UNIVERSITY OF CALIFORNIA
Los Angeles

An Awareness Model for a Two-sided Matching Market

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Xinyuan Zhang

2022

ABSTRACT OF THE THESIS

An Awareness Model for a Two-sided Matching Market

by

Xinyuan Zhang
Master of Science in Statistics
University of California, Los Angeles, 2022
Professor Mark S. Handcock, Chair

In demographic studies, the latent preferences for partners during a partnership formation process has long been a crucial problem. However, in practical setting, individuals are usually only aware of a certain subset of potential partners, and these should be separated from preferences when modelling this matching process. In this thesis, we address this issue by constructing an awareness model for matching where individuals are aware of a subset of potential partners, and have nontransferable utilities for them based on certain characteristics. We extend Goyal et al.'s framework (2022) to estimate preference and awareness parameters, where awareness parameters are estimated based on continuous covariates. We conduct simulation studies in multiple theoretical scenarios to show that under various settings, the model accurately recovers preference and awareness parameters. We also conduct a case study based on recent marriages observed in the 2019 American Community Survey (ACS) data.

The thesis of Xinyuan Zhang is approved.

Nicolas Christou

Jennie E. Brand

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Many social and economic pairing processes can be viewed as two-sided matching processes. In fields like demography, sociology, education and economics, these processes are especially common, and could be in the form of heterosexual marriages, admissions of students to graduate schools, job searching and other practical scenarios. These processes all require a mutual agreement from agents on both sides to achieve a partnership, usually referred to as the "match" in the matching market.

The underlying mechanisms of two-sided matching markets have been exclusively studied by researchers in various fields, with a particular interest in how partnerships are formed under different contexts and how observable and unobservable factors influence partnership formation processes. A large amount of literature has focused on identifying and estimating individual and societal preferences in different matching markets, but this task remains challenging for several reasons. Firstly, it is in most cases infeasible to have all necessary characteristics of observed agents recorded or have complete information about all available alternatives to each agent when making choices. Secondly, there is an interdependence between individuals' preferences, the available matching opportunities, and the individual's awareness of potential partners. Taking the heterosexual marriage market as an example, women may prefer men with higher education experience. However, there might be limited availability of men with this characteristic, or they could not be aware of all men with higher education in a large market due to factors like geographical constraints. Under these circumstances, some women would have to choose a less educated partner or remain single.

Hence, separating individual preferences and availability of potential partners from the individual's perspective in the final matching achieved is a crucial problem, which has long been recognized but has not been resolved satisfyingly ([CS06] ; [Pol86]; [Pol97]; [DBF01]). Another challenge researchers face is the methodological difficulty regarding the available quantitative methods to handle continuous characteristics when examining the two-sided matching behavior. There have been very limited scholars addressing this problem, among which the work by Dupuy and Galichon [DG14] seems to appear as the only one that provides an extension of the model by Choo and Siow [CS06] to include continuous multivariate attributes.

In 2015, Menzel [Men15] proposed a list of mathematical results that produce a tractable asymptotic approximation to the distributions of observable characteristics in a two-sided matching market with stable matching results. Goyal et al. [Goy+20] later developed Menzel's findings [Men15] and proposed a revealed preferences model to recover latent preference parameters over observable attributes in the population, that is primarily applied in demographic studies. The revealed preferences model is similar to most previous works in this field and is still restricted to only discrete attributes. In this thesis, we further develop Goyal et al.'s work [Goy+20] and propose an *awareness model* that addresses the challenge of limited availability and awareness simultaneously. Our model not only recovers the latent preference parameters, but also recovers an *awareness* parameter over agents' limited awareness towards potential matching partners. Same as the mechanism in Goyal et al.'s work [Goy+20], the parameters estimate the total utility over a stable partnership, given the characteristics of the paired agents on both sides. Moreover, our model also contributes on the modeling front, where we make use of a continuous characteristic in the model to examine awareness. We conduct simulation studies to show the estimation stability and accuracy under different scenarios, and the importance of the awareness factor for decision making. We also conduct a case study using the 2019 ACS data to demonstrate the applicability of the awareness model to empirical data.

The awareness model has a flexible structure that makes the generalization towards one-to-many and many-to-many markets feasible, as discussed by Yeung [Yeu19]. For the purpose of this thesis, we limit our focus to the one-to-one matching market, where each individual is allowed at most one partner.

The rest of this article is organized as follows. The remainder of this introduction chapter discusses the background information and reviews previous studies on identifying individual preferences in two-sided matching markets. Chapter 3 provides the details of the proposed awareness model, which is an extension of the revealed preferences model by Goyal et al. [Goy+20] to include a continuous variable. We also present the pseudo likelihood approach for parameter inference and a bias correction method. In Chapter 4, two sets of simulation studies recovering known parameters using the awareness model are provided to demonstrate the application of the proposed model. Chapter 4 presents a case study using the 2019 ACS data to address the ability of the model in practical application. Finally, Chapter 5 discusses the implications and findings, along with future research directions.

1.1 Background / Literature Review

The process of two-sided matching is usually complex as relationships in many contexts tend to form and dissolve over time. To study the matching market, one critical assumption is to assume the final matching results recorded at a particular discrete time point is *stable*. A stable matching is achieved when the two individuals in a relationship would not prefer being single or being partnered with another individual other than their current partner. Roth and Vande Vate [RV90] showed that in scenarios like the marriage market when randomness is introduced to the matching process, the process will converge to a stable matching outcome and Roth and Sotomayor [RS90] demonstrated that stable matching can be achieved in large populations.

There are many approaches to study the two-sided matching market problem. Follow-

ing multiple previous works, this thesis takes a two-sided discrete-choice model structure. Discrete-choice models assume a context where an agent makes a choice from a known set of options, primarily based on their characteristics and the characteristics of the alternatives available. Discrete-choice models then attempt to estimate the probability of an agent making a specific type of choice. Under this scenario, decision-makers with similar attributes are expected to make similar decisions, and the variations of decisions made by such agents are assumed to be due to unobserved attributes and factors. Traditionally, discrete-choice models are developed in a one-sided context, where only agents from one side of the partnership make decisions. This is inappropriate in many practical circumstances. For instance, a marriage formation is a joint decision by a woman and a man, and a mutual selection is also required between a student and a graduate school during the admission and enrollment process. Instead of dividing the population into decision makers and their available options, agents from both sides make decisions in the partnership to maximize their own utility. Hence, a two-sided discrete choice model framework has been developed in many previous studies (e.g. [CS06]).

Among the studies on two-sided discrete choice models, another critical assumption is whether the utility gained by individuals during partnership formation is transferable. A large number of previous studies in the two-sided matching market on decision makings have assumed transferable utility (TU) among individuals (e.g. [CS06], [GS21]). Choo and Siow [CS06] proposed a static transferable utility model in the heterosexual marriage market scenario, that generates a nonparametric marriage matching function. Their model advances previous work as their matching function considers spillover effects, meaning that the matching outcome between a certain type of woman and a certain type of man is also related to available populations of other characteristics. Their results assume a strict restriction on the unobserved heterogeneity term, that the idiosyncratic taste shifters are group-specific rather than individual-specific. Several later studies are developed on their structure. For instance, Galichon and Salanie [GS21] provided an extension to their framework that uses more gen-

eral but still separable and known distributions on the unobserved characters. Nonetheless, the unobservables in their framework are still group-specific.

Dupuy and Galichon [DG14] also extended the Choo and Siow structure [CS06] to include continuously distributed characteristics in the framework. Same as Choo and Siow [CS06], Dupuy and Galichon [DG14] assumed transferable utility. In their setting, each individual can only be match with another individual whom they met, and this process is described as a Poisson process. Their model uses a direct continuous analog of the multinomial logit choice model to estimate the probability density of a woman of a certain type matches a man of a certain type.

To our knowledge, the Dupuy and Galichon model [DG14] is the only continuous extension of a two-sided discrete choice model in the matching market. Their model considers the limitation of awareness over potential spouses for individuals. However, the subset of the whole population, which is the set of the population each individual has access to, is randomly selected without further description of the selection mechanism. Moreover, their logit framework does not allow an unobserved random taste shock, and has to assume a small amount of variation of the unobserved heterogeneity.

Another set of studies, for instance, Logan et al. [LHN08], Menzel [Men15]and Goyal et al. (2022) assume a non-transferable utility (NTU) structure, where no exchange of utilities gained remains when individuals from both sides form a partnership and the number of distinct matching results can be very large. Logan et al. [LHN08] used a Bayesian approach and proposed a model to estimate preference parameters in the marriage market under the framework that the utility of each agent in the population being matched or remaining single depends on observed characteristics of them and their potential partner, and an unobserved component. Specifically, in the heterosexual marriage context, all men are assumed the same deterministic utility function and all women are assumed the same deterministic utility function. Logan et al. [LHN08]introduced a fixed effect term to account for the unobserved characteristics in the utility function for each agent. They stated that due to computational

limitation, instead of applying the maximum likelihood estimation (MLE) method, they proposed a Bayesian approximation using the Markov chain Monte Carlo (MCMC) method.

However, the proposed Bayesian approach is still relatively limited in that it works best for small populations, and would be highly complex and computationally heavy with a larger population. In the Logan et al. study [LHN08] , the model is applied to data from the National Survey 136 of Families and Households (NSFH) with 674 individuals in the sample. With the relatively small sample size, the model is able to compute preference parameter estimates with their approach. However, the method would encounter computational difficulty when the data size is large, such as when using the American Community Survey (ACS) data where the number of individuals exceeds 100,000. In such cases, it would be extremely complex and time-consuming to update parameter estimates in each step of the MCMC process. Logan et al. [LHN08] also noted some limitations on parameter identifiability. The posterior distribution could have multiple local maxima that may result in misinterpretations of the data, and the relative strengths between multiple terms in the utility functions cannot be explicitly identified.

Menzel [Men15] examined the two-sided matching problem under the NTU assumption in the large market, which is desired especially in demographic studies. He studied the distribution of matched observable characteristics from pairwise stable matchings. The major result from Menzel [Men15] is that he proposed an asymptotically stable approximation of the said characteristic distribution when the market is assumed large. His result indicates that in a large enough population, preferences over characteristics during the matching process and the availability of partners are asymptotically separable in their relationship to the distribution of matching outcomes.

