A Locational Demand Model for Bike-Sharing

Ang Xu
Department of Industrial and Systems Engineering, University of Washington, Seattle, WA 98105

Chiwei Yan
Department of Industrial and Systems Engineering, University of Washington, Seattle, WA 98105

Chong Yang Goh
Uber Technologies, Inc., San Francisco, CA 94158*

Patrick Jaillet
Department of Electrical Engineering and Computer Science and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139

Micro-mobility systems (bike-sharing or scooter-sharing) have been widely adopted across the globe as a sustainable mode of urban transportation. To efficiently plan and operate such systems, it is crucial to understand the underlying rider demand — where riders come from and the rates of arrivals into the service area. Estimating rider demand is nontrivial as most systems only keep track of trip data which is a biased representation of the underlying demand. In this paper, we develop a locational demand model to estimate rider demand based on only trip and vehicle location and status data. We establish conditions under which our estimators are identifiable and consistent. In addition, we devise an expectation-maximization (EM) algorithm with closed-form updates for efficient estimation. To scale the estimation procedures, this EM algorithm is complemented with a location-discovery procedure that gradually adds new locations in the service region with largest improvements to the log-likelihood. Experiments using both synthetic data and real data from a dockless bike-sharing system in the Seattle area demonstrate the accuracy and scalability of the model and its estimation algorithm.

Key words: locational demand model, bike-sharing, expectation-maximization, location discovery.
History: This version, 05/2023.

1. Introduction

Bike-sharing services have been widely adopted to provide a sustainable mode of urban transportation. In the United States, as of July 2022, there are 61 docked bike-sharing systems operating 8,473 docking stations (Bureau of Transportation Statistics 2023). Perhaps more excitingly, since its debut in Seattle in 2017, dockless bike-sharing and e-scooter systems have been quickly expanding coverage and gaining popularity due to its increased accessibility. As of July 2022, dockless bike-sharing systems serve 35 cities and e-scooters serve 158 cities in the United States (Bureau of Transportation Statistics 2023).

* Formerly affiliated with Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, where the co-author's research was conducted.

One critical aspect of monitoring and operating bike-sharing systems is to understand the rider demand and how accessible the current service is to different communities in the service region. For example, the city of Seattle has been actively monitoring the usage and accessibility of these services using trip data (see Seattle Department of Transportation 2022 for a comprehensive online dashboard). In addition, these systems often experience supply and demand imbalances that require careful bike allocation and rebalancing. The success of these operations crucially depends on the operator's ability to estimate the ridership demand accurately over the planning horizon. These motivate the topic of our paper.

Demand estimation in bike-sharing systems is non-trivial due to the following difficulties. First, the operator often does not know the exact location of riders but only observes the booking data which indicates the location and time a bike is reserved and picked up. Although technically speaking, rider location data can be accessed through GPS on riders' phones, to the best of the authors' knowledge, it is not an industry standard to record these data (Open Mobility Foundation 2022). These rider location data can be harder to collect and is subject to errors themselves — for example, the place rider is checking her phone can be different from where she departs to pick up a bike. Second, the observed trip locations and trip counts can form biased estimates for the underlying demand — the booking location is likely not the actual rider location and a rider may not choose to book or switch to other transportation services if there are no bikes available in the vicinity of her location, and in such a case this demand will be censored in the observed trip data.

A standard approach for modeling such demand is to use a choice model, which specifies the likelihood that a customer makes a certain choice when presented with a set of alternatives. Analogous to the retail settings to which such models are commonly applied, we can view each bike in the system as a product to be considered among a set of available bikes. One characteristic of a bike-sharing service is that products are horizontally differentiated — they are the same in terms of price and quality and are largely differentiated by their proximity to the rider.

We primarily focus on this locational feature to build our demand model which is applicable to both docked and dockless systems. In particular, we consider a model where riders arrive within a set of locations inside the service region according to a Poisson process. Upon arrival, a rider makes a choice of which bike to pick (or leaves the system) based on her walking distance to all available bikes, governed by a *general* choice model. Our goal is to estimate the set of rider locations and their corresponding arrival rates, using only booking and vehicle location and status data.

We summarize our key contributions as follows. We first study the statistical properties of the demand model. Assuming that the rider choice model is known, we give conditions under which the maximum likelihood estimator (MLE) of the location weights is identifiable and consistent. We also discuss identifiability properties when riders make bike choices according to specific choice models

including a multinomial logit model and a model based on distance ranking, under both docked and dockless settings. Second, we derive an efficient expectation-maximization (EM) algorithm with closed-form updates. To scale this algorithm and handle situations when the set of potential rider locations is large or even not known a priori, we develop a location-discovery procedure that iteratively explores and adds new rider locations in the service region. Lastly, we implement our algorithms on a set of synthetic data and real dockless bike-sharing data in the Seattle area. These experiments demonstrate the scalability and accuracy of our proposed demand model and its estimation algorithm.

Our paper is organized as follows. In Section 2, we discuss relevant literature. In Section 3, we describe the data generative process, our demand model, and discuss the identifiability and consistency of the MLE. In Section 4, we introduce the EM algorithm and the location-discovery procedure. We report the performance of our demand model and its estimation algorithm on an extensive set of numerical experiments in Section 5 and conclude in Section 6. Data and code to reproduce all experiments in the paper can be found at `https://github.com/angxu1/bike_sharing`.

2. Literature Review

In this section, we briefly review related work. There have been numerous works that analyze bike-sharing usage by incorporating data from heterogeneous sources (see, e.g., Rixey 2013, Singhvi et al. 2015, El-Assi et al. 2017), among them being demographic characteristics, built environment factors, weather and usage data of other connecting transportation services (see El-Assi et al. 2017 for a summary of the recent literature and references therein). In contrast, our work is more related to the literature of structurally understanding demand censoring and substitution due to the service (un)availability of bike-sharing systems. Close to our work are O'Mahony and Shmoys (2015), Mellou and Jaillet (2019), Kabra et al. (2019), Freund et al. (2019) and He et al. (2021). Although the main focuses of O'Mahony and Shmoys (2015), Freund et al. (2019) and Mellou and Jaillet (2019) are on improving the operations of bike allocation and transshipment, they all emphasize the importance of demand correction. Specifically, O'Mahony and Shmoys (2015) and Freund et al. (2019) filter out time periods when the station runs out of bikes to correct for demand censoring. Kabra et al. (2019) focus primarily on the question of how the accessibility and the availability of a docked bike-sharing service impact ridership. To that end, the authors proposed a structural demand model in which the rider arrival rate at a location is assumed to vary with several covariates, such as the local population density and metro usage. The pick-up model is constructed using a logit choice model where the utility is a piecewise linear function of walking distance with a break-point at 300 meters. He et al. (2021) study the network effect of bike-sharing demand,

i.e., riders choose to pick up a bike because both the origin and destination stations are attractive. They developed an instrumental variable method to tackle the endogeneity issue of choice set in estimating demand and applied the method to a London bike share system to estimate the rider demand for network products. Our paper differs from these work as we primarily focus on the locational aspect where the riders' choices are based on the distance between their origin location and bike locations. We look at a parsimonious and operational setting where the rider demands are inferred solely based on trip and vehicle status data that fleet operators collect in their daily operations. We reveal statistical properties of estimating rider arrival location and intensities and develop scalable estimation procedures that are applicable to both docked and dockless systems. These add to the toolkit of fleet operators or municipal agencies to effectively monitor, operate and plan micro-mobility systems.

Our work is also closely related to the abundant list of literature on estimating the demand of substitutable products using choice models based on possibly censored transaction data (Talluri and van Ryzin 2004, Vulcano et al. 2012, Newman et al. 2014, Abdallah and Vulcano 2020). Assuming customer demand follows a multinomial logit (MNL) choice model, Vulcano et al. (2012) propose an expectation-maximization (EM) method to estimate customers' demand for substitutable products from (censored) sales transaction data. Newman et al. (2014) develop a two-step strategy that is served as an alternative to the EM method for parameter estimation. Their approach entails breaking down the log-likelihood function into separate marginal and conditional components. To improve upon these two methods, Abdallah and Vulcano (2020) propose a minorization-maximization (MM) algorithm that can achieve a unique global maximum of the log-likelihood function under some data requirements of the transaction data. However, it is less efficient in our setting due to a lack of closed-form updates. van Ryzin and Vulcano (2014) propose a market discovery algorithm to estimate customer demand based on ranking preferences. The algorithm starts with a small set of customer types and enlarges the set by iteratively generating new customer type (ranking preference) that increase the likelihood value. We draw inspiration from this algorithm when developing our location-discovery procedure, though the underlying generative process and the structure of the subproblem to generate new customer type (rider location in our context) are vastly different.

Structurally speaking, our demand model and the EM algorithm share some similarities with the latent-class logit (LCL) model discussed in Bhat (1997) and Greene and Hensher (2003). A latent class in our setting corresponds to a particular rider location, but unlike LCL, we do not restrict rider choice behaviors to follow MNL models. Another important distinguishing factor of our model is that the features of different alternatives (walking distances to different bikes) depend

on the latent class (rider location), which is assumed to be invariant across latent classes in LCL. This renders existing identifiability results (e.g., Grün and Leisch 2008) or enhanced estimation algorithms (e.g., Jagabathula et al. 2020) developed for LCL not applicable in our setting.

3. Model and Preliminaries

In this section, we introduce our model and preliminaries. We first discuss our data generative process and describe the observed data from an operator's perspective. Then we derive the likelihood function of our statistical model. We conclude the section by discussing various identifiability and consistency results regarding our estimators.

Generative Process. We consider a set of rider locations $\mathcal{L} := \{1, \cdots, L\}$ distributed over a bounded space $\mathcal{P} \subset \mathbb{R}^2$ where potential riders come from. We assume that riders arrive at the area according to a Poisson process with a total rate λ. For each arriving rider, its arrival location follows a multinomial distribution with probabilities $\boldsymbol{w} = \{w_1, \cdots, w_L\}$ satisfying $\sum_{l\in\mathcal{L}} w_l = 1$. One can think of this arrival process as one that is dedicated to a particular hour of day or week. We consider a total length of arrival period T. We define $\mathcal{B} := \{1, \cdots, B\}$ as the set of bikes in the system. At time $t \in [0, T]$, we define the coordinate of bike $b \in \mathcal{B}$ as $(x_{b,t}, y_{b,t})$. Let $z_{b,t} = 1$ if bike b is available to be booked at time t, and $z_{b,t} = 0$ if bike b is not available at time t because it is booked or occupied. Let $\mathcal{B}_t := \{j \in \mathcal{B} : z_{b,t} = 1\} \subset \mathcal{B}$ be the set of bikes that are available for booking at time t.

When a rider arrives at a location $l \in \mathcal{L}$ at time t, she is presented with a *bike pattern* $S_t := \{(x_{b,t}, y_{b,t})\}_{b\in\mathcal{B}_t}$ which contains the coordinates of all available bikes at time t. This captures both dock-based and dockless systems. Let p_{l,b,S_t} be the probability that a rider at location l chooses bike $b \in \mathcal{B}_t$ at time t. In addition, let $p_{l,0,S_t}$ denote the probability that a rider at location l leaves without choosing a bike at time t. We require that $\sum_{b\in\mathcal{B}_t\cup\{0\}} p_{l,b,S_t} = 1$. One example of such riders' choice behavior is a multinomial logit (MNL) choice model. Walking distance to the bike is arguably the most important feature. Let d_{l,b,S_t} be the distance from rider location $l \in \mathcal{L}$ to bike $b \in \mathcal{B}$ in bike pattern S_t. For each rider location $l \in \mathcal{L}$ and available bike $b \in \mathcal{B}_t$,

$$p_{l,b,S_t} = \frac{\exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}{1 + \sum_{b\in\mathcal{B}_t} \exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}, \quad p_{l,0,S_t} = \frac{1}{1 + \sum_{b\in\mathcal{B}_t} \exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}, \tag{1}$$

where $\beta_{0,l} \in \mathbb{R}, \beta_{1,l} \in \mathbb{R}_{<0}$ are parameters measuring rider tolerance for walking distance at different locations. Another example is a distance-ranking choice model in which the preference is determined by the ranking of distances to the available bikes. In specific,

$$p_{l,b,S_t} = \frac{\mathbf{1}\{d_{l,b,S_t} \le d_{l,b',S_t},\ \forall b' \in \mathcal{B}_t,\ d_{l,b,S_t} \le \bar{r}_l\}}{\left|\left\{b' \in \mathcal{B}_t : d_{l,b',S_t} = \min\{d_{l,b'',S_t} : b'' \in \mathcal{B}_t\}\right\}\right|}, \quad p_{l,0,S_t} = \mathbf{1}\{d_{l,b,S_t} > \bar{r}_l,\ \forall b \in \mathcal{B}_t\}, \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $\bar{r}_l$ is the consideration radius at rider location l which specifies how far a rider is willing to walk to a bike. In words, a rider at location l chooses to pick up an available bike $b \in \mathcal{B}_t$ if and only if bike b is the closest available bike to rider location l and its distance does not exceed $\bar{r}_l$. When a tie occurs, each bike with the shortest distance within the consideration radius has the same probability to be booked by a rider.

Data and Observations. The data available to the operator consists of records of bookings and returns as well as the statuses and locations of all bikes in real-time. The statuses of a bike include "available" or "occupied". When a rider books a bike, the status changes from "available" to "occupied". When she drops off the bike at her destination, the status changes back to "available". As a key feature of our problem, the operator cannot observe riders' arriving locations. As a consequence, it cannot distinguish between no arrival with a rider arriving without choosing a bike.

3.1. Model Formulation

Given the above generative process and observations, our goal is to estimate the following two quantities: (1) riders' total arrival rate λ into the service region; (2) the probability/weight vector $\boldsymbol{w}$ distributed over the set of rider locations $\mathcal{L}$. The choice behavior of picking up a bike at each rider location l is assumed to be known, and we will provide an extension in Section B.2 of the appendix where the arrival rates at different rider locations and the parameters governing the choice behavior are jointly estimated. We proceed to derive the likelihood of a given set of observations. To simplify the notation, define the *observed* arrival rate

$$\tilde{\lambda}(t) := \lambda \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t}\right).$$

This quantity takes out the portion of riders who choose not to pick up any bike upon arrival from the total arrival rate λ. Suppose there are N bookings in total during the arrival period $[0,T]$. We denote the sequence $\boldsymbol{t} := \{t_n\}_{n=1}^N$ where $t_1, \ldots, t_N \in [0,T]$, $t_1 < t_2 < \cdots < t_N$ as the time epochs that bookings occur and $\boldsymbol{b} := \{b_n\}_{n=1}^N$ where $b_1, \ldots, b_N \in \mathcal{B}$ as the bikes booked by the riders in the corresponding booking times. Define $t_0 = 0$ and $t_{N+1} = T$. Consider a short enough time period $\delta > 0$ around each booking times $\{t_n\}_{n=1}^N$ such that bike patterns S_t do not change during these intervals $t \in [t_n, t_n + \delta]$, $n = 1, \cdots, N$. The incomplete data log-likelihood function is then given by

$$\begin{aligned} l_I(\boldsymbol{w}, \lambda) :=& \lim_{\delta \downarrow 0} \log \left(\prod_{n=0}^{N} \mathbb{P}(\text{no rider books bikes from } t_n + \delta \text{ to } t_{n+1}) \right. \\ & \left. \cdot \prod_{n=1}^{N} \frac{\mathbb{P}(\text{a rider books bike } b_n \text{ from } t_n \text{ to } t_n + \delta)}{\delta} \right) \\ =& \lim_{\delta \downarrow 0} \left(\sum_{n=0}^{N} \log \mathbb{P}(\text{no rider books bikes from } t_n + \delta \text{ to } t_{n+1}) \right. \end{aligned}$$

$$
\begin{aligned}
&+\sum_{n=1}^{N}\log\mathbb{P}(\text{a rider books bike } b_n \text{ from } t_n \text{ to } t_n+\delta)-\sum_{n=1}^{N}\log(\delta)\Bigg)\\
&=\lim_{\delta\downarrow 0}\Bigg(\sum_{n=0}^{N}\log\Bigg(\underbrace{\exp\left(-\int_{t_n+\delta}^{t_{n+1}}\tilde{\lambda}(t)dt\right)}_{\substack{\text{Prob. of no booking}\\ \text{in } [t_n,t_n+\delta]}}\Bigg)\\
&+\sum_{n=1}^{N}\log\Bigg(\underbrace{\int_{t_n}^{t_n+\delta}\tilde{\lambda}(t)dt\cdot\exp\left(-\int_{t_n}^{t_n+\delta}\tilde{\lambda}(t)dt\right)}_{\substack{\text{Prob. of one booking}\\ \text{in } [t_n,t_n+\delta]}}\cdot\underbrace{\frac{\sum_{l\in\mathcal{L}}w_l p_{l,b_n,S_{t_n}}}{1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_{t_n}}}}_{\substack{\text{Prob. of booking } b_n \text{ conditional}\\ \text{on one booking in } [t_n,t_n+\delta]}}\Bigg)-\sum_{n=1}^{N}\log(\delta)\Bigg)\\
&=-\int_0^T\tilde{\lambda}(t)dt+\sum_{n=1}^{N}\log\tilde{\lambda}(t_n)+\sum_{n=1}^{N}\log\frac{\sum_{l\in\mathcal{L}}w_l p_{l,b_n,S_{t_n}}}{1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_{t_n}}}. \qquad (3)
\end{aligned}
$$

Equation (3) holds because $\int_{t_n}^{t_n+\delta}\tilde{\lambda}(t)dt=\tilde{\lambda}(t_n)\delta$ as S_t does not change within $[t_n,t_n+\delta]$. Taking the first-order condition with respect to λ, it can be seen that the total arrival rate λ has a unique closed-form maximizer.

$$
-\int_0^T\left(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t}\right)dt+\sum_{n=1}^{N}\frac{1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_{t_n}}}{\tilde{\lambda}(t_n)}=0\ \Rightarrow\ \lambda=\frac{N}{\int_0^T\left(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t}\right)dt}. \qquad (4)
$$

Plugging the closed form of λ into (3), we can rewrite the incomplete log-likelihood function as a function of $\boldsymbol{w}$ only,

$$
\begin{aligned}
l_I(\boldsymbol{w}) :&= -N+\left(N\log N+\sum_{n=1}^{N}\log\left(\frac{1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_{t_n}}}{\int_0^T\left(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t}\right)dt}\right)\right)+\sum_{n=1}^{N}\log\frac{\sum_{l\in\mathcal{L}}w_l p_{l,b_n,S_{t_n}}}{1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_{t_n}}}\\
&=-N+N\log N-N\log\int_0^T\left(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t}\right)dt+\sum_{n=1}^{N}\log\left(\sum_{l\in\mathcal{L}}w_l p_{l,b_n,S_{t_n}}\right). \qquad (5)
\end{aligned}
$$

A key quantity involved in the above equation is $\int_0^T(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t})dt$. Remarkably, $\int_0^T(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t})dt/T$ measures the average percentage of riders who enter the system and pick up a bike. Typically, bike pattern S_t changes at events such as: (1) a rider books a bike; (2) a rider drops off her bike at her destination; (3) the operator relocates bikes. On the other hand, for the likelihood function (5) to be valid, there is no requirement on how bike patterns S_t change. Note that since S_t does not change continuously over time, the integral over t in (5) can be reorganized as a finite sum. Assume that the pattern changes at time epochs $t'_1,\cdots,t'_Q$ where Q is the total number of changes within $[0,T]$. Then we have $\int_0^T\left(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t}\right)dt=\sum_{q=0}^{Q}\left(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_{t_q}}\right)\left(t'_{q+1}-t'_q\right)$ with $t'_0=0$ and $t'_{Q+1}=T$.

The log-likelihood function (5) is non-concave in general. The non-concavity stems from the fact that the operator is not able to observe rider arrival locations. In the case of complete data where the operator is able to observe every rider's arrival and their arriving locations, the likelihood function becomes strictly concave, as we will show in Section 4.

3.2. Consistency of Location Weights Estimator

Let $\hat{\boldsymbol{w}} \in \arg\max_{\boldsymbol{w}\in\Delta^L} l_I(\boldsymbol{w})$ be the MLE. In this subsection, we investigate the consistency of $\hat{\boldsymbol{w}}$ — whether the estimator $\hat{\boldsymbol{w}}$ converges to true values in an asymptotic limit when the length of the arrival period $T \to \infty$. Note that the consistency of the arrival rate estimator $\hat{\lambda}$ follows directly from the consistency of the location weights through equation (4).

