THE UNIVERSITY OF CHICAGO

ESTIMATION OF SEQUENTIAL SEARCH MODELS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
JAE HYEN CHUNG

CHICAGO, ILLINOIS
JUNE 2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I first would like to thank two of my advisors, Pradeep K. Chintagunta and Sanjog Misra, for collaborating with me on this dissertation. I gratefully acknowledge their great insights that have made it possible to complete this dissertation. I also would like to express my gratitude to Ruey S. Tsay for leading me into the field of the quantitative marketing when I was lost amidst the overwhelming number of possibilities and to Raluca Ursu for helpful conversations. Lastly, I thank my friends and family whose unwavering support I relied upon in time of needs.

ABSTRACT

We propose a new likelihood-based estimation method for the sequential search model. By allowing search costs to be heterogeneous across consumers and products, we can directly compute the joint probability of the search sequence and the purchase decision when consumers are searching for the idiosyncratic preference shocks in their utility functions. Under this procedure, one recursively makes random draws for each dimension that requires numerical integration to simulate the probabilities associated with the purchase decision and the search sequence under the sequential search algorithm. We then present details from an extensive simulation study that compares the proposed approach with existing estimation methods recently used for sequential search model estimation, viz., the kernel-smoothed frequency simulator (KSFS) and the crude frequency simulator (CFS). In the empirical application, we apply the proposed method to the Expedia dataset from Kaggle which has previously been analyzed using the KSFS estimator and the assumption of homogeneous search costs. We demonstrate that the proposed method has a better predictive performance associated with differences in the estimated effects of various drivers of clicks and purchases, and highlight the importance of the heterogeneous search costs assumption even when KSFS is used to estimate the sequential search model. Lastly, from a managerial perspective, we show that sorting products by their expected utilities can enhance consumer welfare and increase the number of transactions.

CHAPTER 1

INTRODUCTION

Applications of the traditional discrete choice model in marketing rely on the assumption that the consumer is aware of all products on the market and their characteristics (i.e., prices, product attributes, etc.). Since the work of Howard and Sheth (1969), however, the concept of the consideration set has been well recognized in the marketing literature. Instead of assuming omniscient consumers, the notion of the consideration set posits that consumers restrict their attention to a subset of products on the market and purchase a product within that set (e.g., by picking the alternative with the highest utility). This large and growing literature includes, among many others, studies by Ratchford (1980); Roberts (1989); Hauser and Wernerfelt (1990); Roberts and Lattin (1991); Siddarth et al. (1995).

In the recent literaure, the formation of the consideration set has been modeled as the result of costly consumer search; i.e., consumers are *a priori* uncertain about certain characteristics of the product (e.g., price) and spend resources (temporal, psychological, etc.) to resolve that uncertainty. The presence of such search costs to resolve uncertainty results in the consumer obtaining information only on a subset of products in the market—the consideration set (Mehta et al., 2003; Kim et al., 2010, 2016; Honka and Chintagunta, 2016)—since the consumer is trading off the benefits and costs of searching. Following the search literature in economics and marketing (Stigler, 1961; Ratchford, 1980; Weitzman, 1979), consideration set formation based on search theory has used the assumptions of either simultaneous (Stigler, 1961) or sequential (Weitzman, 1979) search.

In the context of online search, for example, consumers submit search queries on online retail websites (e.g. Expedia.com or Amazon.com) and receive lists of products. Since such lists often do not reveal all the information on the listed products, the consumer has to browse the list and click on a subset of products in order to obtain all the information on those products (as clicking takes the consumer to a page with more information on the product). In terms of the search model, the act of clicking corresponds to searching and

the subset of products the consumer clicks on is viewed as the consideration set (Kim et al., 2010; Ursu, 2018; Chen and Yao, 2016; De Los Santos et al., 2012). The consumer then makes a choice from the set of clicked alternatives on the list, or chooses to forego making a choice (i.e, the consumer picks the "outside" option of no-purchase). A majority of the recent literature on online search has used the assumptions of sequential search in order to obtain the consideration set.[1]

A key challenge faced by researchers is with taking the search model-based formulation of consideration sets and observed choice behavior to the data. In the traditional discrete choice literature, the outcomes observed in the data are the actual choices made by consumers. Consequently, constructing the probability of the observed choice, conditional on the chosen structure of the choice model (e.g., logit), is straightforward under the assumption that the consumer is fully informed and considers all the products in the category. Thus, one can formulate the likelihood function for a series of choices made by, say, a panel of consumers, and then estimate the parameters of the model by maximizing the likelihood function. On the other hand, in the case where consideration sets are modeled as the result of the sequential search algorithm by Weitzman (1979), there is no closed-form solution for the search set probability or the conditional purchase probability. Thus, researchers have typically resorted to one of two types of simulation-based estimators to construct the likelihood of observed consumer search and purchase decisions. These simulators include the crude frequency simulator (CFS) and kernel-smoothed frequency simulator (KSFS).

For instance, Chen and Yao (2016) use online click-stream data of consumers searching for hotel reservations in order to study the effects of the sorting and filtering options presented to consumers. In order to construct the simulated likelihood, they rely on CFS. That is, given the current parameter estimates, they make a large number of random draws for variables that require numerical integration (e.g. consumer heterogeneity and search costs in their

1. A point to note in this literature is that as the consideration set is modeled as the result of consumer search, the terms consideration set and search set are often used interchangeably, as we do in this dissertation.

case). Then, they approximate the joint probability of the search sequence and purchase decision by the proportion of random draws that satisfy conditions stipulated by the order in which alternatives are searched and the hotel ultimately purchased. The final parameter estimates are those that maximize the average frequency that such conditions are satisfied.

However, CFS has two problems. First, as CFS approximates a probability by the proportion of random draws that satisfy the search set conditions, the simulated likelihood is technically a step-function. So one needs to make a large number of random draws to have well-behaved objective function. And even so, one cannot use gradient-based optimization routines to maximize the likelihood from CFS, as they require smooth objective functions. Second, the likelihood of an unpopular search sequence (i.e., one that has a low probability of being observed in the data) can be inadequately approximated. As an illustration, consider the very simple case in which we need to make only one-dimensional random draws. If one makes n random draws, the smallest possible value of simulated probability is $\frac{1}{n}$. If there is an observation whose true probability lies between 0 and $\frac{1}{n}$, then the estimation routine will approximate the probability to be $\frac{1}{n}$. Therefore, with a finite number of draws, CFS has a risk of inadequately approximating the true probabilities in the data. Increasing the number of draws might help address this problem, but, even so, the curse of dimensionality would slow down the computations, and the increased computational burden will eclipse the computational simplicity of the CFS approach. This renders CFS as an unappealing option in many practical situations.

Next, Honka and Chintagunta (2016) aim to identify the search strategy employed by consumers in the auto insurance market, i.e., whether they adopt sequential or simultaneous search. For both types of search strategies, they use the KSFS with the kernel from a multivariate logistic CDF to construct the simulated likelihood. Specifically, search set conditions (corresponding to the inclusion of an item in the consideration set and to being chosen conditional on inclusion in the consideration set) given in the form of inequalities are first converted to expressions that return positive values when those conditions are met

and negative values otherwise, given the current estimates and random draws for variables that require numerical integration. These expressions are each weighted by a scaling factor. Then, they apply a kernel to the values returned from these weighted expressions and take an average of such kernel-smoothed values across random draws. The parameter estimates are chosen to maximize the average kernel-smoothed value since this value is constructed to increase when the search conditions are met and decrease when they are not. In their application, they assign equal scaling factors to the expressions associated with each of the inequalities.[2]

Applying KSFS to construct the simulated likelihood provides researchers a way to circumvent difficulties of estimating a model. Instead of directly calculating the probabilities of search set conditions, one assumes that the search set conditions follow a distribution that is the same as the chosen kernel. However, one important shortcoming of KSFS is that the chosen kernel is not a part of the model itself. While we often use the error distribution (e.g. logit or probit) to calculate the likelihood, KSFS constructs the simulated likelihood from the chosen kernel, not from the error distribution. Therefore, the estimates are sensitive to the choice of kernel. Also, by applying a kernel to the values returned from the expressions characterizing the search set conditions, KSFS penalizes (rewards) the parameter values that violate (satisfy) these conditions. So the choice of scaling factor in KSFS dictates the balance between the search set conditions being satisfied or violated—a larger value penalizes and rewards more than a smaller value does. As a result, the estimates will be sensitive to the choice of scaling factor. In principle, each scaling factor should be calibrated to reflect the contribution of the corresponding expression to the consideration set and purchase probabilities (Ursu et al., 2018). Since there is no fixed rule to choose the kernel and the scaling factors, KSFS should be used with caution when applied to actual real-world data (although they can be calibrated in simulation settings).

2. Another implementation of the KSFS is in Elberg et al. (2019). An interesting feature of that study is that the authors derive a closed-form expression for the reservation utility (described later in this dissertation) under the assumption of a logistic distribution for the unobserved terms in the utility functions of consumers.

In this dissertation, we present a likelihood-based estimator for sequential search models that addresses the issues associated with current methods used to estimate the parameters of search models of consideration and choice. Instead of using the aforementioned simulators, the proposed method directly simulates the likelihood, forcing the random draws to satisfy the search set conditions. The proposed method has several advantages vis-a-vis the frequency simulators currently in use. First, it requires a relatively small number of random draws. Second, it produces an approximate likelihood associated directly with the probabilities of the actual events of interest. Lastly, it produces more precise estimates with better predictive performance. After explaining the technical details of the estimation procedure, we present extensive simulation results and compare the performance of the proposed method and the two frequency simulators. Next, we apply this approach to a real-world dataset to demonstrate its application and to highlight the advantages of the proposed method.[3]

The remainder of the dissertation is organized as follows. Chapter 2 briefly discusses the sequential search algorithm (Weitzman, 1979) and the concepts of search cost and reservation utilities. We also discuss the setup of the model in Chapter 2. Chapter 3 presents the construction of the simulated likelihood. Chapter 4 briefly discusses the identification strategy. Chapter 5 presents the simulation studies. Chapter 6 presents the empirical application of the proposed estimation strategy. Chapter 7 concludes the dissertation.

3. Other likelihood-based approaches suggested in the literature include those by Kim et al. (2016) (using aggregate-level search and market share data) and Moraga-González et al. (2015) (using a combination of individual and aggregate data). Our approach is based on the commonly observed individual data structure wherein consumers' search and purchase behavior are recorded at the individual level and represents a direct translation of Weitzman (1979)'s result into an empirical likelihood.

CHAPTER 2

SEQUENTIAL SEARCH AND MODEL

2.1 Utility

We model the consumer's decision to search and purchase. As will be discussed in the next section in more detail, consumers are assumed to search across products according to the optimal sequential search algorithm. That is, a consumer will continue searching as long as the expected marginal benefit is greater than the marginal cost of one additional search. Once the search process is terminated, the consumer makes the purchase decision among the searched products according to their realized utilities (i.e., utilities revealed after resolving any uncertainty associated with elements of the utility function).

Following the literature, consumer i has utility, $u_{i,j}$ for product $j = 1, \cdots, J$ each with a (row) vector of characteristics X_j:

$$u_{i,j} = X_j\theta_i + \epsilon_{i,j} = \mu_{i,j} + \epsilon_{i,j} \tag{2.1}$$

with

$$\epsilon_{i,j} \sim N(0, \sigma^2), \quad \theta_i \sim N(\bar{\theta}, \sigma_\theta^2)$$

where $\mu_{i,j}$ denotes consumer i's expected utility from product j, and $\epsilon_{i,j}$ refers to the consumer's idiosyncratic preference or "match value", which is assumed to be identically and independently distributed across consumers and products. Consumers are uncertain about this preference, and search to resolve the uncertainty.

In our setting, as X_j is available prior to search (prior to click, in the context of online search), the expected utility $\mu_{i,j}$ is also known to consumers prior to search. But consumer's utility from each product depends both on $\mu_{i,j}$ and on $\epsilon_{i,j}$, and the latter is assumed to be unknown prior to search. Therefore, consumers engage in the search process in order

to reveal the $\epsilon_{i,j}$ and the true utility. Consumers, however, do know that the idiosyncratic preference follows the distribution above. We assume that the consumer has an outside option; the utility of this option has the following distribution.

$$u_{i,0} \sim N(\mu_0, \sigma^2)$$

In the empirical analysis, the outside option's realized utility is assumed either to be known prior to the first search or to be revealed by the first search. One should pick an assumption that fits the structure of data. In the case of online search, if the data include consumers who submit a search query and leave the website without clicking on any option, then one should assume that $u_{i,0}$ is known prior to the first search because such consumers find that the marginal benefit of the first search is not greater than the cost of search and they choose the outside option without clicking or searching. On the other hand, if the data include only consumers who make at least one click, then one should assume that $u_{i,0}$ is revealed after the first search. Next, we discuss three components of the model: (1) the optimal sequential search algorithm by Weitzman (1979), (2) the relationship between search cost and reservation utility, and (3) other model assumptions.

2.2 Optimal Sequential Search Algorithm

According to Weitzman (1979), the optimal sequential search strategy is composed of the following three stages.

1. Selection : Consumers compute the reservation utilities of all products based on the search cost and the expected utility, which are available prior to search. Then, they begin searching from the product with the highest reservation utility. With each search, the uncertainty regarding the match value is resolved for that product; the utility for the product revealed after search is referred to as the realized utility.

2. Stopping : Consumers stop searching once the maximum of realized utilities in the

searched set exceeds the maximum of the reservation utilities of the non-searched options.

3. Choice : After stopping, consumers purchase the option in the search set that has the highest realized utility. Note that the searched set also includes the outside option.

Using the framework of Kim et al. (2010), under the sequential search model, the expected marginal benefit from searching product j, given the current maximum realized utility u^*, is given by

$$B_j(u^*) = \int_{u^*}^{\infty} (u_j - u^*) f_j(u_j) du_j \tag{2.2}$$

where $f_j(u_j)$ is the probability density function of u_j.

Weitzman (1979) shows that if the stochastic components of the utility are uncorrelated across products, consumers make search decisions based on the reservation utilities of products. Product j's reservation utility z_j is defined to be the hypothetical level of utility that makes a consumer indifferent between searching and not searching the alternative j. That is, if a consumer has already found an option k with realized utility higher than the reservation utility of option j, then a consumer would not search option j. And a consumer will search option j only if its reservation utility is higher than the current maximum of already-searched options' realized utilities.

According to this definition, reservation utility of product j, z_j, and the (positive) search cost of product j, c_j, satisfy the following equation.

$$0 < c_j = B_j(z_j) = \int_{z_j}^{\infty} (u_j - z_j) f_j(u_j) du_j \tag{2.3}$$

In words, a consumer is indifferent between searching and not searching product j even if the expected marginal benefit is positive because searching incurs a positive search cost, c_j, that is equal to the benefit of searching, $B_j(z_j)$.

2.3 Search Cost and Reservation Utility

Kim et al. (2010) show that, under Normality assumption on u_j, one can rewrite (2.3) as

$$\begin{aligned} c_j &= \int_{z_j}^{\infty} (u_j - z_j) f_j(u_j) du_j \\ &= (1 - F_j(z_j)) \left[\int_{z_j}^{\infty} (u_j - z_j) \frac{f_j(u_j)}{1 - F_j(z_j)} du_j \right] \\ &= (1 - F_j(z_j)) E[u_j | u_j > z_j] \\ &= \left(1 - \Phi\left(\frac{z_j - \mu_j}{\sigma} \right) \right) \left[\mu_j - z_j + \sigma \frac{\phi(\frac{z_j - \mu_j}{\sigma})}{1 - \Phi(\frac{z_j - \mu_j}{\sigma})} \right] \end{aligned} \tag{2.4}$$

where $F_j(z_j)$ denotes the cumulative distribution function of u_j evaluated at z_j, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, $\phi(\cdot)$ denotes the probability density function of the standard normal distribution, and $u_j \sim N(\mu_j, \sigma^2)$. In principle, one can derive the relationship between the search cost and the reservation utility upto the third line of (2.4) for any arbitrary distribution assumption on u_j. The normality assumption allows us to further simplify the (implicit) equation. If we define $\gamma_j \equiv \frac{c_j}{\sigma}$ and $\eta_j \equiv \frac{z_j - \mu_j}{\sigma}$, (2.4) can be rewritten as

$$\begin{aligned} c_j &= \left(1 - \Phi\left(\frac{z_j - \mu_j}{\sigma} \right) \right) \left[\mu_j - z_j + \sigma \frac{\phi(\frac{z_j - \mu_j}{\sigma})}{1 - \Phi(\frac{z_j - \mu_j}{\sigma})} \right] \\ &= (1 - \Phi(\eta_j)) \left[\sigma \frac{\phi(\eta_j)}{1 - \Phi(\eta_j)} - \sigma \eta_j \right] \\ \Rightarrow \gamma_j &= (1 - \Phi(\eta_j)) \left[\frac{\phi(\eta_j)}{1 - \Phi(\eta_j)} - \eta_j \right] \\ &= \phi(\eta_j) - \eta_j (1 - \Phi(\eta_j)) \\ &= g(\eta_j) \end{aligned} \tag{2.5}$$

$g(\cdot)$ is a mapping from the standardized reservation utility η_j to a scaled version of the search cost γ_j. A closer examination of the function g reveals another important advantage

of the normality assumption. Taking the derivative of γ_j with respect to η_j or taking a look at Figure 2.1, it can be inferred that function g is a monotonically decreasing one-to-one mapping within a reasonable range of η. It suggests that there is an inverse of the function g and that even if there is no closed-form solution for the inverse of function g, there is a unique η_j that corresponds to each γ_j.

$$\begin{aligned}\frac{d\gamma_j}{d\eta_j} &= -\eta_j\phi(\eta_j) - (1-\Phi(\eta_j)) - \eta_j(-\phi(\eta_j)) \\ &= -(1-\Phi(\eta_j)) \in (-1,0)\end{aligned}$$

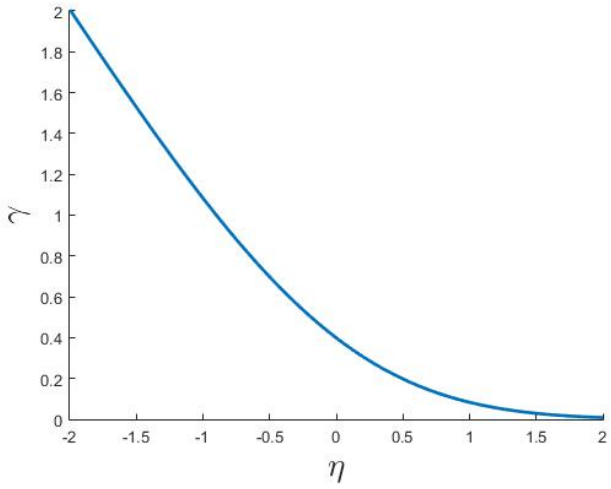


Figure 2.1: Function $g : \eta \to \gamma$

Using the function g, given μ_j, z_j, and σ, one can find c_j that corresponds to z_j by using the following expression

$$c_j = g\left(\frac{z_j - \mu_j}{\sigma}\right)\sigma$$

Also, using the inverse of function g, one can find z_j that corresponds to some search cost c_j by

$$z_j = \mu_j + \sigma g^{-1}\left(\frac{c_j}{\sigma}\right)$$

In order to use the inverse of function g, Kim et al. (2010) create a table of γ values that corresponds to fine grids of η prior to the estimation step and use the table in the estimation step. Instead, in this dissertation, we approximate the function g^{-1} via cubic spline interpolation prior to the estimation step and use it in the estimation.

The function g plays a crucial role in the construction of the likelihood. The difficulty of estimating a sequential search model stems from the fact that there is no closed-form distribution of the reservation utility under the usual probit or logit settings. But the aforementioned features of function g allow us to use the search cost distribution to compute the distribution of reservation utilities. To illustrate, let us assume that we want to compute the probability that reservation utility, z, is less than some value, a. Then, we can write the probability of z in terms of the search cost as follows.

$$\begin{aligned}
\Pr(z < a) &= \Pr(\mu + \sigma\eta < a) \\
&= \Pr\left(\eta < \frac{a-\mu}{\sigma}\right) \\
&= \Pr\left(g(\eta) > g\left(\frac{a-\mu}{\sigma}\right)\right) \\
&= \Pr\left(\gamma > g\left(\frac{a-\mu}{\sigma}\right)\right) \\
&= \Pr\left(\sigma\gamma > \sigma g\left(\frac{a-\mu}{\sigma}\right)\right) \\
&= \Pr\left(c > \sigma g\left(\frac{a-\mu}{\sigma}\right)\right) \qquad (2.6)
\end{aligned}$$

Note that after the third equality sign in (2.6), the inequality changes its direction as g is a decreasing function. Using this conversion and the distributional assumption on c, one can compute the above probability. This conversion of probability plays a crucial role in the construction of the likelihood, which we will discuss in a subsequent section.[1]

1. In this dissertation, we present a method to estimate the sequential search model with preference shocks assumed to follow the normal distribution as in Kim et al. (2010). However, a recent work by Elberg et al. (2019) shows another closed-form conversion between the reservation utility and the search cost under the logistic distribution assumption on the preference shock. As their conversion also allows the reservation utility to be the sum of expected utility and some function of search cost, the proposed method in this

2.4 Model Setup

In this section, we lay out the setup of the model, which does not deviate much from the standard model in the consumer search literature. As discussed in the previous section, we assume that consumers engage in costly search in order to find the match value by revealing the value of the idiosyncratic preference shock, $\epsilon_{i,j}$. Prior to search, consumers are aware of the mean of realized utility, $\mu_{i,j}$, of all products available on the market and their search costs, $c_{i,j}$, as well as the distribution of the idiosyncratic preference, $\epsilon_{i,j}$.[2] Based on this information, consumers compute the reservation utilities and begin searching from products with highest reservation utilities, as described in the previous section.

As the dataset used in the empirical application of this dissertation contains only consumers who make at least one click, it is assumed that the first search is free and it reveals the match value of the clicked option and that of the outside option (which usually refers to no purchase). From this point, the outside option serves just as another inside option. That is, when deciding to continue searching or not, consumers compare the highest realized utility of the searched options, including clicked products and the outside option, to the highest reservation utility of the products that have not yet been searched. Once the stopping criterion is met, consumers make purchase decisions according to the realized utilities of the searched products (including the outside option).

On the other hand, if a dataset contains consumers who make no clicks, this assumption would have to change. In this case, one needs to assume that prior to clicking any product, consumers are aware of the realized utility of the outside option. So the decision to make a first click depends on the realized utility of the outside option and the reservation utilities of the inside products. For example, if a consumer leaves the website without making any click, it suggests that the reservation utilities are all lower than the outside option's realized

dissertation can easily accommodate the specification in Elberg et al. (2019) with small modifications and without increasing the computational burden.

2. In the context of online shopping, consumers know the expected utility and search cost of all products presented on the search result page.

utility. And if consumers click on a product, then one can infer that the clicked product's reservation utility was higher than the outside option's realized utility.

Following Moraga-González et al. (2015) and Yao and Mela (2011), we assume that search costs are heterogeneous at the consumer-product level and are drawn from a common distribution across consumers and products. Given positive search costs, we assume a log-normal distribution for search costs.

$$\log(c_{i,j}) \sim N(\bar{c}, \sigma_c^2)$$

Not only does this assumption allow us to compute the probability of the search sequence, but it also rationalizes the variation in search sequences observed among consumers, an implication that will be discussed in the empirical application section.

CHAPTER 3

LIKELIHOOD

Before we discuss the likelihood, for the rest of this dissertation, we fix the variance of $\epsilon_{i,j}$ to 1 and let S_i denote the (ordered) search set of consumer i, D_i denote the product purchased by consumer i (by assumption, $D_i \in \{S_i \cup 0\}$, where 0 is the index for the outside good), r_k denote the index of a product that has k-th highest reservation utility of all products on the market (e.g. r_1 returns the index of product with highest reservation utility), K_i denote the length of search sequence of consumer i, and J denote the total number of products available on the market.