This is a significant finding, as it solves the challenge that observed matching outcomes could be as much the result of personal preferences as the result of limited availability in desired partners. Goyal et al. [Goy+20] extended on Menzel's work [Men15] and proposed the revealed preference model that estimates the latent parameters which produce the ob-

served matching outcomes. They reparameterized Menzel's mathematical results [Men15] and presented equations describing the relationship between the preference parameters, the availabilities of each type of agent from the two sides in the market, and the limiting distribution characteristics in the matching outcomes. Goyal et al. [Goy+20] showed that their model is applicable to a large population, and the preference estimates are consistent across different availability scenarios. Additionally, Goyal et al. [Goy+20] allowed different opportunity sets for different types of individuals.

Under the NTU setting, no previous studies have considered the role of continuous covariates in decision making and matching outcomes. Menzel [Men15] briefly mentioned the possibility of including a continuous variable in the joint surplus function, also as an explanation of individuals' limited awareness of potential partners. In this thesis, we propose a two-sided discrete choice model under the NTU assumption, which we refer to as the *awareness model*, that considers continuous characteristics. Following previous studies, we focus on bipartite networks where agents in the population are divided into two disjoint groups, and a partnership can only occur between two agents from different groups. We extend Goyal et al.'s framework [Goy+20] to account for continuous characteristics in the utility function. Our model explicitly describes the case when utility is also affected by the possibility of individuals being aware of potential partners with certain characteristics. We utilize key results from Goyal et al. [Goy+20] and estimate not only the preference parameters, but also an awareness parameter based on an observed distribution.

CHAPTER 2

The Awareness Model

2.1 The Revealed Preferences Model

The setting is a one-to-one, bipartite matching model with non-transferable utility (NTU). In this thesis, we specify the context of heterosexual marriages within a two-sex population where individuals are either male or female. Individuals can be observed as matched, or "married" in this context, where they have at most one partner of the opposite sex, and they also have the option to remain single or unpartnered.

By the nature of marriage as the form of partnership in this thesis, a two-sided evaluation process occurs during partnership formation. Individuals use a utility function to evaluate their potential partnership options, which is assumed to be specified identically for individuals of the same gender. The utility function contains a deterministic component and an unobserved random component, that accounts for the fact that individuals' characteristics are partially observed. The latent parameters, or commonly called "preference" parameters of the utility function govern this pair formation process, that represent how individuals would choose between potential partners with different characteristics of interest [LHN08].

We denote the number of women and men in the population as N_w and N_m, which gives the total population size as $N = N_w + N_m$. Following Goyal et al.'s notation [Goy+20], we let variables x_i and z_j represent the information of some observed characteristics on woman $i = 1, ..., N_w$ and man $j = 1, ..., N_m$, respectively. It is worth noting that while the original equation setting by Menzel [**menzel**] allows both discrete and continuous covariates

contained by x and z, the previous work by Goyal et al. [Goy+20] specified a restriction of the model application with discrete covariates only. In this thesis, we apply our awareness model to a combination of discrete and continuous covariates, which is allowed under the methodological development that will be discussed in this chapter.

We assume that all individuals follow the utility-maximizing behavior when choosing partners. In this thesis, following Goyal et al. [Goy+20], we focus on the case in which the deterministic components of the utility functions have an addictive linear structure. Finally, we define the utility gained by woman i with observed characteristics x_i from partnering with man j with observed characteristics z_j as

$$U_{ij} = \underbrace{\theta_{w0} + \sum_{k=1}^{K_w} \theta_{wk} X^k(x_i, z_j)}_{\substack{\text{deterministic} \\ \text{component}}} + \underbrace{\eta_{ij}}_{\substack{\text{unobserved random} \\ \text{component}}} \tag{2.1}$$

where $\theta_W = [\theta_{w0}, \theta_{w1}, \dots \theta_{wK_w}]^T$ is the parameters denoting the woman's preferences, which are vectors of the coefficients in the woman's utility function. Hence, the deterministic component of the woman's utility contains an intercept term θ_{w0} and K_w additive linear functions, where each function represents a proportion of the woman i's total utility gained from matching the man j with certain characteristics. Since we assumed that characteristics are only partially observed, the random component of the utility function accounts for unobserved information about individuals that may potentially affect partnership formation. Following Menzel [Men15], it is assumed to be independently and identically distributed and drawn from a distribution in the domain of attraction of the extreme-value type-I (Gumbel) distribution.

Similarly, we define the utility gained by man j with observed characteristics z_j from

partnering with woman i with observed characteristics x_i as

$$V_{ji} = \underbrace{\theta_{m0} + \sum_{k=1}^{K_m} \theta_{mk} Z^k(x_i, z_j)}_{\substack{\text{deterministic} \\ \text{component}}} + \underbrace{\zeta_{ji}}_{\substack{\text{unobserved random} \\ \text{component}}} \tag{2.2}$$

where $\theta_M = [\theta_{m0}, \theta_{m1}, \dots \theta_{mK_m}]^T$ is the set of parameters representing men's preferences.

Additionally, individuals have the option to remain single, and the random utility for this choice is defined as

$$\begin{aligned} U_{i0} &= 0 + \max_{k=1,\dots,N_m^\delta} \{\eta_{i0,k}\} \\ V_{j0} &= 0 + \max_{k=1,\dots,N_w^\delta} \{\zeta_{j0,k}\} \end{aligned} \tag{2.3}$$

for females and males, respectively. Following Goyal et.al. [Goy+20], the individual's deterministic utility of staying single is 0. The non-deterministic component in Equation (2.3) is defined as the maximum of N_m^δ and N_w^δ independent draws of $\eta_{i,k}$ and $\zeta_{j,k}$, the Gumbel-domain-of-attraction distributed term for women and men respectively.

2.2 Large population approximation and parametrization

Following Goyal et.al. 's notation [Goy+20], we let $w(x)$ be the number of women in the population with characteristics x and $\bar{w}(x) = w(x)/N$. Similarly for men, let $m(z)$ represents the number of men with characteristics z in the population and $\bar{m}(z) = m(z)/N$. Consider a scenario where the population has utilities drawn from the model (1), (2) and (3), from which a probability distribution over the observed characteristics is induced. Randomly sample a person from this population, the densities of single women with observed characteristics x and single men with observed characteristics z can be represented by $f(x, *)$ and $f(*, z)$ respectively. Let $f(x, z)$ define the joint density of the matches between women of type x and men of type z. The densities satisfy the overall normalization constraint:

$$\int f(x,z)dxdz + \int f(x,*)dx + \int f(*,z)dz = 1 \tag{2.4}$$

Moreover, from $f(*,z)$ we deduce $f(x,\diamond)$ and $f(\diamond,z)$ as the probability of being partnered for women of type x and men of type z respectively, which further gives:

$$\begin{aligned} \bar{w}(x) &= f(x,*) + f(x,\diamond) \\ \bar{m}(x) &= f(*,z) + f(\diamond,z) \end{aligned} \tag{2.5}$$

Additionally, we define parameters $g(x,*)$ and $g(*,z)$ as the log-odds that a women or man with observed characteristics x or z is single respectively via the equations

$$\begin{aligned} f(x,*) &= \frac{\bar{w}(x)e^{g(x,*)}}{(1+e^{g(x,*)})} \\ f(*,z) &= \frac{\bar{m}(z)e^{g(*,z)}}{(1+e^{g(*,z)})} \end{aligned} \tag{2.6}$$

This way, the log-odds both have range in the real line. We also note that

$$\begin{aligned} f(x,\diamond) &= \frac{\bar{w}(x)}{(1+e^{g(x,*)})} \\ f(\diamond,z) &= \frac{\bar{m}(z)}{(1+e^{g(*,z)})} \end{aligned}$$

so that (2.5) is satisfied.

Finally, a major result of Menzel [Men15] reparametrized by Goyal et.al. (2022) states that under mild regularity conditions, with a large population and stable matching, the frequencies approximately satisfy the relations:

$$f(x,z) = 2\frac{e^{W(x,z)+g(x,*)+g(*,z)}}{[1+e^{g(*,z)}][1+e^{g(x,*)}]}\bar{w}(x)\bar{m}(z) \qquad \forall x,z \tag{7}$$

where

$$W(x,z|\boldsymbol{\beta}) = U(x,z|\theta_W(\boldsymbol{\beta})) + V(z,x|\theta_M(\boldsymbol{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

is the sum of the discrete deterministic components of the utilities as specified in Goyal et.al 's model [Goy+20], and $\theta_W(\boldsymbol{\beta})$ and $\theta_M(\boldsymbol{\beta})$ are functions such that $\boldsymbol{\beta}$ parameterizes $W(x,z|\cdot)$.

Moreover, the solution should satisfy the equilibrium conditions on the parameter values $\boldsymbol{\beta}$, as reparametrized by Goyal et.al. (2022):

$$
\begin{aligned}
e^{-g(x,*)} &= \int \frac{e^{W(x,s)+g(*,s)}\bar{m}(s)}{1+e^{g(*,s)}}ds \quad \forall\ x \\
e^{-g(*,z)} &= \int \frac{e^{W(s,z)+g(s,*)}\bar{w}(s)}{1+e^{g(s,*)}}ds \quad \forall\ z
\end{aligned}
\tag{8}
$$

(2.5) and (7) together make it possible to obtain estimates $\hat{\boldsymbol{\beta}}$ of the desired preference parameters.

2.3 Matching with Restricted Awareness

In the original revealed preferences model [Goy+20], individuals are assumed to have preferences over observed discrete characteristics of the opposite side of the market. However, in practical settings, individuals may not be aware of all potential partners in a large population context because of various factors, which could also be a plausible explanation of why many individuals remain single in a large population. To illustrate, this thesis develops an extension of Goyal et al.'s model [Goy+20], in which each individual is only aware of a subset of all potential partners and the probability of meeting a potential spouse can be described as a function of observed characteristics. A natural example of such characteristics can be geographic location, that individuals are only aware of potential partners who are less geographically distanced. In this case, the probability of meeting becomes a function of spatial distance, which is usually observed as a continuous covariate. Hence, we propose an awareness model, which provides a mixture framework that incorporates both discrete and continuous attributes based on the work by Goyal et al. [Goy+20].