We first consider the following asymptotic regime when the length of the arrival period T gets large. We assume that there are K bike patterns (K is finite), which is denoted by $\mathcal{S} := \{S_1, \cdots, S_K\}$. We assume that $S_k \neq \emptyset, \forall k \in \{1, \cdots, K\}$. The long-run average fraction of time that we observe bike pattern S_k follows $\lim_{T\to\infty} \int_0^T \mathbf{1}(S_t = S_k)dt/T = \alpha_k > 0$ for all $k \in \{1, \cdots, K\}$ with $\alpha_1 + \cdots + \alpha_K = 1$. This setting typically models dock-based systems. For example, O'Mahony (2015) and Banerjee et al. (2022) use a continuous-time Markov chain (CTMC) to model state evolution where each state (bike pattern) is defined as the number of bikes in each station. Then $\{\alpha_1, \cdots, \alpha_K\}$ can be thought of as the steady-state distribution of this CTMC. Let $\boldsymbol{w}^*$ denote the underlying true weight vector. We first establish the identifiability of our estimator $\hat{\boldsymbol{w}}$ (i.e., different parameter values correspond to different data-generating distributions)[1], which is a necessary condition for consistency. Its proof relies on showing that the long-run average expected likelihood function $\lim_{T\to\infty} \mathbb{E}[l_I(\boldsymbol{w})]$ (the expectation is taken over bookings) has a unique maximizer at $\boldsymbol{w} = \boldsymbol{w}^*$, an equivalent condition for identifiability (see Lemma 5.35 in van der Vaart 2000).

THEOREM 1 **(Identifiability)**. *The location weights $\boldsymbol{w}$ are identifiable if the set of vectors $\left\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}] : b \in \mathcal{B},\ k \in \{1, \cdots, K\}\right\}$ spans the vector space $\mathbb{R}^L$. Moreover, the condition becomes necessary and sufficient when we have $w_l^* = 0$ for at most one $l \in \mathcal{L}$.*

Theorem 1 shows that a sufficient condition for identifiability is to have L linearly independent vectors from the vectors of riders' choice probabilities originating from different locations. This condition becomes necessary when the operator has relatively precise prior knowledge of where rider locations are — at most one location can be redundant in the candidate set $\mathcal{L}$. This result also implies that if riders' choice behavior depends only on distances, a minimum of L distinct bike locations across all bike patterns are required to possibly have identifiability of the estimator. We now provide a few concrete examples below to illustrate this result.

EXAMPLE 1 (IDENTIFIABILITY). Suppose that we have two rider arrival locations shown in Figure 1 (red circles). We consider a case with only one bike pattern $\mathcal{S} = \{S\}$. We adopt Euclidean distance when calculating walking distances. In Figure 1a, we begin by examining a scenario in which a single bike (green circle) remains fixed at a specific location throughout the arrival period

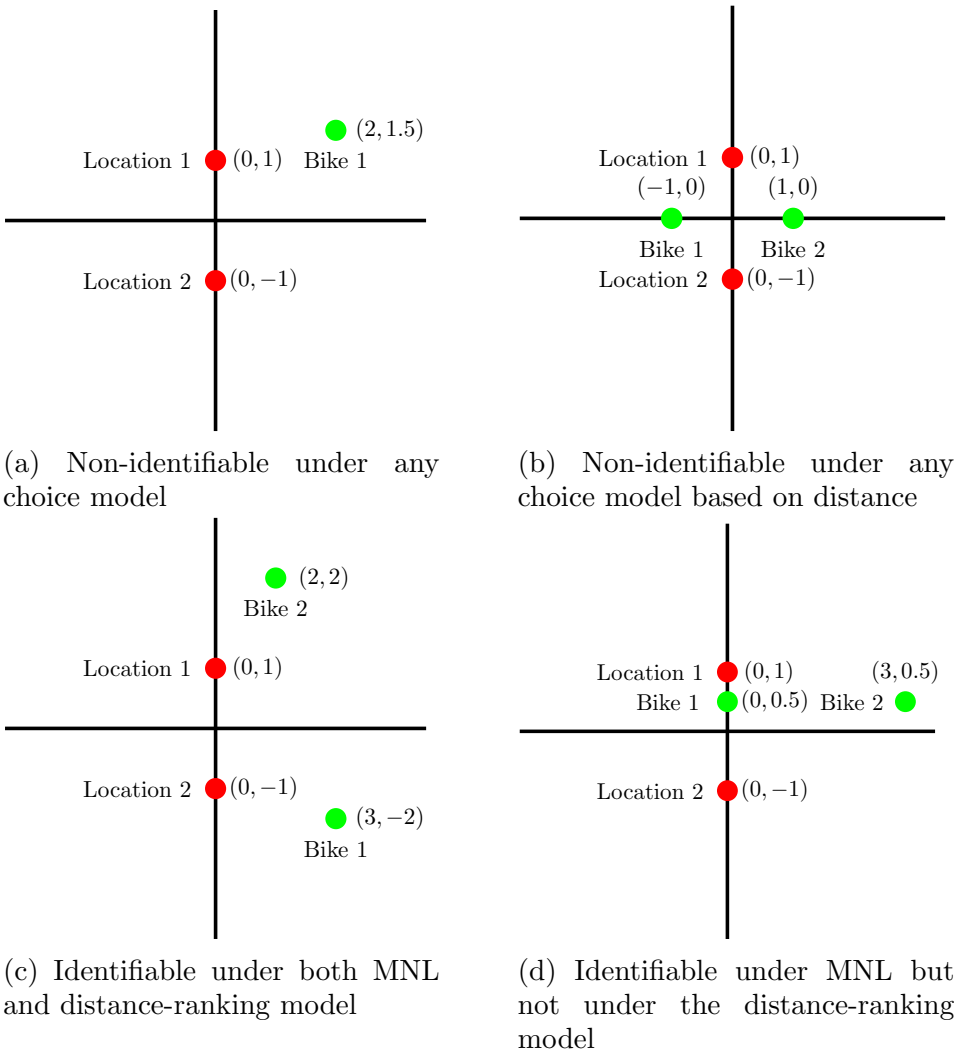
[1] We consider the data-generating distribution of bookings over any length of arrival period T such that $\int_0^T \mathbf{1}(S_t = S_k)dt/T = \alpha_k > 0, \forall k \in \{1, \cdots, K\}$.

until it is booked. In this scenario, the location weights cannot be consistently estimated with any choice model that only depends on distances. Intuitively, riders only have two options — book the bike or leave. Therefore, the only observation here is the booking time, which results in the weights being non-identifiable. On the other hand, when we have two bikes fixed at two distinct locations, the identification depends on the specification of the choice model. By Theorem 1, the weights can be identified if and only if $p_{1,1,S}p_{2,2,S} \neq p_{1,2,S}p_{2,1,S}$, i.e., vectors $[p_{1,1,S}, p_{2,1,S}]$ and $[p_{1,2,S}, p_{2,2,S}]$ are linearly independent. For the MNL model specified in equation (1), this can be simplified to $\beta_{1,1}(d_{1,1,S} - d_{1,2,S}) \neq \beta_{1,2}(d_{2,1,S} - d_{2,2,S})$, where $d_{l,b,S}$ is the distance from location l to bike b under pattern S. For the distance-ranking choice model specified in equation (2), the necessary and sufficient conditions for the location weights to be identifiable are: (1) the closer bike to locations 1 and 2 are different; (2) for both locations 1 and 2, at least one bike is within their consideration radiuses. To ease the presentation, we assume that $\beta_{0,1} = \beta_{0,2}$ and $\beta_{1,1} = \beta_{1,2}$ in the MNL choice model, and the consideration radius r is infinite in the distance-ranking choice model. Figure 1b illustrates a scenario in which non-identifiability of location weights occurs under any choice model since bikes 1 and 2 both have the same distance to locations 1 and 2. Figure 1c gives a scenario where location weights are identifiable under both MNL and distance-ranking choice models. Finally, in Figure 1d, location weights are identifiable under the MNL model but not the distance-ranking choice model. This is because, in the latter, a rider always chooses bike 1 over bike 2 regardless of where she arrives. It is not hard to prove that in this example of two rider locations and two bike locations, if location weights are identifiable under the distance-ranking model, they must be identifiable under the MNL model as well. Interestingly, this is *not* true in general. In Appendix B.1, we give a counter-example with three bikes and three rider locations where the rider location weights are identifiable under the distance-ranking model but not the MNL model, though it is generally expected that the MNL model possesses better identifiability properties because of the smoothness in choice probabilities.

In general, establishing consistency requires not only identifiability but also the uniform convergence of the log-likelihood function (see, e.g., Theorem 5.7 in van der Vaart 2000). We give two sufficient conditions below that ensure strong consistency of the MLE, that is $\hat{\boldsymbol{w}} \to \boldsymbol{w}^*$ with probability one as $T \to \infty$.

PROPOSITION 1 **(Strong Consistency of the MLE)**. *Suppose that the location weights $\boldsymbol{w}$ are identifiable, then the MLE $\hat{\boldsymbol{w}}$ converges to $\boldsymbol{w}^*$ with probability one if one of the following conditions holds:*

1. *choice probability $p_{l,b,S_k} > 0$ for all $l \in \mathcal{L}$, $b \in \mathcal{B}_k$, $k \in \{1, \cdots, K\}$;*
2. *there exists $\epsilon > 0$ such that $w_l^* \geq \epsilon$ for all $l \in \mathcal{L}$.*



(a) Non-identifiable under any choice model

(b) Non-identifiable under any choice model based on distance

(c) Identifiable under both MNL and distance-ranking model

(d) Identifiable under MNL but not under the distance-ranking model

Figure 1 **Examples of identifiability of location weights with only one bike pattern.** In the graphs, red circles represent arrival locations and green circles represent bike locations.

The first condition holds under an MNL choice model such as (1). The second condition holds when the operator has precise knowledge of the set of true locations. Either assumption guarantees a finite log-likelihood value for all $\boldsymbol{w}$ in a compact parameter space that contains the true location weights $\boldsymbol{w}^*$, where the log-likelihood function is dominated by an integrable function. Then by the uniform law of large numbers, the uniform convergence of the log-likelihood can be established by the dominance condition together with the continuity of the log-likelihood function (see Theorem 7.48 in Shapiro et al. 2009).

In most dockless systems, bikes are located at any place where parking is allowed. We thus provide a generalization to scenarios where bike patterns are continuously distributed over space. Let $(\mathcal{S}, \mathcal{F})$ be a measurable space of bike patterns. Let π and μ both be measures on $\mathcal{F}$. Furthermore, π is an

invariant probability measure that is absolutely continuous with respect to μ, which measures the long-run average portion of each bike pattern, $\pi(A) = \lim_{T\to\infty}(1/T)\int_{t=0}^{T}\int_{\mathcal{S}}\mathbb{I}_A(S_t)d\mu(S_t)dt,\ \forall A \in \mathcal{F}$. We have the following corollary regarding the identifiability of the MLE. Consistency can be established in a similar fashion as Proposition 1. We omit the discussion here.

COROLLARY 1. *The location weights $\boldsymbol{w}$ are identifiable if the set of vectors $\{(p_{1,b,S},\cdots,p_{L,b,S}) : b\in\mathcal{B}, S\in\mathcal{S}'\}$ spans the vector space $\mathbb{R}^L$ for any $\mathcal{S}'\subseteq\mathcal{S}$ such that $\pi(\mathcal{S}')=1$. The condition becomes necessary and sufficient when we have $w_l^*=0$ for at most one $l\in\mathcal{L}$.*

We now strengthen our analysis in dock-based systems where bikes are located in a finite set of stations. In particular, we analyze cases where riders make choices according to the distance-ranking model (2). These results bypass the general identifiability and consistency results stated in Theorem 1 and Proposition 1, and give intuitive conditions under which the MLE of location weights has strong consistency. Under this setup, we only need to define a *bike station pattern* as the set of available stations at some time (a station is available if it contains at least one available bike) to replace bike patterns.

We first analyze a stylized case where the service region $\mathcal{P}\subset\mathbb{R}$ is a one-dimensional line segment. For any location $x\in P$, let $r(x)$ be the consideration radius of riders arriving at location x. We assume that $r(\cdot)$ is large enough so that when every bike station has at least one available bike, riders will always choose a bike regardless of where they arrive in $\mathcal{P}$. For convenience, we assume that all possible bike station patterns are observed with a positive fraction of time as $T\to\infty$, although strictly speaking this can be relaxed as we show in its proof.

THEOREM 2. *In a one-dimensional service region, suppose that the consideration radius $r(\cdot)$ is Lipschitz continuous with constant one, i.e., $\|r(x)-r(x')\|\le\|x-x'\|$ for all $x,x'\in\mathcal{P}$. Then for each rider location $l\in\mathcal{L}$, if the sequence of stations within its consideration radius is uniquely ordered based on the distance to location l (i.e., without any ties), and this sequence is distinct from the sequences of other rider locations, then with probability one, $\hat{w}_l\to w_l^*$ as $T\to\infty$.*

Theorem 2 shows that in a one-dimensional service region, under smoothly changing consideration radius, we can consistently estimate the location weight whose distance ranking is unique. This result is crisper than the previous ones as we give consistency results for estimating each *individual* rider location weight. It also complements Proposition 1 to shed light on the consistency of location weights under a distance-ranking model. The proof of Theorem 2 uses very different techniques from the proofs of Theorem 1 and Proposition 1 and requires deliberately constructing a unique solution of location weights from choice probabilities of certain stations being picked

among a set of available stations, which are shown to be consistently estimated. It relies on recovering interesting structures through which the unique distance rankings differ in a one-dimensional space (see Lemmas EC.2 and EC.3).[2] In service regions of higher dimensions, counter-examples can be established that even rider locations with unique distance rankings can be non-identifiable. Nevertheless, we have the following generic result.

PROPOSITION 2. *For each rider location $l \in \mathcal{L}$, if the ranking of the first two closest stations (or the only station) in its consideration radius is unique, then with probability one, we have $\hat{w}_l \to w_l^*$ as $T \to \infty$.*

4. Estimation Procedures

We show how we can obtain the set of rider locations and the MLE of their weights $\hat{\boldsymbol{w}}$ in this section. First, given a set of candidate rider locations, we show how we can computationally approach the MLE using an expectation-maximization (EM) algorithm with closed-form updates. We then introduce a location-discovery procedure to iteratively explore new rider locations until convergence.

4.1. Estimation of Location Weights

Given a candidate set of rider locations $\mathcal{L}$, we consider the optimization problem $\max_{\boldsymbol{w}} l_I(\boldsymbol{w})$. In this subsection, we develop an EM algorithm with closed-form updates to optimize the weight vector $\boldsymbol{w}$. The EM algorithm, proposed by Dempster et al. (1977), is an iterative algorithm that consists of an *expectation* step (E-step) and a *maximization* step (M-step) in each iteration. The algorithm computes the conditional expected log-likelihood with respect to unobserved data (E-step) and then updates the estimates through maximizing the expected log-likelihood in the E-step (M-step). This procedure is repeated until convergence. In our case, the *observed* data contains sequences of booking times $\boldsymbol{t}$, booked bikes $\boldsymbol{b}$, and bike patterns S_t at any time $t \in [0, T]$. The *unobserved* data include: (1) the arrival times of riders who leave without choosing a bike (unobserved riders); and (2) the arrival locations of riders. Define the total number of arrivals $\tilde{N} := N + N'$ where N' is the number of unobserved riders. We first consider the *complete data log-likelihood function*, which is the log-likelihood function derived under the full observation of all arrival times, arrival locations and booked bikes. Let $\tilde{\boldsymbol{l}} := \{\tilde{l}_n\}_{n=1}^{\tilde{N}}$ be the arrival locations of all riders in the sequence, $\tilde{\boldsymbol{t}} := \{\tilde{t}_n\}_{n=1}^{\tilde{N}}$ be the arrival times and $\tilde{\boldsymbol{b}} := \{\tilde{b}_n\}_{n=1}^{\tilde{N}}$, $\tilde{b}_n \in \mathcal{B} \cup \{0\}$ be the sequence of bikes booked by the riders. Here, $\tilde{b}_n = 0$ means that the n^{th} rider arrives without picking up a bike. Using these notations, we

[2] Although this one-dimensional result is arguably stylized in the bike-sharing setting, it can be suitable to model consumer choice behavior in horizontal product differentiation where prices and quality levels are equal across all products, such as yogurt with different flavors and shoes of different colors (see, e.g., Hotelling 1929, Lancaster 1966, 1975 in the economics literature and Gaur and Honhon 2006 in the operations literature on related locational choice models).

can write down the complete data log-likelihood function by following a similar procedure as in equation (3).

$$\begin{aligned} l_C(\boldsymbol{w},\lambda) =& \lim_{\delta\downarrow 0} \log\left(\prod_{n=0}^{\tilde{N}} \mathbb{P}(\text{no rider books bikes from } t_n+\delta \text{ to } t_{n+1})\right. \\ & \left. \cdot \prod_{n=1}^{\tilde{N}} \frac{\mathbb{P}(\text{a rider arriving at } \tilde{l}_n \text{ books bike } \tilde{b}_n \text{ (or leave) from } t_n \text{ to } t_n+\delta)}{\delta}\right) \\ =& \log\left(\exp\left(-\lambda T\right)\right) + \sum_{n=1}^{\tilde{N}} \log\left(\lambda w_{\tilde{l}_n} p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}\right) \\ =& -\lambda T + \tilde{N}\log\lambda + \sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) + \sum_{n=1}^{\tilde{N}} \log\left(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}\right). \end{aligned}$$

Similarly, the arrival rate λ can be substituted with the MLE $\hat{\lambda} = \tilde{N}/T$. This simplifies the log-likelihood function to $l_C(\boldsymbol{w})$ which only depends on the location weights $\boldsymbol{w}$,

$$l_C(\boldsymbol{w}) = -\tilde{N} + \tilde{N}\log(\tilde{N}/T) + \sum_{n=1}^{\tilde{N}} \log\left(w_{\tilde{l}_n}\right) + \sum_{n=1}^{\tilde{N}} \log\left(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}\right).$$

E-step. In the E-step, we compute the expectation of the complete data log-likelihood conditional on the observed data $\boldsymbol{b}$ and $\boldsymbol{t}$ and the current estimate $\boldsymbol{w}^{(m)}$ at the m^{th} iteration. The expectation is taken over the randomness of unobserved data, which includes: (1) the number of unobserved riders N'; (2) the arrival times of unobserved riders $\tilde{\boldsymbol{t}}\backslash\boldsymbol{t}$; and (3) the arrival locations of all riders $\tilde{\boldsymbol{l}}$.

$$\mathbb{E}\left[l_C(\boldsymbol{w}) \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right] = \mathbb{E}\left[-\tilde{N} + \tilde{N}\log\left(\frac{\tilde{N}}{T}\right) + \sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) + \sum_{n=1}^{\tilde{N}} \log\left(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}\right) \,\middle|\, \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right]. \quad (6)$$

Notice that the only part of the expectation in equation (6) that depends on $\boldsymbol{w}$ is $\sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n})$. Thus, it is sufficient to solely focus on the conditional expectation $\mathbb{E}\left[\sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right]$. We denote by $\{l_1, l_2, \cdots, l_N\}$ the sequence of the arrival locations of the observed riders, which is a subsequence of $\{\tilde{l}_1, \cdots, \tilde{l}_{\tilde{N}}\}$, the sequence of arrival locations of all riders. For the sake of exposition, we consider two quantities, which are $\mathbb{P}(l_n = l \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)})$ and $\mathbb{E}[N' \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}]$. The first quantity refers to the probability that the n^{th} observed rider arrives at location $l \in \mathcal{L}$, whereas the second quantity refers to the expected number of unobserved riders. They can be computed as follows.

$$\mathbb{P}\left(l_n = l \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right) = \frac{\mathbb{P}\left(b_n \mid l_n = l,\boldsymbol{t},\boldsymbol{w}^{(m)}\right)\mathbb{P}\left(l_n = l \mid \boldsymbol{t},\boldsymbol{w}^{(m)}\right)}{\mathbb{P}\left(b_n \mid \boldsymbol{t},\boldsymbol{w}^{(m)}\right)} = \frac{p_{l,b_n,S_{t_n}} w_l^{(m)}}{\sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',b_n,S_{t_n}}}, \quad (7)$$

$$\mathbb{E}\left[N' \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}\right] = \mathbb{E}[N \mid \boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m)}] \frac{\int_0^T \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t} dt}{\int_0^T \left(1 - \sum_{l'\in\mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t}\right) dt}$$

$$= N \frac{\int_0^T \sum_{l' \in \mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t} dt}{\int_0^T \left(1 - \sum_{l' \in \mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t}\right) dt}. \tag{8}$$

We now separate the conditional expectation $\mathbb{E}\left[\sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right]$ into two parts, which consist of observed and unobserved data. Let the sequence $\{l'_1, l'_2, \cdots, l'_{N'}\}$ denote the arrival locations of the unobserved riders.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{N} \log(w_{l_n}) \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] + \mathbb{E}\left[\sum_{n=1}^{N'} \log(w_{l'_n}) \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] \\
&= \sum_{l \in \mathcal{L}} \sum_{n=1}^{N} \mathbb{P}\left(l_n = l \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right) \log(w_l) \\
&\quad + \mathbb{E}\left[N' \,\middle|\, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] \sum_{l \in \mathcal{L}} \underbrace{\frac{\int_0^T w_l^{(m)} p_{l,0,S_t} dt}{\int_0^T \sum_{l' \in \mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t} dt}}_{\substack{\text{prob. of arriving at } l \\ \text{conditional on leaving}}} \log(w_l) \qquad \text{(Wald's Lemma)} \\
&= \sum_{l \in \mathcal{L}} \left(\sum_{n=1}^{N} \mathbb{P}\left(l_n = l \mid \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right) + \mathbb{E}\left[N' \mid \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}\right] \frac{\int_0^T w_l^{(m)} p_{l,0,S_t} dt}{\int_0^T \sum_{l' \in \mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t} dt} \right) \log(w_l) \\
&= \sum_{l \in \mathcal{L}} \underbrace{\left(\sum_{n=1}^{N} \frac{p_{l,b_n,S_{t_n}} w_l^{(m)}}{\sum_{l' \in \mathcal{L}} w_{l'}^{(m)} p_{l',b_n,S_{t_n}}} + N \frac{\int_0^T w_l^{(m)} p_{l,0,S_t} dt}{\int_0^T (1 - \sum_{l' \in \mathcal{L}} w_{l'}^{(m)} p_{l',0,S_t}) dt} \right)}_{c_l} \log(w_l) \qquad (9) \\
&= \sum_{l \in \mathcal{L}} c_l \log(w_l).
\end{aligned}$$

Equation (9) holds by replacing the terms with equations (7) and (8). We have thus simplified the conditional expectation into a weighted sum of logarithms, which can be maximized in closed form in the M-step described below.