The likelihood of observing search sequences and purchase decisions of N consumers can be expressed as follows.

$$\begin{aligned}
\mathcal{L} &= \prod_i \Pr(S_i = s, D_i = j^* | \Phi, X_i) \\
&= \prod_i E[\mathbb{I}(S_i = s, D_i = j^* | \Phi, X_i)] \\
&= \prod_i \int_\xi \mathbb{I}(S_i = s, D_i = j^* | \vec{u}_i(\xi_i|X_i), \vec{z}_i(\xi_i|X_i)) dF(\xi|\Phi)
\end{aligned} \tag{3.1}$$

where $\Phi = \{\bar{\theta}, \sigma_\theta, \bar{c}, \sigma_c\}$, $\xi_i = \{\vec{c}_i, \theta_i, \vec{\epsilon}_i\}$, and X_i denotes the matrix of characteristics of all products that are available to consumer i prior to search. $\vec{u}_i(\xi_i|X_i)$ and $\vec{z}_i(\xi_i|X_i)$ denote vectors of consumer i's utilities and reservation utilities given ξ_i and X_i. We approximate the likelihood above by constructing the simulated likelihood as follows.

$$\mathcal{L} \approx \prod_i \frac{1}{n_d} \sum_{d=1}^{n_d} \Pr(S_i = s, D_i = j^* | \vec{u}_i^{(d)}, \vec{z}_i^{(d)}) \tag{3.2}$$

where $\vec{u}_i^{(d)}$ and $\vec{z}_i^{(d)}$ denote vectors of utilities and reservation utilities, respectively, computed at the d-th draw of ξ_i components. More specifically, given $\xi_i^{(d)} = \{\vec{c}_i^{(d)}, \theta_i^{(d)}, \vec{\epsilon}_i^{(d)}\}$,

we use

$$\begin{aligned}\vec{u}_i^{(d)} &= X_i\theta_i^{(d)} + \vec{\epsilon}_i^{(d)} \\ \vec{z}_i^{(d)} &= X_i\theta_i^{(d)} + g^{-1}(\vec{c}_i^{(d)})\end{aligned} \tag{3.3}$$

to construct the quantities in (3.2).

We note that search set conditions vary according to the number of searched products and the position of purchased product within the search sequence. In this section, we first translate the joint decision of search and purchase into a set of inequalities for each case. Then, we describe how to construct the simulated likelihood.

3.1 Joint Probability of Search Set and Purchase Decision

The joint probability depends on the sequence of search, the length of search sequence, the purchase decision, and the position of purchased product within the search sequence. For a comprehensive treatment, we need to consider three cases—(1) $K_i = 1$, (2) $1 < K_i < J$, and (3) $K_i = J$.

3.1.1 Case 1—Single Product Searched ($K_i = 1$)

If a consumer terminates his search process after searching, say, product 2 and does not purchase any product (purchase outside option), the joint probability of the search set and the purchase decision is given by the following:

$$\begin{aligned}\Pr(S_i = \{2\}, D_i = 0) &= \Pr(\underbrace{z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{Selection}} \bigcap \underbrace{u_{i,0} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{Stopping}} \bigcap \underbrace{u_{i,0} > u_{i,2}}_{\text{Choice}}) \\ &= \Pr(u_{i,0} > u_{i,2} \bigcap \min(z_{i,2}, u_{i,0}) > \max_{l \notin \{S_i \cup 0\}} z_{i,l})\end{aligned}$$

$$= \Pr(u_{i,0} > z_{i,2} \bigcap u_{i,0} > u_{i,2} \bigcap z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l})$$
$$+ \Pr(z_{i,2} > u_{i,0} \bigcap u_{i,0} > u_{i,2} \bigcap u_{i,0} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}) \quad (3.4)$$

The need to break the expression down to two cases (as done after the last equality above) arises since there is no closed-form for the distribution of the minimum value of realized utility and reservation utility.

3.1.2 Case 2—Multiple, But Not All Products Searched ($1 < K_i < J$)

If a consumer searches product 2 and then 3 and purchases product 3, the joint probability of the search set and the purchase decision is given by the following.

$$\begin{aligned}
& \Pr(S_i = \{2,3\}, D_i = 3) \\
= \; & \Pr(\underbrace{z_{i,2} > z_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{Selection}} \bigcap \underbrace{u_{i,3} > u_{i,2} \bigcap u_{i,3} > u_{i,0}}_{\text{Choice}}) \\
& \bigcap \underbrace{u_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l} \bigcap z_{i,3} > u_{i,0} \bigcap z_{i,3} > u_{i,2})}_{\text{Stopping}} \\
= \; & \Pr(z_{i,2} > z_{i,3} \bigcap \min(z_{i,3}, u_{i,3}) > u_{i,0} \bigcap \min(z_{i,3}, u_{i,3}) > u_{i,2} \\
& \bigcap \min(z_{i,3}, u_{i,3}) > \max_{l \notin \{S_i \cup 0\}} z_{i,l}) \\
= \; & \Pr(z_{i,3} > u_{i,3} \bigcap u_{i,3} > u_{i,0} \bigcap u_{i,3} > u_{i,2} \bigcap u_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l} \bigcap z_{i,2} > z_{i,3}) \\
& + \Pr(u_{i,3} > z_{i,3} \bigcap z_{i,3} > u_{i,0} \bigcap z_{i,3} > u_{i,2} \bigcap z_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l} \bigcap z_{i,2} > z_{i,3})
\end{aligned} \quad (3.5)$$

If the purchased product coincides with the last-searched product, one needs to break down the probability into two cases just as in the $K_i = 1$ case. If $1 < K_i < J$ and the purchased product does not coincide with the last-searched product (e.g. purchase product

2 or 0 in this case), the probability can be written in a form similar to the first probability after last equality sign of (3.5), since it is implied by the search set and purchase decision that the purchased product's realized utility is lower than the reservation utility of the subsequently searched product.

3.1.3 Case 3—All Products Searched ($K_i = J$)

The last case to consider is one in which a consumer searches through all products on the market. For illustration, let us assume that there are only three products on the market. If a consumer is observed to have searched product 2, 3, and then 1 and purchased product 1, then the joint probability for his search sequence and purchase decision is given by the following.

$$
\begin{aligned}
&\Pr(S_i = \{2,3,1\}, D_i = 1) \\
=&\Pr(\underbrace{z_{i,2} > z_{i,3} > z_{i,1}}_{\text{Selection}} \bigcap \underbrace{u_{i,1} > u_{i,2} \bigcap u_{i,1} > u_{i,3} \bigcap u_{i,1} > u_{i,0}}_{\text{Choice}}) \\
&\bigcap \underbrace{z_{i,1} > u_{i,2} \bigcap z_{i,1} > u_{i,3} \bigcap z_{i,1} > u_{i,0}}_{\text{Stopping}}) \\
=&\Pr(z_{i,2} > z_{i,3} > z_{i,1} \bigcap \min(z_{i,1}, u_{i,1}) > u_{i,2} \\
&\bigcap \min(z_{i,1}, u_{i,1}) > u_{i,3} \bigcap \min(z_{i,1}, u_{i,1}) > u_{i,0}) \\
=&\Pr(z_{i,1} > u_{i,1} \bigcap z_{i,2} > z_{i,3} > z_{i,1} \bigcap u_{i,1} > u_{i,2} \bigcap u_{i,1} > u_{i,3} \bigcap u_{i,1} > u_{i,0}) \\
&+ \Pr(u_{i,1} > z_{i,1} \bigcap z_{i,2} > z_{i,3} > z_{i,1} \bigcap z_{i,1} > u_{i,2} \bigcap z_{i,1} > u_{i,3} \bigcap z_{i,1} > u_{i,0})
\end{aligned}
\tag{3.6}
$$

As in the $1 < K_i < J$ case, if a consumer searches all products on the market and purchases a product other than the last-searched product, then the probability is given by a similar form as the first probability after the last equality sign of (3.6).

3.2 Construction of Simulated Likelihood

As the inequalities dictated by the search sequence and the purchase decision are all laid out in the previous section, in this section we present how one can directly simulate the likelihood for the sequential search model without resorting to KSFS or CFS. The proposed method shares the basic principle with the GHK simulator, named after Geweke (1989, 1991), Hajivassiliou (as reported in Hajivassiliou and McFadden, 1998), and Keane (1990, 1994)—recursively make random draws while enforcing inequality constraints. The difference is that, in the GHK simulator, the inequality constraints are implicitly imposed by the choice and covariance matrix, while in the proposed method, they are explicitly imposed by the search set conditions.

Simply put, let us think about the case when a consumer finished his search process and made a purchase decision. As outlined in the previous section, his search process and purchase decision are entirely determined by the order of reservation utilities and realized utilities. So the basic idea behind the proposed method is that beginning with the realized utility of the purchased product, we can simulate other quantities (realized utilities or reservation utilities) that satisfy the inequalities that determine the search and purchase decision. The difficulty in the construction of simulated likelihood for the sequential search model stems from comparing two different types of utilities, one of which does not have a closed form distribution. The proposed method circumvents this problem by utilizing the heterogeneous search cost assumption and the property of function g. The likelihood simulation from the proposed method can be summarized by the following steps.

1. Given the current parameter estimates, make random draws of the realized utility of purchased product, denoted by u^*.

2. Given random draws of u^*, compute the probability that u^* is the highest among realized utilities of searched options.

3. Using the conversion of inequality presented in Section 2.3, compute the probability

that the inequality between u^* and $z_{r_{K_i}}$, the reservation utility of the last-searched product, is satisfied.

4. Make random draws of $z_{r_{K_i}}$ that satisfy the inequality with u^*.

5. Compute the probability that non-searched products have lower reservation utilities than $z_{r_{K_i}}$ or u^*, depending on the search sequence and purchase decision.

6. In the reverse order of the search sequence, for $l = 0, \cdots, K_i - 2$, compute the probability of the inequality $z_{r_{(K_i-(l+1))}} > z_{r_{(K_i-l)}}$, given random draws of $z_{r_{(K_i-l)}}$, and make random draws of $z_{r_{(K_i-(l+1))}}$ such that they satisfy the inequality.

7. Take the product of probabilities from the steps above and take the average across the random draws.

We refer readers who are interested in more details to Appendix A, where we present detailed instructions on simulating the likelihood for each case presented in Section 3.1.

CHAPTER 4

IDENTIFICATION

There are two sets of parameters to be estimated—the utility or preference parameters $(\bar{\theta}, \sigma_\theta^2)$ and the search cost parameters $(\bar{c}, \sigma_c^2)$. The mean preference parameter, $\bar{\theta}$, is identified by the correlation between the relative search popularity and the product attributes. And its heterogeneity, σ_θ^2, is identified via the heterogeneous composition of search sets. For example, if there are two popular search sequences, one which contains high-priced products and the other which contains low-priced products, then this can be indicative of consumers differing in their price sensitivities.

The mean search cost, $\bar{c}$, is identified by the average length of the search sequence. As an illustration, let us assume that there are only two products, A and B and that there are only two search sequences, one sequence with A and the other with A and then B. The first search sequence implies these inequalities, where μ_j is the expected utility of product j and η_j is the standardized reservation utility as defined in Section 2.3.

$$z_A > z_B \Rightarrow \mu_A + \eta_A > \mu_B + \eta_B$$
$$u_A > z_B \Rightarrow \mu_A + \epsilon_A > \mu_B + \eta_B$$

And the second search sequence implies the following inequalities.

$$z_A > z_B \Rightarrow \mu_A + \eta_A > \mu_B + \eta_B$$
$$u_A < z_B \Rightarrow \mu_A + \epsilon_A < \mu_B + \eta_B \Rightarrow \eta_B > \mu_A - \mu_B + \epsilon_A$$

Therefore, if one observes a higher proportion of people who continue searching after product A, one can infer that η_B should have a higher value more often. Accordingly, the average number of searches identifies the mean of η.

Before discussing the search cost variance, we need to point out a difference between our

setting and the usual discrete choice models, which assume omniscient consumers. In the typical discrete choice model, the mean preference parameters are identified by inequalities similar to the following.

$$\mu_j + \epsilon_j > \mu_{j'} + \epsilon_{j'} \Rightarrow \epsilon_{j'} - \epsilon_j < \mu_j - \mu_{j'} \tag{4.1}$$

And under the search setting, the order of the reservation utilities identifies the mean utilities by inequalities similar to the following.

$$\mu_j + \eta_j > \mu_{j'} + \eta_{j'} \Rightarrow \eta_{j'} - \eta_j < \mu_j - \mu_{j'} \tag{4.2}$$

(4.2) is identical to (4.1), except that ϵ's are replaced by η's. But the typical discrete choice model assumes the distribution of ϵ (e.g. a standard normal distribution), and consequently the distribution of $\epsilon_{j'} - \epsilon_j$ is fixed.

Under the search setting, however, the distribution of η depends both on the mean and on the variance of the search cost distribution. More specifically, $\eta_{j'} - \eta_j$ has zero mean, as j and j' are assumed to have the same search cost distribution, but its variance depends on both $\bar{c}$ and σ_c, as the search cost is assumed to follow a log-normal distribution. Therefore, one needs to fix either $\bar{c}$ or σ_c to some value, thereby forcing the average search sequence length to identify the other search cost parameter. The fixed parameter and the parameter identified by the average search sequence length will scale the inequality above. Throughout the rest of this dissertation, in the Monte Carlo simulation and the empirical application, we choose to fix the search cost variance and estimate the mean search cost.

CHAPTER 5

MONTE CARLO SIMULATION

In this chapter, we present estimation results using synthetic datasets. To generate a synthetic dataset, we simulate 1,000 consumers' search decisions and their following purchase decisions according to the sequential search algorithm with 10 products available on the market and an outside option (no purchase). We repeat this process and generate 100 distinct datasets by varying the random draws for the consumer-product search costs ($c_{i,j}$) and idiosyncratic preference shocks ($\epsilon_{i,j}$)[1] with σ_c fixed at 0.25. For each dataset, we estimate the model and obtain parameter estimates.

In order to demonstrate the performance of the proposed method, we also estimate the same model using KSFS and CFS simulators. The first row of Table 5.1 presents the true value of parameters used to generate the data, and we show the mean and standard deviation of estimates obtained with each simulator. For now, we use the true value of σ_c to estimate the model for all three simulators. To present the issues of KSFS and CFS, which will be discussed in detail in subsequent sections, we vary the scaling factor for KSFS and the number of random draws for CFS. For estimation of the proposed method (denoted by CCM) and KSFS, we use $n_d = 1000$. Lastly, for CFS, we use 25, 50, and 100 random draws for each dimension to examine how the number of draws affects the estimation result.[2] While the number of random draws for CFS is smaller than the other two estimators, the total number of random draws for one observation is actually n_d^2 as laid out in more detail in the appendix. And to ensure a positive value of σ_{θ_2}, we estimate its log value.

From Table 5.1, it can be seen that the proposed method recovers the true parameter values and that the estimates show relatively small variance, implying that the identification

1. Note that, given a vector of parameters, the search sequence and the purchase decision are determined by the search cost and the preference shock.

2. The details of KSFS and CFS closely follow Honka and Chintagunta (2016) and Chen and Yao (2016), respectively. The brief description of KSFS and CFS likelihood construction and the densities of estimates are available in the appendix.

	θ_1	$\bar{\theta}_2$	μ_0	$\bar{c}$	$\log \sigma_{\theta_2}$
True Value	0.45	-1	2.5	-0.35	-0.6931
CCM	0.4521	-1.008	2.5401	-0.3782	-0.6830
	(0.0337)	(0.0421)	(0.0835)	(0.0588)	(0.0587)
KSFS_1	0.3815	-1.2707	3.0672	0.1222	-8.5474
	(0.0447)	(0.0854)	(0.1459)	(0.0428)	(6.2145)
KSFS_5	0.338	-0.8658	2.2868	-0.6582	-1.2338
	(0.0340)	(0.0634)	(0.1241)	(0.0507)	(1.5782)
KSFS_{10}	0.3954	-0.9201	2.4142	-0.6538	-0.7980
	(0.0321)	(0.0579)	(0.1184)	(0.0689)	(0.1118)
KSFS_{50}	0.3965	-0.9071	2.2704	-0.5526	-0.7837
	(0.0494)	(0.0752)	(0.2149)	(0.1464)	(0.1629)
KSFS_{100}	0.3638	-0.8465	2.0756	-0.6223	-0.8832
	(0.0592)	(0.0939)	(0.2392)	(0.1816)	(0.2044)
KSFS_{250}	0.3267	-0.7707	1.9059	-0.8052	-0.9364
	(0.0886)	(0.1038)	(0.3389)	(0.2387)	(0.2547)
CFS_{25}	0.2164	-0.7678	-0.0143	-0.9442	-0.1109
	(0.2347)	(0.5010)	(0.7251)	(1.2530)	(0.5269)
CFS_{50}	0.4145	-0.7834	0.0723	-1.1273	-0.2256
	(0.2310)	(0.4443)	(0.8920)	(1.2177)	(0.5152)
CFS_{100}	0.4297	-0.7317	0.3002	-2.2558	-0.5340
	(0.1998)	(0.4315)	(1.6212)	(2.5829)	(1.6889)

$\text{KSFS}_s \Rightarrow$ KSFS with scaling factor s

$\text{CFS}_n \Rightarrow$ CFS with n random draws for each dimension

Table 5.1: Estimation Results on Synthetic Dataset

strategy discussed earlier works well. In contrast, estimates from the other two simulators are biased or have larger variance. In the next subsections, we discuss why these discrepancies are observed.

5.1 Kernel-Smoothed Frequency Simulator (KSFS)

We first discuss why estimates from KSFS are sensitive to the choice of the scaling factor s, as shown in Table 5.1. As an illustration, Honka and Chintagunta (2016) and Ursu (2018)

use KSFS to estimate the sequential search model with the scaled multivariate logistic CDF to smooth the probabilities as follows.

$$F(w_1, \cdots, w_M; s_1, \cdots, s_M) = \frac{1}{1 + \sum_{m=1}^{M} \exp(-s_m w_m)} \tag{5.1}$$

where s_m's are positive scaling parameters and w_m's are a set of equations that return positive values when a search set's conditions are satisfied. For example, if a consumer searches product 2 and then 3 and purchases product 2, his search sequence and purchase decision imply the following set of inequalities.

$$z_{i,2} > z_{i,3}, \quad z_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}, \quad z_{i,3} > u_{i,2}, \quad u_{i,2} > u_{i,3}, \quad u_{i,2} > u_{i,0}$$

To estimate the sequential search model with KSFS, one first needs to convert the inequalities to the following set of equations that return positive values when the inequalities are satisfied and return negative values otherwise as follows.

$$\begin{aligned}
w_{i,1} &= z_{i,2} - z_{i,3} \\
w_{i,2} &= z_{i,3} - \max_{l \notin \{S_i \cup 0\}} z_{i,l} \\
w_{i,3} &= z_{i,3} - u_{i,2} \\
w_{i,4} &= u_{i,2} - u_{i,3} \\
w_{i,5} &= u_{i,2} - u_{i,0}
\end{aligned}$$

If, given current values of estimates, any of w_m's returns a negative value, then the denominator in (5.1) will have larger value, lowering the approximated probability. Setting s_m to a larger value is equivalent to placing a larger penalty (reward) on a set of parameters that violates (satisfies) the search set conditions. Therefore, the estimates from KSFS balance the extent to which some search set conditions are satisfied and others are violated, given

the choice of scaling factor. It does so without forcing all the search set conditions to be satisfied. As previously noted, the choice of the scaling factors can have implications for the other estimated parameters since they provide relative weights to the various search conditions.

In Appendix D.2, we present the kernel densities of estimates from KSFS with different scaling factor values. As from Table 5.1, it can be seen that estimates are centered around values different from true values. Estimates from KSFS have larger variance than those from the proposed method, locating the true parameter values within the 95% confidence interval. An interesting pattern, especially with smaller s values, is that there is more than one peak in the distribution of some parameter estimates. And the fact that the peak located farther away from the true parameter value diminishes as s increases implies that too small a value of s may introduce multiple local minima and makes it harder for the optimization routine to distinguish them from the global minimum. Estimation results with higher s values show a unimodal pattern, but increasing s does not guarantee better results, as the mean values of some estimates get farther away from the true values.[3]

The mechanism underlying KSFS is intuitive. As under the maximum likelihood estimation, parameters that violate the search set conditions are penalized under KSFS. But KSFS and MLE differ in how the penalty is imposed. In MLE, the penalty is imposed in the form of lowering the likelihood implied by the assumed distribution of the model components (e.g., idiosyncratic preference shock distribution and search cost distribution in this case). But, in KSFS, the penalty is imposed by a kernel that requires a choice of scaling parameter, which determines the extent of penalty, but the kernel and the scaling factor are not a part of the underlying model specification. One may find a combination of a kernel and a scaling parameter that produces estimates similar to those from MLE, but the process of finding such a combination can be tedious. Importantly, the process needs to be repeated

3. Previous literature acknowledges that estimates from KSFS are only asymptotically unbiased (see Hajivassiliou et al. (1996); McFadden (1989)).

for each dataset, and one needs to find a new way to calibrate scaling factors in empirical applications, in which true parameter values are unknown.

5.2 Crude Frequency Simulator (CFS)

Crude Frequency Simulator (CFS), also called Accept-Reject Simulator, is another simulator used to estimate the sequential search model. For instance, Chen and Yao (2016) analyze the impact of search refinement in the context of online hotel search by estimating a sequential search model with CFS. As outlined in Appendix D.3 in more detail, they make random draws of search costs (and construct reservation utilities given the current parameter estimates) and random draws of realized utilities. The random draws of realized utilities are drawn from distributions truncated in accordance with the search set conditions that are in the form of inequalities between realized utilities and reservation utilities. Then, they count the proportion of random draws with which the choice rule and the stopping rule (at each position of the search sequence) are both satisfied.

CFS approximates probabilities by the frequencies, and therefore one needs to make a sufficiently large number of random draws in order to ensure the smoothness of the objective function and avoid the inadequate approximation of probabilities as mentioned in the introduction of this dissertation. Table 5.1 and Figures D.10-D.12 show results using three different number of random draws, 25, 50, and 100.

It can be seen that the CFS provides the worst performance among three estimators in terms of both bias and dispersion of estimates. Also, it can be seen that increasing the number of draws from 25 to 100 per dimension does not improve the performance. Making more random draws, 1,000 per dimension, for example, might yield better estimates, but such a modification will impose a much heavier computational burden, and the advantage of computational simplicity compared to the proposed method will accordingly diminish.

5.3 Proposed Method

Table 5.1 shows that, with the correct value for the variance of the search cost distribution, the proposed procedure has better performance than the two other simulators do. In this section, we present, first, the approximation errors of the proposed method and, second, how the estimates change when one assumes a search cost variance that is different from the true value.

To demonstrate that the proposed method adequately approximates the joint probability of search sequence and purchase decision, we create a synthetic dataset of search and purchase decisions by 1 million consumers. As there is no closed-form solution for the joint probability, the frequency of search sequence and purchase decision from this data is considered to be the true probability of search and purchase decision. And using the true parameter values used to generate the data, we compute the probabilities of search sequence and purchase decision using the proposed method and compare them to the frequencies from the data.[4]

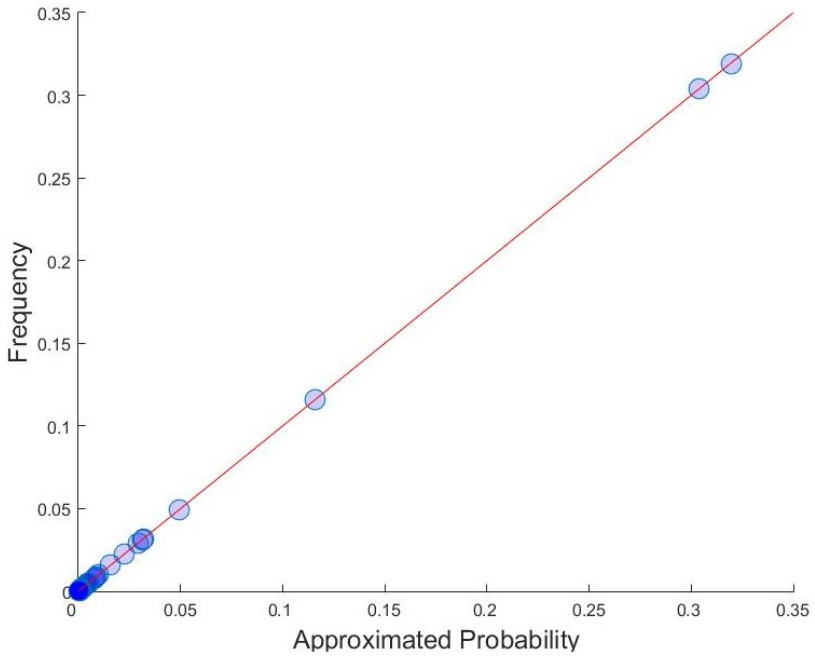


Figure 5.1: Approximated Probability vs. Frequency

4. There are 1,004 unique search sequence and purchase decisions in the synthetic data.

Figure 5.1 compares the approximated probabilities and the true frequencies of search sequences in the data. We take average of approximated probabilities across 500 replications, obtained by varying the random draws for the idiosyncratic preference shocks and search costs. It can be seen that the approximated probability and the true frequency are almost exactly on the solid 45° line, and the correlation coefficient is very close to 1 ($1 - 2.3728 \times 10^{-6}$). Therefore, we can conclude that the proposed method precisely approximates the probability given the true parameters.