While the revealed preferences model [Goy+20] provides a flexible framework that is well suited for different sample sizes and census type data, A direct application of this framework is infeasible. The deterministic component of the utilities gained by partnering with the

individual of a certain type in this context is affected by both discrete and continuous attributes, which consequently requires additional estimation of awareness parameters over continuous characteristics. We take the structure of Proposition 5.1 from Menzel [Men15], let the probability of a woman at location x meeting a man at location z be represented by $r(x,z) \in [0,1]$ and the deterministic components of the utilities now becomes

$$W(x,z|\boldsymbol{\beta}) = U(x,z|\theta_W(\boldsymbol{\beta})) + V(z,x|\theta_M(\boldsymbol{\beta})) + T(x,z|\theta_C(\boldsymbol{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

where

$$T(x,z|\theta_C(\boldsymbol{\beta})) = \theta_c \log r(x,z), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

represents the utility gained by forming a partnership with women and men have location x and z respectively.

The original revealed preferences model requires the defined parameters $g(x,*)$ and $g(*,z)$, the log-odds that a woman or man with characteristics x or z is single. With continuous attributes also observed, when x and z are continuous, $g(x,*)$ and $g(*,z)$ become real-valued functions satisfying the functional constraints (8). It is hence natural to consider smoothing splines models with Bernstein polynomials for the log-odds function $g_c(x,*)$ and $g_c(*,z)$ with continuous characteristics x and z, where

$$g_c(x,*) = \sum_{k=1}^{K} \alpha_k^W b_{k,K}(x)$$
$$g_c(*,z) = \sum_{k=1}^{K} \alpha_k^M b_{k,K}(z)$$

$b_k(x)$ and $b_k(z)$ are the generalized Bernstein polynomials of order k, that is

$$b_{k,K}(x) = \binom{K}{k} x^k (1-x)^{K-k}$$
$$b_{k,K}(z) = \binom{K}{k} z^k (1-z)^{K-k}$$

Then $g_c(x,*)$ is in the forms of linearly combined Bernstein polynomials of order k with $\{\alpha_k^W\}_{k=1}^K$ being its coefficients or weights, and similar cases apply to z. Thus, solving (8') comes down to estimating the preference parameter and the weighting coefficients $\{\alpha_k^W\}_{k=1}^K$ and $\{\alpha_k^M\}_{k=1}^K$. Moreover, to construct the mixture form of the revealed preferences model with both discrete and continuous characteristics, we reweighted the constraints with the log odds that a woman or man with discrete characteristics x_{cat} or z_{cat} is single. We can then rewrite the constraints (8') and it becomes:

$$
\begin{aligned}
e^{-g_c(x,*)} &= \frac{n_m}{n_w+n_m}\bar{m}(z_{cat})\int \frac{e^{W(x,s)+\sum_{k=1}^K \alpha_k^M b_k(s)}\bar{m}(s)}{1+e^{\sum_{k=1}^K \alpha_k^M b_k(s)}}ds \quad \forall\ x \\
e^{-g_c(*,z)} &= \frac{n_w}{n_w+n_m}\bar{w}(x_{cat})\int \frac{e^{W(s,z)+-\sum_{k=1}^K \alpha_k^W b_k(s)}\bar{w}(s)}{1+e^{-\sum_{k=1}^K \alpha_k^W b_k(s)}}ds \quad \forall\ z
\end{aligned}
\tag{8'}
$$

2.4 Inference: Pseudo Likelihood Approach

The preference parameters are $\boldsymbol{\psi} = (\boldsymbol{\beta}, \{\alpha_k^W\}_{k=1}^K, \{\alpha_k^M\}_{k=1}^K)$. If we observe the entire population, the likelihood for $\boldsymbol{\psi}$ would become extremely complex, since each of the matchings between individuals is interdependent. Hence, following Goyal et al.'s work [Goy+20], we use a pseudo likelihood approach, that uses a surrogate for the likelihood for $\boldsymbol{\psi}$. The population likelihood for $\boldsymbol{\psi}$ is:

$$
_{pop}(\boldsymbol{\psi}|\{x_i,z_i,w_i^w\}_{i=1}^{N_w},\{z_j,x_i,w_j^m\}_{j=1}^{N_m}) = \sum_{i=1}^{N_w}\log f(x_i,z_i) + \sum_{j=1}^{N_m}\log f(x_j,z_j) \tag{9}
$$

Since we do not observe the full population, we approximate the population likelihood by the design-based estimators:

$$
\begin{aligned}
&\text{p-log-lik}(\boldsymbol{\psi}|\{x_i,z_i,w_i^w\}_{i=1}^{n_w},\{z_j,x_i,w_j^m\}_{j=1}^{n_m}) \qquad (10)\\
&= \sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} c(x,z)\log f(x,z) + \sum_{x\in\mathcal{X}} c(x,*)\log f(x,*) + \sum_{z\in\mathcal{Z}} c(*,z)\log f(*,z) \qquad (2.7)
\end{aligned}
$$

where $c(x,*)$ and $c(*,z)$ are the design-based estimates of the numbers of single women of type x and single men of type z respectively, and $c(x,z)$ is the design-based estimates of the number of matches between women of type x and single men of type z. Hence, the design-based estimators define the empirical version of the distribution $\bar{f}$ and satisfy:

$$\sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} c(x,z) + \sum_{x\in\mathcal{X}} c(x,*) + \sum_{z\in\mathcal{Z}} c(*,z) = N_h$$

2.5 Bias Correction

The estimators in the above section are based on large population approximation to the generating process, as developed and applied by Menzel [Men15] and Goyal et al. [Goy+20]. Hence, it is likely to be biased in many circumstances where the population is not sufficiently large. Goyal et al. [Goy+20] proposed a series of bootstrap procedures for bias correction and proved its effectiveness, that can be directly applied to our awareness model. The bootstrap procedure is standard. We first sample the households of k individuals from the observed sample with repetition, where we have k number of directly sampled households. This process is repeated b times, which provides us with b sets of bootstrapped samples. Then, we fit the awareness model to all b sets of samples and obtain the the bootstrapped parameter estimates denoted as $\boldsymbol{\psi}^* = [\boldsymbol{\psi}^*_{(1)}, \boldsymbol{\psi}^*_{(2)}, ..., \boldsymbol{\psi}^*_{(b)}]$ for a single parameter $\boldsymbol{\psi}$ as denoted above. Finally, the point estimator correcting the bias is conducted as $\hat{\boldsymbol{\psi}}_{BC} = 2\hat{\boldsymbol{\psi}} - \frac{1}{b}\sum_{i=1}^{b} \boldsymbol{\psi}^*_{(i)}$. The empirical results and effectiveness of this estimator under the awareness model will be discussed in the following section.

CHAPTER 3

Simulation Study

3.1 Method

In this section, we illustrate the properties of the awareness model by conducting two simulation studies in specified theoretical scenarios. In simulation study $\mathcal{A}$ we demonstrate that under different geographic distributions of the population, the awareness model accurately estimates underlying preference and awareness parameters that partially motivate matching outcomes. We will also compare the awareness model estimates with the bootstrap-corrected estimates under the same scenarios. In simulation study $\mathcal{B}$, we show that the model estimates the preference and awareness parameters accurately under a counterfactual scenario where education levels of the population are adjusted corresponding to the population density, and also emphasizes the importance of geographic awareness when recovering preferences over discrete components. In all studies, we show the maximum pseudo-likelihood estimates for the preference and awareness parameters.

The basic procedure for the two simulation studies is the same. We assume a heterosexual marriage market in which males and females base partnership decisions on their education level and the education of prospective spouses and some other unobserved characteristics, as their awareness of potential partners is restricted by spatial distance. We assume that the marginal distributions of gender and education within the population are known and represented as $\{\bar{w}(x), \bar{m}(z)\}$. Additionally, we assume that the form of the partnership utility function $W(x_i, z_j|\boldsymbol{\beta})$ and the preference and awareness parameters $\boldsymbol{\beta}$ for individuals

in the market are both known.

We suppose a population of size N which reflects the gender and education distributions, and the geographical distribution of the population under several counterfactual scenarios and the partnership preferences $\boldsymbol{\beta}$. In all simulation studies, we assume the population has an even gender distribution, one-third of the population is paired and the data is supposed to be a sample of n_h households from the simulated population. We fit the awareness model to the data to produce estimates $\hat{\boldsymbol{\beta}}$ of the original preference parameters.

3.2 Theoretical Scenario Specification

The primary focus of the two simulation studies is to illustrate that the awareness model accurately recovers underlying preference and awareness parameters when a continuous characteristic, which in this context is the spatial distance between a woman and a man calculated from their respective geographic location, is added to the model. Hence for simplicity, gender is assumed to have an even distribution in all scenarios, and one-third of the population is randomly chosen and assumed to be in a stable heterosexual marriage.

We consider several theoretical scenarios for simulation study $\mathcal{A}$ and $\mathcal{B}$, where we specified different marginal distributions for education and different geographic population density distributions. In simulation study $\mathcal{A}_1$, the aim is to illustrate that the awareness model provides accurate estimation under three theoretical population density settings, referred to hereafter as $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$. Education is assumed to have a uniform marginal distribution for both men and women in all scenarios here. In scenario $\mathcal{A}_1$, we assume a simple uniform population in the range of $(0, 1)$ density distribution, meaning that the population is evenly distributed over a certain area. In scenario A_2, a two-cluster Gaussian mixture distribution with $\boldsymbol{\pi} = \{0.35, 0.65\}$, $\boldsymbol{\mu} = \{0.25, 0.75\}$ and $\boldsymbol{\sigma} = \{0.1, 0.1\}$is assumed for population density. This could be viewed as a preliminary geographical simulation of a population from two nearby cities or counties in reality, one of a smaller size and one of a much bigger size.

