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-) optimal parameter values with respect to the chosen loss-minimization (LM) objective function, thereby overcoming a limitation of prior multi-start heuristic approaches that 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. For large LM instances, we derive a nonconvex-convex alternating heuristic that can be used to obtain quick, near-optimal solutions. Partial information, user acceptance criteria, model selection, and regularization techniques can be incorporated to enhance practical efficacy. We test the LM model on simulated and real data, and present results for a variety of single- and multi-item, time varying arrival rate, large-scale demand prediction scenarios, and a dual-layer estimation model extension to learn 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 one's own products or services, but also be able to accurately predict 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. Tracking these lost opportunities may require the purchase of expensive syndicated data regarding competitor sales,

as well as video hardware and software to identify lost transactions, which is not yet available to a vast majority of sellers and service providers. On the other hand, the demand forecasting methods presented in this work are designed to provide a seller a panoramic view of future demand using readily 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). These demand models have a convenient hierarchical structure, which first models the arrivals of interested customers, and then the purchase probability of specific customer choices amongst alternatives. The demand for a customer choice is calculated as the product of the number of arrivals and its corresponding purchase probability. Customer choice scenarios that arise in practice includes (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); and, (3) non-substitutive settings where demands are positively correlated within the assortment (e.g., different T-shirt sizes). 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 been developed (e.g., Talluri and Van Ryzin 2004, Rusmevichientong et al. 2010, Keller et al. 2014).

Calibrating discrete choice models to accurately predict future demand can be challenging when complete historical data regarding all the customer choices is unavailable. This prediction problem has received attention in recent years and our paper proposes novel methods using discrete optimization to solve it effectively in practice.

Contributions: We present an integrated optimization approach to parameter estimation for calibrating discrete choice demand models where historical data regarding one or more choice alternatives are censored. We consider two distinct settings: (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 variations in 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.

For typical (e.g., daily or weekly) demand forecasting instances, a black-box MIP solver can recover proven, optimal prediction parameter values with respect to the loss-minimization objective

under consideration. Furthermore, partial information, user acceptance criteria, model selection and regularization techniques can be incorporated to enhance the efficacy of the proposed MIP in practice. For large scale instances (e.g., thousands of observations), we exploit the special structure of the integrated MIP and propose a nonconvex-convex alternating heuristic (AH) to obtain good quality (near optimal) solutions.

We show that solving the integrated model to optimality (in both aforementioned censored data settings) leads to asymptotically consistent estimators of all the parameters under mild assumptions. It is worth noting that this result does not require additional information for model identification. Also, there is no prior work that we are aware of that considers (1) disambiguation of unobserved arrival rate variations, or (2) the estimation of the unobserved competitor share of lost sales in discrete choice model estimation using censored data. We present several numerical experiments that include arrival rate variations and large scale, 'big data' scenarios to highlight the flexibility of the model and report on empirical results obtained using real world data. We show that the proposed methods dominate EM, achieving better overall solution quality while consuming only a fraction of the EM run time.

We have commercially deployed our method to predict omnichannel demand for retailers for pricing applications (Harsha et al. 2019). We have successfully implemented these forecasting methods over a wide range of industrial settings, from predicting price-sensitive time-of-day electricity load in a 'smart grid' to forecasting demand for chemical and food manufacturers.

1.1. Background and related literature review

Complete information about all customer choices: A 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). MLE in its simplest form maximizes the likelihood of a choice(s) given the assortment, which is a convex optimization problem. A commonly used alternative is the Berkson's method (Berkson 1953, Ben-Akiva and Lerman 1985). This method minimizes the error between the observed and predicted values of the log-transformed market share ratio between any two purchase choices resulting in a simple linear regression problem. The observational data for the MLE method can be as granular as the actual response of an individual and the associated covariate vector. In contrast, the Berkson's method requires grouped data in terms of counts or proportions which consists of the aggregate response of a group of individuals to a common feature vector. Furthermore, even with grouped data, MLE methods can seamlessly handle zero sales, while the Berkson's method requires certain heuristic adjustments to apply the log-transform, making MLE a viable alternative in a broad range of settings. For most parametric choice models, both problems are relatively easy to solve numerically using different

optimization algorithms, yet have the same theoretical asymptotic properties such as consistency and asymptotic normality (Greene 2011).

As we will see next, under incomplete information, most parameter estimation approaches have focused on the extensions of the MLE method, which employ gradient descent methods to solve the resultant non-convex likelihood objectives. In contrast, our parameter estimation approach extends the Berkson's method by employing state-of-art linear and mixed integer programming modeling techniques (Bixby 2012) to find (near-) globally optimal solutions.

Incomplete information about customer choices: Talluri and Van Ryzin (2004), in their work on choice based revenue management, develop an iterative 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. 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, multiple EM runs with different starting points are executed to probabilistically improve the quality of parameter estimates, often resulting in prohibitive run times. However, EM approaches are attractive because of their relative ease of implementation and wide applicability.

We refer to Vulcano et al. (2010) and Newman et al. (2014) for a study of the empirical performance of the EM algorithm, and the former paper for an EM implementation 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 for arrivals following a non-homogeneous (time-dependent) Poisson distribution. However, to avoid multiple optimal solutions or a misidentification of the choice parameters, the market share under the full assortment is required as an additional input. For the same setting, the Minorization-Maximization (MM) approach of Abdallah and Vulcano (2017) with this additional input provably converges to a stationary point. For the pure assortment case in the absence of covariates, the authors state sufficient conditions for identifiability (and hence consistency) of the choice parameters.

Newman et al. (2014) propose a computationally fast two step estimation method by decomposing and optimizing the incomplete data likelihood function assuming a homogeneous Poisson constant arrival rate model. A one-dimensional search is required in the second step to optimize a non-convex objective, and a multi-start procedure is employed to probabilistically overcome local

Property	Proposed LM method	Referenced MLE based methods
Number of observed choices	1 or more	EM: 1 or more; Two step, MM: 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
Zero sales	Needs heuristic adjustments	No adjustment required
Probabilistic assumptions	None	Required
Covariates for arrivals	Modeled	Not modeled
Optimization solution approach	Single step MIP; Alternating heuristic for large instances; (Near-) global optimal solution with a certificate of optimality	At least two steps or iterative; Multi-start heuristic procedures; Local maxima guarantee;
Asymptotic consistency	Proved	No choice covariates and additional market share data: EM, MM 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.

optima. Talluri (2009) also proposes a two-step method wherein a risk-ratio error minimization is instead adopted.

As a remark, except for the classical EM method, other methods in the literature cannot be directly implemented in the binary-choice (buy vs. no-buy) setting, as the first step requires at least two observed purchasing options. This is a basic setting (logistical regression) and is popular in a variety of practical demand forecasting and pricing applications.

The aforementioned MLE based methods, in the presence of share covariates, employ multi-start heuristic procedures to overcome local maxima when solving the resultant non-convex MLE problem. We propose a single step MIP approach that solves a non-convex loss minimization problem to provable (near-) optimality. Another modeling drawback of MLE based methods is that customer arrivals are modeled with a Poisson random variable, independent of covariates. In reality, arrival rates vary over time and additional covariates have to be incorporated to capture such trends. Examples include, the increase in store traffic and sales during holidays and the lifecycle behavior to bookings and sales for perishable products. By modeling these trends, we do not require any additional market share data for model identification and are able to disambiguate the observed sales fluctuations into arrival variations and purchase probability changes. Furthermore, our paper also estimates the unobserved purchase probability of competitors. There is no prior work that addresses either of these disambiguation problems. There are two practical limitations of the proposed MIP model: it requires heuristic adjustments to manage zero sales, and requires grouped data. Our proposed methods can be deployed within applications compatible with these limitations. 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.

In contrast to the above literature on censored choices, Berry (1994) uses uncensored aggregate market share data to estimate mixed-logit models using a log-ratio transformation (similar to the Berkson's approach) where there are unobservable features, an aspect we do not cover in this paper.

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 purchase probability. 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:

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

$$= \quad \tau(\mathbf{Y}_t) \ \frac{\mathcal{A}_m(\mathbf{X}_{mt})}{1+\sum_{m' \in S_t} \mathcal{A}_{m'}(\mathbf{X}_{m't})}, \quad m \in S_t, \tag{2.2}$$

$$D_{mt}(\mathbf{Y}_t, \mathbf{X}_t, S_t) = \quad 0, \qquad\qquad\qquad\qquad m \in M \backslash S_t, \text{ and} \tag{2.3}$$

$$D_{\emptyset t}(\mathbf{Y}_t, \mathbf{X}_t, S_t) = \quad \tau(\mathbf{Y}_t) \ \frac{1}{1+\sum_{m' \in S_t} \mathcal{A}_{m'}(\mathbf{X}_{m't})}, \tag{2.4}$$

where

- $\tau(\mathbf{Y}_t)$ is the model for arrivals or market size that is a measure of consumers interested in any of the choices, including the no-purchase option, as a function of a vector of the market size attributes $\mathbf{Y}_t$ at time t,
- $\mathcal{A}_m(\mathbf{X}_{mt})$ 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 features, covariates and attributes throughout the paper, to refer to regressors 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 (e.g., temporal and 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.

Purchase/choice probability models how consumers choose between alternatives, and is defined by the relative attractiveness of an alternative (including no purchase with a default attraction of 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 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 $\mathcal{A}_m(X_{mt}) = e^{\boldsymbol{\beta}_m^T \mathbf{X}_{mt}}$, in the case of the MCI model is $\mathcal{A}_m(\mathbf{X}_{mt}) = \prod_k X_{mtk}^{\beta_{mk}}$, and in the case of a linear attraction model is $\mathcal{A}_m(\mathbf{X}_{mt}) = \boldsymbol{\beta}_m^T \mathbf{X}_{mt}$ 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, and 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).

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 the 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, and discuss model enhancements and extensions in Section 3.3.

3.1. Proposed estimation method

Table 2 describes the notation used. We assume that when a choice is not in the offer set, the corresponding observed sales are zero, i.e., $m \notin S_t$, $\bar{s}_{mt} = 0$, and that the observed sales are strictly positive otherwise, i.e., $\bar{s}_{mt} > 0 \ \forall m \in S_t$. The latter assumption is reasonable in most instances because we work with aggregated data. In Appendix B we discuss we discuss how we relax this

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 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 arrivals in period t
$f_k,$	lost 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 share in period t is f_k.

Table 2 Notation used.

assumption for settings with low levels of aggregation and employ an adjustment method when zero sales are present and comprise a significant portion of the training data. We 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 an MNL choice 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:

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

Transforming Eqs. (3.1–3.2) using log and ratio-transformations, we obtain:

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

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

The first equation employs the log-ratio transformations like the Berkson's estimator and the second is the log-sum of the Eqs. (3.1–3.2). We now employ a piecewise linear (PWL) transformation to tractably model the unknown lost sales λ_t. To do so, we introduce a discrete sampling of possible lost

shares in $[\epsilon, 1-\epsilon]$, where ϵ is a small positive number and denote them by $f_k \in K$. The underlying assumption here is that we do not observe either extreme of lost shares. We use an auxiliary decision variable set $z_{kt}\ \forall \in K$ for each observation where z_{kt} corresponds to the probability that the lost share in period t is f_k and model them as a Special-Ordered-Set Type 2 (SOS2) variables wherein at most two adjacent members of this ordered set (assuming the f_k's are ordered) can be non-zero. Therefore, the PWL representation of the true share in any period can be written exactly as $f_k z_{kt}$, while the true lost sales λ_t is approximated with a PWL function $\bar{\lambda}_{kt} z_{kt}$ where $\bar{\lambda}_{kt} = \frac{f_k}{1-f_k} \sum_{m \in S_t} \bar{s}_{mt}$. Substituting these expressions, we get:

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

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

where $\bar{\theta}_{kt} = \frac{1}{1-f_k} \sum_{m \in S_t} \bar{s}_{mt}$.

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. Towards this, 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 LM objective function 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 $x \to \pm\infty$. We focus on the L1 or the L2 norms. The constraints (3.7–3.9) imposes the SOS2 nature of the lost share variables. We do not expand on the SOS2 form as all standard optimization packages allow 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 additional binary variables required to model the SOS2 conditions in the presence of non-convexity.

[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).

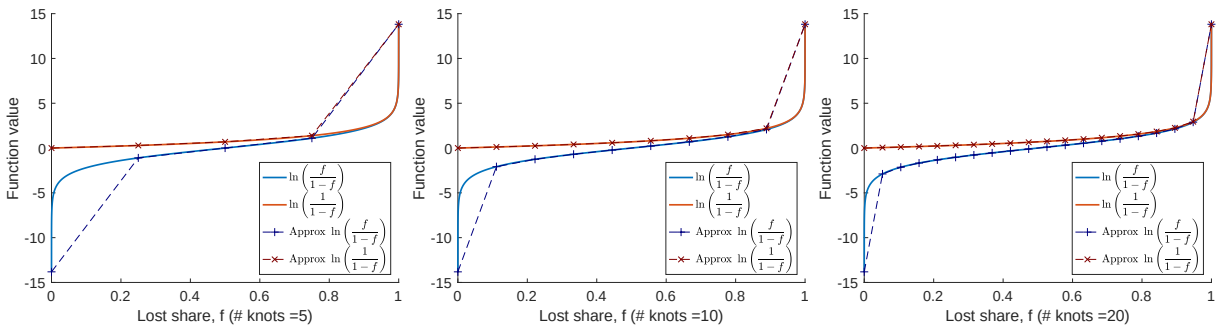


Figure 1 **True and approximate value of the projected lost sales and market size per unit of observed sales as a function of the lost share, f, for varying number of knots uniformly chosen in** $[10^{-6}, 1-10^{-6}]$.

