Estimating Primary Demand in Bike-Sharing Systems

Chong Yang Goh
Operations Research Center, Massachusetts Institute of Technology, Cambridge MA, USA

Chiwei Yan
Uber Technologies, Inc., San Francisco CA, USA
Department of Industrial and Systems Engineering, University of Washington, Seattle WA, USA

Patrick Jaillet
Operations Research Center, Massachusetts Institute of Technology, Cambridge MA, USA

1. Introduction

Over the past decade, bike-sharing services have seen widespread adoption in cities across the world in response to urban congestion. Most bike-sharing services operate by providing a fleet of rentable bikes that are distributed over a fixed network of *stations*, each capable of holding a certain number of bikes at a given time. A commuter can pick up a bike from any station for temporary use and drop it off later by parking it in an available dock at the destination station.

In this paper, we focus on the problem of estimating the primary demand in a bike-sharing system, which is the crucial input to various operational models such as bike inventory allocation and rebalancing. Here, the primary demand refers to the demand that would be observed if the service never goes out anywhere—or equivalently, if every pick-up or drop-off request by a potential commuter is fulfilled at the station of her first choice. For operational purposes, the demand is usually quantified by the bike outflow and inflow rates at individual stations, or the flow rates between origin-destination pairs. Estimating these flow rates by naively aggregating past trip data, however, can be problematic for two reasons: *demand censoring* and *choice substitution*—both can introduce significant biases in the estimates if not properly corrected for. Demand censoring occurs when stations are temporarily out of bikes or docks, during which there is apparently zero demand. Choice substitution occurs when commuters substitute their first choices for other alternatives due to service unavailability. Choice substitution in bike-sharing can happen in two ways: commuters may decide to use available stations nearby instead of their first choices—in which case, the demand is *recaptured* elsewhere; otherwise, commuters may leave the system for external alternatives, such

as other modes of transportation, hence the demand is said to be *spilled* or lost. Both factors can contribute to either overestimation (by recaptures) or underestimation (by spills) of the primary demand.

2. Methodologies and Contributions

We consider a bike-sharing system with N stations, denoted as $\mathcal{N} \triangleq \{1, ..., N\}$. We define the commuter's choice set as $\mathcal{N} \cup \{0\}$, where 0 denotes the choice of not using the service. To model the demand, we assume that commuters arrive to the system according to a Poisson process with rate λ per unit time. Each arriving commuter is characterized by a *type* $\sigma : \{0, \dots, N\} \mapsto \{0, \dots, N\}$, a bijection that specifies a strict ordering over the stations as follows: for any two stations $i, j \in \{0, \dots, N\}$, i is preferred over j if and only if $\sigma(i) < \sigma(j)$, i.e., choice i is ranked *higher* than choice j. We assume that the type of each arriving commuter is drawn independently from a probability distribution over the set of all possible types, $\Sigma \triangleq \{\sigma_1, \dots, \sigma_K\}$. We parametrize the probability mass function of this distribution with a vector $x \triangleq (x_1, \dots, x_K)$, where x_k is the probability that the commuter is of the type σ_k. Overall, the demand model is parametrized by the arrival rate λ and the type probability vector x, both of which are unknown and have to be estimated from data. Suppose we observe a pick-up at station j while the set of available stations is $S \subseteq \mathcal{N}$ (S is called the *offer set*). We say that a type σ is *compatible* with the observed choice $j \in S$ given the offer set S if $\sigma(j)$ corresponds to the highest rank among the alternatives in S. The set of all such types, $\mathcal{M}_j(S)$ is defined as follows,

$$\mathcal{M}_j(S) \triangleq \{k \in \{1, \dots, K\} : \sigma_k(j) < \sigma_k(i), \forall i \in S, i \neq j\}.$$