Next, we present how estimates vary as the assumed value of σ_c changes and how the proposed method can be useful despite the varying estimates. Conceptually, the heterogeneity in search costs at the consumer-product level may reflect, for example, the randomness in the level of psychological reluctance of executing one additional search and mentally processing additional information from the search. However, there is no way to measure or quantify such heterogeneity, so, in empirical applications, researchers would need to choose the value for the search cost variance parameter.[5]

	θ_1	$\bar{\theta}_2$	μ_0	$\bar{c}$	$\log \sigma_{\theta_2}$	$\frac{\bar{\theta}_2}{\theta_1}$
True Value	0.45	-1	2.5	-0.35	-0.6931	-2.2222
$\sigma_c = 0.25$	0.4521	-1.008	2.5401	-0.3782	-0.683	-2.2366
	(0.0337)	(0.0421)	(0.0835)	(0.0588)	(0.0587)	(0.1126)
$\sigma_c = 0.5$	0.7051	-1.6006	3.7889	-0.5093	-0.3386	-2.2787
	(0.0555)	(0.0625)	(0.1193)	(0.0699)	(0.0558)	(0.1213)
$\sigma_c = 1$	1.0737	-2.5046	5.6501	-0.6124	-0.0239	-2.3436
	(0.0946)	(0.1040)	(0.1978)	(0.0991)	(0.0695)	(0.1363)

Table 5.2: Simulation Results with Different σ_c Assumptions

Table 5.2 shows the estimates from the proposed method using different values of search cost heterogeneity for estimation. The common estimation sample is generated with $\sigma_c = 0.25$. The usefulness of the proposed method and the relative harmlessness of an arbitrary

5. Due to log-normal distribution assumption on the search cost, the assumed value of search cost variance should not be too large. We recommend assuming a value not greater than 1.

choice of the search cost variance can be found in the last column, which shows the mean and standard deviation of the ratio between estimates of θ_1 and $\bar{\theta}_2$. It can be seen that the ratio between two search preference parameters are close to the true ratio regardless of the assumed value for σ_c. This implies that the proposed method can inform researchers of how consumers respond to different product characteristics and marketing activities and their relative efficacy in terms of attracting more consumer searches and more sales, no matter what value is assumed for σ_c. This simulation result implies that in the search setting σ_c serves as a scaling term just as σ_ϵ does in usual discrete choice models. Thus, we conclude that the choice of σ_c is relatively innocuous.

CHAPTER 6

EMPIRICAL APPLICATION—ONLINE HOTEL SEARCH

In this chapter, we apply the proposed method to the online hotel search data and present simulation studies to evaluate its data replication performance and compare it with that of KSFS. Further, we show the implications of the homogeneous or deterministic search cost assumption. Then, through a counterfactual, we examine the welfare effects of different sorting algorithms.[1]

6.1 Data

This dataset is provided by one of the leading online travel agencies (OTA), Expedia.[2] On its website, consumers, upon submitting search queries, are presented with impressions, which are lists of hotels, sorted either by an internal algorithm developed by Expedia or in random order. Consumers are randomly assigned to either one of the two types of impressions.

As the general details of the data can be found in Ursu (2018), which uses the same dataset, we briefly explain each hotel characteristic variable, outline the data cleaning process, and present summary statistics for four major destinations used in the estimation. The dataset contains the hotel characteristics (star rating, review score, price per night, location score, brand, and promotion), which are available to consumers prior to clicking, and each consumer's clicking and purchase decisions, including no purchase. Note that this dataset does not include search impressions on which consumers make no click and therefore each impression has at least one click.

Star ratings are assigned by Expedia according to the type of hotel (e.g. motel, hostel, dormitory, hotel, upscale hotel), the level of luxury, and the amenities provided. Review

1. In Appendix F, we also show how to correct for omitted variable bias such as quality.

2. This dataset was released through the International Conference on Data Mining (ICDM 2013) and Kaggle.com (an online platform for data mining competitions posted by companies). The dataset is available at `www.kaggle.com/c/expedia-personalized-sort/data`.

score variable is the average of the review scores from consumers who made reservations for the hotel on Expedia in the past. Location score (from 0 to 7) is assigned by Expedia to summarize how central a hotel's location is, what amenities surround it, and so on. The brand variable indicates whether the hotel belongs to a chain (e.g. Hilton, Marriott), but we do not have access to which chain it belongs to. These four variables are invariant over the sample period. Price per night shows the pre-tax per-night price presented to consumers. Lastly, the promotion indicator shows whether the hotel has an ongoing promotion.

As outlined in Ursu (2018), (1) we only consider impressions of hotels that are randomly sorted in order to remove the endogeneity of Expedia's sorting algorithm. (2) We remove impressions that include hotels with unrealistically low or high price per night. We exclude impressions in which any hotel's price per night is lower than \$10 or higher than \$1000. (3) We correct potential price errors by removing impressions that include the actual total price paid exceeding 130% of price per night multiplied by the number of nights.[3] And as the original data contain over 20,000 destinations, (4) we follow Ursu (2018) and consider only the four largest destinations in the estimation. Lastly, in order to mitigate the effect of varying number of hotels per impression, (5) we include impressions that have the same number of hotels as the two most frequent lengths of impressions for each destination (e.g. if the two most frequent lengths of hotel lists are 31 and 32 for destination 1, we only consider impressions with 31 or 32 hotels). Table 6.1 presents the summary statistics of the data included in the estimation, and Figure 6.1 shows the number of clicks for each position within an impression.

One noticeable pattern from Table 6.1 is that the purchase rate is very low for all of four destinations. The proportion of search sessions that conclude with a purchase is as low as 2.6%. This pattern suggests that most consumers fail to find hotels more appealing than not making a reservation. Another pattern from Figure 6.1 is that the number of clicks decreases

3. Hotel taxes in the United States range from 7% to 20%. As in Ursu (2018), we set the upper limit on tax to 30% to be more conservative. https://www.consumerreports.org/cro/news/2014/06/booking-a-hotel-these-cities-have-the-highest-hotel-taxes/index.htm.

Destination		1	2	3	4	Destination		1	2	3	4
Hotel Level						*Impression Level*					
Observations		19079	33846	12800	14435	Sessions		611	1043	394	444
Price per night	Mean	2.42	1.41	1.36	1.90	# Click	Mean	1.08	1.12	1.08	1.11
($100)	SD	1.29	0.88	0.97	1.04		SD	0.40	0.45	0.44	0.50
	Median	2.19	1.20	1.12	1.69	# Purchase	Mean	0.03	0.08	0.03	0.06
Stars	Mean	3.43	3.93	3.17	3.15		SD	0.16	0.27	0.17	0.24
	SD	0.87	0.87	0.93	0.90						
	Median	4	4	3	3						
Review	Mean	3.88	4.04	3.90	3.91						
	SD	0.90	0.49	0.89	0.71						
	Median	4	4	4	4						
Location Score	Mean	3.95	4.03	2.75	3.42						
	SD	1.80	0.30	0.97	1.58						
	Median	4.30	4.13	3	3.56						
Brand	Mean	0.67	0.79	0.80	0.71						
	SD	0.47	0.41	0.40	0.45						
Promotion	Mean	0.26	0.60	0.27	0.18						
	SD	0.44	0.49	0.44	0.83						

Table 6.1: Data Summary Statistics

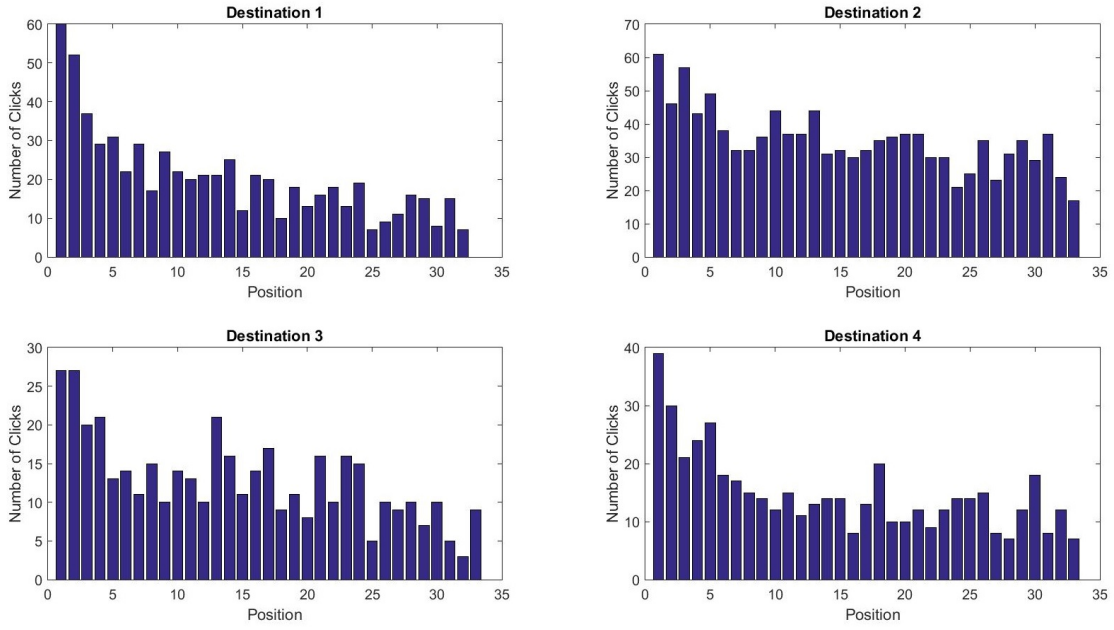


Figure 6.1: Number of Clicks vs. Position

for hotels located in lower positions (closer to the bottom of an impression). So, if hotels are randomly sorted and thus hotel characteristics are also randomly sorted, then there must be another explanation for the decreasing pattern, which Ursu (2018) attributes to the position effect.

Using the identical dataset, Ursu (2018) focuses on the experimental variation of types of impressions presented to consumers and applies the sequential search model to demonstrate two aspects of consumer search behavior. First, she shows that the internal sorting algorithm (developed by Expedia.com) affects consumers via clicking decisions, not via purchase decisions. That is, how hotels are sorted does not affect the purchase decision of consumers, given that a consumer has already gathered information on hotels via search.

Second, Ursu (2018) shows how the position of hotels within an impression affects the click decision. By estimating the sequential search model separately for four major destination cities, she shows that there exist positive and significant position effects, which suggest that consumers are less willing to click hotels at lower positions. Thus, that study represents an important contribution to the search literature by quantifying the position effect on clicks and purchases.

6.2 Model Setup

Following Ursu (2018), consumer i's utility from purchasing product j is specified as:

$$u_{i,j} = X_j\theta + \epsilon_{i,j}$$

where $\epsilon_{i,j}$ is assumed to follow the standard normal distribution, and X_j contains product characteristics available prior to clicking (i.e., searching). As this dataset contains consumers who clicked at least one hotel on the list, the first search is assumed to be free and to reveal the realized utility of the outside option as well as that of the clicked hotel. Most of the assumptions and the model setup used in Ursu (2018) are identical to those laid out in

Section 2.4, except one important difference regarding search costs.

That study assumes that search costs are position-specific and homogeneous. That is, all consumers have the same search cost c_k for hotels located at the k-th position in the list of hotel as given by

$$\log(c_k) = \bar{c} + k\gamma$$

where $\bar{c}$ is the base search cost and γ is the position effect. If γ is estimated to be positive, then consumers are less willing to click the hotels in lower positions. And to construct the simulated likelihood, she uses KSFS with scaled multivariate logistic CDF with scaling factor chosen according to the following steps. 1) She first creates a synthetic dataset that resembles the actual dataset in terms of size and the number and the distribution of characteristics. 2) She estimates the model with several candidate scaling parameters. 3) She chooses the scaling factor that yields estimates close to parameter values used to create the synthetic dataset. 4) She employs the scaling factor (which is equal to 3) to the estimation with the actual dataset.

In order to use the proposed estimation procedure, we change the assumption on search cost so search costs follow a log-normal distribution given by

$$\log(c_{i,k}) \sim N(\bar{c} + k\gamma, \sigma_c^2)$$

Note that the search cost now bears a subscript i for consumer to accommodate the search cost heterogeneity at the position-consumer level. For estimation, we apply the proposed method with three different σ_c values (0.5, 1, and 2) to estimate the sequential search model.

6.3 Results

Table 6.2 summarizes the estimation results from the proposed procedure and compares them to the results from KSFS that are directly imported from Ursu (2018). For each destination, the first three columns show the estimation results under the proposed method

Destination	1				2				3				4			
	$\sigma_c = 0.5$	$\sigma_c = 1$	$\sigma_c = 2$	KSFS	$\sigma_c = 0.5$	$\sigma_c = 1$	$\sigma_c = 2$	KSFS	$\sigma_c = 0.5$	$\sigma_c = 1$	$\sigma_c = 2$	KSFS	$\sigma_c = 0.5$	$\sigma_c = 1$	$\sigma_c = 2$	KSFS
Star	0.0560***	0.1121***	0.2091***	0.0879**	0.1621***	0.3164***	0.5867***	0.3753***	0.0730***	0.1388***	0.2769***	0.0983***	0.1048***	0.2095***	0.4072***	0.1943***
	(0.0110)	(0.0225)	(0.0399)	(0.0284)	(0.0225)	(0.0428)	(0.0720)	(0.0139)	(0.0148)	(0.0298)	(0.0571)	(0.0276)	(0.0200)	(0.0411)	(0.0797)	(0.0235)
Review	0.0069	0.0146	0.0286	-0.0452*	-0.0782***	-0.1564***	-0.3023***	-0.2465***	0.0090	0.0200	0.0433	-0.0486	0.0266*	0.0521*	0.0978*	-0.0808**
	(0.0069)	(0.0135)	(0.0250)	(0.0223)	(0.0204)	(0.0395)	(0.0760)	(0.0336)	(0.0100)	(0.0205)	(0.0410)	(0.0256)	(0.0134)	(0.0257)	(0.0485)	(0.0291)
Location	0.0523***	0.1002***	0.1858***	0.1022***	0.0194	0.0367	0.0753	-0.1477***	0.0287***	0.0562**	0.1114***	-0.0040	0.0363***	0.0703***	0.1315***	0.0438**
Score	(0.0079)	(0.0161)	(0.0281)	(0.0159)	(0.0151)	(0.0293)	(0.0567)	(0.0301)	(0.0085)	(0.0173)	(0.0337)	(0.0254)	(0.0077)	(0.0148)	(0.0285)	(0.0161)
Brand	0.0023	0.0053	0.0108	-0.0097	-0.0191	-0.0335	-0.0558	-0.0089	-0.0509**	-0.0991**	-0.1896**	-0.181**	-0.0087	-0.0154	-0.0308	-0.0276
	(0.0111)	(0.0216)	(0.0411)	(0.0468)	(0.0121)	(0.0235)	(0.0452)	(0.0353)	(0.0158)	(0.0322)	(0.0626)	(0.0562)	(0.0175)	(0.0346)	(0.0652)	(0.0567)
Promotion	0.0255*	0.0504*	0.0945*	0.0645	0.0750***	0.1474***	0.2767***	0.1584***	0.0390**	0.0782**	0.1503**	0.0278	0.0179	0.0444	0.1112	0.0226
	(0.0125)	(0.0245)	(0.0469)	(0.0452)	(0.0133)	(0.0249)	(0.0462)	(0.0312)	(0.0144)	(0.0288)	(0.0569)	(0.0622)	(0.0173)	(0.0347)	(0.0684)	(0.0549)
Price	-0.1041***	-0.2041***	-0.3787***	-0.2312***	-0.1449***	-0.2878***	-0.5252***	-0.2867***	-0.0465**	-0.0872**	-0.1728**	-0.1206**	-0.1030***	-0.2077***	-0.3840***	-0.1857***
	(0.0177)	(0.0359)	(0.0632)	(0.0285)	(0.0222)	(0.0428)	(0.0758)	(0.0291)	(0.0165)	(0.0330)	(0.0619)	(0.0419)	(0.0237)	(0.0459)	(0.0873)	(0.0327)
$\bar{c}$	-2.1261***	-1.7711**	-0.7651	-1.4404***	-1.1359**	-0.5978	0.8500*	-1.0305***	-1.8704**	-1.3076*	0.0196	-1.1467***	-1.2125*	-0.7637	0.5693	-1.0546***
	(0.6455)	(0.5701)	(0.5458)	(0.0028)	(0.3921)	(0.3799)	(0.3721)	(0.0034)	(0.6265)	(0.6158)	(0.5828)	(0.0039)	(0.5018)	(0.4825)	(0.4627)	(0.0046)
Position	0.0112***	0.0220***	0.0441***	0.0185***	0.0043***	0.0087***	0.0176***	0.0044**	0.0073***	0.015***	0.0297***	0.0121***	0.0070***	0.0146***	0.0290***	0.0109***
Effect	(0.0014)	(0.0032)	(0.0064)	(0.0027)	(0.0009)	(0.0018)	(0.0035)	(0.0017)	(0.0014)	(0.0028)	(0.0060)	(0.0027)	(0.0015)	(0.0031)	(0.0064)	(0.0029)
Outside	2.8818***	3.3865***	3.8605***	0.6244***	2.1362***	2.7228***	2.9823***	0.0718	2.9851***	3.2083***	3.8891***	0.2378*	2.6502***	2.9732***	3.7637***	0.4159***
Option	(0.4400)	(0.4640)	(0.4804)	(0.1064)	(0.4743)	(0.5057)	(0.4783)	(0.0473)	(0.5001)	(0.5017)	(0.6074)	(0.0936)	(0.4824)	(0.4838)	(0.6180)	(0.1138)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.2: Estimation Results from Online Hotel Search Data

with various assumed values of σ_c, and the last column shows the estimation result from KSFS with homogeneous search cost assumption. It can be seen that the choice of σ_c does not affect the significance of estimates in most cases. Standard errors are obtained from bootstrap resampling method (Efron and Tibshirani, 1993) with $B = 200$. And supporting the hypotehsis of Ursu (2018), results from both estimation methods suggest that there exists a significant positive position effect. While the estimates from the two methods are not directly comparable given the distinctions alluded to earlier, we do note some (directional) differences in a subset of estimates corresponding to reviews, location, etc.

6.4 Prediction Performance

In this section, we evaluate the merit of using the proposed method as compared to KSFS based on the data replication and out-of-sample prediction performance of the two methods.

6.4.1 Data Replication

First, using the parameter estimates under the two estimation methods, we undertake 100 replications by varying the search cost and the idiosyncratic preference shock to simulate the search sequences and purchase decisions of consumers in the estimation sample. In Figure 6.2, the left column shows the confidence intervals for predicted number of clicks on hotels at each position. In the left column, the dotted lines show the confidence intervals obtained with the estimates from the proposed method with $\sigma_c = 1$[4], and the confidence intervals in dashed lines are obtained with parameter estimates from KSFS. The solid line shows the number of clicks on hotels at each position from the actual data.

The second column shows the actual number of purchases of hotels at each position by the solid line and the mean of the predicted number of purchases from the proposed method's estimates by the dotted line. Lastly, the right column shows the mean predicted number of

4. Data replication results with other assumed values of σ_c are available in the Appendix E.

purchases computed from KSFS estimates in dotted line. Table 6.3 shows some summary statistics of the data replication results. Note that RMSE is root-mean-squared-error of the predicted number of clicks for each position.[5]

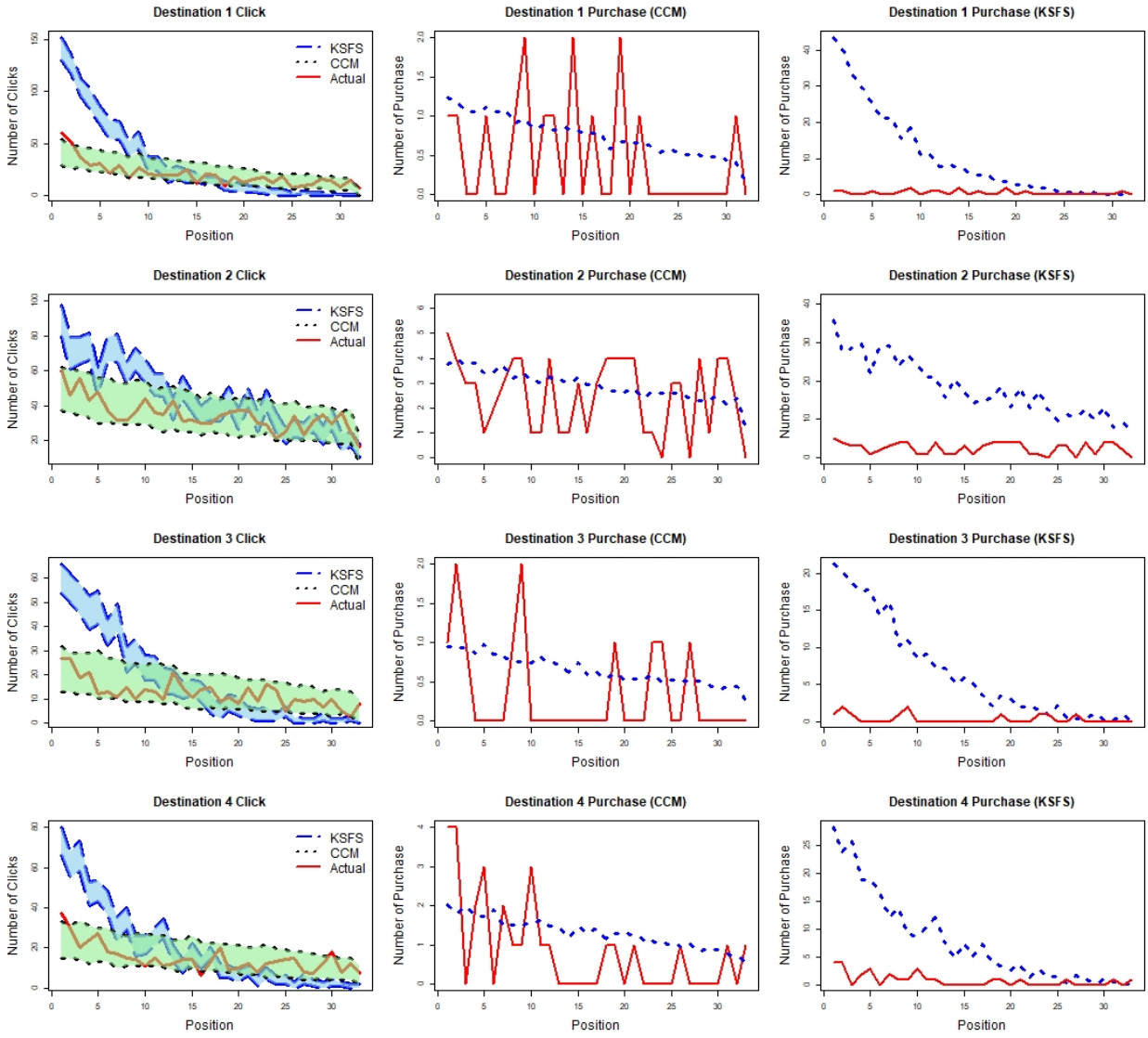


Figure 6.2: Data Replication Results

One can see that the KSFS estimates show higher number of both clicks and purchases than in the actual data. This difference can be, partially, not entirely, attributed to the low outside option mean utility estimate from KSFS. Given the search preference parameters (e.g.