Instead if we adopt an L2 loss function, the problem reduces to a mixed-integer convex quadratic program. These formulations can be directly solved by generic optimization software packages like IBM CPLEX to produce a (near-) optimal solution within a user-specified optimality tolerance (e.g., within 1-10% of optimality). For large scale LM instances, we exploit the special structure of this model in Section 3.3 and propose a mathematical programming based nonconvex-convex alternating heuristic (AH) that can be used to quickly obtain good quality (near optimal) solutions.

A key insight in the proposed model is that the lost sales imputations are done endogenously within 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 $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. As these lost sales fractions are modeled as SOS2 variables, the search is practically on the entire range of the PWL projections of lost sales, $\ln\left(\frac{f}{1-f}\right)$, and 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 discrete sampling points (also commonly referred to as knots) of the PWL segments uniformly chosen between $[10^{-6}, 1-10^{-6}]$. As the number of knots increases, we obtain better approximations. The slope of these functions rapidly increases when the lost shares get closer to 0 or 1 and is relatively invariant in the mid range. In Section 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 (e.g., log-spaced, i.e., uniform in the range of $\ln\left(\frac{f}{1-f}\right)$ instead of uniform-spacing in f).

We conclude the model discussion with two remarks on the impact of problem structure on the difficulty of optimization. Our experiments in Section 4 shed more light on these observations.

1. For $|M|$ choices, there are $|M|$ share terms and one size term per observation in the objective. As $|M|$ increases, more share information per observation is available. As a result, the single purchase choice case, $|M| = 1$, can be relatively harder to solve to provable optimality, resulting

in longer runtimes. Note that besides our method, EM is the only method in the literature that addresses this case and we observe that EM also appears to face a similar challenge.

2. A constant mean arrival rate has been the popular setting studied in the literature, wherein all the observed sales variations are attributed to the changes in the purchase probability. The setting when observed sales variations have to be disambiguated into its size and share components, increases the problem difficulty, and necessitates a joint non-convex optimization of share and size models.

3.2. Consistency of the proposed estimator

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 bin tends to infinity. We use the notation N to denote the number of unit time intervals present within any single aggregate time bin t (in other words, replications) 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 an arrival rate as opposed to market size. For finite data, we conveniently set $N=1$ to identify market size.

In the censored data setting, a joint optimization of the two types of loss terms is required. For example, in the single purchase choice case, for any feasible value of the purchase probability 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

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

Note that when $N \to \infty$, the unobserved arrivals in each time bin $t \in \mathcal{T}$ tends to infinity. 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 across time bins satisfies the following:*

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

This is a mild and natural necessary assumption on the minimum number of time bins having linearly independent covariates for identifiability of the model given the number of unknowns. 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 $\boldsymbol{\gamma}$:*

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

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

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

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 a unique optimal solution using the chosen covariates and assortments across time bins $\mathcal{T}$ 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 bins and variation of covariates and assortments $(\bar{\mathbf{X}}_t, \bar{\mathbf{Y}}_t, S_t)$ across the time bins, 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 assumption, while simulation experiments show that a small number of time bins (even just l) with a reasonable variation of covariates or assortments suffices to identify the parameters.

For a pure assortment problem with no covariates for both choice and arrival rate, the assumption implies every choice is offered at least once and there is at least one pair of time bins 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 $t_1 \neq t_2 \in \mathcal{T}$. in the presence of 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.1. 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. 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 depends on the integrated optimization of the share and size loss terms 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 bins.

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 in the uncensored data settings (Ben-Akiva and Lerman 1985).

Corollary 1. *If the number of feature vectors are finite then the parameters estimated with the LM-C model, after aggregation of observed sales across identical feature vectors and assortments, are consistent with increasing observations, i.e., as $|\mathcal{T}| \to \infty$.*

If $|\mathcal{T}| \to \infty$, aggregating observed sales with identical feature vectors and assortments (which is finite for a finite M) results in increasing arrivals rates, and applying Theorem 1 trivially proves the result. Consistency of MLE-based methods with EM or MM has been established (Abdallah and Vulcano 2017) for the case of varying assortments when products have fixed utilities. The corollary generalizes this result to allow a finite number of possible size and share covariate vectors. The most general case of infinite possibilities of feature vectors is left for future research, along with other statistical properties of the LM estimators such as asymptotic normality and efficiency.

Remark 1. *Assume that the true no purchase probability in any time bin 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 bin 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 Section 4.

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 syndicated data sources and can be included as constraints in the LM model. 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 $60 \pm 5\%$ in a particular month. This can be used to impose a constraint that says $35 \leq \sum_{k \in K, t \in \mathcal{T}_{month}} \frac{f_k z_{kt}}{|\mathcal{T}_{month}|} \leq 45$. Prior information can also be used to optimize the placement and number of PWL knots or prune the lost share range.
2. *Model selection, regularization, sign-constraints, prior values and weighted optimization*: Constraints on prediction parameters (e.g., negative price coefficients) are useful in an automated demand estimation environment and easy to incorporate within an MIP along with desired prior values. 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 (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:* For the MCI and linear attraction demand models, market share terms in the objective of the LM model can be replaced as follows:

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

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

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 $\mathcal{A}_m(\mathbf{X}_{mt})$ of the form $\mathcal{A}_m(\boldsymbol{\beta}_m^T \mathbf{X}_{mt})$.

2. *Non-substitutive and positively correlated demands:* Consider the case where demand for an item is positively correlated with that of other items in the assortment (e.g., different sizes of a Tshirt). Customers buy their preferred size if it is available or walk away otherwise. This scenario can be represented using the following demand model using covariates for the arrival rate alone:

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

The purchase probabilities comprise the size profile and its accurate estimation is a critical requirement in short lifecycle supply chains (e.g. fast fashion, consumer electronics). Stock-outs are common here and result in censored historical data. Eqs. (3.5–3.6) can be appropriately modified to derive a formulation here that is similar to the LM model.

3. *A practical approach for large scale instances: nonconvex-convex alternating heuristic*: In the absence of partial information, an out-of-the-box MIP solver approach may not be practical for large problem instances (e.g., thousands of observations and SOS2 variable sets). To manage such cases, we exploit the special marketsize and market share structure of the LM model and propose a *nonconvex-convex alternating heuristic* (AH). The feasible solution generated by AH can provide an advance start to the exact MIP approach. We show, based on extensive computational experiments in Section 4, that this AH method can be used to quickly obtain provably good quality solutions to the LM model. The details of the AH are described in Procedure 1.

In step 1, all the variables are initialized. When $|M| > 1$ then the relative shares between the observable choices and a default choice d are initialized with model PP. In step 2, the market size coefficients $\boldsymbol{\gamma}$ are estimated jointly with the choice intercepts β_m^0, fixing $\boldsymbol{\beta}_m^{-0}$ and restricting $\beta_m^0 = \beta_d^0 + \Delta_{md}\ \forall m \in M$. The AH-LM model is obtained from the resultant LM model by replacing the observation-specific SOS-2 variable sets with just one SOS-2 variable set, by evaluating the resultant purchase probabilities for every observation for the various discretized feasible values of β_d^0. Consequently, the non-convex AH-LM model can be quickly solved to global optimality even for large instances with a finely discretized β_d^0 (e.g., 100 knots). In step 3, given the market size parameters, all the market share parameters are jointly estimated with model PP-L. For the special case of a constant mean arrival rate, this step can be iteratively invoked with an updated mean arrival rate. The convergence criterion is based on the improvement in the LM objective function and the number of iterations. Models PP-L and PP are simple LPs or convex QPs for L1 and L2 loss functions respectively.

A fundamental difference between the AH and EM is the novel step 2 update using the AH-LM model. In AH, the $\boldsymbol{\gamma}$ covariates are jointly estimated along with the marketshare intercepts

Procedure 1 Nonconvex-Convex Alternating Heuristic (AH)

Input: Observations $Y_t, \mathbf{X}_{mt}, \bar{s}_{mt} \forall m \in S_t, t \in \mathcal{T}$; LM objective tolerance τ; Iteration limit $I_{\max}$.

Notation: We denote the intercept and the non-intercept terms of the purchase probability coefficients $\boldsymbol{\beta}_m$ by β_m^0 and $\boldsymbol{\beta}_m^{-0}$ respectively. For a default choice $d \in M$, let $\Delta_{md} = \beta_m^0 - \beta_d^0$, i.e., the difference in the intercept terms. Let $Z(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{f}(\boldsymbol{\beta}))$ denote the LM objective evaluated using input parameter values $\boldsymbol{\beta}, \boldsymbol{\gamma}$, and purchase probabilities $f_t(\boldsymbol{\beta}) = \left(1 + \sum_{m \in M} e^{\boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt}}\right)^{-1}$.

1: **Initialization:** Choose $d \in M$. Set $i = 0$, $\boldsymbol{\gamma}^i = \mathbf{0}$, $\Delta_{md}^i = 0$, $\boldsymbol{\beta}_m^i = \mathbf{0}$. If $|M| > 1$, update $\left(\Delta_{md}^i, \boldsymbol{\beta}_m^{-0,i}\right)$ by solving the following purchase-probability (PP) model:

$$\min_{\Delta_{md}, \boldsymbol{\beta}_m^{-0}} \sum_{t \in \mathcal{T}} \sum_{m \in S_t} \mathcal{L}\left[\ln(\bar{s}_{mt}) - \ln(\bar{s}_{dt}) - \Delta_{md} - \boldsymbol{\beta}_m^{-0,T} \bar{\mathbf{X}}_{mt}^{-0} + \boldsymbol{\beta}_d^{-0,T} \bar{\mathbf{X}}_{dt}^{-0}\right] \tag{PP}$$

Set $Z^{*i} = Z\left(\boldsymbol{\beta}^i, \boldsymbol{\gamma}^i, \mathbf{f}^i(\boldsymbol{\beta}^i)\right)$.

2: **Estimate market-size coefficients $\boldsymbol{\gamma}^{i+1}$ and intercepts $\beta_m^{0,i}$ given $\left(\Delta_{md}^i, \boldsymbol{\beta}_m^{-0,i}\right)$:** Compute partial attractions $R_{mt}^i = \Delta_{md}^i + \boldsymbol{\beta}_m^{-0,i} \mathbf{X}_{mt}^{-0}$. Use $\epsilon \leq (1 + e^{\beta_d^0} \sum_{m \in M} e^{R_{mt}^i})^{-1} \leq 1 - \epsilon$ to obtain bounds $\beta_{d,\min}^0$, $\beta_{d,\max}^0$ on β_d^0. Let $\bar{\beta}_{dl}^{0,i}$ for $l \in L$ be discretely sampled values of β_d^0 in $[\beta_{d,\min}^0, \beta_{d,\max}^0]$. For each $l \in L$, set lost share at $f_{lt}^i = (1 + e^{\bar{\beta}_{dl}^{0,i}} \sum_{m \in M} e^{R_{mt}^i})^{-1}$, market size at $\bar{\theta}_{lt}^i = \frac{1}{1 - f_{lt}^i} \sum_{m \in S_t} \bar{s}_{mt}$ and lost sales at $\bar{\lambda}_{lt}^i = f_{lt}^i \bar{\theta}_{lt}^i$. Solve the restricted LM model denoted by AH-LM, assuming $\beta_m^0 = \beta_d^0 + \Delta_{md}$, to estimate $\boldsymbol{\gamma}^{i+1}$ and y:

$$\min_{y, \boldsymbol{\gamma}} \sum_{t \in \mathcal{T}} \sum_{m \in S_t} \mathcal{L}\left[\ln(\bar{s}_{mt}) - \sum_{l \in L} \ln(\bar{\lambda}_{lt}^i) y_l - \bar{\beta}_{dl}^{0,i} y_l - R_{mt}^i\right] + \sum_{t \in \mathcal{T}} \mathcal{L}\left[\sum_{l \in L} \ln(\bar{\theta}_{lt}^i) y_l - \boldsymbol{\gamma}^T \bar{\mathbf{Y}}_t\right] \tag{AH-LM}$$

$$\sum_{l \in L} y_l = 1 \tag{3.17}$$

$$y_l \geq 0 \qquad \forall\, l \in L \tag{3.18}$$

$$\{y_l \;\forall\; l \in L\} \in \text{ SOS2}. \tag{3.19}$$

Set $\tilde{\boldsymbol{\beta}}^{i+1} = \left[\sum_{l \in L} \bar{\beta}_{dl}^{0,i} y_l + \Delta_{md}^i, \boldsymbol{\beta}_m^{-0,i}\right]$ and $Z_1^{i+1} = Z\left(\tilde{\boldsymbol{\beta}}^{i+1}, \boldsymbol{\gamma}^{i+1}, \mathbf{f}\left(\tilde{\boldsymbol{\beta}}^{i+1}\right)\right)$.

