Demand Estimation under the Multinomial Logit Model from Sales Transaction Data

Tarek Abdallah
NYU Stern School of Business, IOMS Department, New York, abdalla@stern.nyu.edu

Gustavo Vulcano
NYU Stern School of Business, IOMS Department, New York, gvulcano@stern.nyu.edu
School of Business, Torcuato di Tella University, Buenos Aires, Argentina

May 17, 2016

We study a multinomial logit (MNL) model of demand when customers arrive over time in accordance to a non-homogeneous Poisson process. The model is suitable for retail settings where only sales and product availability data are recorded, not all products are displayed in all periods, and the seller has information about her own aggregate market share. We characterize conditions under which the model is identifiable and the maximum likelihood estimates are consistent.

Then, we propose a Minorization-Maximization (MM) algorithm that inherits the strongest convergence properties of this type of framework: All its limit points are provably stationary points of the associated incomplete data log-likelihood function. In addition, the algorithm is guaranteed to converge to the global optimal solution when the model is identifiable. Through an exhaustive set of numerical comparisons, we benchmark the MM procedure versus state-of-the-art alternative methods: general nonlinear optimization techniques, and an Expectation-Maximization (EM) method. We conclude that the MM-based estimates are of similar quality to the ones obtained by the benchmarks, but the convergence of the MM algorithm is orders of magnitude faster.

Key words: demand uncensoring; maximum likelihood estimation; MM algorithm; EM algorithm; revenue management; retail operations

1. Introduction

Demand forecasts are necessary inputs in bricks-and-mortar retail operations where firms cannot simply wait for the demand to emerge and then respond to it. Instead, they must anticipate and plan for future orders so that they can react immediately to customer requests as they unfold. Similarly, in revenue management (RM) settings, the demand for different products (e.g., airline itinerary-fare-class combinations) is revealed over time, and firms must lock part of their capacity for high valuation customers that may show up later in the selling horizon. In both settings, the quality of the demand estimates play a major role towards effective inventory control and pricing decisions.[1]

Over the last fifteen years, in both the academia and the industry practice there has been a trend of departing from the traditional independent demand model -which assumed that each product had its own captive flow of demand-: first, by accounting for the demand censoring due to stock-outs or deliberate scarcity introduced by the firm, and later, by internalizing substitution effects that occur when customers choose among the available products. The latter substitution phenomenon is common within product families or subcategories in retailing (e.g., scarves within the "women's accessories" category), where the number of SKUs could typically be in the order of dozens.

Even though the classical economic theory on substitution effects (e.g., see Nicholson (1992)) addresses the estimation of demand shifts due to cross price elasticities, the practical operations-related problem stems from the changing availability of products over time in bricks-and-mortar settings where the limited sources of relevant information include sales transaction data, product attributes, and on-hand inventory levels.

Nowadays, the use of discrete choice models is a common approach for estimating demand for different SKUs within a set of comparable items (e.g., Ben-Akiva and Lerman (1994), Train (2003)). Among them, the multinomial logit (MNL) model has captured the most significant attention among practitioners and researchers. The MNL has major restrictions in terms of modeling choice behavior, most notably the property of independence of irrelevant alternatives (IIA), coming from the fact that the ratio of purchase probabilities for two available alternatives is constant regardless of the choice set containing them. More sophisticated choice models (e.g., the nested logit (NL) and mixed MNL models) provide additional flexibility in modeling substitution patterns, but then the tradeoff between specification and estimation error arises. A complex model with many attributes, nests, or latent classes may be able to better approximate a wide range of choice behavior and reduce the specification error. On the other hand, estimating the parameters of such a model from a finite set of data may lead to significant estimation error. A more parsimonious model like the MNL, while potentially less faithful in terms of representing the underlying choice behavior, is less prone to estimation error. Several successful implementations of the MNL model in retail and RM settings have been reported (e.g., Guadagni and Little (1983), and more recently, Ratliff et al. (2008), Vulcano et al. (2010), Vulcano et al. (2012), Newman et al. (2014), Dai et al. (2014)), suggesting that it could provide an interesting balance between specification and estimation errors. In addition, the associated assortment optimization problem is computationally tractable (e.g., see Talluri and van Ryzin (2004), Liu and van Ryzin (2008)), which makes the model particularly attractive from an operational viewpoint.

In this paper we revisit a demand model that has received attention in the recent years (e.g., see Ratliff et al. (2008), Vulcano et al. (2012), Dai et al. (2014)): it combines a multinomial logit (MNL) choice model with nonhomogeneous Poisson arrivals over multiple periods. The problem we address is how to jointly estimate the preference weights of the products and the arrival rates of customers under maximum likelihood (ML) criteria. The only required inputs are observed historical sales, product availability data, and market share information. Our contribution is two-fold. From a theoretical perspective, we characterize a condition under which the model is identifiable. This condition is obtained from a directed graph representation of the model, and reduces to checking if such graph is strongly connected. When the model is indeed identifiable, the estimates inherit the statistical properties of ML estimates: they are asymptotically normal, and asymptotically efficient (i.e., asymptotically unbiased and attaining equality of the Cramér-Rao lower bound for the variance, asymptotically). Regarding the consistency of the estimates (i.e., their convergence in probability to the true parameter values), we discuss the *incidental parameters problem*. This problem emerges when the number of parameters grows as the sample size does. In our case, having more periods implies the increase in the number of Poisson arrival rate parameters. Nonetheless, we show that the MNL estimates are still consistent in this case, and we characterize conditions under which the arrival rate estimates are as well.

Our second contribution is practical and refers to the development of a new minorization-maximization (MM) algorithm for finding maximum likelihood estimates of the demand model under study. The MM algorithm is a generic, iterative procedure for maximizing an objective function (in our case, a likelihood function) by successively maximizing a surrogate function that minorizes the true objective function. Optimizing the surrogate function drives the objective function upward until a local optimum is reached. MM algorithms have been studied under various names for over 40 years, though the initials MM originate with a rejoinder of Hunter and Lange (2000). Our contribution is to specialize the MM procedure to our formulation and show that it is indeed straightforward to implement in any simple procedural language or numerical computing environment via iterations involving closed-form expressions. Our MM inherits the strongest convergence properties of this type of framework: All its limit points are provably stationary points of the associated incomplete data log-likelihood function. The initials "MM" emphasize the close tie between MM algorithms and its best-known special cases, expectation-maximization (EM) algorithms. An EM algorithm operates by identifying a theoretical complete data space. In its E-step, the conditional expectation of the complete data loglikelihood is calculated with respect to the observed data. The surrogate function created by the E-step is, up to a constant, a minorizing function. Implementations of the EM algorithm have shown its potential over this type of demand model regarding computational speed and quality of the solutions obtained, configuring a state-of-the-art method (e.g., see Vulcano et al. (2012), who showed that EM was orders of magnitude faster than general nonlinear optimization methods). Here, through an exhaustive set of numerical comparisons, we benchmark the MM procedure versus the EM method and conclude that the MM-based estimates are of similar quality to the EM-based ones, but they are obtained in orders of magnitude faster computational time.

2. Literature Review

During the last decade the Operations-related literature has shown an increasing interest in choice-based demand estimation and assortment planning. One of the early papers to address both problems was Talluri and van Ryzin (2004), who develop an EM method to jointly estimate arrival rates and parameters of an MNL choice model based on sales transaction data under unobservable no-purchases. We refer the reader to Kok et al. (2008) for a review of the early literature on demand estimation and assortment planning, and to Section 2 in Vulcano et al. (2012) for a discussion on demand uncensoring, MNL-based substitution, and EM-based proposals previous to 2009. In the paragraphs below we discuss few recent references.

Talluri (2009) proposes a method to jointly estimate MNL parameters and market size for a retailer subcategory. He uses a semi-parametric approach where the market size is first realized as a random draw from an arbitrary distribution, and then the realized number of consumers arrive uniformly during the selling horizon. Upon arrival, a consumer chooses among the available products according to an MNL model with covariates. He proposes a two-step approach where in the first step the MNL parameters are estimated based on the classical ML approach conditioned on a purchase, which poses a globally concave optimization problem. In the second step, he proposes a method that estimates the intercept parameter (i.e., the utility of the no-purchase alternative) using a risk ratio defined as the ratio of the purchasing probability for different assortments. He minimizes the squared difference of the realized risk ratio and the average observed risk ratio to estimate the intercept parameter. Talluri (2009) states that deriving statistical properties of his estimation procedure is potentially challenging, but argues that the numerical simulations provide reasonable validation for his proposal.

Musalem et al. (2010) propose a new approach to demand uncensoring by studying sales data of a retailer with partial information on product availability. Their method assumes a periodic review inventory system with infrequent replenishment. Consumer choice is captured by a random-coefficients MNL model. Instead of using a ML estimation approach, they devise a new methodology based on data augmentation which benefits from the beginning and ending per-period inventory to estimate the missing information (the products available for each consumer upon his arrival). Then, they use a computationally intensive estimation procedure that combines Markov chain Monte Carlo (MCMC) with sampling, using Bayesian methods to compute the likelihood function.

Newman et al. (2014) consider an underlying MNL model with covariates, where the no purchase alternative is not observable, and the arrivals occur in accordance with a homogeneous Poisson process. They propose a two-step approach based on decomposing the log-likelihood function into marginal and conditional components. The first step is identical to the first step in Talluri (2009) where the choice parameters of each item are estimated based on the classical MNL, maximum likelihood approach conditioned on a purchase. The second step is to plug-in the estimates from the first step into the joint log-likelihood problem in order to estimate the utility intercept and the mean arrival rate. However, their second step is not guaranteed to be a concave optimization problem and hence, consistency of the estimates cannot be claimed in general. Moreover, they note that their estimates are not necessarily efficient. Through an extensive computational study, they show that their method is much faster than the EM proposal by Talluri and van Ryzin (2004).

The estimation problem posed by Talluri and van Ryzin (2004) was also studied by Subramanian and Harsha (2015), who propose a MIP formulation that is claimed to work extremely fast on multiple real world data sets.

Ding and Kleywegt (2015) focuses on the critical identification issue that may arise in the demand model that is the focus of our study: Poisson arrivals and MNL choice. They state necessary and sufficient conditions for the demand to be identifiable based on the infeasibility of a set of linear inequalities. Our theoretical contribution follows a different perspective: it relates to the characterization of necessary and sufficient conditions for identifiability based on assessing the strong connectivity of an appropriately defined directed graph.

From a methodological angle, our paper contributes to the literature on MM algorithms by providing another successful implementation of the technique. Surveys on the early work using this type of methods can be found in Heiser (1995) and Lange et al. (2000). Hunter (2004) proposes an MM algorithm to estimate Bradley-Terry models for paired and multiple comparisons. In Section 4 therein, he discusses general convergence properties of MM algorithms that we borrow for the convergence results of our proposal.

3. Model description

3.1. Basics

Consider a retailer selling substitutable products from a category with items $\mathcal{N} = \{1, \ldots, n\}$. The firm offers a sequence of assortments $S_t \subset \mathcal{N}$ over a selling horizon of T purchase periods, indexed $t = 1, 2, \ldots, T$. The periods could potentially be of different lengths.

The retailer keeps track of the number of purchase transactions for each of the products in S_t. We assume that a product is either fully available or not available throughout a given period t. The number of purchases of product i observed in period t is denoted z_{it}, and define $\boldsymbol{z}_t = (z_{1t}, \ldots, z_{nt})$. We will assume that $z_{it} \geq 0$ for all i, t (i.e., we do not consider returns). If $i \in \mathcal{N} \setminus S_t$, then $z_{it} = 0$. We will further assume, without loss of generality, that for each product i, there exists at least one period t such that $z_{it} > 0$; else, we can drop product i from the analysis. We denote by $m_t = \sum_{i=1}^{n} z_{it}$ the total number of observed purchases in period t.

In each period, the number of customers arriving to the store is described by a Poisson random variable with mean λ_t. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)$ denote the vector of arrival rates. A major incompleteness in the data is that there is no information about the number of customers who arrive in a period but did not make a purchase; that is, the total number of transactions in a given period is a censored approximation to the true demand of the period. This is a common limitation in retail settings in which sales transactions and item availability are frequently the only data available. We treat the no-purchase option as a separate product (labeled zero) that is always available.

Customers choose among the alternatives in S_t according to a time-homogeneous MNL model. Under the MNL model, the choice probability of a customer is defined based on a preference vector $\boldsymbol{v} \in \mathcal{R}^n$, $\boldsymbol{v} > 0$, that indicates the customer "preference weights" or "attractiveness" for the different products. This vector, together with a no-purchase preference weight v_0, determine a customer's choice probabilities as follows: let $P_i(S, \boldsymbol{v})$ denote the probability that a customer chooses product $i \in S$ when S is offered and preference weights are given by vector $\boldsymbol{v}$. Then,

$$P_i(S_t, \boldsymbol{v}) = \frac{v_i}{\sum_{j \in S_t} v_j + v_0} \quad \text{for} \quad i \in S_t, \ \text{and} \ P_i(S_t, \boldsymbol{v}) = 0 \ \ \text{otherwise.} \tag{1}$$

The no-purchase probability, capturing the fact that when set S is offered, a customer may either buy a product from a competitor or not buy at all, is given by

$$P_0(S, \boldsymbol{v}) = \frac{v_0}{\sum_{j \in S_t} v_j + v_0}.$$

In our presentation, we directly focus on the vector $\boldsymbol{v}$ of attractiveness of the products without requiring attribute selection to define the product utilities. This approach is common in operations-related applications, where the product universe is fixed and attribute selection is non-trivial.[2]

The objective is to jointly estimate the MNL parameters $\boldsymbol{v}$ along with the non-homogenous arrival rates $\boldsymbol{\lambda}$, using the classical maximum likelihood approach. We note here that even if the classical estimation problem of the MNL choice model with complete data (i.e., when the modeler observes the no purchases) is relatively simple, the joint estimation problem of $(\boldsymbol{v}, \boldsymbol{\lambda})$ with censored data can be complicated. In this paper, we show that the estimation problem can be reformulated as an estimation problem of only the choice model parameters $\boldsymbol{v}$. In other words, the joint estimation problem with incomplete data is as "hard" as estimating the classical MNL with complete data.

3.2. Incomplete data likelihood function

The first natural estimation approach would be to attempt to solve directly the estimation problem using maximum likelihood (ML). Following Vulcano et al. (2012), the incomplete data likelihood function can be expressed as:

$$\mathcal{L}_I(\boldsymbol{v}, \boldsymbol{\lambda}) \ = \Pi_{t=1}^{T} \left(\mathbb{P}\left(m_t \text{ customers buy in period } t | \boldsymbol{v}, \boldsymbol{\lambda}\right) \frac{m_t!}{\Pi_{i \in S_t} z_{it}!} \Pi_{i \in S_t} \left(\frac{v_i}{v_0 + \sum_{j \in S_t} v_j} \right)^{z_{it}} \right), \tag{2}$$

where

$$\mathbb{P}(m_t \text{ customers buy}) = \frac{\left[\lambda_t \sum_{i\in S_t} P_i(S_t,\boldsymbol{v})\right]^{m_t} \exp(-\lambda_t \sum_{i\in S_t} P_i(S_t,\boldsymbol{v}))}{m_t!}.$$

Taking logarithm, we can write

$$\begin{aligned}
\ell_I(\boldsymbol{v},\boldsymbol{\lambda}) &= \sum_{t=1}^{T} \log\left[\frac{\left[\lambda_t \sum_{i\in S_t} P_i(S_t,\boldsymbol{v})\right]^{m_t} \exp\left(-\frac{\lambda_t \sum_{i\in S_t} v_i}{v_0+\sum_{i\in S_t} v_i}\right)}{z_{1t}!\ldots z_{nt}!} \prod_{j\in S_t}\left(\frac{v_j}{\sum_{i\in S_t} v_i}\right)^{z_{jt}}\right] \\
&= \sum_{t=1}^{T}\left[\log\left(\frac{\left[\frac{\lambda_t \sum_{i\in S_t} v_i}{v_0+\sum_{i\in S_t} v_i}\right]^{m_t} \exp\left(-\frac{\lambda_t \sum_{i\in S_t} v_i}{v_0+\sum_{i\in S_t} v_i}\right)}{z_{1t}!\ldots z_{nt}!}\right) + \sum_{j\in S_t} \log\left(\frac{v_j}{\sum_{i\in S_t} v_i}\right)^{z_{jt}}\right] \\
&= \sum_{t=1}^{T}\left[m_t \log\left(\frac{\lambda_t}{v_0+\sum_{j\in S_t} v_j}\right) + m_t \log(\sum_{i\in S_t} v_i) - \frac{\lambda_t \sum_{i\in S_t} v_i}{v_0+\sum_{i\in S_t} v_i}\right. \\
&\qquad\qquad \left. - \sum_{i=1}^{n} \log(z_{it}!) + \sum_{j\in S_t}\left(z_{jt}\log v_j - z_{jt}\log(\sum_{i\in S_t} v_i)\right)\right] \\
&= \sum_{t=1}^{T}\left[m_t \log\left(\frac{\lambda_t}{v_0+\sum_{j\in S_t} v_j}\right) - \frac{\lambda_t \sum_{i\in S_t} v_i}{v_0+\sum_{i\in S_t} v_i} - \sum_{i=1}^{n} \log(z_{it}!) + \sum_{j\in S_t} z_{jt}\log(v_j)\right].
\end{aligned} \tag{3}$$