5. $\text{RMSE}_d = \sqrt{\frac{1}{S}\frac{1}{P}\sum_{s=1}^{S}\sum_{p=1}^{P}(\hat{n}_{d,p,s} - n_{d,p})^2}$, where d is index for destination, p for position, and s for replication.

	Destination 1			Destination 2			Destination 3			Destination 4		
	Actual	CCM	KSFS	Actual	CCM	KSFS	Actual	CCM	KSFS	Actual	CCM	KSFS
Total # Click	656	686.06	974.29	1148	1166.20	1459.42	417	446.52	566.26	483	506.97	624.30
Total # Purchase	15	24.07	331.48	83	96.12	598.97	11	21.33	220.28	28	43.29	258.54
Mean Clicked Position	13.38	14.18	10.19	16.75	16.98	18.49	14.93	15.99	11.11	15.38	16.33	11.20
Mean 1st Clicked Position	12.06	12.48	5.48	14.73	15.10	12.90	13.49	13.93	6.88	13.66	14.17	7.45
RMSE	-	7.81	30.21	-	8.03	17.25	-	5.34	15.82	-	6.19	15.94

Table 6.3: Data Replication Results Summary Statistics

coefficients for product characteristics) stay the same, if the mean outside option utility is higher, then consumers will tend to stop searching after searching fewer hotels because it is more likely for them to find that reservation utilities of non-searched options are all lower than the outside option's realized utility, which is revealed after the first search. However, the number of predicted clicks and purchases being larger than the actual data suggests that the outside option mean utility estimate under KSFS is not high enough to replicate the actual data pattern.

Another discrepancy between the actual data and the results from KSFS is that the number of clicks in the upper part of the page is over-predicted, while it is under-predicted in the lower part. The dashed lines in the left column of Figure 6.2 lie above the solid line in the higher positions and below the solid line in the lower positions. This can be seen also in Table 6.3. In the table, the mean of first clicked hotels' position for the destination 1, for example, is around 12 in the actual data, while KSFS predicts the quantity to be around only 5. This discrepancy between KSFS results and the actual data stems from the combination of the positive position effect and the homogeneous search cost assumption. We discuss this in greater detail in Section 6.4.3.

6.4.2 Out-of-Sample Prediction

Besides the data replication performance, in this section, we present the out-of-sample prediction performance of the proposed method and compare it to KSFS. As previously mentioned, the original dataset contains the experimental variation that a consumer is presented with

either a list of hotels sorted randomly or a list of hotels sorted by Expedia's sorting algorithm. Here, we take estimates obtained from random ranking impressions and make predictions on consumers' search sequences and purchase decisions using the hotel characteristics, including the position within an impression, in the Expedia-ranked impressions.

Destination		1	2	3	4			1	2	3	4
Hotel Level						*Impression Level*					
Observations		33801	45816	15849	31063	Sessions		1065	1406	487	946
Price per night	Mean	2.69	1.35	1.33	2.10	# Click	Mean	1.16	1.08	1.14	1.12
($100)	SD	1.17	0.85	0.69	0.99		SD	0.76	0.38	0.97	0.60
	Median	2.56	1.15	1.20	1.89	# Purchase	Mean	0.86	0.89	0.83	0.90
Stars	Mean	3.79	4.03	3.50	3.55		SD	0.35	0.31	0.38	0.30
	SD	0.74	0.80	0.66	0.83						
	Median	4	4	4	4						
Review	Mean	4.07	4.06	4.06	4.12						
	SD	0.77	0.47	0.55	0.55						
	Median	4	4	4	4						
Location Score	Mean	5.11	4.04	2.89	4.51						
	SD	1.24	0.30	0.83	1.14						
	Median	5.69	4.13	2.94	5.12						
Brand	Mean	0.56	0.79	0.65	0.65						
	SD	0.50	0.41	0.48	0.48						
Promotion	Mean	0.42	0.62	0.59	0.35						
	SD	0.49	0.49	0.49	0.48						

Table 6.4: Data Summary Statistics (Expedia Ranking)

Table 6.4 shows the data summary statistics for the Expedia-ranked impressions of hotels. From the table, one can see that hotels in the random ranking data are comparable to those in the Expedia ranking data in terms of hotel characteristics, suggesting that there are no systematic differences in characteristics of hotels in the two rankings. However, there is one distinct difference in the proportion of impressions that conclude with purchase (e.g. conversion rate). In the random ranking, the highest conversion rate among destinations is only about 8% while in the Expedia-ranking the lowest conversion rate is above 80%. According to Google Ads management platform Wordstream, the travel industry's average

conversion rates for the Google Search Network and the Google Display Network are 3.5% and 0.5%, respectively.[6] Here, we do not argue which type of impression has a conversion rate that is closer to the true value, but we merely acknowledge that the Expedia-ranking impressions are not likely to be chosen in the same way as the random ranking impressions are chosen to be included in the dataset. As parameter estimates are obtained from the random ranking, the predicted conversion rate will not be close to the actual conversion rate in the Expedia-ranking, so we present only the out-of-sample prediction results (250 replications) for the number of clicks on hotels at each position in Table 6.5 and Figure 6.3.[7]

	Destination 1			Destination 2			Destination 3			Destination 4		
	Actual	CCM	KSFS	Actual	CCM	KSFS	Actual	CCM	KSFS	Actual	CCM	KSFS
Total # Click	1236	1239.81	1183.38	1512	1593.62	1489.67	555	571.30	514.96	1064	1127.46	976.83
Mean Clicked Position	11.64	13.61	3.95	11.95	14.84	6.04	12.63	15.60	4.47	11.76	15.26	3.76
Mean 1st Clicked Position	9.50	11.46	3.21	10.88	12.99	5.55	10.75	13.15	4.07	9.92	12.63	3.58
RMSE	-	15.75	47.90	-	13.74	43.36	-	7.96	19.44	-	14.58	46.21

Table 6.5: Out-of-Sample Prediction Summary Statistics

Table 6.5 shows summary statistics for the prediction results, and Figure 6.3 shows the 95% confidence interval for the predicted number of clicks on each position in the Expedia-ranked impressions. From the figure and RMSE from the table, one can see that the proposed method has better out-of-sample prediction than KSFS as the confidence intervals obtained with the proposed method (dotted lines) are located around the actual data. On the other hand, KSFS predicts that most of the clicks occur at higher positions, a discrepancy that is also observed in the data replication results. As the proposed method shows out-of-sample prediction that closely follows the actual data pattern, we conclude that the proposed method has better predictive performance than KSFS does, both in terms of data replication and out-of-sample prediction.

6. https://www.wordstream.com/blog/ws/2016/02/29/google-adwords-industry-benchmarks

7. Prediction results with other σ_c values are available in Appendix E.

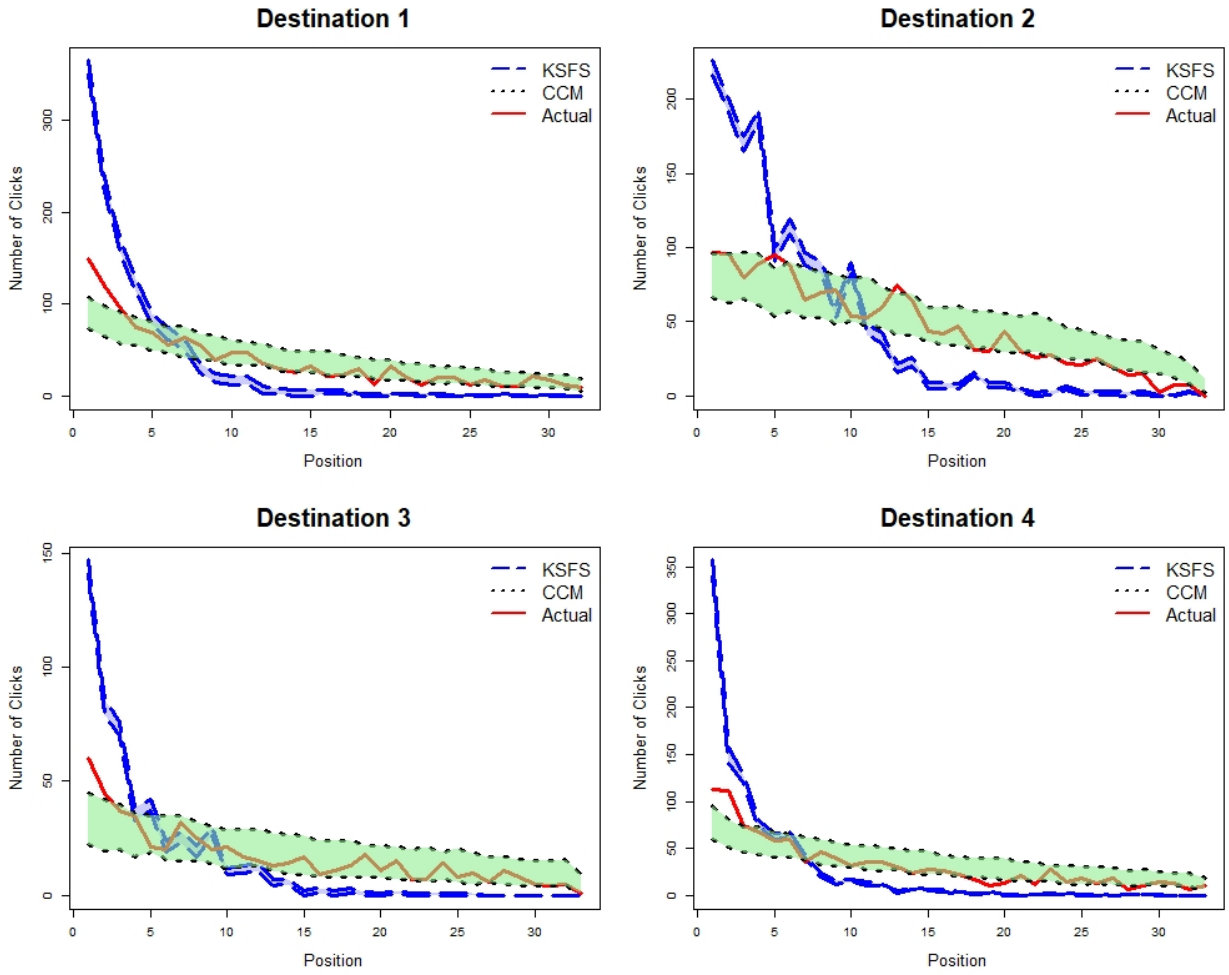


Figure 6.3: Out-of-Sample Prediction of Clicks ($\sigma_c = 1$)

6.4.3 What Is Causing the Discrepancy, the KSFS Method or the Homogeneous Search Cost Assumption?

The setup used in Ursu (2018) to obtain the KSFS estimates differs from the proposed method in two ways. First, the proposed method computes the probability of a consumer's search sequence and purchase decision by using the distributions assumed for each model component (e.g. for the search costs and idiosyncratic preference shocks), while KSFS applies a kernel, potentially arbitrarily chosen, to inequalities obtained from search set conditions. Second, one needs to assume a distribution for search costs, while KSFS does not require

such an assumption. Specifically, Ursu (2018) assumes position-specific homogeneous, or deterministic, search costs. In this section, we present another simulation study to understand which assumption is more responsible for the large discrepancy between the KSFS predictions and the actual data patterns.

The first simulation is to examine the validity of the position-specific homogeneous search cost. Ideally, we should compare the dotted lines of Figure 6.2 against the data replication results obtained with estimates from our proposed method with position-specific *homogeneous* search cost assumption. However, the proposed method requires the search cost to be heterogeneous at the consumer-product level, rendering such a comparison infeasible. Therefore, we take the estimates from the first column ($\sigma_c = 0.5$) for each destination in Table 6.2 and simulate search sequences and purchase decisions with search cost heterogeneity ignored (e.g. $\sigma_c = 0$). For the second simulation, to examine the validity of using KSFS, we estimate the same model using KSFS ($s = 20$) with the heterogeneous search cost assumption ($\sigma_c = 0.5$) and, using the resulting estimates, simulate search sequences and purchase decisions by making random draws of the idiosyncratic preference shocks and search costs.[8]

To determine the main source of the discrepancy, we compare the data fit of these two simulations. That is, if the first simulation shows a better data replication performance than the second simulation does, then we infer that KSFS bears more responsibility for the discrepancy. On the other hand, if the second simulation shows a better performance than the first one does, we then infer that position-specific homogeneous search cost assumption bears more responsibility for the discrepancies between the actual data pattern and the prediction.

In Figure 6.4, the dotted lines show the confidence intervals for the first simulation, under the position-specific homogeneous search cost assumption and with the proposed method estimates, and the dashed lines show the confidence intervals for the second simulation, under the heterogeneous search cost assumption with KSFS estimates. And as before, the

8. Estimation results using KSFS with different scaling factors are available in Table E.1 in Appendix E.

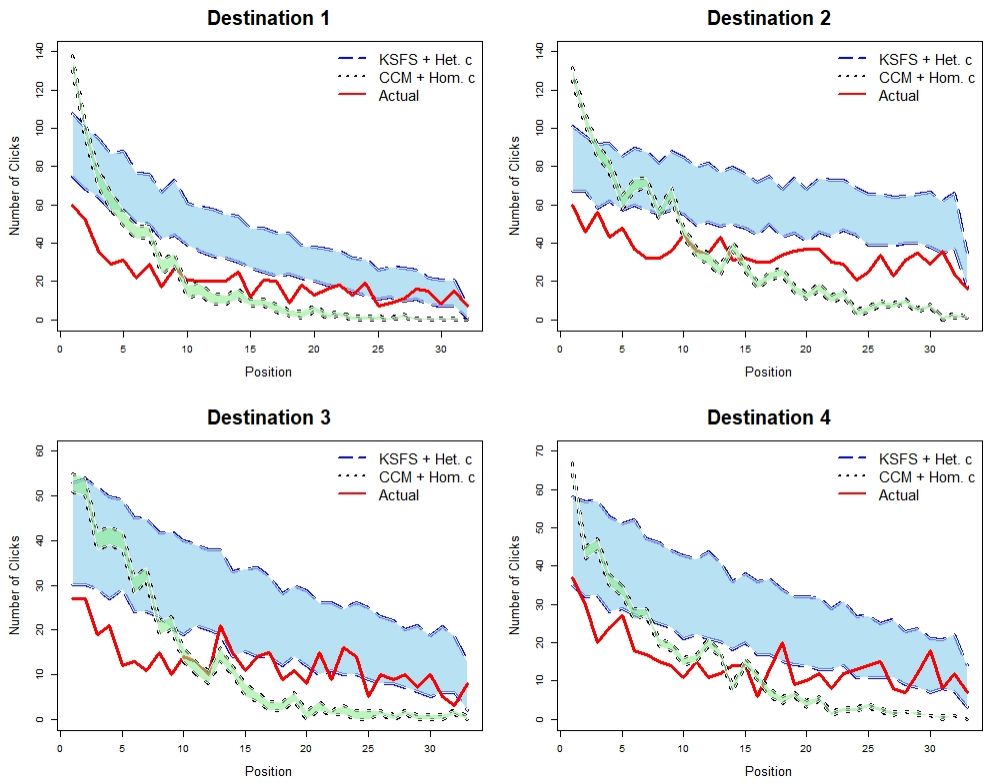


Figure 6.4: Simulated Number of Clicks vs. Position

solid lines show the number of clicks for each position from the actual data.

The dotted lines in Figure 6.4 show far less variability in the predicted number of clicks for hotels at each position, but the number of clicks is over-predicted for the higher positions while it is under-predicted in the lower positions. On the other hand, the dashed lines appear to reflect the fluctuation across positions from the real data, even if the number of clicks are over-predicted across positions compared to the actual data. The fact that both simulation results deviate from the actual data suggests that the inadequate data replication performance in the dashed lines of Figure 6.2 results from a combination of the search cost specification and the estimation method. However, Figure 6.4 suggests that the homogeneous search cost assumption bears more responsibility for the discrepancy, and the KSFS estimates show better prediction performance under the heterogeneous search cost assumption than under the homogeneous search cost assumption.

Now that we have presented various simulation studies, we discuss the implications of the search cost assumption. First, let us assume that the search cost is homogeneous within a consumer (i.e., $c_{i,j} = c_i$). Then, consumer i's reservation utility for product j (see (2.1)) is given by

$$\begin{aligned} z_{i,j} &= \mu_{i,j} + \eta_{i,j} \\ &= \mu_{i,j} + \sigma g^{-1}\left(\frac{c_{i,j}}{\sigma}\right) \\ &= \mu_{i,j} + \sigma g^{-1}\left(\frac{c_i}{\sigma}\right) \\ &= \mu_{i,j} + \eta_i \end{aligned} \tag{6.1}$$

(6.1) shows that if the search cost is assumed to be homogeneous within a consumer, $\eta_{i,j} = \eta_i$, i.e., $\eta_{i,j}$'s will have the same value across alternatives. This implies that a consumer's ranking of reservation utilities is entirely determined by the order of the expected utilities, $\mu_{i,j}$, i.e., he searches product j before product j' if and only if $\mu_{i,j} > \mu_{i,j'}$. It also implies that if a researcher knows a consumer's μ_j, then he can predict the consumer's search order with certainty. Furthermore, if $\mu_{i,j} = \mu_j$ for all i, i.e., if the utility preference parameters are assumed to be homogeneous across consumers, then every consumer will have the same search order. Thus, the search sequences will be identical across consumers, although they could differ in when they terminate their search processes.

On the other hand, if one assumes heterogeneous search costs at the consumer-product level, $c_{i,j} \neq c_{i,j'}$, then $\sigma g^{-1}\left(\frac{c_{i,j}}{\sigma}\right)$ will be heterogeneous across products, and this heterogeneity allows consumers to search a product j before searching j' even if product j has lower expected utility than product j' ($\mu_j < \mu_{j'}$). In this way, the reservation utility order is only partially determined by the ranking of expected utilities, and consumers may have different search sequences. Therefore, the heterogeneous search cost assumption makes it feasible to model heterogeneity in search sequences.

Next, consider the prediction pattern shown in dashed lines in Figure 6.2 and Figure

6.3 and in the dotted lines in Figure 6.4. Assume that consumers have homogeneous search costs and that there is a positive position effect. Then, the reservation utilities of hotels at the k-th position and $(k+1)$-th position in an impression are given respectively by

$$z_k = \mu_k + \eta_k = \mu_k + \sigma g^{-1}\left(\frac{c_k}{\sigma}\right)$$
$$z_{k+1} = \mu_{k+1} + \eta_{k+1} = \mu_{k+1} + \sigma g^{-1}\left(\frac{c_{k+1}}{\sigma}\right)$$

Because there exists a positive position effect (i.e., $\gamma > 0$) and since the search costs are assumed to be homogeneous, it is always true that $c_{k+1} > c_k$, which in turn implies $\eta_{k+1} < \eta_k$ as $g^{-1}(\cdot)$ is a monotonically decreasing function.

Therefore, a consumer will click the hotel in the $(k+1)$-th position before the hotel in the k-th position if and only if the following inequality is satisfied.

$$\mu_{k+1} - \mu_k > \eta_k - \eta_{k+1} \equiv \Delta\eta_{k,k+1}$$

The hotel in the $(k+1)$-th position must have expected utility higher than μ_k at least by $\Delta\eta_{k,k+1}$ for it to be clicked before the hotel in a higher position is clicked. Extending this example to the hotel at the top position ($k = 1$) and the hotel at the 30th position ($k = 30$), the latter must have much higher expected utility than the expected utility of the former, since the position effect accumulates, i.e., $\log(c_{30}) = \log(c_1) + 29\gamma$, and $\Delta\eta_{1,30}$ will be larger too. This leads to the over-prediction of clicks in higher positions and under-prediction in lower positions under the homogeneous search cost assumption and the presence of positive position effect.

On the other hand, the specification of heterogeneous search costs under the proposed method allows for the possibility that, even in the presence of the position effect, lower-positioned hotels can have lower search costs than higher-positioned hotels. As $\Delta\eta_{k,k+1}$ may have a positive or negative value, it becomes possible for a lower-positioned hotel to be clicked before a higher-positioned hotel even if it has a lower expected utility. This flexibility

allows predictions that are comparable to the patterns in the actual data, as shown in the dotted lines in Figure 6.2 and 6.3.

6.4.4 Impact of Sorting Algorithm

In this section, we use the Expedia-ranking impressions and the estimation result from the random ranking impressions in order to examine the impact of the sorting algorithm on consumer welfare and total revenue. First, note that Expedia has an incentive to present hotels in a particular order to consumers to boost sales. We do not have knowledge on how Expedia sorts its hotel partners, so we take the Expedia-ranking impressions from the dataset as the benchmark and examine whether Expedia's sorting algorithm is more beneficial than other sorting schemes, from both Expedia's and consumer's perspective.

To compare with the Expedia-ranking, we apply two additional sorting schemes to Expedia-ranking impressions, namely random ranking and utility-based ranking, which is also explored in Ursu (2018). To create the random ranking, we take hotels in Expedia-ranking impressions and shuffle their orders in each impression. And to create the utility-based ranking, we first compute the expected utilities of hotels in the Expedia-ranking impressions with parameter estimates presented in Table 6.2 with $\sigma_c = 1$ and second, sort hotels by their expected utilities in a descending order so that the hotel with the highest expected utility is located on the top of impression.

With three types of impressions, each sorted (i) randomly, (ii) by Expedia, or (iii) by expected utility, we simulate the search sequence and purchase decisions of consumers and compare predicted revenues and consumer welfare for 200 replications.[9] Note that the idiosyncratic preference shocks and search costs are randomly drawn but fixed across the types of ranking so that the difference across impression types reveals only the impact of the ranking on consumers' search sequences and purchase decisions.

9. For each consumer in the random ranking, we create 200 shuffled impressions, and for each shuffled impression, we make random draws of the search costs and the preference shocks to simulate the consumer's search and purchase decision. This process is replicated 200 times for each shuffled impression.

Destination	1			2			3			4		
# Consumers	$N = 1065$			$N = 1406$			$N = 487$			$N = 946$		
Type	Random	Expedia	U-based	Random	Expedia	U-based	Random	Expedia	U-based	Random	Expedia	U-based
Total	1219	1237.4	1242.6	1590.6	1588.2	1593	567.31	569.56	570.22	1120.4	1122.4	1129.4
Click	(7.07)	(28.29)	(27.00)	(23.88)	(19.83)	(21.58)	(18.19)	(17.89)	(17.79)	(24.45)	(20.28)	(21.05)
Total	48.59	48.46	51.14	134.52	135.08	138.54	30.82	30.74	32.54	111.08	114.42	114.81
Purchase	(7.90)	(6.85)	(6.92)	(11.29)	(11.14)	(10.63)	(5.70)	(5.90)	(5.56)	(9.70)	(10.55)	(9.99)
Total	111.08	111.22	113.57	146.00	145.58	147.02	39.13	39.49	41.14	216.44	222.68	219.95
Revenue	(19.23)	(16.81)	(15.82)	(14.15)	(14.03)	(13.36)	(7.77)	(8.08)	(7.83)	(20.55)	(21.48)	(20.27)
($100)												
Consumer	2.02	2.01	2.03	1.73	1.72	1.74	1.98	1.97	2.00	2.06	2.12	2.09
Welfare	(0.71)	(0.72)	(0.71)	(0.79)	(0.78)	(0.78)	(0.72)	(0.74)	(0.74)	(0.76)	(0.78)	(0.77)

Table 6.6: Impact of Sorting Algorithm

Table 6.6 summarizes the results. For each destination, the first column shows the prediction for the random ranking, the second column Expedia ranking, and the third column utility-based ranking. The first three rows show the mean, across replications, of total predicted clicks, purchases, and revenue (in $100), respectively, made by N consumers. And the last row shows the average consumer welfare for consumers who make purchases after their search sequences. Standard deviations are given in parentheses.

