (A.5), the feature vector of all choices except those in the assortment offered at time t that resulted in that equation are zero. The corresponding feature vector matrix $[\bar{\mathbf{Y}}_t]$ for the K equations is denoted $\mathbf{Y}$. The matrixes for the remaining equations (at least l in number) are denoted by $\mathbf{X}'$, $\tilde{\mathbf{X}}'$ and $\mathbf{Y}'$. We know that $\mathbf{X^{-1}}$ exists because every row is linearly independent. Also, let $\mathbf{w}(\mathbf{Y}, \tilde{\mathbf{X}}, \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$ be the vector of $\ln\left[\left(e^{(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T \bar{\mathbf{Y}}_t} - 1\right)\left(\sum_{m \in S_t} e^{\boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}} + 1\right) + 1\right]$ for the K equations and rest are denoted by $\mathbf{w}(\mathbf{Y}', \tilde{\mathbf{X}}', \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$.

The selected K equations from Eq. (A.5) therefore have the following form:

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{\beta}^* - \mathbf{w}(\mathbf{Y}, \tilde{\mathbf{X}}, \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \tag{A.6}$$

Multiplying for $\mathbf{X^{-1}}$ on either side, we get,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* - \mathbf{X^{-1}}\mathbf{w}(\mathbf{Y}, \tilde{\mathbf{X}}, \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \tag{A.7}$$

The remaining equations of Eq. (A.5) to identify $\hat{\boldsymbol{\gamma}}$ are

$$\mathbf{X}'\hat{\boldsymbol{\beta}} = \mathbf{X}'\boldsymbol{\beta}^* - \mathbf{w}(\mathbf{Y}', \tilde{\mathbf{X}}', \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \tag{A.8}$$

Substituting $\hat{\boldsymbol{\beta}}$ from Eq. (A.7) in Eq. (A.8) we get,

$$\mathbf{X}'\boldsymbol{\beta}^* - \mathbf{X}'\mathbf{X^{-1}}\mathbf{w}(\mathbf{Y}, \tilde{\mathbf{X}}, \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) = \mathbf{X}'\boldsymbol{\beta}^* - \mathbf{w}(\mathbf{Y}', \tilde{\mathbf{X}}', \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \tag{A.9}$$

$$\implies \mathbf{w}(\mathbf{Y}', \tilde{\mathbf{X}}', \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) = \mathbf{X}'\mathbf{X^{-1}}\mathbf{w}(\mathbf{Y}, \tilde{\mathbf{X}}, \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \tag{A.10}$$

This equation is the same as Eq. (3.10) where $\mathbf{X}'\mathbf{X^{-1}}$ results in the superposition vector because the matrix $\mathbf{X}$ is orthogonal and so $\mathbf{X}'$ can be decomposed into elements of $\mathbf{X}$. Under assumption 2, we are guaranteed uniqueness of this system of equations resulting in $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^*$. In order to provide more insight into this equation and assumption 2, observe that the vector $\mathbf{w}(\mathbf{Y}, \tilde{\mathbf{X}}, \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$ are like basis functions and can potentially

generate other vectors of the form $\mathbf{w}(\mathbf{Y}', \tilde{\mathbf{X}}', \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$. More specifically, the vectors should be generated from the exact superposition vectors of $\mathbf{X}'\mathbf{X}^{-1}$. Furthermore, the RHS is also independent of $\mathbf{Y}'$. With a sufficient variation of the covariates and assortments even for a small number of time buckets, this cannot be true and the assumptions holds relatively easily in practice.