The last equality holds since $\sum_{j\in S_t} z_{jt} = m_t$, and thus $\sum_{j\in S_t} z_{jt}\log(\sum_{i\in S_t} v_i) = m_t \log(\sum_{i\in S_t} v_i)$. For future reference, the incomplete data log-likelihood function (after we drop the constant $-\sum_{i=1}^{n}\log(z_{it}!)$) can be written as

$$\ell_I(\boldsymbol{v},\boldsymbol{\lambda}) = \sum_{t=1}^{T}\left[m_t \log\left(\frac{\lambda_t}{v_0+\sum_{j\in S_t} v_j}\right) - \lambda_t \frac{\sum_{i\in S_t} v_i}{v_0+\sum_{j\in S_t} v_j} + \sum_{j\in S_t} z_{jt}\log v_j\right]. \tag{4}$$

Let $K_j := \sum_{t=1}^{T} z_{jt}$ be the total purchases of item j over the selling horizon, then we have

$$\sum_{t=1}^{T}\sum_{j\in S_t} z_{jt}\log v_j = \sum_{j=1}^{n}\sum_{t=1}^{T} z_{jt}\log v_j = \sum_{j=1}^{n} K_j \log v_j.$$

The optimization of $\ell_I(\boldsymbol{v},\boldsymbol{\lambda})$ can be reduced to the following maximization problem

$$\max_{\boldsymbol{v}>0,\boldsymbol{\lambda}>0} \sum_{t=1}^{T}\left[m_t \log\left(\frac{\lambda_t}{v_0+\sum_{j\in S_t} v_j}\right) - \lambda_t \frac{\sum_{i\in S_t} v_i}{v_0+\sum_{j\in S_t} v_j}\right] + \sum_{j=1}^{n} K_j \log v_j. \tag{5}$$

The above log-likelihood function is hard to solve in general due to the complicating terms $\lambda_t \frac{v_i}{v_0+\sum_{j\in S_t} v_j}$. However, we next show that maximizing the log-likelihood function $\ell_I(\boldsymbol{v},\boldsymbol{\lambda})$ can be reduced to a problem of simply estimating the parameters $\boldsymbol{v}$ of the MNL choice model while eliminating the no-purchase alternative.

PROPOSITION 1. *If there is a solution* $(\boldsymbol{v}^*,\boldsymbol{\lambda}^*)$ *to problem* (5)*, then it verifies:* $(\boldsymbol{v}^*,\boldsymbol{\lambda}^*) \in \arg\max_{\boldsymbol{v},\boldsymbol{\lambda}} \ell_I(\boldsymbol{v},\boldsymbol{\lambda})$ *if and only if* $\lambda_t^* = m_t \frac{v_0+\sum_{j\in S_t} v_j^*}{\sum_{j\in S_t} v_j^*}$ *and* $\boldsymbol{v}^* \in \arg\max_{\boldsymbol{v}} \ell_{MNL}(\boldsymbol{v})$*, where*

$$\ell_{MNL}(\boldsymbol{v}) := \sum_{j=1}^{n} K_j \log v_j - \sum_{t=1}^{T} m_t \log\left(\sum_{i\in S_t} v_i\right), \tag{6}$$

and where we define

$$\mathcal{L}_{MNL}(\boldsymbol{v}) := \exp(\ell_{MNL}(\boldsymbol{v})) = \Pi_{t=1}^{T}\Pi_{i\in S_t}\left(\frac{v_i}{\sum_{j\in S_t} v_j}\right)^{z_{it}}. \tag{7}$$

We note that $\ell_{MNL}(\cdot)$ and $\mathcal{L}_{MNL}(\cdot)$ are respectively the log-likelihood and likelihood functions of the classical MNL choice model conditional on the observed transactions. Also both $\ell_{MNL}(\cdot)$ and $\mathcal{L}_{MNL}(\cdot)$ are independent of the no-purchase alternative weight v_0. Hence, Proposition 1 implies that the joint estimation problem of $(\boldsymbol{v},\boldsymbol{\lambda})$ is equivalent to the estimation of the classical MNL choice model conditional on the observed transactional data and ignoring the no purchase option.

If we define $\beta_i := \log(v_i)$ and substitute in (6), we get the reduced log-likelihood function

$$\ell_{MNL}(\boldsymbol{\beta}) := \sum_{j=1}^{n} K_j\beta_j - \sum_{t=1}^{T} m_t \log\left(\sum_{i\in S_t} \exp(\beta_i)\right). \tag{8}$$

The function is concave but does not necessarily have a global maximum; it could be unboundedly increasing or have multiple solutions. In the next section we characterize necessary and sufficient conditions under which there is a unique optimal solution, among other statistical properties.

4. Statistical properties of the demand model

As noted by Vulcano et al. (2012), the likelihood function (4) does not have much structure in general. This implies that the demand model is not guaranteed to be identifiable, which prevents the borrowing of the desirable statistical properties of ML. In this section, we leverage the characterization of statistical properties of the demand model.

4.1. Identifiability

The demand model is identifiable if there is a unique set of underlying parameters $(\boldsymbol{v},\boldsymbol{\lambda})$ that maximizes the likelihood function (4). In order to investigate the identification conditions, we start by noting that given the one-to-one correspondence between β_i and v_i, it would be enough to establish conditions under which there is a unique $\boldsymbol{\beta}^*$ that solves (8). Proposition 1 guarantees that the associated $\boldsymbol{\lambda}^*$ would also be unique.

Observe that the function $\ell_{MNL}(\boldsymbol{\beta})$ is independent of β_0. Therefore, the first step is to normalize $\boldsymbol{\beta}$ by setting $\beta_0 = 0$. However, this is not sufficient to ensure that $\boldsymbol{\beta}^*$ is unique. This identification issue has also been noticed by (Vulcano et al. 2012, Section 3.3). It can be easily verified that the same misidentification problem persists in the reduced log-likelihood problem $\ell_{MNL}(\boldsymbol{\beta})$, where for any feasible solution $\boldsymbol{\beta}$ there exists a continuum of solutions which achieve the same objective function value. We summarize this observation in the following lemma.

Lemma 1. *Consider the reduced log-likelihood function $\ell_{MNL}(\boldsymbol{\beta})$. For any $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, the solution $T(\boldsymbol{\beta}) := \{\alpha + \beta_i\}_{i=1}^n$ achieves the same reduced log-likelihood value, i.e., $\ell_{MNL}(\boldsymbol{\beta}) = \ell_{MNL}(T(\boldsymbol{\beta}))$.*

As pointed out by (Vulcano et al. 2012, Section 3.3), this continuum of log-likelihood values can be resolved by imposing an additional constraint on the parameter values related to market share, defined as $s := \frac{\sum_{j=1}^n \exp(\beta_j)}{1+\sum_{j=1}^n \exp(\beta_j)}$. In principle, this could be problematic for the maximization of $\ell_{MNL}(\boldsymbol{\beta})$ since it adds a non-convex constraint, but the simple EM algorithm that they propose can easily overcome this difficulty. We will follow the same approach of fixing the market share here.[3]

We introduce two definitions. Let the restricted log-likelihood problem be denoted as

$$L_R(\tilde{s}) := \max_{\boldsymbol{\beta}} \{\ell_{MNL}(\boldsymbol{\beta}) : \sum_{i=1}^n \exp(\beta_i) = \tilde{s}\}, \quad \text{where} \quad \tilde{s} := \frac{s}{1-s}.$$

Also define the unrestricted log-likelihood problem $L := \sup_{\boldsymbol{\beta}} \ell_{MNL}(\boldsymbol{\beta})$ with $\Psi := \{\boldsymbol{\beta} \in \mathbb{R}^n : \ell_{MNL}(\boldsymbol{\beta}) = L\}$ as the corresponding set of bounded solutions. In principle, Ψ could be empty if the solution is unbounded. Later on, we will provide conditions under which Ψ is not empty.

We next show that any oracle which provides an optimal solution $\boldsymbol{\beta}^*$ to the unrestricted MNL estimation problem can be used to provide an optimal solution to the restricted market share MNL estimation problem without any loss of computational efficiency. Assume for now that such oracle exists. Let $c_1 := \sum_{i=1}^n \exp(\beta_i^*)$, and define the following transformation:

$$T(\boldsymbol{\beta}, c_0) := \log\left(\frac{c_0}{c_1}\right) + \boldsymbol{\beta}. \tag{9}$$

In the next proposition, we show that $T(\boldsymbol{\beta}^*, \tilde{s})$ is indeed an optimal solution for the restricted market share problem for any given $\tilde{s}$.

Proposition 2. *Assume there exists an oracle that provides an optimal solution $\boldsymbol{\beta}^*$ to the unrestricted problem, i.e. $\boldsymbol{\beta}^* \in \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \ell_{MNL}(\boldsymbol{\beta})$. Then, $T(\boldsymbol{\beta}^*, \tilde{s}) \in \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \{\ell_{MNL}(\boldsymbol{\beta})\ s.t\ \sum_{i=1}^n \exp(\beta_i) = \tilde{s}\}$.*

In what follows, we characterize sufficient and necessary conditions for the identifiability of the choice parameters $\boldsymbol{\beta}$ under the market share constraint. In the next lemma, we establish the connection between the identification conditions of $L_R(\tilde{s})$ and the structure of Ψ.

Lemma 2. *Assume that the unrestricted optimal solution set Ψ has at least one bounded solution, then the following statements are equivalent:*

(i) There is a unique bounded optimal solution to $L_R(\tilde{s})$, where $0 < \tilde{s} < \infty$.

(ii) For any bounded solutions $\boldsymbol{\gamma}, \boldsymbol{\xi} \in \Psi$, we have

$$\gamma_i - \xi_i = \log\left(\frac{\tilde{s}_\gamma}{\tilde{s}_\xi}\right), \qquad \textit{for all } i \in \mathcal{N},$$

where $\tilde{s}_y := \sum_{i=1}^n \exp(y_i)$.

Note that the lemma does not rule out the possibility of an unbounded optimal solution for both the restricted and unrestricted problems.

The usual arguments in the literature to derive sufficient identification conditions are based on establishing strict concavity of the log-likelihood function. However, based on Lemma 2, we have established that identifiability under the restricted market share problem $L_R(\tilde{s})$, which includes a non-convex constraint, is equivalent to the case where the classical MNL log-likelihood problem has multiple solutions with a specific structure. Next, by adapting the arguments described in Hunter

(2004) to exploit the special structure of the optimal solution set Ψ of the unrestricted problem, we show that analogous identification conditions discussed by him also hold for the restricted market share problem (see Assumption 3 and Lemma 2 therein).

Given a set of transactional data with availability information, $\{(\boldsymbol{z}_t, S_t)\}_{t=1}^{T}$, we define the following *directed* graph $G(\mathcal{N}, E)$, where the nodes are the items and the edges are constructed by the following procedure: for each time period t, if $z_{it} > 0$, add a directed edge $i \to j$ for all $j \in S_t, j \neq i$.

PROPOSITION 3. *Given the data $\{(\boldsymbol{z}_t, S_t)\}_{t=1}^{T}$ and a fixed market share s, $0 < s < 1$, then the following statements are equivalent for a normalized $\beta_0 = 0$ and $\tilde{s} = s/(1-s)$:*

(i) $G(\mathcal{N}, E)$ is strongly connected.

(ii) There exists a unique stationary point $\boldsymbol{\beta}$ of $\ell_{MNL}(\boldsymbol{\beta})$, i.e. $\nabla \ell_{MNL}(\boldsymbol{\beta}) = 0$, that is feasible for the market share restriction. Moreover, this $\boldsymbol{\beta}$ is the unique optimal solution of $L_R(\tilde{s})$.

(iii) The log-likelihood function $\ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda})$ -obtained by replacing v_i with $\exp(\beta_i)$ in (4)- with a fixed market share s, $0 < s < 1$, has a unique stationary point $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, i.e. $\nabla \ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda}) = 0$, that is feasible for the market share restriction. Moreover, this $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is the unique optimal solution of $\ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda})$ subject to the market share constraint.

A few observations follow. First, note that checking if the demand model is identifiable reduces to checking if a directed graph is strongly connected, which can be done in linear time. A trivial approach would be to run depth-first-search (DFS) from every node, and verify if all other nodes are reachable, which runs in $O(n^2)$. Second, Proposition 3 streamlines the main theoretical result in Vulcano et al. (2012) (see Theorem 1 therein). Since in general the log-likelihood function is not unimodal, then the EM algorithm is not guaranteed to converge; but if it does converge, it does so to a stationary point of the incomplete data log-likelihood function $\ell_I(\boldsymbol{v}, \boldsymbol{\lambda})$. Now, Proposition 3 states that if the aforementioned directed graph $G(\mathcal{N}, E)$ is indeed strongly connected, then the demand model is identifiable, and the optimal solution is a unique stationary point. Hence, the EM algorithm is also guaranteed to converge to the globally optimal solution. This is due to the continuously differentiable conditional expected log-likelihood function in the E-step therein, and Corollary 1 in Wu (1983).

4.2. Consistency

The maximum likelihood estimates are not necessarily unbiased but are consistent (i.e., they converge in probability to the true parameter values), asymptotically normal, and asymptotically efficient (i.e., asymptotically unbiased and attaining equality of the CramérRao lower bound for the variance, asymptotically) under some mild regularity conditions (see for example (Greene 2011, Chapter 14)). A necessary condition for the ML estimates to be consistent is the identifiability of the model.

The natural regime to study the asymptotic properties of the estimators is to scale the number of periods T. However, it is worth noting that in this case the original log-likelihood problem $\ell_I(\boldsymbol{\lambda}, \boldsymbol{\beta})$ suffers from the *incidental parameters problem*, where the number of parameters λ_t grows to infinity as T goes to infinity. In this case, the ML estimates need not be consistent despite satisfying the regularity conditions Neyman and Scott (1948). However, we show next that our estimates for $\boldsymbol{\beta}$ are indeed consistent regardless of the incidental parameters problem. Yet, we cannot establish the consistency of the estimate of $\boldsymbol{\lambda}$. We later propose a modification to the Poisson arrival process to eliminate the incidental parameters problem and get consistent estimates for $\boldsymbol{\lambda}$.

4.2.1. Incidental parameters problem

The non-homogeneous Poisson arrival process leads to the *incidental parameters problem*. In particular, for each t we have a different λ_t, which implies that the estimator for λ_t will be solely based on a single period observation m_t. As a result, it is not possible to use a law of large number argument in order to establish consistency. Parameters such as $\boldsymbol{\lambda}$ for which it is not possible to obtain consistent estimators are usually referred to as *nuisance* parameters. However, the major concern in the presence of the incidental parameters problem is that even if we find the globally optimal solution to the MLE problem, the estimators of the parameters $\hat{\boldsymbol{\beta}}$ are in general not guaranteed to be consistent.

An alternative approach is to model the arrivals as a Poisson process with a constant arrival rate λ, as in Newman et al. (2014). These authors propose a two-step estimation algorithm where the first step is to estimate the MNL parameters based on a partial likelihood function while normalizing for example $\beta_1 = 0$. Unfortunately, the second step in their estimation is not a concave optimization, therefore it is not possible to guarantee the identifiability (and hence, the consistency) of the ML estimates. In particular, they can only guarantee the consistency of estimates $\{\hat{\beta}_i\}_{i=2}^n$ but not of $\hat{\beta}_0$ and $\hat{\lambda}$. Also, since their estimates $\{\hat{\beta}_i\}_{i=2}^n$ are based on the conditional likelihood, Newman et al. (2014) note that the estimates using the two-step estimation process are not necessarily efficient.

We first argue that the non-homogeneous Poisson process can guarantee the consistency of $\{\hat{\beta}_i\}_{i=1}^n$ despite the incidental parameters problem. In addition, unlike Newman et al. (2014), the estimators are efficient with respect to the constrained Cramér-Rao bound Moore et al. (2008). The only drawback is that we cannot guarantee the consistency of $\hat{\boldsymbol{\lambda}}$. Later, we propose a slight modification to the model that eliminates the incidental parameters issue, so as to extend the consistency to the arrival rates $\hat{\boldsymbol{\lambda}}$.