M-step. Let $c := \sum_{l \in \mathcal{L}} c_l$ be the sum of c_l for all $l \in L$. By weighted AM–GM inequality, we have

$$\begin{aligned}
& 0 = \log\left(\sum_{l \in \mathcal{L}} \frac{c_l}{c} \left(\frac{c w_l}{c_l}\right)\right) \geq \log\left(\prod_{l=1}^{L} \left(\frac{c w_l}{c_l}\right)^{\frac{c_l}{c}}\right) = \sum_{l \in \mathcal{L}} \frac{c_l}{c} (\log(c w_l) - \log c_l) \\
\iff \quad & \sum_{l \in \mathcal{L}} c_l \log(w_l) \leq \sum_{l \in \mathcal{L}} c_l \log(c_l) - \sum_{l \in \mathcal{L}} c_l \log(c).
\end{aligned}$$

The equality holds if and only if $w_l = c_l / c$, which shows that the optimal $\boldsymbol{w}$ has a closed-form solution $w_l = c_l / (\sum_{l \in \mathcal{L}} c_l)$ for $l \in \mathcal{L}$. We summarize our EM algorithm below. To simplify the exposition, define $s = \int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t}\right) dt$. Recall that s/T measures the average percentage of riders who arrive and book a bike.

Algorithm 1 EM Algorithm for Location Weights Estimation

Given a set of candidate locations $\mathcal{L}$ with coordinates $(x_1, y_1), \cdots, (x_L, y_L)$.

Initialize the location weights $\boldsymbol{w} \leftarrow \boldsymbol{w}_0 \in \Delta^L$.

while $\boldsymbol{w}$ does not converge **do**

$s \leftarrow \int_0^T \big(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t}\big) dt$.

$c_l \leftarrow \sum_{n=1}^N \big(p_{l,b_n,t_n} w_l \big/ \big(\sum_{l'=1}^L w_{l'} p_{l',b_n,t_n}\big)\big) + N(T-s)/s$ for all $l \in \mathcal{L}$.

$w_l \leftarrow c_l / \sum_{l'=1}^L c_{l'}$ for all $l \in \mathcal{L}$.

end while

Output location weights $\boldsymbol{w}$.

We show that our EM algorithm meets the regularity conditions in Wu (1983) which leads to the convergence result below.

THEOREM 3. *Let $\boldsymbol{w}^{(m)}$ be the location weights at the m^{th} iteration of Algorithm 1. All limit points of the sequence $\{\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \cdots\}$ are stationary points of the incomplete-data log-likelihood $l_I(\boldsymbol{w})$ and the sequence $\{l_I(\boldsymbol{w}^{(1)}), l_I(\boldsymbol{w}^{(2)}), \cdots\}$ converges monotonically to $l_I(\boldsymbol{w}^*)$ for some stationary point $\boldsymbol{w}^*$.*

Though our EM method does not promise to converge to global maxima, as we will discuss in Section 5, the sequence of points estimated by our EM method always converges to a limit point and empirically gives a close approximation to the ground truth. In Appendix B.2, we extend the EM method to jointly estimate the location weights and parameters governing riders' choice behaviors. When arriving riders pick up bikes according to an MNL model, we show that the EM procedures can be adapted to efficiently estimate the choice model parameters as well (see Proposition EC.3 in Appendix B.2).

We conclude this subsection by commenting that a broader class of algorithms for maximizing the log-likelihood used in the related literature is the Minorization-Maximization (MM) algorithm (Hunter and Lange 2000). MM algorithms maximize simple concave surrogate functions that minorize the log-likelihood function. Different MM algorithms differ in the way the surrogate function is constructed. EM can be viewed as a special case of an MM algorithm using a surrogate function that is constructed by Jensen's inequality (see Section 3.1 of Hunter and Lange 2004). In Section 5, we compare the performance of our EM algorithm with another MM algorithm that is constructed by viewing the log-likelihood function (5) as a difference of two concave functions, and constructing surrogate functions using supporting hyperplanes of the second concave function. We show that our EM algorithm has superior performance due to its simple closed-form updates.

4.2. Estimation of Location Coordinates

In practice, the operator often does not have full knowledge of the possible rider arrival locations. As a result, the set of candidate rider locations $\mathcal{L}$ is usually underdetermined and has to be estimated from data. A naïve approach is to enumerate a large set of *all possible* candidate locations, for example, all of the residential buildings in a service area. We then implement our EM algorithm on this set to estimate each location's weight. We refer to this algorithm as the *all-in* algorithm. If the underlying arrival locations are entirely contained in the initial set, then by various results in Section 3.2, the MLE can be identifiable and consistent under certain conditions. In such cases, the MLE converges to the ground truth for candidate locations that coincide with the true underlying locations, and 0 for all other locations. However, this algorithm has the following drawbacks due to a large cardinality of all possible rider locations: (1) it significantly increases the number of iterations necessary for convergence, resulting in a computationally demanding task; (2) it renders the data requirement for identifiability (e.g., Theorem 1) harder to satisfy, making the estimates less accurate.

A location-discovery procedure. Motivated by van Ryzin and Vulcano (2014) where they proposed a market-discovery algorithm to estimate ranking-based customer preferences, we develop a *location-discovery* algorithm to address the aforementioned drawbacks of the all-in algorithm. The algorithm iteratively and adaptively explores new locations until convergence. Suppose $\mathcal{L}$ is the set of all possible rider locations. We begin with a parsimonious set of locations $\mathcal{L}_0 \subset \mathcal{L}$, a discovery procedure is employed to gradually enlarge our location set. The main idea is to maximize the log-likelihood by iteratively adding new locations with the largest potential for improvements and then executing the EM algorithm (Algorithm 1) to update their weights.

To start the derivation, we define the restricted Lagrangian function $\Theta^{\mathcal{L}_0}(\boldsymbol{w}, \mu)$ with location set $\mathcal{L}_0$ by relaxing the equality constraint $\sum_{l\in\mathcal{L}_0} w_l = 1$ with a Lagrangian multiplier μ.

$$\Theta^{\mathcal{L}_0}(\boldsymbol{w}, \mu) = -N \log \int_0^T \left(1 - \sum_{l\in\mathcal{L}_0} w_l p_{l,0,S_t}\right) dt + \sum_{n=1}^N \log \sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}} + \mu \left(1 - \sum_{l\in\mathcal{L}_0} w_l\right). \quad (10)$$

Consider a local optimum $\bar{\boldsymbol{w}}$ of the problem $\max_{\boldsymbol{w}\in\Delta^{|\mathcal{L}_0|}} l_I(\boldsymbol{w})$ with location set $\mathcal{L}_0$. Recall that $s = \int_0^T \left(1 - \sum_{l\in\mathcal{L}_0} w_l p_{l,0,S_t}\right) dt$. Suppose that $(\bar{\boldsymbol{w}}, \bar{\mu})$ satisfies KKT conditions of the restricted Lagrangian function $\Theta^{\mathcal{L}_0}(\boldsymbol{w}, \mu)$. We have for all $l \in \mathcal{L}_0$ such that $\bar{w}_l > 0$,

$$\left.\frac{\partial \Theta^{\mathcal{L}_0}(\boldsymbol{w}, \mu)}{\partial w_l}\right|_{\boldsymbol{w}=\bar{\boldsymbol{w}}, \mu=\bar{\mu}} = \frac{N}{s} \int_0^T p_{l,0,S_t} dt + \sum_{n=1}^N \frac{p_{l,b_n,S_{t_n}}}{\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}} - \bar{\mu} = 0,$$

which gives that

$$\bar{\mu} = \frac{N}{s} \int_0^T p_{l,0,S_t} dt + \sum_{n=1}^N \frac{p_{l,b_n,S_{t_n}}}{\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}}, \text{ for any } l \in \mathcal{L}_0 \text{ such that } \bar{w}_l > 0.$$

The essence of our location-discovery procedure is to discover a new location that potentially defines the largest improvement direction for the log-likelihood value. To do so, we consider the full Lagrangian function $\Theta^{\mathcal{L}}(\boldsymbol{w}, \mu)$ with the full set of all possible rider locations $\mathcal{L}$. Define $\bar{\bar{\boldsymbol{w}}} \in \Delta^{|\mathcal{L}|}$ such that $\bar{\bar{w}}_l = \bar{w}_l, \forall l \in \mathcal{L}_0$ and $\bar{\bar{w}}_l = 0, \forall l \in \mathcal{L} \setminus \mathcal{L}_0$. Suppose that

$$\left.\frac{\partial \Theta^{\mathcal{L}}(\boldsymbol{w}, \mu)}{\partial w_l}\right|_{\boldsymbol{w}=\bar{\bar{\boldsymbol{w}}}, \mu=\bar{\mu}} = \frac{N}{s}\int_0^T p_{l,0,S_t} dt + \sum_{n=1}^{N} \frac{p_{l,b_n,S_{t_n}}}{\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}} - \bar{\mu} \leq 0, \ \forall l \in \mathcal{L} \setminus \mathcal{L}_0, \tag{11}$$

then $(\bar{\bar{\boldsymbol{w}}}, \bar{\mu})$ satisfies the KKT conditions of the full Lagrangian function $\Theta^{\mathcal{L}}(\boldsymbol{w}, \mu)$ as well. If not, then there exists some $l \in \mathcal{L} \setminus \mathcal{L}_0$ such that $\partial \Theta^{\mathcal{L}}(\boldsymbol{w}, \mu)/\partial w_l > 0$ evaluated at $\boldsymbol{w} = \bar{\bar{\boldsymbol{w}}}$ and $\mu = \bar{\mu}$. Note that since the log-likelihood function $l_I(\boldsymbol{w})$ is not concave in $\boldsymbol{w}$, KKT conditions are not sufficient for local optimum. This means that including such a new location l is not guaranteed but may lead to an improved estimate with greater likelihood. Nevertheless, this gives a principled procedure to gradually include new locations. We summarize the estimation algorithm with a location discovery procedure in Algorithm 2 below.

Finding the location with the largest partial derivative of the Lagrangian. At each iteration of Algorithm 2, a new location is selected by solving

$$\max_{l' \in \mathcal{L} \setminus \mathcal{L}_0} \quad \frac{N}{s}\int_0^T p_{l',0,S_t} dt + \sum_{n=1}^{N} \frac{p_{l',b_n,S_{t_n}}}{\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}} \tag{12}$$

to maximize the partial derivative of the full Lagrangian function $\Theta^{\mathcal{L}}(\boldsymbol{w}, \mu)$. The terms in the objective function have quite intuitive interpretations. It finds a location that strikes a balance between explaining riders' leaving without picking up bikes (the first term) and the observed booking sequence (the second term). Based on the current estimates, if s is small or $\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}$ is large, then the first term possesses a heavier weight than the second term. The new location l' tends to improve the explanation of riders' leaving behaviors by uplifting the leaving probability $p_{l',0,\mathcal{S}_t}$. Conversely, with a large s or small booking probability $\sum_{l\in\mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}$, the new location l' focuses more on explaining the booking behaviors by increasing the value of booking probabilities $p_{l',b_n,S_{t_n}}$.

As we will illustrate later, this objective as a function of the new location coordinates $(x_{l'}, y_{l'})$ is not concave in general and may exist multiple local optima (see Figure 3 in Section 5 for an example). Problem (12) thus, unfortunately, does not possess much structure. We could apply a general nonlinear programming algorithm such as gradient-based methods to approach its local optimum. On the other hand, since only two variables $x_{l'}$ and $y_{l'}$ need to be optimized, an alternative simple and effective method is to use *grid search*. The grid search method optimizes the objective function in a full factorial sampling plan, which places a grid of evenly spaced points across the

search area. We implement a multi-round variant of the search method to speed up the search process. In each round, the granularity of the search grid becomes finer and finer within a shrinking and more targeted search region. More details are provided in Section 5.

Another benefit of running a grid-search method is to easily enable *batch* addition of new locations. To be specific, instead of exploring one location at a time, we can simultaneously discover a batch of locations in each iteration. For example, all local maxima, in addition to the global maximum, can be included in each iteration. In the grid search procedure, we identify local maxima by discovering new locations whose partial derivatives are greater than those of eight neighboring locations in the search grid. As we will show later in Section 5, batch addition often speeds up the convergence and improves the accuracy of the estimator.

Algorithm 2 Estimation Algorithm with a Location-Discovery Procedure

Initialize a parsimonious location set $\mathcal{L}_0 \subset \mathcal{L}$.

Initialize the location weights vector $\boldsymbol{w}$ by running Algorithm 1 with location set $\mathcal{L}_0$.

while stopping criteria are not met **do**

$s \leftarrow \int_0^T \left(1 - \sum_{l \in \mathcal{L}_0} w_l p_{l,0,S_t}\right) dt.$

$l^* \in \arg\max_{l' \in \mathcal{L} \setminus \mathcal{L}_0} \left(\frac{N}{s} \int_0^T p_{l',0,S_t} dt + \sum_{n=1}^N \frac{p_{l',b_n,S_{t_n}}}{\sum_{l \in \mathcal{L}_0} w_l p_{l,b_n,S_{t_n}}} \right).$ ▷ discover a new location

$\mathcal{L}_0 \leftarrow \mathcal{L}_0 \cup \{l^*\}.$

Update $\boldsymbol{w}$ by running Algorithm 1 with location set $\mathcal{L}_0$.

end while

Output the location set $\mathcal{L}_0$ and its corresponding weight vector $\boldsymbol{w}$.

Convergence and stopping criteria. A natural stopping criterion is to terminate Algorithm 2 when conditions (11) hold. In other words, the resulting location set $\mathcal{L}_0$ and its corresponding weight vector $\boldsymbol{w}$, together with the Lagrangian multiplier $\bar{\mu}$, satisfy the KKT conditions. With this stopping criterion, when the set of all possible rider locations $\mathcal{L}$ is finite, it can be seen that Algorithm 2 terminates within a finite number of iterations. This is because, at each iteration, either the stopping criteria are met or a new location will be added. As there are only a finite number of them, the algorithm will converge. However, as we mentioned before, the converging estimate does not necessarily correspond to the MLE estimate as KKT conditions are not sufficient for optimality due to the non-concavity of the log-likelihood function $l_I(\boldsymbol{w})$. Furthermore, with this stopping criterion, the algorithm usually discovers much more locations than necessary since including new location parameters into the likelihood function always improves the log-likelihood value. It is thus

practical to explore other stopping criteria that penalize model complexity to prevent over-fitting. In particular, we use Bayesian Information Criterion (BIC), proposed by Schwarz (1978). BIC addresses this issue by adding a penalty term for the number of parameters in the model. The quantity is computed as

$$\text{BIC} = -l_I(\boldsymbol{w}) + 0.5 \cdot L \cdot \ln(N),$$

where $l_I(\boldsymbol{w})$ is the log-likelihood function, L is the number of locations discovered and N is the number of observed bookings. We terminate Algorithm 2 when the BIC value increases after we add a new location or a batch of new locations.

5. Numerical Experiments

In this section, we present numerical experiments of our demand model and estimation procedures. These are broadly divided into two parts, experiments based on synthetic data and real-world bike-sharing data. In Section 5.1, we demonstrate the performance of the algorithm over synthetically generated data on a square. We benchmark the EM algorithm and give evidence that the location-discovery procedure significantly improves the accuracy of the estimator. In Section 5.2, we illustrate our algorithm and its estimation results on a set of real-world dockless bike-sharing data in Seattle. We show that our methods produce a high accuracy in predicting ridership in out-of-sample tests. In addition, based on the estimation results, we provide managerial insights regarding bike allocations to increase service levels in the Seattle area. All algorithms are implemented in Python 3.8.10 with NumPy 1.23.3 on a virtual machine with 8 vCPUs using Microsoft Azure. Data and code to reproduce all experiments presented in this section can be found at `https://github.com/angxu1/bike_sharing`.

5.1. Experiments Based on Synthetic Data

We describe the generating process of our synthetic data. We consider an arrival period of length T. For each simulation run, we sample the initial locations of each bike (at $t = 0$) independently and uniformly from a 10×10 square. In specific, the (x, y) coordinates live in $[-5, 5]^2$. We then sample a sequence of arrivals based on a homogeneous Poisson process with rate λ within $[0, T]$ representing the rider arrival times. We use a discrete-event simulation that processes arrival times one by one in chronological order. Each rider arrives at a location $l \in \mathcal{L}$ sampled from a multinoulli distribution with probability w_l^* such that $\sum_{l \in \mathcal{L}} w_l^* = 1$. A rider chooses to book an available bike based on an underlying MNL choice model with model parameters $\beta_{0,l} = 1$ and $\beta_{1,l} = -1$ for all $l \in \mathcal{L}$. We use Euclidean distance (the L_2-norm between two location coordinates) as our measurement of distance. When a rider books a bike, we randomly sample her destination uniformly from the

10×10 square. We also generate her booking duration (in hours) according to a rectified Gaussian distribution $\max\{N(\text{walking time} + \text{traveling time}, 0.1), 0.05\}$ where walking time is computed as the distance from the rider's arrival location to the bike location divided by 4 km per hour (walking speed) and the traveling time is computed as the distance from the bike location to the rider's destination divided by 18 km per hour (cycling speed). If the rider chooses to leave, we simply move on to the next arrival. Bike patterns will be updated accordingly. We repeat this procedure until all rider arrivals are processed.

5.1.1. Performance of the EM algorithm We first benchmark the performance of our EM algorithm in maximizing the log-likelihood function by comparing it with an MM algorithm described at the end of Section 4.1. This approach iteratively optimizes a concave surrogate function which minorizes the log-likelihood function. In particular, ignoring the constant, we re-arrange the log-likelihood function (5) into a difference of two concave functions $g(\boldsymbol{w}) - h(\boldsymbol{w})$ where $g(\boldsymbol{w}) = \sum_{n=1}^{N} \log \sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}}$ and $h(\boldsymbol{w}) = N \log \int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_{t_n}}\right) dt$. To find a concave surrogate, we replace $h(\cdot)$ with its first-order approximation. Then the problem reduces to iteratively solving the following concave program with a simplex constraint

$$\boldsymbol{w}^{(m)} \in \underset{\boldsymbol{w} \geq 0, \sum_{l \in \mathcal{L}} w_l = 1}{\arg\max} \; g(\boldsymbol{w}) - h\left(\boldsymbol{w}^{(m-1)}\right) - \nabla h\left(\boldsymbol{w}^{(m-1)}\right)^T \left(\boldsymbol{w} - \boldsymbol{w}^{(m-1)}\right). \tag{13}$$

We solve this program using CVXPY 1.3 through the splitting conic solver (SCS). To compare their performances, we position a 5×5 Cartesian grid in the aforementioned 10×10 square and randomly select certain points from the 25 intersections of the Cartesian grid as the underlying true rider arrival locations. Their corresponding weights are independently sampled from a symmetric Dirichlet distribution with parameter 1. We fix the total number of bikes $B = 30$ and consider the number of ground-truth rider arrival locations $L^* \in \{2, 5, 10\}$, respectively. The length of the period T is set to be 100, and the underlying rider arrival rate is set to be $\lambda = 10$. We consider all 25 intersections on the 5×5 Cartesian grid as the set of potential rider arrival locations $\mathcal{L}$.

To evaluate the accuracy of the estimator, we compute the Wasserstein distance between the predicted location weights and the true location weights. In our model, given the estimated rider locations and their weights $(\hat{\mathcal{L}}, \hat{\boldsymbol{w}})$ and the underlying truth $(\mathcal{L}^*, \boldsymbol{w}^*)$, the Wasserstein 2−distance is defined as

$$W\left((\hat{\mathcal{L}}, \hat{\boldsymbol{w}}), (\mathcal{L}^*, \boldsymbol{w}^*)\right) = \inf_{\lambda_{i,j}} \left\{ \sum_{i=1}^{|\hat{\mathcal{L}}|} \sum_{j=1}^{|\mathcal{L}^*|} \lambda_{i,j} \|\hat{l}_i - l_j^*\|_2^2 : \sum_{i=1}^{|\hat{\mathcal{L}}|} \lambda_{i,j} = w_j^*, \sum_{j=1}^{|\mathcal{L}^*|} \lambda_{i,j} = \hat{w}_i, \lambda_{i,j} \geq 0 \right\}^{1/2},$$

where $\|\hat{l}_i - l_j^*\|_2 := \sqrt{(\hat{x}_i - x_j^*)^2 + (\hat{y}_i - y_j^*)^2}$ refers to the Euclidean distance between locations $\hat{l}_i$ and l_j^*.