The last scenario $\mathcal{A}_3$ assumes a relatively extreme case, where the population is assumed to display a beta distribution with $\alpha = 0.2, \beta = 0.2$ over a certain spacial scale. This theoretical scenario imitates a population from two equal size cities that are located rather far apart.

In simulation study $\mathcal{B}$, we show that the awareness model provides an accurate estimation of preference and awareness parameters when the marginal distribution of education is adjusted. We first adjust the marginal education distribution for men and women accordingly to two of the population distributions identified in simulation study $\mathcal{A}$ as shown in Table 3.1. For each of the two population distributions, we adjust the marginal distribution of education level in two ways, one moderate adjustment and one relatively extreme adjustment. For instance, under the mixture Gaussian population distribution, we specify that the largest portion of the population located between the location value 0.15 to 0.35 complete high school, while for people located between the location value 0.65 and 0.85, the largest population group has completed some years of college. With a moderate adjustment in $\mathcal{B}_{1.1}$, these two groups each take 40% of the population in their respective areas. We have the same specification in scenario $\mathcal{B}_{1.2}$, but with a more extreme adjustment, 70% of individuals in the two areas complete high school and some college respectively. Similar settings apply to scenarios $\mathcal{B}_{2.1}$ and $\mathcal{B}_{2.2}$, where we specify a beta location distribution. Figure 3.1 provides a visualization of the 4 settings described in the above table. The area is split according to the specifications in Table 3.1, into 5 and 3 subareas under the two population location distributions respectively. Within each subarea, the 4 bars represent the number of population with each level of education. The black lines in each plot represents a scaled population density over the location range from 0 to 1. Note that for illustration purpose, Figure 3.1 displays the average counts and distributions from each simulation setting.

Moreover, we illustrate the importance of geographical information for estimating the preference parameters by comparing preference parameter estimates under the same theoretical scenario produced by the awareness model and the original revealed preferences model. Specifically, we use the same simulation data from the four scenarios specified above

Table 3.1: Adjusted Education Distributions under the two theoretical scenarios

	Scenario $\mathcal{B}_{1.1}$: Gaussian Mixture Location Distribution		
Location	**[0.15,0.35]**	**[0.65,0.85]**	**else**
Education level	% Population	% Population	% Population
1 (< high school)	25	15	25
2 (high school)	40	25	25
3 (some college)	20	40	25
4 (≥ bachelors)	15	20	25
	Scenario $\mathcal{B}_{1.2}$: Gaussian Mixture Location Distribution		
Location	**[0.15,0.35]**	**[0.65,0.85]**	**else**
Education level	% Population	% Population	% Population
1 (< high school)	10	10	25
2 (high school)	70	10	25
3 (some college)	10	70	25
4 (≥ bachelors)	10	10	25
	Scenario $\mathcal{B}_{2.1}$: Beta Location Distribution		
Location	**[0,0.25]**	**[0.75,1]**	**else**
Education level	% Population	% Population	% Population
1 (< high school)	25	15	25
2 (high school)	40	25	25
3 (some college)	20	40	25
4 (≥ bachelors)	15	20	25
	Scenario $\mathcal{B}_{2.2}$: Beta Location Distribution		
Location	**[0,0.25]**	**[0.75,1]**	**else**
Education level	% Population	% Population	% Population
1 (< high school)	10	10	25
2 (high school)	70	10	25
3 (some college)	10	70	25
4 (≥ bachelors)	10	10	25
Total	**75.0**	**100.0**	**25.0**

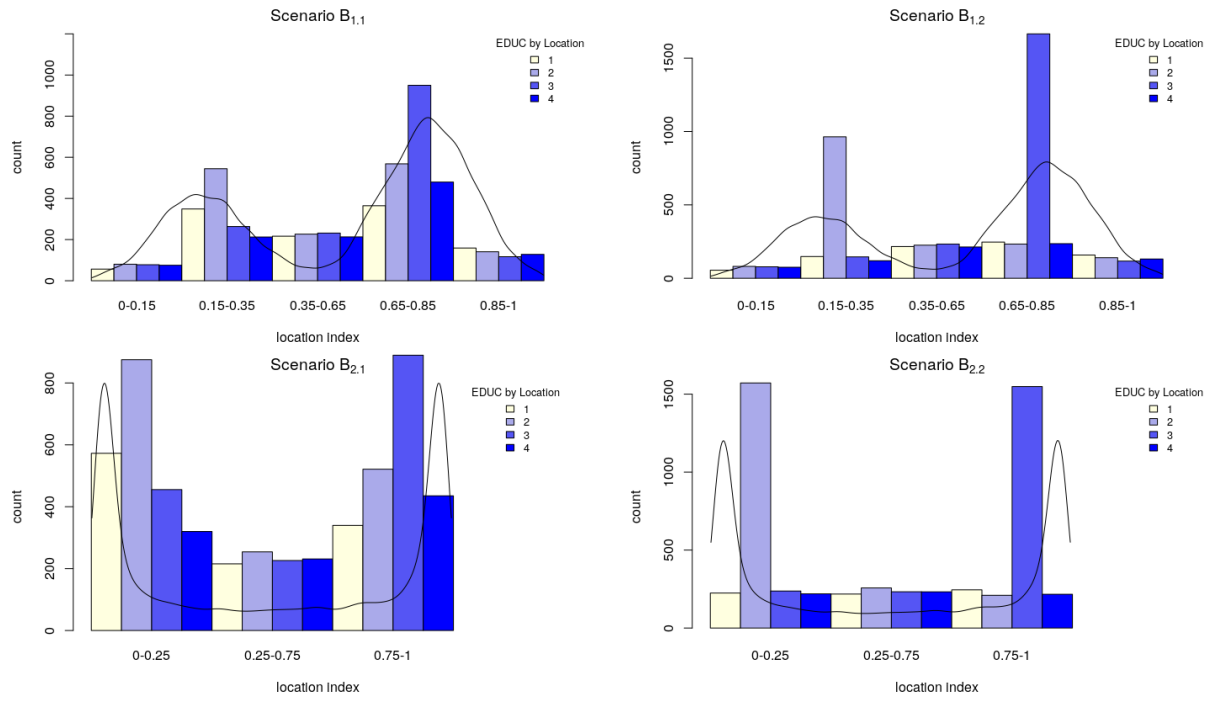

Figure 3.1: Average marginal education distribution by location in Simulation study A; 100 simulations, $N_h = 4000$

and further specify four more scenarios as shown in Table **??** in the Appendix for comparison. The only difference of scenario set $\mathcal{C}$ from $\mathcal{B}$ is that the assigned proportion of the population with high school and come college education in each scenario are swapped. For example, in $\mathcal{C}_{1.1}$, 40% of the population located between the location value 0.15 to 0.35 now have some college education. Similarly, we provide a visualization of the 4 settings in set $\mathcal{C}$ with Figure A.1 in the Appendix. We estimate the preference parameters for the eight scenarios only with the revealed preference model by Goyal et al. [Goy+20], which ignored the geographical data. We show that geographic restriction on awareness over potential spouses is crucial and affects decision making when forming a partnership, and accurate parameter estimation requires all information.

In both sets of simulation studies, we simulate 100 populations each of size $N_h = 4000$ from a population of $N = 8000$, assuming that the education and gender distribution are

even and one-third of the population are married. Moreover, since all scenarios are artificially created, the underlying preference parameters over education level are hence specified as all 1's and the awareness parameter is specified as -2.

We test the performance of the awareness model under the same model specification for both simulation studies. We first manually specified a set of preference and awareness parameters $\boldsymbol{\beta}_0$ which we assume is the underlying truth. In each simulated population, the known preferences $\boldsymbol{\beta}_0$ are applied to calculate the total household utility for every potential partnership and form a stable matching. We fit the awareness model on the observed stable matching outcome from the simulated population and compare the parameter estimates $\hat{\boldsymbol{\beta}}$ to the true underlying preferences $\boldsymbol{\beta}_0$.

In this thesis, we consider the simple model specification assuming that the utility a woman derives from a partnership is based on her own education level and whether her partner shares that same education level, and the awareness of potential partners restricted by spatial distance. There is a corresponding utility function for males. We refer to this as a *type-based match model,* because preference is based on an individual's own type and whether or not their partner's type matches theirs.

Let

$$X^k(x_i, z_j) = Z^k(z_j, x_i) = \mathbb{I}\{x_i = z_j = k\}.$$

The deterministic component of woman i's utility when she is partnered with man j is

$$U(x_i, z_j|\theta_W(\boldsymbol{\beta})) = \theta_{w0} + \sum_{k=1}^{4} \theta_{wk} X^k(x_i, z_j). \tag{3.1}$$

Similarly, the deterministic component of the utility of man j when partnered with woman i is

$$V(z_j, x_i|\theta_M(\boldsymbol{\beta})) = \theta_{m0} + \sum_{k=1}^{4} \theta_{mk} Z^k(z_j, x_i). \tag{3.2}$$

Additionally, the geographic availability between woman i and her partner can be simply

represented as the absolute distance between woman i and man j. Hence, the continuous component of the utility of woman i and man j being partnered is

$$T(x_i, z_j|\theta_{cts}(\boldsymbol{\beta})) = \theta_{cts}|x_i^{loc} - z_j^{loc}|. \tag{3.3}$$

Then, the total utility of woman i and man j if they partnered with each other is given by the sum of Equations 3.1 and 3.2:

$$\begin{aligned} W_{ij}(x_i, z_j|\boldsymbol{\beta}) &= \theta_{w0} + \theta_{m0} + \sum_{k=1}^{4}(\theta_{wk} + \theta_{mk})\mathbb{I}\{x_i = z_j = k\} + \theta_{cts}|x_i^{loc} - z_j^{loc}| \\ &= \beta_0 + \sum_{k=1}^{4}\beta_k\mathbb{I}\{x_i = z_j = k\} + \beta_{cts}|x_i^{loc} - z_j^{loc}|, \end{aligned} \tag{3.4}$$

where $\beta_t = \theta_{wt} + \theta_{mt}$.