For a given probability vector x that parametrizes the type distribution, the probability that a commuter chooses option j when presented with the offer set S is given by

$$\mathbb{P}(j \mid S\,; x) \triangleq \begin{cases} \sum_{k \in \mathcal{M}_j(S)} x_k, & \text{if } j \in S \cup \{0\}, \\ 0, & \text{otherwise.} \end{cases}$$

The function $\mathbb{P}(\cdot \mid \cdot\,; x)$ fully specifies a choice model that predicts the fraction of commuters picking a bike from any station or not using the service under any service outage pattern. In particular, $\mathbb{P}(j \mid \mathcal{N}\,; x)$ and $\lambda_j \triangleq \lambda \mathbb{P}(j \mid \mathcal{N}\,; x)$ correspond to the *first-choice* probability and *primary demand* for station j, respectively, while $\rho(S; x) \triangleq \sum_{j \in S} \mathbb{P}(j \mid S; x) = 1 - \mathbb{P}(0 \mid S; x)$ is the probability that given the offer set S, an arriving commuter will stay and use the service.

Based on this rank-based demand model, we summarize the main contributions of this paper in two areas: *modeling* and *estimation*.

Modeling. We propose to augment the aforementioned rank-based choice model with a *substitution graph*, which allows us to precisely characterize the set of substitutable alternatives for each

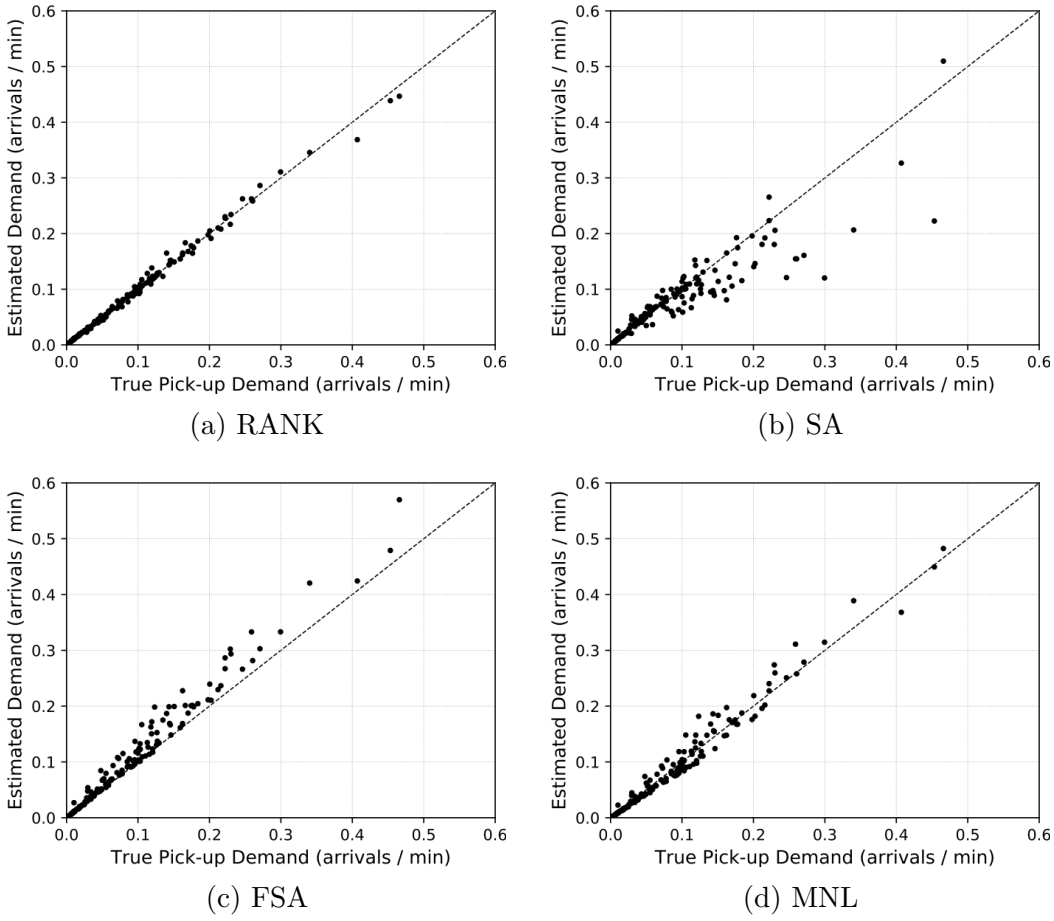
station in the bike-sharing network. The introduction of these substitution constraints serve two purposes: (i) to model commuters' tendency to explore only alternatives that are in close proximity to their first-choice stations, and (ii) to reduce the number of parameters of the model by eliminating preference rankings that imply unrealistic substitution behaviors. Further, we extend the rank-based choice model to jointly account for pick-up and drop-off substitutions using *paired* rankings of OD choices. This extended model allows us estimate the primary demand for trips between all origin-destination (OD) station pairs, which are useful for inventory planning operations (e.g., Shu et al. (2013)).