Table 1 **Performance comparison of the EM and MM algorithms in estimating location weights.**

	EM algorithm				MM algorithm			
L^*	Iterations	WD	Lkd	Time	Iterations	WD	Lkd	Time
2	665.4	3.37	-1,619	7.04	15.3	3.27	-1,619	460.05
5	473.7	3.41	-2,990	10.00	8.1	3.25	-2,990	269.41
10	325.8	2.71	-4,109	9.40	13.3	2.85	-4,109	140.52

Table 1 reports the number of iterations until convergence, Wasserstein distance (WD) between the predicted location weights and the true weights, the log-likelihood values upon convergence (Lkd), and the CPU times in seconds of the algorithms averaged over 10 simulation runs. When computing the log-likelihood value, we exclude the constant term $-N + N \log N$. We observe that both algorithms produce very similar prediction accuracy and the converging location weights are in close proximity. However, the EM algorithm requires significantly less computation time — it benefits from closed-form updates, which makes each iteration hundreds of times faster than the MM algorithm. On the other hand, the MM algorithm requires much fewer iterations, but each iteration is expensive as it requires solving a convex optimization problem (13).

5.1.2. Performance of the location-discovery procedure We now evaluate the performance of the location discovery procedure. Similarly to Section 5.1.1, we randomly generate rider arrival locations and their corresponding weights on Cartesian grids of sizes 5×5 and 10×10 in the same square service region. Rider arrival locations can only be sampled from the intersections of the Cartesian grids. For each grid size, we consider two scenarios with different numbers of ground-truth rider locations and bikes: $(L^*, B) \in \{(5, 20), (10, 40)\}$. We test with different amount of data by setting the length of the arrival period to $T \in \{100, 500\}$. We compare the following methods.

- **All-in algorithm.** For Cartesian grids of sizes 5×5 and 10×10, we choose all 25 and 100 intersections as the set of candidate rider locations $\mathcal{L}$, respectively. We then run Algorithm 1 to obtain their location weights.

- **Location-discovery algorithm.** For Cartesian grids of sizes 5×5 and 10×10, we start with two rider arrival locations uniformly sampled from the 25 and 100 intersections. We consider two variants of the implementation: a single mode and a batch mode. In the single mode, rider location is added one by one. To find a new location to include, we simply compute the partial derivative of all remaining rider locations in the set and select the one with the largest value. For the batch mode, we discover and include all "local maximal" rider locations in the set whose partial derivatives are greater than those of their neighboring locations. The location-discovery algorithm (Algorithm 2) is stopped when the BIC value in the current iteration is greater than that derived in the previous iteration.

• **Clustering algorithm.** We also test a simple baseline method based on clustering. In particular, we simply choose all intersections of the Cartesian grids as the cluster centroids and assign bookings to its closest intersection. Location weights are then computed by normalizing the number of bookings belonging to each cluster so that they sum up to 1.

Table 2 Prediction performance with a finite and known set of candidate rider locations

			5×5 Cartesian Grid					10×10 Cartesian Grid				
Algorithm	T	(L^*, B)	Locs	WD	Lkd	BIC	Time	Locs	WD	Lkd	BIC	Time
Clustering	100	(5, 20)	24.9	4.35	-2,354	2,426	< 0.01	43.2	3.89	-2,468	2,594	< 0.01
		(10, 40)	25.0	3.33	-4,765	4,845	< 0.01	43.7	3.46	-4,795	4,935	< 0.01
	500	(5, 20)	25.0	4.43	-13,572	13,663	0.01	42.3	4.13	-14,104	14,260	0.01
		(10, 40)	25.0	3.34	-27,999	28,098	0.02	40.2	3.70	-28,424	28,585	0.02
All-in	100	(5, 20)	11.4	3.39	-2,249	2,282	0.83	25.6	3.30	-2,355	2,429	3.97
		(10, 40)	14.2	2.68	-4,666	4,711	2.19	27.6	2.92	-4,669	4,758	14.93
	500	(5, 20)	11.7	3.46	-13,068	13,111	4.40	27.0	3.26	-13,449	13,549	20.41
		(10, 40)	14.5	2.77	-27,502	27,560	13.09	29.4	2.95	-27,590	27,708	63.04
Loc. Disc. (Single)	100	(5, 20)	6.8	3.28	-2,260	2,279	0.80	6.8	3.14	-2,369	2,389	0.98
		(10, 40)	9.5	2.67	-4,676	4,707	2.49	9.2	2.59	-4,681	4,711	3.91
	500	(5, 20)	7.6	3.19	-13,080	13,108	5.87	8.5	2.93	-13,467	13,498	8.98
		(10, 40)	11.0	2.40	-27,512	27,556	26.23	11.5	2.36	-27,596	27,642	37.41
Loc. Disc. (Batch)	100	(5, 20)	8.5	3.06	-2,252	2,277	0.82	9.3	2.82	-2,358	2,385	0.96
		(10, 40)	11.4	2.47	-4,668	4,705	2.74	12.6	2.34	-4,672	4,712	3.41
	500	(5, 20)	9.2	3.03	-13,074	13,107	6.08	11.3	2.62	-13,450	13,492	7.48
		(10, 40)	12.6	2.36	-27,505	27,556	22.38	15.5	2.16	-27,580	27,642	29.28

Table 2 reports the performances of different algorithms by averaging over 100 simulation runs. We exclude all rider locations with predicted weights smaller than 0.01 to ensure numerical stability. The table includes several metrics: "Locs" refers to the number of predicted locations; "WD" refers to the Wasserstein distance between the predicted location weights and the underlying true location weights; "Lkd" and "BIC" refer to the log-likelihood and BIC values respectively. We also report computation time measured in seconds. We have the following key observations. (1) All algorithms significantly outperform the baseline clustering algorithm in terms of the Wasserstein distance and the BIC value. (2) The all-in algorithm significantly overestimates the number of rider locations especially when the set of candidate locations is large (10×10 Cartesian grid). This leads to worse predictive accuracy in terms of Wasserstein distance and BIC value. Another observation about the all-in algorithm is that the Wasserstein distance does not significantly decrease as one gets more data, which is likely related to stricter identifiability conditions caused by a large set of candidate locations (Theorem 1). (3) The location-discovery algorithms have overall the best performance and the batch mode improves over the single mode as evidenced by the lower Wasserstein distance and similar BIC values. The batch mode does discover more locations than the single mode.

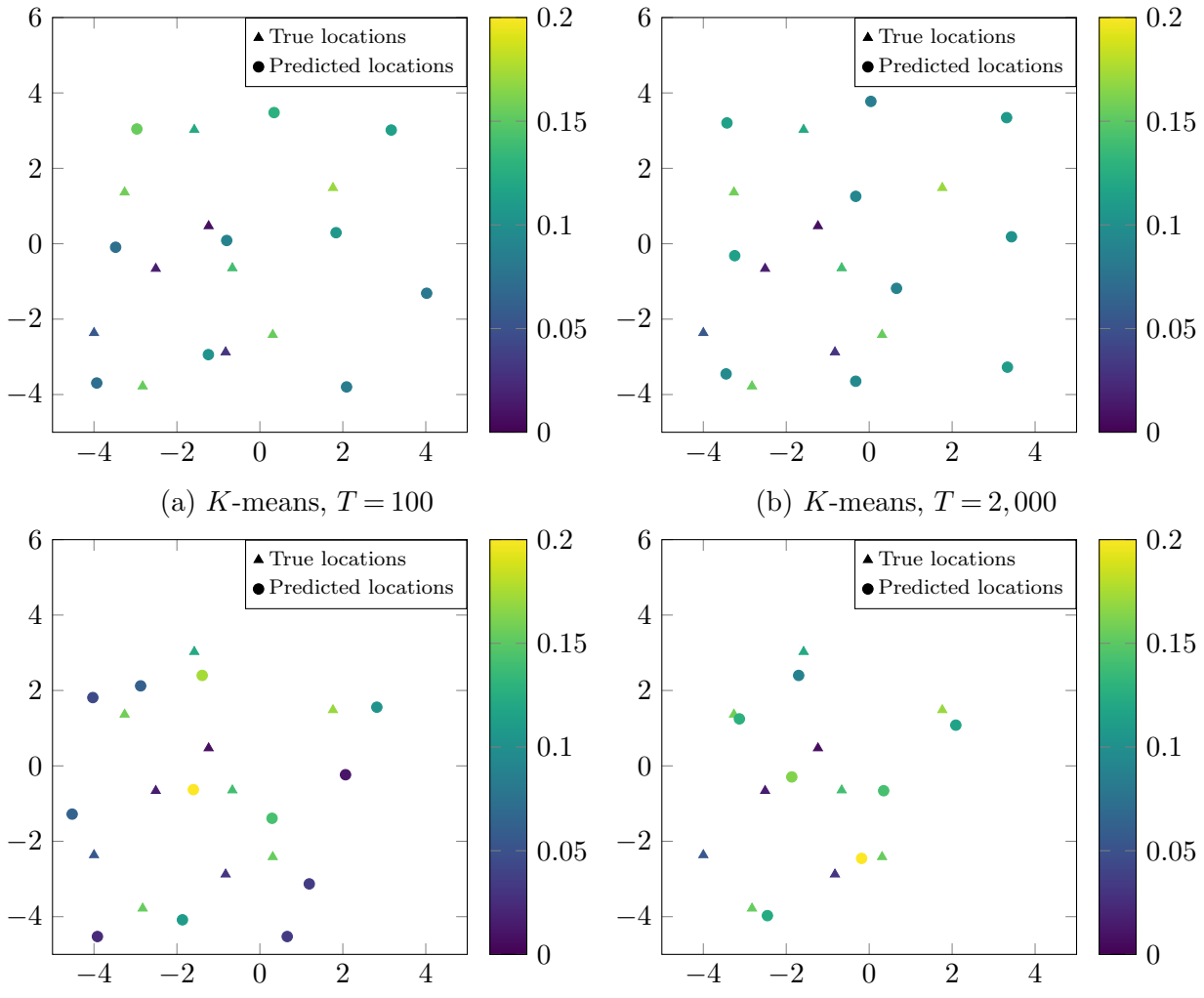
We now consider a different set of experiments where no prior information is available about rider candidate locations other than they live in a square service region. In other words, we allow rider arrival locations to be arbitrary places within the service region. We make the following adjustments to various aforementioned algorithms to fit this setting. (1) We change the clustering algorithm to a K-means clustering algorithm. We give this K-means clustering some advantages by assuming that the number of underlying rider arrival locations L^* is known beforehand. That is, we set the number of clusters $K = L^*$. (2) For the all-in algorithm, we continue to use a 10×10 Cartesian grid overlaid onto the square service region, creating a set of 100 candidate rider locations. (3) For the location-discovery algorithms, we implement a *two-round* grid search. In the single mode, we begin our search by initializing a coarse Cartesian grid of dimensions 10×10 onto the square service region. We find the rider location in the grid that maximizes the partial derivative. We then perform a second grid search that is confined to a smaller square region whose boundary is defined by the neighboring locations of the one selected from the first round. We again overlay a 10×10 Cartesian grid onto this smaller square region and identify the location having the largest partial derivative. In the batch mode, we select all local maxima in the first round, and a second grid search is conducted near all locations selected in the first round.

Table 3 Prediction performance when the set of candidate rider locations is unknown.

T	(L^*, B)	Algorithm	Locs	WD	Lkd	BIC	Time	Algorithm	Locs	WD	Lkd	BIC	Time
100	(5,20)	K-means	5.0	2.96	-2,302	2,316	0.04	All-in	26.4	2.95	-2,183	2,259	3.54
	(10,40)		10.0	2.48	-4,607	4,639	0.07		27.6	2.53	-4,488	4,576	14.23
500	(5,20)		5.0	3.05	-12,822	12,840	0.04		28.1	2.89	-12,244	12,346	16.60
	(10,40)		10.0	2.46	-26,577	26,616	0.17		29.5	2.58	-25,988	26,105	57.56
100	(5,20)	Loc. Disc.	6.4	3.04	-2,204	2,222	1.54	Loc. Disc.	9.1	2.73	-2,185	2,211	2.42
	(10,40)	(Single)	8.6	2.58	-4,505	4,532	7.32	(Batch)	11.8	2.27	-4,491	4,528	10.40
500	(5,20)		7.6	2.86	-12,269	12,297	11.27		10.2	2.43	-12,242	12,280	14.83
	(10,40)		10.5	2.34	-26,005	26,047	58.10		14.2	1.99	-25,979	26,036	67.49

Table 3 reports the performance of various algorithms under this setting averaged over 100 simulation runs. Qualitatively it is very similar to Table 2. Given the advantage of knowing the exact number of ground-truth rider arrival locations L^*, the K-means algorithm produces a relatively lower Wasserstein distance. However, similar to Table 2, it fails to converge to the true location as the length of the arrival period T increases. In Figure 2, we visualize an instance of the predicted locations and their corresponding weights with $L^* = 10$, $B = 40$ and $\lambda = 10$, under the K-means algorithm and the location-discovery algorithm with batch mode when $T = 100$ and $T = 2,000$. We observe that as T increases from 100 to $2,000$, the predicted locations in the discovery algorithm approach the true locations more closely, whereas the K-means algorithm does not exhibit any significant improvement. The all-in algorithm performs quite poorly in this setting. Similar to

Table 2, it significantly overestimates the number of rider locations. Moreover, it struggles at balancing the granularity and complexity of the model — increasing the size of the grid makes location estimates more granular but at the expense of a model harder to identify and expensive to estimate. This is reflected by the observation that its performance is quite insensitive to sample sizes. In contrast, our location discovery algorithms, especially the one with batch addition, achieve improved results with larger sample sizes, as demonstrated by the significantly lower Wasserstein distance obtained for $T = 500$ compared to $T = 100$.



(a) K-means, $T = 100$

(b) K-means, $T = 2,000$

(c) Loc. Disc. (Batch), $T = 100$

(d) Loc. Disc. (Batch), $T = 2,000$

Figure 2 **Predicted and true rider locations and their corresponding weights.** We use triangles to represent the underlying true locations and circles to represent the predicted locations. Different colors represent different weights of the corresponding locations.

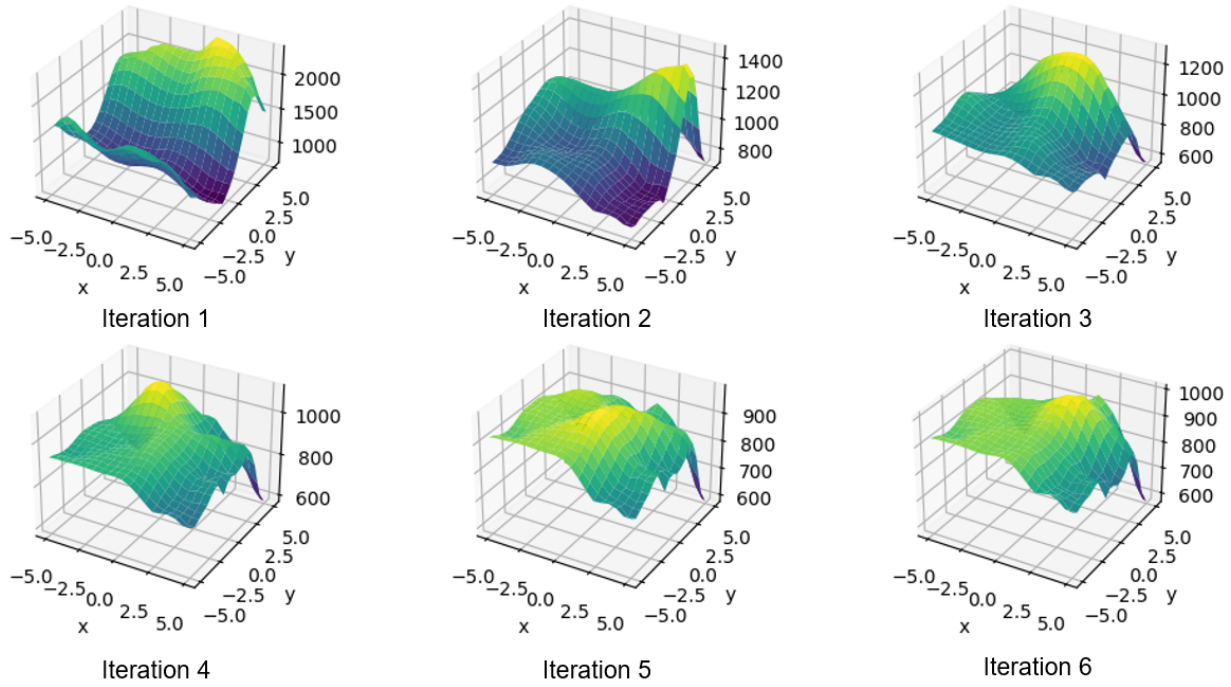


Figure 3 **Objective values** (12) **of all locations in the square service region at each iteration of the location-discovery algorithm with batch addition.**

Finally, Figure 3 displays the objective values (12) of all locations in the square region during the first six iterations of the location discovery algorithm with batch addition, on an instance with $B = 40$, $L^* = 10$, $\lambda = 10$ and $T = 100$. The peak values undergo a sharp slump in the first three iterations and then fluctuate around $1,000$. Moreover, iteration by iteration, the function becomes more multi-modal — the first graph only depicts two local maxima whereas the last graph exhibits more than five local maxima.

In Section B.2 of the appendix, we provide additional computational results when the choice model parameter $\beta_{1,l}$ is unknown. By slightly extending the EM algorithm to jointly estimate the model parameter in each iteration, both the location weights and the value of $\beta_{1,l}$ can be estimated with high accuracy (see Table EC.1 of the appendix).

5.2. Experiments Based on Seattle Bike-Sharing Data

We now run experiments using data from Seattle's dockless bike-sharing system. The data set records all bookings and real-time bike locations and statuses from a dockless bike-sharing company in the Seattle region during July 2019. The data set and code can be accessed at `https://github.com/angxu1/bike_sharing` which contains a detailed description. We confine our analysis to the downtown Seattle area (see Figure 4 for the service region) and look at time periods from 5 pm to 7 pm every day in July 2019. There are $3,497$ bookings in total and 438 bikes.

We use an MNL model to fit rider choice behaviors. Kabra et al. (2019) estimates a structural demand model using Paris bike-sharing data. Their user choice model is specified as an MNL model with walking distance (in km) as the feature and fixed time and location effects. They specify the disutility of walking distance using a piecewise linear structure where the coefficient of walking distance is estimated to be -2.229 for walking distance less than 300 meters and -15.445 for walking distance greater than 300 meters. In our model, we set $\beta_{0,l} = 1$ and $\beta_{1,l} = -5$ for all rider locations l. These two values are selected by searching within the neighborhood informed by the aforementioned model parameters estimated in Kabra et al. (2019) to find values that lead to the lowest BIC value.

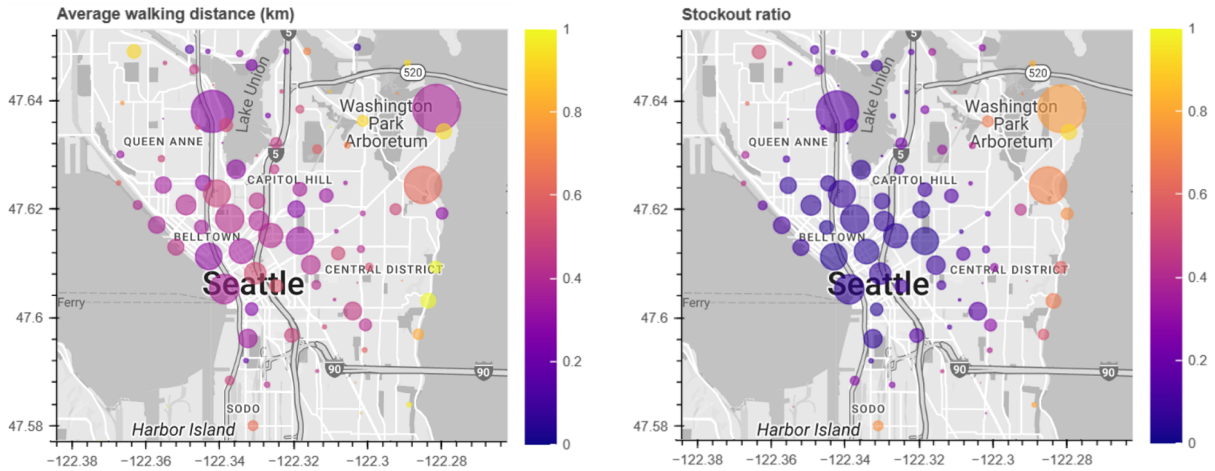
We compare the following four different algorithms: (1) an all-in algorithm with 20×20 candidate locations; (2) an all-in algorithm with 40×40 candidate locations; (3) a location-discovery algorithm of single mode with a two-round grid search each with a granularity of 20×20; (4) a location-discovery algorithm of batch mode with a two-round grid search each with a granularity of 20×20, and we set the maximum number of discovered locations at each iteration to be 10. We initialize the location-discovery algorithms by randomly generating two coordinates on the service region as initial locations. We also require a minimum of 75 locations to be generated as our service region is comprised of approximately 75 census tracts (GeoData 2020).