In order to establish the consistency and efficiency of the ML estimates $\{\hat{\beta}_i\}_{i=1}^n$, we first introduce a definition related to the log-likelihood function $\ell_I(\boldsymbol{v}, \boldsymbol{\lambda})$ in (4), where we replace v_i by $\exp(\beta_i)$, and v_0 by 1:

Definition 1 (Profile log-likelihood). The profile log-likelihood function of $\ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is given by

$$\ell_{I,\text{Profile}}(\boldsymbol{\beta}) := \max_{\boldsymbol{\lambda}} \ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda}).$$

The profile log-likelihood function is obtained by replacing $\boldsymbol{\lambda}$ with the optimal $\boldsymbol{\lambda}^*(\boldsymbol{\beta})$, for a given $\boldsymbol{\beta}$, where according to Proposition 1,

$$\lambda_t^*(\boldsymbol{\beta}) = m_t \frac{1 + \sum_{j \in S_t} \exp(\beta_j)}{\sum_{j \in S_t} \exp(\beta_j)}.$$

It can be verified now that $\ell_{MNL}(\boldsymbol{\beta})$ in (8) is proportional to $\ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda}^*(\boldsymbol{\beta}))$.

In general, in the presence of a large number of nuisance parameters λ_t, inference based on the profile likelihood $\ell_{I,\text{Profile}}(\cdot)$ is not guaranteed to provide consistent estimators $\{\hat{\beta}_i\}_{i=1}^n$ Barndorff-Nielsen (1983), Cox and Reid (1987). However, we claim that in our case, inference based on the profile log-likelihood still provides consistent estimates for $\{\beta_i\}_{i=1}^n$ since the profile log-likelihood function is at the same time proportional to a "conditional" likelihood function which provides consistent estimates. In particular, let $f(\boldsymbol{z}_t;\boldsymbol{\beta},\lambda_t)$ denote the density of the data generating process where $\mathcal{L}_I(\boldsymbol{\beta},\boldsymbol{\lambda})$ is the corresponding likelihood function (consider (2), where we replace v_0 by 1 and v_i by $\exp(\beta_i)$). We have that

$$f(\boldsymbol{z}_t;\boldsymbol{\beta},\lambda)=\left(\frac{\lambda_t\sum_{i\in S_t}\exp(\beta_i)}{1+\sum_{j\in S_t}\exp(\beta_j)}\right)^{m_t}\exp\left(-\frac{\lambda_t\sum_{i\in S_t}\exp(\beta_i)}{1+\sum_{j\in S_t}\exp(\beta_j)}\right)\frac{1}{\Pi_{i\in S_t}z_{it}!}\Pi_{i\in S_t}\left(\frac{\exp(\beta_i)}{1+\sum_{j\in S_t}\exp(\beta_j)}\right)^{z_{it}}$$

Notice that $f(\boldsymbol{z}_t;\boldsymbol{\beta},\lambda)$ can be decomposed into a marginal and a conditional density on the number of purchases m_t such that

$$\begin{aligned} f(\boldsymbol{z}_t;\boldsymbol{\beta},\lambda)&=f(\boldsymbol{z}_t|\boldsymbol{m}_t;\boldsymbol{\beta},\lambda_t)f(\boldsymbol{m}_t;\boldsymbol{\beta},\lambda_t),\ \text{where}\\ f(m_t;\boldsymbol{\beta},\lambda_t)&:=\left(\frac{\lambda_t\sum_{i\in S_t}\exp(\beta_i)}{1+\sum_{j\in S_t}\exp(\beta_j)}\right)^{m_t}\exp\left(-\frac{\lambda_t\sum_{i\in S_t}\exp(\beta_i)}{1+\sum_{j\in S_t}\exp(\beta_j)}\right),\ \text{and}\\ f(\boldsymbol{z}_t|m_t;\boldsymbol{\beta},\lambda_t)&=\frac{1}{\Pi_{i\in S_t}z_{it}!}\Pi_{i\in S_t}\left(\frac{\exp(\beta_i)}{1+\sum_{j\in S_t}\exp(\beta_j)}\right)^{z_{it}}. \end{aligned}$$

We can now define both the marginal and the conditional likelihood functions.

Definition 2 (Marginal likelihood). The marginal likelihood function is given by

$$\mathcal{L}_1(\boldsymbol{m};\boldsymbol{\beta},\boldsymbol{\lambda}):=\prod_{t=1}^T f(m_t;\boldsymbol{\beta},\lambda_t).$$

Definition 3 (Conditional likelihood). The conditional likelihood function is given by

$$\mathcal{L}_2(\boldsymbol{Z}|\boldsymbol{m};\boldsymbol{\beta}):=\prod_{t=1}^T f(\boldsymbol{z}_t|m_t;\boldsymbol{\beta},\lambda_t).$$

Notice that, by construction, the incomplete likelihood is already written as the product of the marginal and the conditional likelihood, such that

$$\mathcal{L}_I(\boldsymbol{\beta},\boldsymbol{\lambda})=\mathcal{L}_1(\boldsymbol{m};\boldsymbol{\beta},\boldsymbol{\lambda})\mathcal{L}_2(\boldsymbol{Z}|\boldsymbol{m};\boldsymbol{\beta}),$$

In our case, we are interested in the statistical properties of the estimators of the conditional likelihood $\mathcal{L}_2(\boldsymbol{Z}|\boldsymbol{m};\boldsymbol{\beta})$ since $\mathcal{L}_{MNL}(\boldsymbol{\beta})\propto\mathcal{L}_2(\boldsymbol{Z}|\boldsymbol{m};\boldsymbol{\beta})$, where $\mathcal{L}_{MNL}(\boldsymbol{\beta})$ is the reduced likelihood (7) given the reparameterization $v_i=\exp(\beta_i)$, $v_0=1$.

Notice that $\mathcal{L}_2(\boldsymbol{Z}|\boldsymbol{m};\boldsymbol{\beta})$ is independent of the nuisance parameters $\boldsymbol{\lambda}$, therefore, it follows from Andersen (1970) that the inference based on the conditional likelihood provides consistent and asymptotically normal estimators for the structural parameters $\{\beta_i\}_{i=1}^n$. Since $\mathcal{L}_{MNL}(\boldsymbol{\beta}) \propto \mathcal{L}_2(\boldsymbol{Z}|\boldsymbol{m};\boldsymbol{\beta})$ and $\mathcal{L}_{MNL}(\boldsymbol{\beta}) \propto \mathcal{L}_{I,\text{Profile}}(\boldsymbol{\beta})$, then inference from $\mathcal{L}_{MNL}(\boldsymbol{\beta})$ inherits the asymptotic properties of the estimators from both the conditional likelihood (i.e. consistent and asymptotically normal Andersen (1970)) and the profile likelihood (i.e efficient[4] Murphy and Van der Vaart (2000)). Finally, since our inference is based on a constrained market share likelihood, the efficiency result with respect to the constrained Cramer-Rao bound for the constrained MLE follows from Moore et al. (2008).

4.3. Arrival rates with assortment fixed effect

As mentioned before, both the non-homogeneous and the homogeneous Process processes cannot guarantee consistent estimates for the arrival rates. For the homogeneous arrival rate case, the estimation problem is hard and a globally optimal solution to the MLE problem is not guaranteed, whereas for the non-homogeneous arrival rate case, we cannot get a consistent estimate for the arrival rates due to the incidental parameters problem. For this reason, we adopt a middle ground that retains the global optimality of the MLE estimates in the case of a non-homogenous Poisson arrival process yet avoids the incidental parameters problem.

We start the analysis by assuming that the data generating process of the assortments is described by a probability mass function $p(S)$, for all $S \in \mathcal{S}$, where $\mathcal{S} \subset 2^{|\mathcal{N}|}$ the set of all observed assortments in the data. Instead of assuming a non-homogeneous arrival rate, we define a different rate for each assortment denoted by λ_S such that

$$\begin{aligned} &\lambda_{\mathcal{N}} = \lambda, \text{ and} \\ &\lambda_S = \lambda + \alpha_S, \qquad \text{for all } S \in \mathcal{S}. \end{aligned}$$

where $\alpha_S \in \mathbb{R}$ is the *assortment fixed effect* on the arrival rate (with $\lambda_S \in \mathbb{R}_+$). In words, λ is the baseline arrival rate when all the items are available, and α_S is the assortment fixed effect in the arrival rate. If the full assortment is not observed in the data then we can define a baseline arrival rate for a different assortment. Each assortment $S \in \mathcal{S}$ is observed for $\tau_S(T)$ time periods, which depends on the length of the selling horizon, such that $\lim_{T\to\infty}|\frac{\tau_S(T)}{T}| = p(S)$ for all S.

We now study the asymptotic properties of our estimators as T grows to infinity. In order, to keep our notation simple, we suppress the dependence of τ_S on T. We denote by z_{itS} the observed purchases of item i in period t given an offered assortment S. We also denote by $m_{tS} := \sum_{i=1}^N z_{itS}$ the overall observed purchases across time horizon when S is offered. Furthermore, let $m_S = \sum_{t=1}^{\tau_S} m_{tS}$ and $z_{iS} = \sum_{t=1}^{\tau_S} z_{itS}$ be, respectively, the overall purchases and item specific purchases over the selling horizon. The likelihood with assortment fixed effect can be written as

$$\begin{aligned}
\mathcal{L}_{FE}(\boldsymbol{v},\boldsymbol{\lambda}) &= \Pi_{S\in\mathcal{S}}\Pi_{t=1}^{\tau_S}\mathbb{P}\left(m_{tS} \text{ customers buy in period } t|\boldsymbol{v},\boldsymbol{\lambda}\right)\frac{1}{\Pi_{i\in S} z_{itS}!}\Pi_{i\in S}\left(\frac{v_i}{v_0+\sum_{j\in S} v_j}\right)^{z_{itS}} \\
&= \Pi_{S\in\mathcal{S}}\Pi_{t=1}^{\tau_S}\left[\frac{\lambda_S\sum_{i\in S} v_i}{v_0+\sum_{i\in S} v_i}\right]^{m_{tS}}\exp\left(-\frac{\lambda_S\sum_{i\in S} v_i}{v_0+\sum_{i\in S} v_i}\right)\frac{1}{\Pi_{i\in S} z_{itS}!}\Pi_{i\in S}\left(\frac{v_i}{v_0+\sum_{j\in S} v_j}\right)^{z_{itS}} \\
&= \Pi_{S\in\mathcal{S}}\left[\frac{\lambda_S\sum_{i\in S} v_i}{v_0+\sum_{i\in S} v_i}\right]^{m_S}\exp\left(-\frac{\lambda_S\sum_{i\in S} v_i}{v_0+\sum_{i\in S} v_i}\right)\Pi_{i\in S}\left(\frac{v_i}{v_0+\sum_{j\in S} v_j}\right)^{z_{iS}}\Pi_{t=1}^{\tau_s}\frac{1}{\Pi_{i\in S} z_{itS}!} \\
&\propto \Pi_{S\in\mathcal{S}}\left[\frac{\lambda_S\sum_{i\in S} v_i}{v_0+\sum_{i\in S} v_i}\right]^{m_S}\exp\left(-\frac{\lambda_S\sum_{i\in S} v_i}{v_0+\sum_{i\in S} v_i}\right)\Pi_{i\in S}\left(\frac{v_i}{v_0+\sum_{j\in S} v_j}\right)^{z_{iS}}. \qquad (10)
\end{aligned}$$

Notice that the likelihood problem in (10) has the same structure as the likelihood problem with non-homogeneous Poisson arrival process in (2). However, the main difference is that as $T \to \infty$, the overall number of arrival rates λ_S remains fixed and there is no incidental parameters problem. This implies that computing ML estimates with a fixed market share and a strongly connected graph $G(\mathcal{N}, E)$, produces consistent and asymptotically normal estimators for both $(\boldsymbol{\lambda}, \boldsymbol{\beta})$. Moreover, given non-negative estimates for each λ_S, we could recover the estimates of the assortment arrival fixed effects $\boldsymbol{\alpha}$, by solving a well defined ordinary system of linear equations.

5. Computing ML Estimates

Proposition 2 guarantees that any software package which can estimate the classical MNL model, could be used to solve the joint estimation problem (2) with restricted market share via a re-scale of the output to match the market share constraint. However, despite the fact that the reduced log-likelihood function $\ell_{MNL}(\boldsymbol{\beta})$ is a known concave maximization problem, most general purpose solvers have difficulty in solving this problem due to the presence of the "log-sum-exp" term in the objective function, which slows down the convergence behavior. For instance, CVX (a Matlab-based convex modeling framework) does not offer native support to functions like exp, log, and log-sum-exp. These functions are handled using a successive approximation method which makes multiple calls to the underlying solver.

On a related note, Hunter (2004) observes the inconvenience of applying general nonlinear methods such as Newton-Raphson to their Bradley-Terry model. Despite its convergence in relatively few iterations, it requires the computation and then the inversion of a square matrix at each iteration, operations that may be extremely time-consuming, particularly if the number of parameters is large. Furthermore, there is no guarantee that a Newton-Raphson step will increase the value of the objective function, so any well-designed Newton-Raphson algorithm must contain safeguards against erratic behavior.

This was also observed by Vulcano et al. (2012) in their Section 5, where the computational time achieved by the built-in MATLAB function *fminsearch* was two-orders of magnitude slower than EM. The function *fminsearch* implements the simplex search method of Lagarias et al. (1998), which is a direct search method that does not use numerical or analytic gradients. Yet, and despite its appealing simplicity, the EM algorithm has been criticized for its slow convergence behavior (e.g., see Newman et al. (2014) for an example of estimation of linear-in-parameters utility on hotel data).

For this reason, we next propose a Maximize-Minorize (MM) optimization algorithm that exploits our reduction from an inference based on $\ell_I(\boldsymbol{v}, \boldsymbol{\lambda})$ to an inference based on the reduced, classical MNL estimation of $\ell_{MNL}(\boldsymbol{\beta})$.

5.1. The generic MM approach

The name "MM algorithm" was first coined by Hunter and Lange (2000) and can be thought of as a general algorithm design technique. The broad idea behind MM algorithms is to define in each iteration a surrogate function that *minorizes* the original objective function. In particular, the surrogate minorizing function must be less than or equal to the original function in all the domain, and has to be equal at the current estimate point. The next step is to maximize the surrogate function and obtain a new estimate and continue iteratively.

More formally, suppose that the problem is to solve for $\arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$. Assume that our current estimate at iteration k is given by $\boldsymbol{\beta}^{(k)}$, we say that a function $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ minorizes $f(\cdot)$ at point $\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}$ if and only if

$$f(\boldsymbol{\beta}^{(k)}) = g(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}), \quad \text{ and } f(\boldsymbol{\beta}) \geq g\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}\right), \text{ for all } \ \boldsymbol{\beta}.$$

Assuming that maximizing the minorizing function $g(\cdot)$ is relatively simple, then we obtain our next iterate $\boldsymbol{\beta}^{(k+1)}$ by maximizing the minorizing function $g(\cdot)$ such that

$$\boldsymbol{\beta}^{(k+1)} \in \arg\max_{\boldsymbol{\beta}} g\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}\right)$$

The MM algorithm belongs to the family of ascent algorithms since by construction we have

$$f(\boldsymbol{\beta}^{(k+1)}) \geq g(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \geq g(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}) = f(\boldsymbol{\beta}^{(k)}),$$

where the first inequality follows from the fact that $g(\cdot)$ is a minorizor, the second inequality follows from the fact that $\boldsymbol{\beta}^{(k+1)}$ is a maximizer of $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$, and the last equality follows from the minorizing property of $g(\cdot)$ at $\boldsymbol{\beta}^{(k)}$.

The MM algorithm is indeed a general framework. Our contribution is to specialize it for the maximization of $\ell_I(\boldsymbol{\beta},\boldsymbol{\lambda})$ subject to the market share constraint, and show that it is straightforward to implement in any simple procedural language or numerical computing environment (e.g., Matlab) via iterations involving closed-form expressions.

5.2. A specific MM algorithm

The practical challenge of MM lies in finding a minorizing function $g(\cdot)$ that is relatively easy to maximize. In our case, the goal is to find a minorizing $g(\cdot)$ which is still concave but does not include a log-sum-exp term. In fact, in the case of $\ell_{MNL}(\boldsymbol{\beta})$, we can get a closed form solution for the maximizer of our proposed $g(\cdot)$ as we demonstrate next.