3: **Estimate corresponding purchase probability coefficients $\boldsymbol{\beta}_m^{i+1}$ given $\boldsymbol{\gamma}^{i+1}$:** Compute $\bar{\theta}_t^{i+1} = \min\left\{\frac{Sales_t}{\epsilon}, \max\left\{\frac{Sales_t}{1-\epsilon}, \boldsymbol{\gamma}^{i+1,T} Y_t\right\}\right\}$ $\forall t \in \mathcal{T}$ where $Sales_t = \sum_{m \in S_t} \bar{s}_{mt}$. Set $\bar{\lambda}_t^{i+1} = \bar{\theta}_t^{i+1} - Sales_t$. Obtain $\boldsymbol{\beta}_m^{i+1}$ $\forall m \in M$ by solving model PP-L given below:

$$\min_{\boldsymbol{\beta}_m} \sum_{t \in \mathcal{T}} \sum_{m \in S_t} \mathcal{L}\left[\ln(\bar{s}_{mt}) - \ln(\bar{\lambda}_t^{i+1}) - \boldsymbol{\beta}_m^T \bar{\mathbf{X}}_{mt}\right] \tag{PP-L}$$

Set $Z_2^{i+1} = Z\left(\boldsymbol{\beta}^{i+1}, \boldsymbol{\gamma}^{i+1}, \mathbf{f}\left(\boldsymbol{\beta}^{i+1}\right)\right)$.

4: **Stopping criteria:** Compute $Z^{*i+1} = \min\{Z^{*i}, Z_1^i, Z_2^i\}$ and denote the corresponding solution by ($\boldsymbol{\gamma}^{*i+1}$, $\boldsymbol{\beta}_m^{*i+1}$). If $|Z^* - \max\{Z_1, Z_2\}| < \tau Z^*$ or $i > I_{\max}$ then estimate the optimality gap $\hat{\psi}$ of ($\boldsymbol{\gamma}^{*i+1}$, $\boldsymbol{\beta}_m^{*i+1}$ $\forall m \in M$) to the best available lower bound (e.g. obtained from a root node analysis of the LM model) and stop. Else go to Step 2 with $\Delta_{md}^{i+1} = \beta_m^{0,i+1} - \beta_d^{0,i+1}$ and $\boldsymbol{\beta}_m^{-0,i+1}$ and set i to $i+1$.

Output: $\boldsymbol{\gamma}^{*i+1}$, $\boldsymbol{\beta}_m^{*i+1}$ $\forall m \in M$, $\hat{\psi}$.

by solving a non-convex restricted LM model. In contrast, EM relies entirely on convex local optimization updates. As a result, AH's step 2 can overcome some local-optima that existing EM approaches cannot, thereby accelerating the convergence to a good quality stable point without depending on randomized starting points. The computational results in Table 9 in Section 4 comparing AH and EM demonstrate the achieved performance gain. Moreover, existing methods, including EM, are restricted to constant mean arrival rate models unlike the AH. As we will see in Section 4, introducing additional arrival rate covariates is computationally more challenging. Our non-convex joint optimization approach is important for disambiguating observed sales fluctuations into its unobserved win rate, and arrival rate variation components. Lastly, operating on the underlying LM model, AH provides an optimality gap upon termination unlike EM. AH can provide warm starts to the MIP solver and can additionally be invoked within an MIP solver to periodically generate feasible solutions by initializing step 1 (of the AH) using the beta-covariate values from an incumbent MIP lower bound (LP relaxation).

4. Computational experiments with simulated and real data

Our computational experiments aim to highlight the modeling and estimation capability of our proposed method on simulated and real data so as to provide guidelines on using and solving the LM model in practice. We present the following experiments: (i) empirical unbiasedness, (ii) consistency of the estimation procedure with respect to the number of arrivals and the number of observations, (iii) the sensitivity bias of the LM model to PWL segments and (iv) to optimality tolerance, (v) the achieved performance when arrival covariates are incorporated, (vi) a performance analysis as the data becomes truly large-scale, and (vii) close with an experiment using real-data to highlight the ability of the LM model to incorporate partial information. In some of these experiments, for benchmarking purposes, we provide comparison on the performance of our method against EM.

Table 3 provides a consolidated view of the settings for each of the experiments. The key LM levers or hyperparameters include the lost share range (specified with ϵ), the number and spacing of the PWL knots and the optimality tolerance of the MIP solver. We simulate data in many of the experiments to analyze the performance of the proposed method in a controlled environment. Experiments with simulated data employ an MNL choice model, and an exponential arrival rate model, along with varying choice and arrival rate attributes that feed their respective models. We generate a probabilisitic (Poisson-distributed unless stated otherwise) number of arrivals based on the mean size, and the purchase choice of each arrival is simulated using the MNL choice probabilities. The resultant average sales rate per product for each data set is also included in Table 3 to describe the portion of the arrivals that are actually observed. As needed, additional experiment-specific details are provided.

Setting #	1	2	3	4	5	6	7
Hyperparameters	Unbiasedness	Consistency	PWL segments	Optimality Tolerance	Arrival Covariates*	Large Instances	Partial Information
Lost share ϵ	10^{-3}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-3}	10^{-3}
PWL knots	20	20	**3-100**	10	20*	20	10
Opt. Tolerance	1%	1%	1%	**1-40%**	1%	10%	1%
Data & Setting							
# Purchase choices	8	1	1	5	1-10	**1-100**	1
# Share features	3	1	1	5	10-100	**100-100K**	7
# Observations	100	**50-200**	20	50	**500**	**200-50K**	30
# Instances	50	30	30	100	10	10	1
e^{γ_0} (Base arrival rate)	4000	**5-20K**	5000	100-10K	5200	50-1000	**Real data** for 4 products with **arrival trend**
$\boldsymbol{\gamma}_{-0}$	**Given in Tbl. 4**	-	-	-	**2,6**	- or same as $\boldsymbol{\beta}_0$	
$\boldsymbol{\beta}_0$		3	U[0.5,2.5]	U[-1.5, 2]	$\lvert M\rvert$U[-0.5, -0.1]	$\lvert M\rvert$U[-0.55, -0.05]	
$\boldsymbol{\beta}_{-0}$		-0.03	10^{-2}U[-4, -2]	U[-0.025,0]	$(-1)^{i-1}$U[1,2]$\lvert M\rvert^{-1}$	$(-1)^{i-1}$U[0.25,1.25]	
$\mathbf{X}$		U[0,200]	U[0,200]	U[0,200]	U[-3, 3]	U[-0.25,0.25]	
Avg. Product Sales Rate	50	39	1460	12-1170	23-185	5-500	

Table 3 The key hyperparameters, setting and data generation process used in experiments in Section 4. The bold entries highlight a feature or the parameter varied in the experiment. The notation U[.,.] refers to the uniform distribution. *The PWL knots are uniformly spaced except in #6 where it is log-spaced.

In all the experiments, the mean market size (same as arrival rate) and the purchase probability parameters are estimated using the LM model and a L1 loss function as a default choice, either using the MIP solver or the AH based on observed sales data, without any additional partial information on lost share. We solve our MIPs (even those within AH) using the CPLEX solver run on a Windows Notebook PC having eight cores and 16GB RAM. We tabulate the computational times and report on the solution quality in terms of model fit and achieved (out-of-sample, OOS, if specified) accuracy of predicting observed sales and censored lost sales using the weighted mean absolute percentage error (WMAPE) metric that is commonly used by practitioners.

$$WMAPE = \frac{\sum_t |\text{predicted value}(t) - \text{actual value}(t)|}{\sum_t \text{actual value}(t)} * 100 \tag{4.1}$$

We sometimes also report the percentage error in estimating the censored arrival rate and purchase probability parameters. When we have thousands of parameters to estimate, we only confirmed that the estimated coefficients were of the right sign.

1. **Unbiasedness and comparison to EM:** (Setting #1) 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). We use their parameters and assortment simulation strategy to test the performance of the LM model.

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 (time) nested price effect terms common to

Parameter	True Value	No assortment changes: LM Method Estimate	LM Method % Error	EM Method Estimate	EM Method % 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
γ (per day)	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 4 Estimated parameters of LM and EM for one instance.

	No assortment changes: Mean estimate	% Error in mean	CoV	With assortment changes: Mean estimate	% Error in mean	CoV	Assortment & arrival trend: Mean estimate	% Error in mean	CoV
β_{Nobuy}	0	0.0	–	0	0.0	–	0	0.0	–
β_{King1}	5.3413	0.8	0.040	5.3513	1.0	0.047	5.3780	1.5	0.055
β_{King3}	4.3858	0.9	0.046	4.3923	1.1	0.053	4.4138	1.5	0.066
β_{King4}	5.3910	0.8	0.038	5.3923	0.8	0.044	5.4170	1.3	0.053
β_{Queen1}	4.0251	1.0	0.051	4.0267	1.0	0.059	4.0551	1.7	0.073
$\beta_{Special}$	4.2464	0.9	0.048	4.2543	1.1	0.055	4.2688	1.5	0.067
β_{Suite1}	7.6643	0.7	0.031	7.6648	0.7	0.037	7.7021	1.2	0.042
β_{Suite2}	5.2071	0.6	0.042	5.2185	0.8	0.050	5.2428	1.3	0.057
β_{TwoDbl}	4.2648	0.9	0.048	4.2615	0.8	0.058	4.2948	1.6	0.067
β_{price}	−0.01725	0.3	−0.021	−0.01721	0.1	−0.020	−0.01725	0.3	−0.019
$\beta_{price,day\geq 1}$	−0.00362	0.3	−0.064	−0.00364	0.8	−0.038	−0.00365	1.2	−0.038
$\beta_{price,day\geq 14}$	−0.00188	−2.7	−0.029	−0.00195	0.8	−0.020	−0.00194	0.4	−0.048
γ (per day)	40.28	0.7	0.134	39.76	−0.6	0.166	40.35	0.9	0.209
Cum. Time (secs)							0.49	−1.8	0.117
	11.4 secs per instance			34 secs per instance			64 secs per instance		
Cum. # Iterations				1 per instance					

Table 5 Empirical unbiasedness with 50 instances of LM with and without assortment effects.

all rooms associated with 14-day ahead, 7-day ahead, and 1-day ahead purchases, respectively. The simulated historical data is in the form of 28-day booking curves for one year's worth of check-in days. We simulate Poisson arrivals for every booking day, which are aggregated across 100 bins. Every bin is associated with one of the 28 booking days, and a random price vector which is uniformly distributed between their minimum and maximum values as specified in Newman et al. (2014). The resultant average lost share was approximately 90%. Assuming an average of 40 daily transactions, we obtained approximately 40,000 (∼40 x 360 x 28 x 0.1) purchases in all. If stock-outs are simulated, all rooms are available for booking days 1 through 7 and thereafter, a room independently closes with 6% probability and remains closed for the remainder of the booking curve (see Newman et al. 2014), which is then aggregated to the bin associated with that assortment. This results in a room being available for booking 65-75% of the time.

We formulate the LM model, with the no-buy transactions censored, and compare the solution quality of the LM against the EM. We employed the EM algorithm described in Talluri and Van Ryzin (2004) with minor modifications to account for aggregated bins (see Appendix C). For a simulation instance, Table 4 summarizes the true and estimated parameter values and the cumulative run times for the two methods. As 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.

We observe that, for the given model settings, the solution quality of the LM method is better than that achieved by the EM using 20 different starting points. We present the best EM parameter estimates, 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 solve this nonconvex problem and produce a

solution having the same or better quality. Importantly, the MIP approach also returns a certificate of (near-)global optimality, which the earlier multi-start methods are unable to provide.

We additionally analyze the empirical unbiasedness of the LM model by re-running the estimation method multiple times, using a different randomization seed each time to generate different data sets. In Table 5, we report the average parameter estimate and run times for the model with, and without stock-outs. We additionally consider a marketsize model variant with an arrival rate trend: $4000e^{0.5d}$ where d is the booking day, and report its results with simulated stockouts. All parameter estimates are within 3% of the true value, with a coefficient of variation (CoV) of less than 0.21. The average run time per instance is 11.4 seconds without assortment changes. This increases to 34 and 64 seconds respectively with assortment changes, and additionally with arrival covariates. We observe that the mean error and CoV in the stock-outs scenario are slightly higher than those obtained in the uncapacitated setting. This is to be expected because there are fewer loss terms in the objective function due to stock-outs. Note that the run time doubles after adding just a single size covariate. The challenge here is that the LM model has to contend with the simultaneous impact of assortment changes, nested price effects, and an increasing arrival trend. Overall, these findings are encouraging from a practical perspective, particularly, given that we are solving a non-convex optimization problem to near global optimality.

2. **Empirical bias and consistency with arrivals and with observations:** (Setting #2) We first analyze the impact of increasing arrival rate on the quality of estimation of a three-parameter binary choice model with 50 observations. The model parameters produce win rates between 40 to 60% for any instance. 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. We observe that the parameters converge to the true values relatively quickly (no more than 40 arrivals, and within the box after 20 arrivals) with lower biases in the parameter estimates achieved at higher arrival rates. We observe high errors at 10 or lower arrivals because a significant percentage of observations consisted of zero sales, and were discarded, which adversely impacted the estimation quality (see the discussion and the proposed heuristic to manage the zero sales issue in Appendix B). We empirically show asymptotic normality in Appendix D.6. In Fig. 3 we present the model fit WMAPEs for the observed and unobserved quantities. The above results provide an empirical substantiation to the results in Theorem 1 and Remark 1.