Following the procedure taken by Goyal et al. [Goy+20], we refer to this model as the *uniform homophily mixture model* because of the shift in the discrete part of the deterministic component of the utility is uniform for all types (education levels) of individuals. Note that while it is not the focus of this thesis, as Goyal et al. [Goy+20] shown, the uniform homophily mixture model can be easily extended if assuming different utility functions of the discrete component.

3.3 Results

3.3.1 Simulation study $\mathcal{A}$: Geographic Distribution Variation

For simulation study $\mathcal{A}$, we simulate populations of size $\mathbf{N}$ and sample size of household N_h for all three theoretical scenarios, and used the Gale-Shapley algorithm in each simulated population to conduct stable matching on individuals. The plots in Figure 3.2 show the distribution of the **30** parameter estimates under three population density distribution for

theoretical scenario $\mathcal{A} \in \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3\}$. The read lines in the plots represent the prespecified true preference and awareness parameters.

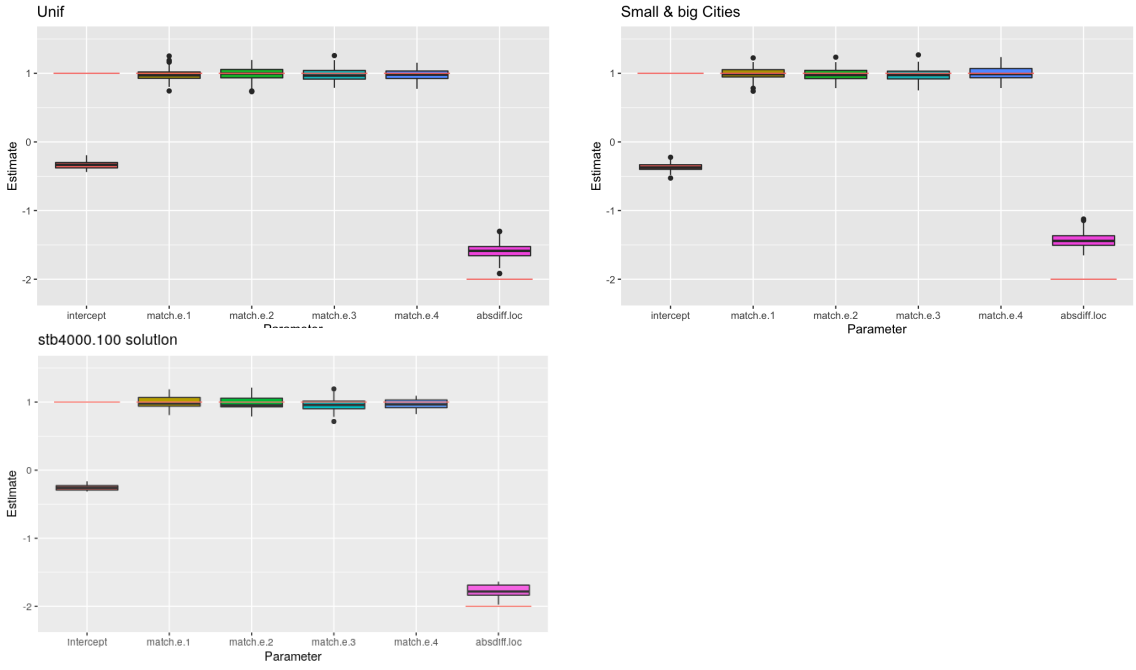

Figure 3.2: Distribution of parameter estimates in Simulation study A (Geographic Distribution Variation); 100 simulations, $N_h = 4000$

From Figure 3.2, we can interpret that under different population density distributions, the awareness model shows similar accuracy in producing estimations of the parameters. The estimates of all parameters resemble a normal distribution. The three plots indicate that the mean estimates of all awareness model parameters have decent accuracy except for the intercept parameters, while the estimation accuracy for the matching preference parameters seems slightly than for the awareness parameters. The performance of the awareness parameter estimation has no noticeable variation under the three hypothetical population density distributions.

Figure 3.3 shows the mean of log-odds of women and men being single by location value and education level under the three population density distributions from each of the 100 simulations. The lines in each plot represent the mean log odds of remaining unmatched

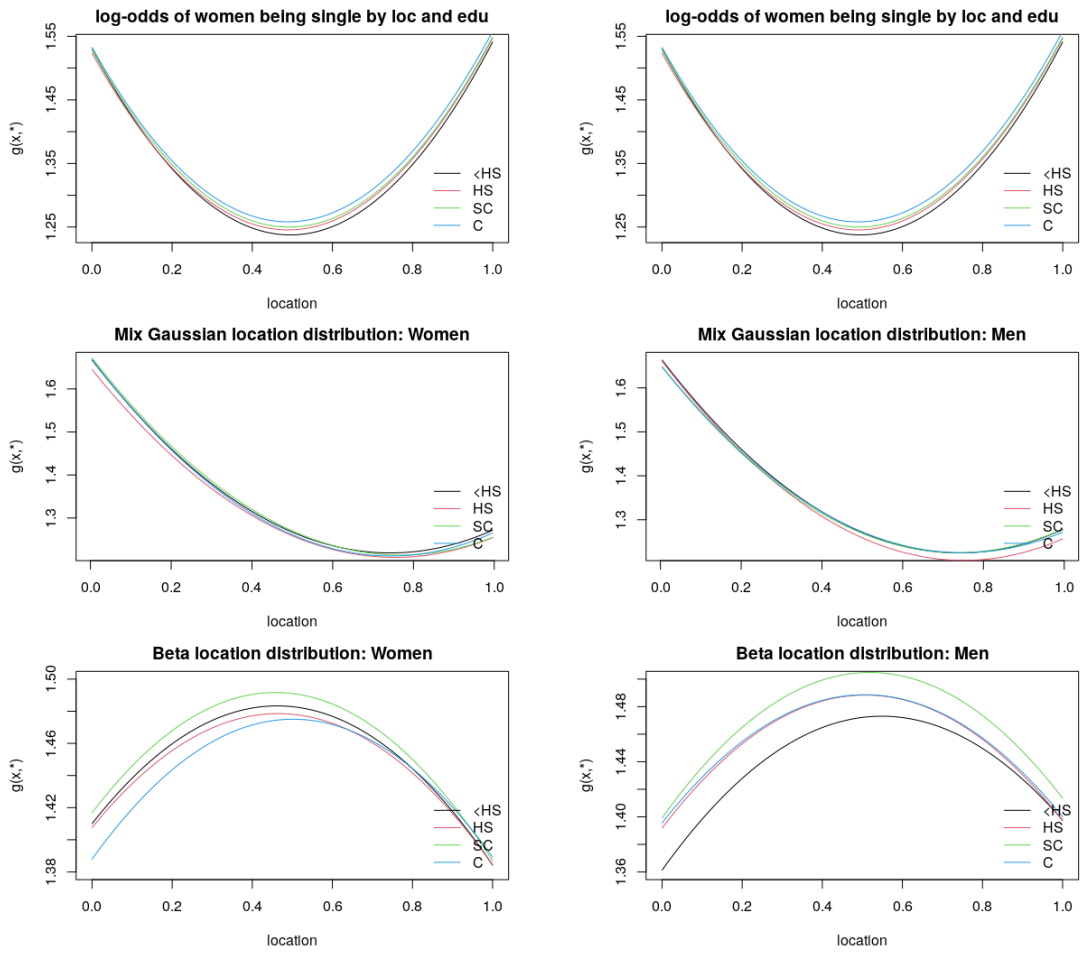

Figure 3.3: Mean Log-Odds of being single in Simulation study A (Geographic Distribution Variation) 100 simulations

at the four education levels. There are no significant gaps between the four lines in each plot, indicating that under all hypothetical scenarios, the probability of being single at each location between different education levels has no significant difference. This is reasonable as an even marginal education distribution was specified. Further, we can interpret from the shape of the lines in each plot that the log-odds of being single at each location is more affected by the population distribution. Under the uniform location distribution, individuals located in the middle area are the least likely to be single with a large population on both sides. With the mixture Gaussian population density distribution, individuals least likely

to be single are located around the area with the highest population density. The shape of log-odds curves under Beta population density distribution resembles a bell shape, meaning that with most population concentrated at the two ends of the area, the chance of remaining single is the highest in the middle area where the population is small.

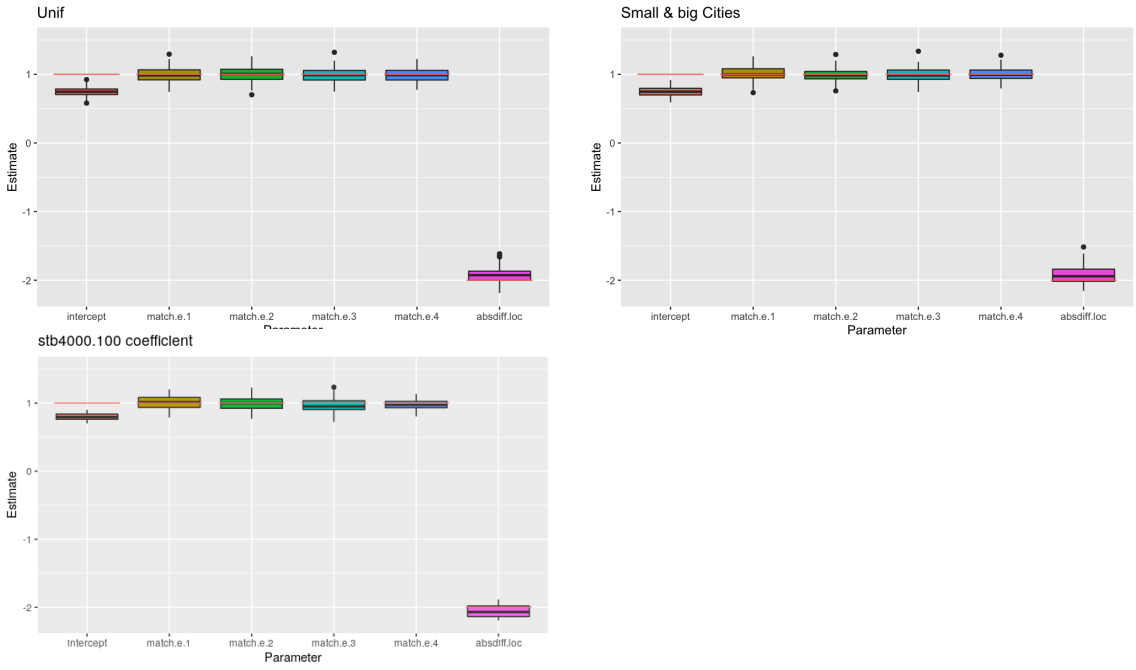

Figure 3.4: Bootstrap Corrected Distribution of parameter estimates in Simulation study A (Geographic Distribution Variation); 100 simulations, $N_h = 4000$

We also show the bootstrap corrected estimates in Figure 3.4. Due to computational limitations, the number of bootstrap processes repeated is specified as twenty.**revise** While the preference parameter estimates remain relatively accurate, the performance of the intercept and awareness parameter estimates improves evidently after the adjustment. All estimates resemble a normal distribution, and the accuracy of the bootstrap-corrected estimates in the three scenarios are relatively on the same level. It is reasonable to conclude the effectiveness of the bootstrap correction method and the results in the next sections are bootstrap-corrected.