In order to find the minorizing function $g(\cdot)$ relative to $\ell_{MNL}(\boldsymbol{\beta})$, we will use the fact that $-\log(\cdot)$ is a convex differentiable function on $\mathbb{R}_{++}$. A first order approximation to the convex function $-\log(\cdot)$ for all $x, y \in \mathbb{R}_{++}$ gives

$$-\log(x) \geq -\log(y) - \frac{1}{y}(x-y) = 1 - \log(y) - \frac{x}{y}, \tag{11}$$

with equality at $x = y$. Using the definition of $\ell_{MNL}(\boldsymbol{\beta})$ in (8), we get

$$\begin{aligned} \ell_{MNL}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} K_j \beta_j - \sum_{t=1}^{T} m_t \log\left(\sum_{i \in S_t} \exp(\beta_i)\right) \\ &= \sum_{i=1}^{n} K_j \beta_j + \sum_{t=1}^{T} m_t \left(-\log\left(\sum_{i \in S_t} \exp(\beta_i)\right)\right). \end{aligned}$$

Plugging in (11) using $y = \boldsymbol{\beta}^{(k)}$ as the anchoring point, we obtain the minorizing function $g(\cdot)$ as follows:

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) := \sum_{j=1}^{N} K_j \beta_j + \sum_{t=1}^{T} m_t \left(1 - \log\left(\sum_{i \in S_t} \exp(\beta_i^{(k)})\right) - \frac{\sum_{i \in S_t} \exp(\beta_i)}{\sum_{i \in S_t} \exp(\beta_i^{(k)})}\right). \tag{12}$$

Recall, that the new iterates $\boldsymbol{\beta}^{(k+1)}$ in the MM-algorithm are obtained by solving for the maximizer $\boldsymbol{\beta}^{(\boldsymbol{k+1})} \in \arg\max_{\boldsymbol{\beta}} g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$. Notice that $\boldsymbol{\beta}^{(k)}$ is only a constant relative to the minorizing function $g(\cdot|\boldsymbol{\beta}^{(k)})$. Therefore, it is clear that $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ is strictly concave in $\boldsymbol{\beta}$. Consequently, the new iterates $\boldsymbol{\beta}^{(k+1)}$ are uniquely defined by the first order condition as follows

$$\begin{aligned} & K_l - \textstyle\sum_{t=1}^{T} m_t \mathbf{1}_{\{l \in S_t\}} \left(\frac{\exp(\beta_l^{(k+1)})}{\sum_{i \in S_t} \exp(\beta_i)^{(k)}}\right) && = 0 \\ \iff & K_l - \frac{\exp(\beta_l^{(k+1)})}{\exp(\beta_l^{(k)})} \textstyle\sum_{t=1}^{T} m_t \mathbf{1}_{\{l \in S_t\}} \left(\frac{\exp(\beta_l^{(k)})}{\sum_{i \in S_t} \exp(\beta_i)^{(k)}}\right) && = 0 \\ \iff & K_l - \exp(\beta_l^{(k+1)} - \beta_l^{(k)}) \textstyle\sum_{t=1}^{T} m_t q_{lt}(\boldsymbol{\beta}^{(k)}) && = 0, \end{aligned}$$

where $q_{lt}(\boldsymbol{\beta}^{(k)}) := \mathbf{1}_{\{l \in S_t\}} \left(\frac{\exp(\beta_l^{(k)})}{\sum_{i \in S_t} \exp(\beta_i^{(k)})}\right)$.

Thus, we get the following closed form solution for the maximizer $\boldsymbol{\beta}^{(k+1)}$:

$$\beta_l^{(k+1)} = \beta_l^{(k)} + \log\left(\frac{K_l}{\sum_{t=1}^{T} m_t q_{lt}(\boldsymbol{\beta}^{(k)})}\right). \tag{13}$$

Notice that $q_{lt}(\boldsymbol{\beta}^{(k)})$ represents the probability that item l is chosen from the assortment S_t given that the MNL specification is $\boldsymbol{\beta}^{(k)}$. Therefore, $\sum_{t=1}^{T} m_t q_{lt}(\boldsymbol{\beta}^{(k)})$ represents the expected number of times that item l will be chosen under the current estimate $\boldsymbol{\beta}^{(k)}$.

Define the following mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$H_l(\boldsymbol{\beta}) := \beta_l + \log\left(\frac{K_l}{\sum_{t=1}^{T} m_t q_{lt}(\boldsymbol{\beta})}\right), \qquad \text{for } l = 1, \ldots, n.$$

We can rewrite the expression in (13) more compactly as

$$\tilde{\boldsymbol{\beta}}^{(k+1)} = H(\boldsymbol{\beta}^{(k)}).$$

It is clear that $H(\boldsymbol{\beta}^{(k)})$ is the unique maximum of $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ for a given $\boldsymbol{\beta}^{(k)}$. However, $H(\boldsymbol{\beta}^{(k)})$ does not necessarily satisfy the market share restriction. For this reason, we further apply the transformation $T(H(\boldsymbol{\beta}^{(k)}), \tilde{s})$, for $T(\cdot, \tilde{s})$ defined in (9), to obtain a feasible solution to $L_R(\tilde{s})$.

We define our MM iteration as the composition M of two continuous mappings,

$$M := T \circ H,$$

and where at each iteration we have

$$\boldsymbol{\beta}^{(k+1)} = M\left(\boldsymbol{\beta}^{(k)}\right).$$

The MM algorithm for joint estimation of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ with a fixed market share s is summarized in Table 1.

Note that our MM algorithm falls outside the traditional framework of the MM algorithm described by Hunter (2004) for the following reasons: (i) the main restricted market share problem $L_R(\tilde{s})$ is a non-convex constrained problem, (ii) the minorizing function $g(\cdot|\boldsymbol{\beta}^{(k)})$ is only a minorizing function for the unrestricted market share problem, and (iii) the solution for $g(\cdot|\boldsymbol{\beta}^{(k)})$ may not satisfy the market share equation. Therefore, the general convergence results by Hunter (2004) do not hold immediately. For this reason, we use a direct approach in the following proposition to establish the main convergence result for our MM proposal.

Table 1 MM Algorithm for the joint estimation problem.

```
• Input: Transactional data $\{(\boldsymbol{z_t}, S_t)\}_{t=1}^{T}$, market share $s$
• Set $\tilde{s} \leftarrow s/(1-s)$
• Initialize $\boldsymbol{\beta}$ (e.g., $\beta_i = \log(1/n)$, for all $i \in \mathcal{N}$)
• While No Convergence on the $\boldsymbol{\beta}$
    ◦ $q_{it}(\boldsymbol{\beta}) \leftarrow \mathbf{1}_{\{i \in S_t\}} \left( \frac{\exp(\beta_i)}{\sum_{j \in S_t} \exp(\beta_j)} \right)$, for all $i \in \mathcal{N}, t = 1.....T$
    ◦ $\beta_i \leftarrow \beta_i + \log \left( \frac{K_i}{\sum_{t=1}^{T} m_t q_{it}(\boldsymbol{\beta})} \right)$, for $i \in \mathcal{N}$
    ◦ $\tilde{s}_1 \leftarrow \sum_{i=1}^{n} \exp(\beta_i)$
    ◦ $\beta_i \leftarrow \beta_i + \log \left( \frac{\tilde{s}}{\tilde{s}_1} \right)$, for $i \in \mathcal{N}$
• End While
• Set: $\lambda_t \leftarrow m_t \frac{1 + \sum_{i \in S_t} \exp(\beta_i)}{\sum_{i \in S_t} \exp(\beta_i)}$, for $t = 1, \ldots, T$
• End
```

Proposition 4. *Any limit point of the sequence $\{\boldsymbol{\beta}^{(k)}, k = 1, 2, \ldots\}$ generated by the MM algorithm in Table 1, starting from an arbitrary $\boldsymbol{\beta}^{(1)}$, is a stationary point of $\ell_{MNL}(\boldsymbol{\beta})$.*

In principle, the result does not guarantee the convergence of the algorithm to a stationary point. It states that *if* the method converges, it does so to a stationary point of the unconstrained log-likelihood function. Nevertheless, the convergence of the sequence of points $\{\boldsymbol{\beta}^{(k)}, k = 1, 2, \ldots\}$ can be checked numerically as part of the MM procedure. Moreover, if the underlying model is indeed identifiable, then the algorithm is indeed guaranteed to converge, as stated below.

Proposition 5. *If the associated graph $G(\mathcal{N}, E)$ is strongly connected, then, for a given market share $0 < s < 1$ and starting from an arbitrary $\boldsymbol{\beta}^{(1)}$, the MM algorithm in Table 1 is guaranteed to* monotonically *converge to the unique globally optimal solution of the incomplete log-likelihood problem $\ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda})$, subject to the market share constraint.*

Even though the focus of our exposition is on a utility model with alternative specific coefficients only, our MM algorithm allows to accommodate linear-in-parameters utilities. Section A2 in the Appendix provides the details.

6. Numerical Examples

The focus of our numerical experiments is the comparison between our MM proposal and the EM algorithm studied in Vulcano et al. (2012). We highlight here that Vulcano et al. (2012) provide evidence of the clear dominance of EM versus an alternative, standard nonlinear optimization method (in their reported numerics, the comparison was versus a Matlab built-in implementation of the simplex search method of Lagarias et al. (1998)) in terms of computational speed, whereas the quality of the estimates was similar. They also report the dominance of EM versus two standard uncensoring methods, positioning EM as a challenging benchmark.

In this section, we report on three different sets of numerical examples using synthetic data. In Section 6.1, we replicate the same example in Section 5.1.1 of Vulcano et al. (2012). In Section 6.2, we study the impact of the data volume and the market share inaccuracy on both the MM and the EM estimates, as in Section 5.1.2 therein. Finally, in Section 6.3 we provide an extensive computational comparison between the MM and the EM algorithms. Both algorithms were coded using the Matlab procedural language.

6.1. Preliminary estimation case

First, we replicate the preliminary estimation problem as outlined in Section 5.1.1 by Vulcano et al. (2012). We consider five different items and a selling horizon of fifteen time periods. The observable and non-observable data is summarized in Table 2.

In order to ensure a fair comparison between both algorithms, we slightly modify the stopping criterion of Vulcano et al. (2012). In particular, we define the stopping criterion for both algorithms based on the l_∞ norm of the difference between vectors $\mathbf{v}$ coming from two consecutive iterations, with a tolerance of $1e-4$.

In Table 3 we report the estimated MNL parameters and the percentage bias relative to the true parameters along with the log-likelihood, computational time, and iterations. In terms of the percentage bias, we notice that the EM algorithm achieves slightly lower bias whereas the MM algorithm converges to a higher log likelihood value within fewer iterations. It is worth mentioning that

Table 2 Purchases and no-purchases for the preliminary example.

Observed Data: Purchases and Nonavailability (NA)																
	Periods															
Products	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	Total
1	10	15	11	14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	50
2	11	6	11	8	20	16	NA	NA	NA	NA	NA	NA	NA	NA	NA	72
3	5	6	1	11	4	5	14	7	11	NA	NA	NA	NA	NA	NA	64
4	4	4	4	1	6	4	3	5	9	9	6	9	NA	NA	NA	64
5	0	2	0	0	1	0	1	3	0	3	3	5	2	3	3	26
Unobserved Data																
No Purchase	8	17	15	12	29	24	40	35	32	37	40	32	48	45	52	466
Arrivals	38	50	42	46	60	49	58	50	52	49	49	46	50	48	55	742

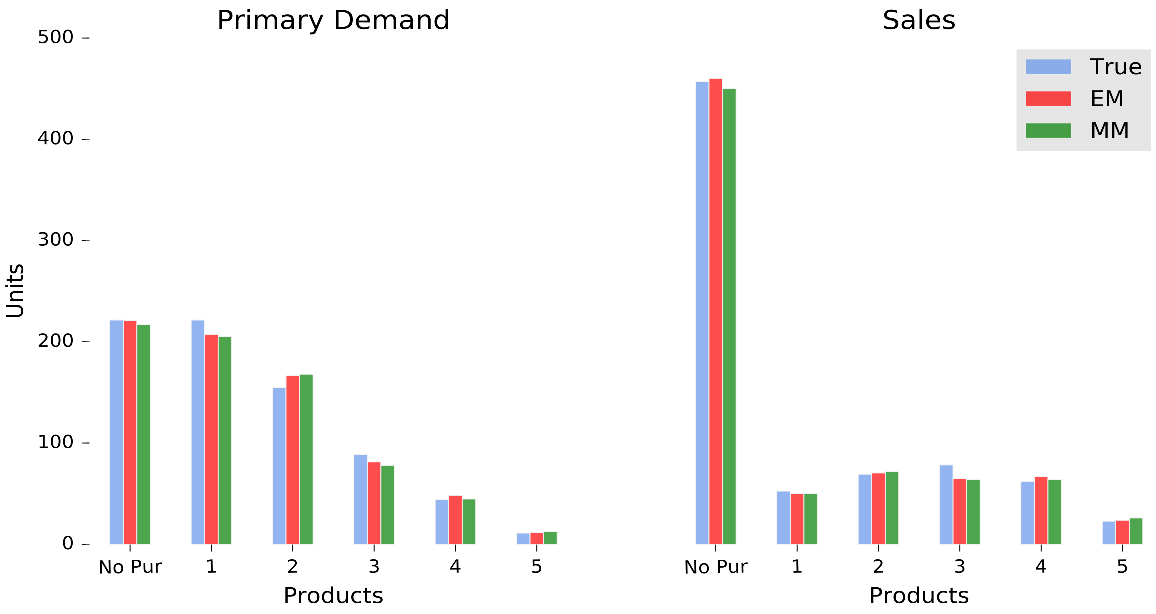
the MM algorithm converges to the same log-likelihood value achieved by the standard nonlinear algorithm that slightly dominated the EM algorithm as reported in Vulcano et al. (2012). However, we highlight the order of magnitude difference in the computational performance: the MM algorithm converges in 0.002 seconds and 12 iterations, compared to 0.033 seconds and 19 iterations of EM (which was in turn 200 times faster than the alternative nonlinear algorithm according to Vulcano et al. (2012)).

Finally, in Figure 1 we compare the predicted primary demand and sales under the true parameters $(\boldsymbol{v}_{true}, \boldsymbol{\lambda}_{true})$, and under the estimates obtained by the EM and MM algorithms. The primary (or first-choice) demand is defined as the demand that would have been observed if all products had been available in all periods. We notice that the performance of both algorithms is comparable in terms of receiving the primary demand and predicting the sales.

Table 3 Estimated Parameters under MM and EM

	True	EM		MM	
Parameters	v_{true}	v_{EM}	Bias (%)	v_{MM}	Bias (%)
v_1	1.00	0.94	-6.1%	0.94	-5.9%
v_2	0.70	0.76	7.9%	0.77	10.2%
v_3	0.40	0.37	-7.9%	0.36	-10.4%
v_4	0.20	0.22	9.6%	0.21	2.7%
v_5	0.05	0.05	3.3%	0.06	15.5%
$\ell_I(\boldsymbol{v}, \boldsymbol{\lambda})$	-96.73	-92.63		-92.38	
Time (sec)		0.033		0.002	
Iterations		19		12	

Figure 1 Estimated Primary Demand (left) and Sales (right) using the EM and the MM algorithm

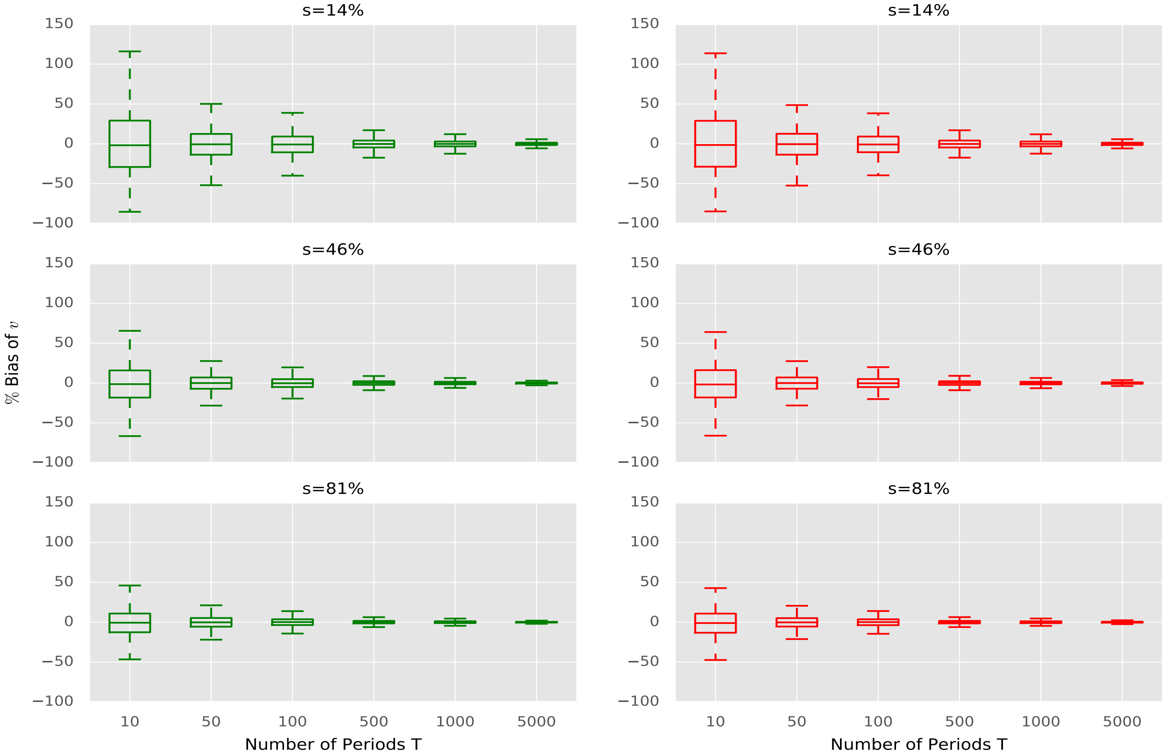


6.2. Impact of data volume and biases

We now study the impact of the data volume and inaccurate market share assumptions on the performance of the MM and the EM algorithms. Following Vulcano et al. (2012), we consider a scenario with 10 items where customers arrive in accordance with a Poisson process with average rate $\lambda = 50$. We set the probability of an item to be available at 0.7 and assume that it is independent across time periods.