Using the same binary choice model, we next analyze the impact of increasing the number of observations (from 5 to 200) on the quality of estimation assuming a mean arrival rate of 100 with an L2 loss function. The box plot of some of those estimated parameters for the different observations are presented in Fig. 4. We observe that the parameters converge to the true values quickly along

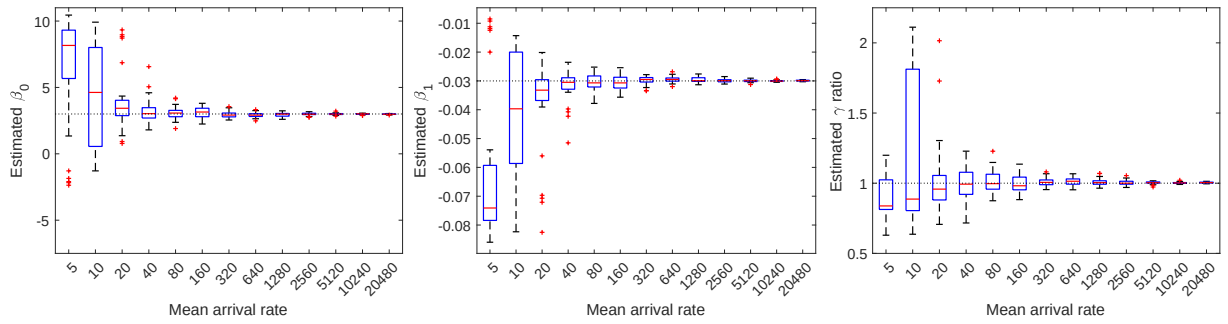


Figure 2 **Box plots of the estimated parameters as a function of increasing mean arrival rate with 50 observations per instance across 30 instances. The true values are marked with dotted lines.**

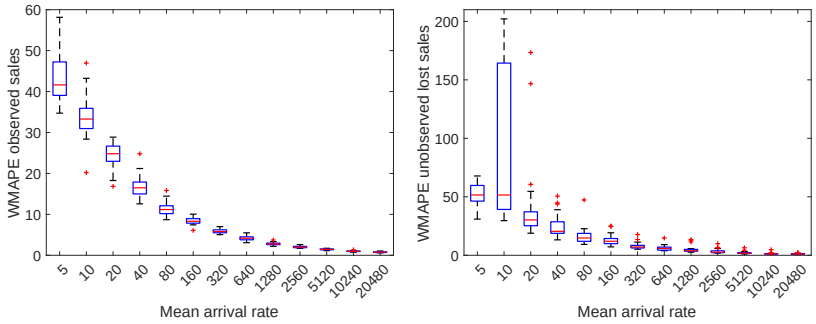


Figure 3 **Average WMPAE of the observed and unobserved sales as a function of the increasing arrival rates.**

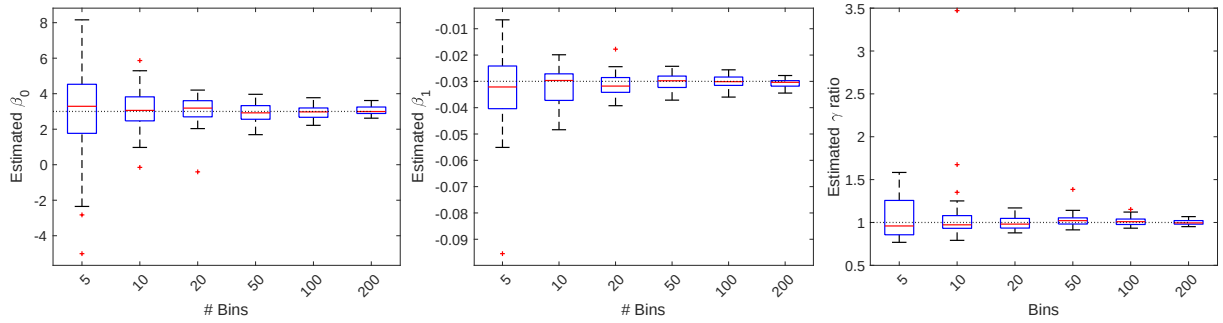


Figure 4 **Box plots of the estimated parameters across 30 instances as a function of increasing the number of observations with a mean arrival rate of 100. The true values are marked with dotted lines.**

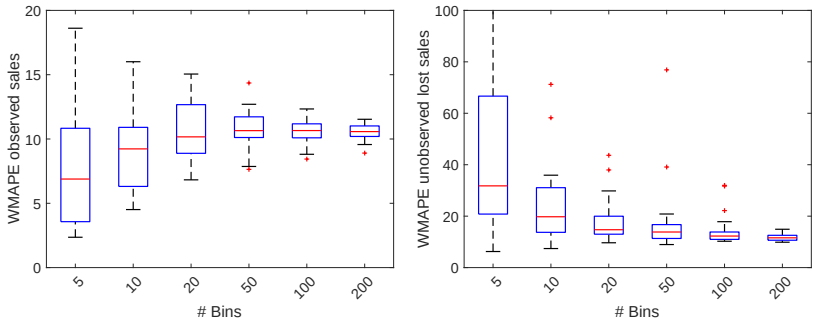


Figure 5 **Average WMPAE of the observed and unobserved sales with increasing number of observations.**

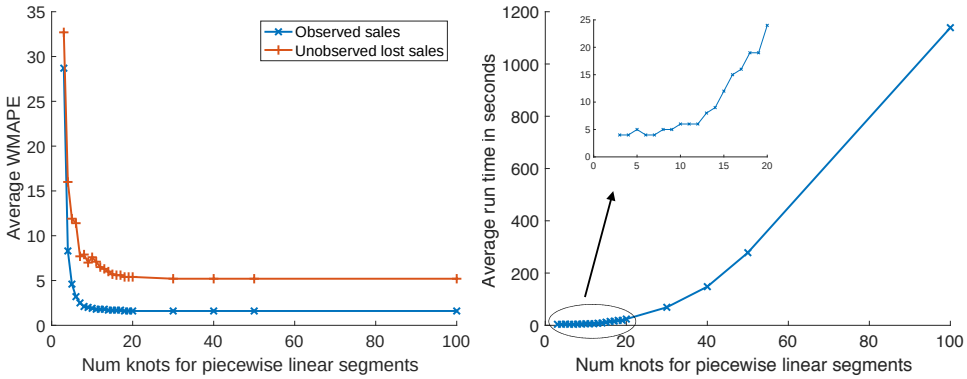


Figure 6 **Average WMAPEs and run times as a function of the number of piecewise linear lost sales segments.**

with a reduction in sample variance as the number of observations increase. The convergence is quicker with a higher mean arrival rate. Fig. 5 tracks the WMAPEs for observed sales and unobserved lost sales. These absolute deviation-based errors converge to a nonzero value depending on the prevailing sales and lost sales rates, respectively. Interestingly, with more observations, the overfitting with observed sales is reduced at the cost of a better fit with unobserved lost sales. The above result empirically substantiates Corollary 1 of Theorem 1.

3. **Model sensitivity to the granularity of piecewise-linearization:** (Setting #3) In this experiment, we report on the impact of varying the number of uniformly spaced PWL knots on the solution quality and run times. Fig. 6 presents the average in-sample model fit WMAPEs of the observed and unobserved quantities, as well as the average run time per instance as a function of the number of knots. The key observation here is that the model performance significantly improves with an initial increase in the number of PWL segments, while a further increase yields diminishing returns. There is no discernible improvement beyond 20 knots while the run times grow exponentially. *Employing fewer, judiciously located knots (e.g., even log-spaced) further reduces run times in practice without sacrificing solution quality.*

4. **Model sensitivity to the optimality tolerance setting:** (Setting #4) In this experiment, we analyze the model sensitivity to the user-tunable 'MIP optimality tolerance' hyper-parameter. We can balance LM run time and solution quality in practice by choosing an appropriate value for this hyper-parameter that specifies an acceptable worst-case relative gap between the best upper bound (feasible objective value) and lower bound (best continuous relaxation objective). We test three arrival rate settings (10, 100, 1000 arrivals per bin), which for the sake of comparison, we denote as sparse, medium, and dense, respectively. We model 5 purchasing choices having an average win rate of 30%, 10% and 4-6% for the remaining 3 choices.

Table 6 summarizes the results. We observe that a tighter optimality tolerance leads to an improvement in predictive performance but yields diminishing returns beyond a 5% (near-optimal)

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 6 Impact of optimality tolerance on the average WMAPEs and run times

range. Although the sparser instances converge faster, the run times for the dense data are lower for a given performance level. As we move from a 5% to a 1% tolerance, the run times increases by a factor of 2-10 for dense to medium instances without really improving solution quality, suggesting that the CPU time is largely expended on generating a certificate of optimality. Solving to near-optimality (say 5-10%) significantly improves upon the unobserved lost sales estimate compared to observed sales. *While a tighter optimality tolerance may not visibly improve WMAPE on a holdout sample, it can result in better model identification and longer-term performance within forecasting applications.* We refer the reader to Appendix C to understand the sensitivity of the model to different loss functions (L1, L2) for the same setting discussed here.

5. **Estimating arrival rate covariates with LM model with the MIP and the AH solution approaches:** (Setting #5) In this experiment, we stress test the LM model by introducing additional arrival rate covariates. The challenge here is to accurately disambiguate the observed sales fluctuations into its market size and market share variation components, all of which vary over time now due to their respective features. We increase the number of observations, the number of choices, and features compared to the previous experiments in order to compare the performance of the MIP approach against the AH.

We assume a mean arrival rate that has a product-life cycle (PLC) form: $\log \theta_t = \gamma_o + \gamma_1 Y_t^1 + \gamma_2 Y_t^2$ where $Y_t^1 = \log\left(0.1 + \frac{w_t}{w_{\max}}\right)$ and $Y_t^2 = \log\left(1.25 - \frac{w_t}{w_{\max}}\right)$ are time cyclical functions that represent the completed and remaining proportion of the current PLC cycle, respectively. Here $w_{\max}$ is the known PLC or season length and is fixed at 52, while $\frac{w_t}{w_{\max}}$ is the proportion of time elapsed within a cycle. The high variation in arrival rates (0-650) due to the PLC function produces very low rates for both the observed sales and unobserved lost sales in many bins, which further increases the degree of difficulty.

We solve the resulting LM instances, each of which contain about 10 cycles of training data, using the exact MIP and the AH, and report the OOS prediction quality achieved. For one data instance, Fig. 7 depicts how the MIP method disambiguates the observed sales variations for a product to estimate the PLC arrival variation and predict the censored lost sales and demand

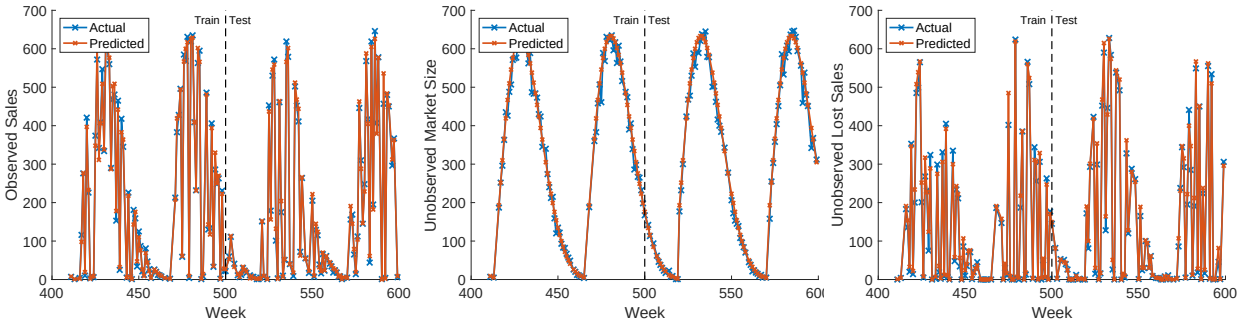


Figure 7 **Predicted and actual observed sales, unobserved market size, and unobserved lost sales obtained using a MIP solver in a single purchase choice setting for a time-cyclical arrival rate model.**

		LM with (cold start) MIP				LM with AH			
$\|M\|$	# Choice Features	Avg run time (seconds)	OOS sales WMAPE %	OOS lost sales WMAPE %	% Error in arrival rate	Avg run time (seconds)	OOS sales WMAPE %	OOS lost sales WMAPE %	% Error in arrival rate
1	10	44	8.7	17.6	5	114	18	37.7	8
10	100	92	24.1	10.4	1	15	24.2	12.1	1

Table 7 **Comparing the performance of two solution methods of LM model in a setting with arrival covariates.**

components. For two settings (1 and 10 purchase choices), Table 7 compares the run time, and the OOS WMAPE averaged over 10 random data instances between the MIP and AH. We observe that both approaches solve the LM instances accurately enough to disambiguate the size from the share variations with a maximum error in the size parameters (not shown in table) being 8% and 5% respectively. The AH is faster for the multi-choice setting but requires more iterations to converge for the single purchase choice and hence slower than the MIP. We find that the MIP performance is better when there is sufficient variation in the choice covariates. When this variation is limited, a tighter MIP optimality tolerance setting is required to achieve comparable performance whereas the AH does better here as it exploits the increased separability in the size and share estimators.

We provide more illustrative examples of mean market size trends that can be modeled using LM without any distributional assumptions on the arrivals in Fig. 11 in the Appendix. We refer the reader to Harsha et al. (2019) for omnichannel pricing applications run on censored sales data from a Fortune 500 company. For hundreds of thousands of products, the LM models were optimized on a weekly basis using the exact MIP approach to jointly estimate the channel switching behavior of consumers with factors such as channel prices and promotions and whose arrival rates varied not only due to a PLC effect, but also due to seasonality and holiday traffic.