We partition our data into training and testing sets. We implement the aforementioned algorithms for the training set and evaluate their performances on the testing set. The training and testing sets consist of bookings ranging from July 1$^{\text{st}}$ to July 21$^{\text{st}}$ and from July 22$^{\text{nd}}$ to July 31$^{\text{st}}$, respectively. Based on our predictions of locations $\hat{\mathcal{L}}$ and weights $\hat{\mathbf{w}}$ using the training set, then the arrival rate $\hat{\lambda}$ can be estimated as $\hat{\lambda} = N_{\text{train}} / \int_0^{T_{\text{train}}} \left(1 - \sum_{l \in \hat{\mathcal{L}}} \hat{w}_l p_{l,0,S_t}\right) dt$ by equation (4) where N_{train} is the number of bookings observed in the training set and T_{train} is the length of the training period. Let T_{test} be the length of the testing period. Then an estimate of the number of bookings in the testing period $\hat{N}_{\text{test}}$ can be computed as $\hat{N}_{\text{test}} = \hat{\lambda} \int_{T_{\text{train}}}^{T_{\text{train}}+T_{\text{test}}} \left(1 - \sum_{l \in \hat{\mathcal{L}}} \hat{w}_l p_{l,0,S_t}\right) dt$. We measure the predictive accuracy of the number of ridership in terms of the mean absolute percentage error (MAPE), which can be calculated as $\text{MAPE} = |\hat{N}_{\text{test}} - N_{\text{test}}| / N_{\text{test}} \times 100\%$. We also measure the log-likelihood value and BIC value on the testing data.

Table 4 reports the performance of all tested algorithms. The all-in algorithm with 40×40 candidate locations performs the worst — increasing the grid size from 20×20 to 40×40 incurs a larger BIC value for the all-in algorithm. The two location-discovery algorithms achieve better performance in both training and testing data. The single and batch modes are comparable in their prediction accuracies while the batch mode has significantly lower computation time. This is likely due to the saved EM iterations thanks to the batch addition of new locations.

Table 4 Prediction performance on Seattle data.

	All-in (20×20)		All-in (40×40)		Loc. Disc. (Single)		Loc. Disc. (Batch)	
	Train	Test	Train	Test	Train	Test	Train	Test
Locs	147	–	224	–	75	–	81	–
Lkd	-43,862	-23,229	-43,837	-23,252	-43,869	-23,282	-43,900	-23,288
BIC	44,426	23,749	44,697	24,046	44,159	23,547	44,212	23,575
MAPE (%)	–	12.9	–	13.0	–	5.6	–	5.4
Time (min:sec)	29:34	–	179:18	–	92:01	–	42:25	–



(a) Average walking distance (km)

(b) Predicted stockout ratio for riders

Figure 4 The average walking distance and bike stockout ratio for each arrival location in the Seattle area. Different colors represent different average walking distances (left) or stockout ratios (right). Each circle represents a discovered location. For a clear view, the radius of each circle is computed as a log-linear transformation of the weight in that location (a larger radius corresponds to a larger weight). Specifically, the radius for location l in the plot is $\log(\hat{w}_l) + 2.3$.

Figure 4 depicts two important service level metrics based on the estimation results using the batch version of the location-discovery algorithm based on all data in July 2019. Figure 4a shows the average walking distance at each rider's arrival location conditional on riders choosing to pick up bikes. Formally, for each location $l \in \mathcal{L}$, this quantity can be computed as $(1/N) \cdot \sum_{n=1}^{N} \sum_{b=1}^{B} p_{l,b,t_n} d_{l,b,t_n} / (1 - p_{l,0,t_n})$. Figure 4b depicts the stockout ratio of each location, which is the probability that an arriving rider chooses to leave without picking up a bike. These two metrics do not have to be perfectly (positively) correlated. For example, the arrival location on the east of Washington Park Arboretum has a relatively low average walking distance but a high stockout ratio. This likely suggests that there is a scarcity of bikes but once there is a bike, it is often close to the rider's location. Some operational insights can be inferred from these results. Rider locations along Lake Washington (the east side of the service region) and in Queen Anne have

particularly low service levels and large potential demand, calling for more bike allocations. These bikes can potentially be relocated from the downtown area where bikes are well stocked and service levels are high. These performance metrics reveal hidden insights and add value to the existing monitoring tools used by municipal agencies which only visualize trip data (Seattle Department of Transportation 2022).

6. Concluding Remarks

In this paper, we present a locational demand model for a bike-sharing system (free-floating or docked-based) that aims to recover rider arrival locations and their corresponding intensities (weights) using only rider booking and vehicle availability data. We give conditions under which the location weights are identifiable and can be consistently estimated and devise a simple and efficient EM algorithm with closed-form updates. This EM algorithm is further complemented by a location-discovery procedure that allows it to scale to larger instances and handle cases where the set of candidate rider locations is unknown a priori. Computational experiments based on both synthetic data and real data from the Seattle area demonstrate its effectiveness over various benchmarks.

Acknowledgments

The first and second authors gratefully acknowledge the support of Pacific Northwest Transportation Consortium (PacTrans) and the Seattle Department of Transportation (SDOT) which provide research funding and support for this project.

References

Abdallah T, Vulcano G (2020) Demand estimation under the multinomial logit model from sales transaction data. *Manufacturing & Service Operations Management* 23(5):1196–1216.

Banerjee S, Freund D, Lykouris T (2022) Pricing and optimization in shared vehicle systems: An approximation framework. *Operations Research* 70(3):1783–1805.

Bhat CR (1997) An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science* 31(1):34–48.

Bureau of Transportation Statistics (2023) Bikeshare and e-scooters in the u.s. `https://data.bts.gov/stories/s/Bikeshare-and-e-scooters-in-the-U-S-/fwcs-jprj/`, accessed: 2023-05-04.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.

El-Assi W, Mahmoud MS, Habib KN (2017) Effects of Built Environment And Weather on Bike Sharing Demand: A Station Level Analysis of Commercial Bike Sharing In Toronto. *Transportation* 44(3):589–613.

Freund D, Henderson SG, Shmoys DB (2019) Bike sharing. Hu M, ed., *Sharing Economy: Making Supply Meet Demand*, volume 6 of *Springer Series in Supply Chain Management*, 435 – 459 (Springer).

Gaur V, Honhon D (2006) Assortment planning and inventory decisions under a locational choice model. *Management Science* 52(10):1528–1543.

GeoData S (2020) 2020 Census Tract Seattle. `https://data-seattlecitygis.opendata.arcgis.com/datasets/SeattleCityGIS::2020-census-tract-seattle-redistricting-data-1990-2020/`, accessed: 2023-03-17.

Greene WH, Hensher DA (2003) A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological* 37(8):681–698.

Grün B, Leisch F (2008) Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification* 25(2):225–247.

He P, Zheng F, Belavina E, Girotra K (2021) Customer preference and station network in the london bike-share system. *Management Science* 67(3):1392–1412.

Hotelling H (1929) Stability in competition. *The Economic Journal* 39(153):41–57.

Hunter DR, Lange K (2000) Optimization transfer using surrogate objective functions: Rejoinder. *Journal of Computational and Graphical Statistics* 9(1):52–59.

Hunter DR, Lange K (2004) A tutorial on MM algorithms. *The American Statistician* 58(1):30–37.

Jagabathula S, Subramanian L, Venkataraman A (2020) A conditional gradient approach for nonparametric estimation of mixing distributions. *Management Science* 66(8):3635–3656.

Kabra A, Belavina E, Girotra K (2019) Bike-share systems: Accessibility and availability. *Management Science* 66(9):3803–3824.

Lancaster K (1975) Socially optimal product differentiation. *American Economic Review* 65(4):567–85.

Lancaster KJ (1966) A new approach to consumer theory. *Journal of Political Economy* 74(2):132–157.

Mellou K, Jaillet P (2019) Dynamic resource redistribution and demand estimation: An application to bike sharing systems. *Available at SSRN 3336416* .

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

O'Mahony E (2015) *Smarter tools for (Citi) bike sharing*. Ph.D. thesis, Cornell University.

O'Mahony E, Shmoys DB (2015) Data Analysis and Optimization for (Citi) Bike Sharing. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 687–694.

Open Mobility Foundation (2022) Mobility-data-specification: A data standard to enable right-of-way regulation and two-way communication between mobility companies and local governments. `https://github.com/openmobilityfoundation/mobility-data-specification/`, accessed: 2022-11-29.

Rixey RA (2013) Station-level forecasting of bikesharing ridership: Station network effects in three us systems. *Transportation research record* 2387(1):46–55.

Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* 461–464.

Seattle Department of Transportation (2022) Scooter share data & permitting - transportation. `https://www.seattle.gov/transportation/projects-and-programs/programs/new-mobility-program/scooter-share`, accessed: 2023-05-14.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures On Stochastic Programming: Modeling and Theory* (SIAM).

Singhvi D, Singhvi S, Frazier PI, Henderson SG, O'Mahony E, Shmoys DB, Woodard DB (2015) Predicting Bike Usage for New York City's Bike Sharing System. *AAAI Workshop On Computational Sustainability*.

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

van der Vaart AW (2000) *Asymptotic statistics*, volume 3 (Cambridge university press).

van Ryzin G, Vulcano G (2014) A Market Discovery Algorithm to Estimate a General Class of Nonparametric Choice Models. *Management Science* 61(2):281–300.

Vulcano G, Van Ryzin G, Ratliff R (2012) Estimating Primary Demand for Substitutable Products from Sales Transaction Data. *Operations Research* 60(2):313–334.

Wu CF (1983) On the convergence properties of the EM algorithm. *The Annals of Statistics* 11(1).

Online Appendix

Appendix A: Proofs of Technical Results

Proof of Theorem 1. To prove the sufficiency, we show that there exists a unique maximizer of $\boldsymbol{w}$ and λ corresponding to the true parameters for the *long-run average expected log-likelihood* function, under the condition that the set of vectors $\left\{[p_{1,b,S_k},\cdots,p_{L,b,S_k}]^\intercal : b\in\mathcal{B}_k, k\in\{1,\cdots,K\}\right\}$ spans $\mathbb{R}^L$. That is, we want to show that any maximizer $(\lambda,\boldsymbol{w})$ of the problem $\max_{\lambda,\boldsymbol{w}}\lim_{T\to\infty}(1/T)\mathbb{E}\left[l_I(\lambda,\boldsymbol{w})\right]$ satisfies $(\lambda,\boldsymbol{w})=(\lambda^*,\boldsymbol{w}^*)$ where λ^* and $\boldsymbol{w}^*$ are the ground truths of arrival rate and location weight vector. Here, the expectation is taken over bookings. We start with the incomplete data log-likelihood function given bookings over a period of length T,

$$l_I(\lambda,\boldsymbol{w})=-\lambda\int_0^T\left(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t}\right)dt+N\log\lambda+\sum_{n=1}^{N}\log\sum_{l\in\mathcal{L}}w_l p_{l,b_n,S_{t_n}}. \tag{EC.1}$$

Recall that $s=\int_0^T(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t})dt=\int_0^T\sum_{l\in\mathcal{L}}\sum_{b\in\mathcal{B}}w_l p_{l,b,S_t}dt$ and we define a similar term $s^*:=\int_0^T(1-\sum_{l\in\mathcal{L}}w_l^* p_{l,0,S_t})dt$ corresponding to the ground-truth weights. The expected log-likelihood can then be computed as follows

$$\begin{aligned}\mathbb{E}\left[l_I(\lambda,\boldsymbol{w})\right]&=-\lambda s+\mathbb{E}[N]\left(\log\lambda+\mathbb{E}\left[\log\sum_{l\in\mathcal{L}}w_l p_{l,b_n,S_{t_n}}\right]\right) && \text{(Wald's Lemma)}\\ &=-\lambda s+\lambda^* s^*\left(\log\lambda+\int_0^T\sum_{b\in\mathcal{B}}\left(\left(\frac{\sum_{l\in\mathcal{L}}w_l^* p_{l,b,S_t}}{s^*}\right)\left(\log\sum_{l\in\mathcal{L}}w_l p_{l,b,S_t}\right)\right)dt\right). && \text{(EC.2)}\end{aligned}$$

By following the first-order condition on λ, the unique maximizer $\hat{\lambda}$ of equation (EC.2) satisfies $\hat{\lambda}=\lambda^* s^*/s$. Again, for convenience, we define $\mathbb{E}[l_I(\boldsymbol{w})]:=\mathbb{E}[l_I(\hat{\lambda},\boldsymbol{w})]$,

$$\begin{aligned}\mathbb{E}[l_I(\boldsymbol{w})]&=-\lambda^* s^*+\lambda^* s^*\log\lambda^*-\lambda^* s^*\log\frac{s}{s^*}+\lambda^*\int_0^T\sum_{b\in\mathcal{B}_t}\left(\left(\sum_{l\in\mathcal{L}}w_l^* p_{l,b,S_t}\right)\log\left(\sum_{l\in\mathcal{L}}w_l p_{l,b,S_t}\right)\right)dt\\ &=\lambda^*\int_0^T\sum_{b\in\mathcal{B}_t}\left(\left(\sum_{l\in\mathcal{L}}w_l^* p_{l,b,S_t}\right)\log\left(s^*\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_t}}{s}\right)\right)dt-\lambda^* s^*+\lambda^* s^*\log\lambda^*\\ &=\lambda^* s^*\int_0^T\sum_{b\in\mathcal{B}_t}\left(\left(\sum_{l\in\mathcal{L}}\frac{w_l^* p_{l,b,S_t}}{s^*}\right)\log\left(s^*\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_t}}{s}\right)\right)dt-\lambda^* s^*+\lambda^* s^*\log\lambda^*.\end{aligned}$$

Recall that for each $k\in\{1,\cdots,K\}$, bike pattern S_k appears with a positive fraction of time $\alpha_k>0$ such that $\sum_{k=1}^K\alpha_k=1$. Let $\bar{s}:=\lim_{T\to\infty}(1/T)\int_0^T(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_t})dt=\sum_{k=1}^K\alpha_k\left(1-\sum_{l\in\mathcal{L}}w_l p_{l,0,S_k}\right)$ and $\bar{s}^*:=\lim_{T\to\infty}(1/T)\int_0^T(1-\sum_{l\in\mathcal{L}}w_l^* p_{l,0,S_t})dt=\sum_{k=1}^K\alpha_k\left(1-\sum_{l\in\mathcal{L}}w_l^* p_{l,0,S_k}\right)$ be the long-run averages. It is clear that $\lim_{T\to\infty}s^*/s=\bar{s}^*/\bar{s}$. Let $\mathcal{B}_k$ be the set of available bikes under bike pattern S_k. We can then write the long-run average expected log-likelihood function as

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[l_I(\boldsymbol{w})\right]\\ &=\lim_{T\to\infty}\lambda^*\bar{s}^*\int_0^T\sum_{b\in\mathcal{B}_k}\left(\left(\sum_{l\in\mathcal{L}}\frac{w_l^* p_{l,b,S_t}}{T\bar{s}^*}\right)\log\left(\bar{s}^*\sum_{l\in\mathcal{L}}\frac{w_l p_{l,b,S_t}}{\bar{s}}\right)\right)dt-\lambda^*\bar{s}^*+\lambda^*\bar{s}^*\log\lambda^*\end{aligned}$$

$$
\begin{aligned}
&= \lim_{T\to\infty} \lambda^* \bar{s}^* \frac{1}{T} \int_0^T \sum_{b\in\mathcal{B}_t}^{B} \left(\left(\sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S_t}}{\bar{s}^*} \right) \left(\log \bar{s}^* + \log \left(\sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S_t}}{\bar{s}} \right) \right) \right) dt - \lambda^* \bar{s}^* + \lambda^* \bar{s}^* \log \lambda^* \\
&= \lim_{T\to\infty} \lambda^* \bar{s}^* \frac{1}{T} \int_0^T \sum_{b\in\mathcal{B}_t} \left(\left(\sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S_t}}{\bar{s}^*} \right) \log \left(\sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S_t}}{\bar{s}} \right) \right) dt + \underbrace{\lambda^* \left(-\bar{s}^* + \bar{s}^* \log \lambda^* + \bar{s}^* \log \bar{s}^* \right)}_{c_0} \\
&= \lambda^* \bar{s}^* \sum_{k=1}^{K} \left(\alpha_k \sum_{b\in\mathcal{B}_k} \left(\sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \left(\sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S_k}}{\bar{s}} \right) \right) \right) + c_0, \qquad \text{(EC.3)}
\end{aligned}
$$

where c_0 is a constant that is independent of $\boldsymbol{w}$. Due to the fact that

$$
\sum_{k=1}^{K} \left(\alpha_k \sum_{b\in\mathcal{B}_k} \left(\sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \right) = \sum_{k=1}^{K} \left(\alpha_k \sum_{b\in\mathcal{B}_k} \left(\sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S_k}}{\bar{s}} \right) \right) = 1,
$$

we can use Gibb's inequality to derive an upper bound of the likelihood function by adding and subtracting a term α_k inside the logarithm function in equation (EC.3),

$$
\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[l_I(\boldsymbol{w}) \right] &= \lambda^* \bar{s}^* \sum_{k=1}^{K} \sum_{b\in\mathcal{B}_k} \left(\alpha_k \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \left(\alpha_k \sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S_k}}{\bar{s}} \right) \right) \\
&\quad - \lambda^* \bar{s}^* \sum_{k=1}^{K} \sum_{b\in\mathcal{B}_k} \left(\alpha_k \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \alpha_k \right) + c_0 \\
&\le \lambda^* \bar{s}^* \sum_{k=1}^{K} \sum_{b\in\mathcal{B}_k} \left(\alpha_k \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \left(\alpha_k \sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \right) \\
&\quad - \lambda^* \bar{s}^* \sum_{k=1}^{K} \sum_{b\in\mathcal{B}_k} \left(\alpha_k \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \alpha_k \right) + c_0 \qquad \text{(Gibb's Inequality)} \\
&= \lambda^* \bar{s}^* \sum_{k=1}^{K} \sum_{b\in\mathcal{B}_k} \left(\alpha_k \sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) \log \sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S_k}}{\bar{s}^*} \right) + c_0.
\end{aligned}
$$

The equality holds if and only if

$$
\sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S_k}}{\bar{s}} = \sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S_k}}{\bar{s}^*}, \qquad \text{(EC.4)}
$$

for all $b \in \mathcal{B}_k$ and $k \in \{1, \cdots, K\}$. Now consider equations (EC.4) as a system of linear equations with respect to $\boldsymbol{w}/\bar{s}$. Since the set of vectors $\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}] : b \in \mathcal{B}_k, k \in \{1, \cdots, K\}\}$ spans the space $\mathbb{R}^L$, then the coefficient matrix for this system of linear equations has full column rank L. Thus, (EC.4) has no solution or exactly one solution. On the other hand, observe that $\boldsymbol{w}/\bar{s} = \boldsymbol{w}^*/\bar{s}^*$ is a solution. This implies that it is a unique solution. Since $1/\bar{s} = \sum_{l\in\mathcal{L}} w_l/\bar{s} = \sum_{l\in\mathcal{L}} w_l^*/\bar{s}^* = 1/\bar{s}^*$, this implies $\bar{s} = \bar{s}^*$ and thus $\boldsymbol{w} = \boldsymbol{w}^*$ is the unique solution to (EC.4) satisfying $\sum_{l\in\mathcal{L}} w_l = 1$. Since the unique maximizer of λ is $\lambda^* s^*/s$, we have proved that $(\lambda^*, \boldsymbol{w}^*)$ is the unique maximizer for $\lim_{T\to\infty}(1/T)\mathbb{E}\left[l_I(\lambda, \boldsymbol{w}) \right]$, which implies the identifiability of the model. The sufficiency is trivial. To show the necessity under the condition $w_l^* = 0$ for at most one $l \in \mathcal{L}$, we establish the existence of some $\boldsymbol{w} \neq \boldsymbol{w}^*$ such that $\boldsymbol{w}$ corresponds to the same generating distribution as $\boldsymbol{w}^*$ if the set of vectors $\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}] : b \in \mathcal{B}_k, k \in \{1, \cdots, K\}\}$ cannot span the space $\mathbb{R}^L$. We start by finding a solution set $\{w_1, \cdots, w_N, \bar{s}\}$ other than the true values regarding the following system of linear equations

$$
\begin{cases}
\sum_{l\in\mathcal{L}} w_l p_{l,b,S_k} - \bar{s} \sum_{l\in\mathcal{L}} w_l^* p_{l,b,S_k}/\bar{s}^* = 0, & \forall b \in \mathcal{B}, k \in \{1, \cdots, K\}, \\
\sum_{l\in\mathcal{L}} w_l = 1.
\end{cases} \qquad \text{(EC.5)}
$$

Note that the constraint $\bar{s} = \sum_{k=1}^{K} \left(\alpha_k \sum_{b \in \mathcal{B}_k} \sum_{l \in \mathcal{L}} w_l p_{l,b,S_k} \right)$ is implied from (EC.5). Since $\sum_{l \in \mathcal{L}} w_l^* p_{l,b,S_k} / \bar{s}^*$ is a linear combination of $p_{1,b,S_k}, \cdots, p_{L,b,S_k}$, we know the dimension of the vector space generated by $\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}, \sum_{l \in \mathcal{L}} w_l^* p_{l,b,S_k} / \bar{s}^*] : b \in \mathcal{B}_k, k \in \{1, \cdots, K\}\}$ is strictly less than L. Hence, the rank of the coefficient matrix of the system of linear equations (EC.5) is strictly less than $L+1$. This implies that the system has infinitely many solutions — there exists a nonzero $\theta \in \mathbb{R}^L$ such that $\boldsymbol{w}^* + \beta \boldsymbol{\theta}$ satisfies (EC.5) for all $\beta \in \mathbb{R}$. Since we have $w_l^* = 0$ for at most one $l \in L$, there always exists some small $\beta \neq 0$ such that the resulting solution $\boldsymbol{w}' = \boldsymbol{w}^* + \beta \boldsymbol{\theta}$ is feasible but differs from the underlying truth $\boldsymbol{w}^*$. We now show that $\boldsymbol{w}'$ and $\boldsymbol{w}^*$ imply the same data-generating distribution of booking data. We consider the data-generating distribution of bookings over any length of arrival period T such that $\int_0^T \mathbf{1}(S_t = S_k) dt / T = \alpha_k > 0$, $\forall k \in \{1, \cdots, K\}$.

$$
\begin{aligned}
l_I(\boldsymbol{w}) &= -N + N \log N - N \log \int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t} \right) dt + \sum_{n=1}^{N} \log \left(\sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}} \right) \\
&= -N + N \log N + \sum_{n=1}^{N} \log \frac{\sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}}}{\int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t} \right) dt} \\
&= -N + N \log N + \sum_{n=1}^{N} \log \frac{\sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}}}{\bar{s}}. \qquad \text{(EC.6)}
\end{aligned}
$$

Since $\boldsymbol{w}'$ and $\boldsymbol{w}^*$ both satisfy the system of linear equations (EC.5), from (EC.6), we have $l_I(\boldsymbol{w}') = l_I(\boldsymbol{w}^*)$ for any booking data, which violates the identifiability of the model. This completes the proof. □

Proof of Proposition 1. We show the result by invoking Theorem 5.3 in Shapiro et al. (2009). In particular, we show that there exists a compact set C that satisfies all four conditions in Theorem 5.3 of Shapiro et al. (2009): (i) the true location weights $\boldsymbol{w}^*$ is contained in C; (ii) the long-run-average expected log-likelihood function $\lim_{T \to \infty} (1/T) \mathbb{E}[l_I(\boldsymbol{w})]$ is finite valued and continuous on C; (iii) the empirical time-averaged log-likelihood function $(1/T) l_I(\boldsymbol{w})$ converges to $\lim_{T \to \infty} (1/T) \mathbb{E}[l_I(\boldsymbol{w})]$ with probability one as $T \to \infty$, uniformly in $\boldsymbol{w} \in C$; and (iv) with probability one, for T large enough the set of maximizers of $(1/T) l_I(\boldsymbol{w})$, $\hat{S}_T$, is nonempty and $\hat{S}_T \subset C$.