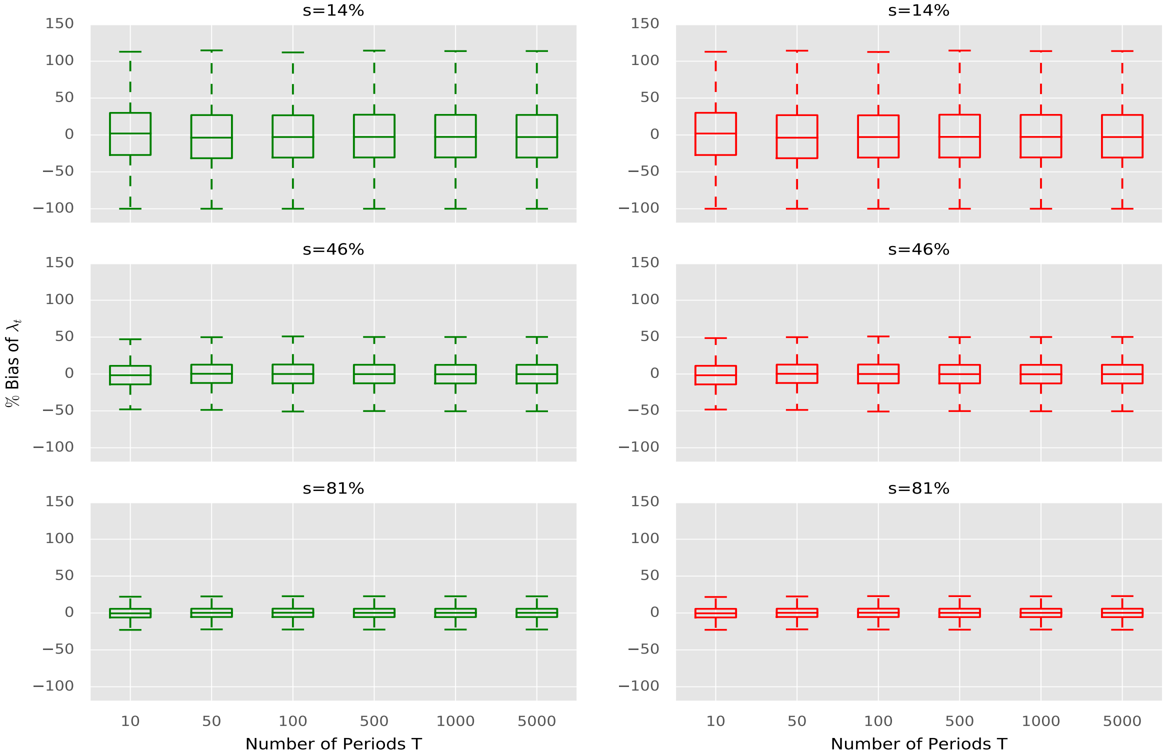
We start by studying the impact of data volume while assuming accurate information of the market share. We consider six different selling horizons $T \in \{10, 50, 100, 500, 1,000, 5,000\}$, and three different values for the market share $s \in \{14\%, 46\%, 81\%\}$. For each parameter combination, we simulated 100 instances of underlying, true demand. In Figure 2, we use box plots to summarize the bias in the estimated $\hat{\boldsymbol{v}}_{MM}$ and $\hat{\boldsymbol{v}}_{EM}$ compared to the true underlying parameters $\boldsymbol{v}_{true}$ for different market shares. The boxplots look similar to Figure 3 in Vulcano et al. (2012): As expected for our consistent estimators, for each market share scenario, as we increase the number of periods, the biases decrease.

Figure 2 Biases of the estimated weights $\hat{\boldsymbol{v}}$ using the MM (left) and the EM(right) algorithm under different market shares for different selling horizon lengths.



In Figure 3, we use box plots to summarize the bias in the estimated $\hat{\boldsymbol{\lambda}}_{MM}$ and $\hat{\boldsymbol{\lambda}}_{EM}$ compared to the true underlying parameters $\boldsymbol{\lambda}_{true}$ for different market shares. Here, increasing T has no effect on the bias, reflecting the incidental parameter problem by which we cannot get consistent estimators of λ_t. This is aligned with the observation in Section 5.1.2 of Vulcano et al. (2012) that points out that the bias of the arrival rate is independent of T.

Figure 3 Biases of the estimated arrival rates $\hat{\boldsymbol{\lambda}}$ using the MM (left) and the EM(right) algorithm under different market shares for different selling horizon lengths.

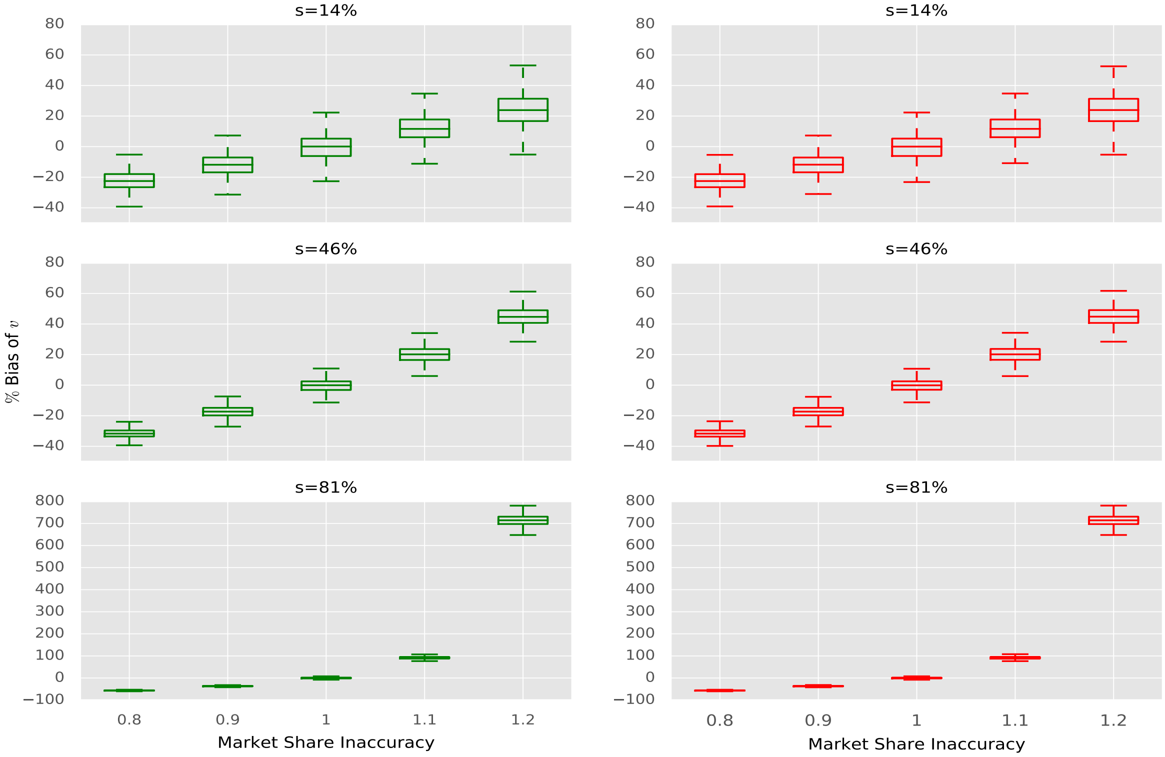


The market share information is a required input for our MM procedure. Of course, the user may be concerned about the impact of the accuracy of this metric. We next study the impact of such piece of information on the performance of both algorithms. We fix $T = 500$ and consider five different values of the market share inaccuracy defined by multiplicative factors $\rho \in \{0.8, 0.9, 1.0, 1.1, 1.2\}$. The case of accurate market information is represented by $\rho = 1$. For $\rho = 0.8$, the market share is underestimated by 20%, and for $\rho = 1.2$ the market share is overestimated by 20%.

We first plot the biases in the estimated value of $\boldsymbol{v}$ under different market share inaccuracies in Figure 4. Both algorithms have comparable performance and are relatively robust to market share inaccuracy for low ($s = 14\%$) or medium ($s = 46\%$) values. However, for a large market share

($s = 81\%$), we notice that both the MM and the EM algorithms produce highly biased estimates, which can scale up to 700% when the market share is overestimated by 20%.

Figure 4 Biases of the preference weights $\boldsymbol{v}$ using the MM (left) and the EM (right) algorithm under different market shares for different market share biases.



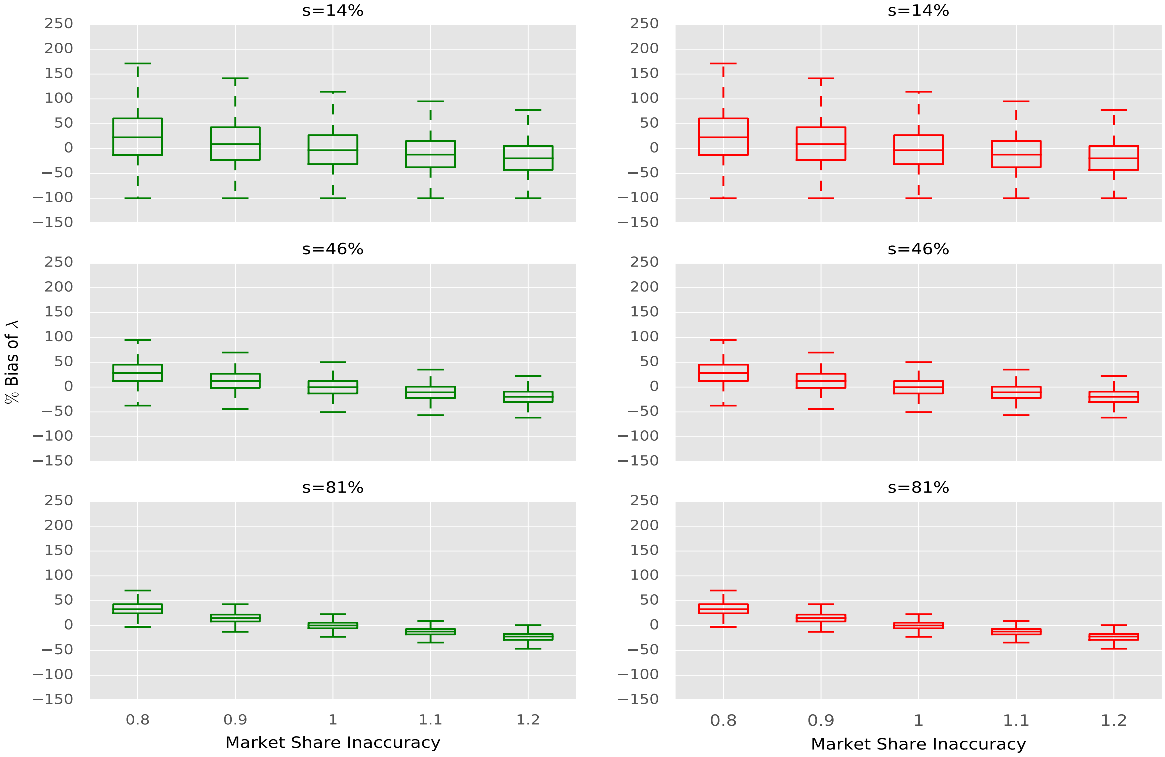
We also plot the biases in estimating the arrival rates λ_t in Figure 5. We notice that the estimates of the MM and EM algorithms are relatively robust in general, and very similar for both algorithms.

Finally, we study the impact of the market share inaccuracy on the biases in estimating the primary demand. Figure 6 summarizes the results. We notice that the estimates have significantly less bias compared to the estimates of $\boldsymbol{v}$ and $\boldsymbol{\lambda}$, indicating that the market share misspecification would impact more the preference values of the products than the primary demand estimates.

6.3. Computational comparison

We now compare the computational performance of the MM algorithm and the EM algorithm. We generated transaction data for $n \in \{20, 50, 100\}$, $T \in \{100, 1,000, 10,000, 50,000\}$, and three

Figure 5 Biases of the arrival rates λ_t using the MM (left) and the EM (right) algorithm under different market share misspecifications.

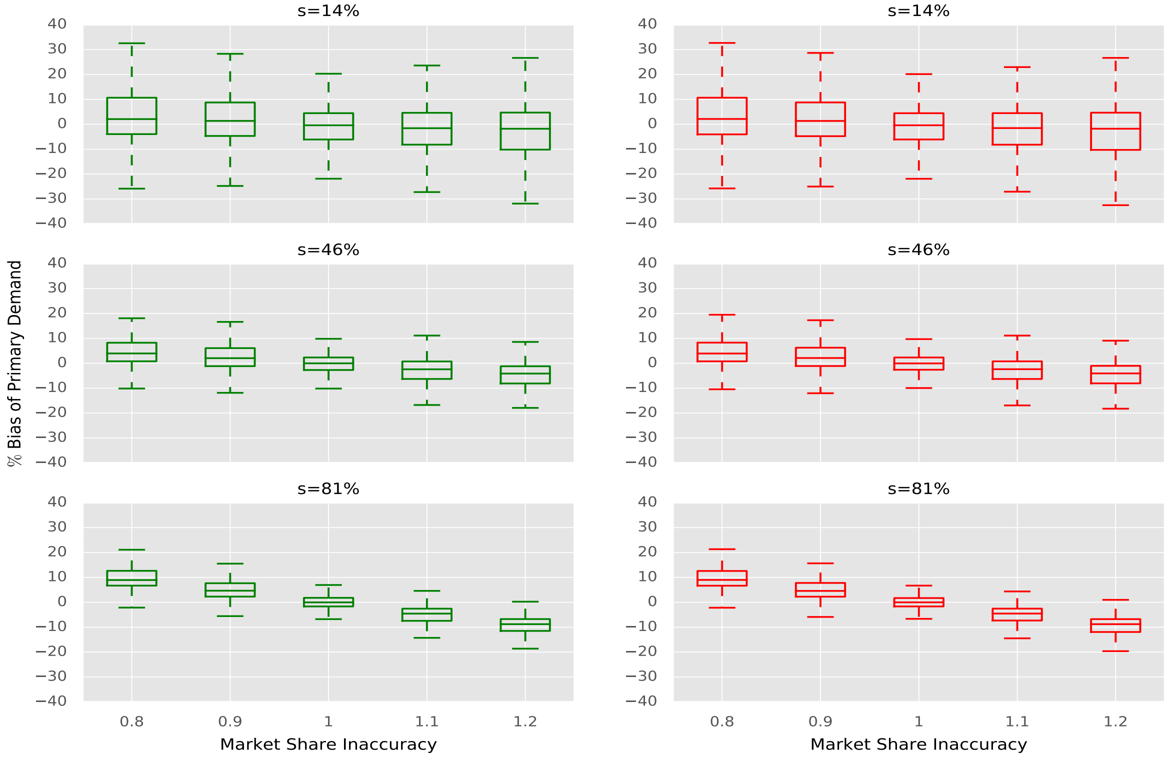


product availability settings, with $\delta \in \{0.5, 0.7, 0.9\}$, where δ denotes the probability of any item to be available in a period and is independent across items and periods. This parameterization leads to 36 different scenarios. We simulated 100 instances for each of those scenarios based on an underlying MNL demand model with preference weights $v_j \sim \text{Unif}[0.05, 1]$, $j = 1, \ldots, n$, $v_0 = 1$, and where the Poisson mean arrival rates are given by $\lambda_t \in \text{Unif}[10, 100]$.

For each of the 3,600 instances, we run both the MM and the EM algorithms using the same initialization of the $\mathbf{v}$ vectors as the fraction of observed demand for each product across the time horizon. To ensure a fair comparison, we set the same convergence criteria for both MM and EM on the l_∞ norm of the difference between $\mathbf{v}$s coming from two consecutive iterations, with a tolerance of $1e-4$.