Comparing the random ranking with the Expedia ranking, one can see that Expedia has some success in gaining more purchases from consumers by sorting impressions with its own algorithm. However, destination 1 and 3 show slightly lower number of total purchases under the Expedia ranking than under the random ranking. And across destinations, consumers are predicted to make more purchases under the utility-based ranking than under other sorting algorithms. Also, it is predicted that utility-based ranking enhances consumer welfare for all destinations. It implies that applying utility-based ranking algorithm actually not only enhances consumer welfare but also helps Expedia to increase the number of transactions.

Table 6.7 compares the utility-based ranking against the random ranking and Expedia ranking. The first column (U-R) of each destination shows the difference between the utility-based ranking and the random ranking, and second column (U-E) shows the difference

Destination	1		2		3		4	
	U-R	U-E	U-R	U-E	U-R	U-E	U-R	U-E
Consumer								
Δ Consumer Welfare ($)	6.23	10.54	5.21	8.95	23.08	33.90	14.70	-12.92
Δ Utility ($)	6.56	6.59	3.37	8.39	20.37	23.53	14.30	-10.45
Δ Total Search Cost ($)	0.32	-3.95	-1.84	-0.56	-2.71	-10.38	-0.39	2.47
Δ Transaction Position	-4.02	-3.23	-5.85	-3.36	-3.63	-2.95	-4.59	-2.11
Expedia								
Δ Transaction (%)	5.25	5.52	2.99	2.56	5.57	5.86	3.35	0.34
Δ Total Revenue (%)	5.13	4.85	0.75	1.07	6.54	5.37	3.16	-2.39

Table 6.7: Comparison of Sorting Algorithms

between the utility-based ranking and the Expedia ranking. The first row of Table 6.7 shows the difference in consumer welfare from applying the utility-based ranking expressed in dollar term. The next two rows show the amount of difference from each source, the change in realized utility or the change in total search cost incurred during the search process. The fourth row shows the change in the average position, within an impression, of purchased hotel. And the last two rows show the percentage change in the total number of transactions and total revenue.

Except the comparison between Expedia-ranking and utility-based ranking for destination 4, the utility-based ranking is predicted to enhance the consumer welfare. And tracking down the source of the change in consumer welfare, under the utility-based ranking, consumers tend to find better options, that is, hotels with higher realized utility, while incurring smaller search cost. Except destination 1, on average, about 16% of gain in consumer welfare is attributed to the lower search cost from the utility-based ranking compared to the random ranking. Also, except destination 4, on average, the reduction in search cost under the utility-based ranking contributes to 25% of consumer welfare gain compared to the Expedia ranking. This larger reduction in search cost caused by the switch from Expedia-ranking to utility-based ranking suggests that under the Expedia ranking, consumers tend to make more clicks before they make purchases. One possible explanation for this difference is that

Expedia places hotels with higher commission rate for itself on the upper part of the results page, leading consumers to make clicks on those hotels due to the position effect. However, consumers may find hotels with higher realized utilities after a few more clicks.

The utility-based ranking also benefits Expedia with an increase in the number of transactions. Under the Expedia ranking, hotels in the lower part of impressions could have higher expected utility and thereby a higher chance to have the highest realized utilities. However, because of the positive position effect, consumers might be discouraged from clicking on hotels in the lower part of the page and leave the website without making any purchase. In contrast, under the utility-based ranking, hotels in the upper part of the list have higher expected utilities and thereby a higher chance to have the highest realized utilities. One can infer that the utility-based ranking lowers the chance that consumers do not click on hotels with higher expected utilities merely because they are located in the lower part of results page.

CHAPTER 7

CONCLUSION

In this dissertation, we present a new likelihood-based estimation procedure for the sequential search model. First, we provide a comprehensive, step-by-step procedure for the construction of the simulated likelihood. Also, we discuss issues with other simulators that researchers have previously used in order to circumvent the difficulties in estimating the sequential search model. By demonstrating that the proposed method addresses such issues and yields more precise estimates than other simulators do, we provide researchers with a new way to estimate the sequential search model. We also demonstrate the empirical application of the proposed method and present simulation studies to demonstrate the ability to replicate the data and the out-of-sample prediction performance of the proposed method. Lastly, we demonstrate that sorting products by their expected utilities will benefit both consumers and the search intermediary.

However, the proposed procedure is not free of limitations. One needs to fix the standard deviation of the search cost distribution to some value. Although we explain that the proposed method can be useful even under an arbitrary choice of σ_c and that σ_c serves as the scaling parameter as the σ_ϵ does in the usual probit discrete choice model, one might still be reluctant to fix it to some value. Therefore, the approximation of σ_c or the estimation of its ratio to any of other model component (σ_ϵ or $\bar{c}$) can further enhance the value of the proposed method. Second, the proposed method requires data on individual consumers' search sequences. It is true that such data have become more accessible than ever, but a model that would allow the application of the sequential search model to the dataset on search set membership only (without search order) can be another breakthrough in the literature. Lastly, as there are some cases in which researchers cannot observe the purchase decision following the search sequence of an individual consumer, modification of the proposed method to incorporate such a case appears to be a good avenue for future research.

REFERENCES

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, July 1995.

Hector Chade and Lones Smith. Simultaneous search. *Econometrica*, 74(5):1293–1307, 2006.

Yuxin Chen and Song Yao. Sequential search with refinement: Model and application with click-stream data. *Management Science*, 2016.

Michael Choi, Anovia Yifan Dai, and Kyungmin Kim. Consumer search and price competition. *Econometrica*, 86(4):1257–1281, 2018.

Babur De Los Santos, Ali Hortaçsu, and Matthijs R. Wildenbeest. Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review*, 102(6):2955–80, May 2012.

Jean-Pierre Dubé, Jeremy T. Fox, and Che-Lin Su. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–2267, 2012.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Andrés Elberg, Pedro M. Gardete, Rosario Macera, and Carlos Noton. Dynamic effects of price promotions: field evidence, consumer search, and supply-side implications. *Quantitative Marketing and Economics*, 17(1):1–58, Mar 2019.

John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57:1317–1339, 1989.

John Geweke. Efficient simulation from the multivariate normal and student -t distributions subject to linear constraints. *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, pages 571–578, 1991.

Vassilis Hajivassiliou and Daniel McFadden. Efficient simulation from the multivariate normal and student -t distributions subject to linear constraints. *Econometrica*, 66:863–96, 1998.

Vassilis Hajivassiliou, Daniel McFadden, and Paul Ruud. Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results. *Journal of Econometrics*, 72(1):85 – 134, 1996.

John R. Hauser and Birger Wernerfelt. An evaluation cost model of consideration sets. *Journal of Consumer Research*, 16(4):393–408, 1990.

Han Hong and Matthew Shum. Using price distributions to estimate search costs. *The RAND Journal of Economics*, 37(2):257–275, 2006.

Elisabeth Honka. Quantifying search and switching costs in the us auto insurance industry. *The RAND Journal of Economics*, 45(4):847–884, 2014.

Elisabeth Honka and Pradeep Chintagunta. Simultaneous or sequential? search strategies in the u.s. auto insurance industry. *Marketing Science*, pages 21–42, August 2016.

Joel L Horowitz and Jordan J Louviere. What is the role of consideration sets in choice modeling? *International Journal of Research in Marketing*, 12(1):39 – 54, 1995.

John Howard and Jagdish Sheth. *The Theory of Buyer Behavior*. John Wiley & Sons, 1969.

Michael Keane. *Four Essays in Empirical Macro and Labor Economics*. PhD thesis, Brown University, 1990.

Michael Keane. A computationally practical simulation estimator for panel data. *Econometrica*, 62:95–116, 1994.

Jun B. Kim, Paulo Albuquerque, and Bart J. Bronnenberg. Online demand under limited consumer search. *Marketing Science*, 29(6):1001–1023, 2010.

Jun B. Kim, Paulo Albuquerque, and Bart J. Bronnenberg. The probit choice model under sequential search with an application to online retailing. *Management Science*, 2016.

Sergei Koulayev. Estimating demand in search markets: the case of online hotel bookings. *FRB of Boston Working Paper*, (09-16), 2009.

Steven R. Lerman and Charles F. Manski. *On the Use of Simulated Frequencies to Approximate Choice Probabilities*. The MIT Press, 1981.

Daniel McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026, 1989.

Nitin Mehta, Surendra Rajiv, and Kannan Srinivasan. Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing Science*, 22(1): 58–84, 2003.

José-Luis Moraga-González, Zsolt Sándor, and Matthijs Wildenbeest. Consumer Search and Prices in the Automobile Market. CEPR Discussion Papers 10487, C.E.P.R. Discussion Papers, March 2015.

Peter Morgan and Richard Manning. Optimal search. *Econometrica*, 53(4):923–944, 1985.

Mike Palazzolo and Fred Feinberg. Modeling consideration set substitution.

Amil Petrin and Kenneth Train. A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47(1):3–13, 2010.

Brian T. Ratchford. The value of information for selected appliances. *Journal of Marketing Research*, 17(1):14–25, 1980.

John Roberts. A grounded model of consideration set size and composition. *Advances in Consumer Research*, 16:749–757, 1989.

John H. Roberts and James M. Lattin. Development and testing of a model of consideration set composition. *Journal of Marketing Research*, 28(4):429–440, 1991.

S. Siddarth, Randolph E. Bucklin, and Donal G. Morrison. Making the cut: Modeling and analyzing choice set restriction in scanner panel data. *Journal of Marketing Research*, 32 (3):255–266, Aug. 1995.

George Stigler. The economics of information. *Journal of Political Economy*, 69, 1961.

Che-Lin Su and Kenneth L. Judd. Constrained optimization approaches to estimation of structural models. *Econometrica*, 80(5):2213–2230, 2012.

Raluca Ursu, Qingliang Wang, and Pradeep K. Chintagunta. Search duration. Available at SSRN: https://ssrn.com/abstract=3250051 or http://dx.doi.org/10.2139/ssrn.3250051, August 2018.

Raluca M. Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552, 2018.

Martin Weitzman. Optimal search for the best alternative. *Econometrica*, 47(3):641–54, 1979.

Song Yao and Carl F. Mela. A dynamic model of sponsored search advertising. *Marketing Science*, 30(3):447–468, 2011.

APPENDIX A

CONSTRUCTING THE SIMULATED LIKELIHOOD: DETAILED INSTRUCTIONS

In this appendix, we present the detailed instructions on how to construct the simulated likelihoods for the search sequences and purchase decisions of consumers using the same examples given in Section 3.1.

A.1 Notation

As we begin the construction of simulated likelihood by making a random draw of the realized utility of the purchased product, we write probability statements conditional on the realized utility of the purchased product. For example, in the subsequent section, we show how to simulate the following probability, which represents the first term of Equation 3.5.

$$\Pr\left(u_{i,0} > u_{i,2} \bigcap u_{i,0} > z_{i,2} \bigcap z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}\right)$$

And in subsequent sections, the inequalities are simplified as following for better readability.

$$\begin{aligned}
&\Pr\left(u_{i,0} > u_{i,2} \bigcap u_{i,0} > z_{i,2} \bigcap z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}\right) \\
=&\Pr\left(\underbrace{u_{i,0} > u_{i,2}}_{\text{A}} \bigcap \underbrace{u_{i,0} > z_{i,2}}_{\text{B}} \bigcap \underbrace{z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{C}}\right) \\
=&\Pr(A \cap B \cap C)
\end{aligned}$$

Technically, the probability above can be calculated by evaluating the following integral.

$$\Pr\left(u_{i,0} > u_{i,2} \bigcap u_{i,0} > z_{i,2} \bigcap z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}\right)$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \int_{-\infty}^{u_{i,0}} \int_{-\infty}^{u_{i,0}} \int_{-\infty}^{z_{i,2}} f_2(u_{i,2}) h_2(z_{i,2}) h_m(m_i) dm_i dz_{i,2} du_{i,2} du_{i,0} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{u_{i,0}} f_2(u_{i,2}) du_{i,2} \int_{-\infty}^{u_{i,0}} \int_{-\infty}^{z_{i,2}} h_2(z_2) h_m(m_i) dm_i dz_{i,2} du_{i,0} \\
&= \int_{-\infty}^{\infty} F_2(u_{i,0}) \int_{-\infty}^{u_{i,0}} h_2(z_{i,2}) \int_{-\infty}^{z_{i,2}} h_m(m_i) dm_i dz_{i,2} du_{i,0} \\
&= \int_{-\infty}^{\infty} F_2(u_{i,0}) \int_{-\infty}^{u_{i,0}} h_2(z_{i,2}) H_m(z_{i,2}) dz_{i,2} du_{i,0} \qquad \text{(A.1)}
\end{aligned}
$$

where m_i denotes $\max_{l \notin \{S_i \cup 0\}} z_{i,l}$, and f_2, h_2, and h_m denote the probability density functions of $u_{i,2}$, z_2, and m_i, respectively, and F_2 and H_m denote the cumulative distribution function of $u_{i,2}$ and m_i, respectively.

Next, let us assume that we know the value of the realized utility of the purchased product, $u^*_{i,0}$, and the value of the reservation utility of product 2, $z^*_{i,2}$, such that inequality B is satisfied. Then, the integrand in Equation A.1 can be expressed by

$$
\begin{aligned}
&F_2(u_{i,0}) \int_{-\infty}^{u_{i,0}} h_2(z_{i,2}) H_m(z^*_{i,2}) dz_{i,2} \\
=&F_2(u^*_{i,0}) H_2(u^*_{i,0}) H_m(z^*_{i,2})
\end{aligned}
$$

where H_2 denotes the cdf of $z_{i,2}$. And the probability in Equation A.1 can be numerically approximated by following.

$$
\begin{aligned}
&\Pr\left(\underbrace{u_{i,0} > u_{i,2}}_{\text{A}} \bigcap \underbrace{u_{i,0} > z_{i,2}}_{\text{B}} \bigcap \underbrace{z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{C}} \right) \\
\approx &\frac{1}{n_d} \sum_{d=1}^{n_d} F_2(u^{(d)}_{i,0}) H_2(u^{(d)}_{i,0}) H_m(z^{(d)}_{i,2} | u^{(d)}_{i,0} > z^{(d)}_{i,2}) \qquad \text{(A.2)}
\end{aligned}
$$

One can see that given a random draw of $u_{i,0}$, inequalities A and B are independent as it is assumed that the preference shock (ϵ) and the search cost are independent, while inequalities B and C are correlated since we make a random draw of $z_{i,2}$ that satisfies inequality B and the same random draw is plugged into the cdf of m_i to calculate the probability of inequality

C. Following this structure, we simplify the inequalities and reflect such dependency as following.

$$\Pr\left(\underbrace{u_{i,0} > u_{i,2}}_{\text{A}} \bigcap \underbrace{u_{i,0} > z_{i,2}}_{\text{B}} \bigcap \underbrace{z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{C}}\right)$$

$$= \Pr(A)\Pr(B)\Pr(C|B)$$

In the following instructions, the probabilities of inequalities represented by alphabets are actually conditional on the random draw of the purchased product's realized utility.

A.2 Case 1—$K_i = 1$

The joint probability of the search sequence and the purchase decision can be expressed as the product of conditional probabilities as following.

$$\begin{aligned}\Pr(S_i = \{2\}, D_i = 0) = &\Pr(\underbrace{u_{i,0} > u_{i,2}}_{\text{A}} \bigcap \underbrace{u_{i,0} > z_{i,2}}_{\text{B}} \bigcap \underbrace{z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{C}}) \\ &+ \Pr(\underbrace{u_{i,0} > u_{i,2}}_{\text{A}} \bigcap \underbrace{z_{i,2} > u_{i,0}}_{\text{D}} \bigcap \underbrace{u_{i,0} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{E}}) \\ = &\Pr(A)\Pr(B)\Pr(C|B) + \Pr(A)\Pr(D)\Pr(E) \qquad \text{(A.3)}\end{aligned}$$

The first term of the right hand side of Equation A.3 is approximated by the following steps.

1. Make n_d draws of $\theta_i^{(d)}$ and compute $\mu_{i,j}^{(d)} \quad \forall\, j$ for $d = 1, \cdots, n_d$.
2. For each d, make a random draw of $\epsilon_{i,0}^{(d)}$ to form $u_{i,0}^{(d)}$.
3. Compute the probability of inequality A.

$$\Pr\left(A^{(d)}\right) = \Pr\left(\epsilon_{i,2} < u_{i,0}^{(d)} - \mu_{i,2}^{(d)}\right)$$

4. For each $u_{i,0}^{(d)}$ and $\mu_{i,2}^{(d)}$, compute $\bar{\eta}_{i,2}^{(d)}$, the implied upper bound of $\eta_{i,2}$, and then corresponding $\underline{\gamma}_{i,2}^{(d)}$, the implied lower bound of $\gamma_{i,2}$ using function g.

$$\bar{\eta}_{i,2}^{(d)} = u_{i,0}^{(d)} - \mu_{i,2}^{(d)}, \qquad \underline{\gamma}_{i,2}^{(d)} = g\left(\bar{\eta}_{i,2}^{(d)}\right)$$

Note that the standard deviation of ϵ, σ, is fixed to 1 and that the upper bound of η corresponds to the lower bound of γ as function g is a decreasing function.

5. Compute the probability of inequality B using the distribution assumption on search cost.

$$\Pr\left(B^{(d)}\right) = \Pr\left(c_{i,2} > \underline{\gamma}_{i,2}^{(d)}\right)$$

6. For each $\underline{\gamma}_{i,2}^{(d)}$, make a random draw of $c_{i,2}^{(d)}$ from its distribution left-truncated at $\underline{\gamma}_{i,2}^{(d)}$, and calculate corresponding $\eta_{i,2}^{(d)}$ by using the inverse of function g to form the simulated reservation utility of product 2, $z_{i,2}^{(d)}$, which satisfies inequality B by construction.

7. For each $z_{i,2}^{(d)}$, and $\forall l \notin \{S_i \cup 0\}$, calculate $\bar{\eta}_{i,l}^{(d)}$, the implied upper bound of $\eta_{i,l}$, and then corresponding $\underline{\gamma}_{i,l}^{(d)}$, the implied lower bound for $c_{i,l}$, and calculate the probability of inequality C.

$$\bar{\eta}_{i,l}^{(d)} = z_{i,2}^{(d)} - \mu_{i,l}^{(d)}, \qquad \underline{\gamma}_{i,l}^{(d)} = g\left(\bar{\eta}_{i,l}^{(d)}\right)$$

$$\Pr\left(C^{(d)}|B^{(d)}\right) = \prod_{l \notin \{S_i \cup 0\}} \Pr\left(c_{i,l} > \underline{\gamma}_{i,l}^{(d)}\right)$$

8. Taking product of probabilities and taking average across the random draws yield the approximated probability of the first term.

$$\Pr(u_{i,0} > u_{i,2} \bigcap u_{i,0} > z_{i,2} \bigcap z_{i,2} > \max_{l \notin \{S_i \cup 0\}} z_{i,l})$$
$$\approx \frac{1}{n_d} \sum_d \Pr\left(A^{(d)}\right) \Pr\left(B^{(d)}\right) \Pr\left(C^{(d)} | B^{(d)}\right)$$

The second term is approximated by a similar process.

1. Compute the probability of inequality $D^{(d)}$ as the complementary of inequality $B^{(d)}$.

$$\Pr\left(D^{(d)}\right) = 1 - \Pr\left(B^{(d)}\right)$$

2. Retain the random draws of $u_{i,0}^{(d)}$ and $\mu_{i,l}^{(d)}$ from the first term approximation. For each $u_{i,0}^{(d)}$ and $\mu_{i,l}^{(d)}$, $\forall l \notin \{S_i \cup 0\}$, calculate $\bar{\eta}_{i,l}^{(d)}$, the implied upper bound of $\eta_{i,l}$, and corresponding $\underline{\gamma}_{i,l}^{(d)}$, the implied lower bound for $c_{i,l}$, and calculate the probability of inequality E.

$$\bar{\eta}_{i,l}^{(d)} = u_{i,0}^{(d)} - \mu_{i,l}^{(d)}, \qquad \underline{\gamma}_{i,l}^{(d)} = g\left(\bar{\eta}_{i,l}^{(d)}\right)$$
$$\Pr\left(E^{(d)}\right) = \prod_{l \notin \{S_i \cup 0\}} \Pr\left(c_{i,l} > \underline{\gamma}_{i,l}^{(d)}\right)$$

3. Taking product of probabilities and taking average across the random draws yield the approximated probability of the second term.

$$\Pr(u_{i,0} > u_{i,2} \bigcap z_{i,2} > u_{i,0} \bigcap u_{i,0} > \max_{l \notin \{S_i \cup 0\}} z_{i,l})$$
$$\approx \frac{1}{n_d} \sum_d \Pr\left(A^{(d)}\right) \Pr\left(D^{(d)}\right) \Pr\left(E^{(d)}\right)$$

The sum of two approximated probabilities represents the simulated joint probability of the search sequence and the purchase decision.

$$\Pr(S_i = \{2\}, D_i = 0) \approx \frac{1}{n_d} \sum_d \Pr\left(A^{(d)}\right) \Pr\left(B^{(d)}\right) \Pr\left(C^{(d)} | B^{(d)}\right) + \frac{1}{n_d} \sum_d \Pr\left(A^{(d)}\right) \Pr\left(D^{(d)}\right) \Pr\left(E^{(d)}\right)$$

A.3 Case 2—$1 < K_i < J$

$$\begin{aligned}
\Pr(S_i = \{2,3\}, D_i = 3) = \Pr(& \underbrace{z_{i,3} > u_{i,3}}_{\text{B}} \bigcap \underbrace{u_{i,3} > u_{i,0} \bigcap u_{i,3} > u_{i,2}}_{\text{A}} \\
& \bigcap \underbrace{u_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{C}} \bigcap \underbrace{z_{i,2} > z_{i,3}}_{\text{D}}) \\
+ \Pr(& \underbrace{u_{i,3} > z_{i,3}}_{\text{E}} \bigcap \underbrace{z_{i,3} > u_{i,0} \bigcap z_{i,3} > u_{i,2}}_{\text{F}} \\
& \bigcap \underbrace{z_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l}}_{\text{G}} \bigcap \underbrace{z_{i,2} > z_{i,3}}_{\text{H}}) \\
= & \Pr(A) \Pr(B) \Pr(C) \Pr(D|B) + \Pr(E) \Pr(F|E) \Pr(G|E) \Pr(H|E)
\end{aligned} \tag{A.4}$$

The first term of the right hand side of Equation A.4 is approximated by the following steps.