3.3.2 Simulation Study $\mathcal{B}$: Adjusted Marginal Education Distribution

Simulation study $\mathcal{B}$ is conducted to compare the performance of the awareness model in more complex and often more realistic scenarios, and to show the importance of individual location information when estimating preference parameters. In all eight scenarios, we simulate 100 populations each of sizes $N_h = 4000$ from the eight different population and education settings, each size of $N = 8000$. Same as in the previous simulation study, we use the Gale-Shapley algorithm to generate a stable matching in the population with true parameters pre-specified. Figure 3.5 and Figure 3.6 show the distribution and distribution comparison of the 100 bootstrap-corrected parameter estimates from the population with adjusted marginal education distributions under Gaussian Mixture location distribution and beta distribution respectively. The red lines in each plot indicate the true parameter value specified. The left column shows the awareness model estimates and the right column shows the comparison between the preference parameter estimates from the awareness model and from the revealed preferences model.

The estimates for all parameters resemble a normal distribution with a few observable outliers. Based on the plots in the left column in Figure 3.5 and Figure 3.6, the mean estimates of the preference parameters in all eight scenarios appear to align with the true values relatively well. It is noticeable that the model produces a slightly more accurate estimation of parameters in scenario $\mathcal{B}_{1.1}$, $\mathcal{B}_{2.1}$, $\mathcal{C}_{1.1}$ and $\mathcal{C}_{1.1}$, where the marginal education distributions are moderately adjusted. This pattern is more observable for the awareness parameter.

Furthermore, the plots in the right column in both figures show that if the location information is ignored, the preference parameter estimates from the two models behave differently. Several noticeable patterns are discovered. Firstly, the intercept estimate is more biased under the discrete model. Second, certain preference parameters are more biased under the discrete model. For instance in scenario $\mathcal{B}_{1.1}$, while the parameter estimates for $\beta_0, \beta_1, \beta_2$ and

β_4 from both model have similar performance, the estimate of β_3 is more biased from the true value under the discrete model. This pattern is more evident in $\mathcal{B}_{1.2}$, where the marginal education distribution is adjusted more extremely. The plots for scenario $\mathcal{B}_{2.1}$ and $\mathcal{B}_{2.2}$ show that with more extreme education distribution adjustment, the estimation performance of β_2 and β_3 under the discrete model declined in accuracy significantly. Moreover, comparing the results from set $\mathcal{B}$ to the corresponding settings in set $\mathcal{C}$, swapping the education distribution settings for high school and some college education in each scenario results in a reverse estimation performance from the discrete model under the Gaussian Mixture location distribution. In scenario $\mathcal{C}_{1.1}$ and $\mathcal{C}_{1.2}$, the estimates of β_2 become the most biased among the four preference parameters. No significant performance difference can be observed under the Beta location distribution between simulation set $\mathcal{B}$ and $\mathcal{C}$. Comparing the location and education specifications and the estimation results under each of the scenarios, the estimate of a certain preference parameter under the discrete model is mode biased when the amount of population with the corresponding level of education is large. This indicates that without the necessary geographical information, the discrete model fails to separate preference and availability during the estimation process.

3.3.3 Discussion

The awareness model performs consistently and satisfyingly in both simulation studies, demonstrating the ability of the model to accurately recover the parameters under various theoretical conditions. The bootstrap bias correction method is proven effective and necessary in this context where the number of population is rather limited, especially for the intercept parameter. Under more complex settings where the marginal education distribution is adjusted according to the location distribution, the awareness model shows a stable and decent estimation accuracy. In comparison, the revealed preferences model that ignores the geographical distribution appears to perform poorly and failed to separate preference and availability. This further illustrates the importance of awareness for preference parameter

estimation.

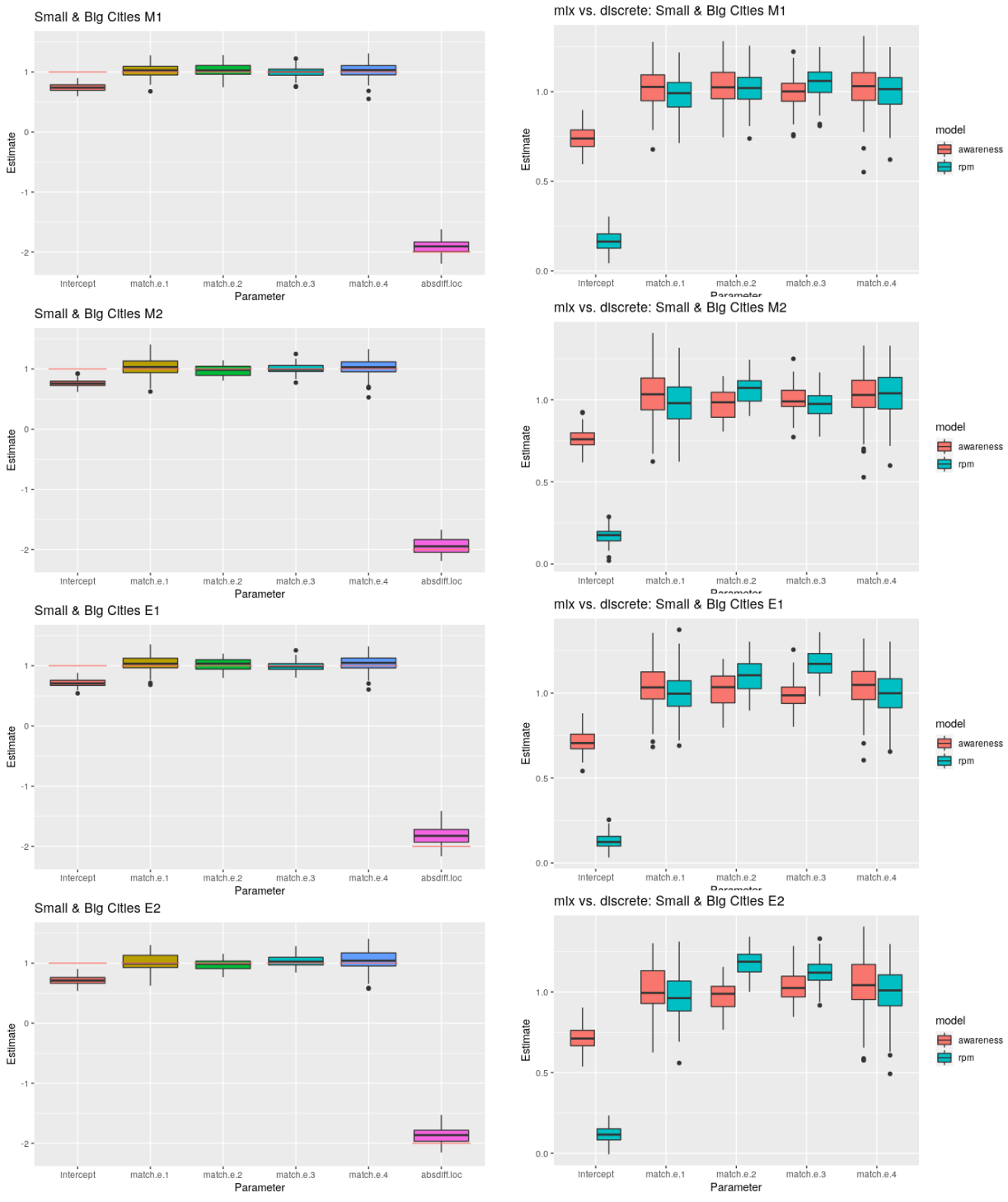

Figure 3.5: Distribution of parameter estimates in Simulation study B: Part 1; 100 simulations, $N_h = 4000$