6. **Computational experiments for large-scale instances**: (Setting #6) We analyze the performance of the AH (sometimes followed by the MIP) in the following censored data scenarios:

(a) Big-data setting and regularization: Up to 100,000 noisy features and up to 1000 observations.

(b) Other large scale instances: Up to 5000 active features and up to 50,000 observations.

(c) Comparing against EM: Up to 1000 active features and 1000 observations

Our goal is to obtain a good prediction model within a limited computation time that scores well on OOS data (100 observations) and recovers a good estimate of the hidden arrival rate. Here, we use CPLEX's Barrier solver rather than the default Simplex for solving the underlying LPs, which results in an order of magnitude improvement in solver performance. The average lost sales rate in the training data set as a result of the parameters chosen varied between 20 to 80 percent and every product contributed to at least 0.1% of the total observed sales.

(a) **Big Data Instances:** In the first series of tests, we simulate censored data instances having up to 100,000 features, and up to 100 purchase choices. Between 0.5% to 10% of these features are active and generated from a uniform distribution, while the remaining features are Normal(0,0.35/# noisy features) random variables having a fixed feature coefficient value of 1.0. The number of training observations ($|T|$) are thousand or fewer, and a L1-penalty of the form $\alpha_L|\mathcal{T}|$ is applied on the parameter value to achieve sparsity, where the coefficient $\alpha_L = 10^{-2}$. As depicted in Fig. 8, the lost sales WMAPE is relatively less sensitive to the L1-penalty. Therefore, the best α_L value can be chosen based on the observed sales WMAPE itself.

Rows 1-4 in Table 8 summarizes the performance of the AH-initialized MIP solution approach for select big-data scenarios. The entire table of results and additional discussion is available in Appendix D.2. In all but the largest scenario, we obtain provable near-optimal solutions. We predict the OOS lost sales within 12% WMAPE, and estimate the hidden arrival rate with 2.6% of the true value for all scenarios. The average OOS observed sales WMAPE across all problem sizes is 13.5% (and is relatively higher because it computed at product level). The achieved WMAPE values can be further tuned down by varying the L1 penalty. Overall, the AH method converged within 30 minutes on average for all but the largest instance (80 minutes) while we obtain provable near-optimal solutions within three hours for the largest MIP-instance. We observe a superlinear increase in MIP run time for the larger instances when the number of observations or the total (but not the active proportion) features are increased. A disproportionate fraction of the MIP run time is expended on improving the lower bound and proving solution quality (e.g. 87% of the run time for scenario 8, comparing rows 2 and 3), a behavior which we previously remarked on in the optimality-tolerance sensitivity experiment.

The 'buy vs. no-buy' instances are particularly tough to solve to near-optimality since a good quality initial solution for the choice covariates via AH is unavailable as there are no cross-product equations available. In contrast, the multiple purchase choice instances, although more time consuming due to a larger number of variables, are solved to provable near-optimality.

(b) **Empirical performance on other large scale instances:** Next we test censored data instances having a large number of active features (up to 5000), and up to 50,000 observations and

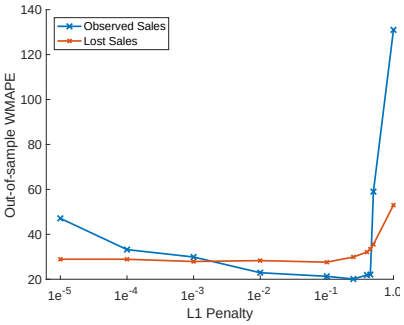


Figure 8 **Illustrative plot for tuning L1 regularization penalty**

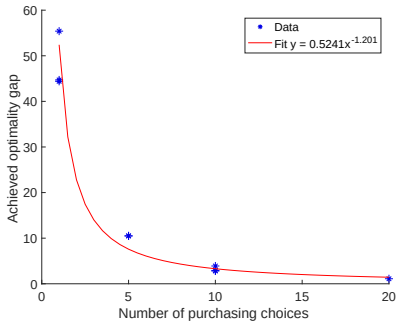


Figure 9 **Achieved optimality gap of AH for scenarios in Table 9; The R^2 of the fit is 0.976.**

	Method	Scenario	$\|M\|$	# Features	# Active features	# Observations	Arrival Rate	Avg sales rate per product	Avg run time (minutes)	OOS sales WMAPE %	OOS lost sales WMAPE %	% Error in arrival rate
Big-data	AH+LM with L1	3	1	1000	100	400	1000	491	0.7	8	7.7	0.3
		8	10	10000	1000	1000	1000	78	157	11	10	1.6
	AH with L1	8	10	10000	1000	1000	1000	78	14	12	9.1	1.8
		10	100	100000	1000	400	1000	6.8	79	26	11	2.6
Large-scale	AH	4	10	1000	1000	10000	100	4.8	26	29	15	6.0
		8	50	5000	5000	1000	1000	12	7.7	30	6.7	2.1
		11	100	5000	5000	1000	1000	6.1	87	35	7.6	2.7
		12	5	400	400	50000	50	6.3	55	30	34	9
		13	100	200	200	1000	$500e^{\boldsymbol{\beta}_{-0}\mathbf{X}_{-0}}$	64	0.9	14.1	6	1.7

Table 8 **Results of sample scenarios of Big-data and Large-scale experiments. See Tables 14, 15 in Appendix for complete results. In scenario 13, $\boldsymbol{\gamma}_{-0}, \mathbf{Y}_{-0}$ are generated from the same distribution as $\boldsymbol{\beta}_{-0}, \mathbf{X}_{-0}$ respectively.**

100 product choices with the AH method. We run one scenario using 100 arrival covariates. A brief summary is presented here for a sample of results (rows 5-9 in Table 8). The complete results are reported in Appendix D.2. We obtain good quality (within 15% of optimality obtained up on warm starting the MIP solver) parameter estimates within 87 minutes of average training time for the most difficult instance. We also observe that the problem with arrival covariates is solved within a minute on average by AH. The average OOS lost sales WMAPE achieved is less than 15% for all problem sizes, except for scenario 12 where it is 34% because of the relatively lower rate of lost sales, and the estimated mean arrival rate is within 9% of the true value. We also obtain reasonable observed sales WMAPE.

We test multiple scenarios that exhibit a low rate of sales by applying the heuristic method described in Appendix B to the training data with zero-sales (scenarios 4 and 11 have more than 94% observations affected by zero sales). All adjusted observations are retained for training. While we observe a moderate degradation in solution quality, the proposed method on average performs well. Appendix B also reports results on the quality of the parameter estimates for a publicly available Amazon DVD model that results in 99.5% of the 5000 observations impacted by zero sales.

				Avg sales	Concurrent 8-multistart EM				AH			
Scenario	$\|M\|$	# Features	# Observations	rate per product	Avg run time (seconds)	OOS sales WMAPE %	OOS lost sales WMAPE %	% Error in arrival rate	Avg run time (seconds)	OOS sales WMAPE %	OOS lost sales WMAPE %	% Error in arrival rate
1	1	100	500	487	107	34	24	13	18	13	4.9	6.9
2	1	100	1000	401	168	41	44	-13	35	4.3	10	5.9
3*	1	100	500	40	60	18	28	-17	16	15	16	-3.8
4*	5	100	500	7.0	26	30	169	109	2	30	25	-7.0
5	5	100	500	73	66	10	127	83	2	10	3.9	-0.5
6	10	100	500	38	225	15	43	23	3	15	9.0	0.7
7	10	200	500	41	172	18	60	36	4	14	6.8	-1.3
8*	10	50	500	4.4	36	42	91	-57	6	37	26	-0.3
9	20	400	500	40	82	22	53	-6.6	9	22	6.6	2.2
10†	20	1000	500	43	636	34	41	-28	15	17	7.5	-0.1

Table 9 Comparison of the average results between EM and the AH across difference scenarios with 10 instances each (†1 instance). The arrival rate is 1000 except for scenarios marked with * where it is 100.

(c) **Comparing against EM:** Here we compare the solution quality and run-time between the AH and the EM method enhanced to handle larger data sets (see Appendix C) using simulated data. As EM remained time consuming due to the expensive MLE gradient computations and poor tail-end convergence, we limit the analysis to random instances having up to 100 product choices, and up to 1000 choice coefficients. The results for the specific scenarios are shown in Table 9. We report the best EM solution, based on OOS observed sales WMAPE, obtained from the 8 different starting points. EM usually converges to a stationary point within 20-100 MLE iterations and produces reasonable observed sales forecasts but yields a poor lost sales estimate (68% avg WMAPE) and arrival rate predictions (39% absolute error on avg). We see that in each of the problem instances, the AH OOS lost sales and arrival rate prediction dominates EM. As far as the average observed sales WMAPE, AH is strictly better in six instances and on par for the other four, while being 4-66 times faster. Fig. 9 plots the achieved (worst-case) optimality gap of the AH warm started MIP for the scenarios in Table 9. We observe that the achieved gap improves as the number of product choices increase, with scenario 9 yielding AH solutions within 1.1% of global optimality.

Overall, AH dominates EM, achieving better overall solution quality while consuming only a fraction of the EM run time. This wide gap can be attributed to the nonconvex optimization step of AH that enables it to quickly converge to the near-optimal solutions, solving no more than 5 PP-L LPs (up to 20 LPs for single purchase choice instances). In contrast, EM sometimes required more than a 100 MLE iterations to converge to an overly suboptimal stationary point.

7. **LM with partial information to improve win rate estimation using real-life data**: (Setting #7) In this section, we apply the LM model to analyze Request-for-Quote (RFQ) data that arise in a Business-to-Business (B2B) setting for procuring high-end computer hardware products. We have historical RFQ data for four different hardware products recorded over a duration of 30 weeks. Each record shows whether a particular quote was won or lost, and is associated with a

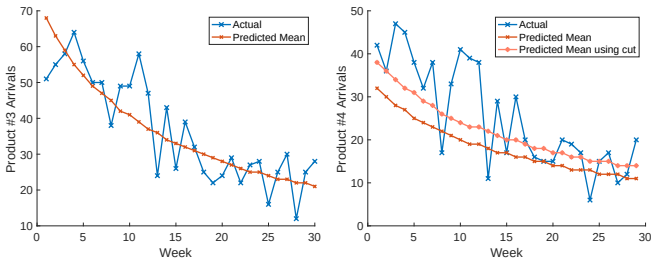


Figure 10 **Weekly actual vs. predicted censored arrivals.**

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 partial info.)	735	28.3	30.4

Table 10 **Average actual and predicted win rate.**

feature vector whose components include the order date (week), normalized price, product quality, manufacturing cost, list price, and other derived features. The weekly total sales across all products reveals a declining trend over the last 30 weeks. We would like to identify the reason for this trend using only the aggregated historical purchase data (losses are assumed to be hidden) in order to predict the weekly arrival rate and average win rate and compare them with the actual values.

Toward this, we aggregate the winning transaction data into weekly sales bins. Doing so captures the aggregate wins and average buyer response to the prevailing product attributes for a week. The LM model with regularization penalty, and a week-index trend as an additional market-size covariate is optimized using settings in #7 in Table 3). The cumulative CPU run time for all products is 129 seconds. In Fig. 10 we plot the mean weekly predicted arrivals against the censored actual values for two of the analyzed products. The LM model correctly identifies a seasonal decline in weekly RFQ arrivals that affects all products, thereby confirming that the observed negative trend in wins is less likely to be due to any adverse trends in component features (quality, price).

We used LM to learn the average win rate for three of the products as shown in Table 10. However, the relatively low weekly sales rate for the fourth product reduces the prediction accuracy. In practice, we are often able to obtain an approximate range for the average market share or "wallet share". To demonstrate the beneficial effect of such 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 is added to the MIP to restrict the average lost share 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 enables us to improve the win rate by 6 percentage points from 36.6% to 30.4% (see Fig. 10).

4.1. Key-take aways and recommendations for real-world implementation

We recommend the use of the exact MIP solver as a default choice for forecasting sales at the daily level and higher when the number of observations are less than a thousand. Any partial information regarding the lost share can be incorporated within the MIP solver to improve performance. We suggest switching to the nonconvex-convex AH to solve larger instances, especially when there is

no partial information to guide and speed up the exact MIP solver. This heuristic can also be invoked within an MIP solver to provide a 'warm start', periodically polish the solution, and obtain a certificate of quality. A practical default setting for the three key hyperparameters are: a lost sales tolerance of 10^{-3} with 20 PWL segments and the optimality tolerance set between 5-10%. We recommend L1 loss function as a default choice, although L2 maybe use useful in low sale rate settings. Tuning these settings can improve performance for difficult instances.

We observe when the number of observations (and hence lost sales imputations) and the number of choice parameters to be estimated are in the hundreds, the run time is in the order of seconds. In large-scale settings where the noisy parameters are up to hundred thousand, given no more than a thousand training observations, the run time is in the order of minutes. Finally, to estimate thousands of choice probability parameter values and perform up to ten thousand lost sales imputations, requires up to three hours to identify near-optimal solutions.

The proposed LM model requires grouped observations and in general, multiple observations with identical covariates can be aggregated into a single bin. When there are frequently changing features over time, the choice of time bins can involve a practical tradeoff between more disaggregate data (more zero sales, variance and observations) and aggregate bins (increased error-in-variables and fewer observations). A brief discussion of this tradeoff for LM and EM methods is available in Appendix E. When zero sales affect a significant portion of the training observations, say 20%, the Poisson heuristic in Appendix B can be employed to adjust the data. We recommend a judicious use of these censored demand estimation models when they are compatible with the application and the quality of the data.