1. Make n_d draws of $\theta_i^{(d)}$ and compute $\mu_{i,j}^{(d)} \quad \forall\, j$ for $d = 1, \cdots, n_d$.

2. For each d, make a random draw of $\epsilon_{i,3}^{(d)}$ to form $u_{i,3}^{(d)}$.

3. Compute the probability of inequality A.

$$\Pr\left(A^{(d)}\right) = \prod_{j \in \{S_i, 0\}, j \neq D_i} \Pr\left(\epsilon_{i,j} < u_{i,3}^{(d)} - \mu_{i,j}^{(d)}\right)$$

4. For each $u_{i,3}^{(d)}$ and $\mu_{i,3}^{(d)}$, compute $\underline{\eta}_{i,3}^{(d)}$, the implied lower bound of $\eta_{i,3}$, and corre-

sponding $\bar{\gamma}_{i,3}^{(d)}$, the implied upper bound of $\gamma_{i,3}$ using function g.

$$\underline{\eta}_{i,3}^{(d)} = u_{i,3}^{(d)} - \mu_{i,3}^{(d)}, \qquad \bar{\gamma}_{i,3}^{(d)} = g\left(\underline{\eta}_{i,3}^{(d)}\right)$$

5. Compute the probability of inequality B.

$$\Pr\left(B^{(d)}\right) = \Pr\left(c_{i,3} < \bar{\gamma}_{i,3}^{(d)}\right)$$

6. For each $u_{i,3}^{(d)}$, $\mu_{i,3}^{(d)}$, and $\forall l \notin \{S_i \cup 0\}$, calculate $\bar{\eta}_{i,l}^{(d)}$, the implied upper bound for $\eta_{i,l}$, and corresponding $\underline{\gamma}_{i,l}^{(d)}$, the implied lower bound for $c_{i,l}$, and calculate probability of inequality C.

$$\bar{\eta}_{i,l}^{(d)} = u_{i,3}^{(d)} - \mu_{i,l}^{(d)}, \qquad \underline{\gamma}_{i,l}^{(d)} = g\left(\bar{\eta}_{i,l}^{(d)}\right)$$

$$\Pr\left(C^{(d)}\right) = \prod_{l \notin \{S_i \cup 0\}} \Pr\left(c_{i,l} > \underline{\gamma}_{i,l}^{(d)}\right)$$

7. For each $\bar{\gamma}_{i,3}^{(d)}$, make a random draw of $c_{i,3}^{(d)}$ from its distribution right-truncated at $\bar{\gamma}_{i,3}^{(d)}$, computed in Step 4, and calculate corresponding $\eta_{i,3}^{(d)}$ by using the inverse of function g to form the simulated reservation utility of product 3, $z_{i,3}^{(d)}$, which satisfies inequality B by construction.

$$\eta_{i,3}^{(d)} = g^{-1}\left(c_{i,3}^{(d)}\right) \qquad z_{i,3}^{(d)} = \mu_{i,3}^{(d)} + \eta_{i,3}^{(d)}$$

8. For each random draw of $z_{i,3}^{(d)}$, compute $\underline{\eta}_{i,2}^{(d)}$, the implied lower bound for $\eta_{i,2}$, and corresponding $\bar{\gamma}_{i,2}^{(d)}$, the implied upper bound of search cost of product 2. And compute the probability of inequality D using the distribution assumption on search cost.

$$\underline{\eta}_{i,2}^{(d)} = z_{i,3}^{(d)} - \mu_{i,2}^{(d)}, \qquad \bar{\gamma}_{i,2}^{(d)} = g\left(\underline{\eta}_{i,2}^{(d)}\right)$$

$$\Pr\left(D^{(d)} | B^{(d)}\right) = \Pr\left(c_{i,2} < \bar{\gamma}_{i,2}^{(d)}\right)$$

The construction of simulated probability of the order of reservation utility begins with making random draws of last-searched product's search cost that satisfy the inequality between the purchased product's realized utility and the last-searched product's reservation utility. And it proceeds to the calculation of implied upper bound for second-to-last searched product's search cost and the calculation of the probability that the second-to-last searched product's search cost falls below the upper bound. Note that if there are more than 2 products in a search sequence, one should repeat this step until one obtains the upper bound of the first searched product's search cost and the corresponding probability. Such case will be presented in the next case.

9. The first probability of Equation A.4 is approximated by the following.

$$\begin{aligned}\Pr(z_{i,3} > u_{i,3} \bigcap u_{i,3} > u_{i,0} \bigcap u_{i,3} > u_{i,2} \bigcap u_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l} \bigcap z_{i,2} > z_{i,3}) \\ \approx \frac{1}{n_d} \sum_d \Pr\left(A^{(d)}\right) \Pr\left(B^{(d)}\right) \Pr\left(C^{(d)}\right) \Pr\left(D^{(d)} | B^{(d)}\right) \end{aligned} \tag{A.5}$$

If a consumer's purchased product does not coincide with his last-searched product, then Step 9 concludes the approximation of the joint probability of the consumer's search sequence and purchase decision.

The approximation of the second probability should use the same random draws of the purchased product's realized utility obtained in Step 1 and 2 of the approximation of the first probability.

1. Retain the random draws of $u_{i,3}^{(d)}$ and $\mu_{i,3}^{(d)}$. Compute $\bar{\eta}_{i,3}^{(d)}$ and corresponding $\underline{\gamma}_{i,3}^{(d)}$, the implied lower bound for search cost of product 3, and calculate the probability of inequality E.

$$\bar{\eta}_{i,3}^{(d)} = u_{i,3}^{(d)} - \mu_{i,3}^{(d)}, \qquad \underline{\gamma}_{i,3}^{(d)} = g\left(\bar{\eta}_{i,3}^{(d)}\right)$$

$$\Pr\left(E^{(d)}\right) = \Pr\left(c_{i,3} > \underline{\gamma}_{i,3}^{(d)}\right)$$

2. For each d, make a random draw of $c_{i,3}^{(d)}$ from its distribution left-truncated at $\underline{\gamma}_{i,3}^{(d)}$ and construct corresponding reservation utility of product 3, $z_{i,3}^{(d)}$.

$$\eta_{i,3}^{(d)} = g^{-1}\left(c_{i,3}^{(d)}\right) \qquad z_{i,3}^{(d)} = \mu_{i,3}^{(d)} + \eta_{i,3}^{(d)}$$

3. Probability of inequality F, for each $z_{i,3}^{(d)}$, can be computed by the following.

$$\Pr\left(F^{(d)}|E^{(d)}\right) = \prod_{j\in\{S_i\cup 0\}, j\neq D_i} \Pr\left(\epsilon_{i,j} < z_{i,3}^{(d)} - \mu_{i,j}^{(d)}\right)$$

4. For each $z_{i,3}^{(d)}$, $\mu_{i,l}^{(d)}$, and $\forall l \notin \{S_i \cup 0\}$, calculate $\bar{\eta}_{i,l}^{(d)}$ and $\underline{\gamma}_{i,l}^{(d)}$, the implied lower bound for non-searched product's search cost, and compute the probability of inequality G.

$$\bar{\eta}_{i,l}^{(d)} = z_{i,3}^{(d)} - \mu_{i,l}^{(d)}, \qquad \underline{\gamma}_{i,l}^{(d)} = g\left(\bar{\eta}_{i,l}^{(d)}\right)$$

$$\Pr\left(G^{(d)}|E^{(d)}\right) = \prod_{l\notin\{S_i\cup 0\}} \Pr\left(c_{i,l} > \underline{\gamma}_{i,l}^{(d)}\right)$$

5. For each $z_{i,3}^{(d)}$, compute $\underline{\eta}_{i,2}^{(d)}$, the implied lower bound for $\eta_{i,2}$, and corresponding $\bar{\gamma}_{i,2}^{(d)}$, the implied upper bound of search cost of product 2. And compute the probability of inequality H using the distribution assumption on search cost.

$$\underline{\eta}_{i,2}^{(d)} = z_{i,3}^{(d)} - \mu_{i,2}^{(d)}, \qquad \bar{\gamma}_{i,2}^{(d)} = g\left(\underline{\eta}_{i,2}^{(d)}\right)$$

$$\Pr\left(H^{(d)}|E^{(d)}\right) = \Pr\left(c_{i,2} < \bar{\gamma}_{i,2}^{(d)}\right)$$

6. The second probability of Equation A.4 can be approximated by the following.

$$\begin{aligned}\Pr(u_{i,3} > & z_{i,3} \bigcap z_{i,3} > u_{i,0} \bigcap z_{i,3} > u_{i,2} \bigcap z_{i,3} > \max_{l \notin \{S_i \cup 0\}} z_{i,l} \bigcap z_{i,2} > z_{i,3}) \\ \approx & \frac{1}{n_d} \sum_d \Pr\left(E^{(d)}\right) \Pr\left(F^{(d)}|E^{(d)}\right) \Pr\left(G^{(d)}|E^{(d)}\right) \Pr\left(H^{(d)}|E^{(d)}\right) \quad \text{(A.6)}\end{aligned}$$

The sum of two approximated probabilities represents the simulated joint probability of the search sequence and the purchase decision.

$$\begin{aligned}&\Pr(S_i = \{2,3\}, D_i = 3) \\ &\quad \approx \frac{1}{n_d} \sum_d \Pr\left(A^{(d)}\right) \Pr\left(B^{(d)}\right) \Pr\left(C^{(d)}\right) \Pr\left(D^{(d)}|B^{(d)}\right) \\ &\quad + \frac{1}{n_d} \sum_d \Pr\left(E^{(d)}\right) \Pr\left(F^{(d)}|E^{(d)}\right) \Pr\left(G^{(d)}|E^{(d)}\right) \Pr\left(H^{(d)}|E^{(d)}\right) \quad \text{(A.7)}\end{aligned}$$

A.4 Case 3—$K_i = J$

Let us assume that there are only three products on the market and a consumer searched through all of them, 2, 3, and then 1, and purchased product 1.

$$\begin{aligned}\Pr(S_i = \{2,3,1\}, D_i = 1) = \Pr(& \underbrace{z_{i,1} > u_{i,1}}_{\text{A}} \bigcap \underbrace{z_{i,2} > z_{i,3} > z_{i,1}}_{\text{C}} \\ & \bigcap \underbrace{u_{i,1} > u_{i,2} \bigcap u_{i,1} > u_{i,3} \bigcap u_{i,1} > u_{i,0}}_{\text{B}}) \\ + \Pr(& \underbrace{u_{i,1} > z_{i,1}}_{\text{D}} \bigcap \underbrace{z_{i,2} > z_{i,3} > z_{i,1}}_{\text{F}} \\ & \underbrace{z_{i,1} > u_{i,2} \bigcap z_{i,1} > u_{i,3} \bigcap z_{i,1} > u_{i,0}}_{\text{E}}) \\ = & \Pr(A) \Pr(B) \Pr(C|A) + \Pr(D) \Pr(E|D) \Pr(F|D) \quad \text{(A.8)}\end{aligned}$$

The first probability in Equation A.8 is approximated by the following process.

1. Make n_d draws of $\theta_i^{(d)}$ and compute $\mu_{i,j}^{(d)} \quad \forall\, j$ for $d = 1, \cdots, n_d$.

2. For each d, make a random draw of $\epsilon_{i,1}^{(d)}$ to form $u_{i,1}^{(d)}$.

3. For each $u_{i,1}^{(d)}$ and $\mu_{i,1}^{(d)}$, compute $\underline{\eta}_{i,1}^{(d)}$ and corresponding $\bar{\gamma}_{i,1}^{(d)}$, the implied upper bound for search cost of product 1, and compute the probability of inequality A.

$$\underline{\eta}_{i,1}^{(d)} = u_{i,1}^{(d)} - \mu_{i,1}^{(d)}, \qquad \bar{\gamma}_{i,1}^{(d)} = g\left(\underline{\eta}_{i,1}^{(d)}\right)$$

$$\Pr\left(A^{(d)}\right) = \Pr\left(c_{i,1} < \bar{\gamma}_{i,1}^{(d)}\right)$$

4. Compute probability of inequality B.

$$\Pr\left(B^{(d)}\right) = \prod_{j \neq 1} \Pr\left(\epsilon_{i,j} < u_{i,1}^{(d)} - \mu_{i,j}^{(d)}\right)$$

5. For each $\bar{\gamma}_{i,1}^{(d)}$ obtained in Step 3, make a random draw of $c_{i,1}^{(d)}$ from its distribution right-truncated at $\bar{\gamma}_{i,1}^{(d)}$. And compute the corresponding reservation utility of product 1 using the inverse of function g.

$$\eta_{i,1}^{(d)} = g^{-1}\left(c_{i,1}^{(d)}\right), \qquad z_{i,1}^{(d)} = \mu_{i,1}^{(d)} + \eta_{i,1}^{(d)}$$

6. Compute $\underline{\eta}_{i,3}^{(d)}$ and corresponding $\bar{\gamma}_{i,3}^{(d)}$, the implied upper bound of product 3's search cost. And calculate the probability of inequality $z_{i,3} > z_{i,1}$ (denoted by inequality $C_{(1)}$).

$$\underline{\eta}_{i,3}^{(d)} = z_{i,1}^{(d)} - \mu_{i,3}^{(d)} \qquad \bar{\gamma}_{i,3}^{(d)} = g\left(\underline{\eta}_{i,3}^{(d)}\right)$$

$$\Pr\left(C_{(1)}^{(d)} | A^{(d)}\right) = \Pr\left(c_{i,3} < \bar{\gamma}_{i,3}^{(d)}\right)$$

7. For each $\bar{\gamma}_{i,3}^{(d)}$, make a random draw of $c_{i,3}^{(d)}$ from its distribution right-truncated at $\bar{\gamma}_{i,3}^{(d)}$. Use the inverse of function g to compute the corresponding reservation utility of product 3.

$$\eta_{i,3}^{(d)} = g^{-1}\left(c_{i,3}^{(d)}\right), \qquad z_{i,3}^{(d)} = \mu_{i,3}^{(d)} + \eta_{i,3}^{(d)}$$

8. Compute $\underline{\eta}_{i,2}^{(d)}$ and corresponding $\bar{\gamma}_{i,2}^{(d)}$, the implied upper bound of product 2's search cost. And calculate the probability of inequality $z_{i,2} > z_{i,3}$ (denoted by inequality $C_{(2)}$).

$$\underline{\eta}_{i,2}^{(d)} = z_{i,3}^{(d)} - \mu_{i,2}^{(d)}, \qquad \bar{\gamma}_{i,2}^{(d)} = g\left(\underline{\eta}_{i,2}^{(d)}\right)$$

$$\Pr\left(C_{(2)}^{(d)} | A^{(d)}, C_{(1)}^{(d)}\right) = \Pr\left(c_{i,2} < \bar{\gamma}_{i,2}^{(d)}\right)$$

9. The first term in Equation A.8 can be approximated by the following.

$$\begin{aligned}\Pr(z_{i,1} > u_{i,1} \bigcap z_{i,2} > z_{i,3} > z_{i,1} \bigcap u_{i,1} > u_{i,2} \bigcap u_{i,1} > u_{i,3} \bigcap u_{i,1} > u_{i,0}) \\ \approx \frac{1}{n_d} \sum_d \Pr\left(A^{(d)}\right) \Pr\left(B^{(d)}\right) \Pr\left(C_{(1)}^{(d)} | A^{(d)}\right) \Pr\left(C_{(2)}^{(d)} | A^{(d)}, C_{(1)}^{(d)}\right)\end{aligned} \tag{A.9}$$

If a consumer searches all products on the market and purchases a product other than the last-searched product, the process above concludes the approximation of the joint probability of search and purchase decision.

The second probability is approximated by the following process.

1. Compute the probability of inequality D by

$$\Pr\left(D^{(d)}\right) = 1 - \Pr\left(A^{(d)}\right)$$

2. Retain the random draws of $\theta_i^{(d)}$ and $\epsilon_{i,1}^{(d)}$ from the first probability approximation. Compute $\bar{\eta}_{i,1}^{(d)}$ and corresponding $\underline{\gamma}_{i,1}^{(d)}$, the implied lower bound for product 1's search

cost. For each $\underline{\gamma}_{i,1}^{(d)}$, make a random draw of $c_{i,1}^{(d)}$ from its distribution left-truncated at $\underline{\gamma}_{i,1}^{(d)}$. And construct product 1's reservation utility, $z_{i,1}^{(d)}$.

$$\bar{\eta}_{i,1}^{(d)} = u_{i,1}^{(d)} - \mu_{i,1}^{(d)}, \qquad \underline{\gamma}_{i,1}^{(d)} = g\left(\bar{\eta}_{i,1}^{(d)}\right)$$

$$\eta_{i,1}^{(d)} = g^{-1}\left(c_{i,1}^{(d)}\right), \qquad z_{i,1}^{(d)} = \mu_{i,1}^{(d)} + \eta_{i,1}^{(d)}$$

3. Compute the probability of inequality E.

$$\Pr\left(E^{(d)}|D^{(d)}\right) = \prod_{j \neq D_i} \Pr\left(\epsilon_{i,j} < z_{i,1}^{(d)} - \mu_{i,j}^{(d)}\right)$$

4. Retaining $z_{i,1}^{(d)}$ from Step 2 and repeat Step 6-8 to obtain approximated probability $\Pr\left(F_{(1)}^{(d)}|D^{(d)}\right)$ and $\Pr\left(F_{(2)}^{(d)}|D^{(d)}, F_{(1)}^{(d)}\right)$.

5. The second term in Equation A.8 can be approximated by the following.

$$\begin{aligned}\Pr(u_{i,1} > z_{i,1} \bigcap z_{i,2} > z_{i,3} > z_{i,1} \bigcap z_{i,1} > u_{i,2} \bigcap z_{i,1} > u_{i,3} \bigcap z_{i,1} > u_{i,0}) \\ \approx \frac{1}{n_d} \sum_d \Pr\left(D^{(d)}\right) \Pr\left(E^{(d)}|D^{(d)}\right) \Pr\left(F_{(1)}^{(d)}|D^{(d)}\right) \Pr\left(F_{(2)}^{(d)}|D^{(d)}, F_{(1)}^{(d)}\right)\end{aligned} \tag{A.10}$$

The sum of two approximated probabilities represents the simulated joint probability of the search sequence and the purchase decision.

$$\begin{aligned}&\Pr(S_i = \{2,3,1\}, D_i = 1) \\ &\approx \frac{1}{n_d} \sum_d \Pr\left(A^{(d)}\right) \Pr\left(B^{(d)}\right) \Pr\left(C_{(1)}^{(d)}|A^{(d)}\right) \Pr\left(C_{(2)}^{(d)}|A^{(d)}, C_{(1)}^{(d)}\right) &\text{(A.11)} \\ &+ \frac{1}{n_d} \sum_d \Pr\left(D^{(d)}\right) \Pr\left(E^{(d)}|D^{(d)}\right) \Pr\left(F_{(1)}^{(d)}|D^{(d)}\right) \Pr\left(F_{(2)}^{(d)}|D^{(d)}, F_{(1)}^{(d)}\right) &\text{(A.12)}\end{aligned}$$

APPENDIX B

KERNEL-SMOOTHED FREQUENCY SIMULATOR LIKELIHOOD

The construction of likelihood using KSFS closely follows Honka and Chintagunta (2016).

1. For consumer i, given current estimates, make random draws of $\theta_i^{(d)}$ and $c_{i,j}^{(d)}$ for all product j.

2. Make random draws of $\epsilon_{i,j}^{(d)}$ for all j and i without truncation and construct the random draws of realized utility, $u_{i,j}^{(d)}$.

3. Using inversion from Kim et al. (2010), calculate the reservation utility $z_{i,j}^{(d)}$ for all product j.

4. Compute $w_{i,1}^{(k,d)}$ for $k = 1, \cdots K_i$.

$$w_{i,1}^{(k,d)} = z_{i,r_k}^{(d)} - \max_{l \notin \{r_1, \cdots, r_k\}} z_{i,l}^{(d)}$$

5. Using the simulated utility of purchased product from Step 2, $u_{i,D_i}^{(d)}$, compute $w_{i,2}^{(d)}$.

$$w_{i,2}^{(d)} = u_{i,D_i}^{(d)} - \max_{l \notin \{S_i \cup 0\}} z_{i,l}^{(d)}$$

6. Using the simulated utilities of searched products, including outside option, compute $w_{i,3}^{(d)}$.

$$w_{i,3}^{(d)} = u_{i,D_i}^{(d)} - \max_{l \in \{S_i \cup 0\}, l \neq D_i} u_{i,l}^{(d)}$$

7. For $K_i > 1$, compute $w_{i,4}^{(q,d)}$ for $q = 2, \cdots, K_i$.

$$w_{i,4}^{(q,d)} = z_{i,r_q}^{(d)} - \max_{j \in \{r_1, \cdots, r_{q-1}\}} u_{i,j}^{(d)}$$

8. Construct the simulated joint probability of consumer i's search sequence and purchase decision.

$$P_i^{(d)} = \frac{1}{1 + \sum_k \exp(-sw_{i,1}^{(k,d)}) + \exp(-sw_{i,2}^{(d)}) + \exp(-sw_{i,3}^{(d)}) + \sum_q \exp(-sw_{i,4}^{(q,d)})}$$

$$P_i \approx \frac{1}{n_d} \sum_d P_i^{(d)}$$

APPENDIX C

CRUDE FREQUENCY SIMULATOR LIKELIHOOD

The construction of CFS likelihood closely follows Chen and Yao (2016).

1. For consumer i, given current estimates, make random draws of θ_i and $c_{i,j}$ for all product j.

2. Using inversion from Kim et al. (2010), calculate the reservation utility $z_{i,j}$ for all product j.

3. For $\forall j \in \{S_i \cup 0\}$, make random draws of $\epsilon_{i,j}$ from truncated distribution.
   For example, if $S_i = \{2, 3\}$ and $D_i = 2$,

$$u_{i,0} < z_{i,3}, \qquad \max_{l \notin S_i} z_{i,l} < u_{i,2} < z_{i,3}, \qquad -\infty < u_{i,3} < \infty$$

4. Across the random draws, calculate the proportion that the purchased product has the highest realized utility among searched options. (e.g. $u_{i,2} = \max_{j \in \{S_i, 0\}} u_{i,j}$)

5. For $k = 1, \cdots, K_i$, calculate the proportion that the max of non-searched products' reservation utility is lower than z_{i,r_k}.

$$z_{i,2} > \max_{l \neq 2} z_{i,l} \quad \text{for } k = 1, \qquad z_{i,3} > \max_{l \notin \{2,3\}} z_{i,l} \quad \text{for } k = 2$$

6. Take the product of proportions obtained from Step 4 and Step 5 to compute the simulated joint probability of search sequence and purchase decision.