If we have $p_{l,b_n,S_{t_n}} > 0$ for all $l \in \mathcal{L}$ and $n \in \{1, \cdots, N\}$, we know that $\sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}} > 0$ for all $\boldsymbol{w}$ belonging to the simplex $\{\boldsymbol{w} \geq 0 : \sum_{l \in \mathcal{L}} w_l = 1\}$. Then C can simply be this simplex which is a compact set. Similarly, if we have $w_l^* \geq \epsilon$, $\forall l \in \mathcal{L}$ for some $\epsilon > 0$, the compact set C can be $C := \{w_l \geq \epsilon, \forall l \in \mathcal{L} : \sum_{l \in \mathcal{L}} w_l = 1\}$. Then conditions (i), (ii), and (iv) in Theorem 5.3 are satisfied. It remains to show the uniform convergence of the time-averaged log-likelihood function (condition (iii)).

To show the uniform convergence, we invoke Theorem 7.48 of Shapiro et al. (2009) which gives sufficient conditions of uniform convergence of $(1/T) l_I(\boldsymbol{w})$ to the long-run average expected log-likelihood function $\lim_{T \to \infty} (1/T) \mathbb{E}[l_I(\boldsymbol{w})]$ as $T \to \infty$. It requires that: (a) for any $\boldsymbol{w} \in C$, $l_I(\boldsymbol{w})$ is continuous at $\boldsymbol{w}$ for almost all booking data; (b) $l_I(\boldsymbol{w})$, $\boldsymbol{w} \in C$ is dominated by an integrable function that only depends on data; (c) the sample is IID. Condition (a) clearly holds. We then show that the log-likelihood is dominated by another integrable function on the compact set C. For the case where $p_{l,b,S_k} > 0$ for all $l \in \mathcal{L}, b \in \mathcal{B}_k, k \in \{1, \cdots, K\}$, let $\delta = \min_{l \in \mathcal{L}, b \in \mathcal{B}_k, k \in \{1, \cdots, K\}} p_{l,b,S_k}$. We have

$$
|l_I(\boldsymbol{w})| = \left| -N + N \log N - N \log \int_0^T \left(1 - \sum_{l \in \mathcal{L}} w_l p_{l,0,S_t} \right) dt + \sum_{n=1}^{N} \log \left(\sum_{l \in \mathcal{L}} w_l p_{l,b_n,S_{t_n}} \right) \right|
$$

$$< N + N\log(N) + N|\log T| + N\left|\log\delta\right| + N\left|\log\delta\right|, \tag{EC.7}$$

which holds for all $\boldsymbol{w} \in C$. In the case that $w_l^* \geq \epsilon$, $\forall l \in \mathcal{L}$, we have $C = \{w_l \geq \epsilon,\ \forall l \in \mathcal{L} : \sum_{l\in\mathcal{L}} w_l = 1\}$. With a slight abuse of notation, let $\delta = \min_{l\in\mathcal{L}, b\in\mathcal{B}_k, k\in\{1,\cdots,K\}: p_{l,b,S_k}>0} p_{l,b,S_k}$. We have for all $\boldsymbol{w} \in C$,

$$T\delta\epsilon \leq \int_0^T \left(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t}\right) dt \leq T.$$

This gives that, for all $\boldsymbol{w} \in C$,

$$\begin{aligned} |l_I(\boldsymbol{w})| &= \left| -N + N\log N - N\log\int_0^T \left(1 - \sum_{l\in\mathcal{L}} w_l p_{l,0,S_t}\right) dt + \sum_{n=1}^N \log\left(\sum_{l\in\mathcal{L}} w_l p_{l,b_n,S_{t_n}}\right)\right| \\ &< N + N\log(N) + N|\log T| + N\left|\log\delta\epsilon\right| + N\left|\log\delta\epsilon\right|, \end{aligned} \tag{EC.8}$$

Note that (EC.7) and (EC.8) are both integrable with respect to the data-generating distribution, so the dominance property holds. To meet the IID condition, we can always reshuffle the booking data into IID episodes with equal length T' such that within each episode, the data-generating distribution of bookings satisfies $\int_0^{T'} \mathbf{1}(S_t = S_k)dt/T' = \alpha_k > 0$, $\forall k \in \{1,\cdots,K\}$. In such a way, the asymptotic limits of the length of the arrival period approaching infinity and the number of episodes approaching infinity become equivalent. This completes proving the uniform convergence of the time-averaged log-likelihood function. Since in Theorem 1 we already show that the long-run average expected log-likelihood has a unique maximizer at its true value $\boldsymbol{w}^*$, this completes the proof that $\hat{\boldsymbol{w}}$ converges to $\boldsymbol{w}^*$ with probability one by invoking Theorem 5.3 in Shapiro et al. (2009). □

Proof of Corollary 1. To show the sufficiency, by the Radon–Nikodym Theorem, there exists a density function for the invariant probability measure π, $f(S) > 0, S \in \mathcal{S}$, such that $\pi(A) = \int_A f(S)d\mu$ for any $A \in \mathcal{F}$. With a little abuse of notation, let $\mathcal{B}_S$ be the set of bikes in bike pattern S. Recall that $c_0 = \lambda^*(-\bar{s}^* + \bar{s}^*\log\lambda^* + \bar{s}^*\log\bar{s}^*)$ as defined in the proof of Theorem 1. Similar to equation (EC.3), we have

$$\begin{aligned} \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[l_T(\boldsymbol{w})\right] &= \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} f(S) \sum_{b\in\mathcal{B}_S} \left(\sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log\sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S}}{\bar{s}}\right) d\mu + c_0 \\ &= \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S)\sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log\sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S}}{\bar{s}}\right) d\mu + c_0 \\ &= \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S)\sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log\left(f(S)\sum_{l\in\mathcal{L}} \frac{w_l p_{l,b,S}}{\bar{s}}\right)\right) d\mu \\ &\quad - \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S)\sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log f(S)\right) d\mu + c_0 \end{aligned} \tag{EC.9}$$

Note that the Gibbs inequality is not applicable here due to the continuity of bike pattern distribution. However, a similar result can be derived from the almost positive definite property of Kullback–Leibler divergence. By Lemma 3.1 of Kullback and Leibler (1951),

$$\text{(EC.9)} \leq \lambda^*\bar{s}^* \int_{S\in\mathcal{S}} \sum_{b\in\mathcal{B}_S} \left(f(S)\sum_{l\in\mathcal{L}} \left(\frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right) \log\left(f(S)\sum_{l\in\mathcal{L}} \frac{w_l^* p_{l,b,S}}{\bar{s}^*}\right)\right) d\mu$$

$$
\begin{aligned}
&-\lambda^*\bar{s}^*\int_{S\in\mathcal{S}}\sum_{b\in\mathcal{B}_S}\left(f(S)\sum_{l\in\mathcal{L}}\left(\frac{w_l^*p_{l,b,S}}{\bar{s}^*}\right)\log f(S)\right)d\mu+c_0\\
&=\lambda^*\bar{s}^*\int_{S\in\mathcal{S}}\sum_{b\in\mathcal{B}_S}\left(f(S)\sum_{l\in\mathcal{L}}\left(\frac{w_l^*p_{l,b,S}}{\bar{s}^*}\right)\log\sum_{l\in\mathcal{L}}\frac{w_l^*p_{l,b,S}}{\bar{s}^*}\right)d\mu+c_0.
\end{aligned}
$$

The equality holds if and only if

$$
f(S)\sum_{l\in\mathcal{L}}\frac{w_lp_{l,b,S}}{\bar{s}}=f(S)\sum_{l\in\mathcal{L}}\frac{w_l^*p_{l,b,S}}{\bar{s}^*}\iff\sum_{l\in\mathcal{L}}\frac{w_lp_{l,b,S}}{\bar{s}}=\sum_{l\in\mathcal{L}}\frac{w_l^*p_{l,b,S}}{\bar{s}^*}\tag{EC.10}
$$

holds for all $b\in\mathcal{B}$ and almost all $S\in\mathcal{S}$. By the conditions in the statement of the corollary, we know $\boldsymbol{w}=\boldsymbol{w}^*$ is the unique solution for (EC.10), and thus the sufficiency holds.

To show the necessity, following the proof of Theorem 1, there exists $\boldsymbol{w}'\neq\boldsymbol{w}^*$ such that when the observed period $T\to\infty$, $(1/T)l_I(\boldsymbol{w}')-(1/T)l_I(\boldsymbol{w}^*)\to 0$ with probability 1. This contradicts to the identifiability condition, which completes the proof. □

Proof of Theorem 2. We let $\mathcal{M}=\{1,2,\cdots,M\}$ to be the set of all bike stations. We first give the formal definition of *distance rankings*. For each rider location $l\in\mathcal{L}$ and bike station $s\in\mathcal{M}$, let $d_{l,s}$ be the distance from rider location l to bike station s.

DEFINITION EC.1. A distance ranking σ_l for a rider location $l\in\mathcal{L}$ is a sequence of distinct stations $s^{(1)},s^{(2)},\cdots,s^{(|\sigma_l|)}\in\mathcal{M}$ defined as follows:

(1) $d_{l,s^{(1)}},\cdots,d_{l,s^{(|\sigma_l|)}}$ is non-decreasing;

(2) $d_{l,s^{(1)}},\cdots,d_{l,s^{(|\sigma_l|)}}\le\bar{r}_l$ and $d_{l,s}>\bar{r}_l$ for $s\in\mathcal{M}\backslash\{s^{(1)},s^{(2)},\cdots,s^{(|\sigma_l|)}\}$.

We let Σ to be the set of unique distance rankings. We comment that the cardinality $|\Sigma|$ is bounded by $\mathcal{O}(M^{2d+1})$, where d is the dimension of the service region. This is established by Skala (2009) where the author shows that the number of unique distance permutations is bounded by $\mathcal{O}(M^{2d})$. Here we have one degree higher as each distance ranking does not have to be of length M since the consideration radius is limited.

Based on the definition, we further define $\Sigma(\sigma)$ as the set of distance rankings whose first $|\sigma|$ choices are exactly the same as σ. With a slight abuse of notation, we refer to w_σ as the corresponding probability that a rider arrives at a location with distance ranking σ. Then to prove the theorem, it is equivalent to show that the MLE $\hat{w}_\sigma$ is consistent for all $\sigma\in\Sigma$. We use $\mathcal{S}$ to denote all possible station patterns, i.e., a set of available stations. For each $S\in\mathcal{S}$, we further define $\mathbb{P}_j(S)$ as the probability that a rider chooses to pick up a bike at station j under some station pattern S. Similarly, $\mathbb{P}_0(S)$ is defined as the probability that a rider chooses to leave under station pattern S. Note that $\mathbb{P}_j(S)$ can be written as the sum of all rider location weights who choose to pick up a bike at station j under bike station pattern S. We first prove three useful lemmas. The first one concerns the consistency of the MLE of the arrival rate $\hat{\lambda}$ and choice probabilities $\hat{\mathbb{P}}_j(S)$.

LEMMA EC.1. *The MLEs $\hat{\lambda}\to\lambda^*$ and $\hat{\mathbb{P}}_j(S)\to\mathbb{P}_j^*(S)$ with probability one for all $S\subset\mathcal{M}$ and $j\in S\cup\{0\}$ as $T\to\infty$, where λ^* and $\mathbb{P}_j^*(S)$ are the corresponding true values.*

Proof. Since each station pattern is observed with a positive fraction when $T \to \infty$, the station pattern where all bikes are available $S = \mathcal{M}$ is observed with an infinite amount of time. Since $\mathbb{P}_0(\mathcal{M}) = 0$, it is clear that λ can be consistently estimated by simply applying the strong law of large numbers to the renewal process (see, e.g., Proposition 3.3.1 in Ross 1996, note that the MLE of λ is simply the number of total arrivals over the length of the arrival period with $S = \mathcal{M}$). For the same reason, for any other station pattern $S \in \mathcal{S}$, the observed arrival rate under that pattern $\tilde{\lambda}_S$ can be consistently estimated. We know that $\mathbb{P}_0(S) = 1 - \tilde{\lambda}_S/\lambda$, which can also be consistently estimated. Finally, for any $j \in S$, we have $\mathbb{P}_j(S) = \tilde{\lambda}_{S,j}(1 - \mathbb{P}_0(S))/\tilde{\lambda}_S$ where $\tilde{\lambda}_{S,j}$ is the observed arrival rate for riders who choose station j under station pattern S, which can be consistently estimated as well. This completes the proof. □

Lemma EC.1 states that the MLEs of choice probabilities and the arrival rate are strongly consistent. What is left to be shown is thus one can uniquely determine weights of distance rankings from the choice probabilities. We first discuss a few interesting observations regarding how distance rankings vary in a one-dimensional space.

LEMMA EC.2. *Given a subset of bike stations $S \subset \mathcal{M}$ with $|S| = k$, for any permutation of S, $\left(s^{(1)}, \cdots, s^{(k)}\right)$ such that $\Sigma\left(\left(s^{(1)}, \cdots, s^{(k)}\right)\right) \neq \emptyset$, there exists at most two distinct bike stations $s^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)}, \cdots, s^{(k)}, s^{(k+1)}\right)\right) \neq \emptyset$. Moreover, there exists at most one permutation $s^{(1)}, \cdots, s^{(k)}$ where $s^{(k+1)}$ can have two distinct options, the rest of the permutations have at most one such $s^{(k+1)}$.*

Proof. Figure EC.1 is an illustration of the proof. Let x_s denote the coordinate of a station $s \in \mathcal{M}$. It is easy to check the first part of the statement. First of all, for any $s^{(1)} \neq \cdots \neq s^{(k)}$ such that $\Sigma((s^{(1)}, \cdots, s^{(k)})) \neq \emptyset$, they have to form a group such that for all $s \neq s^{(1)}, \cdots, s^{(k)}$, either $x_s < x_{s^{(1)}}, \cdots, x_{s^{(k)}}$ or $x_s > x_{s^{(1)}}, \cdots, x_{s^{(k)}}$. Moreover, $s^{(k+1)}$ has to be the closest one on the left or on the right, i.e., $s^{(k+1)}$ has to be either $s_{\text{right}} = \arg\min_{s \in \mathcal{M}: x_s > x_{s^{(1)}}, \cdots, x_{s^{(k)}}}$ or $s_{\text{left}} = \arg\max_{s \in \mathcal{M}: x_s < x_{s^{(1)}}, \cdots, x_{s^{(k)}}}$.

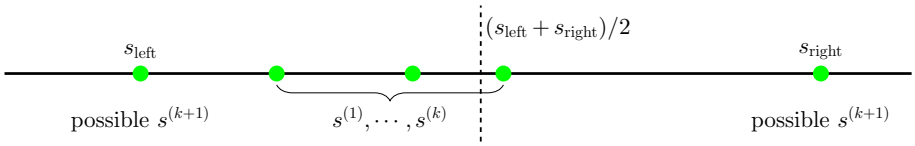


Figure EC.1 Identifiability proof under a one-dimensional space.

To see the second part of the statement, suppose that there exists such two choices of $s^{(k+1)} \in \{s_{\text{left}}, s_{\text{right}}\}$ (if not, the statement is trivial). We draw the middle point of the two potential stations of $s^{(k+1)}$, $(s_{\text{left}} + s_{\text{right}})/2$, depicted by the dashed vertical line. The main idea is to show that there exists at most one permutation $(s^{(1)}, s^{(2)}, \cdots, s^{(k)})$, such that the subregion within which riders' first k nearest stations are $(s^{(1)}, s^{(2)}, \cdots, s^{(k)})$, covers the point $(s_{\text{left}} + s_{\text{right}})/2$.

For ease of presentation, we define $\mathcal{R}(\sigma)$ as the subregion of the rider locations corresponding to distance ranking σ in the space and $\overline{\mathcal{R}}(\sigma) \supseteq \mathcal{R}(\sigma)$ for the subregion whose first $|\sigma|$ closest bike stations is exactly σ. There are two situations to discuss:

(1) The middle point $(s_{\text{left}} + s_{\text{right}})/2$ falls into a subregion corresponding to $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ where $s^{(1)}, \cdots, s^{(k)}$ is a particular permutation of S. Then subregion $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ is separated into two parts. The part on the left of $(s_{\text{left}} + s_{\text{right}})/2$ would possibly correspond to distance rankings in $\Sigma((s^{(1)}, \cdots, s^{(k)}, s_{\text{left}}))$ while the part on the right of $(s_{\text{left}} + s_{\text{right}})/2$ would possibly correspond to rankings in $\Sigma((s^{(1)}, \cdots, s^{(k)}, s_{\text{right}}))$. We say "possibly" here because it could be the case that $\Sigma((s^{(1)}, \cdots, s^{(k)}, s_{\text{left}})) = \emptyset$ or $\Sigma((s^{(1)}, \cdots, s^{(k)}, s_{\text{right}})) = \emptyset$ due to limited consideration radius. For any other permutation $s'^{(1)}, \cdots, s'^{(k)}$, such that $\Sigma((s'^{(1)}, \cdots, s'^{(k)})) \neq \emptyset$, subregions corresponding to any $\sigma \in \Sigma((s'^{(1)}, \cdots, s'^{(k)}))$ will be either on the left side of $(s_{\text{left}} + s_{\text{right}})/2$ or on the right side of $(s_{\text{left}} + s_{\text{right}})/2$. Thus $s^{(k+1)}$ will be either s_{left} or s_{right} depending on which side it locates. It can also be the case that $s^{(k+1)}$ does not exist if it is outside their consideration radiuses.

(2) The middle point $(s_{\text{left}} + s_{\text{right}})/2$ does not fall into subregion $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ for any permutation of S: $s^{(1)}, \cdots, s^{(k)}$. Since the union of subregions $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ over all permutations $s^{(1)}, \cdots, s^{(k)}$ form a convex set. This convex set thus locates either on the left side of $(s_{\text{left}} + s_{\text{right}})/2$ or on the right side of $(s_{\text{left}} + s_{\text{right}})/2$, which means that for all permutations, $s^{(k+1)}$ is either s_{left} or s_{right} depending on which side it locates, or $s^{(k+1)}$ simply does not exist due to limited consideration radius. This completes the proof of Lemma EC.2. □

LEMMA EC.3. *For the consideration radius function $r(\cdot)$ such that $\|r(x) - r(x')\| \leq \|x - x'\|, \forall x, x' \in \mathcal{P}$, given a subset of stations $S \subset \mathcal{M}$ with $|S| = k$, there exists at most two distinctive permutations of S, $s^{(1)}, \cdots, s^{(k)}$ and $s'^{(1)}, \cdots, s'^{(k)}$ such that:*

- $\left(s^{(1)}, \cdots, s^{(k)}\right) \in \Sigma$, *and there exists* $s^{(k+1)}$ *with* $\Sigma\left(\left(s^{(1)}, \cdots, s^{(k+1)}\right)\right) \neq \emptyset$.
- $\left(s'^{(1)}, \cdots, s'^{(k)}\right) \in \Sigma$, *and there exists* $s'^{(k+1)}$ *with* $\Sigma\left(\left(s'^{(1)}, \cdots, s'^{(k+1)}\right)\right) \neq \emptyset$.