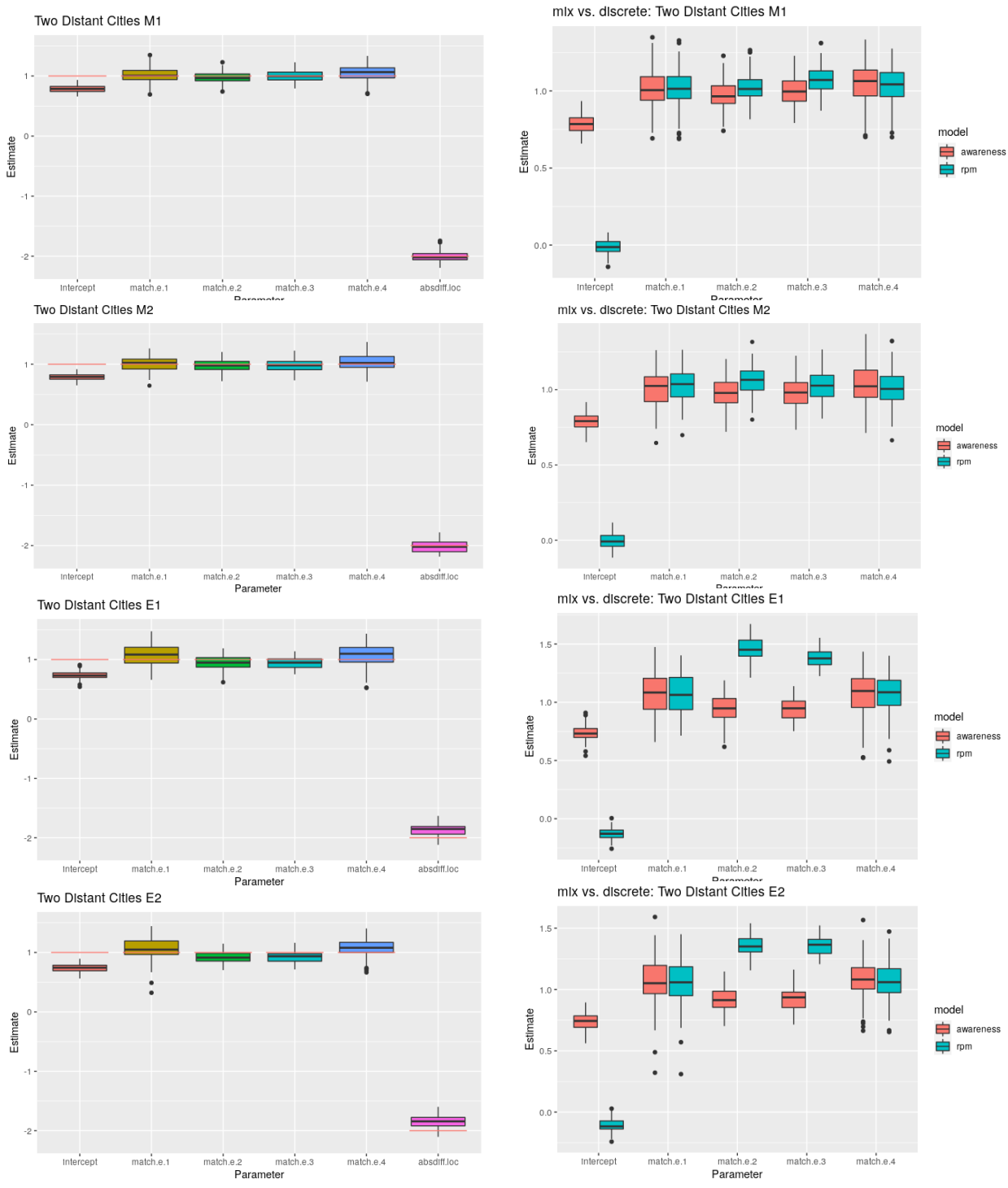

Figure 3.6: Distribution of parameter estimates in Simulation study B: Part 2; 100 simulations, $N_h = 4000$

CHAPTER 4

Case Study: American Community Survey 2019

4.1 Data

In this section, we apply the awareness model to the 2019 American Community Survey (ACS) data from IPUMS [Rug+22] to estimate preferences and awareness in the marriage market. A detailed description of the data can be found in the handbook guides [Bur20]. The ACS is a nationwide survey distributed to over 3.5 million addresses each year to provide reliable and timely demographic, social, economic and housing data every year. A 1-year, a 5-year and a 1-year supplemental estimates are released by the Census Bureau in regard to the ACS each year.

While various census data have been used in many marriage market studies, our study has specific requirements for the data, as the awareness model requires the spatial distance between individuals as a continuous covariate. Since the process of choosing a partner over potential partners is prior to the formation of the marriage, the geographical location of married individuals in the data should be recorded before their marriages. Hence, the ACS data is attractive for our purposes since it not only reports the geographical location of each individual, but also provides data for 1 or 5 consecutive years.

Due to the limitation on the precision of the PUMA codes recorded, instead of analyzing the 2019 ACS data for the whole country, here we focus on North Carolina's data. North Carolina is currently the 28th largest and most populous state in the US, with 100 counties and 45 MIGPUMAs reported in the dataset. The population distribution in North Carolina

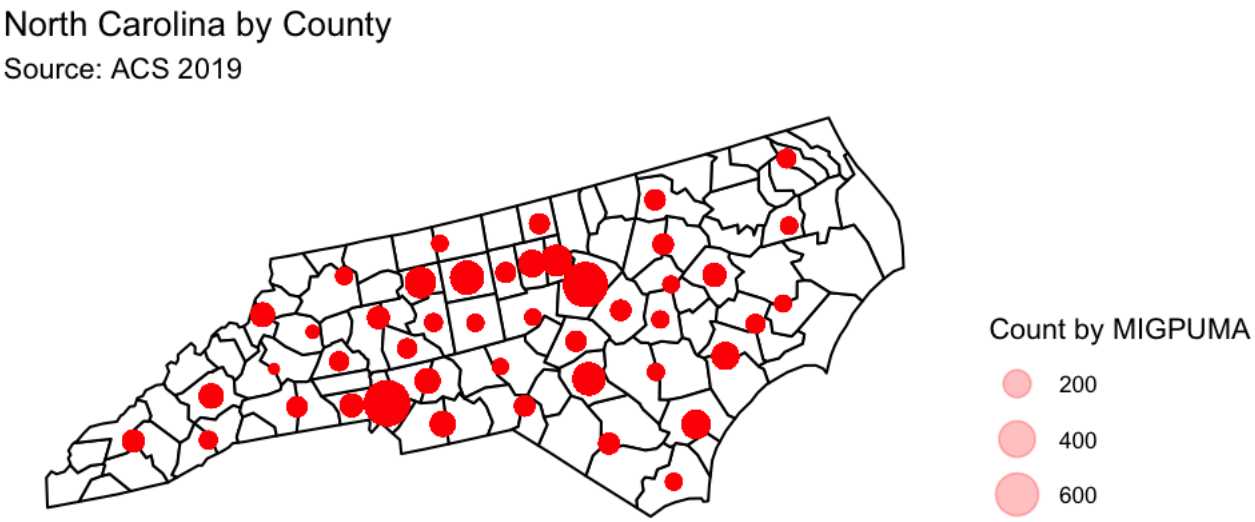

Figure 4.1: North Carolina by County

is particularly interesting for the purpose of our research. Figure 4.1 provides a visualization of the population density given by the data across the state. As a historically rural state, the North Carolina population is concentrated among several urban areas, with a large proportion of counties remaining rather rural. One area in North Carolina that is particularly attractive for our study is the Research Triangle area. The Research Triangle consists of nine counties and is anchored by the cities of Raleigh, Durham and the town of Chapel Hill, which are home to North Carolina State University, Duke University and the University of North Carolina at Chapel Hill respectively. This area also has one of the highest population densities in North Carolina, and with the three research universities located within the Triangle, the population could be more educated than the state average. Hence, in this case study, the awareness model will be applied to the North Carolina state data, and the Research Triangle will be further studied as one major urban area within the state.

With the limitations in data selection discussed above, the analytic sample from North Carolina ACS 2019 data consists of 4970 individuals, among which 4690 remain single and 280 got married in the last year. By survey design, the 280 newly-married samples are married to other individuals in the sample. Hence, there are 140 couples in our sample. The 9-county Research Triangle accounts for 23.84% percent of the North Carolina population, with 54 newly-married samples among 1185 individuals. Table 4.1 describes the gender and education distributions for the whole state and the Research Triangle. It should be mentioned that since the sample size is relatively small, the education level is divided into three instead of four categories, as shown below.

Table 4.1: Gender and Education Distributions under the three availability scenarios

	Males		Females	
Education Level	% Population	% of Males	% Population	% of Females
	North Carolina			
1 ($\leq$ high school)	29.4	56.7	21.2	44.1
2 (some college)	11.5	22.1	13.3	27.7
3 ($\geq$ bachelors)	11.0	21.2	13.6	28.2
Total	**51.9**	**100.0**	**48.1**	**100.0**
	the Research Triangle			
1 ($\leq$ high school)	24.7	49.3	19.5	39.1
2 (some college)	10.8	21.6	12.8	25.5
3 ($\geq$ bachelors)	14.6	29.1	17.6	35.4
Total	**50.1**	**100.0**	**49.9**	**100.0**

Table 4.2: Summary of model fit

$\mathcal{F}_1$: **Awareness model: North Carolina**			
	Estimate	**Std. Error**	**p-value**
Intercept	5.182	0.305	$< 1e-04^{***}$
match.edu.1	-0.0884	0.238	0.711
match.edu.2	-0.221	0.355	0.533
match.edu.3	1.007	0.247	$< 1e-04^{***}$
geo	-0.223	0.012	$< 1e-04^{***}$
$\mathcal{F}_2$: **Awareness model: the Research Triangle**			
	Estimate	**Std. Error**	**p-value**
Intercept	-3.344	0.248	$< 1e-04^{***}$
match.edu.1	1.146	0.411	0.005
match.edu.2	1.591	0.555	0.004
match.edu.3	1.410	0.490	0.004
geo	0.0357	0.004	$< 1e-04^{***}$
$\mathcal{F}_3$: **RPM model: the Research Triangle**			
	Estimate	**Std. Error**	**p-value**
Intercept	-3.321	0.311	$< 1e-04^{***}$
match.edu.6	-1.255	4.358	0.773
match.edu.7	-0.703	3.780	0.853
match.edu.10	1.853	0.439	$< 1e-04^{***}$

4.2 Results

We apply the awareness model to the North Carolina state data and the data of the Research Triangle. We also apply the revealed preferences model to the Research Triangle dataset. In both models, the utility function has the same structure as in the simulation studies, meaning that individuals will only gain utility in a match where both sides have the same education experience. Table 4.2 reported the bootstrap corrected estimates of the three fitted models $\mathcal{F}_1$ to $\mathcal{F}_3$ respectively. The intercept coefficient in all models represents the preference of remaining single, and the three matching coefficients followed represent the preference of choosing a partner with the same amount of education for individuals at each education level respectively. The geo coefficient in $\mathcal{F}_1$ and $\mathcal{F}_2$ represents the estimate of the awareness parameter over geographical distance in miles, which is calculated by the centroid distance between MIGPUMAs.