APPENDIX D

SUPPLEMENTS TO MONTE CARLO SIMULATION

D.1 Simulation Estimation of Proposed Method

Figure D.1: CCM Estimation Results with Correct σ_c

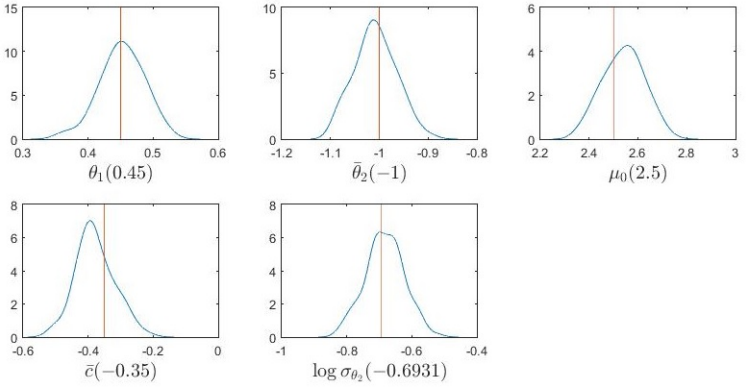


Figure D.2: CCM Estimation Results with $\sigma_c = 0.5$

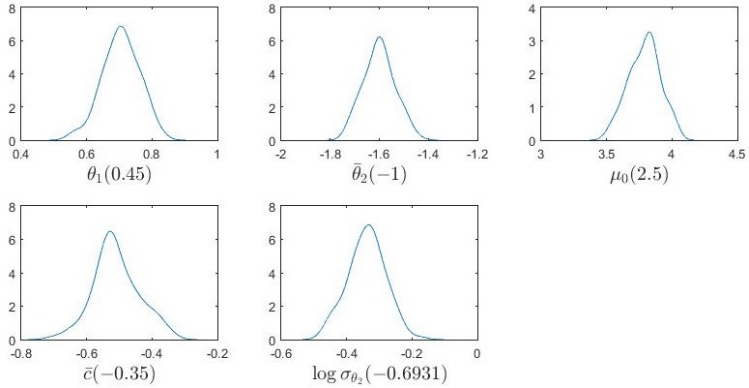
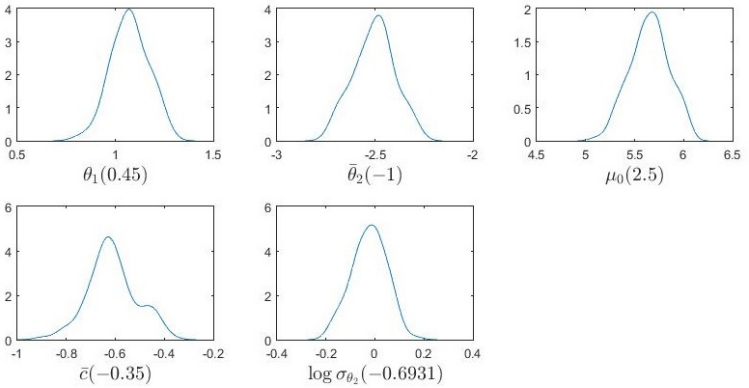


Figure D.3: CCM Estimation Results with $\sigma_c = 1$



D.2 Simulation Estimation of KSFS

Figure D.4: KSFS Estimation Results with $s = 1$

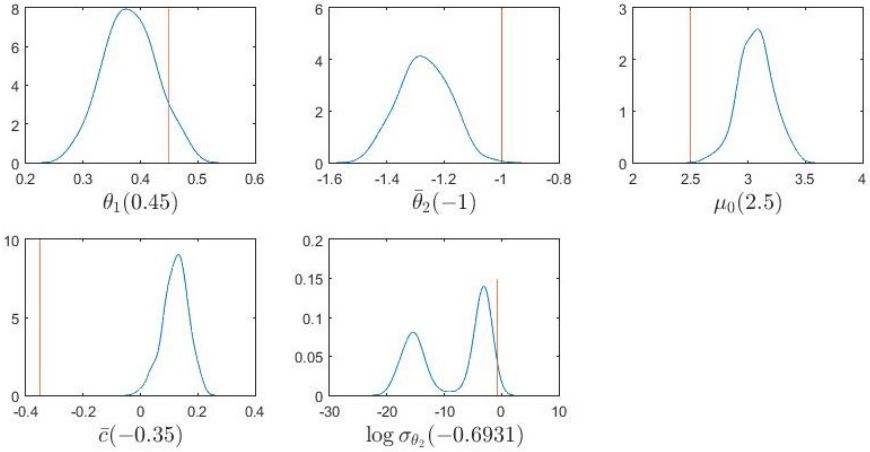


Figure D.5: KSFS Estimation Results with $s = 5$

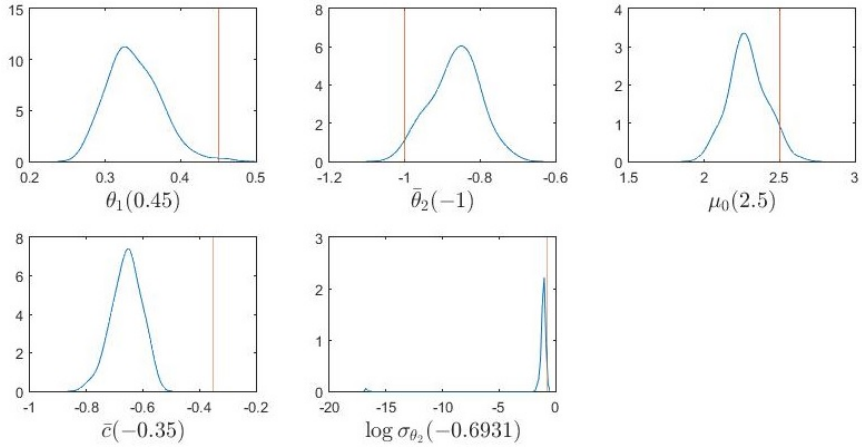


Figure D.6: KSFS Estimation Results with $s = 10$

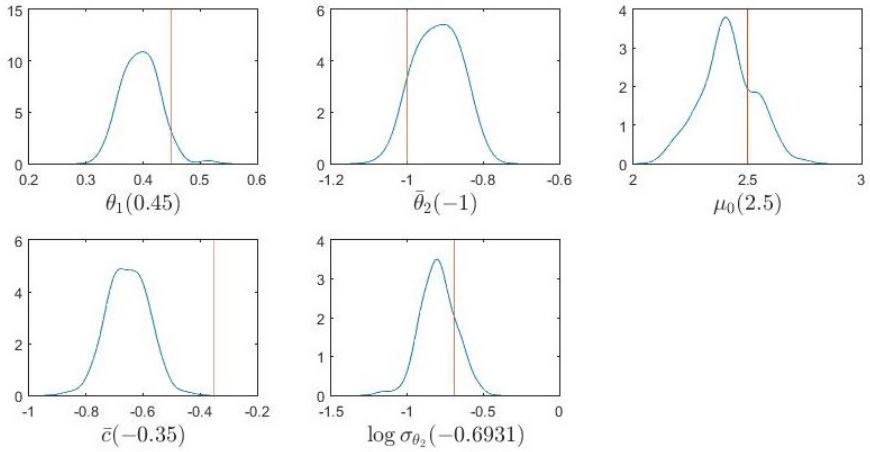


Figure D.7: KSFS Estimation Results with $s = 50$

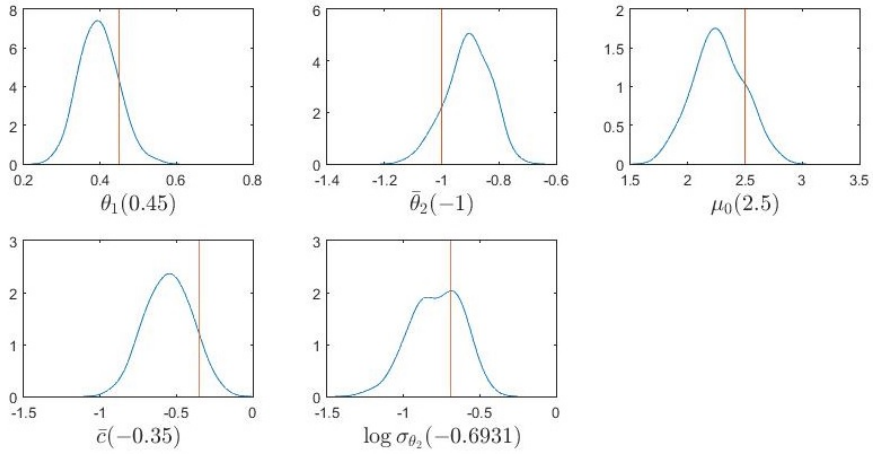


Figure D.8: KSFS Estimation Results with $s = 100$

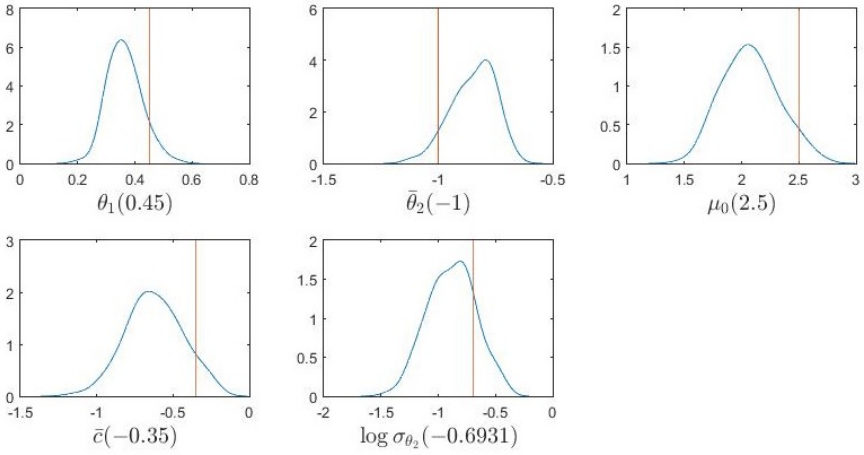
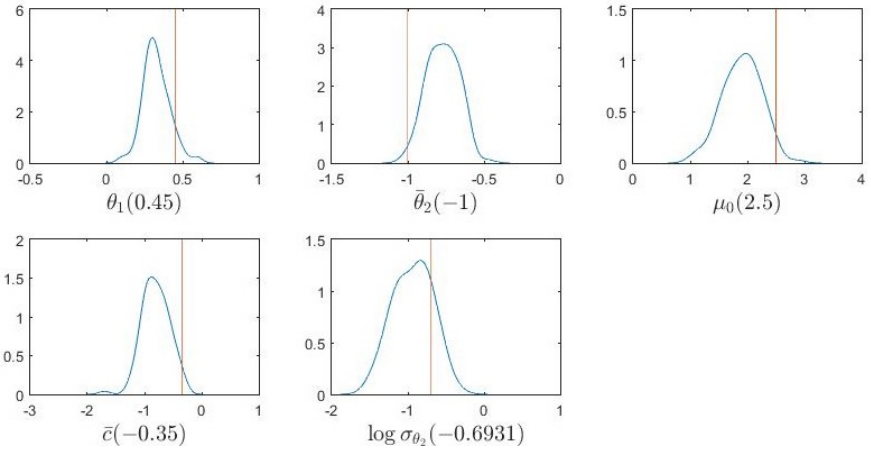


Figure D.9: KSFS Estimation Results with $s = 250$



D.3 Simulation Estimation of CFS

Figure D.10: CFS Estimation Results with $n_d = 25$

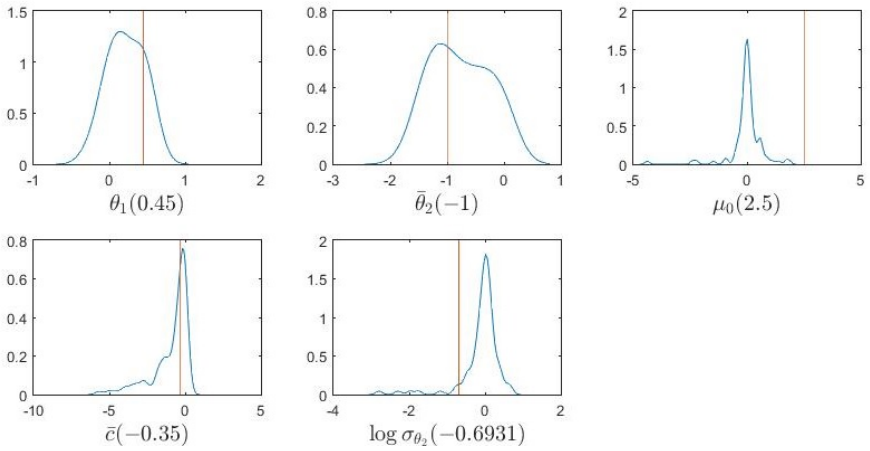


Figure D.11: CFS Estimation Results with $n_d = 50$

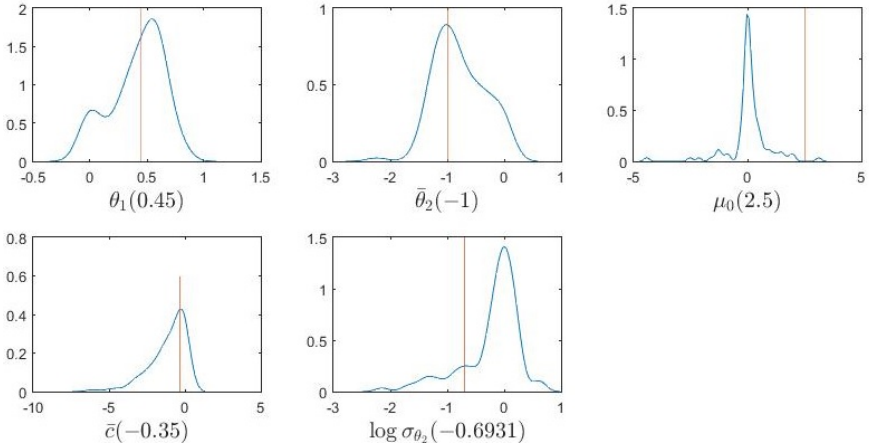
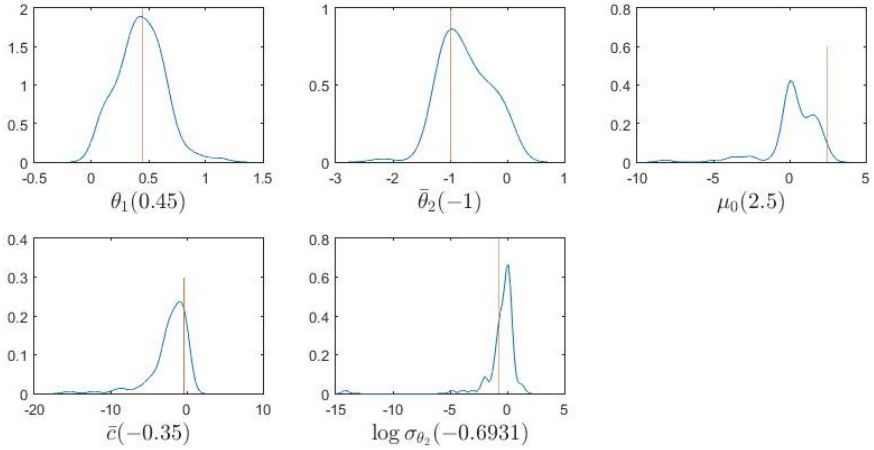


Figure D.12: CFS Estimation Results with $n_d = 100$



APPENDIX E

SUPPLEMENTS TO EMPIRICAL APPLICATION

E.1 Data Replication

Figure E.1: Simulated Number of Clicks vs. Position with $\sigma_c = 0.5$

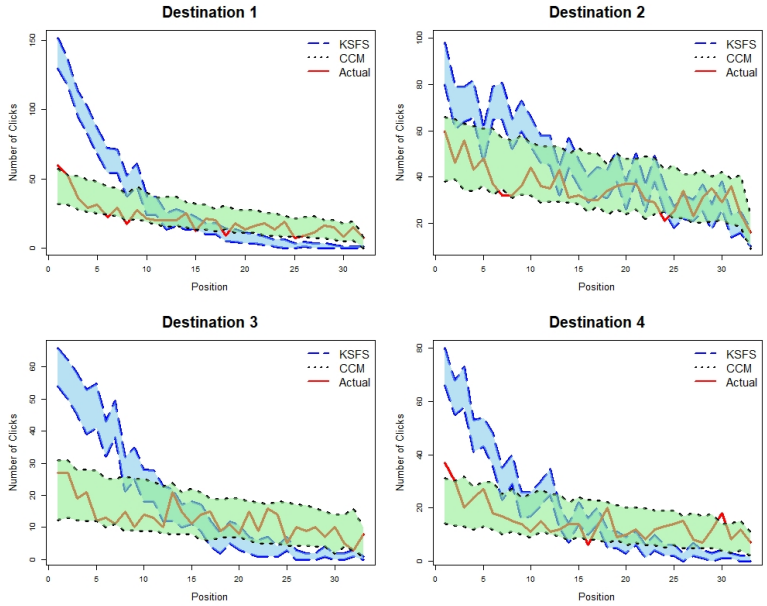


Figure E.2: Simulated Number of Clicks vs. Position with $\sigma_c = 2$

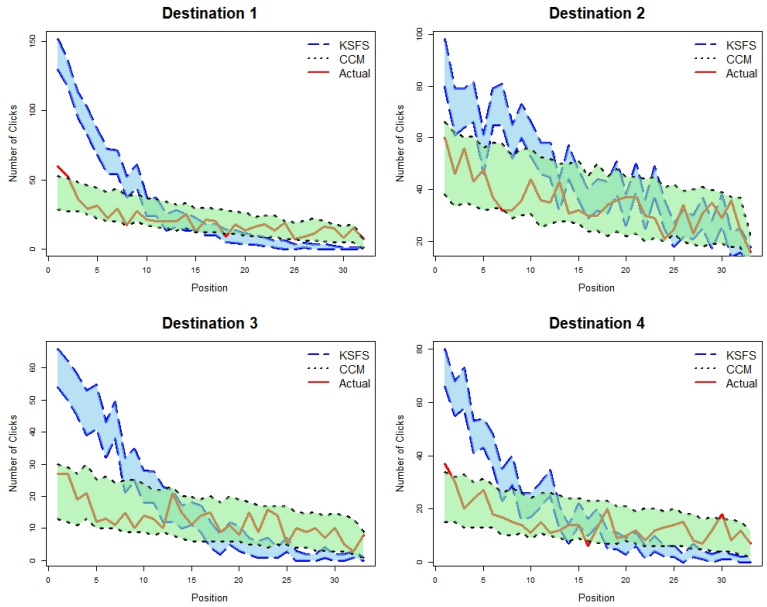


Figure E.3: Simulated Number of Purchase vs. Position with $\sigma_c = 0.5$

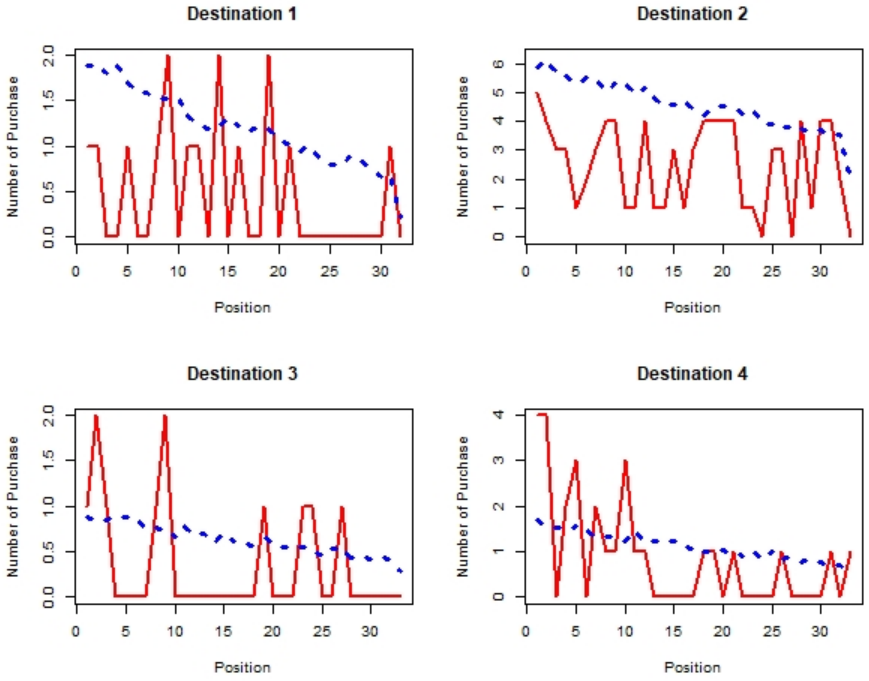
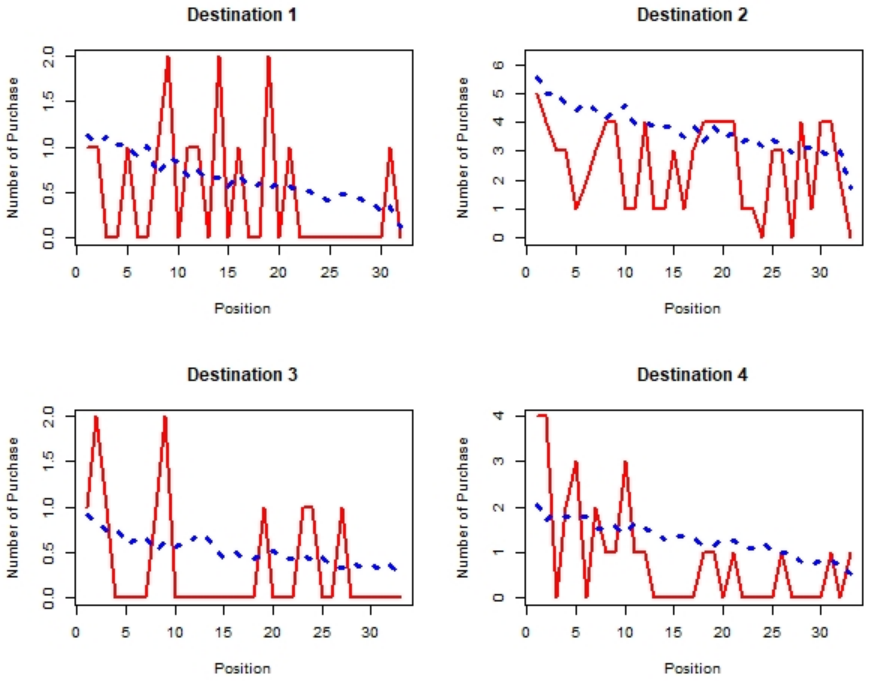


Figure E.4: Simulated Number of Purchase vs. Position with $\sigma_c = 2$



Destination	1				2				3				4			
	$s=3$	$s=5$	$s=10$	$s=20$	$s=3$	$s=5$	$s=10$	$s=20$	$s=3$	$s=5$	$s=10$	$s=20$	$s=3$	$s=5$	$s=10$	$s=20$
Star	0.0418***	0.0436***	0.0647***	0.0684***	0.1511***	0.1705***	0.2116***	0.2195***	0.0459***	0.0484***	0.0806***	0.0874***	0.1051***	0.1171***	0.1433***	0.1381***
	(0.0005)	(0.0008)	(0.0039)	(0.0087)	(0.0019)	(0.0036)	(0.0102)	(0.0165)	(0.0006)	(0.0009)	(0.0051)	(0.0124)	(0.0018)	(0.0036)	(0.0089)	(0.0160)
Review	0.0040***	0.0048***	0.0086***	0.0139	−0.0625***	−0.0750***	−0.0873***	−0.0918***	0.0043***	0.0045***	0.0099***	0.0108***	0.0212***	0.0251***	0.0313***	0.0429***
	(0.0004)	(0.0005)	(0.0020)	(0.0072)	(0.0024)	(0.0042)	(0.0122)	(0.0259)	(0.0004)	(0.0005)	(0.0029)	(0.0095)	(0.0015)	(0.0027)	(0.0051)	(0.0103)
Location	0.0365***	0.0382***	0.0614***	0.0691***	0.0220***	0.0228***	0.0374***	0.0590***	0.0170***	0.0182***	0.0321***	0.0407***	0.0311***	0.0359***	0.0441***	0.0513***
Score	(0.0003)	(0.0004)	(0.0029)	(0.0048)	(0.0022)	(0.0036)	(0.0099)	(0.0218)	(0.0004)	(0.0005)	(0.0026)	(0.0070)	(0.0006)	(0.0013)	(0.0034)	(0.0057)
Brand	0.0040***	0.0045***	0.0094***	0.0061	−0.0140***	−0.0149***	−0.0264***	−0.0438***	−0.0313***	−0.0335***	−0.0598***	−0.0809***	−0.0026***	−0.0015***	−0.0069***	−0.0162***
	(0.0006)	(0.0009)	(0.0028)	(0.0104)	(0.0015)	(0.0026)	(0.0082)	(0.0149)	(0.0007)	(0.0011)	(0.0062)	(0.0194)	(0.0019)	(0.0033)	(0.0086)	(0.0195)
Promotion	0.0142***	0.0154***	0.0309***	0.0479***	0.0617***	0.0712***	0.0933***	0.0952***	0.0188***	0.0193***	0.0351***	0.0348***	0.0100***	0.0162***	0.0359***	0.0563***
	(0.0006)	(0.0008)	(0.0036)	(0.0124)	(0.0015)	(0.0026)	(0.0069)	(0.0141)	(0.0008)	(0.0012)	(0.0056)	(0.0139)	(0.0021)	(0.0039)	(0.0108)	(0.0228)
Price	−0.0776***	−0.0793***	−0.1195***	−0.1299***	−0.1500***	−0.1643***	−0.1965***	−0.2148***	−0.0311***	−0.0332***	−0.0596***	−0.0578***	−0.1056***	−0.1179***	−0.1386***	−0.1368***
	(0.0006)	(0.0009)	(0.0055)	(0.0089)	(0.0023)	(0.0039)	(0.0104)	(0.0183)	(0.0006)	(0.0009)	(0.0045)	(0.0076)	(0.0019)	(0.0037)	(0.0077)	(0.0101)
$\bar{c}$	−23.5696***	−21.3980***	−7.6828***	−3.4494***	−7.3936***	−6.0054***	−3.1763***	−1.6672***	−21.8597***	−19.7577***	−6.8317***	−3.4797***	−9.1525***	−7.2217***	−3.9418***	−2.2881***
	(0.0391)	(0.0668)	(0.7334)	(0.6221)	(0.1250)	(0.1744)	(0.3187)	(0.3026)	(0.0400)	(0.0768)	(0.7881)	(0.7886)	(0.1767)	(0.3228)	(0.4968)	(0.4322)
Position	0.0172***	0.0175***	0.0165***	0.0107***	0.0039***	0.0038***	0.0027***	0.0020***	0.0118***	0.0122***	0.0108***	0.0083***	0.0103***	0.0105***	0.0124***	0.0119***
Effect	(0.0002)	(0.0003)	(0.0007)	(0.0015)	(0.0003)	(0.0005)	(0.0009)	(0.0017)	(0.0002)	(0.0004)	(0.0010)	(0.0020)	(0.0004)	(0.0006)	(0.0013)	(0.0024)
Outside	7.9987***	7.5440***	4.6321***	3.4123***	4.5703***	4.1522***	3.3648***	2.5224***	7.7275***	7.2958***	4.6339***	3.3456***	5.2328***	4.6763***	3.7441***	2.9037***
Option	(0.0246)	(0.0404)	(0.1980)	(0.2447)	(0.0401)	(0.0583)	(0.1332)	(0.1928)	(0.0307)	(0.0544)	(0.2262)	(0.3370)	(0.0517)	(0.0955)	(0.1814)	(0.2216)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table E.1: Estimation Results from KSFS with Heterogeneous Search Cost