5. 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 purchase probabilities and attractiveness of their competitors by analyzing their own censored historical sales data and competitor product attributes such as competitive assortment, prices and promotions. A heuristic approach would be to incorporate these factors within the LM model and treat the entire lost sales as just one unobserved choice. However, this approach provides no 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 the censored lost sales to competition as an additional substitutive choice. Essentially, this requires the total unobserved sales by the seller to be disambiguated based on the competitor's observed attributes.

In addition to the no-purchase option $\emptyset$, let $M_u \subset M$ be the set of censored competing purchase choices. 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 Section 3, using availability

sets S_t. Let α_{mt} denote 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 $\left(D_{mt} = \left[\sum_{j \in J} \bar{\alpha}_{mjt} w_{mjt}\right] \sum_{k \in K} \bar{\lambda}_{kt} z_{kt}\ \right)$ for imputed sales and substitute $Q_{mt} = \ln D_{mt}$ to obtain constraint (5.2) in the formulation below. We now formalize the full model:

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

$$Q_{mt} = \ln(\bar{s}_{mt}) \qquad \forall\ t \in \mathcal{T}, m \in M \backslash M_u \qquad (5.1)$$

$$Q_{mt} - \sum_{k \in K} \ln(\bar{\lambda}_{kt}) z_{kt} - \sum_{j \in J} \ln(\bar{\alpha}_{mjt}) w_{mjt} = 0 \qquad \forall\ t \in \mathcal{T}, m \in M_u \cup \emptyset \qquad (5.2)$$

$$\sum_{m \in M_u \cup \emptyset} \sum_{j \in J} \bar{\alpha}_{mjt} w_{mjt} = 1 \qquad \forall\ t \in \mathcal{T} \qquad (5.3)$$

$$\sum_{j \in J} w_{mjt} = 1 \qquad \forall\ t \in \mathcal{T}, m \in M_u \cup \emptyset \qquad (5.4)$$

$$\sum_{k \in K} z_{kt} = 1 \qquad \forall\ t \in \mathcal{T} \qquad (5.5)$$

$$\{w_{mjt}\ \forall\ j \in J\} \in \text{SOS2} \qquad \forall\ t \in \mathcal{T}, m \in M_u \cup \emptyset \qquad (5.6)$$

$$\{z_{kt}\ \forall\ k \in K\} \in \text{ SOS2} \qquad \forall\ t \in \mathcal{T} \qquad (5.7)$$

$$w_{mjt}, z_{kt} \geq 0. \qquad (5.8)$$

Disambiguation constraint (5.2) delineates the share of each unobserved choice based on its observed attributes, while constraints (5.3–5.4) 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 estimation. The first layer estimates the total unobserved share, and the second layer further disambiguates the no-purchase from the various competitor purchase options based on the relative variations in their attributes over time. Consequently, this problem is harder to solve compared to the no-competitor case. We observe that relatively more time bins, 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. Yet the continuous version of the Comp-LM model 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 $\boldsymbol{\beta}_{M\backslash M_u}$:*

$$w_t = \sum_{t' \in K_m} \eta_{t'}^m w_{t'} \ \forall t \in \mathcal{T} \backslash K_m, \ m \in M_u. \tag{5.9}$$

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 a feasible solution.

Eq. (5.9) are the resultant limiting non-linear equations for identifying the unknown true parameters $\boldsymbol{\gamma}^*$ 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 bins and variation of covariates and assortments $(\bar{\mathbf{X}}_t, \bar{\mathbf{Y}}_t, S_t)$ across the time bins, Eq. (5.9) 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 satisfied.* Eq. (5.9) highlights the higher degree of complexity of the non-linear equations we have to satisfy and hence the need for additional time bins 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.,*

$$\lim_{N \to \infty} Z^{CC}(N, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{f}^*, \boldsymbol{\alpha}^*) = 0, \tag{5.10}$$

   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.,*

$$\lim_{N \to \infty} Z^{CC}\left(N, \boldsymbol{\beta}^N, \boldsymbol{\gamma}^N, \mathbf{f}^N, \alpha^N\right) = 0. \tag{5.11}$$

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 $\alpha_{mt} \ \forall m \in M_u \cup \emptyset, t \in \mathcal{T}$ using the Comp-LM-C problem are consistent with the true values as $N \to \infty$, i.e.,*

$$\lim_{N \to \infty} \left\langle \boldsymbol{\beta}^N, \boldsymbol{\gamma}^N, \mathbf{f}^N, \alpha^N \right\rangle = \left\langle \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{f}^*, \alpha^* \right\rangle. \tag{5.12}$$

The proof of the above theorem is very similar to Theorem 1 with some differences, in particular, the derivation of the limiting Eq. (5.9), the proof of which is provided in Appendix A.2.

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
γ	40	34.6	−13.5	0.111	39.2	−2.0	0.236
Run time		78 secs			137 secs		

Table 11 **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.**

5.1. Experiments with hidden competitor choice using simulated data

The first experiment analyzes the empirical unbiasedness of the Comp-LM model using the hotel data from experiment #1 in Section 4, wherein additionally, the king4 room's sales history as well as the no-buy data are unavailable to the model. The model settings remain the same (#1 of Table 3) with the exception of the optimality tolerance which is relaxed to 5% for this difficult problem setting. The Comp-LM results for the same instances as the prior experiment are reported in Table 11 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). This increase in run-time and empirical bias is indicative of the challenging nature of the Comp-LM's MIP containing multiple estimation layers. Despite these challenges, the performance of the model was reasonable. The average competitor purchase probability estimated by the Comp-LM model was 4.7% (and 4.2% using 50 PWL knots) compared to the unobserved actual value of 4.1%. A detailed sensitivity analysis of the Comp-LM model behavior and performance to increased arrival rates and finer PWL discretization is presented in Appendix D.4.

In our second experiment, we test for empirical consistency with increasing arrival rate, where additionally, the sales of one of the purchase choices are unobserved, along with the no-purchase option. The results are provided in Appendix D.5 and the inferences are substantially similar to experiment 2 in Section 4, although the rate of convergence is slower than the no-competitor case.

Appendix A: Proof of Theorems in the paper

A.1. 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 an 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 means 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 bins, 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 (5.12) and the theorem. □

A.2. 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.1. 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:

$$\ln\left(\frac{1-\hat{f}_t}{\hat{\alpha}_{\emptyset t}\hat{f}_t}\frac{\alpha^*_{\emptyset t} f^*_t}{1-f^*_t}\right) = \left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_m\right)^T \bar{\mathbf{X}}_{mt}, \qquad \forall\, m \in M\backslash M_u, t \in \mathcal{T}, \tag{A.11}$$

$$\ln\left(\frac{\hat{\alpha}_{mt}}{\hat{\alpha}_{\emptyset t}}\right) = \hat{\boldsymbol{\beta}}_m^T \bar{\mathbf{X}}_{mt}, \qquad \forall\, m \in M_u, t \in \mathcal{T}, \tag{A.12}$$

$$\ln\left(\frac{1-f^*_t}{1-\hat{f}_t}\right) = (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^T \bar{\mathbf{Y}}_t, \qquad \forall\, t \in \mathcal{T}, \text{ and} \tag{A.13}$$

$$\sum_{m \in Mu} \hat{\alpha}_{mt} + \hat{\alpha}_{\emptyset t} = 1, \qquad \forall\, t \in \mathcal{T}. \tag{A.14}$$

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. (A.13) and $\hat{\alpha}_{mt}\ \forall m \in M_u$, $\hat{\alpha}_{\emptyset t}$ using Eqs. (A.12–A.14) in Eq. (A.11), we get,

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_m^T \bar{\mathbf{X}}_{mt} - \ln\left(1 + \sum_{m\in M_u} e^{\hat{\boldsymbol{\beta}}_m^T \bar{\mathbf{X}}_{mt}}\right) = \boldsymbol{\beta}^{*T}_m \bar{\mathbf{X}}_{mt} \\
&- \ln\left[\left(e^{(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T \bar{\mathbf{Y}}_t} - 1\right)\left(\sum_{m\in M} e^{\boldsymbol{\beta}^{*T}_m \bar{\mathbf{X}}_{mt}} + 1\right) + 1 + \sum_{m\in M_u} e^{\boldsymbol{\beta}^{*T}_m \bar{\mathbf{X}}_{mt}}\right] \qquad \forall\, m \in M\backslash M_u, t \in \mathcal{T}.
\end{aligned} \tag{A.15}$$

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. □

Appendix B: Adjusting the LM objective function for zero sales

The LM method uses aggregated censored data to jointly calibrate the MNL model as well as the arrival rate model. When the average rate of sales of one or more product choices is low, there can be several observations that record zero product sales. When such observations are few, they can simply be discarded, but when they comprise a significant fraction of training data, we have to retain these observations by replacing the zero values by a small positive quantity since the LM objective is undefined at zero sales. This is a concern even for the Berkson's method in the uncensored data setting and different zero-sales adjustments have been proposed in the literature (Greene 2011).

Cox (1970) prescribes a weighted objective correction for the uncensored data setting and recommends using a value of 0.5 for zero sales values. We adapted this weighting scheme to our censored data setting but adjusted the observed sales for all purchase choices in an affected training observation from $\bar{s}_{mt}$ to $\bar{s}_{mt} + (1 - p^0_m)$, where p^0_m is the probability that choice m has zero sales in an observation. Assuming a Poisson distribution, this is $e^{-\lambda_m}$, where λ_m is the average observed sales rate for choice m. Using this probabilistic

	β_0	β_{price}	β_{votes}	γ
True parameter value	-4.31	-0.038	0.0000354	50
Mean estimate	-3.88	-0.041	0.0000366	49.8
% Error in mean	9.9	6.7	3.4	-0.3
Coeff. Of variation	-0.070	-0.059	0.018	0.252

Table 12 Empirical unbiasedness of the estimated parameters for the Amazon DVD model over 50 instances with 99.5% of the observations impacted with zero sales.

adjustment better preserves the relative sales strength of products in the affected observations and helps limit the estimated parameter bias.

We tested this heuristic approach using the AH method on simulated data set based on the Amazon DVD model presented in Rusmevichientong et al. (2010) and studied in the censored data context in Abdallah and Vulcano (2017). The data represents the sales of 15 DVD packages each of which has two fixed covariates: average price per disc and a number of useful votes. The underlying nominal utility value for each DVD has the same linear model, including the intercept term. The fixed DVD dependent covariate vector isspecified in (Farias et al. 2013). Using a homogeneous mean Poisson arrival rate of 50 per observation, a total of 5000 observations were generated and their choices to assortment changes (uniform discrete between 4 and 11) recorded. This simulation setting resulted in a low average observed sales rate of 0.26 units per DVD, with more than 99.5% of the observations recording at least one zero sale in the offered assortment of DVDs. The tabulated results of our method over 50 instances are shown in Table 12 with the average run time being 63 seconds per instance. We were able to estimate the MNL parameters to within 10% of their true value and accurately estimated the unobserved arrival rate to within 0.3% of the true value. These estimates are noteworthy since they are obtained without additional information, such as the average total observed market share under the full assortment setting used by Abdallah and Vulcano (2017), and instead relies on the homogeneity assumption of the unobserved arrival rate.

In the other large-scale experiments with synthetic models discussed in Section 4, we also report results for low sales rate settings that uses the above adjustment whenever we have zero sales observations.

Appendix C: EM algorithm implemented in the comparisons

Step 0: Initialization

1. Set $k = 0$. Generate an initial random vector of lost sales $f_t^k \ \forall t \in \mathcal{T}$ and obtain an average arrival rate estimate $\theta^k = \frac{\sum_t (S_t + f_t^k)}{|\mathcal{T}|}$, where S_t = total observed sales across all choices.
2. Use the LBFGS algorithm to run MLE on the market shares to estimate MNL coefficients $\boldsymbol{\beta}^k$.

Step 1: Expectation

1. Compute predicted no-purchase probabilities p_t^k from $\boldsymbol{\beta}^k$ for each observation t and calculate predicted lost sales $f_t^{k+1} = \theta^k p_t^k$.
2. Compute expected market size $\theta^{k+1} = \frac{\sum_t (S_t + f_t^{k+1})}{|\mathcal{T}|}$.
3. Optional acceleration step: Recompute $f_t^{k+1} = \theta^{k+1} p_t^k$.

Step 2: Maximization

1. Given f_t^{k+1}, calculate for every observation t the market shares for each purchase choice and the no purchase option.

2. Use the LBFGS algorithm to run MLE on the market shares, starting from $\boldsymbol{\beta}^k$, to estimate MNL coefficients $\boldsymbol{\beta}^{k+1}$.

Step 3: Convergence check:
If $\frac{\sum_{m\in M, i\in I} |\beta_{mi}^{k+1} - \beta_{mi}^{k}| + |\theta^{k+1} - \theta^{k}|}{(|M||I|+1)} < 0.01$, stop. Else $k = k+1$. Go to Step 1.

This algorithm typically converges within 10-50 MLE iterations, but can consume more than 100 iterations, depending on the starting point. Also, a more stringent convergence tolerance can significantly increase the number of MLE iterations.

For the hotel data set, 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.

Implementation enhancements for large datasets: In experiment (1) in Section 4 we reported the EM performance using an exact Newton-Raphson method (described above). This is impractical for larger instances. We therefore enhanced the EM approach in the following ways:

1. Replaced the Newton-Raphson method with the LBFGS algorithm and warm start capability.
2. Limited the number of MLE calls (one per EM iteration) to a maximum of 1000.
3. Concurrent processing spawn multiple EM instances starting from different initial solutions.