If the two permutations co-exist, it must be that $s^{(k+1)} \neq s'^{(k+1)}$.

Proof. We prove this by contradiction. We show the second part of the statement first. Suppose that $s^{(k+1)} = s'^{(k+1)}$. Knowing that $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ and $\overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$ are two disjoint convex sets. Without loss of generality, we assume $s^{(k+1)}$ is on the left of $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$. We now consider two situations:

(1) $\overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$ is on the right side of $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$: Because $(s^{(1)}, \cdots, s^{(k)}) \in \Sigma$, there exists $x_0 \in \overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$ such that $\|x_0 - x_{s^{(k+1)}}\| > r(x_0)$ (since $\mathcal{R}((s^{(1)}, \cdots, s^{(k)}))$ is non-empty). Let $x^* = \sup_{x \in \overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))} x$ denote the rightmost point in the subregion $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$. Given that $\|r(x^*) - r(x_0)\| = r(x^*) - r(x_0) \leq \|x^* - x_0\| = x^* - x_0$, it is clear that $\|x^* - x_{s^{(k+1)}}\| = x^* - x_{s^{(k+1)}} = \|x^* - x_0\| + \|x_0 - x_{s^{(k+1)}}\| > r(x^*)$. For all $x' \in \overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$:

$$\|x' - x_{s^{(k+1)}}\| \geq \|x^* - x_{s^{(k+1)}}\| > r(x^*).$$

since $\overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$ is on the right-hand side of $\overline{\mathcal{R}}((s^{(1)}, \cdots, s^{(k)}))$. In the case of $r(x^*) \geq r(x')$, we immediately get $\|x' - x_{s^{(k+1)}}\| > r(x')$. In the case of $r(x^*) < r(x')$, we know $\|x' - x_{s^{(k+1)}}\| - \|x^* - x_{s^{(k+1)}}\| = \|x' - x^*\| \geq \|r(x') - r(x^*)\| = r(x') - r(x^*)$. Then we have the following:

$$\|x' - x_{s^{(k+1)}}\| - r(x') \geq \|x^* - x_{s^{(k+1)}}\| - r(x^*) > 0.$$

We thus have $\|x' - x_{s^{(k+1)}}\| > r(x')$ for all $x' \in \overline{\mathcal{R}}((s'^{(1)}, \cdots, s'^{(k)}))$, which contradicts to the assumption that there exists $s'^{(k+1)}$ with $\Sigma((s'^{(1)}, \cdots, s'^{(k)}, s'^{(k+1)})) \neq \emptyset$.

(2) $\overline{\mathcal{R}}((s'^{(1)},\cdots,s'^{(k)}))$ is on the left-hand side of $\overline{\mathcal{R}}((s^{(1)},\cdots,s^{(k)}))$: This case can be proved by following the exactly same procedure from the first case by symmetry. We only need to swap all notations of $s^{(i)}$ with $s'^{(i)}$ for $i\in\{1,\cdots,k+1\}$.

The first part of Lemma EC.3 can be shown with similar arguments. Suppose that there exists a third different permutation $s''^{(1)},\cdots,s''^{(k)}$ such that $(s''^{(1)},\cdots,s''^{(k)})\in\Sigma$ and there exists $s''^{(k+1)}$ with $\Sigma((s''^{(1)},\cdots,s''^{(k)},s''^{(k+1)}))\neq\emptyset$. From Lemma EC.2 we know that $s''^{(k+1)}\in\{s^{(k+1)},s'^{(k+1)}\}$. However, we just proved that there does not exist two different permutations with the same $(k+1)^{\text{th}}$ station. This reaches a contradiction and completes the proof. □

We are now ready to prove Theorem 2. We show that there is a unique solution of distance ranking weights from choice probabilities, which can be consistently estimated from Lemma EC.1. In specific, we want to prove for any $1\le k\le M$, and for any $s^{(1)}\neq\cdots\neq s^{(k)}\in\mathcal{M}$, $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)}))}w_\sigma$ and $w_{(s^{(1)},\ldots,s^{(k-1)})}$ are uniquely identifiable by inducting on k. Intuitively, this induction allows us to gradually increase the resolution of the distance rankings we are able to identify. We first prove the initial condition $k=1$. Notice that

$$\sum_{\sigma\in\Sigma((i))}w_\sigma=\mathbb{P}_i(\mathcal{M}), \tag{EC.11}$$

$$\sum_{\sigma\in\Sigma((i,j))}w_\sigma=\sum_{\sigma\in\Sigma((j))}w_\sigma+\sum_{\sigma\in\Sigma((i,j))}w_\sigma-\sum_{\sigma\in\Sigma((j))}w_\sigma=\mathbb{P}_j(\mathcal{M}\backslash\{i\})-\mathbb{P}_j(\mathcal{M}), \tag{EC.12}$$

$$\Longrightarrow\qquad w_{(i)}=\sum_{\sigma\in\Sigma((i))}w_\sigma-\sum_{\sigma\in\Sigma((i,j)),j\in\mathcal{M}}w_\sigma=\mathbb{P}_i(\mathcal{M})-\sum_{j\in\mathcal{M}}\left(\mathbb{P}_j(\mathcal{M}\backslash\{i\})-\mathbb{P}_j(\mathcal{M})\right), \tag{EC.13}$$

for all $i\neq j$ and $i,j\in\{1,\cdots,M\}$. For the induction step at k, suppose that for any positive integer $n\le k$, we have $s^{(1)}\neq\cdots\neq s^{(n)}\in\mathcal{M}$ and $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(n)}))}w_\sigma$ and $w_{(s^{(1)},\cdots,s^{(n-1)})}$ are uniquely identifiable. Then, we want to prove for any $s^{(1)}\neq\cdots\neq s^{(k+1)}\in\mathcal{M}$, $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k+1)}))}w_\sigma$ and $w_{(s^{(1)},\cdots,s^{(k)})}$ are uniquely identifiable. For a given set of k stations $\{s^{(1)},\cdots,s^{(k)}\}$, each permutation of it can be classified into one of the following types:

1. $\left(s^{(1)},\cdots,s^{(k)}\right)\in\Sigma$ and there does not exist $s^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)},\cdots,s^{(k+1)}\right)\right)\neq\emptyset$. Then it is clear that

$$\left(s^{(1)},\cdots,s^{(k)}\right)=\Sigma\left(\left(s^{(1)},\cdots,s^{(k)}\right)\right)$$

Since $\sum_{\sigma\in\Sigma((s^{(1)},\ldots,s^{(k)}))}w_\sigma$ is known according to the induction hypothesis, we thus are able to identify $w_{(s^{(1)},\cdots,s^{(k)})}$ and $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))}w_\sigma$.

2. $\left(s^{(1)},\cdots,s^{(k)}\right)\notin\Sigma$ and there exists only one $s^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)},\cdots,s^{(k+1)}\right)\right)\neq\emptyset$. Then we have

$$\Sigma\left(\left(s^{(1)},\cdots,s^{(k)},s^{(k+1)}\right)\right)=\Sigma\left(\left(s^{(1)},\cdots,s^{(k)}\right)\right)$$

Similarly, we have $w_{(s^{(1)},\cdots,s^{(k)})}=0$ and $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))}w_\sigma=\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)}))}w_\sigma$, which is known according to the induction hypothesis.

3. There exists $s^{(k+1)}$ and $s'^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)},\cdots,s^{(k)},s^{(k+1)}\right)\right)\neq\emptyset$ and $\Sigma\left(\left(s^{(1)},\cdots,s^{(k)},s'^{(k+1)}\right)\right)\neq\emptyset$. With a slight abuse of notation, we define $\Sigma(S,s)$ for some $S\subseteq\mathcal{M}$ and $s\in\mathcal{M}\backslash S$ as the set of distance rankings whose first $|S|$ stations are *any* permutation of S and the $(|S|+1)^{\text{th}}$ station is s. Put set $S^k=\{s^{(1)},\cdots,s^{(k)}\}$. Then we have the following identity:

$$\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))} w_\sigma = \mathbb{P}_{s^{(k+1)}}\left(\mathcal{M}\backslash S^k\right) - \sum_{S\subsetneq S^k}\sum_{\sigma\in\Sigma((S,s^{(k+1)}))} w_\sigma - \sum_{\sigma\in\Sigma((S^k,s^{(k+1)}))\backslash\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))} w_\sigma. \tag{EC.14}$$

The second term on the right-hand side is known according to the induction hypothesis. For the third term, by lemma EC.2, since the permutation $(s^{(1)},\cdots,s^{(k)})$ has two distinct options $s^{(k+1)}$ and $s'^{(k+1)}$, any rest of the permutation of S^k, $s''^{(1)},\cdots,s''^{(k)}$, has at most one $s''^{(k+1)}$ such that $\Sigma\left(\left(s''^{(1)},\cdots,s''^{(k+1)}\right)\right)\neq 0$. Clearly, we only need to consider the case when such $s''^{(k+1)}$ exists. Now consider two subcases: 1) if $(s^{(1)},\cdots,s^{(k)})\notin\Sigma$, we claim that $(s''^{(1)},\cdots,s''^{(k)})\notin\Sigma$. We prove by contradiction. Without loss of generality, suppose that $s^{(k+1)}$ and $((s''^{(1)},\cdots,s''^{(k)}))$ are on the left-hand side of $\overline{\mathcal{R}}((s^{(1)},\cdots,s^{(k)}))$. Then if $(s''^{(1)},\cdots,s''^{(k)})\in\Sigma$ it is clear that for all $x\in\overline{\mathcal{R}}((s^{(1)},\cdots,s^{(k)}))$, we have $|x-x_{s^{(k+1)}}|>r(x)$ which contradicts to the fact that $(s^{(1)},\cdots,s^{(k)},s^{(k+1)})\in\Sigma$. This suggests that the permutation $s''^{(1)},\cdots,s''^{(k)}$ belongs to the second type we discussed above and is proven to be identifiable; if $(s^{(1)},\cdots,s^{(k)})\in\Sigma$, then by Lemma EC.3, we know $s'^{(k+1)}\neq s^{(k+1)}$, which implies that the permutation σ in the third term cannot be $(s'^{(1)},\cdots,s'^{(k)},s'^{(k+1)})$. Hence, by equation (EC.14), we know that $\sum_{\sigma\in\Sigma\left(\left(s^{(1)},\cdots,s^{(k)},s^{(k+1)}\right)\right)} w_\sigma$ can be uniquely identified. By symmetry, we can show that $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s'^{(k+1)}))} w_\sigma$ is identifiable as well. We can further deduce $w_{(s^{(1)},\cdots,s^{(k)})}$ by

$$w_{(s^{(1)},\cdots,s^{(k)})} = \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)}))} w_\sigma - \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))} w_\sigma - \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s'^{(k+1)}))} w_\sigma.$$

4. The permutation $(s^{(1)},\cdots,s^{(k)})\in\Sigma$ and there exists only one $s^{(k+1)}$ such that $\Sigma\left(\left(s^{(1)},\cdots,s^{(k+1)}\right)\right)\neq\emptyset$. Similarly, we can show that $\sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k+1)}))} w_\sigma$ is identifiable using the same approach in type 3. We can further deduce $w_{(s^{(1)},\cdots,s^{(k)})}$ by

$$w_{(s^{(1)},\cdots,s^{(k)})} = \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)}))} w_\sigma - \sum_{\sigma\in\Sigma((s^{(1)},\cdots,s^{(k)},s^{(k+1)}))} w_\sigma.$$

This completes the induction and proof of Theorem 2. We comment that the proof does not necessarily require all possible station patterns to be observed with a positive fraction of time in the long run. We only need $\mathcal{M}$ and $\mathcal{M}\setminus S^k$ to be observed for all possible S^k that are constructed by looking at all distance rankings in the set of rider locations $\mathcal{L}$. □

Proof of Proposition 2 The result can be directly shown by equations (EC.11), (EC.12) and (EC.13) in the proof of Theorem 2. □

Proof of Theorem 3 Since the closed-form solution $\hat{\boldsymbol{w}}$ derived in the M-step is a unique maximizer, we know that in the EM algorithm, the M-step generates a sequence of vectors $\{\boldsymbol{w}^{(m)}, m=1,2,\cdots\}$ where $\boldsymbol{w}^{(m)}$ is the unique maximizer for the expected complete log-likelihood function $\mathbb{E}[l_C(\boldsymbol{w})\mid\boldsymbol{b},\boldsymbol{t},\boldsymbol{w}^{(m-1)}]$ for

$m = 1, 2, \cdots$. Here, $\boldsymbol{w}^{(0)}$ refers to the initial weight vector to start the EM algorithm. Moreover, from equation (6), it is clear that the expected complete log-likelihood function $\mathbb{E}[l_C(\boldsymbol{w}) \mid \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}]$ is continuous in both $\boldsymbol{w}$ and $\boldsymbol{w}^{(m)}$. Then by Theorem 2 in Wu (1983), we can infer that the sequence $\{l(\boldsymbol{w}^{(1)}), l(\boldsymbol{w}^{(2)}), \cdots\}$ converges monotonically to $l(\boldsymbol{w}^*)$ for some stationary point $\boldsymbol{w}^*$. □

Appendix B: Additional Results

B.1. Estimating Location Weights

In Example 1 we give instances regarding the identifiability of location weights under the MNL and distance-ranking models. Here, we provide another example that the location weights under the distance-ranking model are identifiable while they are not under the MNL model. Again, we use the Euclidean distance as our distance metric. Consider the bike and rider locations depicted in Figure EC.2. Denote the bike pattern by S. We let $\beta_{1,1} = \beta_{1,2} = -\ln(2)/(2\sqrt{3} - 2)$ and $\beta_{1,3} = (\ln(3) - \ln(4))/(\sqrt{7} - \sqrt{3})$ and consider an asymptotic scenario where $\beta_{0,1}, \beta_{0,2}, \beta_{0,3} \to \infty$. Then for the MNL model, we have $[p_{1,1,S}, p_{2,1,S}, p_{3,1,S}] = [0.4, 0.2, 0.3]$, $[p_{1,2,S}, p_{2,2,S}, p_{3,2,S}] = [0.2, 0.4, 0.3]$ and $[p_{1,3,S}, p_{2,3,S}, p_{3,3,S}] = [0.4, 0.4, 0.4]$, which leads to non-identifiability according to Theorem 1 since these vectors are linearly dependent. On the other hand, consider a distance-ranking model with infinite consideration radius. We have choice probabilities as $[p_{1,1,S}, p_{2,1,S}, p_{3,1,S}] = [0.5, 0, 0]$, $[p_{1,2,S}, p_{2,2,S}, p_{3,2,S}] = [0.0, 0.5, 0]$ and $[p_{1,3,S}, p_{2,3,S}, p_{3,3,S}] = [0.5, 0.5, 1]$ that are linearly independent. This implies that the model is identifiable.

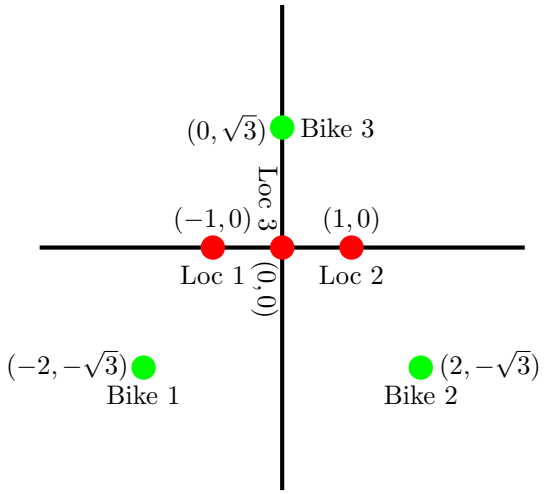


Figure EC.2 **Identifiable under a distance-ranking model but non-identifiable under the MNL model for certain parameter values.**

Theorem 1 describes the identifiability of all location weights. Here we give another result regarding partial identifiability of the weight of a particular location $l \in \mathcal{L}$. Let $\boldsymbol{e}_l \in \{0, 1\}^L$ be a binary vector whose l^{th} entry is one and zero otherwise and $\mathbb{1}_L$ be a L-dimensional vector with all ones.

PROPOSITION EC.1 **(Partial Identifiability)**. *For any $l \in \mathcal{L}$, if both $\boldsymbol{e}_l$ and $\mathbb{1}_L$ are a linear combination of vectors $\{[p_{1,b,S_k}, \cdots, p_{L,b,S_k}] : b \in \mathcal{B}_k,\ k \in \{1, \cdots, K\}\}$, then w_l is identifiable.*

Proof. For any $l \in \mathcal{L}$, by equation (EC.4), it is sufficient to show that the system of linear equations $\sum_{l\in\mathcal{L}} w_l p_{l,b,S_k}/\bar{s} = \sum_{l\in\mathcal{L}} w_l^* p_{l,b,S_k}/\bar{s}^*$, $\forall k \in \{1,\cdots,K\}$ and $b \in \mathcal{B}_k$ has a unique solution $w_l = w_l^*$. Since $\boldsymbol{e}_l$ is a linear combination of vectors $\{[p_{1,b,S_k},\cdots,p_{N,b,S_k}] : b \in \mathcal{B}_k,\ k \in \{1,\cdots,K\}\}$, we know $w_l/\bar{s} = w_l^*/\bar{s}^*$. If $w_l \neq w_l^*$, we have $\bar{s} \neq \bar{s}^*$. However, since $\mathbb{1}_L$ can also be written as a linear combination of these vectors, we have $\sum_{l\in\mathcal{L}} w_i/\bar{s} = \sum_{l\in\mathcal{L}} w_l^*/\bar{s}^*$. Since $\sum_{l\in\mathcal{L}} w_l = \sum_{l\in\mathcal{L}} w_l^* = 1$, we have $\bar{s} = \bar{s}^*$. This leads to $w_l = w_l^*$. □

One way to interpret Proposition EC.1 and its proof is to think of all locations other than location l as a whole. Then the problem can be simplified into a scenario only containing two arrival locations with weights w_l and $1 - w_l$, thereby leading to partial identifiability by invoking Theorem 1. Proposition EC.1 is particularly useful under a distance-ranking model as the choice probabilities are mostly 0 or 1. We now give an example below to illustrate Proposition EC.1.

Example EC.1. Consider a one-dimensional case (Figure EC.3) with three rider arrival locations and one bike. The rider arrival locations are located at -1, 4, and 5 on the axis from left to right. Suppose that there are two possible bike patterns S_1, S_2 where the bike locates at 0 under S_1 and 3 under S_2. Given a constant consideration radius $r = 3$, we have $[p_{1,1,S_1}, p_{2,1,S_1}, p_{3,1,S_1}] = [1,0,0]$ and $[p_{1,2,S_1}, p_{2,2,S_1}, p_{3,2,S_1}] = [0,1,1]$. Then by Theorem 1, the location weights $\boldsymbol{w}$ are non-identifiable since these vectors cannot span $\mathbb{R}^3$. However, by Proposition EC.1, we know that $\boldsymbol{e}_1$ and $\mathbb{1}_3$ are a linear combination of $(1,0,0)$ and $(0,1,1)$, which implies that the w_1 is identifiable.

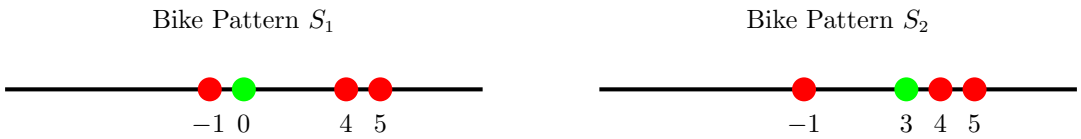


Figure EC.3 **An example of partial identifiability in the distance-ranking model.** As before, red circles represent rider locations and green circles represent bike location. The numbers underneath are coordinates.

B.2. Estimating MNL Model Parameters

In this section, we assume that the rider choices are governed by a multinomial logit (MNL) model stated in equation (1) and consider jointly estimate $\boldsymbol{w}$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$. By following the proof of Theorem 1, $\boldsymbol{w}$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ are identifiable if and only if the system of nonlinear equations $\sum_{l\in\mathcal{L}} w_l p_{l,b,S_k}/\bar{s} = \sum_{l\in\mathcal{L}} w_l^* p_{l,b,S_k}^*/\bar{s}^*$, $b \in \mathcal{B}_k$, $k \in \{1,2,\cdots,K\}$ has a unique solution $\boldsymbol{w} = \boldsymbol{w}^*$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^*$. It is expected that this condition a has higher chance to be satisfied as there are more bike patterns. This condition is hard to simplify in general. Nevertheless, we provide below a set of simplified sufficient conditions for identifiability of $\boldsymbol{\beta}_1$ when there is only one rider location and $\boldsymbol{\beta}_0$ is assumed to be known.