	Destination 1				Destination 2				Destination 3				Destination 4			
	Actual	$\sigma_c = 0.5$	$\sigma_c = 2$	KSFS	Actual	$\sigma_c = 0.5$	$\sigma_c = 2$	KSFS	Actual	$\sigma_c = 0.5$	$\sigma_c = 2$	KSFS	Actual	$\sigma_c = 0.5$	$\sigma_c = 2$	KSFS
# Click	656	737.75	677.78	974.29	1148	1249.44	1194.46	1459.42	417	447.42	429.07	566.26	483	488.68	498.37	624.30
# Purchase	15	38.52	20.88	331.48	83	150.73	124.48	598.97	11	20.53	16.91	220.28	28	36.26	42.55	258.54
Position Clicked	13.38	15.28	13.92	10.19	16.75	18.24	17.32	18.49	14.93	16.18	15.37	11.11	15.38	15.84	16.06	11.20
First Clicked Position	12.06	12.38	12.46	5.48	14.73	15.11	15.05	12.90	13.49	14.05	13.99	6.88	13.66	14.26	14.21	7.45
Purchased Position	12.73	13.46	13.09	6.90	15.75	15.28	15.17	13.38	11.55	14.66	14.33	8.14	9.54	14.45	14.36	8.28
RMSE	-	8.17	7.69	30.21	-	8.92	8.34	17.25	-	5.31	5.15	15.82	-	6.16	6.10	15.94

Table E.2: Data Replication Simulation Summary Statistics for Different σ_c

E.2 Out-of-Sample Prediction

Figure E.5: Out-of-Sample Prediction on Number of Clicks ($\sigma_c = 0.5$)

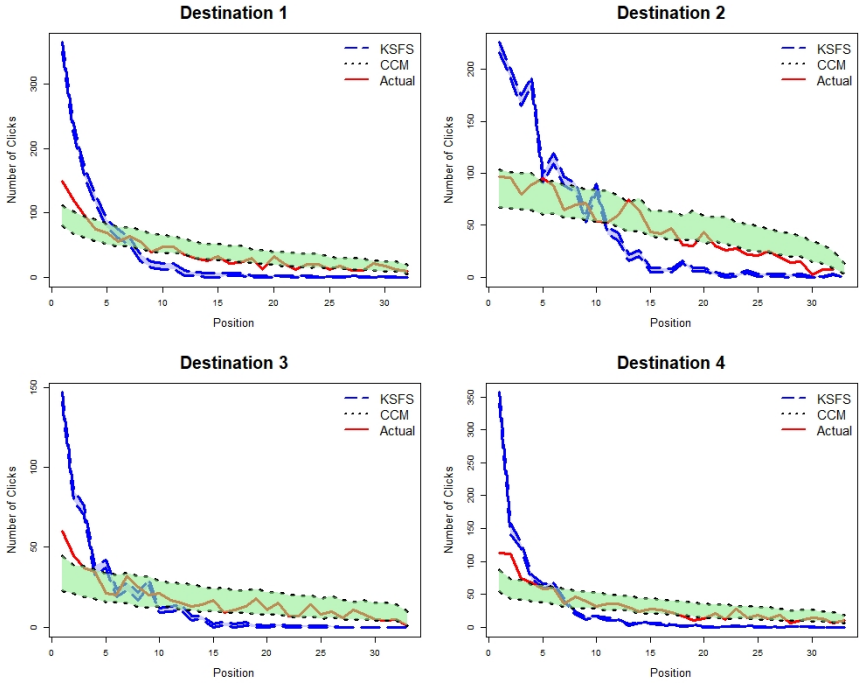
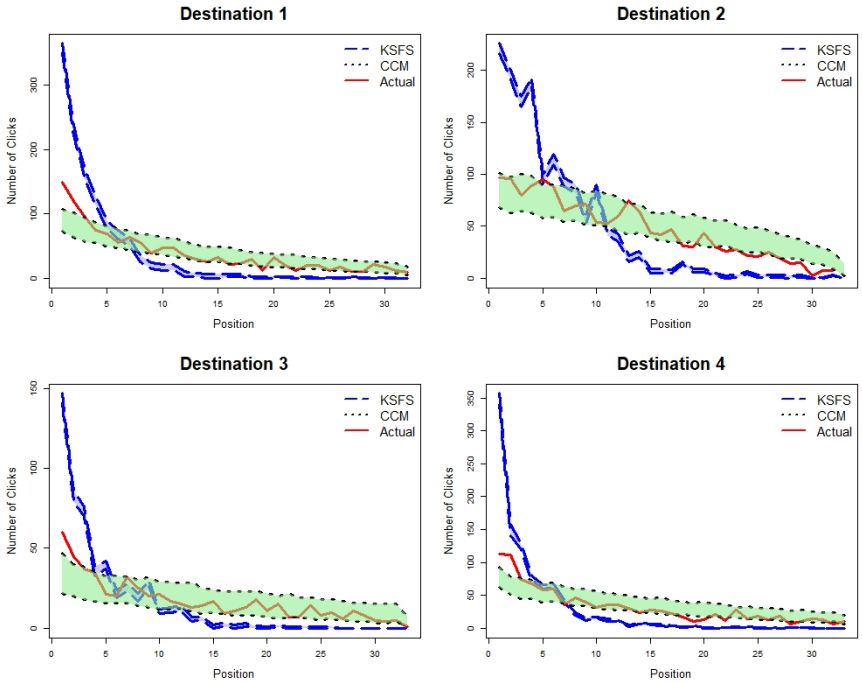


Figure E.6: Out-of-Sample Prediction on Number of Clicks ($\sigma_c = 2$)



APPENDIX F

EXPLORATORY EXERCISE: ACCOUNTING FOR UNOBSERVED QUALITY

The ideal dataset to apply the sequential search model would allow researches to observe every piece of information that consumers observe prior to search. However, this may not be the case sometimes. For instance, consumers may infer product qualities from the pictures of products, some information could be omitted for any reason, or there may exist some type of information that is hard to be stored in the form of variable. Similarly, we can argue that there is some information unobservable to researchers in the case of Expedia. The dataset summarizes each hotel with only 6 characteristics, but consumers' utility may be affected by the pictures of hotels or other amenities provided by hotels (e.g. free breakfast, free Wi-Fi, free shuttle bus, etc.) that are not listed in the data. Therefore, we present one way to take the unobserved quality into account by constructing a proxy variable for the quality from hotels under Expedia ranking. We first briefly discuss how we construct the proxy and present the estimation results and the predictive performance.

F.1 Construction of Proxy Variable

If one can estimate hotel-specific fixed effects in the search model, then such fixed effects will represent the unobserved characteristics. However, in this particular dataset, there are over 300 unique hotels for three destinations and the even the smallest number of hotels for a destination is 47. Therefore, it will drastically prolong the time taken for estimation. Also, infrequent appearance of hotels in impressions would render such approach impractical. Therefore, we instead focus on the set of impressions under the Expedia sorting algorithm for the same four destinations. The construction of the proxy variable consists of two steps. First, we estimate a linear model with hotel fixed effects from Expedia rankings, which are assumed to reflect the hotel quality. Second, we project the estimated fixed effects onto the

hotel characteristics space and treat the residuals as the proxy variable for quality, which is treated as a separate product attribute in the search model estimation for randomly sorted impressions.

According to Expedia's document to hotel partners that explains the visibility of hotels within an impression,[1] the visibility of a hotel is determined by three scores assigned by Expedia. The first score measures the "offer strength" based on the current price, historical average daily rate, active promotion, production, review ratings, location score, and star rating. The second score is named "quality score", which is based on the hotel's competitiveness or rates and availability compared to similar hotels on Expedia, quality and accuracy of hotel description, and guest experience. The last score is called "compensation". If a hotel pays Expedia via a program called Expedia Accelerator, then Expedia will move the hotel to upper position within an impression. But, as the dataset is for period between November 2012 and June 2013 and the Accelerator program was introduced in late 2015, we assume that there is no channel through which hotels can contribute to Expedia's revenue other than commissions on reservations made on its website. These scores motivate us to define variables in the first stage regression as outlined below.

For the first stage, we run a regression on hotels' positions under Expedia rankings for each destination as following.

$$p_{j,t} = \alpha_j + z_{j,t}\gamma + e_{j,t}$$

where $p_{j,t}$ is the (some measure of) position of hotel j in impression t, α_j is hotel-specific fixed effects, and $z_{j,t}$ contains hotel characteristics that Expedia considers when sorting hotels. More specifically, $p_{j,t}$ is defined as following so that higher position (closer to the

1. The document can be accessed at `https://discover.expediapartnercentral.com/wp-content/uploads/2016/12/Expedia_Marketplace-White-Paper-April-2016.pdf`

top) corresponds to larger value.

$$p_{j,t} = \log\left(2 - \frac{pos_{j,t}}{1+N_t}\right)$$

where $pos_{j,t}$ is the raw position of hotel j in impression t, and N_t denotes the total number of hotels in impression t. And $z_{j,t}$ contains (1) price rank, (2) discount rank, (3) negative standardized (within impression) price, and (4) an indicator for promotion. The first two variables are transformed as following so that expected signs of coefficients are positive.

$$r^{(p)}_{j,t} = \log\left(2 - \frac{rr^{(p)}_{j,t}}{1+N_t}\right)$$
$$r^{(d)}_{j,t} = \log\left(1 + \frac{rr^{(d)}_{j,t}}{1+N_t}\right)$$

where $rr^{(p)}_{j,t}$ and $rr^{(d)}_{j,t}$ are the raw ranking of price and discount of hotel j in impression t, respectively, (1 for highest price and highest discount rate). Therefore, a hotel with the lowest price in the impression will get largest value of $r^{(p)}_{j,t}$, and a hotel with the highest rate of discount (compared to historical price) will get the largest value of $r^{(d)}_{j,t}$.

Destination	1	2	3	4
Discount Rank	0.3788***	0.0632***	0.2803***	0.4218***
	(0.0081)	(0.0071)	(0.0112)	(0.0078)
Price Rank	0.0691***	0.0307*	-0.0386	-0.0179
	(0.0186)	(0.0186)	(0.0266)	(0.0161)
-(Std. Price)	-0.0268***	0.0585***	0.0121*	-0.0221***
	(0.0037)	(0.0029)	(0.0052)	(0.0033)
Promotion	0.0636***	0.0581***	0.0423***	0.0926***
	(0.0027)	(0.0019)	(0.0049)	(0.0030)
Adjusted R^2	0.8421	0.8942	0.8486	0.8599

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table F.1: First Stage Regression Result

Table F.1 summarizes this first regression results. We can see that most of estimates are positive and significant and that the specification above captures about 85% of variability in the data for all four destinations. Two destinations have negative and significant estimates for negative standardized price while other two destinations have positive estimates. Different signs of price coefficients may imply the conflicting impact of prices on Expedia's sorting algorithm. Higher priced hotels will bring higher commissions to Expedia when reserved, while lower priced hotels will get more reservations. As we are not aware of the exact mechanism of commissions and we are not trying to understand how Expedia sorts hotels, we accept the estimates as they are. Figure F.1 shows the density of estimated hotel fixed effects.

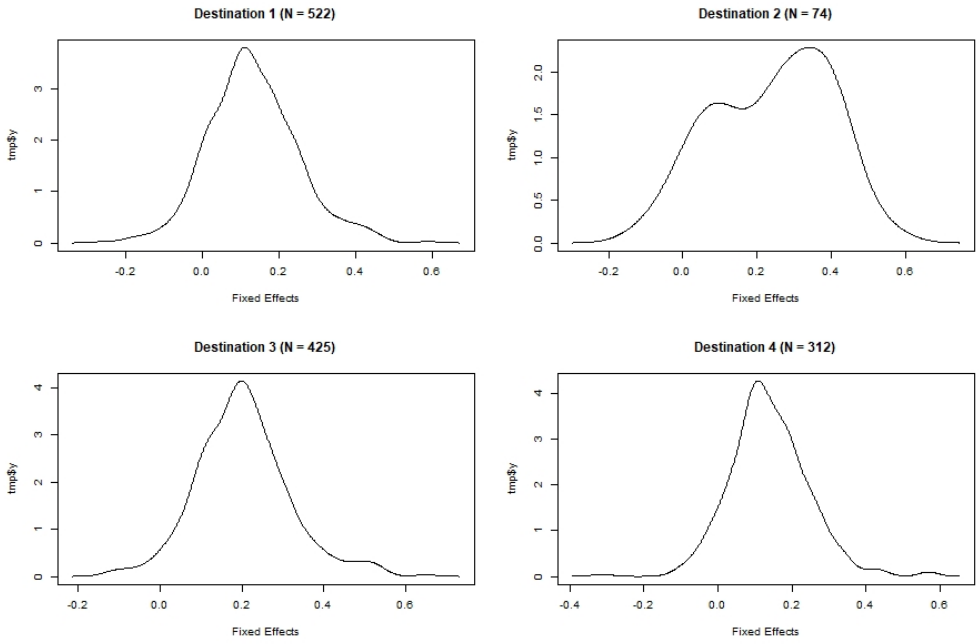


Figure F.1: Density of Estimated Hotel Fixed Effects

Next, the estimated hotel fixed effects $\hat{\alpha}_j$ are projected on the hotel characteristic space as following.

$$\hat{\alpha}_j = Y_j\delta + q_j$$

where Y_j is the vector of hotel characteristics, including star rating, review score, brand

indicator, and location score. As these characteristics stay constant during the data time window, we have to run this second stage regression, instead of running only one regression with $z_{j,t}$ that includes Y_j. We treat the residual q_j as the proxy variable for the hotel quality.

Destination	1	2	3	4
Star	0.0231***	0.0678***	0.0078	-0.0099
	(0.0062)	(0.0196)	(0.0072)	(0.0088)
Review	0.0054	0.0016	0.0312***	0.0298***
	(0.0047)	(0.0204)	(0.0064)	(0.0071)
Brand	0.0011	0.0524	0.0071	-0.0105
	(0.0111)	(0.0396)	(0.0132)	(0.0136)
Location Score	0.0068*	-0.0127	0.0177***	0.0162***
	(0.0034)	(0.0184)	(0.0049)	(0.0040)
Adjusted R^2	0.5349	0.7482	0.7528	0.6092

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table F.2: Second Stage Regression Result

Table F.2 summarizes the second stage regression results. Even though many of estimates are insignificant, most estimates have positive signs as expected, and negative estimates are all insignificant. And as the R^2 ranges from 0.53 to 0.75, there are some variation left after controlling for these impression-invariant variables, implying the existence of the unobserved quality. Figure F.2 shows the distribution of the proxy for quality.

After we obtain the proxy for quality, we apply the search model and run estimation treating the proxy as another variable for hotels in random ranking. One issue with this approach is that some hotels appear on the random ranking but do not appear on the Expedia ranking, so it is infeasible to construct the proxy variable for such hotels. For such hotels, we take the average value of proxy variables of 20 nearest neighbors in terms of L^2 distance in the impression-invariant characteristics space.[2]

2. For destination 2 which has only 47 unique hotels in random ranking, all hotels that appear in random ranking appear also in Expedia ranking.

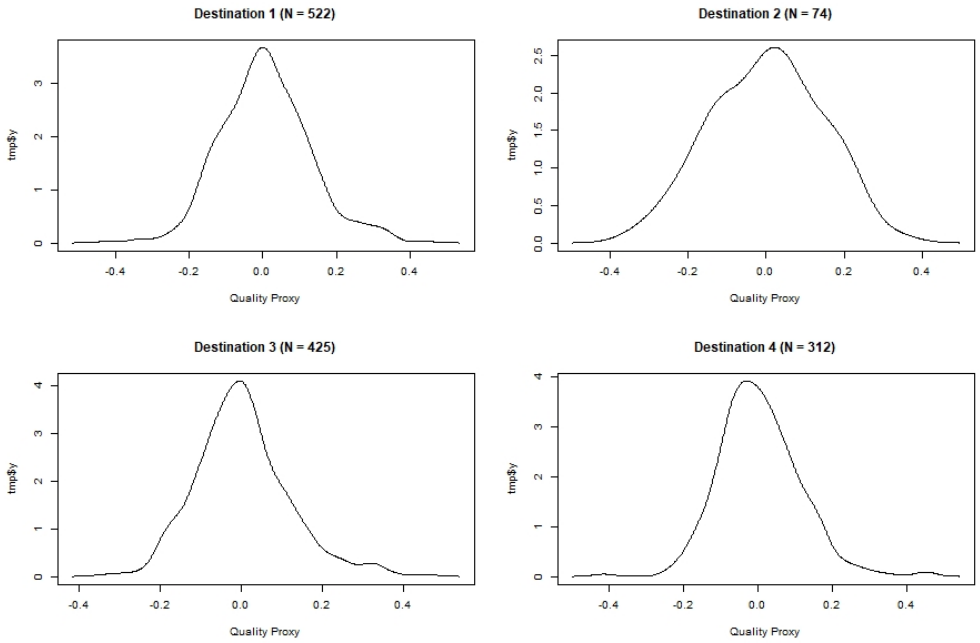


Figure F.2: Distribution of Estimated Proxy for Quality

F.2 Estimation Results

We present the estimation result with quality proxy included in the utility function in Table F.3. Standard errors are obtained by bootstrap resampling method with $B = 200$ (Efron and Tibshirani, 1993).

One can compare Table F.3 and the second column of each destination in Table 6.2 to see the impact of the introduction of the proxy variable for quality. Although the changes in utility parameters are not statistically significant, there is no negative and significant utility parameter (except the coefficient for brand indicator, whose direction of impact on consumer utility is debatable). Destination 2's review score has negative and significant coefficient estimate without quality proxy, but with the proxy, the estimate changes its sign. One change that is consistent across destinations is that the outside option estimate is larger than in Table 6.2, and this change is as expected. As other linear utility coefficients are about the same in terms of magnitude and the coefficient for the newly introduced quality proxy is positive, the overall average of inside goods' utility level is higher than without the

Destination	1	2	3	4
Star	0.1247***	0.2663***	0.1437***	0.1979***
	(0.0204)	(0.0266)	(0.0276)	(0.0482)
Review	0.0235	0.0479	0.0200	0.0719*
	(0.0140)	(0.0397)	(0.0246)	(0.0280)
Location Score	0.1048***	-0.0262	0.0567***	0.0721***
	(0.0108)	(0.0314)	(0.0157)	(0.0172)
Brand	0.0149	0.0261	-0.1021**	-0.0133
	(0.0224)	(0.0240)	(0.0362)	(0.0309)
Promotion	0.0530*	0.0957***	0.0765*	0.0389
	(0.0206)	(0.0221)	(0.0371)	(0.0388)
Price	-0.2060***	-0.3343***	-0.0895**	-0.1925***
	(0.0250)	(0.0340)	(0.0311)	(0.0512)
Quality	0.3312**	1.0393***	-0.0028	0.5927***
	(0.1094)	(0.1075)	(0.1268)	(0.1697)
$\bar{c}$	-1.6341***	-0.5514***	-1.1920***	-0.7199**
	(0.2575)	(0.1356)	(0.3568)	(0.2699)
Position Effect	0.0225***	0.0088***	0.0148***	0.0147**
	(0.0025)	(0.0016)	(0.0026)	(0.0039)
Outside Option	3.3954***	2.9600***	3.2764***	3.174***
	(0.1185)	(0.1769)	(0.1881)	(0.6447)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table F.3: Estimation Results with Quality Proxy ($\sigma_c = 1$)

quality proxy. Therefore, the outside option mean utility estimate has to be higher to reflect such increase in the utility level of inside goods.

F.3 Data Replication

With estimates from Table F.3, we replicate the search sequences and purchase decisions of consumers as done in Section 6.4. Dotted lines in Figure F.3 show the confidence intervals of number of clicks for each position. One can see that the replicated data well captures the actual data pattern presented by the solid line. Although this set of data replication

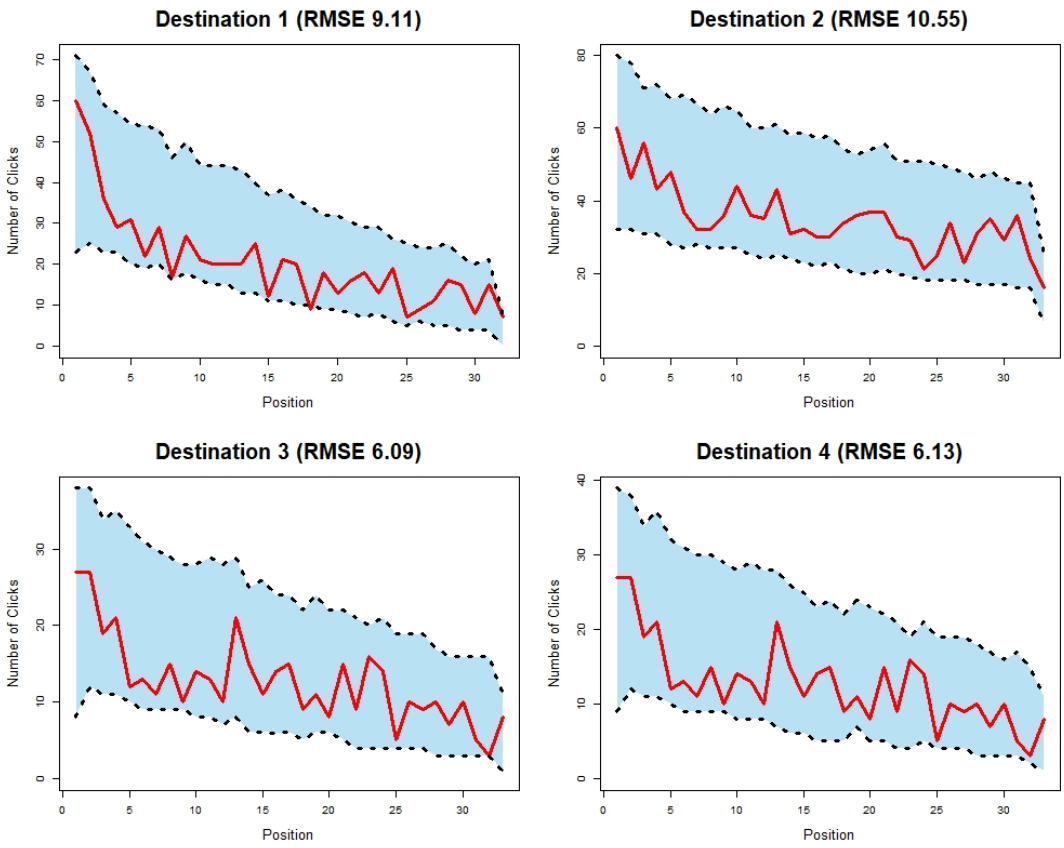


Figure F.3: Simulated Number of Clicks vs. Position with Quality Proxy

shows wider confidence intervals and slightly larger RMSE's than without the quality proxy variable, it appears that the peaks and troughs in the number of clicks are better captured in this set of data replication. These mixed signs from the data replication suggest that there is a room for improvement in the construction of proxy variable and also that the proposed method can accommodate this approach to address the omitted variable bias.

F.4 Discussion

One should note that this section is not intended to make claims that there exists an omitted variable in this particular data. Instead, we intend to demonstrate that the proposed estimation method can easily accommodate the estimation of sequential search model even when there is omitted variable in the data and/or when price is endogenous. Other than the

proxy variable approach we take in this dissertation, one can employ the control function approach by Petrin and Train (2010) to correct for price endogeneity, given that he has access to appropriate instrumental variables. For this particular case, the estimates and the predictions do not show drastic differences, but there will be cases in which researchers have to worry about how to deal with price endogeneity in sequential search model. And the proposed estimation method has the flexibility to deal with the problem without much harm to the predictive performance.