Despite this, the EM approach remained overly time-consuming for larger instances (thousand or more coefficients) as the MLE gradient computations were expensive. So, we had to limit EM testing to some of the smaller problem sizes (and sometimes just one instance).

Appendix D: Additional details on computational experiments in the paper

D.1. Sensitivity of the model to different loss functions (Setting #4)

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 as setting #4 in Table 3. Table 13 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.

D.2. Complete set of results for large-scale instances (Setting #6)

(a) **Big Data Instances:** Table 14 reports the average performance of the AH-initialized MIP solution approach across various simulated big data scenarios for a wide range of problem sizes and features. In all but the largest of these scenarios (marked by * and scenario 4), we obtain provable near-optimal solutions (converges within the optimality tolerance). Scenario 4 terminates upon reaching the preset CPLEX node limit. For the large instances (marked by *), we obtain solutions of similar quality, but are unable to improve the solution and calculate a lower bound using CPLEX-MIP due to computer memory limits.

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

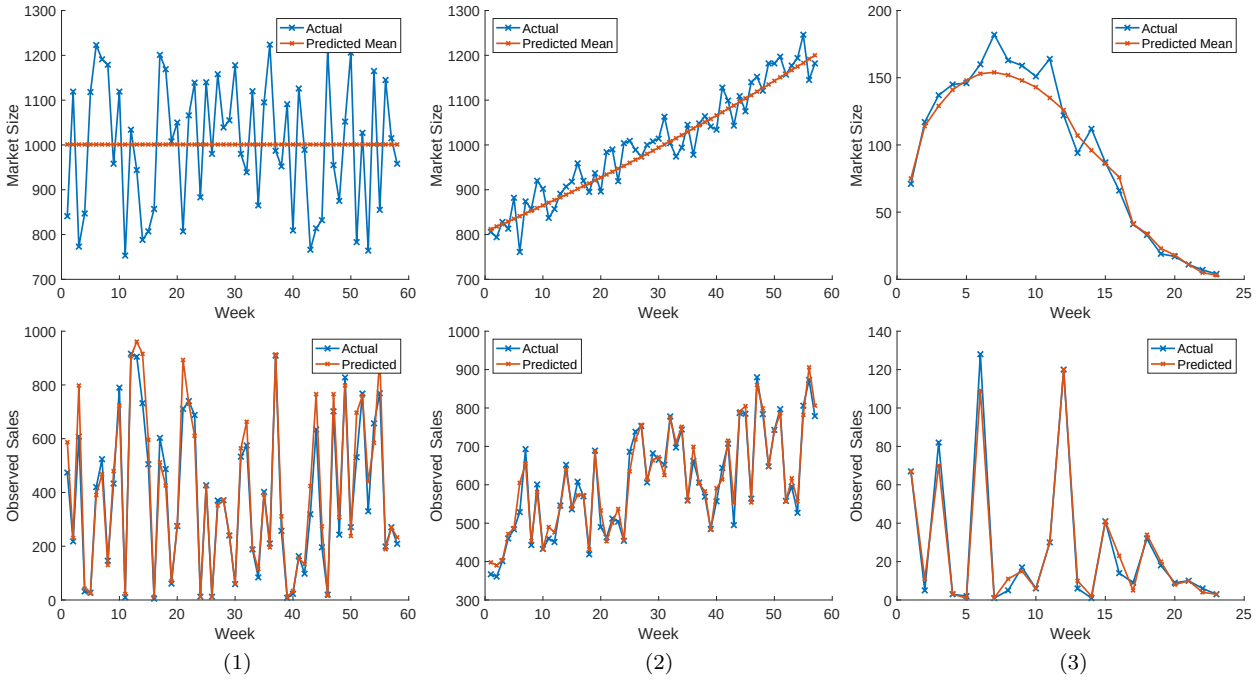


Figure 11 Mean market size and observed sales estimates when the size has the following distribution and mean arrival rate form: (1) uniform and constant, (2) Poisson with a linear trend $\gamma_0 + \gamma_1 Y_t$ and (3) Poisson with a PLC form $\gamma_o (Y_t)^{\gamma_1} (1 - Y_t)^{\gamma_2}$ where $Y_t = \frac{t}{|\mathcal{T}|}$.

Note that the OOS observed sales WMAPE is a product level percentage-based metric that varies with the observed sales rate per product. Consequently, this metric tends to be higher relative to that of the lost sales and the arrival rate, which tend to be of larger magnitudes. The achieved WMAPE values can be further tuned down by varying the L1 penalty.

(b) **Empirical performance on other large scale instances:** The results are reported in Table 15. The prediction results are already described in the main paper. As far as training run times and sensitivity to problem size, the empirical behavior is like that observed in the big data settings. We stress tested the LM method for low rate of sales settings in two ways: (1) by keeping the total arrivals per observation fixed even as we increased the number of product choices and number of observations, and (2) reducing the arrival rate from 1000 to 100 per observation for scenarios 3 and 50 for scenario 12. Doing so reduces the average sales rate per product choice to single digits, which results in a high proportion of observations being impacted by zero sales in the simulated training data (close to 100% in some instances). The heuristic method described

Scenario	$\|M\|$	# Features	# Active features	# Observations	Avg sales rate per product	Avg run time (minutes)	OOS sales WMAPE %	OOS lost sales WMAPE %	% Error in arrival rate
1	1	1000	50	200	440	0.4	4.6	4.4	0.5
2	1	1000	50	400	440	0.9	3.9	4.1	0.7
3	1	1000	100	400	491	0.7	8.0	7.7	0.3
4	1	10000	100	400	487	28	7.1	7.2	0.4
5	10	10000	500	200	80	4.8	12	11	1.9
6	10	10000	500	400	79	26	11	12	2.1
7	10	10000	1000	400	78	35	12	10	1.6
8	10	10000	1000	1000	78	157	11	10	1.6
9	10	100000	500	200	79	227	12	11	1.6
8*	10	10000	1000	1000	78	14	12	9.1	1.8
9*	10	100000	500	200	79	30	14	12	2.2
10*	100	100000	1000	400	6.8	79	26	11	2.6

Table 14 Average results over 10 instances per scenario for the big-data setting using the MIP approach (with L1 regularization and warm started with L1-regularized AH), except scenarios with * where AH (with L1) was used. The mean arrival rate is 1000.

Scenario	$\|M\|$	# Features	# Observations	Arrival rate	Avg sales rate per product	Avg run time (minutes)	OOS sales WMAPE %	OOS lost sales WMAPE %	% Error in arrival rate
1	10	100	1000	1000	83	0.2	6.5	7.0	0.2
2	10	1000	1000	1000	84	1.4	11	7.2	0.1
3	10	1000	5000	100	4.6	4.3	28	14	5.6
4	10	1000	10000	100	4.8	26	29	15	6.0
5	20	1000	1000	1000	12	2.7	30	9.0	4.4
6	20	1000	10000	1000	24	83	21	6.2	2.8
7	40	2000	1000	1000	17	4.4	22	6.6	1.5
8	50	5000	1000	1000	12	7.7	30	6.7	2.1
9	50	5000	2000	1000	8.2	37	33	13	2.1
10	100	5000	500	1000	5.6	1.4	33	6.1	0.4
11	100	5000	1000	1000	6.1	87	35	7.6	2.7
12	5	400	50000	50	6.3	55	30	34	9
13	100	200	1000	$500e^{\gamma_{-0}\mathbf{Y}_{-0}}$	64	0.9	14.1	6	1.7

Table 15 Average results over 10 instances per scenario using AH. In scenario 13, $\boldsymbol{\gamma}_{-0}, \mathbf{Y}_{-0}$ are generated from the same distribution as $\boldsymbol{\beta}_{-0}, \mathbf{X}_{-0}$ respectively.

in Appendix B is applied to the training data in such situations and all such adjusted observations are retained for training. While we observe a moderate degradation in solution quality, the proposed method on average performs well.

D.3. Tangential cutting planes to improve model performance (Setting #3)

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

No. of piecewise linear knots	No tangent cuts			With tangent cuts		
	WMAPE		Runtime	WMAPE		Runtime
	Unobs	Obs	(in secs)	Unobs	Obs	(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 16 Impact of using tangent cuts on the average WMAPEs and run times

concavity in $(0,1)$.

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

$$\ln(1 - f_k) \leq h(f) \leq \ln(1 - f_k) - \frac{1}{1 - f_k}(f - f_k) \qquad \forall k \in K \tag{D.2}$$

The LHS of the inequalities ensure that the functions $g(f), h(f)$ are above or equal to the PWL approximation, while the RHS ensure that they lie 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).

Table 16 provides the results with and without tangential cutting planes for 30 random instances generated for experiment 3 on PWL segments in Section 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.

D.4. Empirical sensitivity analysis of Comp-LM model to increased arrival rates and finer PWL discretization using hotel data

In Table 11, 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 CoV 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 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.

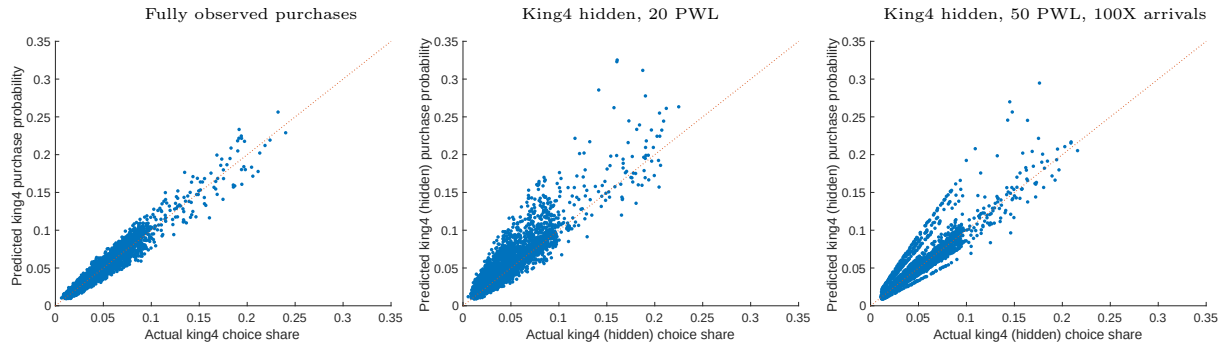


Figure 12 **Scatter plot of the predicted purchase probability versus the actual market share of king4 room.**

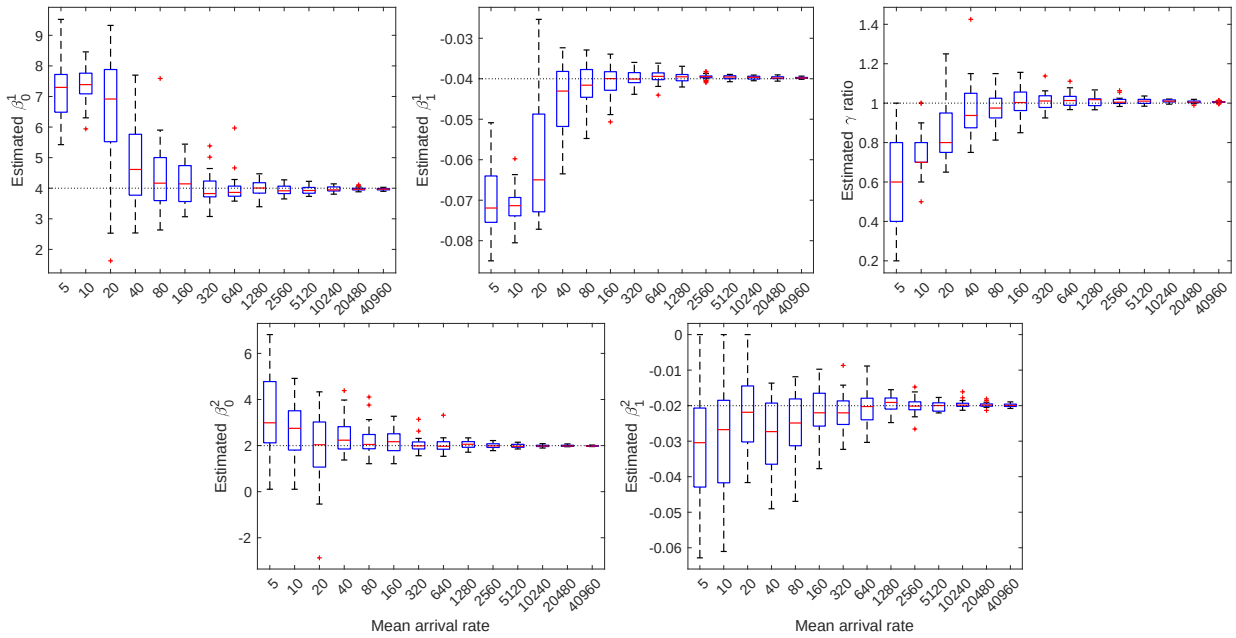


Figure 13 **Box plots of the estimated parameters $\beta_0^1, \beta_1^1, \beta_0^2, \beta_1^2$ and (normalized) γ with 30 distinct instances as a function of increasing mean arrival rate. The true values are marked with dotted lines. Here, the competitor sales (choice 2) is hidden in addition to lost sales.**

Fig. 12 shows the ability of the LM and Comp-LM models to successfully predict the weekly purchase probability of the king4 room choice compared to the actual market shares 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 purchase probability estimated by the Comp-LM model was 4.7% and 4.2% for the 20PWL and 50PWL cases respectively, compared to the unobserved actual value of 4.1%. The average estimated no-purchase probability was 83.9% and 85.6% respectively, compared to the unobserved true value of 86.4%. Overall, the Comp-LM models results are encouraging from a prediction standpoint.