Proposition EC.2. *Assume that riders only arrive at one given location and riders' choice behavior follows an MNL model with known β_0. β_1 is identifiable if at least one of the following conditions holds:*

1. *At least two bikes have different distances to the rider location in some bike pattern;*
2. *$\beta_1^* < 0$ and one bike has different distances to the rider location in at least two bike patterns where each pattern has the same number of available bikes.*

Proof. Let $p_{l,b,S_k}, b \in \mathcal{B}_k, k \in \{1, \cdots, K\}$ be the probability that a rider in the only rider location l chooses bike b in bike pattern S_k and p^*_{l,b,S_k} as the corresponding probability under the true value β_1^*. By the proof of Theorem 1, we want to show the equality $p_{l,b,S_k}/\bar{s} = p^*_{l,b,S_k}/\bar{s}^*$ holds for all $b \in \mathcal{B}_k$, $k \in \{1, \cdots, K\}$ only if $\beta_1 = \beta_1^*$. To show the necessity, we consider the first case where two distinct bikes $b \neq b' \in S_k$ have different distances $d_{l,b,S_k} \neq d_{l,b',S_k}$ to the rider location l in some bike pattern S_k. Without loss of generality, $d_{l,b,S_k} > d_{l,b,S_{k'}}$. This gives,

$$\frac{p_{l,b,S_k}}{p^*_{l,b,S_k}} = \frac{p_{l,b',S_k}}{p^*_{l,b',S_k}} = \frac{\bar{s}}{\bar{s}^*} \quad \Rightarrow \quad \beta_1(d_{l,b,S_k} - d_{l,b',S_k}) = \beta_1^*(d_{l,b,S_k} - d_{l,b',S_k}) \quad \Rightarrow \quad \beta_1 = \beta_1^*.$$

We then consider the second case where one bike b in two different bike patterns $\mathcal{B}_k$ and $\mathcal{B}_{k'}$ have different distances $d_{l,b,S_k} \neq d_{l,b,S_{k'}}$ to the rider location l. If there are more than one bike having different distances to the rider location in either pattern, then it returns to the first case and we have $\hat{\beta}_1 \to \beta_1^*$. Thus, we only need to show the case when all bikes have the same distance to the rider location in each bike pattern. Let B be the number of available bikes in bike patterns $\mathcal{B}_k$ and $\mathcal{B}_{k'}$. Since each pattern has the same number of available bikes, we have

$$\frac{p_{l,b,S_k}}{p^*_{l,b,S_k}} = \frac{p_{l,b,S_{k'}}}{p^*_{l,b,S_{k'}}} \quad \Rightarrow$$
$$\frac{\exp(\beta_0 + \beta_1 d_{l,b,S_k})}{B^{-1} + \exp(\beta_0 + b_k \beta_1 d_{l,b,S_k})} \frac{B^{-1} + \exp(\beta_0 + \beta_1 d_{l,b,S_{k'}})}{\exp(\beta_0 + \beta_1 d_{l,b,S_{k'}})} = \frac{\exp(\beta_0 + \beta_1^* d_{l,b,S_k})}{B^{-1} + \exp(\beta_0 + \beta_1^* d_{l,b,S_k})} \frac{B^{-1} + \exp(\beta_0 + \beta_1^* d_{l,b,S_{k'}})}{\exp(\beta_0 + \beta_1^* d_{l,b,S_{k'}})}.$$

The derivative of the left-hand side of the above equation with respect to β_1 can be written as

$$c_0 \exp(\beta_1(d_{l,b,S_k} + d_{l,b,S_{k'}})) \frac{c_0(d_{l,b,S_k} - d_{l,b,S_{k'}}) + d_{l,b,S_k} \exp(\beta_1 d_{l,b,S_{k'}}) - d_{l,b,S_{k'}} \exp(\beta_1 d_{l,b,S_k})}{\left((c_0 + \exp(\beta_1 d_{l,b,S_k})) \exp(\beta_1 d_{l,b,S_{k'}})\right)^2},$$

where $c_0 := B^{-1} \exp(-\beta_0)$. Since we have $\beta_1 < 0$, it is not hard to see that the derivative would be strictly positive when $d_{l,b,S_k} > d_{l,b,S_{k'}}$. Therefore, the left-hand side is strictly increasing with β_1, which implies the uniqueness of β_1 that satisfies the above equation. This gives the identifiability of β_1. □

We now discuss how the EM algorithm described in Section 4.1 can be adapted to the case where $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and $\boldsymbol{w}$ can be jointly estimated. Our first result below shows that the M-step has the same updating form for $\boldsymbol{w}$ as that in Algorithm 1 and optimizing $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ in the M-step is a well behaved concave maximization problem.

Proposition EC.3. *Assume that rider choice behaviors are governed by an MNL model with parameters $\boldsymbol{\beta} := (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$. Then in the M-step of the EM algorithm, $\boldsymbol{w}$ has a closed-form update and the expected likelihood in the E-step is concave with respect to $\boldsymbol{\beta}$.*

Proof. Analogous to equation (6), in the m^{th} iteration, the conditional expectation can be written as

$$\mathbb{E}\left[l_C(\boldsymbol{w}, \boldsymbol{\beta}_1) \;\middle|\; \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}, \boldsymbol{\beta}^{(m)}\right] = \mathbb{E}\left[-\tilde{N} + \tilde{N}\log\left(\frac{\tilde{N}}{T}\right) + \sum_{n=1}^{\tilde{N}} \log(w_{\tilde{l}_n}) + \sum_{n=1}^{\tilde{N}} \log\left(p_{\tilde{l}_n, \tilde{b}_n, S_{\tilde{t}_n}}\right) \;\middle|\; \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{w}^{(m)}, \boldsymbol{\beta}^{(m)}\right].$$

As in equation (6), the first two terms inside the expectation do not depend on $\boldsymbol{w}$ or $\boldsymbol{\beta}$. The third term depends on $\boldsymbol{w}$ but not $\boldsymbol{\beta}$, and the fourth term depends on $\boldsymbol{\beta}$ but not $\boldsymbol{w}$. Hence, $\boldsymbol{w}$ has the same updating process as before with the only difference lies in using $\boldsymbol{\beta}^{(m)}$ to replace the true value. To optimize for $\boldsymbol{\beta}$, we

only need to consider the fourth term. To simplify the notation, let $\boldsymbol{\Theta}^{(m)} = (\boldsymbol{w}^{(m)}, \boldsymbol{\beta}^{(m)})$ and $\boldsymbol{X} = (\boldsymbol{b}, \boldsymbol{t})$. We first derive a few important quantities. The joint conditional density for a rider arriving at location l at time t given that she chooses the leaving option is

$$\begin{aligned} f\left(l,t \;\middle|\; \boldsymbol{\Theta}^{(m)}, b=0\right) &= \mathbb{P}\left(l \;\middle|\; t, \boldsymbol{\beta}^{(m)}, \boldsymbol{w}^{(m)}, b=0\right) \cdot f\left(t \;\middle|\; \boldsymbol{\beta}^{(m)}, \boldsymbol{w}^{(m)}, b=0\right) \\ &= \frac{p^{(m)}_{l,0,S_t} w^{(m)}_l}{\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t}} \cdot \frac{\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t}}{\int_0^T \sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t} dt} \\ &= \frac{p^{(m)}_{l,0,S_t} w^{(m)}_l}{\int_0^T \sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t} dt}. \end{aligned}$$

where $p^{(m)}_{l,b,S_t}$, $l\in\mathcal{L}$, $b\in\mathcal{B}$, $t\in[0,T]$ is the rider's choice probability under $\beta^{(m)}_{0,l}$ and $\beta^{(m)}_{1,l}$. Similar to equations (7) and (8), we have

$$\mathbb{P}\left(l_n = l \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right) = \frac{p^{(m)}_{l,b_n,S_{t_n}} w^{(m)}_l}{\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',b_n,S_{t_n}}},\ \forall n\in\{1,\cdots,N\},\ \forall l\in\mathcal{L}, \tag{EC.15}$$

$$\mathbb{E}\left[N' \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right] = N\frac{\int_0^T \sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t} dt}{\int_0^T \left(1-\sum_{l'\in\mathcal{L}} w^{(m)}_{l'} p^{(m)}_{l',0,S_t}\right) dt}. \tag{EC.16}$$

Let the sequence $\{t'_1, t'_2, \cdots, t'_{N'}\}$ denote the arrival time of the unobserved riders. Then we can rewrite the third term in the conditional expectation as

$$\begin{aligned} &\mathbb{E}\left[\sum_{n=1}^{\tilde{N}} \log(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}) \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right] \\ =&\mathbb{E}\left[\sum_{n=1}^{N} \log(p_{l_n,b_n,S_{t_n}}) \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right] + \mathbb{E}\left[\sum_{n=1}^{N'} \log(p_{l'_n,0,S_{t'_n}}) \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right] \\ =&\sum_{l\in\mathcal{L}}\sum_{n=1}^{N} \left(\mathbb{P}\left(l_n = l \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right) \log p_{l,b_n,S_{t_n}}\right) + \mathbb{E}\left[N' \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right] \int_0^T \sum_{l\in\mathcal{L}} f\left(l,t \;\middle|\; \boldsymbol{\Theta}^{(m)}, b=0\right) \log p_{l,0,S_t} dt \\ =&\sum_{l\in\mathcal{L}}\sum_{n=1}^{N} \left(\mathbb{P}\left(l_n = l \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right) \log p_{l,b_n,S_{t_n}}\right) + \mathbb{E}\left[N' \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right] \sum_{l\in\mathcal{L}} \int_0^T f\left(l,t \;\middle|\; \boldsymbol{\Theta}^{(m)}, b=0\right) \log p_{l,0,S_t} dt \\ =&\sum_{l\in\mathcal{L}} \left(\sum_{n=1}^{N} \mathbb{P}\left(l_n = l \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right) \log p_{l,b_n,S_{t_n}} + \mathbb{E}\left[N' \;\middle|\; \boldsymbol{\Theta}^{(m)}, \boldsymbol{X}\right] \int_0^T f\left(l,t \;\middle|\; \boldsymbol{\Theta}^{(m)}, b=0\right) \log p_{l,0,S_t} dt\right). \end{aligned}$$

The above equation shows that $\beta_{1,l}$ is separable for $l\in\mathcal{L}$. Thus, to show the concavity, it suffices to show the formula inside the parenthesis is concave regarding $(\beta_{0,l}, \beta_{1,l})$. For each rider location $l\in\mathcal{L}$ and available bike $b\in\mathcal{B}_t$, recall that the choice probability is

$$p_{l,b,S_t} = \frac{\exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}{1+\sum_{b\in\mathcal{B}_t} \exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}, \quad p_{l,0,S_t} = \frac{1}{1+\sum_{b\in\mathcal{B}_t} \exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})}.$$

Therefore, we have

$$\log p_{l,b_n,S_{t_n}} = \beta_{0,l} + \beta_{1,l} d_{l,b_n,S_t} - \log\left(1 + \sum_{b\in\mathcal{B}_{t_n}} \exp(\beta_{0,l} + \beta_{1,l} d_{l,b_n,S_{t_n}})\right),$$

$$\log p_{l,0,S_t} = -\log\left(1 + \sum_{b\in\mathcal{B}_t} \exp(\beta_{0,l} + \beta_{1,l} d_{l,b,S_t})\right).$$

Notice that for any $t \in [0,T]$, the function $\log\left(1+\sum_{b\in\mathcal{B}_t}\exp(\beta_{0,l}+\beta_{1,l}d_{l,b,S_t})\right)$ is a composition of log-sum of exponentials with an affine function, which is convex. This establishes that $\mathbb{E}\left[\sum_{n=1}^{\tilde{N}}\log(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}) \mid \boldsymbol{\Theta}^{(m)},\boldsymbol{X}\right]$ is concave in $\boldsymbol{\beta}$, which completes the proof. □

We now summarize the EM algorithm to estimate $\boldsymbol{w}$ and $\boldsymbol{\beta}$ jointly. Substituting equation (EC.15) and (EC.16) into the third term of the conditional expectation yields

$$\mathbb{E}\left[\sum_{n=1}^{\tilde{N}}\log(p_{\tilde{l}_n,\tilde{b}_n,S_{\tilde{t}_n}}) \;\middle|\; \boldsymbol{\Theta}^{(m)},\boldsymbol{X}\right] = \sum_{l\in\mathcal{L}}\left(\sum_{n=1}^{N}\frac{p^{(m)}_{l,b_n,S_{t_n}}w^{(m)}_l}{\sum_{l'\in\mathcal{L}}w^{(m)}_{l'}p^{(m)}_{l',b_n,S_{t_n}}}\log p_{l,b_n,S_{t_n}} + \frac{N}{s}\int_0^T p^{(m)}_{l,0,S_t}w^{(m)}_l\log p_{l,0,S_t}dt\right),$$

where $s=\int_0^T(1-\sum_{l\in\mathcal{L}}w^{(m)}_l p^{(m)}_{l,0,S_t})dt$. In the M-step, one needs to maximize the formula inside the parenthesis of the above equation for each rider location $l\in\mathcal{L}$. The complete algorithm is presented in Algorithm 3

Algorithm 3 EM Algorithm estimating MNL model parameters

Initialize a location set $\mathcal{L}_0$ with coordinates $(x_1,y_1),\cdots,(x_{L_0},y_{L_0})$.

Initialize the weight vector $\boldsymbol{w}\leftarrow\boldsymbol{w}^{(0)}\in\Delta^L$, $\boldsymbol{\beta}_0\leftarrow\boldsymbol{\beta}_0^{(0)}$, $\boldsymbol{\beta}_1\leftarrow\boldsymbol{\beta}_1^{(0)}$.

Initialize the number of iterations $m\leftarrow 0$.

while stopping criteria are not met **do**

Compute $s\leftarrow\int_0^T(1-\sum_{l\in\mathcal{L}}w^{(m)}_l p^{(m)}_{l,0,S_t})dt$.

For $l\in\mathcal{L}$, find

$$\beta^{(m+1)}_{0,l},\beta^{(m+1)}_{1,l}\in\underset{\beta_{0,l},\beta_{1,l}}{\arg\max}\left(\sum_{n=1}^{N}\frac{p^{(m)}_{l,b_n,S_{t_n}}w^{(m)}_l}{\sum_{l'\in\mathcal{L}}w^{(m)}_{l'}p^{(m)}_{l',b_n,S_{t_n}}}\log p_{l,b_n,S_{t_n}} + \frac{N}{s}\int_0^T p^{(m)}_{l,0,S_t}w^{(m)}_l\log p_{l,0,S_t}dt\right).$$

Compute $c_l\leftarrow\sum_{n=1}^{N}\frac{p^{(m)}_{l,b_n,S_{t_n}}w^{(m)}_l}{\sum_{l'\in\mathcal{L}}w^{(m)}_{l'}p^{(m)}_{l',b_n,S_{t_n}}}+N(T-s)/s$, for $l\in\mathcal{L}$.

Update $w^{(m+1)}_l\leftarrow c_l/\sum_{l'\in\mathcal{L}}c_{l'}$, for $l\in\mathcal{L}$.

$m\leftarrow m+1$.

end while

Output $\boldsymbol{w}^{(m)}$, $\boldsymbol{\beta}_{\mathbf{0}}^{(m)}$ and $\boldsymbol{\beta}_{\mathbf{1}}^{(m)}$.

below. A similar location-discovery procedure can be developed to complement Algorithm 3.

We generate synthetic data to evaluate this algorithm. For simplicity, we assume that $\boldsymbol{\beta}_0$ is known to the operator and $\boldsymbol{\beta}_1$ is to be estimated. We further assume that $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ have the same value across all locations. We thus treat them as scalars (β_0,β_1). We use the same simulation setup described in Section 5.1.2. We test five pairs of ground-truth (β_0^*,β_1^*), which are $(1,-1)$, $(3,-1)$, $(5,-1)$, $(5,-2)$ and $(5,-3)$. When searching the value of β_1, we employ a golden-section search and presume a possible range for β_1 to be $(-10,0)$. For the location-discovery algorithm, we stop the algorithm when *all* of the following three conditions are satisfied: (1) a minimum number of locations $N=5$ has been discovered; (2) the BIC value is larger than that in the previous iteration; and (3) the relative percentage difference of β_1 between two

consecutive iterations is less than 5%. We generate 20 instances for each scenario and output the average performance among all instances. Again, we trim the locations with weights under 0.01 before we output the result. The performance is reported in Table EC.1. In this table, "MAPE" column reports the mean absolute percentage error between the predicted value of β_1 and its underlying truth, and "Time" column records the CPU times in seconds.

Table EC.1 Prediction performance when β_1 is unknown.

(β_0^*, β_1^*)	(L^*, B)	Algorithm	Locs	WD	Lkd	BIC	Time (sec)	MAPE (β_1)
(1,-1)	(5,20)	K-means	5.0	2.94	-2,213	2,228	<1	–
	(10,40)		10.0	2.40	-4,718	4,750	<1	–
(3,-1)	(5,20)		5.0	3.09	-4,197	4,213	<1	–
	(10,40)		10.0	2.25	-6,606	6,640	<1	–
(5,-1)	(5,20)		5.0	3.08	-5,529	5,546	<1	–
	(10,40)		10.0	2.43	-7,295	7,329	<1	–
(5,-2)	(5,20)		5.0	2.86	-1,622	1,635	<1	–
	(10,40)		10.0	2.36	-4,588	4,620	<1	–
(5,-3)	(5,20)		5.0	2.90	-492	503	<1	–
	(10,40)		10.0	2.14	-2,137	2,165	<1	–
(1,-1)	(5,20)	All-in	36.0	3.22	-2,211	2,314	48	138%
	(10,40)		36.4	2.84	-4,733	4,849	210	135%
(3,-1)	(5,20)		35.5	3.20	-4,333	4,447	122	153%
	(10,40)		31.3	2.72	-6,540	6,646	547	59%
(5,-1)	(5,20)		27.6	3.25	-5,578	5,671	294	102%
	(10,40)		29.6	2.75	-7,164	7,266	632	51%
(5,-2)	(5,20)		34.2	3.38	-1,571	1,662	42	126%
	(10,40)		28.8	2.54	-4,413	4,505	404	32%
(5,-3)	(5,20)		34.5	3.25	-439	510	8	118%
	(10,40)		30.9	2.55	-2,063	2,149	80	67%
(1,-1)	(5,20)	Loc. Disc.	7.9	3.02	-2,104	2,126	18	15%
	(10,40)	(Single)	9.1	2.40	-4,618	4,647	71	16%
(3,-1)	(5,20)		6.9	2.78	-4,030	4,052	24	7%
	(10,40)		8.7	2.52	-6,481	6,510	78	10%
(5,-1)	(5,20)		6.7	2.81	-5,333	5,355	38	5%
	(10,40)		8.6	2.52	-7,157	7,187	78	6%
(5,-2)	(5,20)		6.0	3.41	-1,450	1,466	8	10%
	(10,40)		8.8	2.57	-4,386	4,414	47	8%
(5,-3)	(5,20)		5.9	3.15	-424	437	3	17%
	(10,40)		8.8	2.62	-1,978	2,003	21	12%
(1,-1)	(5,20)	Loc. Disc.	11.4	3.01	-2,104	2,137	22	18%
	(10,40)	(Batch)	14.8	2.30	-4,614	4,661	98	15%
(3,-1)	(5,20)		8.7	3.02	-4,038	4,066	28	10%
	(10,40)		13.6	2.19	-6,467	6,513	114	9%
(5,-1)	(5,20)		8.8	3.05	-5,336	5,366	42	6%
	(10,40)		16.0	1.97	-7,160	7,215	142	17%
(5,-2)	(5,20)		8.8	2.86	-1,449	1,473	11	10%
	(10,40)		15.0	2.11	-4,380	4,429	91	14%
(5,-3)	(5,20)		10.0	2.65	-414	435	4	23%
	(10,40)		17.4	1.85	-1,934	1,983	49	9%

We have the following observations from Table EC.1. These observations are similar as those in Table 3. (1) The all-in algorithm gives a poor estimation on β_1 whereas the discovery algorithms have relatively accurate estimation, as indicated by the MAPEs. (2) The computation time for the all-in algorithm significantly increases due to the line search for β_1 in the M-step. Meanwhile, the discovery algorithm shows a less significant increase in computation time, as the embedded EM updates account for only a small portion of the total time. (3) The location-discovery algorithm with batch addition has overall the best performance. It improves over the single mode as evidenced by the lower Wasserstein distance and similar MAPE of β_1, though the batch mode discovers more locations than the single mode.

References

Kullback S, Leibler RA (1951) On information and sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Ross S (1996) *Stochastic processes*. Wiley series in probability and statistics: Probability and statistics (Wiley), ISBN 9780471120629, URL `https://books.google.de/books?id=ImUPAQAAMAAJ`.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures On Stochastic Programming: Modeling and Theory* (SIAM).

Skala M (2009) Counting distance permutations. *Journal of Discrete Algorithms* 7(1):49–61.