Estimation. Our proposed model is characterized by a set of parameters—the Poisson arrival rate and the probability mass function of preference rankings—that can be estimated by maximum likelihood estimation (MLE). One issue, however, is that the number of preference rankings grows factorially with the number of alternatives. Therefore, it is usually not feasible to solve the fully parametrized MLE problem using off-the-shelf optimization solvers. Instead, large-scale optimization methods such as constraint sampling (Farias, Jagabathula, and Shah 2013) and column generation (van Ryzin and Vulcano 2014) have been applied. However, the column generation subproblem requires solving an integer program that is NP-hard, making it intractable when there are hundreds of stations, as is the case in many bike-sharing systems. In this paper, we take a different approach by proposing a dimensionality reduction technique that is based on *sparse representations*. More specifically, given a sequence of observed available stations (offer sets) over time, our algorithm will generate a parsimonious set of preference rankings that sufficiently explain the observations, subject to the constraints imposed by a substitution graph. For our application, we find that this approach can reduce the problem dimension by orders of magnitude, making it feasible to solve the MLE problem without resorting to large-scale optimization methods. Our second contribution is focused on estimation. We show that the MLE problem, while nonconcave in general, can be reduced to a *difference-of-convex* (DC) program. Based on this reduction, we apply an iterative DC programming technique that estimates the parameters by solving a sequence of convex programs and each convex program can be solved efficiently by the Frank-Wolfe method.

3. Results

We evaluate our methods on synthetic and actual data sets of Hubway, a bike-sharing service in the Boston metropolitan area. We demonstrate the efficiency and practicality of our method on a city scale. We further show that across a wide range of simulated conditions, our method consistently outperforms several benchmarks, including the independent demand and MNL models. Specifically, we achieve error reduction by more than 20% when the average stockout frequency is close to the actual levels observed in Boston during the evening peak hours. These improvements translate into

ridership increases of up to 3% when bikes are allocated based on the estimated demands, using a bike deployment model by Shu et al. (2013) with a planning horizon of several hours. When applied to the actual Hubway data set, our method achieves an overall out-of-sample error of 24.8% in ridership prediction, which is 10% lower compared to the best-performing benchmark.



(a) RANK (b) SA

(c) FSA (d) MNL

Figure 1 **Comparison of estimated and true pick-up demand for 171 Hubway stations. The results are averaged over 10 independent simulations with a rebalancing interval of $T_r = 180$ min.**

References

Farias VF, Jagabathula S, Shah D, 2013 *A Nonparametric Approach to Modeling Choice with Limited Data. Management Science* 59(2):305–322.

Shu J, Chou MC, Liu Q, Teo CP, Wang IL, 2013 *Models for Effective Deployment and Redistribution of Bicycles Within Public Bicycle-Sharing Systems. Operations Research* 61(6):1346–1359.

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.