D.5. Empirical consistency of the Comp-LM model

Similar to experiment 1 in Section 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

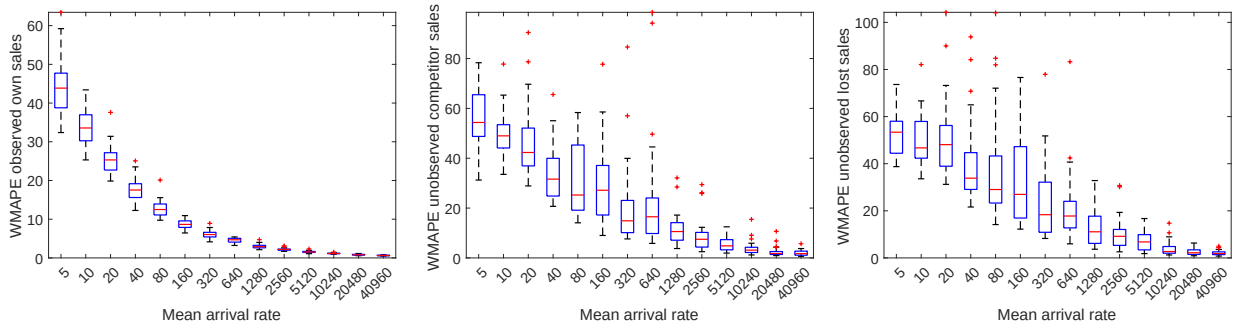


Figure 14 **Average WMPAE of the observed and unobserved sales as a function of the increasing arrival rates. Here, the competitor sales are hidden in addition to lost sales.**

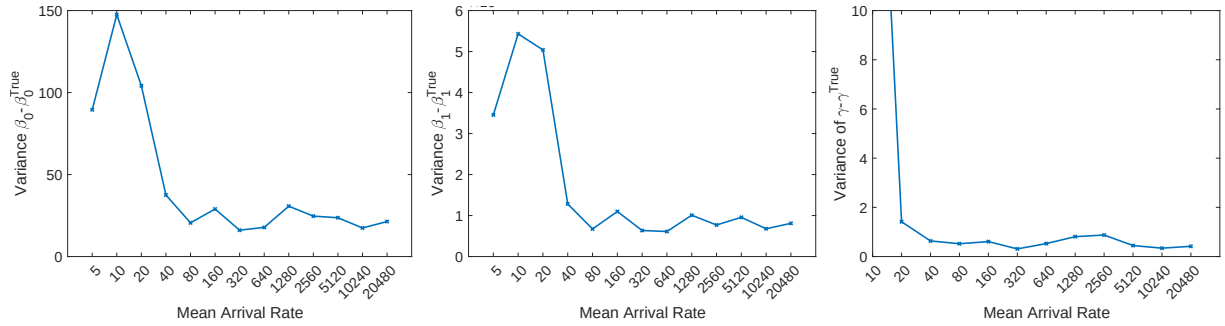


Figure 15 **Finite sample variance scaled by the arrival rate of the error in the estimated parameters β_0, β_1 and (normalized) γ as a function of increasing mean arrival rate.**

is unobserved, along with the no-purchase option. Here, the sales data for choice 2 (competitor's choice) is unobserved. We simulate the arrivals over 50 bins, varying the mean arrival rates from 5 to 40,960. The parameters of the two choices are set to $\beta_0^1 = 4$, $\beta_1^2 = -0.04$ $\beta_0^2 = 2$ and $\beta_1^2 = -0.02$. The prices for each bin are uniformly generated between $[0, 200]$. We generate 30 distinct instances for each arrival rate and estimate the parameters using the Comp-LM model with 20 PWL knots. The 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. 13. We observe that the parameters converge to their true parameters values relatively quickly (within the box after 40 arrivals) with lower biases in the parameter estimates achieved at higher arrival rates. The rate of convergence is slower than the no-competitor case, where an arrival rate of 10 was sufficient to contain the true parameters within the box. We empirically show asymptotic normality in Appendix D.6. In Fig. 14, we present the model fit WMAPEs for the observed and unobserved sales to competitor as well as no-purchase.

D.6. Empirical Asymptotic Normality

We report on an empirical study of asymptotic normality in the following way: (a) plotting the variance of the error in the estimator scaled by the mean arrival rate in Fig. 15 and Fig. 16 respectively for the LM and CompLM estimators respectively and observe that it is roughly a constant; and (b) use a QQ plot (see Fig. 17a,b) to show that for various arrival rates this error is well approximated by a Normal distribution.

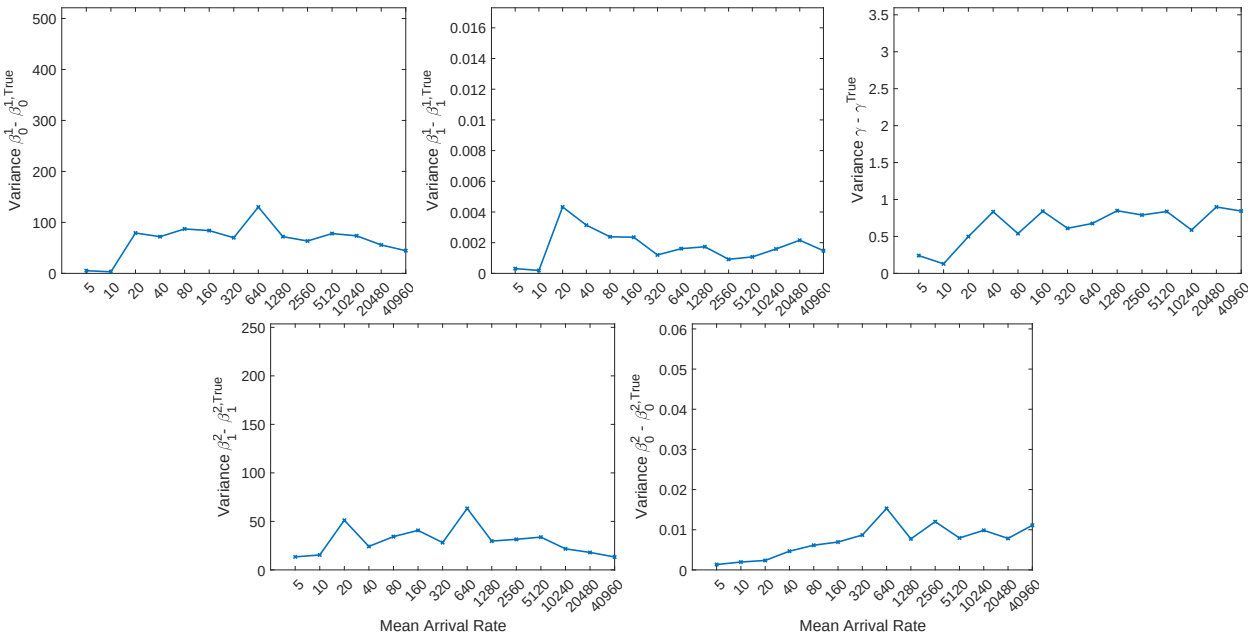
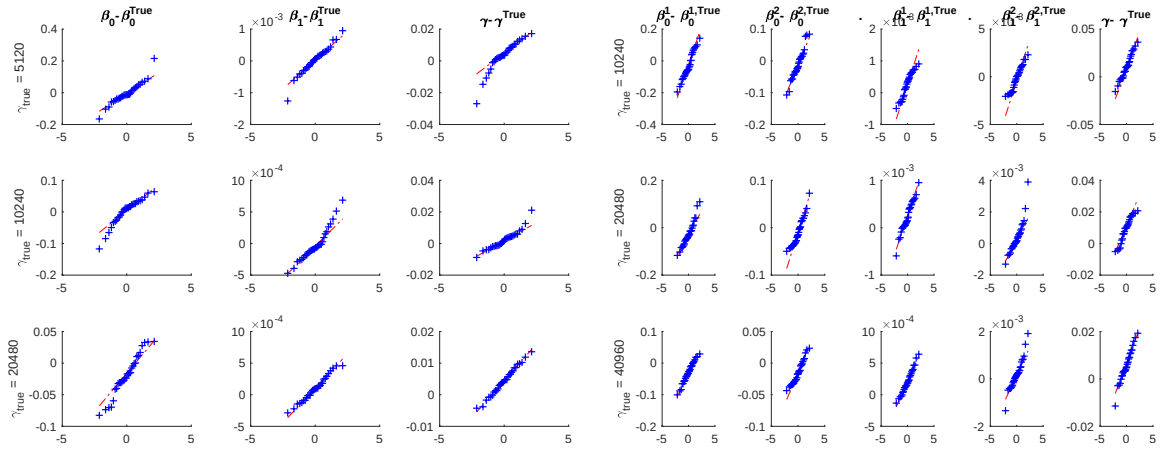


Figure 16 Finite sample variance scaled by the arrival rate of the error in the estimated parameters $\beta_0^1, \beta_1^1, \beta_0^2, \beta_1^2$ and (normalized) γ as a function of increasing mean arrival rate. Here, the competitor sales (choice 2) is hidden in addition to lost sales.



(a) Lost sales censored

(b) Choice 2 and lost sales censored

Figure 17 QQ plot to empirically illustrate normality in two settings for various arrival rate.

Appendix E: Tradeoff underlying the granularity of time bins

When there are frequently changing features over time, the choice of time bins can involve a practical tradeoff between more disaggregate data (more zero sales and more observations) and aggregate bins (increased error-in-variables that dampens the causal effects, and fewer observations). We first discuss the impact of this tradeoff on two different censored estimation approaches (MLE-based EM and LM) given the same aggregated data.

Fewer arrivals per bin (small N_t) and more observations:

1. The probability of zero sales increases. MLE based methods are unaffected by zero sales. On the other hand, the LM model due to its log-transformation may require a heuristic adjustment (such as the Poisson-method discussed in Appendix B) to the training data when the percentage of zero sales is significant (e.g. more than 20%).
2. With more observations, the LM's runtime with an exact MIP solver increases due to the higher number of SOS-2 variables, while the AH converges relatively quickly to good quality solutions. On the other hand, the MLE gradient computations that are repeatedly invoked dominate the EM runtime and become prohibitively expensive for large instances (see Section 4 experiment 6 Table 9 for a comparison).
3. The variance of the market shares increases (inversely proportional to N_t, Cox 1970, Ben-Akiva and Lerman 1985), and this impacts both estimation methods.

More arrivals per bin (large N_t) and fewer observations:
Increased aggregation produces higher error-in-variables (EIV) due to a poorer approximation of the feature variations using one or more aggregate properties such as mean, minimum, maximum, cumulative change, etc. Fewer observations further worsens the prediction accuracy. These issues are likely to affect both estimation approaches, while the reduced number of bins may computationally benefit the exact MIP approach for LM, and faster gradient computations for MLE. Simulations done by Domencich and McFadden (1975) indicate that Berkson's method may provide desirable estimates even in the presence of EIV. This experimental study of EIV is also reported in Ben-Akiva and Lerman (1985).

Empirical Study on tradeoff: We now empirically evaluate this tradeoff on predictive accuracy using real-life B2B data available to us. We briefly summarize our findings. We consider 3 levels of data aggregation: (1) Fine-grained (3-day bins), (2) Weekly groupings, (3) Biweekly groupings. The impact of zero-sales is minimal at the fine-grained aggregation level and the average win rate (actuals in column 3 of Table 10) remains steady at all three levels. The LM results for the weekly-aggregation are discussed in experiment 7 of Section 4, and we briefly compare the performance for the other two levels of aggregation. (a) Fine-grained groupings worsened the win-rate prediction for all four products and the average predicted win-rate error increased from 2.75% at the weekly level to 4.75%, due to the increase in variance even though the observations increase. (b) The biweekly level improved the win-rate prediction for product-4 which had a relatively low weekly sales rate but adversely impacted the other products, resulting in an average win rate error of 4.5%. In practice, in a censored data setting, one cannot measure error in the win rates. We therefore calculated the observed sales WMAPEs and noted that the weekly equivalent average WMAPE for cases (1), (2) and (3) were 25%, 25.1%. and 30% respectively. Clearly, increased aggregation was not beneficial while the reduction in WMAPE at the disaggregate level was insignificant. Based on these findings, we conclude that the weekly granularity represents the best tradeoff for the LM model in this particular setting.

Therefore, in settings where when there are frequently changing features the modeler must choose time bins that are not too fine-grained to limit zero sales and manage runtimes, and not overly aggregated to limit the impact of EIV and preserve predictive accuracy. In our experience, the right level of aggregation has been primarily driven by the underlying decision problem where the forecasting method is employed and/or the users desired frequency of updates.

References

Abdallah T, Vulcano G (2017) 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).

Bixby RE (2012) A brief history of linear and mixed-integer programming computation. *Documenta Mathematica* 107–121.

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), 1 edition.

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.

Harsha P, Subramanian S, Uichanco J (2019) Dynamic pricing of omnichannel inventories. *Manufacturing & Service Operations Management* 21(1):47–65.

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.