We compare the performance of the MM and the EM algorithms in Figure 7 using box plots along four dimensions. The most striking difference is the computational efficiency of the MM algorithm compared to the pronounced growth in the computational time of the EM algorithm for

Figure 6 Biases of the primary demand using the MM (left) and the EM (right) algorithm under different market shares for different market share biases.



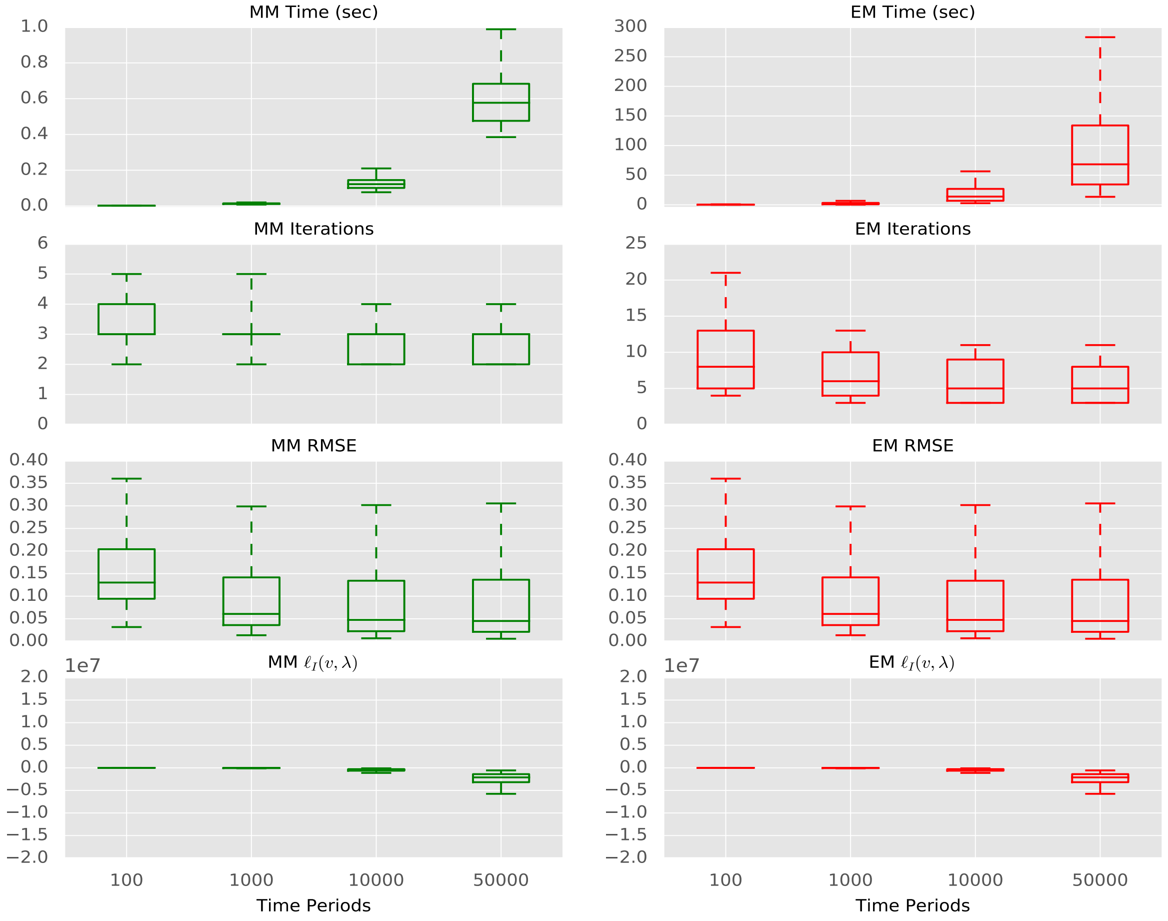
large instances. For $T = 50,000$, the difference could scale from a fraction of a second for MM to more than a minute for EM. The MM algorithm requires roughly half of the iterations of the EM algorithm. Nevertheless, we note here that in most of the cases both algorithms are converging to almost the same solution in terms of RMSE and log-likelihood values computed as in (3).

7. Conclusions

In this paper we revisit the problem of estimating the underlying preferences for substitutable products when customers choose in accordance with a multinomial logit (MNL) model and arrive following a nonhomogeneous Poisson process over multiple periods. It assumes realistic data: observed sales, product availability, and an aggregate estimate of the market share of the subcategory.

Our contribution is two-fold. From a theoretical perspective, we characterize necessary and sufficient conditions under which the demand model is identifiable under maximum likelihood criteria by representing the dataset with a directed graph and checking if it is strongly connected. We also discuss the conditions that guarantee the consistency of the estimates.

Figure 7 Computational comparison of the MM (left) and the EM (right) algorithms. Note that the scale in the right and left panels may be different for the same measure.



From a practical perspective, we propose a minorization-maximization (MM) algorithm to estimate the model parameters that runs remarkably fast by iterating through closed form expressions. Indeed, the algorithm is an order of magnitude faster than the competing EM procedure, which could lead to a dramatic performance difference in practical implementations (e.g., a major airline needs to estimate hundreds of thousands of origin-destination markets on a daily or even more frequent basis).

Endnotes

1. For instance, according to Pölt (1998), a 20% reduction of forecast error in the airline RM practice can translate into a 1% additional revenues.

2. For instance, Sabre Airline Solutions, one of the leading RM software providers for airlines, with a tradition of high quality R&D, implemented a proprietary version of the procedure described in Ratliff et al. (2008).

3. In their Section 3.4, Vulcano et al. (2012) discuss how to interpret s, and in their Section 6.1 discuss how to obtain values of s in the first place when such data is not available.

4. Murphy and Van der Vaart (2000) show that the inference based on the profile likelihood is efficient provided that it is consistent.

References

E. Andersen. Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 283–301, 1970.

O. Barndorff-Nielsen. On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343–365, 1983.

M. Ben-Akiva and S. Lerman. *Discrete Choice Analysis: Theory and Applications to Travel Demand*. The MIT Press, Cambridge, MA, sixth edition, 1994.

D. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–39, 1987.

J. Dai, W. Ding, A. Kleywegt, X. Wang, and Y. Zhang. Choice-based revenue management for parallel flights. Working paper, Georgia Institute of Technology, 2014.

W. Ding and A. Kleywegt. Estimation of arrival rates and choice model using censored data. In *INFORMS Revenue Management and Pricing Section Annual Meeting*. INFORMS, 2015.

W. Greene. *Econometric Analysis*. Pearson, seventh edition, 2011.

P. Guadagni and J. Little. A logit model of brand choice calibrated on scanner data. *Marketing Science*, 2: 203–238, 1983.

W. Heiser. Convergent computation by iterative majorization. In W. Krzanowski, editor, *Recent Advances in Descriptive Multivariate Analysis*, pages 157–189. Oxford University Press, 1995.

D. Hunter. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, pages 384–406, 2004.

D. R Hunter and K. Lange. Rejoinder. *Journal of Computational and Graphical Statistics*, 9(1):52–59, 2000.

G. Kok, M. Fisher, and R. Vaidyanathan. Assortment planning: Review of literature and industry practice. In N. Agrawal and S. Smith, editors, *Retail Supply Chain Management*, International Series in Operations Research and Management Science, chapter 8, pages 99–153. Springer, 2008.

J. Lagarias, J. Reeds, M. Wright, and P. Wright. Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1):112–147, 1998.

K. Lange, D. Hunter, and I. Yang. Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, 9:1–59, 2000.

Q. Liu and G. van Ryzin. On the choice-based linear programming model for network revenue management. *M&SOM*, 10(2):288–310, 2008.

T. Moore, B. Sadler, and R. Kozick. Maximum-likelihood estimation, the cramér–rao bound, and the method of scoring with parameter constraints. *Signal Processing, IEEE Transactions on*, 56(3):895–908, 2008.

S. Murphy and A. Van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95 (450):449–465, 2000.

A. Musalem, M. Olivares, E. Bradlow, C. Terwiesch, and D. Corsten. Structural estimation of the effect of out-of-stocks. *Management Science*, 56(7):1180–1197, 2010.

J. Newman, M. Ferguson, L. Garrow, and T. Jacobs. Estimation of choice-based models using sales data from a single firm. *Manufacturing & Service Operations Management*, 16(2):184–197, 2014.

J. Neyman and E. Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

W. Nicholson. *Microeconomic Theory: Basic Principles and Extensions.* The Dryden Press, 5th edition, 1992.

S. Pölt. Forecasting is difficult –especially if it refers to the future. In *Revenue Management and Distribution Study Group Annual Meeting*, Melbourne, Australia, May 1998. AGIFORS.

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

S. Subramanian and P. Harsha. An integrated estimation method for predicting demand and lost sales in consumer choice models. In *INFORMS Revenue Management and Pricing Section Annual Meeting*. INFORMS, 2015.

K. Talluri. A finite-population revenue management model and a risk-ratio procedure for the joint estimation of population size and parameters. Working paper, Imperial College London, 2009.

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

K. Train. *Discrete choice methods with simulation*. Cambridge University Press, New York, NY, 2003.

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

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

CF Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

Demand Estimation under the Multinomial Logit Model from Sales Transaction Data

Appendix

Tarek Abdallah

NYU Stern School of Business, New York, tabdalla@stern.nyu.edu

Gustavo Vulcano

NYU Stern School of Business, New York, gvulcano@stern.nyu.edu

School of Business, Torcuato di Tella University, Buenos Aires, Argentina

A1. Proofs of Technical Results

Proof of Proposition 1. First, note that λ_t^* is a well defined interior point since from (4), $\ell_I(\boldsymbol{v},\boldsymbol{\lambda}) \to -\infty$ as $\lambda_t \to 0$, for any t. Using the first order condition of the log-likelihood function (5) with respect to λ_t we get

$$\frac{m_t}{\lambda_t} - \frac{\sum_{i\in S_t} v_i}{v_0 + \sum_{j\in S_t} v_j} = 0,$$

which implies that $\lambda_t^*(\boldsymbol{v}) = m_t \frac{v_0+\sum_{j\in S_t} v_j}{\sum_{j\in S_t} v_j}$.

Substituting $\lambda_t^*(\boldsymbol{v})$ back into (5) we get,

$$\ell_I(\boldsymbol{v},\boldsymbol{\lambda}^*(\boldsymbol{v})) := \underbrace{\sum_{t=1}^{T} m_t(\log m_t - 1)}_{\text{Constant}} + \ell_{MNL}(\boldsymbol{v}),$$

where

$$\ell_{MNL}(\boldsymbol{v}) := \sum_{j=1}^{n} K_j \log v_j - \sum_{t=1}^{T} m_t \log\left(\sum_{i\in S_t} v_i\right).$$

Let $\boldsymbol{v}^*$ be a maximizer of $\ell_{MNL}(\boldsymbol{v})$. Then, $(\boldsymbol{v}^*, \lambda^*(\tilde{v}^*))$ is a solution to (5) and taking the exponential of $\ell_{MNL}(\boldsymbol{v})$, we get

$$\mathcal{L}_{MNL}(\boldsymbol{v}) = \left(\Pi_{j\in S_t} v_j^{K_j}\right)\left(\Pi_{t=1}^{T} \frac{1}{\left(\sum_{i\in S_t} v_i\right)^{m_t}}\right) = \left(\Pi_{j\in S_t} v_j^{\sum_{t=1}^{T} z_{jt}}\right)\left(\Pi_{t=1}^{T} \frac{1}{\left(\sum_{i\in S_t} v_i\right)^{\sum_{j\in S_t} z_{jt}}}\right)$$
$$= \Pi_{t=1}^{T} \Pi_{j\in S_t} \left(\frac{v_j}{\sum_{i\in S_t} v_i}\right)^{z_{jt}}.$$

□

Proof of Lemma 1. The reduced log-likelihood function verifies

$$\begin{aligned}\ell_{MNL}(T(\boldsymbol{\beta})) &= \sum_{j=1}^{n} K_j(\alpha+\beta_j) - \sum_{t=1}^{T} m_t \log\left(\sum_{i\in S_t}\exp(\alpha+\beta_i)\right)\\ &= \sum_{j=1}^{n} K_j\alpha + \sum_{j=1}^{n} K_j(\beta_j) - \sum_{t=1}^{T} m_t\alpha - \sum_{t=1}^{T} m_t\log\left(\sum_{i\in S_t}\exp(\beta_i)\right)\\ &= \sum_{j=1}^{n} K_j\beta_j - \sum_{t=1}^{T} m_t\log\left(\sum_{i\in S_t}\exp(\beta_i)\right) = \ell_{MNL}(\boldsymbol{\beta}).\end{aligned}$$

The last equality is due to the fact that $\sum_{j=1}^{n} K_j = \sum_{t=1}^{T} m_t$, which are the total number of purchases in the data. □

Proof of Proposition 2. Assume that the oracle returns an optimal solution to the unrestricted problem $\boldsymbol{\beta}^* \in \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \ell_{MNL}(\boldsymbol{\beta})$. Let $c_1 := \sum_{i=1}^{n}\exp(\beta_i)$. Then, using Lemma 1, we have that $\ell_{MNL}(T(\boldsymbol{\beta}^*,\tilde{s})) = \ell_{MNL}(\boldsymbol{\beta}^*)$. However, $T(\boldsymbol{\beta}^*,\tilde{s})$ is feasible for the restricted market share problem. In particular,

$$\sum_{i=1}^{N}\exp\left(\log\left(\frac{\tilde{s}}{c_1}\right)+\beta_i^*\right) = \frac{\tilde{s}}{c_1}\sum_{i=1}^{N}\exp(\beta_i^*) = \tilde{s},$$

Therefore, $T(\boldsymbol{\beta}^*,\tilde{s})$ is an optimal solution for the restricted market share problem. □

Proof of Lemma 2. $((i)\Rightarrow(ii))$ Let $\boldsymbol{\beta}\in\mathbb{R}^n$ be the unique bounded optimal solution of $L_R(\tilde{s})$. Consider two bounded solutions $\boldsymbol{\gamma},\boldsymbol{\xi}\in\Psi$. Define the following re-scaled solutions

$$\tilde{\boldsymbol{\gamma}} = \log\left(\frac{\tilde{s}}{\tilde{s}_\gamma}\right)+\boldsymbol{\gamma}, \quad \text{and} \quad \tilde{\boldsymbol{\xi}} = \log\left(\frac{\tilde{s}}{\tilde{s}_\xi}\right)+\boldsymbol{\xi}. \tag{A1}$$

Since $\boldsymbol{\gamma},\boldsymbol{\xi}$ are well defined, then we have that $0<\tilde{s}_\gamma,\tilde{s}_\xi<\infty$. Consequently, both $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\xi}}$ are also well defined. In addition, we have

$$\tilde{s}_{\tilde{\gamma}} = \sum_{i=1}^{n}\exp(\gamma_i) = \sum_{i=1}^{n}\exp\left(\log\left(\frac{\tilde{s}}{\tilde{s}_\gamma}\right)+\gamma_i\right) = \frac{\tilde{s}}{\tilde{s}_\gamma}\sum_{i=1}^{n}\exp(\gamma_i) = \tilde{s},$$

and likewise we have $\tilde{s}_{\tilde{\xi}} = \tilde{s}$. Hence, both $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\xi}}$ are feasible solutions for the restricted problem $L_R(\tilde{s})$. We can now establish the following chain of equalities:

$$\ell_{MNL}(\tilde{\boldsymbol{\gamma}}) = \ell_{MNL}(\boldsymbol{\gamma}) = \ell_{MNL}(\boldsymbol{\xi}) = \ell_{MNL}(\tilde{\boldsymbol{\xi}}),$$

where the first and third equalities follow from Lemma 1, and the second one from the fact that both $\boldsymbol{\gamma}, \boldsymbol{\xi} \in \Psi$. Then,

$$\ell_{MNL}(\tilde{\boldsymbol{\gamma}}) = \ell_{MNL}(\tilde{\boldsymbol{\xi}}) = L \geq L_R(\tilde{s}),$$

which jointly with the fact that $\tilde{s}_{\tilde{\gamma}} = \tilde{s}_{\tilde{\xi}} = \tilde{s}$ guarantee the equality $L = L_R(\tilde{s})$. From the uniqueness of $\boldsymbol{\beta}^*$, we get $\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\xi}} = 0$. Rearranging equation (A1):

$$\boldsymbol{\xi} - \boldsymbol{\gamma} = \log\left(\frac{\tilde{s}}{\tilde{s}_{\gamma}}\right) - \log\left(\frac{\tilde{s}}{\tilde{s}_{\xi}}\right) = \log\left(\frac{\tilde{s}_{\xi}}{\tilde{s}_{\gamma}}\right),$$

establishes the result.