For individuals from North Carolina, the estimates from model $\mathcal{F}_1$ indicates that the only significant matching preference parameter is β_3, the one for partnerships between individuals with equal or higher than bachelor's education level. β_3 is estimated to be around 1, indicating that there is a strong preference for the most educated population in North Carolina to marry someone with the same amount of education. Additionally, the significant negative result of the awareness parameter means that North Carolina residents are more aware of potential partners with closer resident locations, which is expected by intuition. It can be explained as with each mile between an individual and his potential spouse, the utility of being partnered with this individual decreases by 0.223.

Limiting the focus to the nine-county Research Triangle, the awareness model coefficients indicate that individuals at each education level in this region all prefer to marry a spouse at the same education level. The awareness parameter is significant and positive, which appears counter-intuitive. However, since the data used in this model contains only 27 couples, the result could be biased since the mechanism of the model requires a large population

for asymptotic approximation. Hence, a possible interpretation may be that geographical awareness has no influence on the matching utility value in this scenario. Moreover, the discrete revealed preferences model estimates indicate that individuals with equal or higher than college education in this area are the most likely to choose a partner with the same amount of education. For individuals at the other three education levels, the model estimates show no significantly matching preference when choosing partners. This is a reasonable result since students, faculties and staffs from the three research universities consist a large amount of the Research Triangle population and within the 54 newly-married individuals in this area, 18 women and 14 men have a bachelor's degree or higher.

CHAPTER 5

Discussion and Concluding Remarks

This thesis proposes a novel two-sided discrete choice model under the NTU assumption that separates preference and availability, and considers individuals' limited awareness of matching opportunities. This model is developed on the revealed preferences model [Goy+20], which we have extended to the mixture case here, where observed characteristics in the model can be both discrete and continuous.

The technical advance of the awareness model is that it allows characteristics with continuous densities. Simulation studies have been conducted to demonstrate the estimation accuracy and consistency under different technical challenges, and show the importance of the awareness factor for preference parameter estimation. In simulation study $\mathcal{A}$, we simulate three small populations (N = 8000) with different distributions of the continuous covariate and run the Gale-Shapley algorithm to obtain stable matchings. We are able to compute parameter estimates with relatively high accuracy under the three scenarios. We further adjust the distribution of a discrete covariate in simulation study $\mathcal{B}$, and simulate eight small populations (N = 8000) with different combinations of distributions of observed characteristics. The model produces parameter estimates that are again very close to the pre-specified value under all scenarios. We also apply Goyal et al.'s revealed preferences model [Goy+20] to the eight simulated datasets and compare the preferences parameter estimates from the two models. We show that the discrete model fails to separate preference and availability when ignoring the awareness information. We also conduct an empirical study by applying the awareness model to the North Carolina dataset from the 2019 ACS data. In this context,

the matching is the formation of heterosexual marriage. The observed discrete characteristic for matching preference is the education level, and the awareness of potential spouses is calculated based on the spatial distance between two individuals. We show the applicability of this model to empirical data.

The awareness model opens up interesting possibilities for future research. The technical contribution of allowing a continuous covariate can be useful in many studies. Although we specify the continuous component as a measurement of the limited awareness, it has a flexible structure that allows other ways of interpretation with minor alternations. For example, in demographic studies, age is a popular and interesting variable which due to methodological constraints, is usually simplified as a categorical variable. Our model is able to include age as a continuous variable, and it can be interpreted as either a parallel term in the utility function or as an awareness term, meaning that people are only aware of potential partners in a certain age range.

APPENDIX A

Appendix

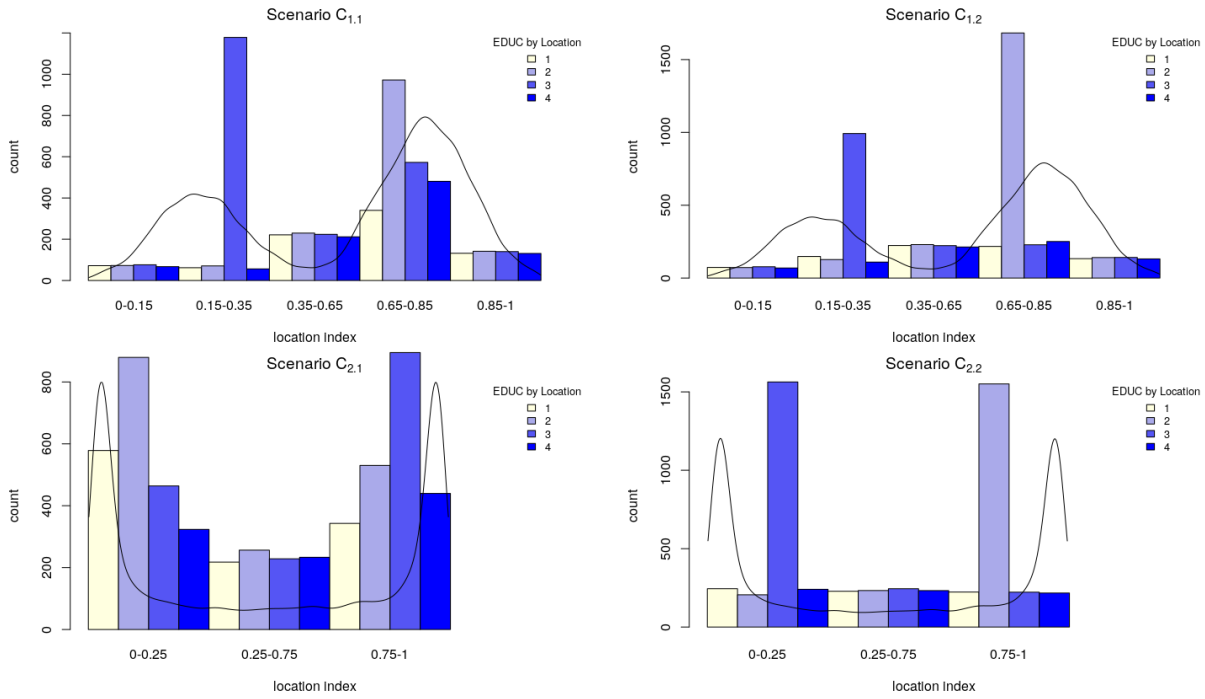

Figure A.1: Average marginal education distribution by location in Simulation study A; 100 simulations, $N_h = 4000$

Table A.1: Adjusted Education Distributions under the two theoretical scenarios

	Scenario $\mathcal{C}_{1.1}$: Gaussian Mixture Location Distribution		
Location	**[0.15,0.35]**	**[0.65,0.85]**	**else**
Education level	% Population	% Population	% Population
1 (< high school)	25	15	25
2 (high school)	20	40	25
3 (some college)	40	25	25
4 (≥ bachelors)	15	20	25
	Scenario $\mathcal{C}_{1.2}$: Gaussian Mixture Location Distribution		
Location	**[0.15,0.35]**	**[0.65,0.85]**	**else**
Education level	% Population	% Population	% Population
1 (< high school)	10	10	25
2 (high school)	10	70	25
3 (some college)	70	10	25
4 (≥ bachelors)	10	10	25
	Scenario $\mathcal{C}_{2.1}$: Beta Location Distribution		
Location	**[0,0.25]**	**[0.75,1]**	**else**
Education level	% Population	% Population	% Population
1 (< high school)	25	15	25
2 (high school)	20	40	25
3 (some college)	40	25	25
4 (≥ bachelors)	15	20	25
	Scenario $\mathcal{C}_{2.2}$: Beta Location Distribution		
Location	**[0,0.25]**	**[0.75,1]**	**else**
Education level	% Population	% Population	% Population
1 (< high school)	10	10	25
2 (high school)	10	70	25
3 (some college)	70	10	25
4 (≥ bachelors)	10	10	25
Total	**75.0**	**100.0**	**25.0**

Bibliography

[Bur20] U.S. Census Bureau. *Understanding and Using American Community Survey Data: What All Data Users Need to Know*. U.S. Government Publishing Office, Washington, DC, 2020.

[CS06] Eugene Choo and Aloysius Siow. "Who Marries Whom and Why". In: *Journal of Political Economy* 114.1 (2006), pp. 175–201.

[DBF01] John K. Dagsvik, Helge Brunborg, and Ane S. Flaatten. "A behavioral two-sex marriage model". In: *Mathematical Population Studies* 9.2 (2001), pp. 97–121.

[DG14] Arnaud Dupuy and Alfred Galichon. "Personality Traits and the Marriage Market". In: *Journal of Political Economy* 122.6 (2014), pp. 1271–1319.

[GS21] Alfred Galichon and Bernard Salanié. "Cupid's Invisible Hand: Social Surplus and Identification in Matching Models". In: *The Review of Economic Studies* (Dec. 2021).

[Goy+20] Shuchi Goyal et al. "A Practical Revealed Preference Model for Separating Preferences and Availability Effects in Marriage Formation". unpublished. 2020.

[LHN08] John Allen Logan, Peter D. Hoff, and Michael A. Newton. "Two-sided estimation of mate preferences for similarities in age, education, and religion". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 559–569.

[Men15] Konrad Menzel. "Large Matching Markets as Two-Sided Demand Systems". In: *Econometrica* 83.3 (2015), pp. 897–941.

[Pol86] Robert A. Pollak. "A reformulation of the two-sex problem". In: *Demography* 23.2 (1986), pp. 247–259.

[Pol97] J.H. Pollard. "Modelling the interaction between the sexes". In: *Mathematical and Computer Modelling* 26.6 (1997), pp. 11–24.

[RS90] Alvin E. Roth and Marilda A. Oliveira Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Econometric Society Monographs. Cambridge University Press, 1990.

[RV90] Alvin E. Roth and John H. Vande Vate. "Random Paths to Stability in Two-Sided Matching". In: *Econometrica* 58.6 (1990), pp. 1475–1480.

[Rug+22] Steven Ruggles et al. *IPUMS USA: Version 12.0 [dataset]*. Tech. rep. Minneapolis, MN: IPUMS, 2022. DOI: https://doi.org/10.18128/D010.V12.0.

[Yeu19] Fiona Yeung. "Statistical Revealed Preference Models for Bipartite Networks". PhD thesis. UCLA, 2019.