Therefore, $\mathbf{w}(\mathbf{Y}, \tilde{\mathbf{X}}, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}}) = 0$ and from Eq. (A.7), $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$. Therefore, prediction parameters are identical to the true parameters and so are the $\hat{f}_t$ values. Hence Eqs (4.12) and the theorem. □

Appendix B: Computational experiments with tangential cutting planes

Table 8 provides the results with and without tangential cutting planes for 30 random instances generated for experiment 2 in Section 3.4. The achieved WMAPE of the model is better in the case with the valid inequalities than without, but there is a runtime tradeoff. The improvement is significant for the cases with few knots and in particular, for unobserved sales over observed sales. As the number of knots increases, the WMAPE for the observed sales stabilizes in both scenarios even though the WMAPE for the unobserved lost sales continues to slightly improve whenever the tangential cuts are employed. We observe that the run times can be reduced while preserving solution quality, especially in the cases with more knots, by including just a few tangents. This tradeoff between solution quality and run times can additionally be exploited in practical settings.

No. of piecewise linear knots	No tangent cuts			With tangent cuts		
	WMAPE Unobs	WMAPE Obs	Runtime (in secs)	WMAPE Unobs	WMAPE Obs	Runtime (in secs)
4	16	8.3	4	5.9	3.3	5
6	11.4	3.2	4	6.4	1.9	5
8	7.9	2.1	5	6.1	1.7	7
10	7.6	1.9	6	5.6	1.7	11
20	5.4	1.6	24	4.7	1.6	50
50	5.2	1.6	278	5.1	1.6	418
100	5.2	1.6	1139	5	1.6	2778

Table 8 Impact of using tangent cuts on the average WMAPEs and run times

Appendix C: Proof of Theorem 2

Part 1, part 2 and the existence of the limit follow similar steps as in the proof of Theorem 1 provided in Section A. The main difference is the derivation of the limiting equations which we show here.

Equations equivalent to Eqs. (A.3–A.4) in this setting are as follows:

$$\ln\left(\frac{1-\hat{f}_t}{\hat{\alpha}_{\emptyset t}\hat{f}_t}\frac{\alpha^*_{\emptyset t} f^*_t}{1-f^*_t}\right) = \left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_m\right)^T \bar{\mathbf{X}}_{mt}, \qquad \forall\, m \in M\backslash M_u, t \in \mathcal{T}, \tag{C.1}$$

$$\ln\left(\frac{\hat{\alpha}_{mt}}{\hat{\alpha}_{\emptyset t}}\right) = \hat{\boldsymbol{\beta}}^T_m \bar{\mathbf{X}}_{mt}, \qquad \forall\, m \in M_u, t \in \mathcal{T}, \tag{C.2}$$

$$\ln\left(\frac{1-f^*_t}{1-\hat{f}_t}\right) = \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right)^T \bar{\mathbf{Y}}_t, \qquad \forall\, t \in \mathcal{T}, \text{ and} \tag{C.3}$$

$$\sum_{m \in Mu} \hat{\alpha}_{mt} + \hat{\alpha}_{\emptyset t} = 1, \qquad \forall\, t \in \mathcal{T}. \tag{C.4}$$

where $f^*_t = \left(1 + \sum_{m\in M_u} e^{\boldsymbol{\beta}^{*T}_m \bar{\mathbf{X}}_{mt}}\right)\left(1 + \sum_{m\in M} e^{\boldsymbol{\beta}^{*T}_m \bar{\mathbf{X}}_{mt}}\right)^{-1}$, $\alpha^*_{\emptyset t} = \left(1 + \sum_{m\in M_u} e^{\boldsymbol{\beta}^{*T}_m \bar{\mathbf{X}}_{mt}}\right)^{-1}$. Substituting

for $\hat{f}_t$ using Eq. (C.3) and $\hat{\alpha}_{mt}\ \forall m \in M_u$, $\hat{\alpha}_{\emptyset t}$ using Eqs. (C.2–C.4) in Eq. (C.1), we get,

$$\hat{\boldsymbol{\beta}}_m^T \bar{\mathbf{X}}_{mt} - \ln\left(1 + \sum_{m \in M_u} e^{\hat{\boldsymbol{\beta}}_m^T \bar{\mathbf{X}}_{mt}}\right) = \boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}$$
$$- \ln\left[\left(e^{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^T \bar{\mathbf{Y}}_t} - 1\right)\left(\sum_{m \in M} e^{\boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}} + 1\right) + 1 + \sum_{m \in M_u} e^{\boldsymbol{\beta}_m^{*T} \bar{\mathbf{X}}_{mt}}\right] \qquad \forall\ m \in M \backslash M_u, t \in \mathcal{T}. \tag{C.5}$$