($(ii) \Rightarrow (i)$) Let $\boldsymbol{\gamma} \in \Psi$ be a bounded optimal solution to the unrestricted market share problem. Note that by the assumption, such a solution exists.

Define the solution $\boldsymbol{\alpha} := T(\boldsymbol{\gamma}, \tilde{s})$ as described in (9). It follows from Proposition 2, that $\boldsymbol{\alpha}$ is a bounded optimal solution to $L_R(\tilde{s})$. Hence, the set of well defined solutions to $L_R(\tilde{s})$ is non-empty. We next show that $\boldsymbol{\alpha}$ is the only bounded optimal solution.

Let $\boldsymbol{\beta} \in \mathbb{R}^n$ be a bounded optimal solution to $L_R(\tilde{s})$, and define the following re-scaled solutions

$$\tilde{\boldsymbol{\alpha}} = \log\left(\frac{\tilde{s}_{\gamma}}{\tilde{s}}\right) + \boldsymbol{\alpha}, \quad \text{and} \quad \tilde{\boldsymbol{\beta}} = \log\left(\frac{\tilde{s}_{\gamma}}{\tilde{s}}\right) + \boldsymbol{\beta}.$$

Therefore, $\tilde{s}_{\tilde{\alpha}} = \tilde{s}_{\tilde{\beta}} = \tilde{s}_{\gamma}$. In addition, from Lemma 1, we have that $\ell_{MNL}(\tilde{\boldsymbol{\alpha}}) = \ell_{MNL}(\tilde{\boldsymbol{\beta}})$. Define

$$\tilde{\boldsymbol{\gamma}} = \log\left(\frac{\tilde{s}}{\tilde{s}_{\gamma}}\right) + \boldsymbol{\gamma}, \quad \text{so that} \quad \tilde{s}_{\tilde{\gamma}} = \tilde{s}.$$

This is a feasible solution of $L_R(\tilde{s})$ with $\ell_{MNL}(\tilde{\boldsymbol{\gamma}}) = \ell_{MNL}(\boldsymbol{\gamma}) = L$ (from Lemma 1). Then, any optimal solution of $L_R(\tilde{s})$ must have log-likelihood value L, and in particular,

$$\ell_{MNL}(\tilde{\boldsymbol{\alpha}}) = \ell_{MNL}(\boldsymbol{\alpha}) = \ell_{MNL}(\boldsymbol{\beta}) = \ell_{MNL}(\tilde{\boldsymbol{\beta}}) = L.$$

The first and third equalities follow from Lemma 1, the second one from the fact that both $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are optimal solutions of $L_R(\tilde{s})$, and the last one from the fact that $L = L_R(\tilde{s})$. Hence, $\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}} \in \Psi$ and, by the hypothesis, they must verify

$$\tilde{\alpha}_i - \tilde{\beta}_i = \log\left(\frac{\tilde{s}_{\tilde{\alpha}}}{\tilde{s}_{\tilde{\beta}}}\right) = 0, \qquad \text{for all } i \in \mathcal{N},$$

since $\tilde{s}_{\tilde{\alpha}} = \tilde{s}_{\tilde{\beta}}$. This leads to $\alpha_i = \beta_i$ for all $i \in \mathcal{N}$, establishing the uniqueness of the solution to $L_R(\tilde{s})$.

□

Proof of Proposition 3. We show a list of implications.

((i)⇒(ii)) We first show that letting $G(\mathcal{N},E)$ be a strongly connected graph ensures that Ψ is non-empty, i.e., there is a bounded optimal solution. We study the case where $||\beta|| \to \infty$.

Consider an arbitrary item $q \in \mathcal{N}$, and define the following partition of $\mathcal{N}$:

$$\mathcal{K}_q := \{i \in \mathcal{N} : \lim_{||\boldsymbol{\beta}|| \to \infty} |\beta_i - \beta_q| < +\infty\},$$

$$\overline{\mathcal{K}}_q := \{i \in \mathcal{N} : \lim_{||\boldsymbol{\beta}|| \to \infty} \beta_i - \beta_q = +\infty\}$$

$$\underline{\mathcal{K}}_q := \{i \in \mathcal{N} : \lim_{||\boldsymbol{\beta}|| \to \infty} \beta_i - \beta_q = -\infty\}$$

For simplicity we assume that the limits are well defined, else we can use the same arguments using any subsequence with a defined limit in the extended reals. We argue that if $\overline{\mathcal{K}}_q \cup \underline{\mathcal{K}}_q \neq \emptyset$ then $\ell_{MNL}(\boldsymbol{\beta}) \to -\infty$. We consider two different cases:

Case 1 ($\overline{\mathcal{K}}_q \neq \emptyset$): Since $G(\mathcal{N},E)$ is strongly connected, then there exists an edge that goes from $\mathcal{K}_q$ to $\overline{\mathcal{K}}_q$. Consider the set S_t which defines that edge. We have that $S_t \cap \mathcal{K}_q \neq \emptyset$, $S_t \cap \overline{\mathcal{K}}_q \neq \emptyset$, and $z_{lt} \geq 1$ for some item $l \in S_t \cap \mathcal{K}_q$. Therefore, we get

$$\begin{aligned}
\lim_{||\boldsymbol{\beta}|| \to \infty} \ell_{MNL}(\boldsymbol{\beta}) &\leq \lim_{||\boldsymbol{\beta}|| \to \infty} \log\left(\frac{\exp(\beta_l)}{\sum_{i \in S_t \cap \mathcal{K}_q} \exp(\beta_i) + \sum_{j \in S_t \cap \overline{\mathcal{K}}_q} \exp(\beta_j) + \sum_{k \in S_t \cap \underline{\mathcal{K}}_q} \exp(\beta_k)}\right) \\
&= \lim_{||\boldsymbol{\beta}|| \to \infty} \log\left(\frac{\exp(\beta_l - \beta_q)}{\sum_{i \in S_t \cap \mathcal{K}_q} \exp(\beta_i - \beta_q) + \sum_{j \in S_t \cap \overline{\mathcal{K}}_q} \exp(\beta_j - \beta_q) + \sum_{k \in S_t \cap \underline{\mathcal{K}}_q} \exp(\beta_k - \beta_q)}\right) \\
&= -\infty,
\end{aligned}$$

where the inequality holds because the term in parenthesis is just one term of the likelihood function (2), the first equality follows from multiplying numerator and denominator by $\exp(\beta_q)$, and the last one is due to the fact that $\sum_{j \in S_t \cap \overline{\mathcal{K}}_q} \exp(\beta_j - \beta_q) \to +\infty$, $\sum_{j \in S_t \cap \underline{\mathcal{K}}_q} \exp(\beta_j - \beta_q) \to 0$, and the other terms are positive and finite.

Case 2 ($\underline{\mathcal{K}}_q \neq \emptyset$ and $\overline{\mathcal{K}}_q = \emptyset$): Following with the same line of argument above, since $G(\mathcal{N},E)$ is strongly connected, then there exists S_t such that $S_t \cap \mathcal{K}_q \neq \emptyset$, $S_t \cap \underline{\mathcal{K}}_q \neq \emptyset$ where $z_{mt} \geq 1$ for some item $m \in S_t \cap \underline{\mathcal{K}}_q$. Therefore, we have

$$\lim_{||\boldsymbol{\beta}|| \to \infty} \ell_{MNL}(\boldsymbol{\beta}) \leq \lim_{||\boldsymbol{\beta}|| \to \infty} \log\left(\frac{\exp(\beta_m)}{\sum_{i \in S_t \cap \mathcal{K}_q} \exp(\beta_i) + \sum_{k \in S_t \cap \underline{\mathcal{K}}_q} \exp(\beta_k)}\right)$$

$$= \lim_{||\boldsymbol{\beta}|| \to \infty} \log \left(\frac{\exp(\beta_m - \beta_q)}{\sum_{i \in S_t \cap \mathcal{K}_q} \exp(\beta_i - \beta_q) + \sum_{k \in S_t \cap \underline{\mathcal{K}}_q} \exp(\beta_k - \beta_q)} \right)$$

$$= -\infty,$$

where the last equality is due to the fact that $\exp(\beta_m - \beta_q) \to 0$ while the other terms are positive and finite.

Therefore, any optimal solution such that $||\boldsymbol{\beta}|| \to \infty$ implies that $\mathcal{K}_q = \mathcal{N}$. However, any log-likelihood value of such solution can be replicated by a bounded solution. Therefore, the corresponding set of optimal solutions Ψ of the unconstrained ℓ_{MNL} is non-empty.

Now, for any $\gamma \in \Psi$, we have that $T(\boldsymbol{\gamma}, \tilde{s}) \in \Psi$. Hence, $T(\boldsymbol{\gamma}, \tilde{s}) \in \Psi$ is a stationary point of $\ell_{MNL}(.)$ where $\nabla \ell_{MNL}(T(\boldsymbol{\gamma}, \tilde{s})) = 0$. Since $T(\boldsymbol{\gamma}, \tilde{s})$ is feasible to the market share restriction, then it is an optimal solution to $L_R(\tilde{s})$. To finish the proof, we are left to show that $L_R(\tilde{s})$ has a unique solution.

In order to do so, we next show that any two solutions in Ψ satisfy the characterization in Lemma 2(ii). Consider any two well-defined optimal solutions $\boldsymbol{\gamma}, \boldsymbol{\xi} \in \Psi$. Notice that by the concavity of $\ell_{MNL}(.)$, any convex combination of $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ is also optimal. Fix $0 < \alpha < 1$, then we have

$$\ell_{MNL}\left(\alpha \boldsymbol{\gamma} + (1-\alpha)\boldsymbol{\xi}\right) = \ell_{MNL}(\boldsymbol{\gamma}) = \ell_{MNL}(\boldsymbol{\xi}) = \alpha \ell_{MNL}(\boldsymbol{\gamma}) + (1-\alpha)\ell_{MNL}(\boldsymbol{\xi}). \tag{A2}$$

At the same time,

$$\begin{aligned}
&\ell_{MNL}(\alpha \boldsymbol{\gamma} + (1-\alpha)\boldsymbol{\xi}) \\
&= \alpha \sum_{j=1}^{n} K_j \gamma_j + (1-\alpha) \sum_{j=1}^{n} K_j \xi_j - \sum_{t=1}^{T} m_t \log \left[\sum_{i \in S_t} \exp(\gamma_i)^{\alpha} \exp(\xi_i)^{(1-\alpha)} \right] \\
&\geq \alpha \sum_{j=1}^{n} K_j \gamma_j + (1-\alpha) \sum_{j=1}^{n} K_j \xi_j - \sum_{t=1}^{T} m_t \log \left[\left(\sum_{i \in S_t} \exp(\gamma_i) \right)^{\alpha} \left(\sum_{i \in S_t} \exp(\xi_i) \right)^{(1-\alpha)} \right] \qquad \text{(A3)} \\
&= \alpha \sum_{j=1}^{n} K_j \gamma_j - \alpha \sum_{t=1}^{T} m_t \log \left(\sum_{i \in S_t} \exp(\gamma_i) \right) + (1-\alpha) \sum_{j=1}^{n} K_j \xi_j - (1-\alpha) \sum_{t=1}^{T} m_t \log \left(\sum_{i \in S_t} \exp(\xi_i) \right) \\
&= \alpha \ell_{MNL}(\boldsymbol{\gamma}) + (1-\alpha)\ell_{MNL}(\boldsymbol{\xi}).
\end{aligned}$$

Inequality (A3) is due to Holder's inequality, which states:

$$\sum_{i \in S} x_i y_i \leq \left(\sum_{i \in S} x_i^p \right)^{1/p} \left(\sum_{i \in S} y_i^q \right)^{1/q}, \quad \text{with} \quad p, q \in (1, \infty), \;\; p + q = 1.$$

In our case, we define $x_i = \exp(\gamma_i)^{\lambda}$, and $y_i = \exp(\xi_i)^{1-\lambda}$, with $p = 1/\lambda$ and $q = 1/(1-\lambda)$. Moreover, in view of (A2), the inequality (A3) must hold with equality, which occurs if and only if there exists $\delta \in \mathbb{R}$ such that $\exp(\gamma_i) = \delta \exp(\xi_i)$ for all $i \in \mathcal{N}$. Therefore,

$$\delta = \frac{\sum_{i=1}^{n} \exp(\gamma_i)}{\sum_{i=1}^{n} \exp(\xi_i)} = \frac{\tilde{s}_\gamma}{\tilde{s}_\xi}.$$

This implies that $\gamma_i - \xi_i = \log\left(\frac{\tilde{s}_\gamma}{\tilde{s}_\xi}\right)$ for all $i \in \mathcal{N}$. Therefore, by Lemma 2, $L_R(\tilde{s})$ has a unique solution.

((ii)⇒(i)) We will show the equivalent implication ($\neg$(ii)⇒ $\neg$(i)). Suppose that $G(\mathcal{N}, E)$ is not strongly connected. Then there exist two nodes i and j such that there is no directed path from i to j. Let V_i denote the set of nodes that can be reached from i, and define $V_i^c = \mathcal{N} \setminus V_i$. Then, we have that $i \in V_i$ and $j \in V_i^c$, and by construction there is no directed edge from V_i to V_i^c. We now consider two possible cases regarding the existence or not of a directed edge from V_i^c to V_i.

Case 1 (No directed edge from V_i^c to V_i). Suppose there is no directed edge from V_i^c to V_i but there exists a unique optimal solution $\boldsymbol{\beta}^*$ to the restricted market share problem $L_R(\tilde{s})$. We will now construct a solution to the unrestricted problem L that violates Lemma 2 and hence violates the uniqueness assumption of $\boldsymbol{\beta}^*$.

Define a new solution $\tilde{\boldsymbol{\beta}}$ such that

$$\tilde{\beta}_l = \begin{cases} \beta_l^* & \text{if } l \in V_i, \\ c + \beta_l^* & \text{if } l \in V_i^c, \end{cases}$$

for any $c \in \mathbb{R}$, $c \neq 0$.

Since V_i and V_i^c are disconnected, then, by construction, for every t, we have $S_t \subseteq V_i$ or $S_t \subseteq V_i^c$. Consequently, we have

$$\begin{aligned} \frac{\mathcal{L}_{MNL}\left(\boldsymbol{\beta}^*\right)}{\mathcal{L}_{MNL}\left(\tilde{\boldsymbol{\beta}}\right)} &= \frac{\Pi_{t=1}^{T} \Pi_{k\in S_t} \left(\frac{\exp(\beta_k^*)}{\sum_{l\in S_t} \exp(\beta_l^*)}\right)^{z_{kt}}}{\Pi_{t=1}^{T} \Pi_{k\in S_t} \left(\frac{\exp(\tilde{\beta}_k)}{\sum_{l\in S_t} \exp(\tilde{\beta}_l)}\right)^{z_{kt}}} \\ &= \Pi_{t=1, S_t \subseteq V_i^c}^{T} \Pi_{k\in S_t} \left(\frac{\frac{\exp(\beta_k^*)}{\sum_{l\in S_t} \exp(\beta_l^*)}}{\frac{\exp(c+\beta_k^*)}{\sum_{l\in S_t} \exp(c+\beta_l^*)}}\right)^{z_{kt}} = 1. \end{aligned}$$

Therefore, both $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ are optimal solutions to the unrestricted problem. Hence, $\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^* \in \Psi$. However, we have

$$\tilde{\beta}_l - \beta_l^* = \begin{cases} 0 & \text{if } l \in V_i, \\ c & \text{if } l \in V_i^c, \end{cases}$$

Therefore, by Lemma 2, $\boldsymbol{\beta^*}$ is not unique which is a contradiction.

Case 2 (Directed edge from V_i^c to V_i) Assume there exists a directed edge (h,k). In this case, there exists S_t such that $h \in (S_t \cap V_i^c)$ and $k \in (S_t \cap V_i)$. However, by construction, for all such S_t we have $z_{kt} = 0$ for all $k \in S_t \cap V_i$. Otherwise, there would be a directed edge from V_i to V_i^c.

Define a new solution $\tilde{\boldsymbol{\beta}}$ such that

$$\tilde{\beta}_l = \begin{cases} \beta_l^* & l \in V_i \\ c + \beta_l^* & l \in V_i^c \end{cases}$$

for any $c > 0$. Therefore we have,

$$\begin{aligned} \frac{\mathcal{L}_{MNL}\left(\boldsymbol{\beta}^*\right)}{\mathcal{L}_{MNL}\left(\tilde{\boldsymbol{\beta}}\right)} &= \frac{\Pi_{t=1}^T \Pi_{k \in S_t} \left(\frac{\exp(\beta_k^*)}{\sum_{l \in S_t} \exp(\beta_l^*)}\right)^{z_{kt}}}{\Pi_{t=1}^T \Pi_{k \in S_t} \left(\frac{\exp(\tilde{\beta}_k)}{\sum_{l \in S_t} \exp(\tilde{\beta}_l)}\right)^{z_{kt}}} \\ &= \Pi^T_{\substack{t=1: S_t \cap V_i^c \neq \phi \\ S_t \cap V_i \neq \phi}} \Pi_{k \in S_t \cap V_i^c} \left(\frac{\frac{\exp(\beta_k^*)}{\sum_{l \in S_t} \exp(\beta_l^*)}}{\frac{\exp(c+\beta_k^*)}{\sum_{m \in S_t \cap V_i} \exp(\beta_m^*) + \sum_{l \in S_t \cap V_i^c} \exp(c+\beta_l^*)}}\right)^{z_{kt}} < 1, \end{aligned}$$

where the last strict inequality is due to the fact that the denominator is strictly increasing in c. Thus, $\boldsymbol{\beta}^*$ is not optimal which is a contradiction.