The proof after this point again follows similar steps as in the earlier proof and using assumption 3 instead of assumption 2 we can conclude that the consistency of the parameters is achieved. □

References

Abdallah T, Vulcano G (2016) Demand estimation under the multinomial logit model from sales transaction data, working Paper.

Ben-Akiva ME, Lerman SR (1985) *Discrete choice analysis: theory and application to travel demand*, volume 9 (MIT press).

Berkson J (1953) A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association* 48(263):565–599.

Berry ST (1994) Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics* 242–262.

Bertsimas D, King A, Mazumder R, et al. (2016) Best subset selection via a modern optimization lens. *The Annals of Statistics* 44(2):813–852.

Bertsimas D, Tsitsiklis JN (1997) *Introduction to linear optimization*, volume 6 (Athena Scientific).

Bodea T, Ferguson M, Garrow L (2009) Data set-choice-based revenue management: Data from a major hotel chain. *Manufacturing & Service Operations Management* 11(2):356–361.

Cox DR (1970) *Analysis of binary data* (Methuen's Statistical Monograph).

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 1–38.

Domencich TA, McFadden D (1975) *Urban travel demand-a behavioral analysis* (North-Holland).

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

Greene WH (2011) *Econometric analysis* (Pearson Education).

Guadagni PM, Little JD (1983) A logit model of brand choice calibrated on scanner data. *Marketing science* 2(3):203–238.

Haensel A, Koole G (2011) Estimating unconstrained demand rate functions using customer choice sets. *Journal of Revenue & Pricing Management* 10(5):438–454.

James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning*, volume 6 (Springer).

Keller PW, Levi R, Perakis G (2014) Efficient formulations for pricing under attraction demand models. *Mathematical Programming* 145(1-2):223–261.

Kök AG, Fisher ML (2007) Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research* 55(6):1001–1021.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics* 105–142.

Newman JP, Ferguson ME, Garrow LA, Jacobs TL (2014) Estimation of choice-based models using sales data from a single firm. *Manufacturing and Service Operations Management* 16(2):184–197.

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

Reibstein DJ, Gatignon H (1984) Optimal product line pricing: The influence of elasticities and cross-elasticities. *Journal of Marketing Research* 21(3).

Rusmevichientong P, Shen ZJM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research* 58(6):1666–1680.

Subramanian S, Sherali HD (2010) A fractional programming approach for retail category price optimization. *Journal of Global Optimization* 48(2):263–277.

Talluri K (2009) A finite-population revenue management model and a risk-ratio procedure for the joint estimation of population size and parameters. Technical report, Universitat Pompeu Fabra.

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

Theil H (1970) On the estimation of relationships involving qualitative variables. *American Journal of Sociology* 103–154.

Tofallis C (2015) A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society* 66(8):1352–1362.

Train KE (2009) *Discrete choice methods with simulation* (Cambridge university press), 2nd edition.

Urban GL (1969) A mathematical modeling approach to product line decisions. *Journal of Marketing Research* 40–47.

van Ryzin G, Vulcano G (2011) An expectation-maximization algorithm to estimate a general class of non-parametric choice models. *preprint* .

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

Vulcano G, Van Ryzin G, Ratliff R (2012) Estimating primary demand for substitutable products from sales transaction data. *Operations Research* 60(2):313–334.

Wu CJ (1983) On the convergence properties of the em algorithm. *The Annals of statistics* 95–103.