((ii) $\Leftrightarrow$ (iii)) Any optimal solution $(\boldsymbol{\beta}^*, \boldsymbol{\lambda}^*)$ for the log-likelihood problem $\ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda})$ with restricted market share should satisfy the KKT conditions given by

$$\begin{aligned} \frac{\partial \ell_I(\boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \beta_l} &= \pi \exp(\beta_l), \qquad \text{for all } l \in \mathcal{N}, \\ \lambda_t^* &= m_t \frac{1 + \sum_{i \in S_t} \exp(\beta_i^*)}{\sum_{j \in S_t} \exp(\beta_j^*)} \qquad \text{for all } t = 1, \ldots, T. \\ \sum_{i=1}^n \exp(\beta_i^*) &= \tilde{s}, \end{aligned}$$

for some $\pi \in \mathbb{R}$ that is the Lagrange multiplier of the market share constraint.

Moreover, the partial derivative with respect to β_l is given by

$$\begin{aligned}\frac{\partial \ell_I(\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \beta_l} &= \sum_{t=1}^{T} \mathbb{1}_{\{l\in S_t\}}\left[-m_t \frac{\exp(\beta_l)}{1+\sum_{j\in S_t}\exp(\beta_j)} - \lambda_t \frac{\exp(\beta_l)}{\left(1+\sum_{j\in S_t}\exp(\beta_j)\right)^2}\right] + K_l \\ &= \sum_{t=1}^{T} \mathbb{1}_{\{l\in S_t\}}\left[-m_t \frac{\exp(\beta_l)}{1+\sum_{j\in S_t}\exp(\beta_j)}\left(1+\frac{\lambda_t}{m_t\left(1+\sum_{j\in S_t}\exp(\beta_j)\right)}\right)\right] + K_l.\end{aligned}$$

Evaluating the partial derivative at $(\boldsymbol{\beta}^*,\boldsymbol{\lambda}^*)$ where from Proposition 1, $\lambda_t^* = m_t \frac{1+\sum_{j\in S_t}\exp(\beta_j^*)}{\sum_{j\in S_t}\exp(\beta_j^*)}$, we get

$$\begin{aligned}\frac{\partial \ell_I(\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \beta_l}\Big|_{(\boldsymbol{\beta}^*,\boldsymbol{\lambda}^*)} &= \sum_{t=1}^{T} \mathbb{1}_{\{l\in S_t\}}\left[-m_t \frac{\exp(\beta_l^*)}{1+\sum_{j\in S_t}\exp(\beta_j^*)}\left(1+\frac{1}{\sum_{j\in S_t}\exp(\beta_j^*)}\right)\right] + K_l \\ &= \sum_{t=1}^{T} \mathbb{1}_{\{l\in S_t\}}\left[-m_t \frac{\exp(\beta_l^*)}{\sum_{j\in S_t}\exp(\beta_j^*)}\right] + K_l,\end{aligned}$$

where the last equality is the same as the partial derivative of $\ell_{MNL}(\boldsymbol{\beta})$. As a result, the KKT conditions of $\ell_I(\boldsymbol{\beta},\boldsymbol{\lambda})$ reduce to

$$\begin{aligned}\frac{\partial \ell_{MNL}(\boldsymbol{\beta}^*)}{\partial \beta_l} &= \pi \exp(\beta_l^*), \qquad \text{for all } l\in\mathcal{N}, \\ \lambda_t^* &= m_t \frac{1+\sum_{i\in S_t}\exp(\beta_i^*)}{\sum_{j\in S_t}\exp(\beta_j^*)} \qquad \text{for all } t=1,\dots,T, \\ \sum_{i=1}^{n}\exp(\beta_i^*) &= \tilde{s},\end{aligned}$$

for some $\pi\in\mathbb{R}$. These are the same KKT conditions for $\ell_{MNL}(\beta^*)$ with the added condition on λ_t^*. Therefore, both problems share the same stationary points with respect to $\boldsymbol{\beta}$. Moreover, since $\mathcal{L}_I(\beta,\boldsymbol{\lambda}^*)\propto\mathcal{L}_{MNL}(\boldsymbol{\beta})$, then their respective log-likelihoods share the same set of maximizers, $\boldsymbol{\beta}^*$, which completes the proof. □

Proof of Proposition 4. If $\boldsymbol{\beta}^* = \lim_{i\to\infty}\boldsymbol{\beta}^{(k_i)}$ for a subsequence $\boldsymbol{\beta}^{(k_1)},\boldsymbol{\beta}^{(k_2)},\dots$, then the result is obtained by taking the limit in

$$\ell_{MNL}(\boldsymbol{\beta}^{(k_i)}) = g(\boldsymbol{\beta}^{(k_i)}|\boldsymbol{\beta}^{(k_i)}) \le g(H(\boldsymbol{\beta}^{(k_i)})|\boldsymbol{\beta}^{(k_i)}) \le \ell_{MNL}(H(\boldsymbol{\beta}^{(k_i)})) = \ell_{MNL}(M(\boldsymbol{\beta}^{(k_i)})) = \ell_{MNL}(\boldsymbol{\beta}^{(k_{i+1})}).$$

The first equality follows from $g(\cdot|\boldsymbol{\beta}^{(k_i)})$ being the minorizer of $\ell_{MNL}(\cdot)$ with equality at $\boldsymbol{\beta}^{(k_i)}$. The maximizer of $g(\cdot|\boldsymbol{\beta}^{(k_i)})$ is $H(\boldsymbol{\beta}^{(k_i)})$, determining the next inequality. Next, $g(\cdot|\boldsymbol{\beta}^{(k_i)})$ is again a minorizer of $\ell_{MNL}(\cdot)$, whereas the second to last equality holds from Lemma 1. Finally, the mapping $M(\boldsymbol{\beta}^{(k_i)}) = \boldsymbol{\beta}^{(k_{i+1})}$ closes the chain.

The fact that $\ell_{MNL}(M(\boldsymbol{\beta})) = \ell_{MNL}(\boldsymbol{\beta})$ implies that $\boldsymbol{\beta}$ is a stationary point since the differentiable minorizing function is tangent to the log-likelihood function at the current iterate. □

Proof. of Proposition 5. Consider a sequence of MM iterations $\{\boldsymbol{\beta}^{(k)}\}$ re-scaled via $M(\cdot)$ that are feasible for the restricted market share problem $L_R(\tilde{s})$. Since $G(\mathcal{N}, E)$ is strongly connected and $0 < s < 1$, it follows from Proposition 3 that $L_R(\tilde{s})$ has a unique optimal solution which is a unique stationary point. From Proposition 4, we know that if there is a convergence point for the algorithm, in this case it must be the unique stationary point. So we are left to show that the algorithm indeed converges.

We first show an auxiliary result:

LEMMA A1. *Given a strongly connected graph $G(\mathcal{N}, E)$ and $0 < \tilde{s} < \infty$, then, for every $c \in \mathbb{R}$, the set $\mathcal{A}(c) := \{\boldsymbol{\beta} : \ell_{MNL}(\boldsymbol{\beta}) \geq c, \sum_{i=1}^{n} \exp(\beta_i) = \tilde{s}\}$ is compact.*

Proof of Lemma A1 We have that

$$\mathcal{A}(c) = \{\boldsymbol{\beta} : \ell_{MNL}(\boldsymbol{\beta}) \geq c\} \cap \{\boldsymbol{\beta} : \sum_{i=1}^{n} \exp(\beta_i) \geq \tilde{s}\} \cap \{\boldsymbol{\beta} : \sum_{i=1}^{n} \exp(\beta_i) \leq \tilde{s}\}.$$

By the continuity of $\ell_{MNL}(\boldsymbol{\beta})$ and $\exp(\beta_i)$, we have that, for every $c \in \mathbb{R}$, the three sets on the right hand side are closed and so is their intersection. Hence, we are left to show that $\mathcal{A}(c)$ is bounded for any $c \in \mathbb{R}$. Assume for a contradiction that it is not, then there exists $\tilde{c} \in \mathbb{R}$ such that $\mathcal{A}(\tilde{c})$ is unbounded. Consequently, there exists a sequence $\{\boldsymbol{\beta}_i : i = 1, 2, \ldots\} \in \mathcal{A}(\tilde{c})$ with $||\boldsymbol{\beta}|| \to \infty$. However, we show that such a sequence cannot exist. We consider three different cases:

Case 1 ($\beta_i \to +\infty$ for some $i \in \mathcal{N}$): It is clear that such a sequence does not belong to $\mathcal{A}(\tilde{c})$ since it violates the market share constraint. In particular, we have

$$\lim_{||\boldsymbol{\beta}|| \to +\infty} \sum_{i=1}^{n} \exp(\beta_i) = +\infty \neq \tilde{s}.$$

Case 2 ($\beta_i \to -\infty$ for all $i \in \mathcal{N}$): Again this sequence violates the market share constraint where

$$\lim_{\beta_i \to -\infty} \sum_{i=1}^{n} \exp(\beta_i) = 0 \neq \tilde{s}.$$

We are left to eliminate the case when only a subset of the elements have $\beta \to -\infty$. Notice that such cases are not ruled out by Case 1 and 2. For this reason, define a new set $\underline{\mathcal{K}} \subsetneq \mathcal{N}$ where

$$\underline{\mathcal{K}} := \{i \in \mathcal{N} : \beta_i \to -\infty\}.$$

Case 3 ($\underline{\mathcal{K}} \neq \emptyset$): Let $\mathcal{K} := \mathcal{N} \backslash \underline{\mathcal{K}}$ where $\mathcal{K} \neq \emptyset$. Since $G(\mathcal{N}, E)$ is strongly connected, then there exists an edge that goes from $\mathcal{K}$ to $\underline{\mathcal{K}}$. Consider the set S_t that defines that edge. We have that $S_t \cap \mathcal{K} \neq \emptyset$, $S_t \cap \underline{\mathcal{K}} \neq \emptyset$, and $z_{lt} \geq 1$ for some item $l \in S_t \cap \underline{\mathcal{K}}$. Therefore, we get

$$\lim_{||\boldsymbol{\beta}|| \to \infty} \ell_{MNL}(\boldsymbol{\beta}) \leq \lim_{||\boldsymbol{\beta}|| \to \infty} \log \left(\frac{\exp(\beta_l)}{\sum_{i \in S_t \cap \mathcal{K}} \exp(\beta_i) + \sum_{k \in S_t \cap \underline{\mathcal{K}}} \exp(\beta_k)} \right) = -\infty,$$

where the inequality holds because the term in parenthesis is just one term of the likelihood function (2) and the equality is due to the fact that $\lim_{||\boldsymbol{\beta}|| \to \infty} \sum_{i \in S_t \cap \mathcal{K}} \exp(\beta_i) > 0$ and $\exp(\beta_l) \to 0$.

Hence, any sequence for which $||\boldsymbol{\beta}|| \to \infty$, cannot belong to $\mathcal{A}(\tilde{c})$ which is a contradiction. Therefore, for every $c \in \mathbb{R}$, $\mathcal{A}(\tilde{c})$ is a closed and bounded set in $\mathbb{R}^n$ and hence compact. □

From Lemma A1 the set $\mathcal{A}(c) := \{\boldsymbol{\beta} : \ell_{MNL}(\boldsymbol{\beta}) \geq c, \sum_{i=1}^n \exp(\beta_i) = \tilde{s}\}$ is compact for every $c \in \mathbb{R}$. Therefore, by construction, the sequence of MM iterations $\{\boldsymbol{\beta}^{(k)}\}$ is bounded and has at least one convergent subsequence.

Now consider any convergent subsequence $\{\boldsymbol{\beta}^{(k_l)}\}$ and denote its limit point by $\mathcal{B}$. We have that the sequence of likelihood functions $\{\ell_{MNL}(\boldsymbol{\beta}^{(k_l)})\}$ is non-decreasing and bounded from above, hence it converges. In particular, we have

$$\lim_{l \to \infty} \ell_{MNL}(M(\boldsymbol{\beta}^{(k_l)})) = \lim_{l \to \infty} \ell_{MNL}(\boldsymbol{\beta}^{(k_l)}),$$

and by the continuity of $\ell_{MNL}(.)$ and $M(.)$ we get

$$\ell_{MNL}(M(\mathcal{B})) = \ell_{MNL}(\mathcal{B}).$$

As a result, $\mathcal{B}$ is a stationary point for ℓ_{MNL} and is feasible to $L_R(\tilde{s})$. Therefore, $\mathcal{B}$ is the unique optimal solution for $L_R(\tilde{s})$. Finally, since the subsequence $\{\boldsymbol{\beta}^{(k_l)}\}$ was chosen arbitrarily, then the whole sequence $\{\boldsymbol{\beta}^{(k)}\}$ converges monotonically to the unique optimal solution. Finally, the fact that the estimates are optimal solutions to $\ell_I(\boldsymbol{\lambda}, \boldsymbol{\beta})$ follows from Proposition 1. □

Table A1 MM Algorithm for the joint estimation problem with covariates.

```
• **Input**: Transactional data $\{(\boldsymbol{z_t}, S_t)\}_{t=1}^{T}$, market share $s$
• Set $\tilde{s} \leftarrow s/(1-s)$
• Initialize $\boldsymbol{\beta}$ with nonzero values
• **While** No Convergence on the $\boldsymbol{\beta}$
    ○ $\beta_{i0} \leftarrow \beta_{i0} + \log\left(\frac{K_i}{\sum_{t=1}^{T} m_t \tilde{q}_{it}(\boldsymbol{\beta})}\right)$, for $i \in \mathcal{N}$
    ○ $\beta_k \leftarrow \beta_k + \log\left(\frac{\sum_{t=1}^{T}\sum_{i\in S_t} z_{it}X_{ikt}}{\sum_{t=1}^{T}\sum_{i\in S_t} z_{it}X_{ikt}\tilde{q}_{kt}(\boldsymbol{\beta})}\right)$, for $k = 1, \ldots, K$
    ○ $\tilde{s}_1 \leftarrow \sum_{i=1}^{n} \exp(\beta_{i0} + \sum_{k=1}^{K} \beta_k X_{ikt})$
    ○ $\beta_{i0} \leftarrow \beta_{i0} + \log\left(\frac{\tilde{s}}{\tilde{s}_1}\right)$, for $i \in \mathcal{N}$
• **End While**
• Set: $\lambda_t \leftarrow m_t \frac{1+\sum_{i\in S_t} \exp(\beta_{i0}+\sum_{k=1}^{k} \beta_k X_{ikt})}{\sum_{i\in S_t} \exp(\beta_{i0}+\sum_{k=1}^{K} \beta_k X_{ikt})}$, for $t = 1, \ldots, T$
• **End**
```

A2. Extension: Linear-in-parameters utility

Assume that there are $K \geq 1$ observable product attributes relevant for the utility specification. We capture unobservable features for product i as the intercept β_{i0}. Then, the utility for product $i \in \mathcal{N}$ in period t can be written as

$$u_i = \beta_{i0} + \sum_{k=1}^{K} \beta_k X_{ikt} + \epsilon_{it}, \quad \text{and} \quad u_0 = \epsilon_{0t},$$

where u_0 is the utility of the no-purchase option whose deterministic part is normalized to 0, and ϵ_{it} are i.i.d draws from the standard Gumbel distribution. This model allows features such as price to be dynamic (i.e., the features can change from period to period).

We define $\tilde{q}_{it}(\boldsymbol{\beta}) := \mathbf{1}_{\{i\in S_t\}} \frac{\beta_{i0}+\exp(\sum_{k=1}^{K} \beta_k X_{ikt})}{\sum_{j\in S_t} \exp(\beta_{j0}+\sum_{k=1}^{K} \beta_k X_{jkt})}$. In the presence of dynamic features, the required market share information is an aggregate, average measure of the subcategory in the market. The MM algorithm for the joint estimation $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is summarized in Table A1.

Notice that the iteration map for the product specific intercepts are the same as the map for the MM algorithm given in Table 1. The only change in the MM algorithm is related to the update in the coefficients of the product attributes. The iteration in this case tries to match the observed and the predicted market share for each feature.